Chapter 3

Analysis of Cross-Sectional Data

Note: The primary reference text for these notes is Hayashi (2000). Other comprehensive treatments are available in Greene (2007) and Davidson and MacKinnon (2003).

Linear regression is the foundation of modern econometrics. While the importance of linear regression in financial econometrics has diminished in recent years, it is still widely employed. More importantly, the theory behind least-squares estimators is useful in broader contexts, and many results of this chapter are special cases of more general estimators presented in subsequent chapters. This chapter covers model specification, estimation, small- and large-sample inference, and model selection.

Linear regression is an essential tool of any econometrician and is widely used throughout finance and economics. Linear regression's success is owed to two key features: the availability of simple, closed-form estimators, and the ease and directness of interpretation. However, despite the regression estimator's superficial simplicity, the concepts presented in this chapter will reappear in the chapters on time series, panel data, Generalized Method of Moments (GMM), event studies, and volatility modeling.

3.1 Model Description

Linear regression expresses a dependent variable as a linear function of independent variables, possibly random, and an error.

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i, \tag{3.1}$$

where Y_i is known as the *regressand*, *dependent variable* or simply the *left-hand-side variable*. The k variables, $X_{1,i}, \ldots, X_{k,i}$ are known as the *regressors*, *independent variables* or *right-hand-side variables*. β_1, β_2, ..., β_k are the *regression coefficients*, ε_i is known as the *innovation*, *shock* or *error* and $i = 1, 2, \ldots, n$ index the observation. While this representation clarifies the relationship between Y_i and the Xs, matrix notation will generally be used to compactly describe models:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.2}$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{3.3}$$

where $\mathbf{X}$ is an n by k matrix, β is a k by 1 vector, and both $\mathbf{y}$ and ε are n by 1 vectors.

Two vector notations will occasionally be used: row,

$$\begin{bmatrix} Y_1 & = & \mathbf{X}_1\beta & +\varepsilon_1 \\ Y_2 & = & \mathbf{X}_2\beta & +\varepsilon_2 \\ \vdots & & \vdots & \vdots \\ Y_n & = & \mathbf{X}_n\beta & +\varepsilon_n \end{bmatrix} \tag{3.4}$$

and column,

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_k\mathbf{x}_k + \varepsilon. \tag{3.5}$$

Linear regression allows coefficients to be interpreted, *all things being equal*. Specifically, the effect of a change in one variable can be examined without changing the others. Regression analysis also allows for models that contain all of the information relevant for determining Y_i, whether these quantities are of primary interest or not. This feature provides the mechanism to interpret the coefficient on a regressor as the unique effect of that regressor (under certain conditions), a feature that makes linear regression very attractive.

3.1.1 What is a model?

What constitutes a model is a difficult question to answer. One view of a model is that of the *data generating process* (DGP). For instance, if a model postulates

$$Y_i = \beta_1 X_i + \varepsilon_i$$

then one interpretation is that the regressand, Y_i, is wholly determined by X_i and some random shock. The alternative view is that X_i is the only relevant variable available to the econometrician that explains variation in Y_i. Everything else that determines Y_i cannot be measured and, in the usual case, cannot be placed into a framework that would allow the researcher to formulate a model.

Consider monthly returns on the S&P 500, a value-weighted index of 500 large firms in the United States. Equity holdings and returns are generated by individuals based on their beliefs and preferences. If one were to take a (literal) data generating process view of the return on this index, data on individual investors' preferences and beliefs would need to be collected and formulated into a model for market returns. Collecting data and building this model would be a substantial challenge.

On the other hand, a model can be built to explain the variation in the market based on observable quantities (such as oil price changes or macroeconomic news announcements) without explicitly collecting information on beliefs and preferences. In a model of this type, explanatory variables can be viewed as inputs individuals consider when forming their beliefs and, subject to their preferences,

taking actions that ultimately affect the price of the S&P 500. The model allows the relationships between the regressand and regressors to be explored and is meaningful even though the model is not plausibly the data generating process.

In the context of time-series data, models often postulate that a series's past values are useful in predicting future values. Consider building a model of monthly returns on the S&P 500 using past returns to explain future returns. Treated as a DGP, this model implies that average returns in the future are determined by returns in the immediate past. Alternatively, if treated as an approximation, then one interpretation postulates that changes in risk aversion, beliefs, or other variables that influence holdings of assets change slowly (possibly in an unobservable manner). These slowly changing "factors" produce predictability in returns. Of course, there are other interpretations, but these should come from finance theory rather than data. The *model as a proxy* interpretation is additionally useful as it allows models to be specified, which are only loosely coupled with theory but that capture essential features of a theoretical model.

Careful consideration of what defines a model is a crucial step in the development of an econometrician, and one should always consider which assumptions and beliefs are needed to justify any specification.

3.1.2 Example: Cross-section regression of returns on factors

The concepts of linear regression will be explored in the context of a cross-section regression of returns on a set of factors thought to capture systematic risk. Cross-sectional regressions in financial econometrics date back at least to the Capital Asset Pricing Model (CAPM, Markowitz (1959), Sharpe (1964) and Lintner (1965)), a model formulated as a regression of individual asset's excess returns on the excess return of the market. More general specifications with multiple regressors are motivated by the Intertemporal CAPM (ICAPM, Merton (1973)) and Arbitrage Pricing Theory (APT, Ross (1976)).

The basic model postulates that *excess* returns are linearly related to a set of systematic risk factors. The factors can be returns on other assets, such as the market portfolio, or any other variable related to intertemporal hedging demands, such as interest rates, shocks to inflation, or consumption growth.

$$R_i - R_i^f = \mathbf{f}_i\beta + \varepsilon_i$$

or more compactly,

$$r_i^e = \mathbf{f}_i\beta + \varepsilon_i$$

where $R_i^e = R_i - R_i^f$ is the excess return on the asset and $\mathbf{f}_i = [F_{1,i}, \ldots, F_{k,i}]$ are returns on factors that explain systematic variation.

Linear factors models have been used in countless studies, the most well known by Fama and French (Fama and French (1992) and Fama and French (1993)) who use returns on specially constructed portfolios as factors to capture specific types of risk. The data set contains the variables listed in table 3.1.

Monthly data from July 1963 until January 2020 is used in the examples. Except for the interest rates, all return data are from the CRSP database. Returns are calculated as 100 times the logarithmic price difference ($R_i = 100(\ln(P_i) - \ln(P_{i-1}))$). Portfolios were constructed by sorting the firms into categories based on market capitalization, Book Equity to Market Equity (BE/ME), or past returns

Variable	Description
VWM	Returns on a value-weighted portfolio of all NYSE, AMEX and NASDAQ stocks
SMB	Returns on the Small minus Big factor, a zero investment portfolio that is long small market capitalization firms and short big caps.
HML	Returns on the High minus Low factor, a zero investment portfolio that is long high BE/ME firms and short low BE/ME firms.
MOM	Returns on a portfolio that is long winners and short losers as defined by their performance over the past 12 months, excluding the last month. Includes the large and small cap stocks but excludes mid-cap stocks.
SL	Returns on a portfolio of small cap and low BE/ME firms.
SM	Returns on a portfolio of small cap and medium BE/ME firms.
SH	Returns on a portfolio of small cap and high BE/ME firms.
BL	Returns on a portfolio of big cap and low BE/ME firms.
BM	Returns on a portfolio of big cap and medium BE/ME firms.
BH	Returns on a portfolio of big cap and high BE/ME firms.
RF	Risk free rate (Rate on a 3 month T-bill).
$DATE$	Date in format YYYYMM.

Table 3.1: Variable description for the data available in the Fama-French data-set used throughout this chapter.

over the previous year. For further details on the construction of portfolios, see Fama and French (1993) or Ken French's website:

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

A general model for the BH portfolio can be specified

$$BH_i - RF_i = \beta_1 + \beta_2(VWM_i - RF_i) + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$$

or, in terms of the excess returns,

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i.$$

The coefficients in the model can be interpreted as the effect of a change in one variable holding the other variables constant. For example, β_3 captures the effect of a change in the SMB_i risk factor holding VWM_i^e, HML_i and MOM_i constant. Table 3.2 contains some descriptive statistics of the factors and the six portfolios included in this data set.

3.2 Functional Form

A linear relationship is fairly specific and, in some cases, restrictive. It is important to distinguish specifications that can be examined in the linear regression framework from those that cannot. Linear

	Mean	Std. Dev.	Skewness	Kurtosis
VWM^e	6.66	15.42	-0.54	4.91
SMB	2.17	10.52	0.43	7.83
HML	3.06	9.95	0.01	5.41
MOM	7.95	14.52	-1.28	13.20
SL^e	6.54	23.55	-0.39	4.74
SM^e	10.21	18.93	-0.54	5.81
SH^e	11.23	19.69	-0.53	6.80
BL^e	6.78	15.94	-0.34	4.84
BM^e	6.47	14.87	-0.48	5.39
BH^e	8.22	17.20	-0.62	6.23

Table 3.2: Descriptive statistics of the six portfolios that will be used throughout this chapter. The data consist of monthly observations from January 1927 until June 2008 ($n = 978$).

regressions require two key features of any model: each term on the right-hand side must have only one coefficient that enters multiplicatively, and the error must enter additively.[1] Most specifications satisfying these two requirements can be treated using the tools of linear regression.[2] Other forms of "nonlinearities" are permissible. Any regressor or the regressand can be nonlinear transformations of the original observed data.

Double log (also known as log-log) specifications, where both the regressor and the regressands are log transformations of the original (positive) data, are frequently used.

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + \varepsilon_i.$$

In the parlance of linear regression, the model is specified

$$\tilde{Y}_i = \beta_1 + \beta_2 \tilde{X}_i + \varepsilon_i$$

where $\tilde{Y}_i = \ln(Y_i)$ and $\tilde{X}_i = \ln(X_i)$. The usefulness of the double log specification can be illustrated by a Cobb-Douglas production function subject to a multiplicative shock

$$Y_i = \beta_1 K_i^{\beta_2} L_i^{\beta_3} \varepsilon_i.$$

Using the production function directly, it is not obvious that, given values for output (Y_i), capital (K_i) and labor (L_i) of firm i, the model is consistent with a linear regression. However, taking logs,

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln K_i + \beta_3 \ln L_i + \ln \varepsilon_i$$

the model can be reformulated as a linear regression on the transformed data. Other forms, such as semi-log (either log-lin, where the regressand is logged but the regressors are unchanged, or lin-log, which logs only the regressor), are often useful to describe nonlinear relationships.

[1] A third but obvious requirement is that neither Y_i nor any of the $X_{j,i}$ may be latent (unobservable), $j = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n$.

[2] There are further requirements on the data, both the regressors and the regressand, to ensure that estimators of the unknown parameters are reasonable, but these are treated in subsequent sections.

Linear regression does, however, rule out specifications that may be of interest. Linear regression is not an appropriate framework to examine a model of the form $Y_i = \beta_1 X_{1,i}^{\beta_2} + \beta_3 X_{2,i}^{\beta_4} + \varepsilon_i$. Fortunately, more general frameworks, such as the generalized method of moments (GMM) or maximum likelihood estimation (MLE), topics of subsequent chapters, can be applied.

Two other transformations of the original data, dummy variables and interactions, are commonly used to generate nonlinear (in regressors) specifications. A *dummy variable* is a special class of regressor that takes the value 0 or 1. In finance, dummy variables (or dummies) are used to model calendar effects, leverage (where the magnitude of a coefficient depends on the sign of the regressor), or group-specific effects. Variable *interactions* parameterize nonlinearities into a model through products of regressors. Common interactions include powers of regressors ($X_{1,i}^2, X_{1,i}^3, \ldots$), cross-products of regressors ($X_{1,i}X_{2,i}$) and interactions between regressors and dummy variables. Variable transformations add significant flexibility to the linear regression models.

The use of nonlinear transformations also changes the interpretation of the regression coefficients. If only unmodified regressors are included,

$$Y_i = \mathbf{x}_i\beta + \varepsilon_i$$

then $\frac{\partial Y_i}{\partial X_{k,i}} = \beta_k$. Suppose a specification includes both X_i and X_i^2 as regressors,

$$Y_i = \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

In this specification, $\frac{\partial Y_i}{\partial X_i} = \beta_1 + \beta_2 X_i$ and the level of the variable enters its partial effect. Similarly, in a simple double log model

$$\ln Y_i = \beta_1 \ln X_i + \varepsilon_i,$$

and

$$\beta_1 = \frac{\partial \ln Y_i}{\partial \ln X_i} = \frac{\frac{\partial Y}{Y}}{\frac{\partial X}{X}} = \frac{\%\Delta Y}{\%\Delta X}$$

Thus, β_1 corresponds to the *elasticity* of Y_i with respect to X_i. In general, the coefficient on a variable that enters the model in in levels corresponds to the effect of a one-unit change in that variable. The coefficient on a variable that appears logged corresponds to the effect of a one percent change in that variable. For example, in a semi-log model where only the regressor is logged,

$$Y_i = \beta_1 \ln X_i + \varepsilon_i,$$

β_1 will correspond to a unit change in Y_i for a % change in X_i. Finally, in the case of discrete regressors, where there is no differential interpretation of coefficients, β represents the effect of a *whole* unit change, such as a dummy going from 0 to 1.

3.2.1 Example: Dummy variables and interactions in cross-section regressions

The January and the December effects are seasonal phenomena that have been widely studied in finance. Simply put, the December effect hypothesizes that returns in December are unusually low

due to tax-induced portfolio rebalancing, mostly to realized losses, while the January effect stipulates returns are abnormally high as investors return to the market. To model excess returns on a portfolio (BH_i^e) as a function of the excess market return (VWM_i^e), a constant, and the January and December effects, a model can be specified

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 I_{1i} + \beta_4 I_{12i} + \varepsilon_i$$

where $I_{1i} = 1$ if the return was generated in January and $I_{12i} = 1$ in December. The model can be reparameterized into three cases:

$$\begin{aligned} BH_i^e &= (\beta_1 + \beta_3) + \beta_2 VWM_i^e + \varepsilon_i && \text{January} \\ BH_i^e &= (\beta_1 + \beta_4) + \beta_2 VWM_i^e + \varepsilon_i && \text{December} \\ BH_i^e &= \beta_1 + \beta_2 VWM_i^e + \varepsilon_i && \text{Otherwise} \end{aligned}$$

Dummy interactions can be used to produce models that have both different intercepts and different slopes in January and December,

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 I_{1i} + \beta_4 I_{12i} + \beta_5 I_{1i} VWM_i^e + \beta_6 I_{12i} VWM_i^e + \varepsilon_i.$$

If excess returns on a portfolio were nonlinearly related to returns on the market, a simple model could be specified

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 (VWM_i^e)^2 + \beta_4 (VWM_i^e)^3 + \varepsilon_i.$$

Dittmar (2002) proposed a similar model to explain the cross-sectional dispersion of expected returns.

3.3 Estimation

Linear regression is also known as ordinary least squares (OLS) or simply least squares. The least-squares estimator minimizes the squared distance between the fit line (or plane if there are multiple regressors) and the regressand. The parameters are estimated as the solution to

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \min_{\beta} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\beta)^2. \tag{3.6}$$

First-order conditions of this optimization problem are

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = -2\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta\right) = -2\sum_{i=1}^{n} \mathbf{x}_i (Y_i - \mathbf{x}_i\beta) = \mathbf{0} \tag{3.7}$$

and rearranging, the least-squares estimator for β can be analytically derived.

Definition 3.1 (OLS Estimator)**.** The ordinary least-squares estimator, denoted $\hat{\beta}$, is defined

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.8}$$

This estimator is only reasonable if $\mathbf{X}'\mathbf{X}$ is invertible, which is equivalent to the condition that $\text{rank}(\mathbf{X}) = k$. This requirement states that no column of $\mathbf{X}$ can be exactly expressed as a combination of the $k-1$ remaining columns and that the number of observations is at least as large as the number of regressors ($n \geq k$). This is a weak condition and is trivial to verify in most econometric software packages: using a less than full rank matrix will generate a warning or error.

Dummy variables create one further issue worthy of special attention. Suppose dummy variables corresponding to the four quarters of the year, $I_{1i}, \ldots, I_{4i}$, are constructed from a quarterly data set of portfolio returns. Consider a simple model with a constant and all four dummies

$$R_i = \beta_1 + \beta_2 I_{1i} + \beta_3 I_{2i} + \beta_4 I_{3i} + \beta_5 I_{4i} + \varepsilon_i.$$

It is not possible to estimate this model with all four dummy variables *and* the constant because the constant is a perfect linear combination of the dummy variables, and so the regressor matrix would be rank deficient. The solution is to exclude either the constant or one of the dummy variables. The choice of variable to exclude makes no difference in estimation, and only the interpretation of the estimated coefficients changes. In the case where the constant is excluded, the coefficients on the dummy variables are directly interpretable as quarterly average returns. If one of the dummy variables is excluded, for example, the first quarter dummy variable, the interpretation changes. In this parameterization,

$$R_i = \beta_1 + \beta_2 I_{2i} + \beta_3 I_{3i} + \beta_4 I_{4i} + \varepsilon_i,$$

β_1 is the average return in Q1, while $\beta_1 + \beta_j$ is the average return in Qj.

It is also important that any regressor, other than the constant, be nonconstant. Suppose a regression that included the number of years since public floatation is fitted on a data set that contains only assets that have been trading for exactly 10 years. Including both this regressor and a constant results in perfect collinearity, but, more importantly, without variability in a regressor, it is impossible to determine whether changes in the regressor (years since float) results in a change in the regressand or whether the effect is simply constant across all assets. The role that that variability of regressors plays in estimating model parameters will be revisited when studying the statistical properties of $\hat{\beta}$.

The second derivative matrix of the minimization,

$$2\mathbf{X}'\mathbf{X},$$

ensures that the solution must be a minimum as long as $\mathbf{X}'\mathbf{X}$ is positive definite, which is equivalent to a condition that $\text{rank}(\mathbf{X}) = k$.

Once the regression coefficients have been estimated, it is useful to define the fit values, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and sample residuals $\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Rewriting the first-order condition in terms of the explanatory variables and the residuals provides insight into the numerical properties of the residuals. An equivalent first-order condition to eq. (3.7) is

$$\mathbf{X}'\hat{\varepsilon} = 0. \tag{3.9}$$

This set of linear equations is commonly referred to as the normal equations or orthogonality conditions. This set of conditions requires that $\hat{\varepsilon}$ is outside the span of the columns of $\mathbf{X}$. Moreover, considering the columns of $\mathbf{X}$ separately, $\mathbf{X}_j'\hat{\varepsilon} = 0$ for all $j = 1, 2, \ldots, k$. When a column contains a

constant (an intercept in the model specification), $\iota'\hat{\varepsilon} = 0$ ($\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$), and the mean of the residuals will be exactly 0.[3]

The OLS estimator of the residual variance, $\hat{\sigma}^2$, can be defined.[4]

Definition 3.2 (OLS Variance Estimator)**.** The OLS residual variance estimator, denoted $\hat{\sigma}^2$, is defined

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} \tag{3.10}$$

Definition 3.3 (Standard Error of the Regression)**.** The standard error of the regression is defined as

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{3.11}$$

The least-squares estimator has two final noteworthy properties. First, nonsingular transformations of X and non-zero scalar transformations of Y have deterministic effects on the estimated regression coefficients. Suppose $\mathbf{A}$ is a k by k nonsingular matrix, and c is a non-zero scalar. The coefficients of a regression of cY_i on $\mathbf{x}_i\mathbf{A}$ are

$$\begin{aligned}
\tilde{\beta} &= [(\mathbf{XA})'(\mathbf{XA})]^{-1}(\mathbf{XA})'(c\mathbf{y}) \\
&= c(\mathbf{A}'\mathbf{X}'\mathbf{XA})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= c\mathbf{A}^{-1}\hat{\beta}.
\end{aligned} \tag{3.12}$$

Second, as long as the model contains a constant, the regression coefficients on all terms except the intercept are unaffected by adding an arbitrary constant to either the regressor or the regressands. Consider transforming the standard specification,

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

to

$$\tilde{Y}_i = \beta_1 + \beta_2 \tilde{X}_{2,i} + \ldots + \beta_k \tilde{X}_{k,i} + \varepsilon_i$$

where $\tilde{Y}_i = Y_i + c_y$ and $\tilde{X}_{j,i} = X_{j,i} + c_{x_j}$. This model is identical to

$$Y_i = \tilde{\beta}_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

where $\tilde{\beta}_1 = \beta_1 + c_y - \beta_2 c_{x_2} - \ldots - \beta_k c_{x_k}$.

[3] ι is an n by 1 vector of 1s.

[4] The choice of $n-k$ in the denominator will be made clear once the properties of this estimator have been examined.

	Constant	VWM^e	SMB	HML	MOM	$\hat{\sigma}$
SL^e	-0.15	1.09	1.02	-0.26	-0.03	0.99
SM^e	0.08	0.96	0.82	0.35	-0.00	0.77
SH^e	0.05	1.00	0.87	0.69	-0.00	0.56
BL^e	0.12	0.99	-0.15	-0.28	-0.00	0.69
BM^e	-0.05	0.98	-0.13	0.31	-0.00	1.15
BH^e	-0.09	1.08	0.00	0.76	-0.04	1.06

Table 3.3: Estimated regression coefficients from the model $R_i^{p_i} = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$, where $R_i^{p_i}$ is the excess return on one of the six size and value sorted portfolios. The final column contains the standard error of the regression.

3.3.1 Estimation of Cross-Section regressions of returns on factors

Table 3.3 contains the estimated regression coefficients as well as the standard error of the regression for the six portfolios in the Fama-French data set in a specification that includes all four factors and a constant. There has been a substantial decrease in the magnitude of the standard error of the regression relative to the standard deviation of the original data. The next section will formalize how this reduction is interpreted.

3.4 Assessing Fit

Once the parameters have been estimated, the next step is to determine whether the model fits the data. The minimized sum of squared errors, the optimization's objective, is an obvious choice to assess fit. However, there is an important drawback to using the sum of squared errors: changes in the scale of Y_i alter the minimized sum of squared errors without changing the fit. It is necessary to distinguish between the portions of $\mathbf{y}$ explained by $\mathbf{X}$ from those that are not to construct a scale-free metric.

The projection matrix, $\mathbf{P_X}$, and the annihilator matrix, $\mathbf{M_X}$, are useful when decomposing the regressand into the explained component and the residual.

Definition 3.4 (Projection Matrix)**.** The projection matrix, a symmetric idempotent matrix that produces the projection of a variable onto the space spanned by $\mathbf{X}$, denoted $\mathbf{P_X}$, is defined

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{3.13}$$

Definition 3.5 (Annihilator Matrix)**.** The annihilator matrix, a symmetric idempotent matrix that produces the projection of a variable onto the null space of $\mathbf{X}'$, denoted $\mathbf{M_X}$, is defined

$$\mathbf{M_X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{3.14}$$

These two matrices have some desirable properties. Both the fited value of $\mathbf{y}$ ($\hat{\mathbf{y}}$) and the estimated errors, $\hat{\varepsilon}$, can be expressed in terms of these matrices as $\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y}$ and $\hat{\varepsilon} = \mathbf{M_X}\mathbf{y}$, respectively. These

matrices are also idempotent: $\mathbf{P_X P_X} = \mathbf{P_X}$ and $\mathbf{M_X M_X} = \mathbf{M_X}$ and orthogonal: $\mathbf{P_X M_X} = \mathbf{0}$. The projection matrix returns the portion of $\mathbf{y}$ that lies in the linear space spanned by $\mathbf{X}$, while the annihilator matrix returns the portion of $\mathbf{y}$ in the null space of $\mathbf{X}$. In essence, $\mathbf{M_X}$ annihilates any portion of $\mathbf{y}$ explainable by $\mathbf{X}$, leaving only the residuals.

Decomposing $\mathbf{y}$ using the projection and annihilator matrices,

$$\mathbf{y} = \mathbf{P_X y} + \mathbf{M_X y}$$

which follows since $\mathbf{P_X} + \mathbf{M_X} = \mathbf{I}_n$. The squared observations can be decomposed

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= (\mathbf{P_X y} + \mathbf{M_X y})'(\mathbf{P_X y} + \mathbf{M_X y}) \\
&= \mathbf{y}'\mathbf{P_X P_X y} + \mathbf{y}'\mathbf{P_X M_X y} + \mathbf{y}'\mathbf{M_X P_X y} + \mathbf{y}'\mathbf{M_X M_X y} \\
&= \mathbf{y}'\mathbf{P_X y} + 0 + 0 + \mathbf{y}'\mathbf{M_X y} \\
&= \mathbf{y}'\mathbf{P_X y} + \mathbf{y}'\mathbf{M_X y}
\end{aligned}$$

noting that $\mathbf{P_X}$ and $\mathbf{M_X}$ are idempotent and $\mathbf{P_X M_X} = \mathbf{0}_n$. These three quantities are often referred to as[5]

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} Y_i^2 \qquad \text{Uncentered Total Sum of Squares } (\text{TSS}_U) \tag{3.15}$$

$$\mathbf{y}'\mathbf{P_X y} = \sum_{i=1}^{n} (\mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \qquad \text{Uncentered Regression Sum of Squares } (\text{RSS}_\text{U}) \tag{3.16}$$

$$\mathbf{y}'\mathbf{M_X y} = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \qquad \text{Uncentered Sum of Squared Errors } (\text{SSE}_\text{U}). \tag{3.17}$$

Dividing through by $\mathbf{y}'\mathbf{y}$

$$\frac{\mathbf{y}'\mathbf{P_X y}}{\mathbf{y}'\mathbf{y}} + \frac{\mathbf{y}'\mathbf{M_X y}}{\mathbf{y}'\mathbf{y}} = 1$$

or

$$\frac{\text{RSS}_\text{U}}{\text{TSS}_U} + \frac{\text{SSE}_\text{U}}{\text{TSS}_U} = 1.$$

This identity expresses the scale-free total variation in $\mathbf{y}$ that is captured by $\mathbf{X}$ ($\mathbf{y}'\mathbf{P_X y}$) and that which is not ($\mathbf{y}'\mathbf{M_X y}$). The portion of the total variation explained by $\mathbf{X}$ is known as the uncentered R^2 (R^2_U),

[5]There is no consensus about the names of these quantities. In some texts, the component capturing the fit portion is known as the Regression Sum of Squares (RSS) while in others, it is known as the Explained Sum of Squares (ESS), while the portion attributable to the errors is known as the Sum of Squared Errors (SSE), the Sum of Squared Residuals (SSR), the Residual Sum of Squares (RSS) or the Error Sum of Squares (ESS). The choice to use SSE and RSS in this text was to ensure the reader that SSE must be the component of the squared observations relating to the error variation.

Definition 3.6 (Uncentered R^2($\mathrm{R}^2_{\mathrm{U}}$))**.** The uncentered R^2, which is used in models that do not include an intercept, is defined

$$\mathrm{R}^2_{\mathrm{U}} = \frac{\mathrm{RSS}_{\mathrm{U}}}{\mathrm{TSS}_U} = 1 - \frac{\mathrm{SSE}_{\mathrm{U}}}{\mathrm{TSS}_U} \tag{3.18}$$

While $\mathrm{R}^2_{\mathrm{U}}$ is scale-free, it suffers from one shortcoming. Suppose a constant is added to $\mathbf{y}$ so that the TSS_U changes to $(\mathbf{y}+c)'(\mathbf{y}+c)$. The identity still holds, and so $(\mathbf{y}+c)'(\mathbf{y}+c)$ must increase (for a sufficiently large c). In turn, one of the right-hand side variables must also grow larger. In the usual case where the model contains a constant, the increase will occur in the $\mathrm{RSS}_{\mathrm{U}}$ ($\mathbf{y}'\mathbf{P}_{\mathbf{X}}\mathbf{y}$), and as c becomes arbitrarily large, uncentered R^2 will asymptote to one. A centered measure *computed using deviations from the mean* rather than on levels overcomes this limitation.

Let $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y} = \mathbf{M}_\iota \mathbf{y}$ where $\mathbf{M}_\iota = \mathbf{I}_n - \iota(\iota'\iota)^{-1}\iota'$ is matrix which subtracts the mean from a vector of data. Then

$$\begin{aligned}
\mathbf{y}'\mathbf{M}_\iota\mathbf{P}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y} + \mathbf{y}'\mathbf{M}_\iota\mathbf{M}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y} &= \mathbf{y}'\mathbf{M}_\iota\mathbf{y} \\
\frac{\mathbf{y}'\mathbf{M}_\iota\mathbf{P}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} + \frac{\mathbf{y}'\mathbf{M}_\iota\mathbf{M}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} &= 1
\end{aligned}$$

or more compactly

$$\frac{\tilde{\mathbf{y}}'\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} + \frac{\tilde{\mathbf{y}}'\mathbf{M}_{\mathbf{X}}\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} = 1.$$

Centered R^2 ($\mathrm{R}^2_{\mathrm{C}}$) is defined analogously to uncentered replacing the uncentered sums of squares with their centered counterparts.

Definition 3.7 (Centered R^2($\mathrm{R}^2_{\mathrm{C}}$))**.** The uncentered R^2, used in models that include an intercept, is defined

$$\mathrm{R}^2_{\mathrm{C}} = \frac{\mathrm{RSS}_{\mathrm{C}}}{\mathrm{TSS}_{\mathrm{C}}} = 1 - \frac{\mathrm{SSE}_{\mathrm{C}}}{\mathrm{TSS}_{\mathrm{C}}} \tag{3.19}$$

where

$$\mathbf{y}'\mathbf{M}_\iota\mathbf{y} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \text{Centered Total Sum of Squares } (\mathrm{TSS}_{\mathrm{C}}) \tag{3.20}$$

$$\mathbf{y}'\mathbf{M}_\iota\mathbf{P}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y} = \sum_{i=1}^{n}(\mathbf{x}_i\hat{\beta} - \bar{\mathbf{x}}\hat{\beta})^2 \qquad \text{Centered Regression Sum of Squares } (\mathrm{RSS}_{\mathrm{C}}) \tag{3.21}$$

$$\mathbf{y}'\mathbf{M}_\iota\mathbf{M}_{\mathbf{X}}\mathbf{M}_\iota\mathbf{y} = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i\hat{\beta})^2 \qquad \text{Centered Sum of Squared Errors } (\mathrm{SSE}_{\mathrm{C}}). \tag{3.22}$$

and where $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$.

The expressions R^2, SSE, RSS, and TSS should be assumed to correspond to the centered version unless further qualified. With two versions of R^2 available that generally differ, which should be used? Centered should be used if the model is centered (contains a constant), and uncentered should

be used when it does not. Failing to select the correct R^2 can lead to incorrect conclusions about the model's fit, and mixing the definitions can lead to a nonsensical R^2 that falls outside of $[0,1]$. For instance, computing R^2 using the centered version when the model does not contain a constant often results in a negative value when

$$R^2 = 1 - \frac{SSE_C}{TSS_C}.$$

Most software will return centered R^2, and caution is warranted if a model is fit without a constant.

R^2 does have some caveats. First, adding an additional regressor will always (weakly) increase the R^2 since the sum of squared errors cannot increase by the inclusion of an additional regressor. This renders R^2 useless in discriminating between two models where one is nested within the other. One solution to this problem is to use the degree of freedom adjusted R^2.

Definition 3.8 (Adjusted R^2 $\left(\bar{R}^2\right)$)**.** The adjusted R^2, which adjusts for the number of estimated parameters, is defined

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{SSE}{TSS}\frac{n-1}{n-k}. \tag{3.23}$$

$\bar{R}^2$ will increase if the reduction in the SSE is large enough to compensate for a loss of one degree of freedom, captured by the $n-k$ term. However, if the SSE does not change, $\bar{R}^2$ will decrease. $\bar{R}^2$ is preferable to R^2 for comparing models, although the topic of model selection will be more formally considered at the end of this chapter. $\bar{R}^2$, like R^2, should be constructed from the appropriate versions of the RSS, SSE, and TSS (either centered or uncentered).

Second, R^2 is not invariant to changes in the regressand. A frequent mistake is to use R^2 to compare the fit from two models with different regressands, for instance, Y_i and $\ln(Y_i)$. These numbers are incomparable, and this type of comparison must be avoided. Moreover, R^2 is even sensitive to more benign transformations. Suppose a simple model is postulated,

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$$

and a model logically consistent with the original model,

$$Y_i - X_i = \beta_1 + (\beta_2 - 1)X_i + \varepsilon_i,$$

is estimated. The R^2s from these models will generally differ. For example, suppose the original coefficient on x_i was 1. Subtracting x_i will reduce the explanatory power of x_i to 0, rendering it useless and resulting in a R^2 of 0 irrespective of the R^2 in the original model.

3.4.1 Example: R^2 and $\bar{R}^2$ in Cross-Sectional Factor models

To illustrate the use of R^2, consider alternative models of BH^e that include one or more risk factors. The R^2 values in the top half of Table 3.4 show that R^2 never declines as additional variables are added. Note that the adjusted measure of fit, $\bar{R}^2_U$, also never declines, although it grows more slowly. The monotonic pattern occurs since the adjustment penalty is small when the sample size n is large, as is the case here. The table only shows the correct version of the R^2 – centered for models that contain a constant and uncentered for those that do not.

Regressand	Regressors	R^2_U	$\bar{R}^2_U$	R^2_C	$\bar{R}^2_C$
BH^e	$1, VME^e$	0.7620	0.7616	–	–
BH^e	$1, VME^e, SMB$	0.7644	0.7637	–	–
BH^e	$1, VME^e, SMB, HML$	0.9535	0.9533	–	–
BH^e	$1, VME^e, SMB, HML, MOM$	0.9543	0.9541	–	–
BH^e	VWM^e	–	–	0.7656	0.7653
$10+BH^e$	$1, VME^e$	0.7620	0.7616	–	–
$10+BH^e$	VME^e	–	–	0.2275	0.2264
$10\times BH^e$	$1, VME^e$	0.7620	0.7616	–	–
$10\times BH^e$	VME^e	–	–	0.7656	0.7653
$BH^e - VME^e$	$1, VME^e$	0.0024	0.0009	–	–
$\sum_Y BH^e$	$1, \sum_Y VME^e$	0.6800	0.6743	–	–

Table 3.4: Centered and uncentered R^2 and $\bar{R}^2$ from models with regressor or regressand changes. Only the correct version of the R^2 is shown – centered for models that contain a constant as indicated by 1 in the regressor list, or uncentered for models that do not. The top rows demonstrate how R^2 and its adjusted version change as additional variables are added. The bottom two rows demonstrate how changes in the regressand – the left-hand-side variable – affect the R^2.

The bottom half of the table shows how R^2 changes when the regressand changes. The R^2 in models that include a constant are invariant to constant shifts in the regressand. The R^2_U of the model that regresses $10+BH^e$ on a constant and the excess market is identical to the same model only using BH^e. This relationship does not hold for models that do not contain a constant and R^2_C changes when 10 is added to the return. Both measures are invariant to multiplicative adjustments. The penultimate line shows that R^2 is *not* invariant to changes in the regressand that do not fundamentally alter the interpretation of the model. In this model, the difference in returns, $BH^e - VMW^e$, is regressed on a constant and the excess market. The coefficient on the excess market, $\hat{\gamma}_2$, in this model

$$BH^e_i - VWM^e = \gamma_1 + \gamma_2 VWM^e_i + \varepsilon_i.$$

will be *exactly* 1 less than the coefficient in the model

$$BH^e_i = \beta_1 + \beta_2 VWM^e_i + \varepsilon_i.$$

While these two models are conceptually identical and describe the same relationship between BH^e, the R^2 changes. In this example, the coefficient on VWM^ein near zero since the coefficient in the original specification is near 1. The R^2 of the return difference is near 0 even though the market is an important determinant of the the Big-High portfolio's return. The final line shows the regression coefficient of the annual return of BH^e ($\sum_Y BH^e$) on the annual return on the market ($\sum_Y VWM^e$). This type of aggregation also changes the R^2. These final two results highlight a common form of misuse of R^2: do not compare the values of R^2 in models with different regressands.

3.5 Assumptions

Thus far, all of the derivations and identities presented are purely numerical. They do not indicate whether $\hat{\beta}$ is a reasonable way to estimate β. It is necessary to make some assumptions about the innovations and the regressors to provide a statistical interpretation of $\hat{\beta}$. Two broad classes of assumptions can be used to analyze the behavior of $\hat{\beta}$: the classical framework (also known as the small-sample or finite-sample framework) and asymptotic analysis (also known as the large-sample framework).

Neither method is ideal. The small-sample framework is precise in that the exact distribution of regressors and test statistics are known. This precision comes at the cost of many restrictive assumptions – assumptions not usually plausible in financial applications. On the other hand, asymptotic analysis requires few restrictive assumptions and is broadly applicable to financial data, although the results are only exact if the number of observations is infinite. Asymptotic analysis is still useful for examining the behavior in finite samples when the sample size is large enough for the asymptotic distribution to approximate the finite-sample distribution reasonably well.

This leads to the most important question of asymptotic analysis: How large does n need to be before the approximation is reasonable? Unfortunately, the answer to this question is "*it depends*". In simple cases, where residuals are independent and identically distributed, as few as 30 observations may be sufficient for the asymptotic distribution to be a good approximation to the finite-sample distribution. In more complex cases, anywhere from 100 to 1,000 may be needed, while in the extreme cases, where the data is heterogenous and highly dependent, an asymptotic approximation may be poor with more than 1,000,000 observations.

The properties of $\hat{\beta}$ will be examined under both sets of assumptions. While the small-sample results are not generally applicable, it is important to understand these results as the *lingua franca* of econometrics, as well as the limitations of tests based on the classical assumptions, and to be able to detect when a test statistic may not have the intended asymptotic distribution. Six assumptions are required to examine the finite-sample distribution of $\hat{\beta}$ and establish the optimality of the OLS procedure(although many properties only require a subset).

Assumption 3.1 (Linearity)**.** $Y_i = \mathbf{x}_i\beta + \varepsilon_i$

This assumption states the obvious condition necessary for least squares to be a reasonable method to estimate the β. It further imposes a less obvious condition, that $\mathbf{x}_i$ must be observed and measured without error. Many applications in financial econometrics include *latent* variables. Linear regression is not applicable in these cases and a more sophisticated estimator is required. In other applications, the *true* value of $x_{k,i}$ is not observed and a noisy proxy must be used, so that $\tilde{x}_{k,i} = x_{k,i} + \nu_{k,i}$ where $\nu_{k,i}$ is an error uncorrelated with $x_{k,i}$. When this occurs, ordinary least-squares estimators are misleading and a modified procedure (two-stage least squares (2SLS) or instrumental variable regression (IV)) must be used.

Assumption 3.2 (Conditional Mean)**.** $\mathrm{E}[\varepsilon_i|\mathbf{X}] = 0, \quad i = 1,2,\ldots,n$

This assumption states that the mean of each ε_i is zero given any $X_{k,i}$, any function of any $X_{k,i}$ or combinations of these. It is stronger than the assumption used in the asymptotic analysis and is not valid in many applications (e.g., time-series data). When the regressand and regressor consist of time-series data, this assumption may be violated and $\mathrm{E}[\varepsilon_i|\mathbf{x}_{i+j}] \neq 0$ for some j. This assumption also implies that the correct form of $X_{k,i}$ enters the regression, that $\mathrm{E}[\varepsilon_i] = 0$ (through a simple application

of the law of iterated expectations), and that the innovations are uncorrelated with the regressors, so that $\mathrm{E}[\varepsilon_{i'} x_{j,i}] = 0, i' = 1,2,\ldots,n,\ i = 1,2,\ldots,n,\ j = 1,2,\ldots,k$.

Assumption 3.3 (Rank)**.** *The rank of* $\mathbf{X}$ *is* k *with probability 1.*

This assumption is needed to ensure that $\hat{\beta}$ is identified and can be estimated. In practice, it requires that the no regressor is perfectly co-linear with the others, that the number of observations is at least as large as the number of regressors ($n \geq k$) and that variables other than a constant have non-zero variance.

Assumption 3.4 (Conditional Homoskedasticity)**.** $\mathrm{V}[\varepsilon_i|\mathbf{X}] = \sigma^2$

Homoskedasticity is rooted in *homo* (same) and *skedannumi* (scattering) and in modern English means that the residuals have identical variances. This assumption is required to establish the optimality of the OLS estimator and it specifically rules out the case where the variance of an innovation is a function of a regressor.

Assumption 3.5 (Conditional Correlation)**.** $\mathrm{E}[\varepsilon_i \varepsilon_j|\mathbf{X}] = 0,\ i = 1,2,\ldots,n,\ j = i+1,\ldots,n$

Assuming the residuals are conditionally uncorrelated is convenient when coupled with the homoskedasticity assumption, and the residuals covariance is $\sigma^2\mathbf{I}_n$. Like homoskedasticity, this assumption is needed for establishing the optimality of the least-squares estimator.

Assumption 3.6 (Conditional Normality)**.** $\varepsilon|\mathbf{X} \sim N(\mathbf{0},\Sigma)$

Assuming a specific distribution is very restrictive – results based on this assumption will only be correct is the errors are actually normal – but this assumption allows for precise statements about the finite-sample distribution of $\hat{\beta}$ and test statistics. This assumption, when combined with assumptions 3.4 and 3.5, provides a simple distribution for the innovations: $\varepsilon_i|\mathbf{X} \xrightarrow{d} N(0,\sigma^2)$.

3.6 Small-Sample Properties of OLS estimators

Using these assumptions, many useful properties of $\hat{\beta}$ can be derived. Recall that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Theorem 3.1 (Bias of $\hat{\beta}$)**.** *Under assumptions 3.1 - 3.3*

$$\mathrm{E}[\hat{\beta}|\mathbf{X}] = \beta. \tag{3.24}$$

While unbiasedness is a desirable property, it is not particularly meaningful without further qualification. For instance, an estimator which is unbiased, but does not increase in precision as the sample size increases is generally not desirable. Fortunately, $\hat{\beta}$ is not only unbiased, it has a variance that goes to zero.

Theorem 3.2 (Variance of $\hat{\beta}$)**.** *Under assumptions 3.1 - 3.5*

$$\mathrm{V}[\hat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.25}$$

Under the conditions necessary for unbiasedness for $\hat{\beta}$, plus assumptions about homoskedasticity and the conditional correlation of the residuals, the form of the variance is simple. Consistency follows since

$$(\mathbf{X}'\mathbf{X})^{-1} = \left(n\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n}\right)^{-1} \approx \frac{1}{n}\mathrm{E}\left[\mathbf{x}_i'\mathbf{x}_i\right]^{-1} \tag{3.26}$$

will be declining as the sample size increases.

However, $\hat{\beta}$ has an even stronger property under the same assumptions. It is BLUE: *B*est *L*inear *U*nbiased *E*stimator. Best, in this context, means that it has the lowest variance among all other linear unbiased estimators. While this is a strong result, a few words of caution are needed to properly interpret this result. The class of Linear Unbiased Estimators (LUEs) is small in the universe of all unbiased estimators. Saying OLS is the "best" is akin to a one-armed boxer claiming to be the best one-arm boxer. While possibly true, she probably would not stand a chance against a two-armed opponent.

Theorem 3.3 (Gauss-Markov Theorem)**.** *Under assumptions 3.1-3.5, $\hat{\beta}$ is the minimum variance estimator among all linear unbiased estimators. That is* $\mathrm{V}[\tilde{\beta}|\mathbf{X}]$ - $\mathrm{V}[\hat{\beta}|\mathbf{X}]$ *is positive semi-definite where* $\tilde{\beta} = \mathbf{C}\mathbf{y}$, $\mathrm{E}[\tilde{\beta}] = \beta$ *and* $\mathbf{C} \neq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Letting $\tilde{\beta}$ be any other *linear*, *unbiased* estimator of β, it must have a larger covariance. However, many estimators, including most maximum likelihood estimators, are nonlinear and so are not necessarily less efficient. Finally, making use of the normality assumption, it is possible to determine the conditional distribution of $\hat{\beta}$.

Theorem 3.4 (Distribution of $\hat{\beta}$)**.** *Under assumptions 3.1 – 3.6,*

$$\hat{\beta}|\mathbf{X} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{3.27}$$

Theorem 3.4 should not be surprising. $\hat{\beta}$ is a linear combination of (jointly) normally distributed random variables and thus is also normally distributed. Normality is also useful for establishing the relationship between the estimated residuals $\hat{\varepsilon}$ and the estimated parameters $\hat{\beta}$.

Theorem 3.5 (Conditional Independence of $\hat{\varepsilon}$ and $\hat{\beta}$)**.** *Under assumptions 3.1 - 3.6, $\hat{\varepsilon}$ is independent of $\hat{\beta}$, conditional on* $\mathbf{X}$.

One implication of this theorem is that $\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\beta}_j|\mathbf{X}) = 0$ $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$. As a result, functions of $\hat{\varepsilon}$ will be independent of functions of $\hat{\beta}$, a property useful in deriving distributions of test statistics that depend on both. Finally, in the small-sample setup, the exact distribution of the sample error variance estimator, $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-k)$, can be derived.

Theorem 3.6 (Distribution of $\hat{\sigma}^2$)**.**

$$(n-k)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$$

where $\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{y}}{n-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$.

Since $\hat{\varepsilon}_i$ is a normal random variable, once it is standardized and squared, it should be a χ^2_1. The change in the divisor from n to $n-k$ reflects the loss in degrees of freedom due to the k estimated parameters.

3.7 Maximum Likelihood

Once the assumption that the innovations are conditionally normal has been made, conditional maximum likelihood is an obvious method to estimate the unknown parameters (β, σ^2). Conditioning on $\mathbf{X}$, and assuming the innovations are normal, homoskedastic, and conditionally uncorrelated, the likelihood is given by

$$f(\mathbf{y}|\mathbf{X};\beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma^2}\right) \tag{3.28}$$

and, taking logs, the log likelihood

$$l(\beta,\sigma^2;\mathbf{y}|\mathbf{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma^2}. \tag{3.29}$$

Recall that the logarithm is a monotonic, strictly increasing transformation, and the extremum points of the log-likelihood and the likelihood will occur at the same parameters. Maximizing the likelihood with respect to the unknown parameters, there are $k+1$ first-order conditions

$$\frac{\partial l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial \beta} = \frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\hat{\beta})}{\sigma^2} = 0 \tag{3.30}$$

$$\frac{\partial l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})}{2\hat{\sigma}^4} = 0. \tag{3.31}$$

The first set of conditions is identical to the first-order conditions of the least-squares estimator ignoring the scaling by σ^2, assumed to be greater than 0. The solution is

$$\hat{\beta}^{\text{MLE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.32}$$

$$\hat{\sigma}^{2\,\text{MLE}} = n^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta}) = n^{-1}\hat{\varepsilon}'\hat{\varepsilon}. \tag{3.33}$$

The regression coefficients are identical under maximum likelihood and OLS, although the divisor in $\hat{\sigma}^2$ and $\hat{\sigma}^{2\,\text{MLE}}$ differ.

It is important to note that the derivation of the OLS estimator does not require an assumption of normality. Moreover, the unbiasedness, variance, and BLUE properties do not rely on the conditional normality of residuals. However, if the innovations are homoskedastic, uncorrelated and normal, the results of the Gauss-Markov theorem can be strengthened using the Cramer-Rao lower bound.

Theorem 3.7 (Cramer-Rao Inequality)**.** *Let $f(\mathbf{z};\theta)$ be the joint density of $\mathbf{z}$ where θ is a k dimensional parameter vector Let $\hat{\theta}$ be an unbiased estimator of θ_0 with finite covariance. Under some regularity condition on $f(\cdot)$*

$$\text{V}[\hat{\theta}] \geq \mathcal{I}^{-1}(\theta_0)$$

where

$$\mathcal{I} = -\text{E}\left[\left.\frac{\partial^2 \ln f(\mathbf{z};\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right] \tag{3.34}$$

and

$$\mathcal{J} = \mathrm{E}\left[\frac{\partial \ln f(\mathbf{z};\theta)}{\partial \theta}\frac{\partial \ln f(\mathbf{z};\theta)}{\partial \theta'}\bigg|_{\theta=\theta_0}\right] \tag{3.35}$$

and, under some additional regularity conditions,

$$\mathcal{I}(\theta_0) = \mathcal{J}(\theta_0).$$

The last part of this theorem is the information matrix equality (IME) and when a model is correctly specified in its entirety, the expected covariance of the scores is equal to negative of the expected hessian.[6] The IME will be revisited in later chapters. The second order conditions,

$$\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial\beta\partial\beta'} = -\frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}^2} \tag{3.36}$$

$$\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial\beta\partial\sigma^2} = -\frac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\beta)}{\sigma^4} \tag{3.37}$$

$$\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial^2\sigma^2} = \frac{n}{2\sigma^4} - \frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{\sigma^6} \tag{3.38}$$

are needed to find the lower bound for the covariance of the estimators of β and σ^2. Taking expectations of the second derivatives,

$$\mathrm{E}\left[\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial\beta\partial\beta'}\right] = -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \tag{3.39}$$

$$\mathrm{E}\left[\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial\beta\partial\sigma^2}\right] = \mathbf{0} \tag{3.40}$$

$$\mathrm{E}\left[\frac{\partial^2 l(\beta,\sigma^2;\mathbf{y}|\mathbf{X})}{\partial^2\sigma^2}\right] = -\frac{n}{2\sigma^4} \tag{3.41}$$

and so the lower bound for the variance of $\hat{\beta} = \hat{\beta}^{\mathrm{MLE}}$ is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Theorem 3.2 show that $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is also the variance of the OLS estimator $\hat{\beta}$ and so the Gauss-Markov theorem can be strengthened in the case of conditionally homoskedastic, uncorrelated normal residuals.

Theorem 3.8 (Best Unbiased Estimator)**.** *Under assumptions 3.1 - 3.6, $\hat{\beta} = \hat{\beta}^{\mathrm{MLE}}$ is the best unbiased estimator of β.*

The difference between this theorem and the Gauss-Markov theorem is subtle but important. The class of estimators is no longer restricted to include only linear estimators and so this result is both broad and powerful: MLE (or OLS) is an ideal estimator under these assumptions (in the sense that no other unbiased estimator, linear or not, has a lower variance). This results does not extend to the variance estimator since $\mathrm{E}[\hat{\sigma}^{2\,\mathrm{MLE}}] = \frac{n}{n-k}\sigma^2 \neq \sigma^2$, and so the optimality of $\hat{\sigma}^{2\,\mathrm{MLE}}$ cannot be established using the Cramer-Rao theorem.

[6] There are quite a few regularity conditions for the IME to hold, but discussion of these is beyond the scope of this course. Interested readers should see White (1996) for a thorough discussion.

3.8 Small-Sample Hypothesis Testing

Most regressions are estimated to test implications of economic or finance theory. Hypothesis testing is the mechanism used to determine whether data and theory are congruent. Formalized in terms of β, the null hypothesis (also known as the maintained hypothesis) is formulated as

$$H_0 : \mathbf{R}(\beta) - \mathbf{r} = \mathbf{0} \tag{3.42}$$

where $\mathbf{R}(\cdot)$ is a function from $\mathbb{R}^k$ to $\mathbb{R}^m$, $m \leq k$ and $\mathbf{r}$ is an m by 1 vector. Initially, a subset of all hypotheses, those in the linear equality hypotheses class, formulated

$$H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0} \tag{3.43}$$

will be examined where $\mathbf{R}$ is a m by k matrix. In subsequent chapters, more general test specifications including nonlinear restrictions on the parameters will be considered. All hypotheses in this class can be written as weighted sums of the regression coefficients,

$$\begin{aligned} R_{11}\beta_1 + R_{12}\beta_2 \ldots + R_{1k}\beta_k &= r_1 \\ R_{21}\beta_1 + R_{22}\beta_2 \ldots + R_{2k}\beta_k &= r_2 \\ &\vdots \\ R_{m1}\beta_1 + R_{m2}\beta_2 \ldots + R_{mk}\beta_k &= r_i \end{aligned}$$

Each constraint is represented as a row in the above set of equations. Linear equality constraints can be used to test parameter restrictions such as

$$\begin{aligned} \beta_1 &= 0 \\ 3\beta_2 + \beta_3 &= 1 \\ \sum_{j=1}^{k} \beta_j &= 0 \\ \beta_1 = \beta_2 = \beta_3 &= 0. \end{aligned} \tag{3.44}$$

For instance, if the unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \varepsilon_i$$

the hypotheses in eq. (3.44) can be described in terms of $\mathbf{R}$ and $\mathbf{r}$ as

H_0	$\mathbf{R}$	$\mathbf{r}$
$\beta_1 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$	0
$3\beta_2 + \beta_3 = 1$	$\begin{bmatrix} 0 & 3 & 1 & 0 & 0 \end{bmatrix}$	1
$\sum_{j=1}^{k} \beta_j = 0$	$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \end{bmatrix}$	0
$\beta_1 = \beta_2 = \beta_3 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

When using linear equality constraints, alternatives are specified as $H_1 : \mathbf{R}\beta - \mathbf{r} \neq 0$. Once both the null and the alternative hypotheses have been postulated, it is necessary to discern whether the data are consistent with the null hypothesis. Three classes of statistics will be described to test these hypotheses: Wald, Lagrange Multiplier and Likelihood Ratio. Wald tests are perhaps the most intuitive: they directly test whether $\mathbf{R}\beta - \mathbf{r}$ is *close* to zero. Lagrange Multiplier tests incorporate the constraint into the least-squares problem using a Lagrangian. If the constraint has a small effect on the minimized sum of squares, the Lagrange multipliers, often described as the shadow price of the constraint in economic applications, should be close to zero. The magnitude of these forms the basis of the LM test statistic. Finally, likelihood ratios test whether the data are less likely under the null than they are under the alternative. If the null hypothesis is not restrictive this ratio should be close to one and the difference in the log-likelihoods should be small.

3.8.1 *t*-tests

T-tests can be used to test a single hypothesis involving one or more coefficients,

$$H_0 : \mathbf{R}\beta = r$$

where $\mathbf{R}$ is a 1 by k vector and r is a scalar. Recall from theorem 3.4, $\hat{\beta} - \beta \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Under the null, $\mathbf{R}(\hat{\beta} - \beta) = \mathbf{R}\hat{\beta} - \mathbf{R}\beta = \mathbf{R}\hat{\beta} - r$ and applying the properties of normal random variables,

$$\mathbf{R}\hat{\beta} - r \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}').$$

A simple test can be constructed

$$z = \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}, \tag{3.45}$$

where $z \sim N(0,1)$. To perform a test with size α, the value of z can be compared to the critical values of the standard normal and rejected if $|z| > C_\alpha$ where C_α is the $1-\alpha$ quantile of a standard normal. However, z is an infeasible statistic since it depends on an unknown quantity, σ^2. The natural solution is to replace the unknown parameter with an estimate. Dividing z by $\sqrt{\frac{s^2}{\sigma^2}}$ and simplifying,

$$\begin{aligned} t &= \frac{z}{\sqrt{\frac{s^2}{\sigma^2}}} \\ &= \frac{\frac{\mathbf{R}\hat{\beta} - r}{\sqrt{\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}}{\sqrt{\frac{s^2}{\sigma^2}}} \\ &= \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}. \end{aligned} \tag{3.46}$$

Note that the denominator $(n-k)\frac{s^2}{\sigma^2} \sim \chi^2_{n-k}$, and so t is the ratio of a standard normal to the square root of a χ^2_ν normalized by it standard deviation. As long as the standard normal in the numerator and the χ^2_ν are independent, this ratio will have a Student's t distribution.

Definition 3.9 (Student's t distribution)**.** Let $z \sim N(0,1)$ (standard normal) and let $w \sim \chi^2_\nu$ where z and w are independent. Then

$$\frac{z}{\sqrt{\frac{w}{\nu}}} \sim t_\nu. \tag{3.47}$$

2

The independence of $\hat{\beta}$ and s^2 – which is only a function of $\hat{\varepsilon}$ – follows from 3.5, and so t has a Student's t distribution.

Theorem 3.9 (t-test)**.** *Under assumptions 3.1 - 3.6,*

$$\frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}} \sim t_{n-k}. \tag{3.48}$$

As $\nu \to \infty$, the Student's t distribution converges to a standard normal. As a practical matter, when $\nu > 30$, the T distribution is close to a normal. While any single linear restriction can be tested with a t-test , the expression t-stat has become synonymous with a specific null hypothesis.

Definition 3.10 (t-stat)**.** The t-stat of a coefficient, β_k, is the t-test value of a test of the null $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k \neq 0$, and is computed

$$\frac{\hat{\beta}_k}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{[kk]}}} \tag{3.49}$$

where $(\mathbf{X}'\mathbf{X})^{-1}_{[kk]}$ is the k$^{\text{th}}$ diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

The previous examples were all two-sided; the null would be rejected if the parameters differed in either direction from the null hypothesis. The T-test is also unique among these three main classes of test statistics in that it can easily be applied against both one-sided alternatives and two-sided alternatives.[7]

However, there is often a good argument to test a one-sided alternative. For instance, in tests of the market premium, theory indicates that it must be positive to induce investment. Thus, when testing the null hypothesis that a risk premium is zero, a two-sided alternative could reject in cases which are not theoretically interesting. More importantly, a one-sided alternative, when appropriate, will have more power than a two-sided alternative since the direction information in the null hypothesis can be used to tighten confidence intervals. The two types of tests involving a one-sided hypothesis are upper tail tests which test nulls of the form $H_0 : \mathbf{R}\beta \leq r$ against alternatives of the form $H_1 : \mathbf{R}\beta > r$, and lower tail tests which test $H_0 : \mathbf{R}\beta \geq r$ against $H_1 : \mathbf{R}\beta < r$.

Figure 3.1 contains the rejection regions of a t_{10} distribution. The dark gray region corresponds to the rejection region of a two-sided alternative to the null that $H_0 : \hat{\beta} = \beta^0$ for a 10% test. The light gray region, combined with the upper dark gray region corresponds to the rejection region of a one-sided upper tail test, and so test statistic between 1.372 and 1.812 would be rejected using a one-sided alternative but not with a two-sided one.

Algorithm 3.1 (t-test)**.**

[7] Wald, LM, and LR tests can be implemented against one-sided alternatives with considerably more effort.

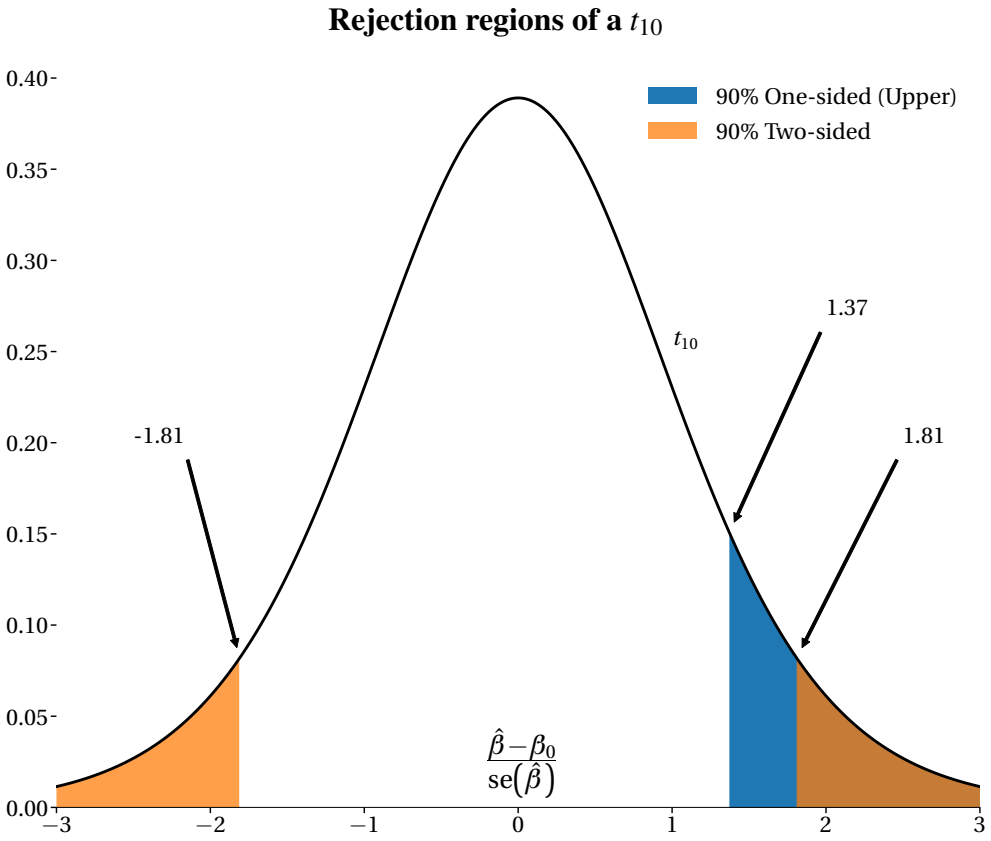

Figure 3.1: Rejection region for a t-test of the nulls $H_0 : \beta = \beta^0$ (two-sided) and $H_0 : \beta \leq \beta^0$. The two-sided rejection region is indicated by dark gray while the one-sided (upper) rejection region includes both the light and dark gray areas in the right tail.

---

1. *Estimate* $\hat{\beta}$ *using least squares.*
2. *Compute* $s^2 = (n-k)^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ *and* $s^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. *Construct the restriction matrix,* $\mathbf{R}$*, and the value of the restriction,* r *from the null hypothesis.*
4. *Compute* $t = \frac{\mathbf{R}\hat{\beta}-r}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}$.
5. *Compare* t *to the critical value,* C_α*, of the* t_{n-k} *distribution for a test size with* α*. In the case of a two tailed test, reject the null hypothesis if* $|t| > F_{t_\nu}(1-\alpha/2)$ *where* $F_{t_\nu}(\cdot)$ *is the CDF of a* t_ν*-distributed random variable. In the case of a one-sided upper-tail test, reject if* $t > F_{t_\nu}(1-\alpha)$ *or in the case of a one-sided lower-tail test, reject if* $t < F_{t_\nu}(\alpha)$.

3.8.2 Wald Tests

Wald test directly examines the distance between $\mathbf{R}\beta$ and $\mathbf{r}$. Intuitively, if the null hypothesis is true, then $\mathbf{R}\beta - \mathbf{r} \approx 0$. In the small-sample framework, the distribution of $\mathbf{R}\beta - \mathbf{r}$ follows directly from the properties of normal random variables. Specifically,

$$\mathbf{R}\beta - \mathbf{r} \sim N(\mathbf{0}, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$$

Thus, to test the null $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$ against the alternative $H_0 : \mathbf{R}\beta - \mathbf{r} \neq 0$, a test statistic can be based on

$$W_{\text{Infeasible}} = \frac{(\mathbf{R}\beta - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\beta - \mathbf{r})}{\sigma^2} \tag{3.50}$$

which has a χ^2_m distribution.[8] However, this statistic depends on an unknown quantity, σ^2, and to operationalize W, σ^2 must be replaced with an estimate, s^2.

$$W = \frac{(\mathbf{R}\beta - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{\sigma^2} \frac{\sigma^2}{s^2} = \frac{(\mathbf{R}\beta - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{s^2} \tag{3.51}$$

The replacement of σ^2 with s^2 has an effect on the distribution of the estimator which follows from the definition of an F distribution.

Definition 3.11 (F distribution)**.** Let $z_1 \sim \chi^2_{\nu_1}$ and let $z_2 \sim \chi^2_{\nu_2}$ where z_1 and z_2 are independent. Then

$$\frac{\frac{z_1}{\nu_1}}{\frac{z_2}{\nu_2}} \sim F_{\nu_1,\nu_2} \tag{3.52}$$

The conclusion that W has a $F_{m,n-k}$ distribution follows from the independence of $\hat{\beta}$ and $\hat{\varepsilon}$, which in turn implies the independence of $\hat{\beta}$ and s^2.

Theorem 3.10 (Wald test)**.** *Under assumptions 3.1 - 3.6,*

$$\frac{(\mathbf{R}\beta - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{s^2} \sim F_{m,n-k} \tag{3.53}$$

Analogous to the t_ν distribution, an F_{ν_1,ν_2} distribution converges to a scaled χ^2 in large samples ($\chi^2_{\nu_1}/\nu_1$ as $\nu_2 \to \infty$). Figure 3.2 contains *failure to reject* (FTR) regions for some hypothetical Wald tests. The shape of the region depends crucially on the correlation between the hypotheses being tested. For instance, panel (a) corresponds to testing a joint hypothesis where the tests are independent and have the same variance. In this case, the FTR region is a circle. Panel (d) shows the FTR region for highly correlated tests where one restriction has a larger variance.

Once W has been computed, the test statistic should be compared to the critical value of an $F_{m,n-k}$ and rejected if the test statistic is larger. Figure 3.3 contains the pdf of an $F_{5,30}$ distribution. Any $W > 2.049$ would lead to rejection of the null hypothesis using a 10% test.

The Wald test has a more common expression in terms of the SSE from both the restricted and unrestricted models. Specifically,

[8] The distribution can be derived noting that $\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-\frac{1}{2}} (\mathbf{R}\beta - \mathbf{r}) \sim N\left(0, \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)$ where the matrix square root makes use of a generalized inverse. A more complete discussion of reduced rank normals and generalized inverses is beyond the scope of this course.

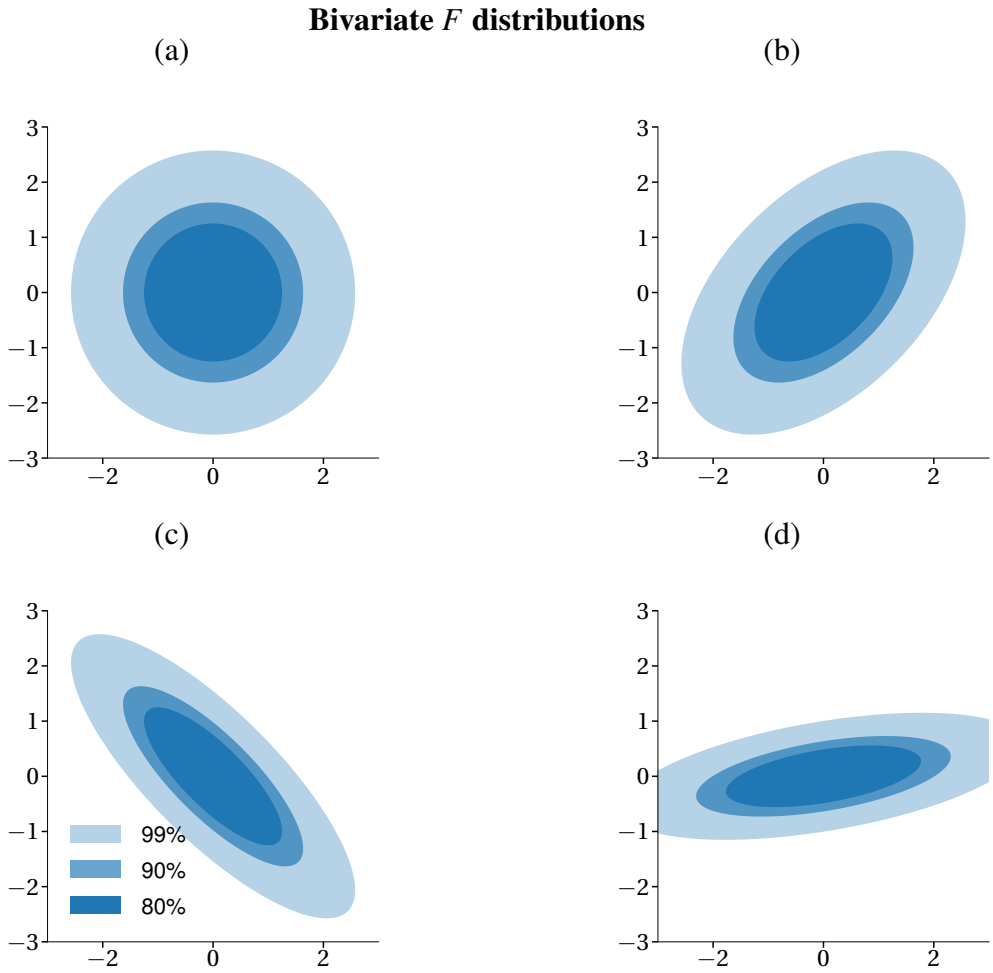

Figure 3.2: Bivariate plot of an F distribution. The four panels contain the failure-to-reject regions corresponding to 20, 10 and 1% tests. Panel (a) contains the region for uncorrelated tests. Panel (b) contains the region for tests with the same variance but a correlation of 0.5. Panel (c) contains the region for tests with a correlation of -.8 and panel (d) contains the region for tests with a correlation of 0.5 but with variances of 2 and 0.5 (The test with a variance of 2 is along the x-axis).

$$W = \frac{\frac{\text{SSE}_\text{R} - \text{SSE}_\text{U}}{m}}{\frac{\text{SSE}_\text{U}}{n-k}} = \frac{\frac{\text{SSE}_\text{R} - \text{SSE}_\text{U}}{m}}{s^2}. \tag{3.54}$$

where SSE_R is the sum of squared errors of the restricted model.[9] The restricted model is the original model with the null hypothesis imposed. For example, to test the null $H_0 : \beta_2 = \beta_3 = 0$ against an alternative that $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$ in a bivariate regression,

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \varepsilon_i \tag{3.55}$$

[9]The SSE should be the result of minimizing the squared errors. The centered should be used if a constant is included and the uncentered versions if no constant is included.

the restricted model imposes the null,

$$\begin{aligned} Y_i &= \beta_1 + 0X_{1,i} + 0X_{2,i} + \varepsilon_i \\ &= \beta_1 + \varepsilon_i. \end{aligned}$$

The restricted SSE, SSE_R is computed using the residuals from this model while the unrestricted SSE, SSE_U, is computed from the general model that includes both X variables (eq. (3.55)). While Wald tests usually only require the unrestricted model to be estimated, the difference of the SSEs is useful because it can be computed from the output of any standard regression package. Moreover, any linear regression subject to linear restrictions can be estimated using OLS on a modified specification where the constraint is directly imposed. Consider the set of restrictions, $\mathbf{R}$, in an augmented matrix with $\mathbf{r}$

$$\begin{bmatrix} \mathbf{R} & \mathbf{r} \end{bmatrix}$$

By transforming this matrix into row-echelon form,

$$\begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{R}} & \tilde{\mathbf{r}} \end{bmatrix}$$

a set of m restrictions can be derived. This also provides a direct method to check whether a set of constraints is logically consistent and feasible or if it contains any redundant restrictions.

Theorem 3.11 (Restriction Consistency and Redundancy)**.** *If* $\begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{R}} & \tilde{\mathbf{r}} \end{bmatrix}$ *is* $\begin{bmatrix} \mathbf{R} & \mathbf{r} \end{bmatrix}$ *in reduced echelon form, then a set of restrictions is logically consistent if* $\text{rank}(\tilde{\mathbf{R}}) = \text{rank}(\begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{R}} & \tilde{\mathbf{r}} \end{bmatrix})$. *Additionally, if* $\text{rank}(\tilde{\mathbf{R}}) = \text{rank}(\begin{bmatrix} \mathbf{I}_m & \tilde{\mathbf{R}} & \tilde{\mathbf{r}} \end{bmatrix}) = m$*, then there are no redundant restrictions.*

1. *Estimate the unrestricted model* $Y_i = \mathbf{x}_i\beta + \varepsilon_i$*, and the restricted model,* $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$*.*
2. *Compute* $\text{SSE}_\text{R} = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ *where* $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ *are the residuals from the restricted regression, and* $\text{SSE}_\text{U} = \sum_{i=1}^n \hat{\varepsilon}_i^2$ *where* $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ *are the residuals from the unrestricted regression.*
3. *Compute* $W = \frac{\frac{\text{SSE}_\text{R} - \text{SSE}_\text{U}}{m}}{\frac{\text{SSE}_\text{U}}{n-k}}$.
4. *Compare* W *to the critical value,* C_α*, of the* $F_{m,n-k}$ *distribution at size* α*. Reject the null hypothesis if* $W > C_\alpha$*.*

Finally, in the same sense that the t-stat is a test of the null $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k \neq 0$, the F-stat of a regression tests whether all coefficients are zero (except the intercept) against an alternative that at least one is non-zero.

Definition 3.12 (F-stat of a Regression)**.** The F-stat of a regression is the value of a Wald test that all coefficients are zero except the coefficient on the constant (if one is included). Specifically, if the unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots \beta_k X_{k,i} + \varepsilon_i,$$

the F-stat is the value of a Wald test of the null $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$ against the alternative $H_1 : \beta_j \neq 0$, for $j = 2, \ldots, k$ and corresponds to a test based on the restricted regression

$$Y_i = \beta_1 + \varepsilon_i.$$

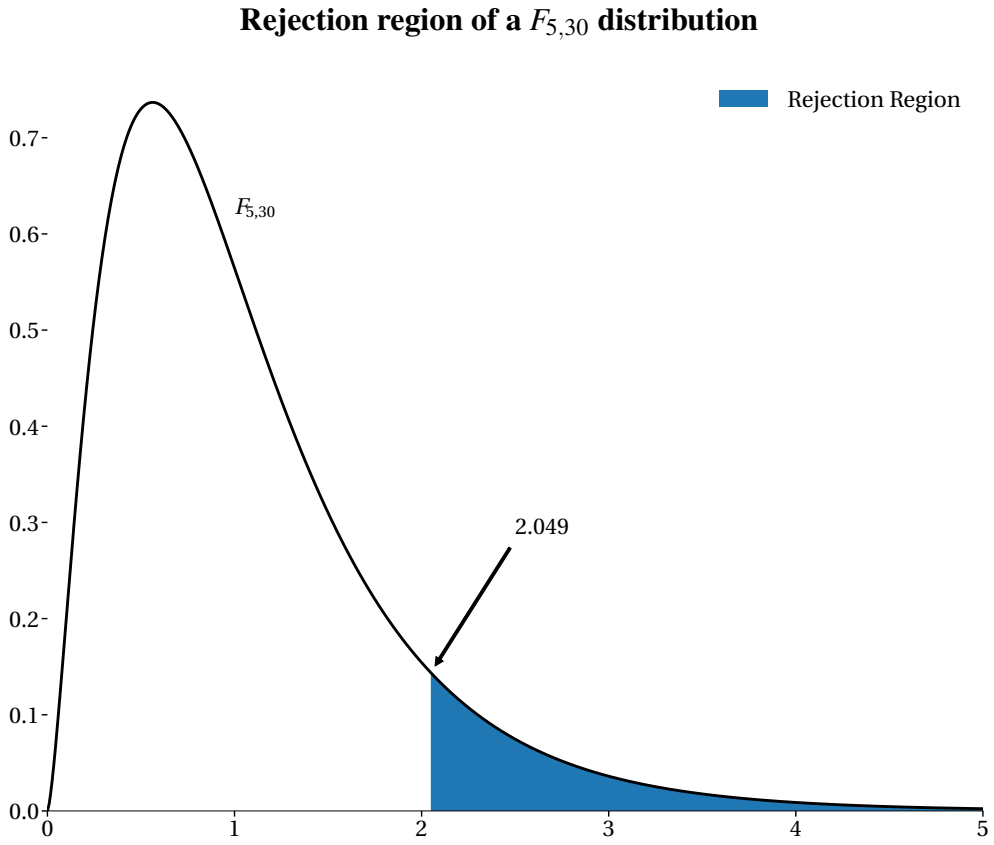

Figure 3.3: Rejection region for a $F_{5,30}$ distribution when using a test with a size of 10%. If the null hypothesis is true, the test statistic should be relatively small (would be 0 if exactly true). Large test statistics lead to rejection of the null hypothesis. In this example, a test statistic with a value greater than 2.049 would lead to a rejection of the null at the 10% level.

3.8.3 Example: T and Wald Tests in Cross-Sectional Factor models

Returning to the factor regression example, the t-stats in the 4-factor model can be computed

$$t_j = \frac{\hat{\beta}_j}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{[jj]}}}.$$

For example, consider a regression of BH^e on the set of four factors and a constant,

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$$

The fit coefficients, t-stats and p-values are contained in table 3.5.

Definition 3.13 (P-value)**.** The p-value is the smallest test size (α) where the null hypothesis may be rejected. The p-value can be equivalently defined as the largest size where the null hypothesis cannot be rejected.

P-values have the advantage that they are independent of the distribution of the test statistic. For example, when using a 2-sided t-test, the p-value of a test statistic t is $2(1-F_{t_\nu}(|t|))$ where $F_{t_\nu}(|\cdot|)$ is the CDF of a t-distribution with ν degrees of freedom. In a Wald test, the p-value is $1-F_{f_{\nu_1,\nu_2}}(W)$ where $F_{f_{\nu_1,\nu_2}}(\cdot)$ is the CDF of an f_{ν_1,ν_2} distribution.

The critical value, C_α, for a 2-sided 10% t-test with 973 degrees of freedom ($n-5$) is 1.645, and so if $|t|>C_\alpha$ the null hypothesis should be rejected, and the results indicate that the null hypothesis that the coefficients on the constant and *SMB* are zero cannot be rejected the 10% level. The p-values indicate the null that the constant was 0 could be rejected at a α of 14% but not one of 13%.

Table 3.5 also contains the Wald test statistics and p-values for a variety of hypotheses, some economically interesting, such as the set of restrictions that the four factor model reduces to the CAPM, $\beta_j=0,\ j=1,3,\ldots,5$. Only one regression, the completely unrestricted regression, was needed to compute all of the test statistics using Wald tests,

$$W=\frac{(\mathbf{R}\beta-\mathbf{r})'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\beta-\mathbf{r})}{s^2}$$

where $\mathbf{R}$ and $\mathbf{r}$ depend on the null being tested. For example, to test whether a strict CAPM was consistent with the observed data,

$$\mathbf{R}=\begin{bmatrix}1&0&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}\text{ and }\mathbf{r}=\begin{bmatrix}0\\0\\0\\0\end{bmatrix}.$$

All of the null hypotheses save one are strongly rejected with p-values of 0 to three decimal places. The sole exception is $H_0:\beta_1=\beta_3=0$, which produced a Wald test statistic of 2.05. The 5% critical value of an $F_{2,973}$ is 3.005, and so the null hypothesis would be not rejected at the 5% level. The p-value indicates that the test would be rejected at the 13% level but not at the 12% level. One further peculiarity appears in the table. The Wald test statistic for the null $H_0:\beta_5=0$ is exactly the square of the t-test statistic for the same null. This should not be surprising since $W=t^2$ when testing a single linear hypothesis. Moreover, if $z\sim t_\nu$, then $z^2\sim F_{1,\nu}$. This can be seen by inspecting the square of a t_ν and applying the definition of an $F_{1,\nu}$-distribution.

3.8.4 Likelihood Ratio Tests

Likelihood Ratio (LR) test are based on the relative probability of observing the data if the null is valid to the probability of observing the data under the alternative. The test statistic is defined

$$LR=-2\ln\left(\frac{\max_{\beta,\sigma^2}f(\mathbf{y}|\mathbf{X};\beta,\sigma^2)\quad\text{subject to}\quad\mathbf{R}\beta=\mathbf{r}}{\max_{\beta,\sigma^2}f(\mathbf{y}|\mathbf{X};\beta,\sigma^2)}\right)\tag{3.56}$$

Letting $\hat{\beta}_R$ denote the constrained estimate of β, this test statistic can be reformulated

t-Tests

	$\hat{\beta}$	s.e. $\left(\hat{\beta}\right)$	t-stat	p-value
Constant	-0.086	0.042	-2.04	0.042
VWM^e	1.080	0.010	108.7	0.000
SMB	0.002	0.014	0.13	0.893
HML	0.764	0.015	50.8	0.000
MOM	-0.035	0.010	-3.50	0.000

Wald Tests

Null	Alternative	W	M	p-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	3558.8	4	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	956.5	3	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	10.1	2	0.000
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2.08	2	0.126
$\beta_5 = 0$	$\beta_5 \neq 0$	12.3	1	0.000

Table 3.5: The upper panel contains t-stats and p-values for the regression of Big-High excess returns on the four factors and a constant. The lower panel contains test statistics and p-values for Wald tests of the reported null hypothesis. Both sets of tests were computed using the small-sample assumptions and may be misleading since the residuals are both non-normal and heteroskedastic.

$$\begin{aligned} LR &= -2\ln\left(\frac{f(\mathbf{y}|\mathbf{X};\hat{\beta}_R,\hat{\sigma}_R^2)}{f(\mathbf{y}|\mathbf{X};\hat{\beta},\hat{\sigma}^2)}\right) \\ &= -2[l(\hat{\beta}_R,\hat{\sigma}_R^2;\mathbf{y}|\mathbf{X};) - l(\hat{\beta},\hat{\sigma}^2;\mathbf{y}|\mathbf{X})] \\ &= 2[l(\hat{\beta},\hat{\sigma}^2;\mathbf{y}|\mathbf{X}) - l(\hat{\beta}_R,\hat{\sigma}_R^2;\mathbf{y}|\mathbf{X})] \end{aligned} \tag{3.57}$$

In the case of the normal log likelihood, LR can be further simplified to[10]

$$\begin{aligned} LR &= -2\ln\left(\frac{f(\mathbf{y}|\mathbf{X};\hat{\beta}_R,\hat{\sigma}_R^2)}{f(\mathbf{y}|\mathbf{X};\hat{\beta},\hat{\sigma}^2)}\right) \\ &= -2\ln\left(\frac{(2\pi\hat{\sigma}_R^2)^{-\frac{n}{2}}\exp(-\frac{(\mathbf{y}-\mathbf{X}\hat{\beta}_R)'(\mathbf{y}-\mathbf{X}\hat{\beta}_R)}{2\hat{\sigma}_R^2})}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}}\exp(-\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\hat{\sigma}^2})}\right) \\ &= -2\ln\left(\frac{(\hat{\sigma}_R^2)^{-\frac{n}{2}}}{(\hat{\sigma}^2)^{-\frac{n}{2}}}\right) \\ &= -2\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}^2}\right)^{-\frac{n}{2}} \end{aligned}$$

[10]Note that $\hat{\sigma}_R^2$ and $\hat{\sigma}^2$ use n rather than a degree-of-freedom adjustment since they are MLE estimators.

$$= n\left[\ln(\hat{\sigma}_R^2) - \ln(\hat{\sigma}^2)\right]$$
$$= n\left[\ln(\text{SSE}_\text{R}) - \ln(\text{SSE}_\text{U})\right]$$

Finally, the distribution of the *LR* statistic can be determined by noting that

$$LR = n\ln\left(\frac{\text{SSE}_\text{R}}{\text{SSE}_\text{U}}\right) = N\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}\right) \tag{3.58}$$

and that

$$\frac{n-k}{m}\left[\exp\left(\frac{LR}{n}\right) - 1\right] = W. \tag{3.59}$$

The transformation between W and LR is monotonic so the transformed statistic has the same distribution as W, a $F_{m,n-k}$.

Algorithm 3.2 (Small-Sample Wald Test)**.**

1. *Estimate the unrestricted model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$, and the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.*
2. *Compute $\text{SSE}_\text{R} = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ where $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ are the residuals from the restricted regression, and $\text{SSE}_\text{U} = \sum_{i=1}^n \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ are the residuals from the unrestricted regression.*
3. *Compute $LR = n\ln\left(\frac{\text{SSE}_\text{R}}{\text{SSE}_\text{U}}\right)$.*
4. *Compute $W = \frac{n-k}{m}\left[\exp\left(\frac{LR}{n}\right) - 1\right]$.*
5. *Compare W to the critical value, C_α, of the $F_{m,n-k}$ distribution at size α. Reject the null hypothesis if $W > C_\alpha$.*

3.8.5 Example: LR Tests in Cross-Sectional Factor models

LR tests require estimating the model under both the null and the alternative. In all examples here, the alternative is the unrestricted model with four factors while the restricted models (where the null is imposed) vary. The simplest restricted model corresponds to the most restrictive null, $H_0 : \beta_j = 0$, $j = 1,\ldots,5$, and is specified

$$Y_i = \varepsilon_i.$$

To compute the likelihood ratio, the conditional mean and variance must be estimated. In this simple specification, the conditional mean is $\hat{\mathbf{y}}_R = \mathbf{0}$ (since there are no parameters) and the conditional variance is estimated using the MLE with the mean, $\hat{\sigma}_R^2 = \mathbf{y}'\mathbf{y}/n$ (the sum of squared regressands). The mean under the alternative is $\hat{\mathbf{y}}_U = \mathbf{x}_i'\hat{\beta}$ and the variance is estimated using $\hat{\sigma}_U^2 = (\mathbf{y} - \mathbf{x}_i'\hat{\beta})'(\mathbf{y} - \mathbf{x}_i'\hat{\beta})/n$. Once these quantities have been computed, the *LR* test statistic is calculated

$$LR = n\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}\right) \tag{3.60}$$

LR Tests				
Null	Alternative	*LR*	*M*	*p*-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	3558.8	4	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	956.5	3	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	10.1	2	0.000
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2.08	2	0.126
$\beta_5 = 0$	$\beta_5 \neq 0$	12.3	1	0.000

LM Tests				
Null	Alternative	*LM*	*M*	*p*-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	163.4	4	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	184.3	3	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	9.85	2	0.000
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2.07	2	0.127
$\beta_5 = 0$	$\beta_5 \neq 0$	12.1	1	0.001

Table 3.6: The upper panel contains test statistics and p-values using LR tests for using a regression of excess returns on the big-high portfolio on the four factors and a constant. In all cases the null was tested against the alternative listed. The lower panel contains test statistics and p-values for LM tests of same tests. Note that the LM test statistics are uniformly smaller than the LR test statistics which reflects that the variance in a LM test is computed from the model estimated under the null, a value that must be larger than the estimate of the variance under the alternative which is used in both the Wald and LR tests. Both sets of tests were computed using the small-sample assumptions and may be misleading since the residuals are non-normal and heteroskedastic.

where the identity $\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2} = \frac{\text{SSE}_\text{R}}{\text{SSE}_\text{U}}$ has been applied. Finally, *LR* is transformed by $\frac{n-k}{m}\left[\exp\left(\frac{LR}{n}\right) - 1\right]$ to produce the test statistic, which is numerically identical to W. This can be seen by comparing the values in table 3.6 to those in table 3.5.

3.8.6 Lagrange Multiplier Tests

Consider minimizing the sum of squared errors subject to a linear hypothesis.

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \text{ subject to } \mathbf{R}\beta - \mathbf{r} = 0.$$

This problem can be formulated in terms of a Lagrangian,

$$\mathcal{L}(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\mathbf{R}\beta - \mathbf{r})'\lambda$$

and the problem is

$$\max_{\lambda} \left\{ \min_{\beta} \mathcal{L}(\beta, \lambda) \right\}$$

The first-order conditions correspond to a saddle point,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta} &= -2\mathbf{X}'(\mathbf{y}-\mathbf{X}\beta)+\mathbf{R}'\lambda = \mathbf{0}\\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \mathbf{R}\beta-\mathbf{r} = \mathbf{0}\end{aligned}$$

pre-multiplying the top FOC by $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$ (which does not change the value, since it is 0),

$$\begin{aligned}2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta-2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}+\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda &= \mathbf{0}\\ \Rightarrow 2\mathbf{R}\beta-2\mathbf{R}\hat{\beta}+\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda &= \mathbf{0}\end{aligned}$$

where $\hat{\beta}$ is the usual OLS estimator. Solving,

$$\tilde{\lambda} = 2\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\beta}-\mathbf{r}) \tag{3.61}$$

$$\tilde{\beta} = \hat{\beta}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\beta}-\mathbf{r}) \tag{3.62}$$

These two solutions provide some insight into the statistical properties of the estimators. $\tilde{\beta}$, the constrained regression estimator, is a function of the OLS estimator, $\hat{\beta}$, and a step in the direction of the constraint. The size of the change is influenced by the distance between the unconstrained estimates and the constraint ($\mathbf{R}\hat{\beta}-\mathbf{r}$). If the unconstrained estimator happened to exactly satisfy the constraint, there would be no step.[11]

The Lagrange multipliers, $\tilde{\lambda}$, are weighted functions of the unconstrained estimates, $\hat{\beta}$, and will be near zero if the constraint is nearly satisfied ($\mathbf{R}\hat{\beta}-\mathbf{r}\approx 0$). In microeconomics, Lagrange multipliers are known as *shadow prices* since they measure the magnitude of the change in the objective function would if the constraint was relaxed a small amount. Note that $\hat{\beta}$ is the only source of randomness in $\tilde{\lambda}$ (like $\tilde{\beta}$), and so $\tilde{\lambda}$ is a linear combination of normal random variables and will also follow a normal distribution. These two properties combine to provide a mechanism for testing whether the restrictions imposed by the null are consistent with the data. The distribution of $\hat{\lambda}$ can be directly computed and a test statistic can be formed.

There is another method to derive the LM test statistic that is motivated by the alternative name of LM tests: Score tests. Returning to the first-order conditions and plugging in the parameters,

$$\begin{aligned}\mathbf{R}'\lambda &= 2\mathbf{X}'(\mathbf{y}-\mathbf{X}\tilde{\beta})\\ \mathbf{R}'\lambda &= 2\mathbf{X}'\tilde{\varepsilon}\end{aligned}$$

where $\tilde{\beta}$ is the constrained estimate of β and $\tilde{\varepsilon}$ are the corresponding estimated errors ($\tilde{\varepsilon} = \mathbf{y}-\mathbf{X}\tilde{\beta}$). Thus $\mathbf{R}'\lambda$ has the same distribution as $2\mathbf{X}'\tilde{\varepsilon}$. However, under the small-sample assumptions, $\tilde{\varepsilon}$ are linear combinations of normal random variables and so are also normal,

$$2\mathbf{X}'\tilde{\varepsilon} \sim N(\mathbf{0}, 4\sigma^2\mathbf{X}'\mathbf{X})$$

[11] Even if the constraint is valid, the constraint will never be exactly satisfied.

and

$$\mathbf{X}'\tilde{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{X}'\mathbf{X}). \tag{3.63}$$

A test statistic that the scores are zero can be constructed in the same manner as a Wald test:

$$LM_{\text{Infeasible}} = \frac{\tilde{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\varepsilon}}{\sigma^2}. \tag{3.64}$$

However, like a Wald test this statistic is not feasible since σ^2 is unknown. Using the same substitution, the LM test statistic is given by

$$LM = \frac{\tilde{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\varepsilon}}{\tilde{s}^2} \tag{3.65}$$

and has a $F_{m,n-k+m}$ distribution where $\tilde{s}^2$ is the estimated error variance from the constrained regression. This is a different estimator than was used in constructing a Wald test statistic, where the variance was computed from the unconstrained model. Both estimates are consistent under the null. However, since $\text{SSE}_\text{R} \geq \text{SSE}_\text{U}$, $\tilde{s}^2$ is likely to be larger than s^2.[12] LM tests are usually implemented using a more convenient – but equivalent – form,

$$LM = \frac{\frac{\text{SSE}_\text{R} - \text{SSE}_\text{U}}{m}}{\frac{\text{SSE}_\text{R}}{n-k+m}}. \tag{3.66}$$

To use the Lagrange Multiplier principle to conduct a test:

Algorithm 3.3 (Small-Sample LM Test)**.**

1. *Estimate the unrestricted model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$, and the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.*
2. *Compute $\text{SSE}_\text{R} = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ where $\tilde{\varepsilon}_i = \tilde{y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ are the residuals from the restricted regression, and $\text{SSE}_\text{U} = \sum_{i=1}^n \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ are the residuals from the unrestricted regression.*
3. *Compute $LM = \frac{\frac{\text{SSE}_\text{R} - \text{SSE}_\text{U}}{m}}{\frac{\text{SSE}_\text{R}}{n-k+m}}$.*
4. *Compare LM to the critical value, C_α, of the $F_{m,n-k+m}$ distribution at size α. Reject the null hypothesis if $LM > C_\alpha$.*

Alternatively, the scores can be directly tested.

Algorithm 3.4 (Alternative Small-Sample LM Test)**.**

1. *Estimate the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.*
2. *Compute $LM = \frac{\tilde{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\varepsilon}}{\frac{m}{s^2}}$ where $\mathbf{X}$ is n by k the matrix of regressors from the unconstrained model and $s^2 = \frac{\sum_{i=1}^n \tilde{\varepsilon}_i^2}{n-k+m}$.*
3. *Compare LM to the critical value, C_α, of the $F_{m,n-k+m}$ distribution at size α. Reject the null hypothesis if $LM > C_\alpha$.*

[12] Note that since the degree-of-freedom adjustment in the two estimators is different, the magnitude estimated variance is not directly proportional to SSE_R and SSE_U.

3.8.7 Example: LM Tests in Cross-Sectional Factor models

Table 3.6 also contains values from LM tests. LM tests have a slightly different distributions than the Wald and LR and do not produce numerically identical results. While the Wald and LR tests require estimation of the unrestricted model (estimation under the alternative), LM tests only require estimation of the restricted model (estimation under the null). For example, in testing the null H_0 : $\beta_1 = \beta_5 = 0$ (that the *MOM* factor has no explanatory power and that the intercept is 0), the restricted model is estimated from

$$BH_i^e = \gamma_1 VWM_i^e + \gamma_2 SMB_i + \gamma_3 HML_i + \varepsilon_i.$$

The two conditions, that $\beta_1 = 0$ and that $\beta_5 = 0$ are imposed by excluding these regressors. Once the restricted regression is fit, the residuals estimated under the null, $\tilde{\varepsilon}_i = Y_i - \mathbf{x}_i\tilde{\beta}$ are computed and the LM test is calculated from

$$LM = \frac{\tilde{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\varepsilon}}{s^2}$$

where $\mathbf{X}$ is the set of explanatory variables from the *unrestricted* regression (in the case, $\mathbf{x}_i = [1 \; VWM_i^e \; SMB_i \; HML_i \; MOM_i]$). Examining table 3.6, the LM test statistics are considerably smaller than the Wald test statistics. This difference arises since the variance used in computing the LM test statistic, $\tilde{\sigma}^2$, is estimated under the null. For instance, in the most restricted case ($H_0 = \beta_j = 0$, $j = 1, \ldots, k$), the variance is estimated by $\mathbf{y}'\mathbf{y}/N$ (since $k = 0$ in this model) which is very different from the variance estimated under the alternative (which is used by both the Wald and LR). Despite the differences in the test statistics, the p-values in the table would result in the same inference. For the one hypothesis that is not completely rejected, the p-value of the LM test is slightly larger than that of the LR (or W). However, .130 and .129 should never make a qualitative difference (nor should .101 and .099, even when using a 10% test). These results highlight a general feature of LM tests: test statistics based on the LM-principle are smaller than Likelihood Ratios and Wald tests, and so less likely to reject.

3.8.8 Comparing the Wald, LR, and LM Tests

With three tests available to test the same hypothesis, which is the correct one? In the small-sample framework, the Wald is the obvious choice because $W \approx LR$ and W is larger than LM. However, the LM has a slightly different distribution, so it is impossible to make an absolute statement. The choice among these three tests reduces to user preference and ease of computation. Since computing SSE_U and SSE_R is simple, the Wald test is likely the simplest to implement.

These results are no longer true when nonlinear restrictions and/or nonlinear models are estimated. Further discussion of the factors affecting the choice between the Wald, LR, and LM tests will be reserved until then. Figure 3.4 contains a graphical representation of the three test statistics in the context of a simple regression, $Y_i = \beta X_i + \varepsilon_i$.[13] The Wald test measures the magnitude of the constraint $R\beta - r$ at the unconstrained estimator $\hat{\beta}$. The LR test measures how much of the sum of squared errors has changed between $\hat{\beta}$ and $\tilde{\beta}$. Finally, the LM test measures the magnitude of the gradient, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)$ at the constrained estimator $\tilde{\beta}$.

[13] Magnitudes of the lines is not to scale, so the magnitude of the test statistics cannot be determined from the picture.

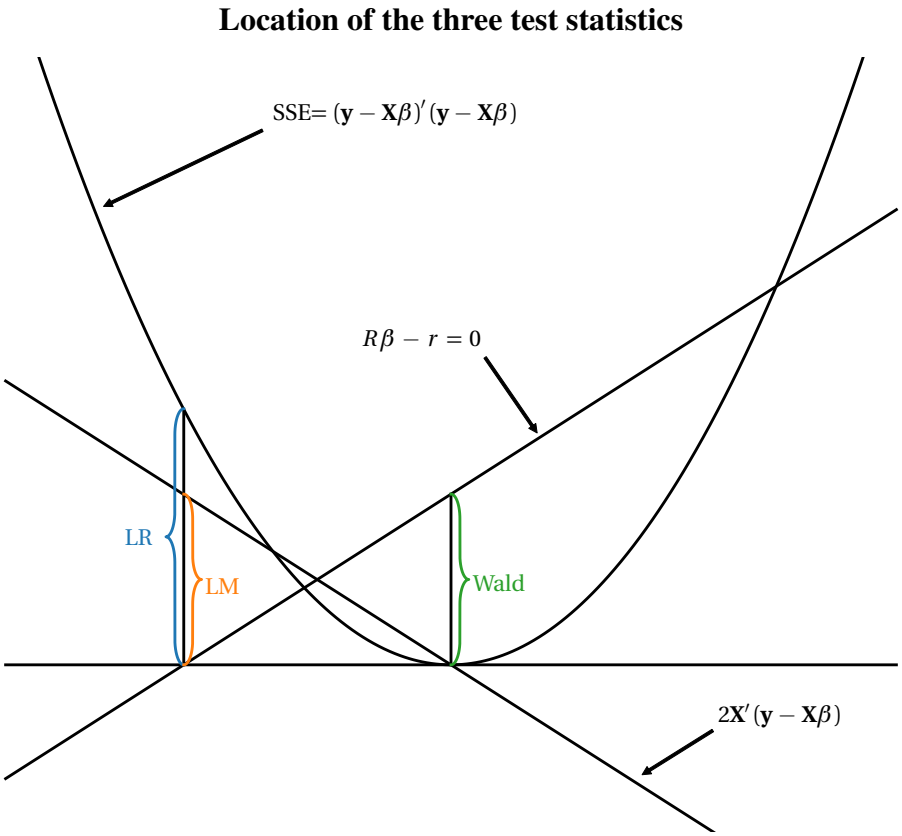

Figure 3.4: Graphical representation of the three major classes of tests. The Wald test measures the magnitude of the constraint, $\mathbf{R}\beta - r$, at the OLS parameter estimate, $\hat{\beta}$. The LM test measures the magnitude of the score at the restricted estimator ($\tilde{\beta}$) while the LR test measures the difference between the SSE at the restricted estimator and the SSET at the unrestricted estimator. Note: Only the location of the test statistic, not their relative magnitudes, can be determined from this illustration.

3.9 Large-Sample Assumption

While the small-sample assumptions allow the exact distribution of the OLS estimator and test statistics to be derived, these assumptions are not realistic in applications using financial data. Asset returns are non-normal (both skewed and leptokurtic), heteroskedastic, and correlated. The large-sample framework allows for inference on β without making strong assumptions about the distribution or error covariance structure. However, the generality of the large-sample framework comes at the loss of the ability to say anything exact about the estimates in finite samples.

Four new assumptions are needed to analyze the asymptotic behavior of the OLS estimators.

Assumption 3.7 (Stationary Ergodicity)**.** $\{(\mathbf{x}_i, \varepsilon_i)\}$ *is a strictly stationary and ergodic sequence.*

This is a technical assumption needed for consistency and asymptotic normality. It implies two properties about the joint density of $\{(\mathbf{x}_i, \varepsilon_i)\}$: the joint distribution of $\{(\mathbf{x}_i, \varepsilon_i)\}$ and $\{(\mathbf{x}_{i+j}, \varepsilon_{i+j})\}$ depends on the time between observations (j) and *not* the observation index (i) and that averages will converge to their expected value (as long as they exist). There are a number of alternative assumptions

that could be used in place of this assumption, although this assumption is broad enough to allow for i.i.d. , i.d.n.d (independent not identically distributed, including heteroskedasticity), and some n.i.n.i.d. data, although it does rule out some important cases. Specifically, the regressors cannot be trending or otherwise depend on the observation index, an important property of some economic time series such as the level of a market index or aggregate consumption. Stationarity will be considered more carefully in the time-series chapters.

Assumption 3.8 (Rank)**.** $\mathrm{E}[\mathbf{x}_i'\mathbf{x}_i] = \Sigma_{\mathbf{XX}}$ *is nonsingular and finite.*

This assumption, like assumption 3.3, is needed to ensure identification.

Assumption 3.9 (Martingale Difference)**.** $\{\mathbf{x}_i'\varepsilon_i, \mathcal{F}_i\}$ *is a martingale difference sequence,*

$$\mathrm{E}\left[\left(X_{j,i}\varepsilon_i\right)^2\right] < \infty,\ j = 1,2,\ldots,k,\ i = 1,2\ldots$$

and $\mathbf{S} = \mathrm{V}[n^{-\frac{1}{2}}\mathbf{X}'\boldsymbol{\varepsilon}]$ *is finite and non singular.*

A martingale difference sequence has the property that its mean is unpredictable using the information contained in the information set ($\mathcal{F}_i$).

Definition 3.14 (Martingale Difference Sequence)**.** Let $\{\mathbf{Z}_i\}$ be a vector stochastic process and $\mathcal{F}_i$ be the information set corresponding to observation i containing all information available when observation i was collected except $\mathbf{Z}_i$. $\{\mathbf{Z}_i, \mathcal{F}_i\}$ is a martingale difference sequence if

$$\mathrm{E}[\mathbf{Z}_i|\mathcal{F}_i] = \mathbf{0}$$

In the context of the linear regression model, it states that the current score is not predictable by any of the previous scores, that the mean of the scores is zero ($\mathrm{E}[\mathbf{X}_i'\varepsilon_i] = 0$), and there is no other variable in $\mathcal{F}_i$ which can predict the scores. This assumption is sufficient to ensure that $n^{-1/2}\mathbf{X}'\boldsymbol{\varepsilon}$ will follow a Central Limit Theorem, and it plays a role in consistency of the estimator. A m.d.s. is a fairly general construct and does not exclude using time-series regressors as long as they are *predetermined*, meaning that they do not depend on the process generating ε_i. For instance, in the CAPM, the return on the market portfolio can be thought of as being determined independently of the idiosyncratic shock affecting individual assets.

Assumption 3.10 (Moment Existence)**.** $\mathrm{E}[X_{j,i}^4] < \infty$, $i = 1,2,\ldots$, $j = 1,2,\ldots,k$ *and* $\mathrm{E}[\varepsilon_i^2] = \sigma^2 < \infty$, $i = 1,2,\ldots$.

This final assumption requires that the fourth moment of any regressor exists and the variance of the errors is finite. This assumption is needed to derive a consistent estimator of the parameter covariance.

3.10 Large-Sample Properties

These assumptions lead to two theorems that describe the asymptotic behavior of $\hat{\beta}$: it is consistent and asymptotically normally distributed. First, some new notation is needed. Let

$$\hat{\beta}_n = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right) \tag{3.67}$$

be the regression coefficient using n realizations from the stochastic process $\{\mathbf{x}_i, \varepsilon_i\}$.

Theorem 3.12 (Consistency of $\hat{\beta}$)**.** *Under assumptions 3.1 and 3.7 - 3.9*

$$\hat{\beta}_n \xrightarrow{p} \beta$$

Consistency is a weak property of the OLS estimator, but it is important. This result relies crucially on the implication of assumption 3.9 that $n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$, and under the same assumptions, the OLS estimator is also asymptotically normally distributed.

Theorem 3.13 (Asymptotic Normality of $\hat{\beta}$)**.** *Under assumptions 3.1 and 3.7 - 3.9*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}) \tag{3.68}$$

where $\Sigma_{\mathbf{xx}} = \mathrm{E}[\mathbf{x}_i'\mathbf{x}_i]$ *and* $\mathbf{S} = \mathrm{V}[n^{-1/2}\mathbf{X}'\boldsymbol{\varepsilon}]$

Asymptotic normality provides the basis for hypothesis tests on β. However, using only theorem 3.13, tests are not feasible since $\Sigma_{\mathbf{xx}}$ and $\mathbf{S}$ are unknown, and so must be estimated.

Theorem 3.14 (Consistency of OLS Parameter Covariance Estimator)**.** *Under assumptions 3.1 and 3.7 - 3.10,*

$$\begin{aligned}\hat{\Sigma}_{\mathbf{xx}} &= n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \Sigma_{\mathbf{xx}} \\ \hat{\mathbf{S}} &= n^{-1}\sum_{i=1}^{n} e_i^2\mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} \mathbf{S} \\ &= n^{-1}\left(\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}\right)\end{aligned}$$

and

$$\hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1} \xrightarrow{p} \Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}$$

where $\hat{\mathbf{E}} = \mathrm{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)$ *is a matrix with the estimated residuals squared along its diagonal.*

Combining these theorems, the OLS estimator is consistent, asymptotically normal, and the asymptotic variance can be consistently estimated. These three properties provide the tools necessary to conduct hypothesis tests in the asymptotic framework. The usual estimator of the residual variance is also consistent for the variance of the innovations under the same conditions.

Theorem 3.15 (Consistency of OLS Variance Estimator)**.** *Under assumptions 3.1 and 3.7 - 3.10 ,*

$$\hat{\sigma}_n^2 = n^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \xrightarrow{p} \sigma^2$$

Further, if homoskedasticity is assumed, then the parameter covariance estimator can be simplified.

Theorem 3.16 (Homoskedastic Errors)**.** *Under assumptions 3.1, 3.4, 3.5 and 3.7 - 3.10,*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2\Sigma_{\mathbf{xx}}^{-1})$$

Combining the result of this theorem with that of theorems 3.14 and 3.15, a consistent estimator of $\sigma^2\Sigma_{\mathbf{xx}}^{-1}$ is given by $\hat{\sigma}_n^2\hat{\Sigma}_{\mathbf{xx}}^{-1}$.

3.11 Large-Sample Hypothesis Testing

All three test types, Wald, LR, and LM, have large-sample equivalents that exploit the estimated parameters' asymptotic normality. While these tests are only asymptotically exact, the use of the asymptotic distribution is justified as an approximation to the finite-sample distribution, although the quality of the CLT approximation depends on how well behaved the data are.

3.11.1 Wald Tests

Recall from Theorem 3.13,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}). \tag{3.69}$$

Applying the properties of a normal random variable, if $\mathbf{z} \sim N(\mu, \Sigma)$, $\mathbf{c}'\mathbf{z} \sim N(\mathbf{c}'\mu, \mathbf{c}'\Sigma\mathbf{c})$ and that if $w \sim N(\mu, \sigma^2)$ then $\frac{(w-\mu)^2}{\sigma^2} \sim \chi_1^2$. Using these two properties, a test of the null

$$H_0 : \mathbf{R}\beta - \mathbf{r} = 0$$

against the alternative

$$H_1 : \mathbf{R}\beta - \mathbf{r} \neq 0$$

can be constructed.

Following from Theorem 3.13, if $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$ is true, then

$$\sqrt{n}(\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} N(0, \mathbf{R}\Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}\mathbf{R}') \tag{3.70}$$

and

$$\Gamma^{-\frac{1}{2}}\sqrt{n}(\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} N(0, \mathbf{I}_k) \tag{3.71}$$

where $\Gamma = \mathbf{R}\Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}\mathbf{R}'$. Under the null that $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$,

$$n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' \left[\mathbf{R}\Sigma_{\mathbf{xx}}^{-1}\mathbf{S}\Sigma_{\mathbf{xx}}^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} \chi_m^2 \tag{3.72}$$

where m is the rank$(\mathbf{R})$. This estimator is not feasible since Γ is not known and must be estimated. Fortunately, Γ can be consistently estimated by applying the results of Theorem 3.14

$$\hat{\Sigma}_{\mathbf{xx}} = n^{-1}\mathbf{X}'\mathbf{X}$$

$$\hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i'\mathbf{x}_i$$

and so

$$\hat{\Gamma} = \hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1}.$$

The feasible Wald statistic is defined

$$W = n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' \left[\mathbf{R}\hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} \chi^2_m. \tag{3.73}$$

Test statistic values can be compared to the critical value C_α from a χ^2_m at the α-significance level and the null is rejected if W is greater than C_α. The asymptotic t-test (which has a normal distribution) is defined analogously,

$$t = \sqrt{n}\frac{\mathbf{R}\hat{\beta}_n - r}{\sqrt{\mathbf{R}\hat{\Gamma}\mathbf{R}'}} \xrightarrow{d} N(0,1), \tag{3.74}$$

where $\mathbf{R}$ is a 1 by k vector. Typically $\mathbf{R}$ is a vector with 1 in its j$^{\text{th}}$ element, producing statistic

$$t = \sqrt{n}\frac{\hat{\beta}_{jN}}{\sqrt{[\hat{\Gamma}]_{jj}}} \xrightarrow{d} N(0,1)$$

where $[\hat{\Gamma}]_{jj}$ is the j$^{\text{th}}$ diagonal element of $\hat{\Gamma}$.

The n term in the Wald statistic (or $\sqrt{n}$ in the t-test) may appear strange at first, although these terms are also present in the classical tests. Recall that the t-stat (null $H_0 : \beta_j = 0$) from the classical framework with homoskedastic data is given by

$$t_1 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}}.$$

The t-stat in the asymptotic framework is

$$t_2 = \sqrt{n}\frac{\hat{\beta}_{jN}}{\sqrt{\hat{\sigma}^2[\hat{\Sigma}_{\mathbf{xx}}^{-1}]_{jj}}}.$$

If t_1 is multiplied and divided by $\sqrt{n}$, then

$$t_1 = \sqrt{n}\frac{\hat{\beta}_j}{\sqrt{n}\sqrt{\hat{\sigma}^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} = \sqrt{n}\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}]_{jj}}} = \sqrt{n}\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[\hat{\Sigma}_{\mathbf{xx}}^{-1}]_{jj}}} = t_2,$$

and these two statistics have the same value since $\mathbf{X}'\mathbf{X}$ differs from $\hat{\Sigma}_{\mathbf{xx}}$ by a factor of n.

Algorithm 3.5 (Large-Sample Wald Test)**.**

1. *Estimate the unrestricted model $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.*
2. *Estimate the parameter covariance using $\hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1}$ where*

$$\hat{\Sigma}_{\mathbf{xx}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i, \;\; \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i$$

3. *Construct the restriction matrix, $\mathbf{R}$, and the value of the restriction, $\mathbf{r}$, from the null hypothesis.*
4. *Compute $W = n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' \left[\mathbf{R}\hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r})$.*
5. *Reject the null if $W > C_\alpha$ where C_α is the critical value from a χ^2_m using a size of α.*

3.11.2 Lagrange Multiplier Tests

Recall that the first-order conditions of the constrained estimation problem require

$$\mathbf{R}'\hat{\lambda} = 2\mathbf{X}'\tilde{\varepsilon}$$

where $\tilde{\varepsilon}$ are the residuals estimated under the null $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$. The LM test examines whether λ is close to zero. In the large-sample framework, $\hat{\lambda}$, like $\hat{\beta}$, is asymptotically normal and $\mathbf{R}'\hat{\lambda}$ will only be close to 0 if $\hat{\lambda} \approx 0$. The asymptotic version of the LM test can be compactly expressed if $\tilde{\mathbf{s}}$ is defined as the average score of the restricted estimator, $\tilde{\mathbf{s}} = n^{-1}\mathbf{X}'\tilde{\varepsilon}$. In this notation,

$$LM = n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m. \tag{3.75}$$

If the model is correctly specified, $n^{-1}\mathbf{X}'\tilde{\varepsilon}$, which is a k by 1 vector with j$^{\text{th}}$ element $n^{-1}\sum_{i=1}^{n} x_{j,i}\tilde{\varepsilon}_i$, should be a mean-zero vector with asymptotic variance $\mathbf{S}$ by assumption 3.7. Thus, $\sqrt{n}(n^{-1}\mathbf{X}'\tilde{\varepsilon}) \xrightarrow{d} N(0,\mathbf{S})$ implies

$$\sqrt{n}\mathbf{S}^{-\frac{1}{2}}\tilde{\mathbf{s}} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \tag{3.76}$$

and so $n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m$. This version is infeasible and the feasible version of the LM test must be used,

$$LM = n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m. \tag{3.77}$$

where $\tilde{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i$ is the estimator of the asymptotic variance *computed under the null*. This means that $\tilde{\mathbf{S}}$ is computed using the residuals from the restricted regression, $\tilde{\varepsilon}$, and that it will differ from the usual estimator $\hat{\mathbf{S}}$ which is computed using residuals from the unrestricted regression, $\hat{\varepsilon}$. Under the null, both $\tilde{\mathbf{S}}$ and $\hat{\mathbf{S}}$ are consistent estimators for $\mathbf{S}$ and using one or the other has no asymptotic effect.

If the residuals are homoskedastic, the LM test can also be expressed in terms of the R^2 of the unrestricted model when testing a null that the coefficients on all explanatory variables except the intercept are zero. Suppose the regression fit was

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{kn}.$$

To test the $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (where the excluded β_1 corresponds to a constant),

$$LM = n\mathrm{R}^2 \xrightarrow{d} \chi^2_k \tag{3.78}$$

is equivalent to the test statistic in eq. (3.77). This expression is useful as a simple tool to test whether the explanatory variables in a regression appear to explain *any* variation in the dependent variable. If the residuals are heteroskedastic, the $n\mathrm{R}^2$ form of the LM test does not have standard distribution and should not be used.

Algorithm 3.6 (Large-Sample LM Test)**.**

1. *Form the unrestricted model,* $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.
2. *Impose the null on the unrestricted model and estimate the restricted model,* $\tilde{Y}_i = \tilde{\mathbf{X}}_i\beta + \varepsilon_i$.

3. *Compute the residuals from the restricted regression,* $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$.

4. *Construct the score using the residuals from the restricted regression from both models,* $\tilde{\mathbf{s}}_i = \mathbf{x}_i\tilde{\varepsilon}_i$ *where* $\mathbf{x}_i$ *are the regressors from the unrestricted model.*

5. *Estimate the average score and the covariance of the score,*

$$\tilde{\mathbf{s}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \quad \tilde{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i'\tilde{\mathbf{s}}_i \tag{3.79}$$

6. *Compute the LM test statistic as* $LM = n\tilde{\mathbf{s}}\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}}'$.

7. *Reject the null if* $LM > C_\alpha$ *where* C_α *is the critical value from a* χ^2_m *using a size of* α.

3.11.3 Likelihood Ratio Tests

A critical distinction between small-sample and large-sample hypothesis testing is the omission of assumption 3.6. Without this assumption, the distribution of the errors is left unspecified. Based on the ease of implementing the Wald and LM tests their asymptotic framework, it may be tempting to think the likelihood ratio is asymptotically valid. It is not. The technical details are complicated, and the validity of the asymptotic distribution of the LR relies crucially on the Information Matrix Equality holding. If the shocks are heteroskedastic, then the IME will generally not hold, and the distribution of LR tests will be nonstandard.[14]

There is, however, a feasible likelihood-ratio like test available. The motivation for this test will be clarified in the GMM chapter. For now, the functional form will be given with only minimal explanation,

$$LR = n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m, \tag{3.80}$$

where $\tilde{\mathbf{s}} = n^{-1}\mathbf{X}'\tilde{\varepsilon}$ is the average score vector when the estimator is computed under the null. This statistic is similar to the LM test statistic, although there are two differences. First, one term has been left out of this expression, and the formal definition of the asymptotic LR is

$$LR = n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} - \hat{\mathbf{s}}'\mathbf{S}^{-1}\hat{\mathbf{s}} \xrightarrow{d} \chi^2_m \tag{3.81}$$

where $\hat{\mathbf{s}} = n^{-1}\mathbf{X}'\hat{\varepsilon}$ are the average scores from the *unrestricted* estimator. Recall from the first-order conditions of OLS (eq. (3.7)) that $\hat{\mathbf{s}} = \mathbf{0}$ and the second term in the general expression of the *LR* will always be zero. The second difference between *LR* and *LM* exists only in the feasible versions. The feasible version of the LR is given by

$$LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m. \tag{3.82}$$

where $\hat{\mathbf{S}}$ is estimated using the scores of the *unrestricted* model (under the alternative),

$$\hat{\mathbf{S}}^{-1} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i. \tag{3.83}$$

[14] In this case, the LR will converge to a weighted mixture of m independent χ^2_1 random variables where the weights are not 1. The resulting distribution is not a χ^2_m.

The feasible LM, $n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$, uses a covariance estimator ($\tilde{\mathbf{S}}$)based on the scores from the restricted model, $\tilde{\mathbf{s}}$.

In models with heteroskedasticity, it is impossible to determine *a priori* whether the LM or the LR test statistic will be larger, although folk wisdom states that LR test statistics are larger than LM test statistics (and hence the LR will be more powerful). If the data are homoskedastic, and homoskedastic estimators of $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$ are used ($\hat{\sigma}^2(\mathbf{X}'\mathbf{X}/n)^{-1}$ and $\tilde{\sigma}^2(\mathbf{X}'\mathbf{X}/n)^{-1}$, respectively), then it must be the case that $LM < LR$. This ordering of the two test statistic occurs since $\hat{\sigma}^2$ must be smaller than $\tilde{\sigma}^2$ because OLS minimizes the squared residuals. The LR is guaranteed to have more power in this case.

Algorithm 3.7 (Large-Sample LR Test)**.**

1. *Estimate the unrestricted model* $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.
2. *Impose the null on the unrestricted model and estimate the restricted model,* $\tilde{Y}_i = \tilde{\mathbf{X}}_i\beta + \varepsilon_i$.
3. *Compute the residuals from the restricted regression,* $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$, *and from the unrestricted regression,* $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$.
4. *Construct the score from both models,* $\tilde{\mathbf{s}}_i = \mathbf{x}_i\tilde{\varepsilon}_i$ *and* $\hat{\mathbf{s}}_i = \mathbf{x}_i\hat{\varepsilon}_i$, *where in both cases* $\mathbf{x}_i$ *are the regressors from the unrestricted model.*
5. *Estimate the average score and the covariance of the score,*

$$\tilde{\mathbf{s}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \quad \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\mathbf{s}}_i'\hat{\mathbf{s}}_i \tag{3.84}$$

6. *Compute the LR test statistic as* $LR = n\tilde{\mathbf{s}}\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}}'$.
7. *Reject the null if* $LR > C_\alpha$ *where* C_α *is the critical value from a* χ^2_m *using a size of* α.

3.11.4 Revisiting the Wald, LM, and LR tests

The previous tests can now be revisited while allowing for heteroskedasticity in the data. Tables 3.7 and 3.8 contain t-tests, Wald tests, LM tests, and LR tests that compare large-sample versions of these test statistics to their small-sample framework equivalents. There is a clear direction in the difference between the small-sample and large-sample test statistics: the large-sample statistics are smaller than the small-sample statistics, often substantially. Examining table 3.7, 4 out of 5 of the t-stats have decreased. Since the estimator of $\hat{\beta}$ is the same in both the small-sample and the large-sample frameworks, all of the difference is attributable to changes in the standard errors, which typically increased by 50%. When t-stats differ dramatically under the two covariance estimators, the likely cause is heteroskedasticity.

Table 3.8 shows that the Wald, LR, and LM test statistics also changed by large amounts.[15] The heteroskedasticity-robust Wald statistics decreased by up to a factor of 2, and the robust LM test statistics decreased by up to 5 times. The LR test statistic values were generally larger than those

[15]The statistics based on the small-sample assumptions have $f_{m,t-k}$ or $f_{m,t-k+m}$ distributions while the statistics based on the large-sample assumptions have χ^2_m distributions, and so the values of the small-sample statistics must be multiplied by m to be compared to the large-sample statistics.

		Homoskedasticity			Heteroskedasticity		
	$\hat{\beta}$	s.e.$(\hat{\beta})$	t-stat	p-value	s.e.$(\hat{\beta})$	t-stat	p-value
Constant	-0.086	0.042	-2.04	0.042	0.043	-1.991	0.046
VWM^e	1.080	0.010	108.7	0.000	0.012	93.514	0.000
SMB	0.002	0.014	0.13	0.893	0.017	0.110	0.912
HML	0.764	0.015	50.8	0.000	0.021	36.380	0.000
MOM	-0.035	0.010	-3.50	0.000	0.013	-2.631	0.009

Table 3.7: Comparing small and large-sample t-stats. The small-sample statistics in the left panel of the table overstate the precision of the estimates. The heteroskedasticity robust standard errors are larger for 4 out of 5 parameters, and one variable which was significant at the 15% level is insignificant.

of the corresponding Wald or LR test statistics. The relationship between the robust versions of the Wald and LR statistics is not clear, and for models that are grossly misspecified, the Wald and LR test statistics are substantially larger than their LM counterparts. However, when the value of the test statistics is smaller, the three are virtually identical, and the decision taken using any of these three tests is the same. All nulls except $H_0 : \beta_1 = \beta_3 = 0$ are rejected using standard sizes (5-10%).

These changes should serve as a warning to conducting inference using covariance estimates based on homoskedasticity. In most applications to financial time-series, heteroskedasticity robust covariance estimators (and often HAC (Heteroskedasticity and Autocorrelation Consistent), which will be defined in the time-series chapter) are automatically applied without testing for heteroskedasticity.

3.12 Violations of the Large-Sample Assumptions

The large-sample assumptions are just that: assumptions. While this set of assumptions is far more general than the finite-sample setup, they may be violated in a number of ways. This section examines the consequences of certain violations of the large-sample assumptions.

3.12.1 Omitted and Extraneous Variables

Suppose that the model is linear but misspecified, and a subset of the relevant regressors are excluded. The model can be specified

$$Y_i = \beta_1 \mathbf{X}_{1,i} + \beta_2 \mathbf{X}_{2,i} + \varepsilon_i \tag{3.85}$$

where $\mathbf{X}_{1,i}$ is 1 by k_1 vector of included regressors and $\mathbf{X}_{2,i}$ is a 1 by k_2 vector of excluded but relevant regressors. Omitting $\mathbf{x}_{2,i}$ from the fit model, the least-squares estimator is

$$\hat{\beta}_{1n} = \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n} \right)^{-1} \frac{\mathbf{X}_1'\mathbf{y}}{n}. \tag{3.86}$$

This misspecified estimator is biased, and the bias depends on the magnitude of the coefficients on the omitted variables and the correlation between the omitted and excluded regressors.

Wald Tests

			Small Sample		Large Sample	
Null	Alternative	M	W	p-value	W	p-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	4	3558.8	0.000	2661.2	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	3	956.5	0.000	583.2	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	2	10.1	0.000	7.35	0.001
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2	2.08	0.126	2.04	0.131
$\beta_5 = 0$	$\beta_5 \neq 0$	1	12.3	0.000	6.92	0.009

LR Tests

			Small Sample		Large Sample	
Null	Alternative	M	LR	p-value	LR	p-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	4	3558.8	0.000	2696.4	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	3	956.5	0.000	589.3	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	2	10.1	0.000	8.11	0.000
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2	2.08	0.126	2.13	0.119
$\beta_5 = 0$	$\beta_5 \neq 0$	1	12.3	0.000	7.40	0.007

LM Tests

			Small Sample		Large Sample	
Null	Alternative	M	LM	p-value	LM	p-value
$\beta_j = 0,\ j = 2,\ldots,5$	$\beta_j \neq 0,\ j = 2,\ldots,5$	4	163.4	0.000	34.8	0.000
$\beta_j = 0,\ j = 3,4,5$	$\beta_j \neq 0,\ j = 3,4,5$	3	184.3	0.000	31.9	0.000
$\beta_j = 0,\ j = 1,5$	$\beta_j \neq 0,\ j = 1,5$	2	9.85	0.000	7.82	0.000
$\beta_j = 0,\ j = 1,3$	$\beta_j \neq 0,\ j = 1,3$	2	2.07	0.127	2.11	0.121
$\beta_5 = 0$	$\beta_5 \neq 0$	1	12.1	0.001	6.50	0.011

Table 3.8: Comparing large- and small-sample Wald, LM, and LR test statistics. The large-sample test statistics are smaller than their small-sample counterparts due to the the heteroskedasticity present in the data. While the decisions of these tests are unaffected by the choice of covariance estimator, this will not always be the case.

Theorem 3.17 (Misspecified Regression)**.** *Under assumptions 3.1 and 3.7 - 3.9 through , if* $\mathbf{X}$ *can be partitioned* $[\mathbf{X}_1 \quad \mathbf{X}_2]$ *where* $\mathbf{X}_1$ *correspond to included variables while* $\mathbf{X}_2$ *correspond to excluded variables with non-zero coefficients, then*

$$\begin{aligned} \hat{\beta}_{1n} &\xrightarrow{p} \beta_1 + \Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\beta_2 \\ \hat{\beta}_1 &\xrightarrow{p} \beta_1 + \delta\beta_2 \end{aligned} \tag{3.87}$$

where

$$\Sigma_{\mathbf{xx}} = \begin{bmatrix} \Sigma_{\mathbf{x}_1\mathbf{x}_1} & \Sigma_{\mathbf{x}_1\mathbf{x}_2} \\ \Sigma'_{\mathbf{x}_1\mathbf{x}_2} & \Sigma_{\mathbf{x}_2\mathbf{x}_2} \end{bmatrix}$$

The bias term, $\delta\beta_2$ is composed of two elements. The first, δ, is a matrix of regression coefficients where the j$^{\text{th}}$ column is the probability limit of the least-squares estimator in the regression

$$\mathbf{X}_{2j} = \mathbf{X}_1\delta_j + \nu,$$

where $\mathbf{X}_{2j}$ is the j$^{\text{th}}$ column of $\mathbf{X}_2$. The second component of the bias term is the original regression coefficients. As should be expected, larger coefficients on omitted variables lead to larger bias.

$\hat{\beta}_{1n} \xrightarrow{p} \beta_1$ under one of three conditions:

1. $\hat{\delta}_n \xrightarrow{p} \mathbf{0}$
2. $\beta_2 = \mathbf{0}$
3. The product $\hat{\delta}_n\beta_2 \xrightarrow{p} \mathbf{0}$.

β_2 has been assumed to be non-zero (if $\beta_2 = \mathbf{0}$ the model is correctly specified). $\delta_n \xrightarrow{p} \mathbf{0}$ only if the regression coefficients of $\mathbf{X}_2$ on $\mathbf{X}_1$ are zero, which requires that the omitted and included regressors to be uncorrelated ($\mathbf{X}_2$ lies in the null space of $\mathbf{X}_1$). This assumption should be considered implausible in most applications and $\hat{\beta}_{1n}$ is biased and inconsistent, in general. Note that certain classes of regressors that are mutually orthogonal by design and can be safely omitted.[16] Finally, if both δ and β_2 are non-zero, the product could be zero, although, without a very peculiar specification and a careful selection of regressors, this possibility should be considered unlikely.

Alternatively, consider the case where some irrelevant variables are included. The correct model specification is

$$Y_i = \mathbf{X}_{1,i}\beta_1 + \varepsilon_i$$

and the model estimated is

$$Y_i = \mathbf{X}_{1,i}\beta_1 + \mathbf{X}_{2,i}\beta_2 + \varepsilon_i$$

As long as the assumptions of the asymptotic framework are satisfied, the least-squares estimator is consistent under theorem 3.12 and

$$\hat{\beta}_n \xrightarrow{p} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \mathbf{0} \end{bmatrix}$$

If the errors are homoskedastic, the variance of $\sqrt{n}(\hat{\beta}_n - \beta)$ is $\sigma^2\Sigma_{\mathbf{xx}}^{-1}$ where $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$. The variance of $\hat{\beta}_{1n}$ is the upper left k_1 by k_1 block of $\sigma^2\Sigma_{\mathbf{xx}}^{-1}$. Using the partitioned inverse,

$$\Sigma_{\mathbf{xx}}^{-1} = \begin{bmatrix} \Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1} + \Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\mathbf{M}_1\Sigma'_{\mathbf{x}_1\mathbf{x}_2}\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1} & -\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\mathbf{M}_1 \\ \mathbf{M}_1\Sigma'_{\mathbf{x}_1\mathbf{x}_2}\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1} & \Sigma_{\mathbf{x}_2\mathbf{x}_2}^{-1} + \Sigma_{\mathbf{x}_2\mathbf{x}_2}^{-1}\Sigma'_{\mathbf{x}_1\mathbf{x}_2}\mathbf{M}_2\Sigma_{\mathbf{x}_1\mathbf{x}_2}\Sigma_{\mathbf{x}_2\mathbf{x}_2}^{-1} \end{bmatrix}$$

[16] Safely in terms of consistency of estimated parameters. Omitting variables will cause the estimated variance to be inconsistent.

where

$$\mathbf{M}_1 = \lim_{n\to\infty} \frac{\mathbf{X}_2'\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2}{n}$$
$$\mathbf{M}_2 = \lim_{n\to\infty} \frac{\mathbf{X}_1'\mathbf{M}_{\mathbf{X}_2}\mathbf{X}_1}{n}$$

and so the upper left block of the variance, $\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1} + \Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\mathbf{M}_1\Sigma_{\mathbf{x}_1\mathbf{x}_2}'\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}$, *must* be larger than $\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}$ because the second term is a quadratic form and $\mathbf{M}_1$ is positive semi-definite.[17] Noting that $\hat{\sigma}^2$ is consistent under both the correct specification and the expanded specification, the cost of including extraneous regressors is an increase in the asymptotic variance.

In finite samples, there is a bias-variance trade-off. Fewer regressors included in a model leads to more precise estimates. Models containing more variables tend to produce coefficient estimated with less bias. Additionally, if relevant variables are omitted then $\hat{\sigma}^2$ is larger than it would be if all relevant variables are included, and so the estimated parameter variance, $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ is also larger. Asymptotically, only the bias remains as it is of a higher order than variance (scaling $\hat{\beta}_n - \beta$ by $\sqrt{n}$, the bias is exploding while the variance is constant), and so when the sample size is large and estimates are precise, a larger model should be preferred to a smaller model. In cases where the sample size is small, there is a justification for omitting a variable to enhance the precision of those remaining, particularly when the effect of the omitted variable is not of interest or when the excluded variable is highly correlated with one or more included variables.

3.12.2 Errors Correlated with Regressors

Bias can arise from sources other than omitted variables. Consider the case where $\mathbf{X}$ is measured with noise and define $\tilde{\mathbf{X}}_i = \mathbf{X}_i + \eta_i$ where $\tilde{\mathbf{X}}_i$ is a noisy proxy for $\mathbf{X}_i$, the "true" (unobserved) regressor, and η_i is an i.i.d. mean 0 noise process which is independent of $\mathbf{X}$ and ε with finite second moments $\Sigma_{\eta\eta}$. The OLS estimator,

$$\hat{\beta}_n = \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{n}\right)^{-1}\frac{\tilde{\mathbf{X}}'\mathbf{y}}{n} \tag{3.88}$$
$$= \left(\frac{(\mathbf{X}+\eta)'(\mathbf{X}+\eta)}{n}\right)^{-1}\frac{(\mathbf{X}+\eta)'\mathbf{y}}{n} \tag{3.89}$$
$$= \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\eta}{n} + \frac{\eta'\mathbf{X}}{n} + \frac{\eta'\eta}{n}\right)^{-1}\frac{(\mathbf{X}+\eta)'\mathbf{y}}{n} \tag{3.90}$$
$$= \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\eta}{n} + \frac{\eta'\mathbf{X}}{n} + \frac{\eta'\eta}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{y}}{n} + \frac{\eta'\mathbf{y}}{n}\right) \tag{3.91}$$

will be biased downward. To understand the source of the bias, consider the behavior, under the asymptotic assumptions, of

[17] Both $\mathbf{M}_1$ and $\mathbf{M}_2$ are covariance matrices of the residuals of regressions of $\mathbf{x}_2$ on $\mathbf{x}_1$ and $\mathbf{x}_1$ on $\mathbf{x}_2$ respectively.

$$\begin{aligned}
\frac{\mathbf{X}'\mathbf{X}}{n} &\xrightarrow{p} \Sigma_{\mathbf{XX}} \\
\frac{\mathbf{X}'\eta}{n} &\xrightarrow{p} \mathbf{0} \\
\frac{\eta'\eta}{n} &\xrightarrow{p} \Sigma_{\eta\eta} \\
\frac{\mathbf{X}'\mathbf{y}}{n} &\xrightarrow{p} \Sigma_{\mathbf{XX}}\beta \\
\frac{\eta'\mathbf{y}}{n} &\xrightarrow{p} \mathbf{0}
\end{aligned}$$

so

$$\left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\eta}{n} + \frac{\eta'\mathbf{X}}{n} + \frac{\eta'\eta}{n}\right)^{-1} \xrightarrow{p} (\Sigma_{\mathbf{XX}} + \Sigma_{\eta\eta})^{-1}$$

and

$$\hat{\beta}_n \xrightarrow{p} (\Sigma_{\mathbf{XX}} + \Sigma_{\eta\eta})^{-1}\Sigma_{\mathbf{XX}}\beta.$$

If $\Sigma_{\eta\eta} \neq \mathbf{0}$, then $\hat{\beta}_n \xrightarrow{p} \beta$ and the estimator is inconsistent.

The OLS estimator is also biased in the case where $n^{-1}\mathbf{X}'\varepsilon \xrightarrow{p} \mathbf{0}_k$, which arises in situations with *endogeneity*. In these cases, $\mathbf{x}_i$ and ε_i are simultaneously determined and correlated. This correlation results in a biased estimator since $\hat{\beta}_n \xrightarrow{p} \beta + \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}\varepsilon}$ where $\Sigma_{\mathbf{X}\varepsilon}$ is the limit of $n^{-1}\mathbf{X}'\varepsilon$. The classic example of endogeneity is simultaneous equation models although many situations exist where the innovation may be correlated with one or more regressors; omitted variables can be considered a special case of endogeneity by reformulating the model.

The solution to this problem is to find an instrument, $\mathbf{z}_i$, which is correlated with the endogenous variable, $\mathbf{x}_i$, but uncorrelated with ε_i. Intuitively, the endogenous portions of $\mathbf{x}_i$ can be annihilated by regressing $\mathbf{x}_i$ on $\mathbf{z}_i$ and using the fit values. This procedure is known as instrumental variable (IV) regression in the case where the number of $\mathbf{z}_i$ variables is the same as the number of $\mathbf{x}_i$ variables and two-stage least squares (2SLS) when the size of $\mathbf{z}_i$ is larger than k.

Define $\mathbf{z}_i$ as a vector of exogenous variables where $\mathbf{z}_i$ may contain any of the variables in $\mathbf{x}_i$ which are exogenous. However, all endogenous variables – those correlated with the error – must be excluded.

First, a few assumptions must be reformulated.

Assumption 3.11 (IV Stationary Ergodicity)**.** $\{(\mathbf{Z}_i, \mathbf{X}_i, \varepsilon_i)\}$ *is a strictly stationary and ergodic sequence.*

Assumption 3.12 (IV Rank)**.** $\mathrm{E}[\mathbf{Z}_i'\mathbf{X}_i] = \Sigma_{\mathbf{ZX}}$ *is nonsingular and finite.*

Assumption 3.13 (IV Martingale Difference)**.** $\{\mathbf{Z}_i'\varepsilon_i, \mathcal{F}_i\}$ *is a martingale difference sequence,*

$$\mathrm{E}\left[\left(Z_{j,i}\varepsilon_i\right)^2\right] < \infty,\ j = 1, 2, \ldots, k,\ i = 1, 2 \ldots$$

and $\mathbf{S} = \mathrm{V}[n^{-\frac{1}{2}}\mathbf{Z}'\varepsilon]$ *is finite and non singular.*

Assumption 3.14 (IV Moment Existence)**.** $\mathrm{E}[X_{ji}^4] < \infty$ *and* $\mathrm{E}[Z_{ji}^4] < \infty$, $j = 1,2,\ldots,k$, $i = 1,2,\ldots$ *and* $\mathrm{E}[\varepsilon_i^2] = \sigma^2 < \infty$, $i = 1,2,\ldots$.

These four assumptions are nearly identical to the four used to establish the asymptotic normality of the OLS estimator. The IV estimator is defined

$$\hat{\beta}_n^{\mathrm{IV}} = \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \frac{\mathbf{Z}'\mathbf{y}}{n} \tag{3.92}$$

where the n term is present to describe the number of observations used in the IV estimator. The asymptotic properties are easy to establish and are virtually identical to those of the OLS estimator.

Theorem 3.18 (Consistency of the IV Estimator)**.** *Under assumptions 3.1 and 3.11-3.13, the IV estimator is consistent,*

$$\hat{\beta}_n^{\mathrm{IV}} \xrightarrow{p} \beta$$

and asymptotically normal

$$\sqrt{n}(\hat{\beta}_n^{\mathrm{IV}} - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{ZX}}^{-1}\ddot{\mathbf{S}}\Sigma_{\mathbf{ZX}}^{-1}) \tag{3.93}$$

where $\Sigma_{\mathbf{ZX}} = \mathrm{E}[\mathbf{x}_i'\mathbf{z}_i]$ *and* $\ddot{\mathbf{S}} = \mathrm{V}[n^{-1/2}\mathbf{Z}'\boldsymbol{\varepsilon}]$.

Additionally, consistent estimators are available for the components of the asymptotic variance.

Theorem 3.19 (Asymptotic Normality of the IV Estimator)**.** *Under assumptions 3.1 and 3.11 - 3.14,*

$$\hat{\Sigma}_{\mathbf{ZX}} = n^{-1}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \Sigma_{\mathbf{ZX}} \tag{3.94}$$

$$\hat{\ddot{\mathbf{S}}} = n^{-1}\sum_{i=1}^{n} \varepsilon_i^2\mathbf{z}_i'\mathbf{z}_i \xrightarrow{p} \ddot{\mathbf{S}} \tag{3.95}$$

and

$$\hat{\Sigma}_{\mathbf{ZX}}^{-1}\hat{\ddot{\mathbf{S}}}\hat{\Sigma}_{\mathbf{ZX}}^{\prime-1} \xrightarrow{p} \Sigma_{\mathbf{ZX}}^{-1}\ddot{\mathbf{S}}\Sigma_{\mathbf{ZX}}^{\prime-1} \tag{3.96}$$

The asymptotic variance can be easily computed from

$$\begin{aligned}\hat{\Sigma}_{\mathbf{ZX}}^{-1}\hat{\ddot{\mathbf{S}}}\hat{\Sigma}_{\mathbf{ZX}}^{-1} =& n\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{z}_i'\mathbf{z}_i\right)\left(\mathbf{X}'\mathbf{Z}\right)^{-1} \\ =& n\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\hat{\mathbf{E}}\mathbf{Z}\right)\left(\mathbf{X}'\mathbf{Z}\right)^{-1}\end{aligned} \tag{3.97}$$

where $\hat{\mathbf{E}} = \mathrm{diag}(\hat{\varepsilon}_1^2,\ldots,\hat{\varepsilon}_n^2)$ is a matrix with the estimated residuals squared along its diagonal.

IV estimators have one further complication beyond those of OLS. Assumption 3.8 requires the rank of $\mathbf{Z}'\mathbf{X}$ to be full (k), and so $\mathbf{z}_i$ must be correlated with $\mathbf{x}_i$. Moreover, since the asymptotic variance depends on $\Sigma_{\mathbf{ZX}}^{-1}$, even variables with non-zero correlation may produce imprecise estimates, especially if the correlation is low. Instruments must be carefully chosen, although substantially deeper treatment is beyond the scope of this course. Fortunately, IV estimators are infrequently needed in financial econometrics.

3.12.3 Monte Carlo: The effect of instrument correlation

While IV estimators are not often needed with financial data[18], the problem of endogeneity is severe and it is important to be aware of the consequences and pitfalls of using IV estimators.[19] To understand this problem, consider a simple Monte Carlo. The regressor (X_i), the instrument (Z_i) and the error are all drawn from a multivariate normal with the covariance matrix,

$$\begin{bmatrix} X_i \\ Z_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \rho_{xz} & \rho_{x\varepsilon} \\ \rho_{xz} & 1 & 0 \\ \rho_{x\varepsilon} & 0 & 1 \end{bmatrix}\right).$$

Throughout the experiment, $\rho_{x\varepsilon} = 0.4$ and ρ_{xz} is varied from 0 to .9. 200 data points were generated from

$$Y_i = \beta_1 X_i + \varepsilon_i$$

where $\beta_1 = 1$. It is straightforward to show that $\mathrm{E}[\hat{\beta}] = 1 + \rho_{x\varepsilon}$ and that $\hat{\beta}_n^{\mathrm{IV}} \xrightarrow{p} 1$ as long as $\rho_{xz} \neq 0$. 10,000 replications were generated and the IV estimators were computed

$$\hat{\beta}_n^{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}).$$

Figure 3.5 contains kernel density plots of the instrumental variable estimator for ρ_{xz} of .2, .4, .6 and .8. When the correlation between the instrument and X is low, the distribution is dispersed (exhibiting a large variance). As the correlation increases, the variance decreases and the distribution become increasingly normal. This experiment highlights two fundamental problems with IV estimators: they have large variance when no "good instruments" – highly correlated with $\mathbf{x}_i$ by uncorrelated with ε_i – are available and the finite-sample distribution of IV estimators may be poorly approximated a normal.

3.12.4 Heteroskedasticity

Assumption 3.7 does not require data to be homoskedastic, which is useful since heteroskedasticity is the rule rather than the exception in financial data. If the data are homoskedastic, the asymptotic covariance of $\hat{\beta}$ can be consistently estimated by

$$\hat{\mathbf{S}} = \hat{\sigma}^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$$

Heteroskedastic errors require the use of a more complicated covariance estimator, and the asymptotic variance can be consistently estimated using

[18] IV estimators are most common in corporate finance when examining executive compensation and company performance.

[19] The intuition behind IV estimators is generally applicable to 2SLS.

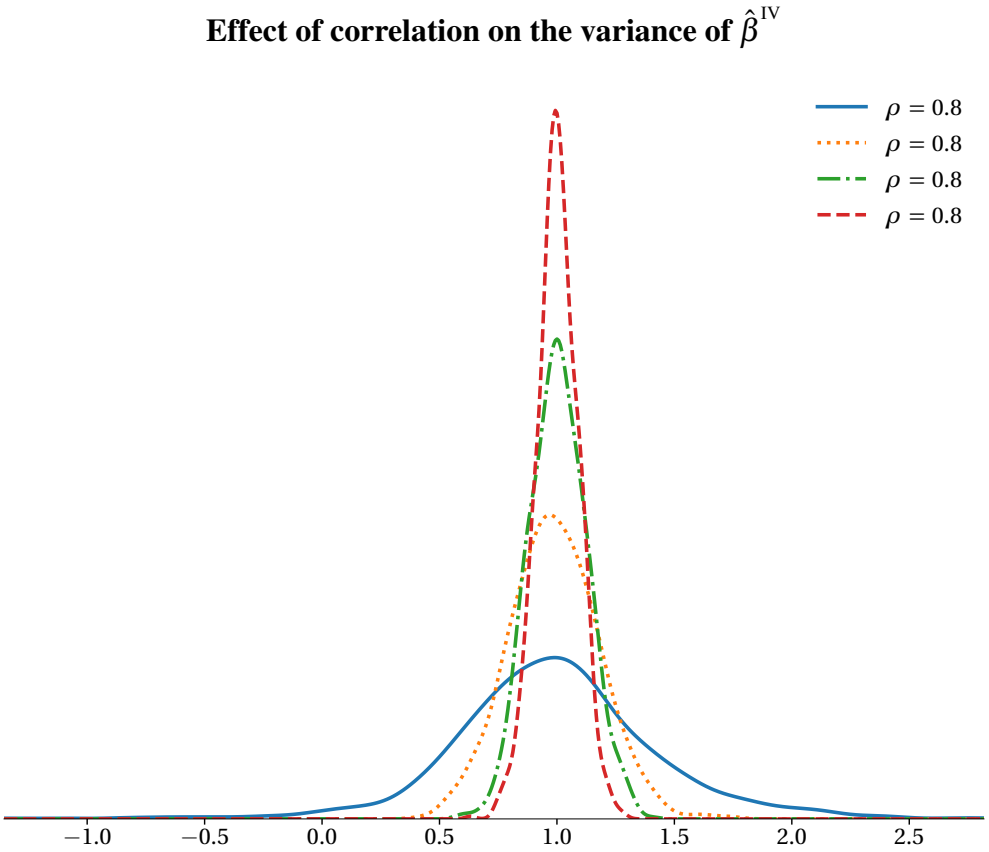

Figure 3.5: Kernel density of the instrumental variable estimator $\hat{\beta}_n^{IV}$ with varying degrees of correlation between the endogenous variable and the instrument. Increasing the correlation between the instrument and the endogenous variable leads to a large decrease in the variance of the estimated parameter ($\beta = 1$). When the correlation is small (.2), the distribution has a large variance and is not well approximated by a normal random variable.

$$\begin{aligned}\hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{xx}}^{-1} &= \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \\ &= n\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= n\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned} \tag{3.98}$$

where $\hat{\mathbf{E}} = \text{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)$ is a matrix with the estimated residuals squared along its diagonal.

Faced with two covariance estimators, one which is consistent under minimal assumptions and one which requires an additional, often implausible assumption, it may be tempting use rely exclusively on the robust estimator. This covariance estimator is known as the White heteroskedasticity consistent covariance estimator and standard errors computed using eq. (3.98) are called heteroskedasticity robust standard errors or White standard errors (White, 1980). Using a heteroskedasticity-consistent

estimator when not needed (homoskedastic data) results in test statistics that have worse small-sample properties. In small samples, hypothesis tests are more likely to have size distortions and so using 5% critical values may lead to rejection of the null 10% or more of the time when the null is true. On the other hand, using an inconsistent estimator of the parameter covariance – assuming homoskedasticity when the data are not – produces tests with size distortions, even asymptotically.

White (1980)also provides a test to determine if a heteroskedasticity robust covariance estimator is required. Each term in the heteroskedasticity-consistent estimator takes the form

$$\varepsilon_i^2 \mathbf{x}_i' \mathbf{x}_i = \begin{bmatrix} \varepsilon_i^2 x_{1,i}^2 & \varepsilon_i^2 x_{1,i} x_{2,i} & \dots & \varepsilon_i^2 x_{1,i} x_{kn} \\ \varepsilon_i^2 x_{1,i} x_{2,i} & \varepsilon_i^2 x_{2,i}^2 & \dots & \varepsilon_i^2 x_{2,i} x_{kn} \\ \vdots & \vdots & \dots & \vdots \\ \varepsilon_i^2 x_{1,i} x_{kn} & \varepsilon_i^2 x_{2,i} x_{kn} & \dots & \varepsilon_i^2 x_{kn}^2 \end{bmatrix},$$

and so, if $\mathrm{E}[\varepsilon_i^2 x_{jn} x_{ln}] = \mathrm{E}[\varepsilon_i^2]\mathrm{E}[x_{jn} x_{ln}]$, for all j and l, then the heteroskedasticity robust and the standard estimator will both consistently estimate the asymptotic variance of $\hat{\beta}$. White's test is formulated as a regression of *squared* estimated residuals on all unique squares and cross products of $\mathbf{x}_i$. Suppose the original regression specification is

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \varepsilon_i.$$

White's test uses an auxiliary regression of $\hat{\varepsilon}_i^2$ on the squares and cross-produces of all regressors, $\{1, X_{1,i}, X_{2,i}, X_{1,i}^2, X_{2,i}^2, X_{1,i} X_{2,i}\}$:

$$\hat{\varepsilon}_i^2 = \delta_1 + \delta_2 X_{1,i} + \delta_3 X_{2,i} + \delta_4 X_{1,i}^2 + \delta_5 X_{2,i}^2 + \delta_6 X_{1,i} X_{2,i} + \eta_i. \tag{3.99}$$

The null hypothesis tested is $H_0 : \delta_j = 0$, $j > 1$, and the test statistic can be computed using $n\mathrm{R}^2$ where the centered R^2 is from the model in eq. (3.99). Recall that $n\mathrm{R}^2$ is an LM test of the null that all coefficients except the intercept are zero and has an asymptotic χ_ν^2 where ν is the number of restrictions – the same as the number of regressors excluding the constant. If the null is rejected, a heteroskedasticity robust covariance estimator is required.

Algorithm 3.8 (White's Test)**.**

1. *Fit the model* $Y_i = \mathbf{X}_i \beta + \varepsilon_i$
2. *Construct the fit residuals* $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i \hat{\beta}$
3. *Construct the auxiliary regressors* $\mathbf{Z}_i$ *where the* $k(k+1)/2$ *elements of* $\mathbf{z}_i$ *are computed from* $X_{i,o} X_{i,p}$ *for* $o = 1, 2, \dots, k$, $p = o, o+1, \dots, k$.
4. *Estimate the auxiliary regression* $\hat{\varepsilon}_i^2 = \mathbf{Z}_i \gamma + \eta_i$
5. *Compute White's Test statistic as* nR^2 *where the* R^2 *is from the auxiliary regression and compare to the critical value at size* α *from a* $\chi_{k(k+1)/2-1}^2$.

3.12.5 Example: White's test on the FF data

White's heteroskedasticity test is implemented using the estimated residuals, $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i'\hat{\beta}$, by regressing the estimated residuals squared on all unique cross products of the regressors. The primary model fit is

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i.$$

and the auxiliary model is specified

$$\begin{aligned}\hat{\varepsilon}_i^2 = \delta_1 &+ \delta_2 VWM_i^e + \delta_3 SMB_i + \delta_4 HML_i + \delta_5 MOM_i + \delta_6 \left(VWM_i^e\right)^2 + \delta_7 VWM_i^e SMB_i \\ &+ \delta_8 VWM_i^e HML_i + \delta_9 VWM_i^e MOM_i + \delta_{10} SMB_i^2 + \delta_1 1 SMB_i HML_i \\ &+ \delta_{12} SMB_i MOM_i + \delta_{13} HML_i^2 + \delta_{14} HML_i MOM_i + \delta_{15} MOM_i^2 + \eta_i\end{aligned}$$

Estimating this regression produces an R^2 of 10.9% and $nR^2 = 74.8$, which has an asymptotic χ^2_{14} distribution (14 regressors, excluding the constant). The p-value of this test statistic is 0.000, and the null of homoskedasticity is strongly rejected.

3.12.6 Generalized Least Squares

An alternative to modeling heteroskedastic data is to transform the data so that it is homoskedastic using generalized least squares (GLS). GLS extends OLS to allow for arbitrary weighting matrices. The GLS estimator of β is defined

$$\hat{\beta}^{\text{GLS}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}, \tag{3.100}$$

for some positive definite matrix $\mathbf{W}$. Without any further assumptions or restrictions on $\mathbf{W}$, $\hat{\beta}^{\text{GLS}}$ is unbiased under the same conditions as $\hat{\beta}$, and the variance of $\hat{\beta}$ can be easily shown to be

$$(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$$

where $\mathbf{V}$ is the n by n covariance matrix of ε.

The full value of GLS is only realized when $\mathbf{W}$ is wisely chosen. Suppose that the data are heteroskedastic but not serial correlated,[20] and so

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{3.101}$$

where $\mathrm{V}[\varepsilon_i|\mathbf{X}] = \sigma_i^2$ and therefore heteroskedastic. Further, assume σ_i^2 is known. Returning to the small-sample assumptions, choosing $\mathbf{W} \propto \mathrm{V}(\varepsilon|\mathbf{X})$[21], the GLS estimator will be efficient.

Assumption 3.15 (Error Covariance)**.** $\mathbf{V} = \mathrm{V}[\varepsilon|\mathbf{X}]$

Setting $\mathbf{W} = \mathbf{V}$, the GLS estimator is BLUE.

[20] Serial correlation is ruled out by assumption 3.9.

[21] $\propto$ is the mathematical symbol for "proportional to".

Theorem 3.20 (Variance of $\hat{\beta}^{\text{GLS}}$)**.** *Under assumptions 3.1 - 3.3 and 3.15,*

$$\mathrm{V}[\hat{\beta}^{\text{GLS}}|\mathbf{X}] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

and $\mathrm{V}[\hat{\beta}^{\text{GLS}}|\mathbf{X}] \leq \mathrm{V}[\tilde{\beta}|\mathbf{X}]$ *where* $\tilde{\beta} = \mathbf{C}\mathbf{y}$ *is any other linear unbiased estimator with* $\mathrm{E}[\tilde{\beta}] = \beta$

To understand the intuition behind this result, note that the GLS estimator can be expressed as an OLS estimator using transformed data. Returning to the model in eq. (3.101), and pre-multiplying by $\mathbf{W}^{-\frac{1}{2}}$,

$$\begin{aligned}\mathbf{W}^{-\frac{1}{2}}\mathbf{y} &= \mathbf{W}^{-\frac{1}{2}}\mathbf{X}\beta + \mathbf{W}^{-\frac{1}{2}}\varepsilon \\ \tilde{\mathbf{y}} &= \tilde{\mathbf{X}}\beta + \tilde{\varepsilon}\end{aligned}$$

and so

$$\begin{aligned}\hat{\beta} &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= \left(\mathbf{X}'\mathbf{W}^{-\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\right)\mathbf{X}'\mathbf{W}^{-\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \\ &= \hat{\beta}^{\text{GLS}}.\end{aligned}$$

In the original model, $\mathbf{W} = \mathrm{V}[\varepsilon|\mathbf{X}]$, and so $\mathrm{V}[\mathbf{W}^{-\frac{1}{2}}\varepsilon|\mathbf{X}] = \mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}} = \mathbf{I}_n$. $\tilde{\varepsilon}$ is homoskedastic and uncorrelated and the transformed model satisfies the assumption of the Gauss-Markov theorem (theorem 3.3).

This result is only directly applicable under the small-sample assumptions and then only if $\mathrm{V}[\varepsilon|\mathbf{X}]$ is known *a priori*. In practice, neither is true: data are not congruent with the small-sample assumptions and $\mathrm{V}[\varepsilon|\mathbf{X}]$ is never known. The feasible GLS (FGLS) estimator solves these two issues, although the efficiency gains of FGLS have only asymptotic justification. Suppose that $\mathrm{V}[\varepsilon|\mathbf{X}] = \omega_1 + \omega_2 x_{1,i} + \ldots + \omega_{k+1}x_{kn}$ where ω_j are unknown. The FGLS procedure provides a method to estimate these parameters and implement a feasible GLS estimator.

The FGLS procedure is described in the following algorithm.

Algorithm 3.9 (Feasible GLS Estimation)**.**

1. *Estimate* $\hat{\beta}$ *using OLS.*
2. *Using the estimated residuals,* $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$*, estimate an auxiliary model by regressing the squared residual on the variables of the variance model.*
3. *Using the estimated variance model parameters* $\hat{\omega}$*, produce a fit variance matrix,* $\widehat{\mathbf{V}}$*.*
4. *Compute* $\tilde{\mathbf{y}} = \widehat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{y}$ *and* $\tilde{\mathbf{X}} = \widehat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{X}$ *compute* $\hat{\beta}^{\text{FGLS}}$ *using the OLS estimator on the transformed regressors and regressand.*

Hypothesis testing can be performed on $\hat{\beta}^{\text{FGLS}}$ using the standard test statistics with the FGLS variance estimator,

$$\tilde{\sigma}^2(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} = \tilde{\sigma}^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$$

where $\tilde{\sigma}^2$ is the sample variance of the FGLS regression errors ($\tilde{\varepsilon} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}^{\text{FGLS}}$).

While FGLS is only formally asymptotically justified, FGLS estimates are often much more precise in finite samples, especially if the data is very heteroskedastic. Estimator accuracy improves the most when some observations have a vastly larger variance than others. The OLS estimator gives these observations too much weight, inefficiently exploiting the information in the remaining observations. FGLS, even when estimated with a diagonal weighting matrix that may be slightly misspecified, can produce substantially more precise estimates.[22]

3.12.6.1 Monte Carlo: A simple GLS

A simple Monte Carlo was designed to demonstrate the gains of GLS. The observed data are generated according to

$$Y_i = X_i + X_i^{\alpha}\varepsilon_i$$

where X_i is i.i.d. U(0,1) and ε_i is standard normal. α takes the values of 0.8, 1.6, 2.8 and 4. When α is low the data are approximately homoskedastic. As α increases the data are increasingly heteroskedastic and the probability of producing a few residuals with *small* variances increases. The OLS and (infeasible) GLS estimators were fit to the data and figure 3.6 contains kernel density plots of $\hat{\beta}$ and $\hat{\beta}^{GLS}$.

When α is small, the OLS and GLS parameter estimates have similar variances, indicated by the similarity in distribution. As α increases, the GLS estimator becomes very precise which is due to GLS's reweighing of the data by the inverse of its variance. In effect, observations with the smallest errors become very influential in determining $\hat{\beta}$. This is the general principle behind GLS: let the data points which are most precise about the unknown parameters have the most influence.

3.12.7 Example: GLS in the Factor model

Even if it is unreasonable to assume that the entire covariance structure of the residuals can be correctly specified in the auxiliary regression, GLS estimates are often much more precise than OLS estimates. Consider the regression of BH^e on the four factors and a constant. The OLS estimates are identical to those previously presented and the GLS estimates will be computed using the estimated variances from White's test. Define

$$\hat{\mathbf{V}} = \text{diag}\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_n^2\right)$$

where $\hat{\sigma}_i^2$ is the fit value from the auxiliary regression in White's test that included only the squares of the explanatory variables. Coefficients were estimated by regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}$ where

$$\tilde{\mathbf{y}} = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{y}$$

[22] If the model for the conditional variance of ε_i is misspecified in an application of FGLS, the resulting estimator is not asymptotically efficient and a heteroskedasticity robust covariance estimator is required.

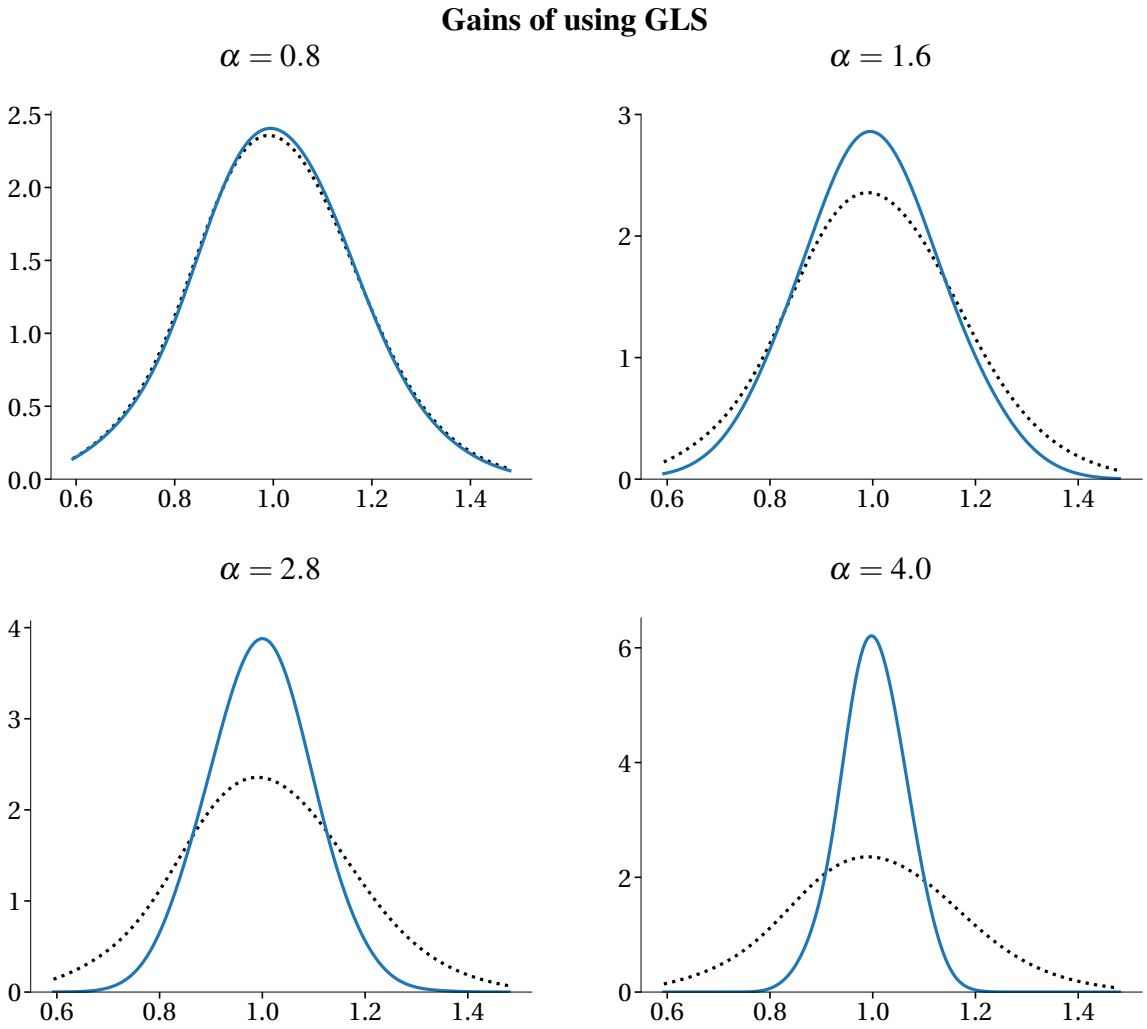

Figure 3.6: The four plots show the gains to using the GLS estimator on heteroskedastic data. The data were generated according to $Y_i = X_i + X_i^{\alpha}\varepsilon_i$ where X_i is i.i.d. uniform and ε_i is standard normal. For large α, the GLS estimator is substantially more efficient than the OLS estimator. However, the intuition behind the result is not that high variance residuals have been down-weighted, but that low variance residuals, some with very low variances, have been up-weighted to produce an accurate fit.

$$\tilde{\mathbf{X}} = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{X}$$

and $\hat{\beta}^{GLS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$. $\hat{\varepsilon}^{GLS} = \mathbf{y} - \mathbf{X}\hat{\beta}^{GLS}$ are computed from the original data using the GLS estimate of β, and the variance of the GLS estimator can be computed using

$$(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}'\widehat{\tilde{\mathbf{E}}}\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}.$$

where $\widehat{\tilde{\mathbf{E}}}$ is a diagonal matrix with the estimated residuals squared, $\left(\hat{\varepsilon}_i^{GLS}\right)^2$, from the GLS procedure along its diagonal. Table 3.9 contains the estimated parameters, t-stats and p-values using both the

	OLS				GLS			
	$\hat{\beta}$	s.e.$(\hat{\beta})$	t-stat	p-values	$\hat{\beta}^{GLS}$	s.e.$(\hat{\beta}^{GLS})$	t-stats	p-values
Constant	-0.09	0.04	-1.99	0.05	-0.09	0.04	-2.26	0.02
VWM^e	1.08	0.01	93.5	0.00	1.08	0.01	101.6	0.00
SMB	0.00	0.02	0.11	0.91	-0.00	0.02	-0.19	0.85
HML	0.76	0.02	36.4	0.00	0.73	0.02	39.3	0.00
MOM	-0.04	0.01	-2.63	0.01	-0.04	0.01	-3.06	0.00

Table 3.9: OLS and GLS parameter estimates and t-stats. t-stats indicate that the GLS parameter estimates are more precise.

OLS and the GLS estimates. The GLS estimation procedure appears to provide more precise estimates and inference. The difference in precision is particularly large for *SMB*.

3.13 Model Selection and Specification Checking

Econometric problems often begin with a variable whose dynamics are of interest and a relatively large set of candidate explanatory variables. The process by which the set of regressors is reduced is known as model selection or building.

Model building inevitably reduces to balancing two competing considerations: congruence and parsimony. A congruent model is one that captures all of the variation in the data explained by the regressors. Obviously, including all of the regressors and all functions of the regressors should produce a congruent model. However, this is also an infeasible procedure since there are infinitely many functions of even a single regressor. Parsimony dictates that the model should be as simple as possible and so models with fewer regressors are favored. The ideal model is the *parsimonious congruent* model that contains all variables necessary to explain the variation in the regressand and nothing else.

Model selection is as much a black art as science and some lessons can only be taught through experience. One principle that should be universally applied when selecting a model is to rely on economic theory and, failing that, common sense. The simplest method to select a poorly performing model is to try any and all variables, a process known as data snooping that is capable of producing a model with an arbitrarily high R^2 even if there is no relationship between the regressand and the regressors.

There are a few variable selection methods which can be examined for their properties. These include:

- General to Specific modeling (GtS)
- Specific to General modeling (StG)
- Information criteria (IC)
- Cross-validation

3.13.1 Model Building

3.13.1.1 General to Specific

General to specific (GtS) model building begins by estimating the largest model that can be justified by economic theory (and common sense). This model is then pared down to produce the smallest model that remains congruent with the data. The simplest version of GtS begins with the complete model. If any coefficients have individual p-values less than some significance level α (usually 5 or 10%), the least significant regressor is dropped from the regression. The procedure is repeated using the remaining included regressors until all coefficients are statistically significant. In each step, the least significant regressor is removed from the model.

One drawback to this simple procedure is that variables that are correlated but relevant are often dropped. This is due to a problem known as multicollinearity and individual t-stats will be small but joint significance tests that all coefficients are simultaneously zero will strongly reject. This suggests using joint hypothesis tests to pare the general model down to the specific one. While theoretically attractive, the scope the of possible joint hypothesis tests is vast even in a small model, and so using joint test is impractical.

GtS suffers from two additional issues. First, it will include an irrelevant variable with positive probability (asymptotically) but will never exclude a relevant variable. Second, test statistics do not have standard distributions when they are used sequentially (as is the case with any sequential model building procedure). The only viable solution to the second problem is to fit a single model, make variable inclusions and exclusion choices, and live with the result. This practice is not typically followed and most econometricians use an iterative procedure despite the problems of sequential testing.

3.13.1.2 Specific to General

Specific to General (StG) model building begins by estimating the smallest model, usually including only a constant. Variables are then added sequentially based on maximum t-stat until there is no excluded variable with a significant t-stat at some predetermined α (again, usually 5 or 10%). StG suffers from the same issues as GtS. First it will asymptotically include all relevant variables and some irrelevant ones and second, tests implemented sequentially do not have correct size. Choosing between StG and GtS is mainly user preference, although they rarely select the same model. One argument in favor of using a GtS approach is that the variance is consistently estimated in the first step of the general specification while the variance estimated in the first step of the an StG selection is too large. The leads StG processes to have t-stats that are smaller than GtS t-stats and so StG generally selects a smaller model than GtS.

3.13.1.3 Information Criteria

The third method of model selection uses Information Criteria (IC). Information Criteria reward the model for producing smaller SSE while punishing it for the inclusion of additional regressors. The two most frequently used are the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC) or Bayesian Information Criterion (BIC).[23] Most Information Criteria are of the form

[23]The BIC and SIC are the same. BIC is probably the most common name but SIC or S/BIC are also frequently encountered.

$$-2l + P$$

where l is the log-likelihood value at the parameter estimates and P is a penalty term. In the case of least squares, where the log-likelihood is not known (or needed), IC's take the form

$$\ln \hat{\sigma}^2 + P$$

where the penalty term is divided by n.

Definition 3.15 (Akaike Information Criterion (AIC))**.** For likelihood-based models the AIC is defined

$$AIC = -2l + 2k \tag{3.102}$$

and in its least squares application,

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{n} \tag{3.103}$$

Definition 3.16 (Schwarz/Bayesian Information Criterion (S/BIC))**.** For likelihood-based models the BIC (SIC) is defined

$$BIC = -2l + k \ln n \tag{3.104}$$

and in its least squares applications

$$BIC = \ln \hat{\sigma}^2 + k\frac{\ln n}{n} \tag{3.105}$$

The obvious difference between these two IC is that the AIC has a constant penalty term while the BIC has a penalty term that increases with the number of observations. The effect of the sharper penalty in the S/BIC is that for larger data sizes, the marginal increase in the likelihood (or decrease in the variance) must be greater. This distinction is subtle but important: using the BIC to select from a finite set of regressors leads to the correct model being chosen while the AIC asymptotically selects a model that includes irrelevant regressors.

Using an IC to select a model is similar to either a GtS or StG search. For example, to use an StG selection method, begin with the smallest model (usually a constant) and compute the IC for this model. Next, consider all possible univariate regressions. If any *reduce* the IC, extend the specification to include the variable that produced the smallest IC. Now, beginning from the selected univariate regression, estimate all bivariate regressions. Again, if any decrease the IC, choose the one which produces the smallest value. Repeat this procedure until the marginal contribution to the IC of adding any additional variable is positive (i.e., when comparing an L and $L+1$ variable regression, including and additional variables increase the IC).

As an alternative, if the number of regressors is sufficiently small (less than 20) it is possible to try every possible combination and choose the smallest IC. This requires 2^L regressions where L is the number of available regressors (2^{20} is about 1,000,000).

3.13.1.4 Cross-validation

Cross-validation uses pseudo-out-of-sample prediction performance to assess model specification. It is most commonly used to select a preferred model from a set of candidate models, for example, the collection of models visited as part of a GtS or StG model selection process. Variables with robust predictive power should be useful both in- and out-of-sample. Cross-validation estimates parameters

using a random subset of the data and then computes the pseudo-out-of-sample SSE on the observations that were not used in estimation. This criterion rewards models include variables with good predictive power and exclude models that incorporate variables with small coefficients that do not improve out-of-sample prediction.

The mutually exclusive and exhaustive subsets used for estimation and evaluation are randomly chosen. This randomization selection is then repeatedly applied to assess the out-of-sample fit of all data points. The most common form of cross-validation used in cross-sectional analysis is as k-fold cross-validation. This method splits the data into k-equal-sized blocks where block assignment is random. Model parameters are then estimated using the data in $k-1$ blocks, and the predictive power is evaluated on the excluded block. This leave-one-block-out strategy is then repeated for each of the remaining $k-1$ blocks. The overall cross-validated SSE is computed from the SSE values calculated on each block held out of the estimation.

Algorithm 3.10 (k-fold Cross-validation)**.**

1. *Split the data randomly into k-equal-sized bins*
2. *For each model* $m = 1, \ldots, M$ *under consideration*
   (a) *For* $i = 1, \ldots, k$
      i. *Estimate model parameters excluding the the observations in block i,*

$$\hat{\beta}_{m,i} = \arg\min \beta_{m,i} \sum_{j=1, j \notin \mathcal{B}_i}^{n} \left(Y_j - \mathbf{x}_{m,j}\beta_{m,i}\right)^2$$

      where $\mathbf{x}_{m,\cdot}$ *are the regressors included in model m and* $\mathcal{B}_i$ *is the set of observation indices in block i.*
      ii. *Compute the block i* SSE *as* $\text{SSE}_{m,i} = \sum_{j \in \mathcal{B}_i} \left(Y_j - \mathbf{x}_{m,j}\hat{\beta}_{m,i}\right)^2$.
   (b) *Compute the overall cross-validated* SSE *as* $\text{SSE}_{m,CV} = \sum_{i=1}^{k} \text{SSE}_{m,i}$.
3. *Select the model that produces the smallest cross-validates* SSE.

3.13.2 Specification Checking

Once a model has been selected, the final step is to examine the specification, where a number of issues may arise. For example, a model may have neglected some nonlinear features in the data, a few outliers may be determining the parameter estimates, or the data may be heteroskedastic. Residuals for the basis of most specification checks, although the first step in assessing model fit is always to plot the residuals. A simple residual plot often reveals problems with a model, such as large (and generally influential) residuals or correlation among the residuals in time-series applications.

Residual Plots and Nonlinearity Plot, plot, plot. Plots of both data and residuals, while not perfect, are effective methods to detect specification problems. Most data analysis should include a plot of the initial unfiltered data where large observation or missing data are easily detected. Once

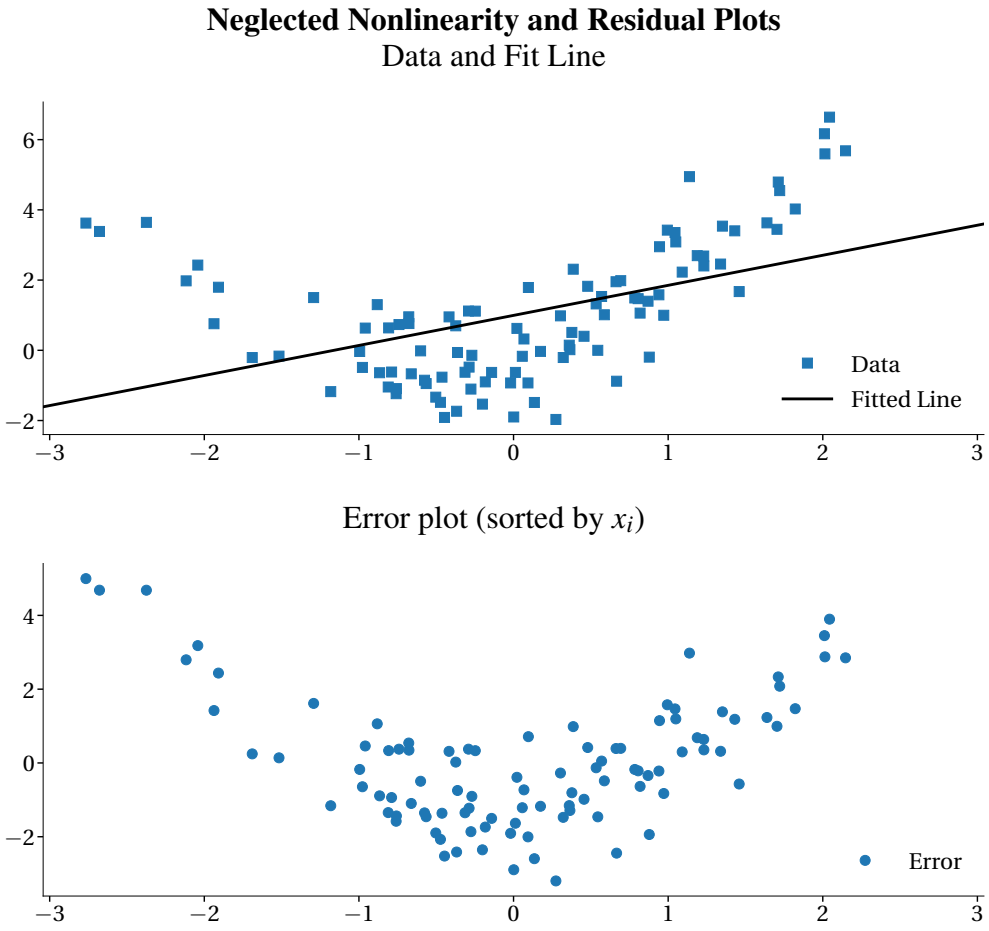

Figure 3.7: The top panel contains data generated according to $Y_i = X_i + X_i^2 + \varepsilon_i$ and a fit from a model $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. The nonlinearity should be obvious, but is even clearer in the ordered (by X_i) residual plot where a distinct "U" shape can be seen (bottom panel).

a model has been estimated the residuals should be plotted, usually by sorting them against the ordered regressors when using cross-sectional data or against time (the observation index) in time-series applications.

To see the benefits of plotting residuals, suppose the data were generated by $Y_i = X_i + X_i^2 + \varepsilon_i$ where X_i and ε_i are i.i.d. standard normal, but an affine specification, $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ was fit. Figure 3.7 contains plots of the data and fit lines (top panel) and errors (bottom panel). It is obvious from the data and fit line that the model is misspecified and the residual plot makes this clear. Residuals should have no discernible pattern in their mean when plotted against any variable (or function of variables) in the data set.

One statistical test for detecting neglected nonlinearity is Ramsey's RESET test. Suppose the model

$$Y_i = \mathbf{X}_i \beta + \varepsilon_i$$

is fit and one desires to test whether there is a neglected nonlinearity present. The RESET test uses

powers of the fit data, $\hat{Y}_i$ as additional regressors to test whether there is evidence of nonlinearity in the data.

Definition 3.17 (Ramsey's RESET Test)**.** The RESET test is a test of the null the null $H_0 : \gamma 1 = \ldots = \gamma_R = 0$ in an auxiliary regression,

$$Y_i = \mathbf{X}_i\beta + \gamma_1\hat{Y}_i^2 + \gamma_2\hat{Y}_i^3 + \ldots + \gamma_R\hat{Y}_i^{R-1}\varepsilon_i$$

where $\hat{Y}_i$ are the fit values of Y_i generated in the initial regression. The test statistic has an asymptotic χ_R^2 distribution.

R is typically 1 or 2 since higher powers may produce numerical problems, imprecise estimates, and size distortions. The biggest difficulty of using a RESET test is that rejection of the null is not informative about the changes needed to the original specification.

3.13.2.1 Parameter Stability

Parameter instability is a common problem in actual data. For example, recent evidence suggests that the market β in a CAPM may be differ across up and down markets Ang, Chen, and Xing (2006). A model fit assuming the strict CAPM would be misspecified since the parameters are not constant.

There is a simple procedure to test for parameter stability if the point where the parameters changes is known. The test is specified by including a dummy for any parameter that may change and testing the coefficient on the dummy variables for constancy.

Returning to the CAPM example, the standard specification is

$$R_i^e = \beta_1 + \beta_2(R_i^M - R_i^f) + \varepsilon_i$$

where R_i^M is the return on the market, R_i^f is the return on the risk free asset and R_i^e is the excess return on the dependent asset. To test whether the slope is different when $(R_i^M - R_i^f) < 0$, define a dummy $I_i = I_{[(R_i^M - R_i^f)<0]}$ and perform a standard test of the null $H_0 : \beta_3 = 0$ in the regression

$$R_i^e = \beta_1 + \beta_2(R_i^M - R_i^f) + \beta_3 I_i(R_i^M - R_i^f) + \varepsilon_i.$$

If the breakpoint is not known *a priori*, it is necessary to test whether there is a break in the parameter at any point in the sample. This test can be implemented by testing at every point and then examining the largest test statistic. While this is a valid procedure, the distribution of the largest test statistic is no longer χ^2 and so inference based on standard tests (and their corresponding distributions) will be misleading. This type of problem is known as a nuisance parameter problem. If the null hypothesis (that there is no break) is correct, then the value of regression coefficients after the break is not well defined. In the example above, if there is no break, then β_3 is not identified (and is a nuisance). Treatment of the issues surrounding nuisance parameters is beyond the scope of this course, but interested readers should start see Andrews and Ploberger (1994).

3.13.2.2 Rolling and Recursive Parameter Estimates

Rolling and recursive parameter estimates are useful tools for detecting parameter instability in cross-section regression of time-series data (e.g., asset returns). Rolling regression estimates use a fixed-length sample of data to estimate β and then "roll" the sampling window to produce a sequence of estimates.

Definition 3.18 (m-sample Rolling Regression Estimates)**.** The m-sample rolling regression estimates are defined as the sequence

$$\hat{\beta}_j = \left(\sum_{i=j}^{j+m-1} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \mathbf{x}_i' Y_i \tag{3.106}$$

for $j = 1, 2, \ldots, n-m+1$.

The rolling window length should be large enough so that parameter estimates in each window are reasonably well approximated by a CLT but not so long as to smooth out any variation in β. 60-months is a common window length in applications using monthly asset price data and window lengths ranging between 3-months and 2-year are common when using daily data. The rolling regression coefficients can be visually inspected for evidence of instability, and approximate confidence intervals (based on an assumption of parameter stability) can be constructed by estimating the parameter covariance on the full sample of n observations and then scaling by n/m so that the estimated covariance is appropriate for a sample of m observations. The parameter covariance can alternatively be estimated by averaging the $n-m+1$ covariance estimates corresponding to each sample, $\hat{\Sigma}_{\mathbf{XX},j}^{-1} \hat{\mathbf{S}}_j \hat{\Sigma}_{\mathbf{XX},j}^{-1}$, where

$$\hat{\Sigma}_{\mathbf{XX},j} = m^{-1} \sum_{i=j}^{j+m-1} \mathbf{x}_i' \mathbf{x}_i \tag{3.107}$$

and

$$\hat{\mathbf{S}}_j = m^{-1} \sum_{i=j}^{j+m-1} \hat{\varepsilon}_{i,j} \mathbf{x}_i' \mathbf{x}_i \tag{3.108}$$

where $\hat{\varepsilon}_{i,j} = Y_i - \mathbf{x}_i' \hat{\beta}_j$, and if the parameters are stable these methods for estimating the parameter covariance should produce similar confidence intervals.

60-month rolling regressions of the *BH* portfolio in the 4-factor model are presented in figure 3.8 where approximate confidence intervals were computed using the re-scaled full-sample parameter covariance estimate. While these confidence intervals cannot directly be used to test for parameter instability, the estimate of the loadings on the market, *SMB* and *HML* vary more than their intervals indicate these parameters should were they stable.

An alternative to rolling regressions is to recursively estimate parameters which uses an expanding window of observations to estimate $\hat{\beta}$.

Definition 3.19 (Recursive Regression Estimates)**.** Recursive regression estimates are defined as the sequence

$$\hat{\beta}_j = \left(\sum_{i=1}^{j} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \mathbf{x}_i' Y_i \tag{3.109}$$

for $j = l, 2, \ldots, n$ where $l > k$ is the smallest window used.

Approximate confidence intervals can be computed either by re-scaling the full-sample parameter covariance or by directly estimating the parameter covariance in each recursive sample. Documenting evidence of parameter instability using recursive estimates is often more difficult than with rolling, as demonstrated in figure 3.9

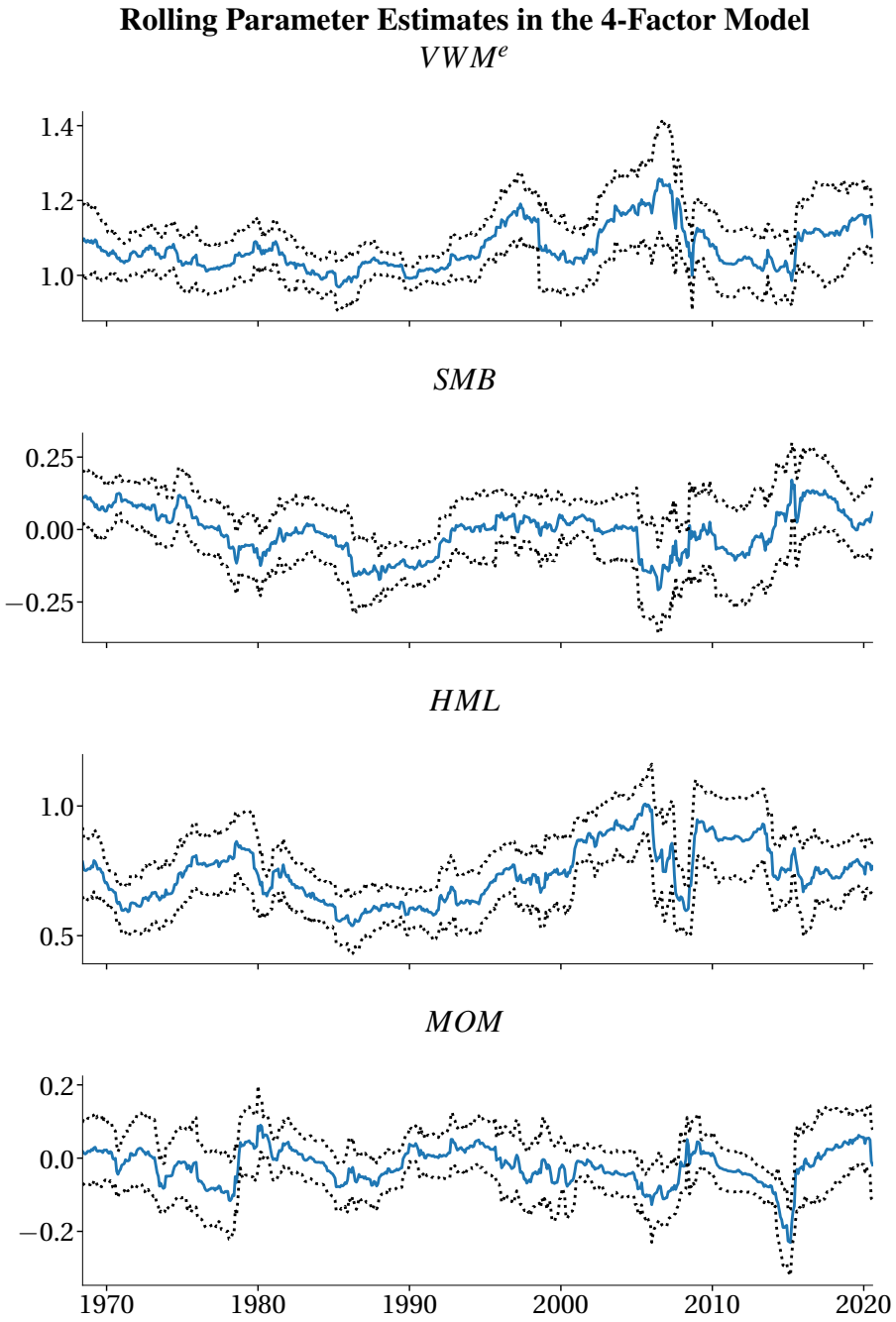

Figure 3.8: 60-month rolling parameter estimates from the model $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. Approximate confidence intervals were constructed by scaling the full sample parameter covariance. These rolling estimates indicate that the market loading of the Big-High portfolio varied substantially at the beginning of the samplefixed-length sample and that the loadings on both *SMB* and *HML* may be time-varying.

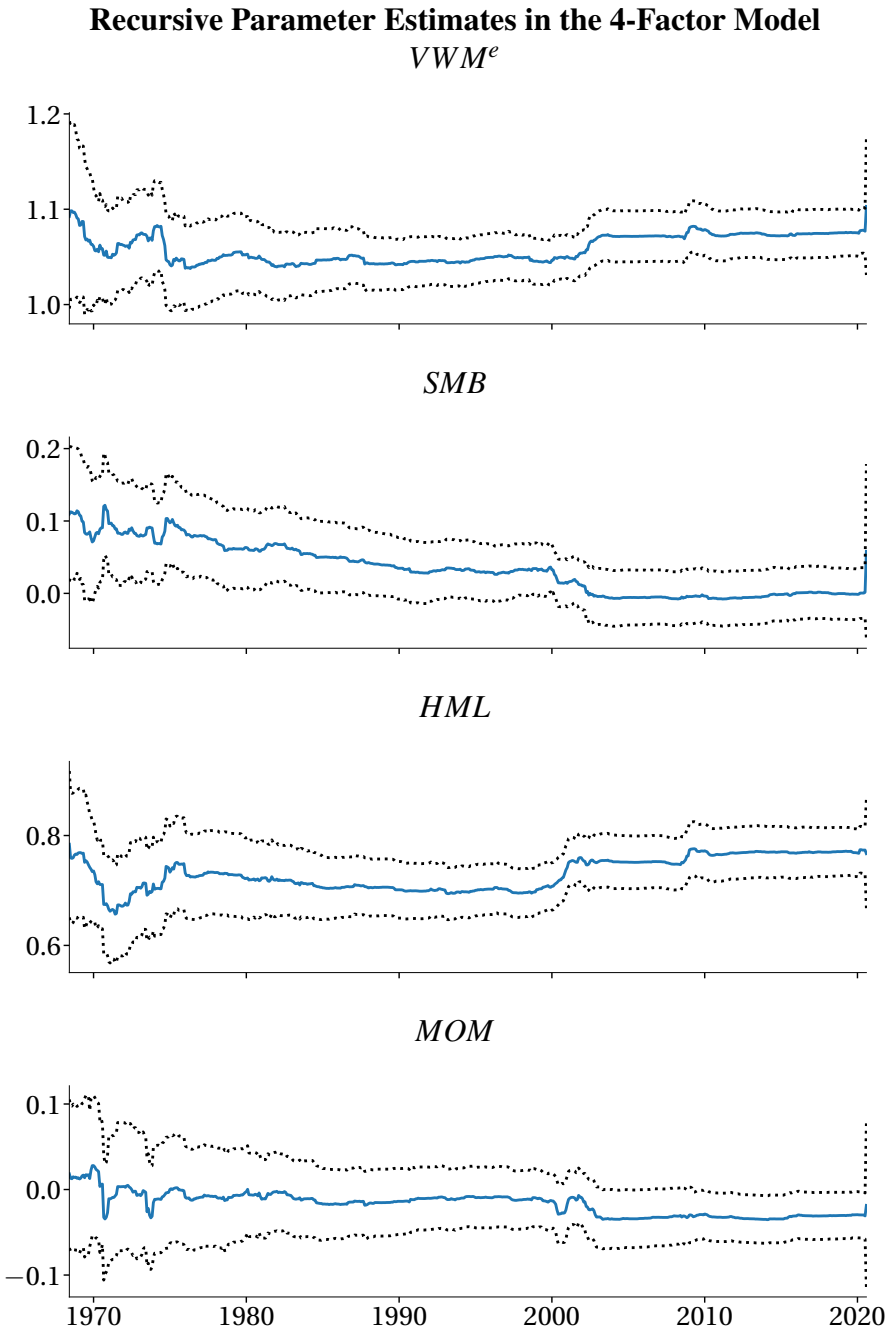

Figure 3.9: Recursive parameter estimates from the model $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. Approximate confidence intervals were constructed by scaling the full sample parameter covariance. While less compelling than the rolling window estimates, these recursive estimates indicate that the loading on the market and on HML may not be constant throughout the sample.

3.13.2.3 Normality

Normality may be a concern if the validity of the small-sample assumptions is important. The standard method to test for normality of estimated residuals is the Jarque-Bera (JB) test which is based on two higher order moments (skewness and kurtosis) and tests whether they are consistent with those of a normal distribution. In the normal, the skewness is 0 (it is symmetric) and the kurtosis is 3. Let $\hat{\varepsilon}_i$ be the estimated residuals. Skewness and kurtosis are defined

$$\hat{sk} = \frac{n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^3}{(\hat{\sigma}^2)^{\frac{3}{2}}}$$

$$\hat{\kappa} = \frac{n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^4}{(\hat{\sigma}^2)^2}$$

The JB test is computed

$$JB = \frac{n}{6}\left(sk^2 + \frac{1}{4}(\kappa - 3)^2\right)$$

and is distributed χ^2_2. If $sk \approx 0$ and $\kappa \approx 3$, then the JB should be small and normality should not be rejected. To use the JB test, compute JB and compare it to C_α where C_α is the critical value from a χ^2_2. If $JB > C_\alpha$, reject the null of normality.

3.13.2.4 Heteroskedasticity

Heteroskedasticity is a problem if neglected. See section 3.12.4.

3.13.2.5 Influential Observations

Influential observations are those which have a large effect on the estimated parameters. Data, particularly data other than asset price data, often contain errors.[24] These errors, whether a measurement problem or a typo, tend to make $\hat{\beta}$ unreliable. One method to assess whether any observation has an undue effect on the sample is to compute the vector of "hat" matrices,

$$h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'.$$

This vector (which is the diagonal of $\mathbf{P}_\mathbf{X}$) summarizes the influence of each observation on the estimated parameters and is known as the influence function. Ideally, these should be similar and no observation should dominate.

Consider a simple specification where $Y_i = X_i + \varepsilon_i$ where X_i and ε_i are i.i.d. standard normal. In this case the influence function is well behaved. Now suppose one x_i is erroneously increased by 100. In this case, the influence function shows that the contaminated observation (assume it is X_n) has a large impact on the parameter estimates. Figure 3.10 contains four panels. The two left panels show the original data (top) and the data with the error (bottom) while the two right panels contain the influence functions. The influence function for the non-contaminated data is well behaved and each observation has less than 10% influence. In the contaminated data, one observation (the big outlier), has an influence greater than 98%.

[24] And even some asset price data, such as TAQ prices.

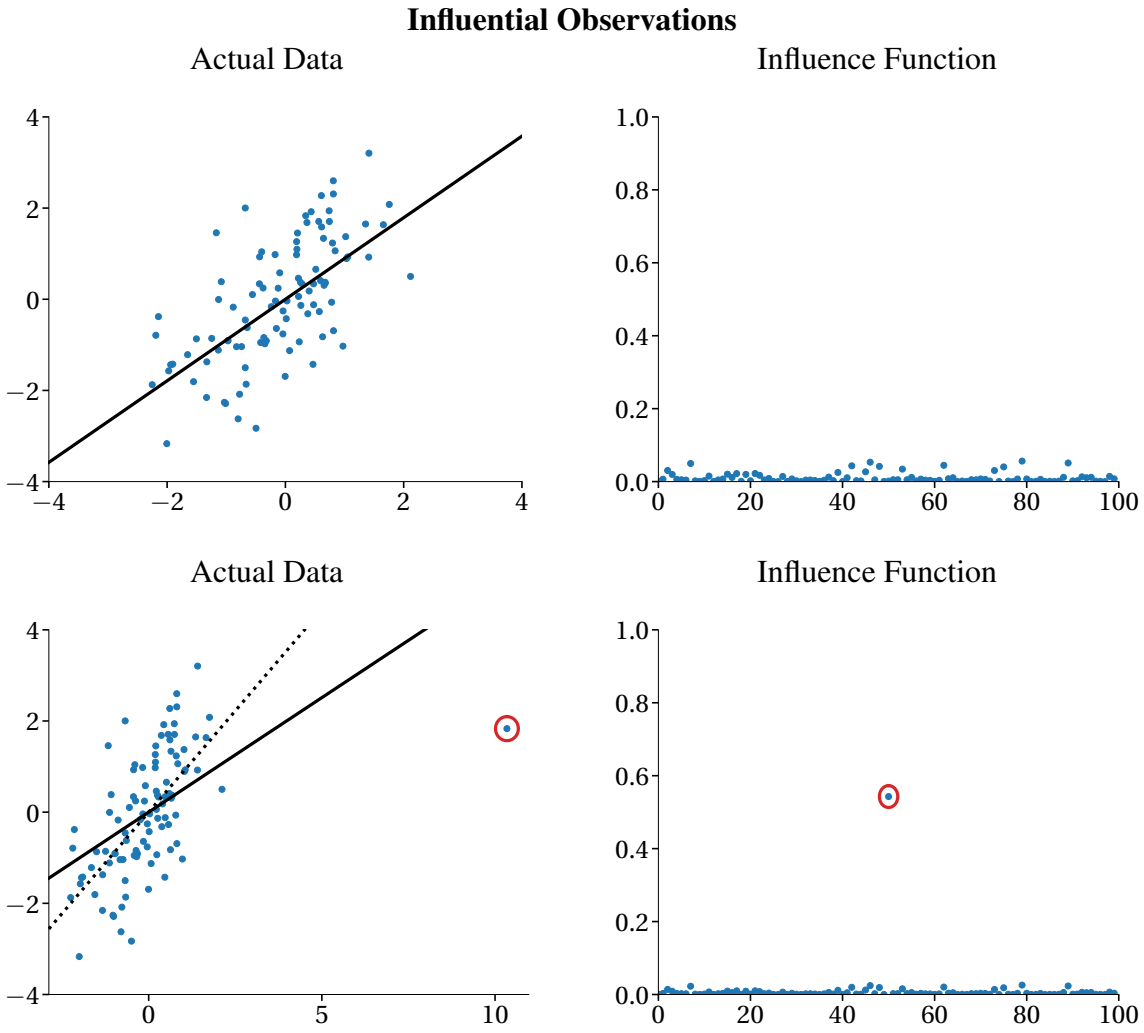

Figure 3.10: The two left panels contain realizations from the data generating process $Y_i = X_i + \varepsilon_i$ where a single X_i has been contaminated (bottom left panel). The two right panels contain the influence functions of the X_i. If all data points were uniformly influential, the distribution of the influence function should be close to uniform (as is the case in the top left panel). In the bottom right panel, it is clear that the entire fit is being driven by a single X_i which has an influence greater than .98.

Plotting the data would have picked up this problem immediately. However, it may be difficult to determine whether an observation is influential when using multiple regressors because the regressors for an observation may be "large" in many dimensions.

3.13.3 Improving estimation in the presence of outliers

Data may contain outliers for many reasons: someone entered an incorrect price on an electronic exchange, a computer glitch multiplied all data by some large constant or a CEO provided an answer out-of-line with other answers due to misunderstanding a survey question. The standard least-squares

estimator is non-robust in the sense that large observations can have a potentially unbounded effect on the estimated parameters. A number of techniques have been developed to produce "robust" regression estimates that use weighted least squares to restrict the influence of any observation.

For clarity of exposition, consider the problem of estimating the mean using data that may be contaminated with a small number of large errors. The usual estimator will be heavily influenced by these outliers, and if outliers occur with any regularity in the data (suppose, for example, 1% of data is contaminated), the effect of outliers can result in an estimator that is biased and in some cases inconsistent. The simplest method to robustly estimate the mean is to use an α-trimmed mean where α represents a quantile of the empirical distribution of the data.

Definition 3.20 (α-Trimmed Mean)**.** The α-quantile trimmed mean is

$$\hat{\mu}_\alpha = \frac{\sum_{i=1}^n Y_i I_{[C_L \leq Y_i \leq C_U]}}{n^*} \tag{3.110}$$

where $n^* = n(1-\alpha) = \sum_{i=1}^n I_{[-C<Y_i<C]}$ is the number of observations used in the trimmed mean.[25]

Usually α is chosen to be between .90 and .99. To use an α-trimmed mean estimator, first compute C_L the $\alpha/2$-quantile and C_U the $1-\alpha/2$-quantile of the of y. Using these values, compute the trimmed mean as

A closely related estimator to the trimmed mean is the Winsorized mean. The sole difference between an α-trimmed mean and a Winsorized mean is the method for addressing the outliers. Rather than dropping extreme observations below C_L and C_U, a Winsorized mean truncates the data at these points.

Definition 3.21 (Winsorized mean)**.** Let Y_i^* denote a transformed version of Y_i,

$$Y_i^* = \max(\min(Y_i, C_U), C_L)$$

where C_L and C_U are the $\alpha/2$ and $1-\alpha/2$ quantiles of Y. The Winsorized mean is defined

$$\hat{\mu}_W = \frac{\sum_{i=1}^n Y_i^*}{n}. \tag{3.111}$$

While the α-trimmed mean and the Winsorized mean are "robust" to outliers, they are not robust to other assumptions about the data. For example, both mean estimators are biased unless the distribution is symmetric, although "robust" estimators are often employed as an ad-hoc test that results based on the standard mean estimator are not being driven by outliers.

Both of these estimators are in the family of linear estimators (L-estimators). Members of this family can always be written as

$$\hat{\mu}^* = \sum_{i=1}^n w_i Y_i$$

for some set of weights w_i where the data, Y_i, are ordered such that $Y_{j-1} \leq Y_j$ for $j = 2, 3, \ldots, N$. This class of estimators obviously includes the sample mean by setting $w_i = \frac{1}{n}$ for all i, and it also includes the median by setting $w_i = 0$ for all i except $w_m = 1$ where $m = (n+1)/2$ (n is odd) or $w_m = w_{m+1} = 1/2$ where $m = n/2$ (n is even). The trimmed mean estimator can be constructed by

[25] This assumes that $n\alpha$ is an integer. If this is not the case, the second expression is still valid.

setting $w_i = 0$ if $n \leq s$ or $i \geq n-s$ and $w_i = \frac{1}{n-2s}$ otherwise where $s = n\alpha$ is assumed to be an integer. The Winsorized mean sets $w_i = 0$ if $n \leq s$ or $n \geq N-s$, $w_i = \frac{s+1}{n}$ if $n = s+1$ or $n = n-s-1$ and $w_i = \frac{1}{n}$ otherwise. Examining the weights between the α-trimmed mean and the Winsorized mean, the primary difference is on the weights w_{k+1} and w_{n-k-1}. In the trimmed mean, the weights on these observation are the same as the weights on the data between these points. In the Winsorized mean estimator, the weights on these observations are $\frac{k+1}{n}$ reflecting the censoring that occurs at these observations.

3.13.3.1 Robust regression-based estimators

Like the mean estimator, the least-squares estimator is not "robust" to outliers. To understand the relationship between L-estimators and linear regression, consider decomposing each observation into its mean and an additive error,

$$\begin{aligned}\hat{\mu}^* &= \sum_{i=1}^{n} w_i Y_i \\ &= \sum_{i=1}^{n} w_i (\mu + \varepsilon_i) \\ &= \sum_{i=1}^{n} w_i \mu + \sum_{i=1}^{n} w_i \varepsilon_i\end{aligned}$$

A number of properties can be discerned from this decomposition. First, in order for μ^* to be unbiased it must be the case that $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} \mathrm{E}[w_i \varepsilon_i] = 0$. All of the linear estimators satisfy the first condition although the second will depend crucially on the distribution of the errors. If the distribution of the errors is symmetric then the Winsorized mean, the α-trimmed mean or even median are unbiased estimators of the mean. However, if the error distribution is not symmetric, then these estimators are likely to be biased. Unlike the usual case where $\mathrm{E}[w_i \varepsilon_i] = w_i \mathrm{E}[\varepsilon_i]$, the weights are functions of the errors and the expectation of the product of the expectations is not the expectation of the product. Second, weights on the observations (Y_i) are the same as weights on the errors, ε_i. This relationship follows from noticing that if $Y_j \leq Y_{j+1}$, then it must be the case that $\varepsilon_j \leq \varepsilon_{j+1}$.

Robust estimators in linear regression models require a two-step or iterative procedure. The difference between robust mean estimators and robust regression arises since if Y_i has a relationship to a set of explanatory variables $\mathbf{x}_i$, then orderings based on Y_i will *not* be the same as orderings based on the residuals, ε_i. For example, consider the simple regression

$$Y_i = \beta X_i + \varepsilon_i.$$

Assuming $\beta > 0$, the largest Y_i are those which correspond either the largest X_i *or* ε_i. Simple trimming estimators will not only trim large errors but will also trim Y_i that have large values of X_i. The left panels of figure 3.11 illustrate the effects of Windsorization and trimming on the raw data. In both cases, the regression coefficient is asymptotically biased (as indicated by the dotted line) since trimming the raw data results in an error that is correlated with the regressor. For example, observations with the largest X_i values and with positive ε_i more likely to be trimmed. Similarly, observations for

the smallest X_i values and with negative ε_i are more likely to be trimmed. The result of the trimming is that the remaining ε_i are *negatively* correlated with the remaining X_i.

To avoid this issue, a two-step or iterative procedure is needed. The first step is used to produce a preliminary estimate of $\hat{\beta}$. OLS is commonly used in this step although some other weighted least-squares estimator may be used instead. Estimated residuals can be constructed from the preliminary estimate of β ($\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$), and the trimming or Windsorizing is done on these preliminary residuals. In the case of α-trimming, observations with the largest errors (in absolute value) are dropped, and the α-trimmed regression is estimated using only the observations with $C_L < \hat{\varepsilon}_i < C_U$.

Winsorized regression also uses the first step regression to estimate $\hat{\varepsilon}$, but, rather than dropping observations, errors larger than C_U are set to $\hat{\varepsilon}_U$ and errors smaller than C_L are set to $\hat{\varepsilon}_L$. Using these modified errors,

$$\hat{\varepsilon}_i^{\star} = \max(\min(\hat{\varepsilon}_i, C_U), C_L)$$

a transformed set of dependent variables is created, $Y_i^{\star} = \mathbf{x}_i\hat{\beta} + \varepsilon_i^{\star}$. The Winsorized regression coefficients are then estimated by regressing $Y_i^{\star}$ on $\mathbf{x}_i$. The correct application of α-trimming and Windsorization are illustrated in the bottom two panels of figure 3.11. In the α-trimming examples, observations marked with an $\mathtt{x}$ were trimmed, and in the Windsorization example, observations marked with a $\bullet$ were reduced from their original value to either C_U or C_L. It should be noted that while both of these estimators are unbiased, this result relies crucially on the symmetry of the errors.

In addition to the two-step procedure illustrated above, an iterative estimator can be defined by starting with some initial estimate of $\hat{\beta}$ denoted $\hat{\beta}^{(1)}$ and then trimming (or Windsorization) the data to estimate a second set of coefficients, $\hat{\beta}^{(2)}$. Using $\hat{\beta}^{(2)}$ and the original data, a different set of estimated residuals can be computed $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}^{(2)}$ and trimmed (or Winsorized). Using the new set of trimmed observations, a new set of coefficients, $\hat{\beta}^{(3)}$, can be estimated. This procedure can be repeated until it converges – $\max \left|\hat{\beta}^{(i)} - \hat{\beta}^{(i-1)}\right|$.[26]

Both α-trimmed and Winsorized regression are special cases of a broader class of "robust" regression estimators. Many of these robust regression estimators can be implemented using an iterative procedure known as Iteratively Re-weighted Least Squares (IRWLS) and, unlike trimmed or Winsorized least squares, are guaranteed to converge. For more on these estimators, see Huber (2004) or Rousseeuw and Leroy (2003).

3.13.3.2 Ad-hoc "Robust" Estimators

It is not uncommon to see papers that use Windsorization (or trimming) in the academic finance literature as a check that the findings are not being driven by a small fraction of outlying data. This is usually done by directly Windsorizing the dependent variable and the regressors. While there is no theoretical basis for these ad-hoc estimators, they are a useful tool to ensure that results and parameter estimates are valid for "typical" observations as well as for the full sample. However, if this is the goal, other methods, such as visuals inspections of residuals or residuals sorted by explanatory variables, are equally valid and often more useful in detecting problems in a regression.

[26] These iterative procedures may not converge due to cycles in $\{\hat{\beta}^{(j)}\}$.

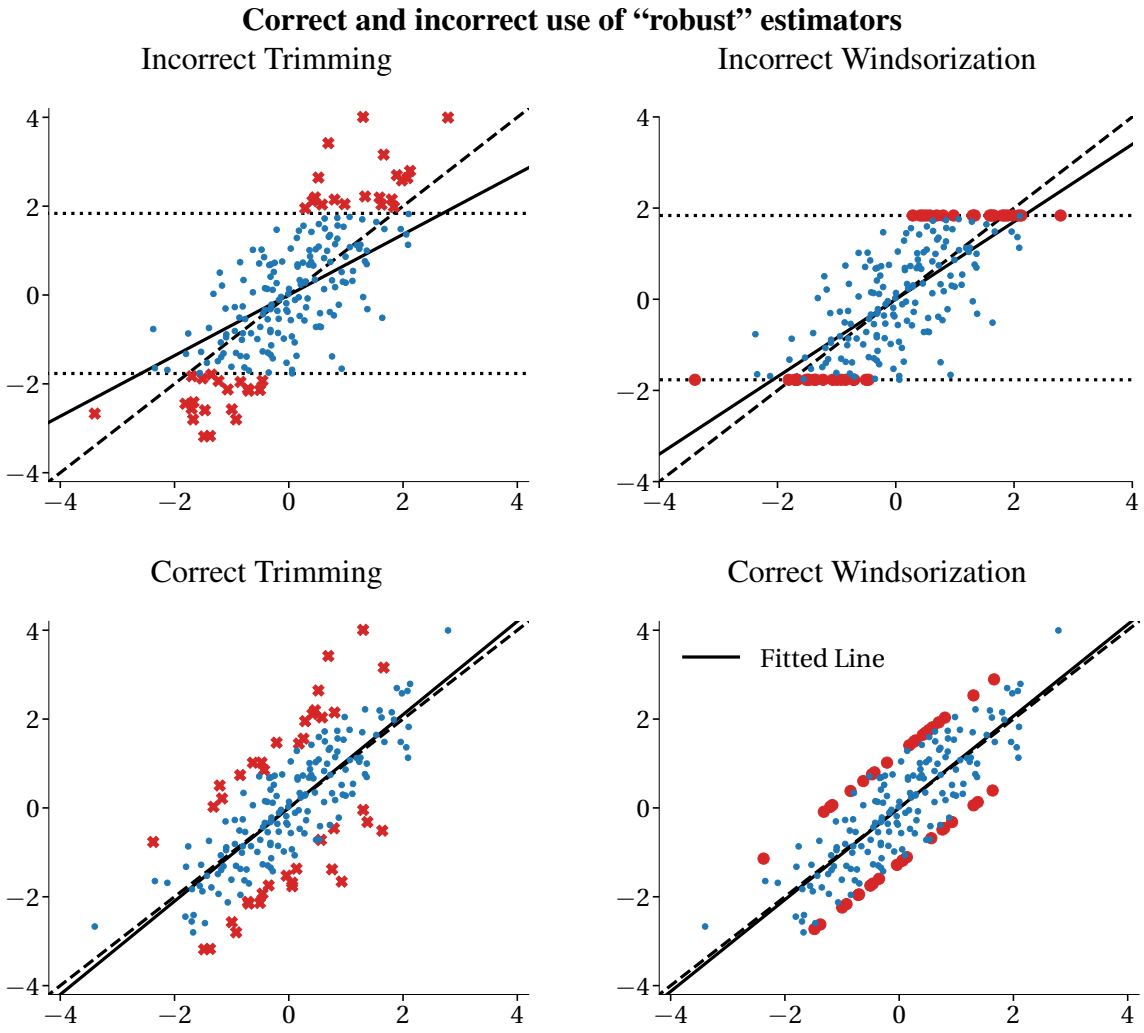

Figure 3.11: These four panels illustrate correct and incorrect α-trimming (left) and Windsorization (right). In both cases, the DGP was $Y_i = X_i + \varepsilon_i$ where X_i and ε_i were independent standard normal random variables. The top panels show incorrect trimming based on the unmodified data, and the bottom panels show correct trimming based on an initial estimate of the slope.

3.13.3.3 Inference on "Robust" Estimators

It may be tempting to use OLS or White heteroskedasticity robust standard errors in "robust" regressions. These regressions (and most L-estimators) appear similar to standard least-squares estimators. However, there is an additional term in the asymptotic covariance of the estimated regression coefficients since the trimming or Windsorization point must be estimated. This term is related to the precision of the trimming point and is closely related to the uncertainty affecting the estimation of a quantile. Fortunately, bootstrapping can be used (under some mild conditions) to estimate the covariance of the regressors.

3.14 Machine Learning

Machine learning approaches to regression, also known as supervised learning, address two key challenges:

- Variable selection when the number of candidate variables is large. In machine learning, variables are often called features, and the collection of all features is called the feature space. Most machine learning algorithms are capable of modeling data sets where the number of variables exceeds the number of observations available.
- Optimizing model parameters to perform well in out-of-sample prediction. In most applications, this optimization makes an explicit trade-off between bias and variance, and most ML approaches to regression use biased estimators that have lower parameter variance than vanilla OLS. This reduction in variance, especially for parameters that have a small effect relative to their uncertainty, improves out-of-sample prediction at the cost of some bias.

ML approaches achieve these goals using cross-validation to select models and parameter values that perform well both in- and out-of-sample. These alternative approaches generally provide methods to jointly select relevant variables and estimate parameters. Some methods make use of bootstrapping to improve the reliability of the models in out-of-sample data. Ultimately these approaches all produce a standard linear regression model where the coefficients are not usually estimated using standard OLS. The most useful strategies tend to introduce a limited amount of bias by *shrinking* regression coefficients toward 0 to mitigate the cost of parameter uncertainty.

3.14.1 Best Subset Regression

Best Subset Regression is the simplest method to construct a model given a set of predictors. Suppose you have p candidate variables $X_{1,i},\ldots,X_{p,i}$. Best Subset Regression finds the combination of variables in this set that optimizes the model's fit according to some criteria, for example, the cross-validated SSE or BIC. Best Subset Regression begins by finding the model that produces the smallest in-sample SSE, or equivalently the largest R^2, using k of the p variables. Let this model be denoted $\mathcal{M}_k$. This step involves fitting $\left(\begin{array}{c} p \\ k \end{array}\right)$ distinct models. The best model is selected for each possible value of $k = 1,2,\ldots,p$. The initial inputs are a set of $p+1$ distinct models $\mathcal{M}_0,\mathcal{M}_1,\ldots,\mathcal{M}_p$ where $\mathcal{M}_0$ is a model that contains no predictors. The Best Subset Regression is chosen by comparing the performance of these $p+1$ models using some criterion, for example, the cross-validated SSE, and selecting the model that performs the best. There are two important issues with Best Subset Regression. First, it can only be used when the set of candidate predictors p is moderate (≤ 30) since there are $2^p - 1$ distinct models that must be estimated. Second, the coefficients of the best model are estimated by OLS. OLS estimates always overfit the sample used to estimate the parameters, and the in-sample overfitting reduces the out-of-sample performance of the models.

Algorithm 3.11 (Best Subset Regression)**.**

1. *For $k \in \{0,1,\ldots,p\}$ estimate each of the $\left(\begin{array}{c} p \\ k \end{array}\right)$ distinct models containing k variables, saving the model that produces the smallest* SSE *as $\mathcal{M}_j$, $j = 0,\ldots,p$.*

2. *Select the Best Subset Regression as the model from the set* $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$ *that minimizes some criterion such as the cross-validated* SSE.

3.14.2 Forward, Backward, and Hybrid Stepwise Regression

Best Subset Regression cannot be used when p is large. Stepwise model building is an alternative the builds the models $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ sequentially. Forward stepwise regression begins with no variables selected. Each of the excluded variables, p in total, are tried one at a time, and the regressor that produces the best fit is retained in $\mathcal{M}_1$. The second model, $\mathcal{M}_2$, is then selected by adding each of the $p-1$ variables that were not included in $\mathcal{M}_1$ and is defined as the model that produces the best in-sample fit. This process is repeated so that $\mathcal{M}_{j+1}$ adds one of the $p-j$ variables to $\mathcal{M}_j$ that were not included in $\mathcal{M}_j$. The output of the first step is a set of $p+1$ models $\mathcal{M}_0, \mathcal{M}_1, \ldots \mathcal{M}_p$ where larger models always nest smaller models. The final model is selected from the set of candidate models by optimizing some criterion such as the cross-validated SSE.

Algorithm 3.12 (Forward Stepwise Regression)**.**

1. *Begin with the empty model,* $\mathcal{M}_0$.
2. *For* $j \in \{0, \ldots, p-1\}$*, construct model* $\mathcal{M}_{j+1}$ *as the model the minimizes the* SSE *by adding each of the* $p-j$ *variables to the variables included in model* $\mathcal{M}_j$.
3. *Select the Forward Stepwise Regression as the model from the set* $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$ *that minimizes some criterion such as the cross-validated* SSE.

Backward stepwise regression operates in the opposite direction. Begin with the model that contains all variables $\mathcal{M}_p$. The next smaller model, $\mathcal{M}_{p-1}$ is defined as the model that minimizes the SSE considering each of the p models that drops a single variable from $\mathcal{M}_p$. This process continues where $\mathcal{M}_j$ is defined as the model that maximizes the in-sample fit using j of the $j+1$ variables included in $\mathcal{M}_{j+1}$. Like forward stepwise regression, backward stepwise regression produces a set of $p+1$ models $\mathcal{M}_0, \mathcal{M}_1, \ldots \mathcal{M}_p$. The best model is then selected from this set of candidate models by optimizing some criterion function.

Algorithm 3.13 (Backward Stepwise Regression)**.**

1. *Begin with the complete model,* $\mathcal{M}_p$.
2. *For* $j \in \{p-1, p-2, \ldots, 0\}$*, construct model* $\mathcal{M}_j$ *as the model the minimizes the* SSE *by removing each of the* j *variables, one at a time, of the variables included in model* $\mathcal{M}_{j+1}$.
3. *Select the Backward Stepwise Regression as the model from the set* $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$ *that minimizes some criterion such as the cross-validated* SSE.

Hybrid approaches combine the two. For example, suppose forward stepwise regression is used to select $\mathcal{M}_k$ where $k < p$. Backward stepwise regression can be used on the k included regressors in $\mathcal{M}_k$ to produce a new sequence of models $\mathcal{M}_j^k$ for $j = k-1, k-2, \ldots 1$. This sequence may be distinct from what forward or backward stepwise regression would arrive at alone. The hybrid approach generally produces a larger set of candidate models while remaining computationally tractable as long as the number of direction switches is small. This larger set of candidate models has an increased chance

of including the Best Subset Regression than either forward or backward stepwise regression alone. The primary challenge of the hybrid approach is determining the number of direction reversals to use, although, in practice, this is often dictated by the computational time available. Like both forward and backward stepwise regression, the final model is selected from the enlarged pool of candidate models by optimizing some criteria.

3.14.3 Ridge Regression

Ridge regression differs from best subset and stepwise regression in two ways: it does not select variables, and coefficients are not estimated using standard OLS.

Definition 3.22 (Ridge Regression)**.**
The ridge regression estimator with tuning parameter ω is defined as the solution to

$$\arg\min_{\beta} (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta) \text{ subject to } \sum_{j=1}^{k}\beta_j^2 \leq \omega. \tag{3.112}$$

This constrained problem is equivalent to the unconstrained problem

$$\arg\min_{\beta} (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{j=1}^{k}\beta_j^2 \tag{3.113}$$

where ω and λ take different values and have an inverse relationship (i.e., large values of ω correspond to small values of λ). The solution to this optimization problem is

$$\hat{\beta}^{\text{Ridge}} = \left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k\right)^{-1}\mathbf{X}'\mathbf{y} \tag{3.114}$$

where k is the number of regressors included in the model.

Recall that the OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The effect of the ridge penalty is simple to deduce from eq. (3.114) since $\lambda > 0$. The term $\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k$ must always be larger, in a matrix sense, than $\mathbf{X}'\mathbf{X}$ since $\lambda\mathbf{I}_k$ is a diagonal matrix with positive values along its diagonal. It must then be the case that $(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k)^{-1}$ is *smaller* than $\mathbf{X}'\mathbf{X}$, again in a matrix sense, and so the ridge coefficient estimates $\hat{\beta}^{\text{Ridge}}$ are always closer to 0 than the OLS estimates $\hat{\beta}$. Ridge regression is known as a *shrinkage* estimator since the parameter estimates pull the parameters towards the shrinkage target of 0. In practice shrinkage introduces some bias in the coefficient but reduces their variance, and ridge regression often outperforms OLS in out-of-sample applications.

Ridge regression depends on a single tuning parameter, λ, which controls how bias and variance are traded off. The optimal value is determined by trying several different values and selecting the value $\lambda^\star$ that produces the smallest cross-validated SSE. Note that ridge regression does not provide any guidance as to which variables to include in the model, and so some form of model selection is usually needed. The optimal choice of λ depends on the number of regressors included in the model, and so it must be re-optimized in each distinct model. There are many variants of ridge regression that change the penalty structure. For example, one variant allows the shrinkage to be applied to only a subset of the included variables. This penalization structure can be useful if some variables are strong predictors, while others are less useful. This penalty structure can be further generalized to

apply different amounts of shrinkage to distinct groups of regressors or even to impose cross-regressor shrinkage where the total magnitude of a set of the regressors in the model is affected.[27]

3.14.4 LASSO, Forward Stagewise Regression, and LARS

LASSO (least absolute shrinkage and selection operator), Forward Stagewise Regression, and LARS (Least Angle Regression) are relatively new methods the embed both variable selection and shrinkage into a unified approach (Tibshirani, 1996; Efron, Hastie, Johnstone, and Tibshirani, 2004). LASSO is similar to ridge regression and can be written as a constrained least square problem.

Definition 3.23 (LASSO)**.** The LASSO estimator with tuning parameter ω is defined as the solution to

$$\underset{\beta}{\arg\min}\ (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\ \text{ subject to }\ \sum_{j=1}^{k}\left|\beta_j\right| < \omega \tag{3.115}$$

The key difference is that the constraint is on the sum of the *absolute value* of the coefficients and not their squared values. The LASSO estimator adds an additional constraint to the least-squares problem that limits the magnitude of regression coefficients that produces an interpretable model. Regressors that have little explanatory power will have coefficients *exactly* equal to 0 (and hence are excluded). This means that LASSO both estimates parameters and selects variables – any variable with a coefficient that is exactly 0 is effectively removed from the model.

The LASSO constrained minimization problem is dual to a penalized least-squares problem,

$$\underset{\beta}{\arg\min}\ (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{j=1}^{k}\left|\beta_j\right| \tag{3.116}$$

where ω and λ have an inverse relationship. While LASSO has a closed form solution for any value of λ,the estimator is not simple to describe in a single equation.

Forward Stagewise Regression is closely related to LASSO and illustrates the fundamental principle used in variable selection. Estimation begins with a model that contains no regressors. The algorithm then uses an iterative method to build the regression in small steps by expanding the regression coefficients (small enough that the coefficient expansions should be virtually continuous).

[27] The complete formulation of a ridge regression is

$$\underset{\beta}{\arg\min}\ (\mathbf{y}-\mathbf{X}\beta)(\mathbf{y}-\mathbf{X}\beta)+(\beta-\beta_0)'\Lambda(\beta-\beta_0)$$

where β_0 is the shrinkage target and Λ is a positive definite matrix that controls the amount of shrinkage. This form nests the classic specification when $\Lambda=\lambda\mathbf{I}_k$ and $\beta_0=0$. If Λ is not diagonal, then the estimator will apply cross-variable penalties. The solution to the general problem is

$$\hat{\beta}^{\text{Ridge}}=(\mathbf{X}'\mathbf{X}+\Lambda)^{-1}(\mathbf{X}'\mathbf{y}+\Lambda\beta_0).$$

This shows that the OLS solution is recovered when $\Lambda=\mathbf{0}$. If Λ is very large, then $\hat{\beta}^{\text{Ridge}}\approx\Lambda^{-1}\Lambda\beta_0=\beta_0$ and the estimate depends only on the shrinkage target β_0.

Algorithm 3.14 (Forward Stagewise Regression)**.** *The Forward Stagewise Regression (FSR) estimator is defined as the sample paths of* $\hat{\beta}$ *defined by*

1. *Begin with* $\hat{\beta}^{(0)} = 0$*, and errors* $\varepsilon^{(0)} = \mathbf{y}$
2. *Compute the correlations of the residual at iteration i with the regressors,* $\mathbf{c}^{(i)} = \text{Corr}\left[\mathbf{X}, \varepsilon^{(i)}\right]$
3. *Define* j *to be the index of the largest element of* $|\mathbf{c}^{(i)}|$ *(the absolute value of the correlations), and update the coefficients where* $\hat{\beta}_j^{(i+1)} = \hat{\beta}_j^{(i)} + \eta \cdot sign\left(c_j\right)$ *and* $\hat{\beta}_l^{(i+1)} = \hat{\beta}_l^{(i)}$ *for* $l \neq j$ *where* η *is a small number (should be much smaller than* c_j*).*[28]
4. *Compute* $\varepsilon^{(i+1)} = \mathbf{y} - \mathbf{X}\hat{\beta}^{(i+1)}$
5. *Repeat steps 2 – 4 until all correlations are 0 (if* $\varepsilon^{(i)} = \mathbf{0}$ *than all correlations are 0 by definition).*

The coefficients of FSR are determined by taking a small step in the direction of the highest correlation between the regressors and the current error, and so the algorithm will always take a step in the direction of the regressor that has the most (local) explanatory power over the regressand. The final stage FSR coefficients will be equal to the OLS estimates as long as the number of regressors under consideration is smaller than the number of observations. The LASSO estimate is usually computed using the LARS algorithm, which simplifies FSR by finding the exact step size needed before the next variable enters the regression.

Algorithm 3.15 (Least Angle Regression)**.** *The Least Angle Regression (LARS) estimator is defined as the sample paths of* $\hat{\beta}$ *defined by:*

1. *Begin with* $\hat{\beta}^{(0)} = 0$*, and errors* $\varepsilon^{(0)} = \tilde{\mathbf{y}}$ *where*

$$\tilde{\mathbf{y}} = \frac{\mathbf{y} - \bar{y}}{\hat{\sigma}_y} \tag{3.117}$$

and

$$\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \bar{x}_i}{\hat{\sigma}_x} \tag{3.118}$$

are studentized versions of the original data.[29]

2. *Compute the correlations of the residual at state i with the regressors,* $\mathbf{c}^{(i)} = \text{Corr}\left[\tilde{\mathbf{X}}^{(i)}, \varepsilon^{(i)}\right]$ *and define* j *to be the index of the largest element of* $|\mathbf{c}^{(i)}|$ *(the absolute value of the correlations).*
3. *Define the active set of regressors* $\tilde{\mathbf{X}}^{(1)} = \tilde{\mathbf{x}}_j$*.*

[28] η should be larger than some small value to ensure the algorithm completes in finitely many steps, but should always be weakly smaller than $|c_j|$.

[29] LARS can be implemented on non-studentized data be replacing correlation with $\mathbf{c}^{(i)} = \mathbf{X}^{(i)\prime}\varepsilon^{(i)}$.

4. *Move* $\hat{\beta}^{(1)} = \hat{\beta}_j$ *towards the least squares estimate of regressing* $\varepsilon^{(0)}$ *on* $\tilde{\mathbf{X}}^{(1)}$ *until the correlation between* $\varepsilon^{(1)} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}^{(1)}\hat{\beta}^{(1)}$ *and some other* $\tilde{\mathbf{x}}_k$ *is equal to the correlation between* $\varepsilon^{(1)}$ *and* $\tilde{\mathbf{x}}_j$.

5. *Add* $\tilde{\mathbf{x}}_k$ *to the active set of regressors so* $\tilde{\mathbf{X}}^{(2)} = \left[\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k\right]$.

6. *Move* $\hat{\beta}^{(2)} = \left[\hat{\beta}_j \hat{\beta}_k\right]$ *towards the least squares estimate of regressing* $\varepsilon^{(1)}$ *on* $\tilde{\mathbf{X}}^{(2)}$ *until the correlation between* $\varepsilon^{(2)} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}^{(2)}\hat{\beta}^{(2)}$ *and some other* $\tilde{\mathbf{x}}_l$ *is equal to the correlation between* $\varepsilon^{(2)}$ *and* $\tilde{\mathbf{X}}^{(2)}$.

7. *Repeat steps 5 – 6 by adding regressors to the active set until all regressors have been added or n steps have been taken, whichever occurs first.*

The algorithm of LARS describes the statistical justification for the procedure – variables are added as soon as they have the largest correlation. Once the active set contains two or more regressors, the maximum correlation between the error and all regressors will be the same since regression coefficients are expanded in a manner that keeps the correlation identical between the error and any regressors in the active set. Efron, Hastie, et al. (2004) proposes a new algorithm that allows the entire path of LASSO, FSR, and LARS estimates to be quickly computed in models that contain a large number of candidate regressors. LASSO differs from LARS in one technical aspect, although they are very similar in practice.

These models are deeply related as shown Efron, Hastie, et al. (2004) and Hastie et al. (2007). All three can be used for model selection once a stopping rule (FSR, LARS) or the penalty (λ, LASSO) has been selected. k-fold cross-validation is commonly used to choose these values. Note that the usual standard OLS errors and t-stats are no longer correct since these estimators are constrained versions of least squares. Tibshirani (1996) proposes a *bootstrap* method that can be used to compute standard errors and make inference on LASSO estimators.[30]

Figure 3.12 illustrates how ridge regression and LASSO estimate parameters. Both show the OLS estimate $\hat{\beta}$ surrounded by ellipsoids the trace iso-SSE curves – that is, values of β_1 and β_2 that produce the same regression fit. The estimators are defined as the point where the smallest SSE is just tangent to the constraint. The ridge regression shrinks the estimate towards zero in a non-uniform way. This happens since the regressors are correlated. Ridge regression produces an estimate where both coefficients are non-zero. LASSO, on the other hand, estimates β_2 to be exactly. This happens since non-zero $\beta_1$provides a larger reduction in the SSE than β_2, at least near the point $(0,0)$. In general, ridge regression will never estimate any coefficients to be exactly 0 except when the OLS coefficient is exactly 0. LASSO frequently estimates coefficients to be zero since the cost of adding a small amount of a coefficient near zero is linear in β while the gain in terms of the SSE is quadratic in β (i.e., $\propto \beta^2$).

Figure 3.13 shows that paths of both the ridge regression and LASSO estimators are the restriction parameter ω is reduced. The model estimated regresses the return on the Big-High portfolio on the four factors, VWM^e, SMB, HML, and MOM. The paths begin with $\omega = 0$. As the constraint is relaxed, the parameters converge towards the OLS estimates, which limit cases as ω increases. There

[30]The standard errors subsequent to a selection procedure using GtS, StG, or IC are also not correct since tests have been repeated. In this regard, the bootstrap procedure should be more accurate since it accounts for the variation due to the selection, something not usually done in traditional model selection procedures.

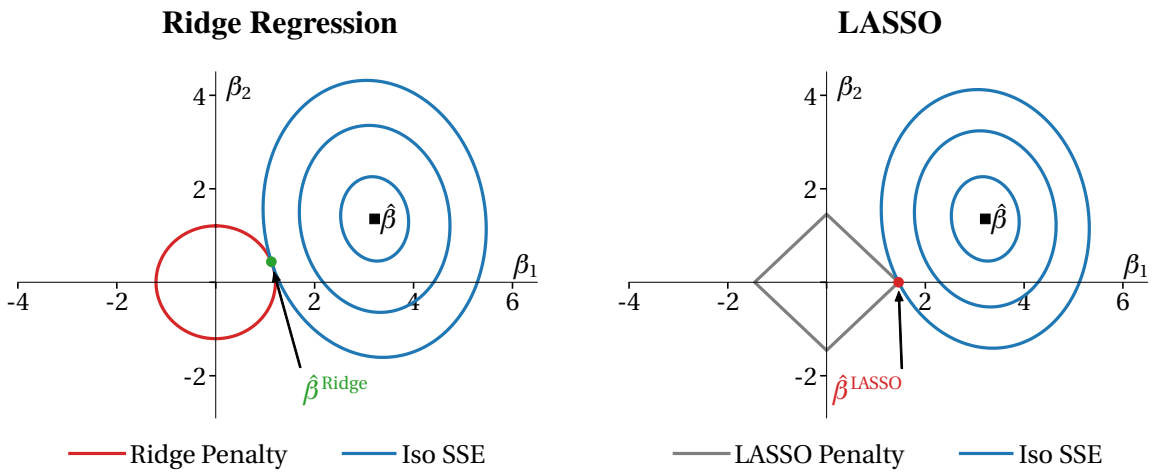

Figure 3.12: The left panel shows the ridge regression restriction for a specific value of ω along with three lines that trace combinations of β_1 and β_2 that produce the same model SSE. The ridge estimate is defined as the point where the SSE is just tangent to a restriction. The right shows the LASSO constraint along with the iso-SSE curves for the same data generating process.

is one clear distinction between the two paths. The paths from ridge regression evolve smoothly as ω increases. All coefficients except *SMB* are different from zero once $\omega > 1/8$. The LASSO paths have a distinct kinked shape. These kinks are points where the correlation between one excluded regressor and the included regressor(s) equalize so that the active set of regressors increases. The market is the strongest predictor, followed by the value factor. Momentum enters the model for small values of the penalty parameter, and size has a non-zero coefficient only at the OLS estimate (and then very small). The dashed line in each plot indicates that optimal choice $\omega^\star$ selected using 5-fold cross-validation. The cross-validated penalty parameter suggests that little shrinkage is needed. This occurs since the sample size is large enough that parameters, even small values, are precisely estimated.

3.14.5 Regression Trees and their Refinements

Regression Trees build models using only dummy variables. Constructing a regression tree begins by splitting the data into two groups using the values in regressors as possible split values. The model is constructed by splitting the observations into two groups using on all possible values of each regressor. The split that minimizes the SSE is retained, and the two groups are called leaves. The algorithm is then rerun on each leaf again, splitting on all possible values in each of the variables included in the model. This process of splitting into two leaves continues until either the homogeneity in the group as measures by the within-group MSE is sufficiently low, or the number of observations in a leaf falls below some prespecified value.

Figure 3.14 shows the first three levels of a model for the returns on the Big-High portfolio on the four factor portfolios. Splitting the data first on the market produced the largest gains, and the optimal split value was very near zero. The two leaves were then split according to the market into four groups corresponding to very low market returns (≤ -7.17), negative market returns ($-7.17 < VMW \leq -0.81$), positive market returns ($-0.81 < VMW \leq 3.78$), and very high market returns

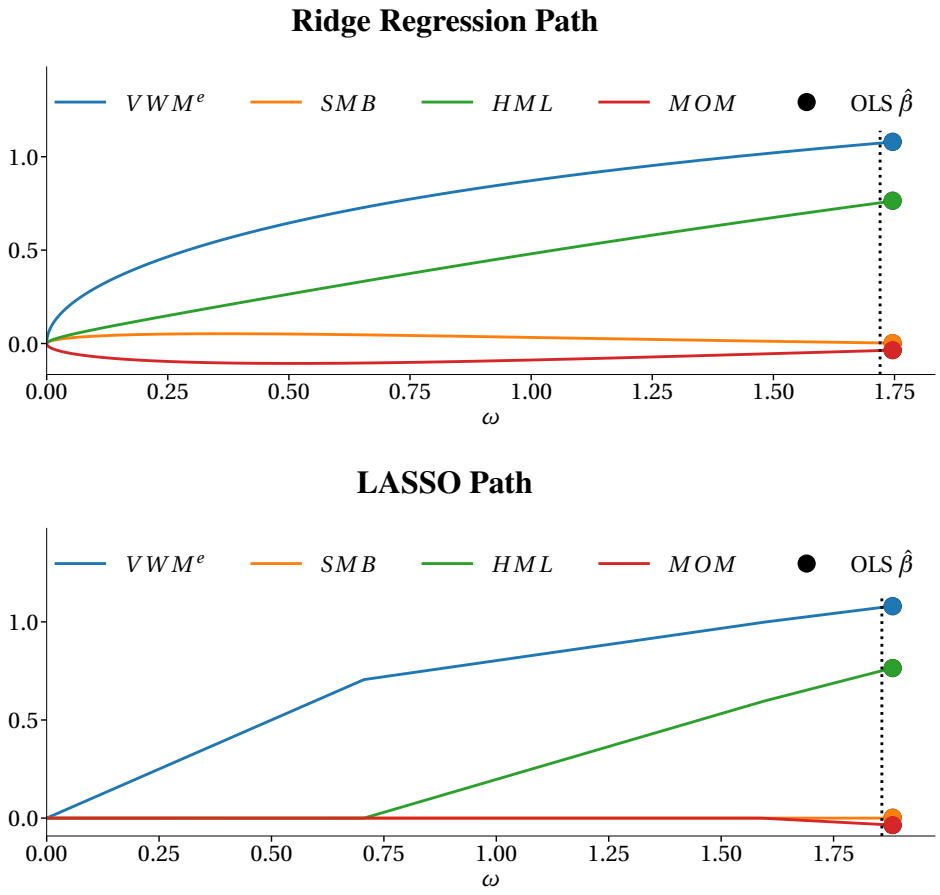

Figure 3.13: The top panel shows the path of the ridge regression estimates from the four factor model $BH^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. The penalty parameter ω is increased from zero to the value that produces the OLS estimate. The bottom panel contains the path of the LASSO estimates as the restriction is decreased. The kinks indicate points where a parameter switches from being exactly zero to a non-zero value.

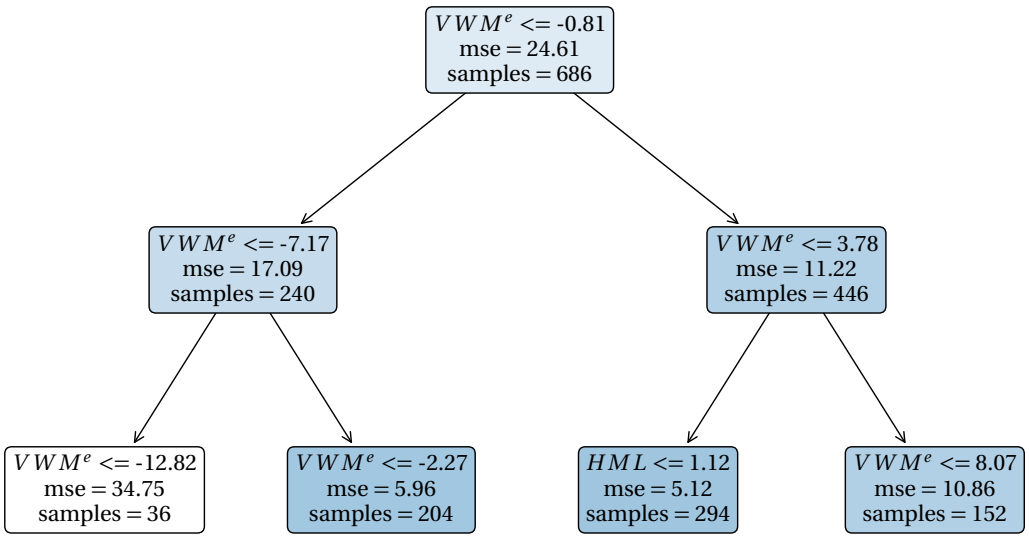

Figure 3.14: A regression tree where the left-hand-side variable is the return on the Big-High portfolio and the model is built using the four factors: VWM^e, SMB, HML, and MOM. The first and second splits used the market portfolio to bin the returns into four regions ranging from very low to very high. The final level splits used different variables so that the terminal leaves depend on both the market and the size factor.

($> 3.78\%$). If the tree was stopped at this node, the regression selected would be

$$BH^e = \beta_1 I_{[VWM_i^e \leq -7.17]} + \beta_2 I_{[-7.17 < VWM_i^e \leq -0.81]} + \beta_3 I_{[-0.81 < VWM_i^e \leq 3.78]} + \beta_4 I_{[VWM^e > 3.78]} + \varepsilon_i$$

The estimates of the parameters are simply the within-group means. The final level further splits the data into eight leaves (not shown). Three of the final level splits used the market return to split the negative returns further and to define an extreme positive return leaf. The other split preferred to use value. This final regression model would have eight terms constructed using combinations of restrictions on the return on the market factor and the return of the value factor.

Regression trees have step-function like behavior and frequently are not well suited to analyzing continuous-valued variables using continuously values regressors. While plain regression threes should usually be avoided, four refinements, pruning, Random Forests, bagging, and boosting all produce improvements in regression-tree models. Figure 3.15 compares a 2-level tree with OLS when modeling the return of the Big-High portfolio using the excess market return. The tree approximates the regression line as a step function. While this fit is not a terrible description of the data near 0, there are obvious deficiencies in the tails.

3.14.5.1 Improving Regression Trees

Three techniques are commonly used to improve regression trees: pruning, bagging, boosting, and Random Forests. Pruning a tree removes nodes that make a negligible improvement to the in-sample fit and often decrease out-of-sample fit. Pruning is implemented by optimizing the modified objective function

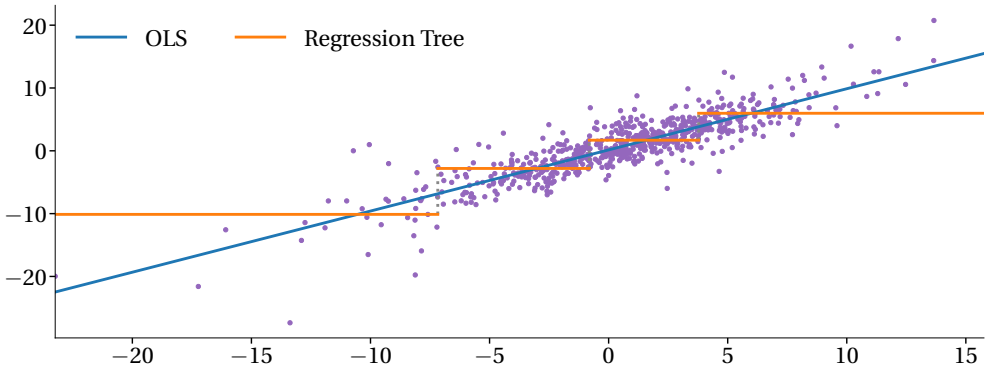

Figure 3.15: The regression tree implied by the first two splits and the OLS fit of the excess returns on the Big-High portfolio on the market.

$$\sum_{i=1}^{n}\left(Y_i-\hat{f}\left(\mathbf{x}_i\right)\right)^2+\alpha\left|T\right|$$

where $\hat{f}\left(\mathbf{x}_i\right)$ is the predicted value for a given tree and $|T|$is the number of terminal nodes in the tree. Pruning starts with a large tree with T_0nodes that is only terminated when either the number of nodes hits some threshold, the maximum number of levels is reached, or a SSE-based stopping criterion is met. For values of α on a grid of plausible values $\left\{\alpha_1<\alpha_2<\ldots<\alpha_q\right\}$ the tree that minimizes the modified objective function is selected. The preferred value of $\hat{\alpha}$ is chosen from this grid using k-fold cross-validation. Finally, the pruned tree is estimated by minimizing the modified objective function using $\hat{\alpha}$ on the original sample.

Bagging makes use of B bootstrap samples to the parameters of multiple trees. Each tree can then be used to generate predictions for any value of the regressor $\mathbf{x}$. These predictions are then be averaged to produce the bagged forecast. Note that each tree may have both a different structure and parameter values. While the forecasts will tend to be similar, they are not perfectly correlated, and the average forecast has a lower variance than any of the individual forecasts.

Algorithm 3.16 (Bagging Regression Trees)**.** *A bagged prediction from a regression tree is constructed following:*

1. *For $i=1,2,\ldots,B$ generate a bootstrap sample from $(Y_i,\mathbf{x}_i)$ and fit a regression tree to the bootstrapped sample.*

2. *Using the B trees, construct the forecast as*

$$\hat{f}\left(\mathbf{x}\right)=1/B\sum_{i=1}^{B}\hat{f}_i\left(\mathbf{x}\right)$$

where $f_i\left(\mathbf{x}\right)$ is the prediction from the tree estimated using bootstrap sample i.

Random Forests make use of randomization by selecting a subset of the available regressors when estimating a tree. When the number of regressors p is large, most trees will tend to have a very

similar structure even when fit to bootstrapped samples. This structure arises since strong predictors will always be selected in the first levels of the tree. The Random Forest solution is to estimate a tree using a bootstrap sample that also random selects $\approx \sqrt{p}$ regressors. This fitting of trees to random subsamples of the data is repeated many times, and the Random Forecast forecast is the average of forecasts of these models. The distinct trees tend to have low correlation, which translates into large gains from averaging.

Algorithm 3.17 (Random Forests)**.** *A Random Forest of regression trees is constructed following:*

1. *For $i = 1,2,\ldots,B$ generate a bootstrap sample of the data with a random subset of $k \approx \sqrt{p}$ regressors and fit a regression tree using the selected subset of the regressors.*

2. *Using the B trees, construct the forecast as*

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^{B} \tilde{f}_i(\mathbf{x})$$

where $\tilde{f}_i$ is the prediction using random regressor subset i.

Note that a Random Forest is identical to a bagged regression tree when $k = p$ regressors are used to build each tree.

Boosting also fits multiple trees, only sequentially to the same data. A boosted tree begins by fitting a small tree with d nodes to the data and computing the residuals. It then fits a new tree to the residuals. This is repeated many times. The trees are then combined using a tuning parameter λ as

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^{B} \lambda \ddot{f}_i(\mathbf{x})$$

where $\ddot{f}_1$is the tree fit to the original data and $\ddot{f}_j$, $j \geq 2$ is the prediction from the tree estimated using the residuals of the form

$$\hat{\varepsilon}_{i,j} = \hat{\varepsilon}_{i,j-1} - \lambda f_{j-1}(\mathbf{x}_i)$$

where $\hat{\varepsilon}_{i,0} = Y_i$.

Algorithm 3.18 (Bagging Regression Trees)**.** *Begin with $\hat{\varepsilon}_{i,0} = \tilde{Y}_i$ where $\tilde{Y}_i$ is the standardized version of Y_i. For $j = 1,\ldots,B$:*

1. *Fit a regression tree using $\left(\varepsilon_{i,j-1}, \mathbf{x}_i\right)$ with d splits and generate $\hat{\varepsilon}_{i,j} = \hat{\varepsilon}_{i,j-1} - \lambda \ddot{f}_j(\mathbf{x}_i)$ where $\ddot{f}_j$is the tree fit in iteration j.*

2. *Produce the boosted forecast as*

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^{B} \lambda \ddot{f}_i(\mathbf{X}).$$

Boosting makes uses of three tuning parameters, λ, d, and B. λ is usually set to some small value in the range $(0.001, 0.10)$. Small values of λ slow the learning since much of the forecast is down-weighted. d, the number of terminal nodes in a tree, is also set to some small number, often

1. d determines the maximum number of interactions allowed between the regressors when building the dummy-variable representation of a regression tree. Finally, B is usually set to some large value, often in the range of 1,000 – 10,000. These three parameters all interact and are substitutes – increases in one should usually be matched by decreases in the others when building optimal predictions. All three can be selected using a grid of values and k-fold cross-validation.

3.15 Projection

Least squares has one further justification: it is the best linear predictor of a dependent variable where best is interpreted to mean that it minimizes the mean square error (MSE). Suppose $f(\mathbf{x})$ is a function of only $\mathbf{x}$ and not Y. Mean squared error is defined

$$\mathrm{E}[(Y - f(\mathbf{x}))^2].$$

Assuming that it is permissible to differentiate under the expectations operator, the solution is

$$\mathrm{E}[Y - f(\mathbf{x})] = 0,$$

and, using the law of iterated expectations,

$$f(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}].$$

If $f(\mathbf{x})$ is restricted to include only linear functions of $\mathbf{x}$ then the problem simplifies to choosing β to minimize the MSE,

$$\mathrm{E}[(Y - \mathbf{x}\beta)^2]$$

and differentiating under the expectations (again, when possible),

$$\mathrm{E}[\mathbf{x}'(Y - \mathbf{x}\beta)] = \mathbf{0}$$

and $\hat{\beta} = \mathrm{E}[\mathbf{x}'\mathbf{x}]^{-1}\mathrm{E}[\mathbf{x}'\mathbf{y}]$. In the case where $\mathbf{x}$ contains a constant, this allows the best linear predictor to be expressed in terms of the covariance matrix of y and $\tilde{\mathbf{x}}$ where the˜indicates the constant has been excluded (i.e., $\mathbf{x} = [1\ \tilde{\mathbf{x}}]$), and so

$$\hat{\beta} = \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}$$

where the covariance matrix of $[Y\ \tilde{\mathbf{x}}]$ can be partitioned

$$\mathrm{Cov}([Y\ \tilde{\mathbf{x}}]) = \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{xy}}' & \Sigma_{yy} \end{bmatrix}$$

Recall from assumptions 3.7 that $\{\mathbf{x}_i, \varepsilon_i\}$ is a stationary and ergodic sequence and from assumption 3.8 that it has finite second moments and is of full rank. These two assumptions are sufficient to justify the OLS estimator as the best linear predictor of Y. Further, the OLS estimator can be used to make predictions for out of sample data. Suppose Y_{n+1} was an out-of-sample data point. Using the OLS procedure, the best predictor of Y_{n+1} (again, in the MSE sense), denoted $\hat{Y}_{n+1}$ is $\mathbf{x}_{n+1}\hat{\beta}$.

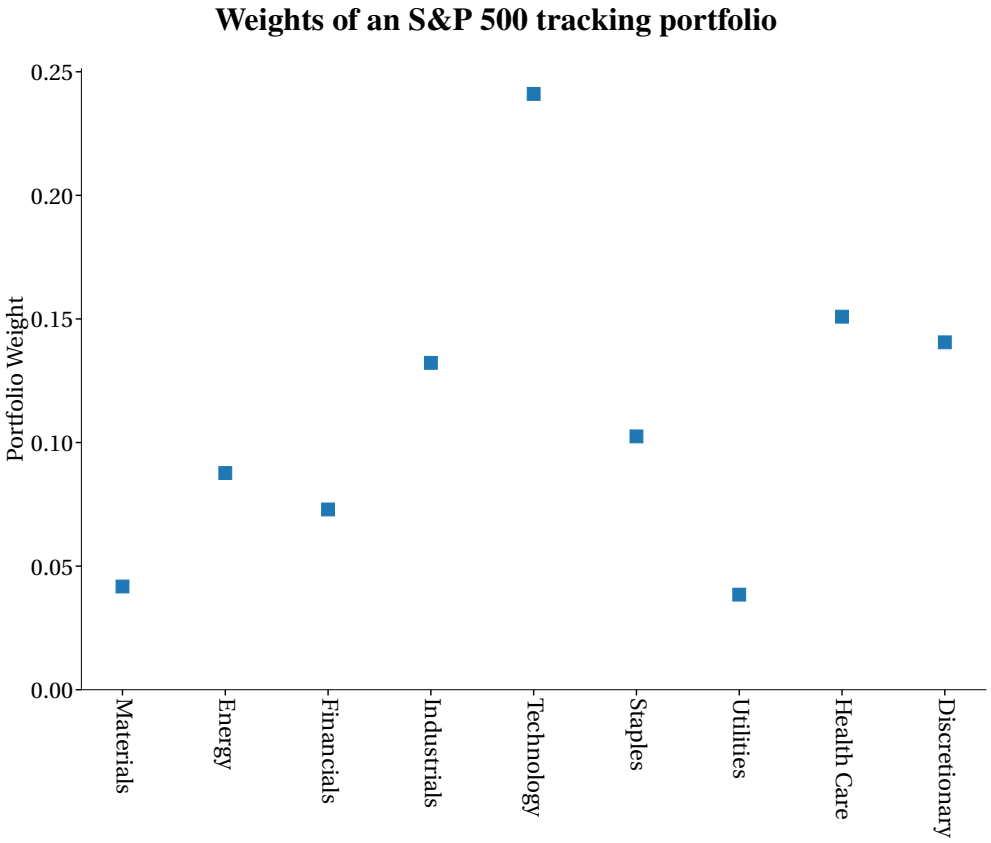

Figure 3.16: Plot of the optimal tracking portfolio weights. The optimal tracking portfolio is long all asset and no weight is greater than 25%.

3.15.1 Tracking Error Minimization

Consider the problem of setting up a portfolio that would generate returns as close as possible to the return on some index, for example, the FTSE 100. One option would be to buy the entire portfolio and perfectly replicate the portfolio. For other indices, such as the Wilshire 5000, which consists of many small and illiquid stocks, complete replication is impossible, and a tracking portfolio consisting of many fewer stocks must be created. One method to create the tracking portfolios is to find the best linear predictor of the index using a set of individual shares.

Let $\mathbf{x}_i$ be the returns on a set of assets and let Y_i be the return on the index. The tracking error problem is to minimize the

$$\mathrm{E}[(Y_i - \mathbf{X}_i\mathbf{w})^2]$$

where $\mathbf{w}$ is a vector of portfolio weights. Portfolio tracking has the same structure as the best linear predictor and the optimal weights are $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Data between January 5, 2010, and December 31, 2019, was used, a total of 2,515 trading days. The regression specification is simple: the return on the S&P is regressed on the returns on the sector

ETF returns,

$$R_i^{SP500} = \sum_{j=1}^{30} w_j R_{ij} + \varepsilon_i$$

where the portfolios are ordered alphabetically (not that this matters). The portfolio weights (which need not sum to 1) are presented in figure 3.16. All funds have positive weights, and the maximum just under 25%. More importantly, this portfolio has a correlation of 99.5% with the return on the S&P 500. Its return tracks the return of the S&P to within 1.4% per year. The tracking error variance is much smaller than the 14.7% annualized volatility of the S&P over this period.

While the regression estimates provide the solution to the unconditional tracking error problem, this estimator ignores two important considerations: how should stocks be selected, and how conditioning information (such as time-varying covariance) can be used. The first issue, which stocks to choose, is difficult and is typically motivated by the cost of trading and liquidity. The second issue will be re-examined using Multivariate GARCH and related models in a later chapter.

3.A Selected Proofs

Theorem 3.1.

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}|\mathbf{X}\right] &= \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}\right] \\
&= \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon|\mathbf{X}\right] \\
&= \beta + \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon|\mathbf{X}\right] \\
&= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathrm{E}\left[\varepsilon|\mathbf{X}\right] \\
&= \beta
\end{aligned}$$

□

Theorem 3.2.

$$\begin{aligned}
\mathrm{V}\left[\hat{\beta}|\mathbf{X}\right] &= \mathrm{E}\left[\left(\hat{\beta} - \mathrm{E}\left[\hat{\beta}|\mathbf{X}\right]\right)\left(\hat{\beta} - \mathrm{E}\left[\hat{\beta}|\mathbf{X}\right]\right)'|\mathbf{X}\right] \\
&= \mathrm{E}\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] \\
&= \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathrm{E}\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{I}_n\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

□

Theorem 3.3. Without loss of generality $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} + \mathbf{D}'$ where $\mathbf{D}'$ must satisfy $\mathbf{D}'\mathbf{X} = \mathbf{0}$ and $\mathrm{E}[\mathbf{D}'\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ since

$$\begin{aligned}\mathrm{E}\left[\tilde{\boldsymbol{\beta}}|\mathbf{X}\right] &= \mathrm{E}[\mathbf{C}\mathbf{y}|\mathbf{X}] \\ &= \mathrm{E}\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}'\right)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\mathbf{X}\right] \\ &= \boldsymbol{\beta} + \mathbf{D}'\mathbf{X}\boldsymbol{\beta} + \mathrm{E}\left[\mathbf{D}'\boldsymbol{\varepsilon}|\mathbf{X}\right]\end{aligned}$$

and by assumption $\mathbf{C}\mathbf{y}$ is unbiased and so $\mathrm{E}[\mathbf{C}\mathbf{y}|\mathbf{X}] = \boldsymbol{\beta}$.

$$\begin{aligned}\mathrm{V}\left[\tilde{\boldsymbol{\beta}}|\mathbf{X}\right] &= \mathrm{E}\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}'\right)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\left(\mathbf{D} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)|\mathbf{X}\right] \\ &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] + \mathrm{E}\left[\mathbf{D}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{D}|\mathbf{X}\right] + \mathrm{E}\left[\mathbf{D}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] + \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}\mathbf{D}|\mathbf{X}\right] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{D}'\mathbf{D} + \sigma^2\mathbf{D}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D} \\ &= \mathrm{V}\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right] + \sigma^2\mathbf{D}'\mathbf{D} + \mathbf{0} + \mathbf{0} \\ &= \mathrm{V}\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right] + \sigma^2\mathbf{D}'\mathbf{D}\end{aligned}$$

and so the variance of $\tilde{\boldsymbol{\beta}}$ is equal to the variance of $\hat{\boldsymbol{\beta}}$ plus a positive semi-definite matrix, and so

$$\mathrm{V}\left[\tilde{\boldsymbol{\beta}}|\mathbf{X}\right] - \mathrm{V}\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right] = \sigma^2\mathbf{D}'\mathbf{D} \geq \mathbf{0}$$

where the inequality is strict whenever $\mathbf{D} \neq \mathbf{0}$. □

Theorem 3.4.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

and so $\hat{\boldsymbol{\beta}}$ is a linear function of normal random variables $\boldsymbol{\varepsilon}$, and so it must be normal. Applying the results of Theorems 3.1 and 3.2 completes the proof. □

Theorem 3.5. □

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}_{\mathbf{X}}\mathbf{y} = \mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}$, and so

$$\begin{aligned}\mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\hat{\boldsymbol{\varepsilon}}'|\mathbf{X}\right] &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}_{\mathbf{X}}|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}\right]\mathbf{M}_{\mathbf{X}} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_{\mathbf{X}} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}_{\mathbf{X}}\mathbf{X})' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{0} \\ &= \mathbf{0}\end{aligned}$$

since $\mathbf{M}_{\mathbf{X}}\mathbf{X} = \mathbf{0}$ by construction. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are jointly normally distributed since both are linear functions of $\boldsymbol{\varepsilon}$, and since they are uncorrelated they are independent.[31]

[31] Zero correlation is, in general, insufficient to establish that two random variables are independent. However, when two random variables are jointly normally distribution, they are independent if and only if they are uncorrelated.

Theorem 3.6. $\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$ and so $(n-k)\,\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}$. $\hat{\varepsilon} = \mathbf{M}_\mathbf{X}\varepsilon$, so $(n-k)\,\hat{\sigma}^2 = \varepsilon'\mathbf{M}_\mathbf{X}'\mathbf{M}_\mathbf{X}\varepsilon$ and $(n-k)\,\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\varepsilon'\mathbf{M}_\mathbf{X}\varepsilon}{\sigma^2} = \frac{\varepsilon}{\sigma}'\mathbf{M}_\mathbf{X}\frac{\varepsilon}{\sigma} = \mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z}$ since $\mathbf{M}_\mathbf{X}$ is idempotent (and hence symmetric) where $\mathbf{z}$ is a n by 1 multivariate normal vector with covariance $\mathbf{I}_n$. Finally, applying the result in Lemma 3.1, $\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z} \sim \sum_{i=1}^{n}\lambda_i\chi^2_{1,i}$ where $\{\lambda_i\}$, $i = 1,2,\ldots,n$ are the eigenvalues of $\mathbf{M}_\mathbf{X}$ and $\chi^2_{1,i}$, $i = 1,2,\ldots,n$ are independent χ^2_1 random variables. Finally, note that $\mathbf{M}_\mathbf{X}$ is a rank $n-k$ idempotent matrix, so it must have $n-k$ eigenvalues equal to 1, $\lambda_i = 1$ for $i = 1,2,\ldots,n-k$ and k eigenvalues equal to 0, $\lambda_i = 0$ for $i = n-k+1,\ldots,n$, and so the distribution is a χ^2_{n-k}. □

Lemma 3.1 (Quadratic Forms of Multivariate Normals)**.** *Suppose* $\mathbf{z} \sim N(\mathbf{0},\Sigma)$ *where* Σ *is a* n *by* n *positive semi-definite matrix, and let* $\mathbf{W}$ *be a* n *by* n *positive semi-definite matrix, then*

$$\mathbf{z}'\mathbf{W}\mathbf{z} \sim N_2(\mathbf{0},\Sigma;\mathbf{W}) \equiv \sum_{i=1}^{n}\lambda_i\chi^2_{1,i}$$

where λ_i *are the eigenvalues of* $\Sigma^{\frac{1}{2}}\mathbf{W}\Sigma^{\frac{1}{2}}$ *and* $N_2(\cdot)$ *is known as a type-2 normal..*

This lemma is a special case of Baldessari (1967) as presented in White (Lemma 8.2, 1996).

Theorem 3.8. The OLS estimator is the BUE estimator since it is unbiased by Theorem 3.1 and it achieves the Cramer-Rao lower bound (Theorem 3.7). □

Theorem 3.9. Follows directly from the definition of a Student's t by applying Theorems 3.4, 3.5, and 3.2. □

Theorem 3.10. Follows directly from the definition of a F_{ν_1,ν_2} by applying Theorems 3.4, 3.5, and 3.2. □

Theorem 3.12.

$$\begin{aligned}\hat{\beta}_n - \beta &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon \\ &= \left(\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\varepsilon_i \\ &= \left(\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n}\right)^{-1}\frac{\sum_{i=1}^{n}\mathbf{x}_i'\varepsilon_i}{n}\end{aligned}$$

Since $\mathrm{E}[\mathbf{x}_i'\mathbf{x}_i]$ is positive definite by Assumption 3.8, and $\{\mathbf{x}_i\}$ is stationary and ergodic by Assumption 3.7, then $\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n}$ will be positive definite for n sufficiently large, and so $\hat{\beta}_n$ exists. Applying the Ergodic Theorem (Theorem 3.21), $\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n} \xrightarrow{a.s.} \Sigma_{\mathbf{XX}}$ and $\frac{\sum_{i=1}^{n}\mathbf{x}_i'\varepsilon_i}{n} \xrightarrow{a.s.} \mathbf{0}$ and by the Continuous Mapping Theorem (Theorem 3.22) combined with the continuity of the matrix inverse function, $\left(\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n}\right)^{-1} \xrightarrow{a.s.} \Sigma_{\mathbf{XX}}^{-1}$, and so

$$\begin{aligned}\hat{\beta}_n - \beta &= \left(\frac{\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i}{n}\right)^{-1}\frac{\sum_{i=1}^{n}\mathbf{x}_i'\varepsilon_i}{n} \\ &\xrightarrow{a.s.} \Sigma_{\mathbf{XX}}^{-1}\cdot\mathbf{0} \\ &\xrightarrow{a.s.} \mathbf{0}.\end{aligned}$$

Finally, almost sure convergence implies convergence in probability and so $\hat{\beta}_n - \beta \xrightarrow{p} 0$ or $\hat{\beta}_n \xrightarrow{p} \beta$. □

Theorem 3.21 (Ergodic Theorem)**.** *If $\{\mathbf{z}_t\}$ is ergodic and its r^{th} moment, μ_r, is finite, then*

$$T^{-1}\sum_{t=1}^{T}\mathbf{z}_t^r \xrightarrow{a.s.} \mu_r$$

.

Theorem 3.22 (Continuous Mapping Theorem)**.** *Given $\mathbf{g}:\mathbb{R}^k \to \mathbb{R}^l$, and any sequence of random k by 1 vectors $\{\mathbf{z}_n\}$ such that $\mathbf{z}_n \xrightarrow{a.s.} \mathbf{z}$ where $\mathbf{z}$ is k by 1, if $\mathbf{g}$ is continuous at $\mathbf{z}$, then $\mathbf{g}(\mathbf{z}_n) \xrightarrow{a.s.} \mathbf{g}(\mathbf{z})$.*

Theorem 3.13. See White (Theorem 5.25, 2000). □

Theorem 3.15.

$$\begin{aligned}
\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} &= \frac{\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n\right)}{n} \\
&= \frac{\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n\right)}{n} \\
&= \frac{\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n+\mathbf{X}\beta-\mathbf{X}\beta\right)'\left(\mathbf{y}-\mathbf{X}\hat{\beta}_n+\mathbf{X}\beta-\mathbf{X}\beta\right)}{n} \\
&= \frac{\left(\mathbf{y}-\mathbf{X}\beta+\mathbf{X}\left(\beta-\hat{\beta}_n\right)\right)'\left(\mathbf{y}-\mathbf{X}\beta+\mathbf{X}\left(\beta-\hat{\beta}_n\right)\right)}{n} \\
&= \frac{\left(\varepsilon+\mathbf{X}\left(\beta-\hat{\beta}_n\right)\right)'\left(\varepsilon+\mathbf{X}\left(\beta-\hat{\beta}_n\right)\right)}{n} \\
&= \frac{\varepsilon'\varepsilon}{n}+2\frac{\left(\beta-\hat{\beta}_n\right)'\mathbf{X}'\varepsilon}{n}+\frac{\left(\beta-\hat{\beta}_n\right)'\mathbf{X}'\mathbf{X}\left(\beta-\hat{\beta}_n\right)}{n}
\end{aligned}$$

By the Ergodic Theorem and the existence of $\mathrm{E}[\varepsilon_i^2]$ (Assumption 3.10), the first term converged to σ^2. The second term

$$\frac{\left(\beta-\hat{\beta}_n\right)'\mathbf{X}'\varepsilon}{n} = \left(\beta-\hat{\beta}_n\right)'\frac{\sum_{i=1}\mathbf{X}'\varepsilon}{n} \xrightarrow{p} \mathbf{0}'\mathbf{0} = 0$$

since $\hat{\beta}_n$ is consistent and $\mathrm{E}[\mathbf{x}_i\varepsilon_i] = 0$ combined with the Ergodic Theorem. The final term

$$\begin{aligned}
\frac{\left(\beta-\hat{\beta}_n\right)'\mathbf{X}'\mathbf{X}\left(\beta-\hat{\beta}_n\right)}{n} &= \left(\beta-\hat{\beta}_n\right)'\frac{\mathbf{X}'\mathbf{X}}{n}\left(\beta-\hat{\beta}_n\right) \\
&\xrightarrow{p} \mathbf{0}'\Sigma_{\mathbf{XX}}\mathbf{0} \\
&= 0
\end{aligned}$$

and so the variance estimator is consistent. □

Theorem 3.17.

$$\begin{aligned}
\hat{\beta}_{1n} &= \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{y}}{n} \\
\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'(\mathbf{X}_1+\mathbf{X}_2+\varepsilon)}{n} &= \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{X}_1}{n}+\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{X}_2}{n}+\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'\varepsilon}{n} \\
&\xrightarrow{p} \beta_1+\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\beta_2+\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\mathbf{0} \\
&= \beta_1+\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}\Sigma_{\mathbf{x}_1\mathbf{x}_2}\beta_2
\end{aligned}$$

where $\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\xrightarrow{p}\Sigma_{\mathbf{x}_1\mathbf{x}_1}^{-1}$ and $\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\xrightarrow{p}\Sigma_{\mathbf{x}_1\mathbf{x}_2}$ by the Ergodic and Continuous Mapping Theorems (Theorems 3.21 and 3.22). Finally note that

$$\begin{aligned}
\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{X}_2}{n} &= \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\left[\mathbf{X}_1\mathbf{x}_{2,1}\,\mathbf{X}_1\mathbf{x}_{2,2}\ldots\mathbf{X}_1\mathbf{x}_{2,k_2}\right] \\
&= \left[\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\mathbf{X}_1\mathbf{x}_{2,1}\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\mathbf{X}_1\mathbf{x}_{2,2}\ldots\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n}\right)^{-1}\mathbf{X}_1\mathbf{x}_{2,k_2}\right] \\
&= \left[\hat{\delta}_{1n}\,\hat{\delta}_{2n}\ldots\hat{\delta}_{k_2n}\right]
\end{aligned}$$

where δ_j is the regression coefficient in $\mathbf{x}_{2,j}=\mathbf{X}\delta_j+\eta_j$. □

Theorem 3.18. See White (Theorem 6.3, 2000). □

Theorem 3.19. See White (Theorem 6.4, 2000). □

Theorem 3.20. By Assumption 3.15,

$$\mathbf{V}^{-\frac{1}{2}}\mathbf{y}=\mathbf{V}^{-\frac{1}{2}}\mathbf{X}\beta+\mathbf{V}^{-\frac{1}{2}}\varepsilon$$

and $\mathrm{V}\left[\mathbf{V}^{-\frac{1}{2}}\varepsilon\right]=\sigma^2\mathbf{I}_n$, uncorrelated and homoskedastic, and so Theorem 3.3 can be applied. □

Shorter Problems

Problem 3.1. Derive the OLS estimator for the model $Y_i=\alpha+\varepsilon_i$.

Problem 3.2. Derive the OLS estimator for the model $Y_i=\beta X_i+\varepsilon_i$.

Problem 3.3. What are information criteria and how are they used?

Problem 3.4. Outline the steps to compute the bootstrap variance estimator for a regression when the data are heteroskedastic.

Problem 3.5. Discuss White's covariance estimator, and in particular when should White's covariance estimator be used? What are the consequences to using White's covariance estimator when it is not needed? How can one determine if White's covariance estimator is needed?

Problem 3.6. Suppose $Z_i=a+bX_i$, and two models are estimated using OLS: $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$ and $Y_i=\gamma_0+\gamma_1Z_i+\eta_i$, What the relationship between γ and β and between $\hat{\varepsilon}_i$ and $\hat{\eta}_i$?

Problem 3.7. Describe the steps to implement k-fold cross-validation in a regression to select a model.

Longer Exercises

Exercise 3.1. Imagine you have been given the task of evaluating the relationship between the return on a mutual fund and the number of years its manager has been a professional. You have a panel data set which covers all of the mutual funds returns in the year 1970-2005. Consider the regression

$$R_{i,t} = \alpha + \beta exper_{i,t} + \varepsilon_{i,t}$$

where r_{it} is the return on fund i in year t and $exper_{it}$ is the number of years the fund manager has held her job in year t. The initial estimates of β and α are computed by stacking all of the observations into a vector and running a single OLS regression (across all funds and all time periods).

1. What test statistic would you use to determine whether experience has a positive effect?
2. What are the null and alternative hypotheses for the above test?
3. What does it mean to make a type I error in the above test? What does it mean to make a type II error in the above test?
4. Suppose that experience has no effect on returns but that unlucky managers get fired and thus do not gain experience. Is this a problem for the above test? If so, can you comment on its likely effect?
5. Could the estimated $\hat{\beta}$ ever be valid if mutual funds had different risk exposures? If so, why? If not, why not?
6. If mutual funds do have different risk exposures, could you write down a model which may be better suited to testing the effect of managerial experience than the initial simple specification? If it makes your life easier, you can assume there are only 2 mutual funds and 1 risk factor to control for.

Exercise 3.2. Consider the linear regression

$$Y_t = \beta X_t + \varepsilon_t$$

1. Derive the least-squares estimator. What assumptions are you making in the derivation of the estimator?
2. Under the classical assumptions, derive the variance of the estimator $\hat{\beta}$.
3. Suppose the errors ε_t have an AR(1) structure where $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ where $\eta_t \xrightarrow{d} N(0,1)$ and $|\rho| < 1$. What is the variance of $\hat{\beta}$ now?
4. Now suppose that the errors have the same AR(1) structure but the x_t variables are i.i.d.. What is the variance of $\hat{\beta}$ now?
5. Finally, suppose the linear regression is now

   $$Y_t = \alpha + \beta X_t + \varepsilon_t$$

   where ε_t has an AR(1) structure and that x_t is i.i.d.. What is the covariance of $[\alpha \ \beta]'$?

Exercise 3.3. Consider the simple regression model $Y_i = \beta X_{1,i} + \varepsilon_i$ where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the $X_{1,i}$.

1. Show that the OLS estimator of β is consistent.
2. Is the previously derived OLS estimator of β still consistent if $Y_i = \alpha + \beta X_{1,i} + \varepsilon_i$? Show why or why not.
3. Now suppose the data generating process is

   $$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

   Derive the OLS estimators of β_1 and β_2.
4. Derive the asymptotic covariance of this estimator using the method of moments approach.
   (a) What are the moment conditions?
   (b) What is the Jacobian?
   (c) What does the Jacobian limit to? What does this require?
   (d) What is the covariance of the moment conditions. Be as general as possible.

In all of the above, clearly state any additional assumptions needed.

Exercise 3.4. Let $\hat{\mathbf{S}}$ be the sample covariance matrix of $\mathbf{z} = [\mathbf{y}\ \mathbf{X}]$, where $\mathbf{X}$ *does not include a constant*

$$\hat{\mathbf{S}} = n^{-1} \sum_{i=1}^{n} (\mathbf{z}_i - \bar{\mathbf{z}})'(\mathbf{z}_i - \bar{\mathbf{z}})$$

$$\hat{\mathbf{S}} = \begin{bmatrix} \hat{s}_{yy} & \hat{\mathbf{s}}_{xy}' \\ \hat{\mathbf{s}}_{xy} & \hat{\mathbf{S}}_{xx} \end{bmatrix}$$

and suppose n, the sample size, is known ($\hat{\mathbf{S}}$ is the sample covariance estimator). Under the small-sample assumptions (including homoskedasticity and normality if needed), describe one method, using only $\hat{\mathbf{S}}$, $\bar{\mathbf{X}}$ (the 1 by $k-1$ sample mean of the matrix $\mathbf{X}$, column-by-column), $\bar{y}$ and n, to

1. Estimate $\hat{\beta}_1, \ldots, \hat{\beta}_k$ from a model

   $$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

2. Estimate s, the standard error of the regression
3. Test $H_0 : \beta_j = 0$, $j = 2, \ldots, k$

Exercise 3.5. Consider the regression model

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the x_i. Also assume that x_i is i.i.d. with mean μ_x and variance σ_x^2, both finite.

1. Using scalar notation, derive the OLS estimators of β_1 and β_2.
2. Show these estimators are consistent. Are any further assumptions needed?
3. Show that the matrix expression for the estimator of the regression parameters, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is identical to the estimators derived using scalar notation.

Exercise 3.6. Let $\mathbf{x}_m\beta$ be the best linear projection of Y_m. Let ε_m be the prediction error.

1. What is the variance of a projected Y?
2. What is the variance if the βs are estimated using regressors that do not include observation m (and hence not $\mathbf{x}_m$ or ε_m)? Hint: You can use any assumptions in the notes, just be clear what you are assuming.

Exercise 3.7. Are Wald tests of linear restrictions in a linear regression invariant to linear reparameterizations? Hint: Let $\mathbf{F}$ be an invertible matrix. Parameterize W in the case where $H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0}$ and $H_0 : \mathbf{F}(\mathbf{R}\beta - \mathbf{r}) = \mathbf{F}\mathbf{R}\beta - \mathbf{F}\mathbf{r} = \mathbf{0}$.

1. Are they the same?
2. Show that $n \cdot \mathrm{R}^2$ has an asymptotic χ^2_{k-1} distribution under the classical assumptions when the model estimated is
$$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$
Hint: What is the does the distribution of c/ν converge to as $\nu \to \infty$ when $c \sim \chi^2_\nu$.

Exercise 3.8. Suppose an unrestricted model is
$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i$$

1. Sketch the steps required to test a null $H_0 : \beta_2 = \beta_3 = 0$ in the large-sample framework using a Wald test and an LM test.
2. Sketch the steps required to test a null $H_0 : \beta_2 + \beta_3 + \beta_4 = 1$ in the small-sample framework using a Wald test, a t-test, an LR test, and an LM test.

In the above questions be clear what the null and alternative are, which regressions must be estimated, how to compute any numbers that are needed and the distribution of the test statistic.

Exercise 3.9. Let Y_i and X_i conform to the small-sample assumptions and let $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. Define another estimator
$$\breve{\beta}_2 = \frac{\bar{Y}_H - \bar{Y}_L}{\bar{X}_H - \bar{X}_L}$$
where $\bar{X}_H$ is the average value of X_i given $X_i >$ median $(\mathbf{x})$, and $\bar{Y}_H$ is the average value of Y_i for n such that $X_i >$ median $(\mathbf{x})$. $\bar{X}_L$ is the average value of X_i given $X_i \leq$ median $(\mathbf{x})$, and $\bar{Y}_L$ is the average value of Y_i for n such that $X_i \leq$ median $(\mathbf{x})$ (both $\bar{X}$ and $\bar{Y}$ depend on the order of X_i, and not Y_i). For example, suppose the X_i were ordered such that $X_1 < X_2 < X_3 < \ldots < X_i$ and n is even. Then,
$$\bar{X}_L = \frac{2}{n}\sum_{i=1}^{n/2} X_i$$

and

$$\bar{X}_H = \frac{2}{n}\sum_{i=n/2+1}^{n} X_i$$

1. Is $\breve{\beta}_2$ unbiased, conditional on $\mathbf{X}$?
2. Is $\breve{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
3. What is the variance of $\breve{\beta}_2$, conditional on $\mathbf{X}$?

Exercise 3.10. Suppose

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and that variable Z_i is available where $\mathrm{V}[Z_i] = \sigma_z^2 > 0$, $\mathrm{Corr}(X_i, Z_i) = \rho \neq 0$ and $\mathrm{E}[\varepsilon_i|\mathbf{z}] = 0$, $n = 1,\ldots,N$. Further suppose the other assumptions of the small-sample framework hold. Rather than the usual OLS estimator,

$$\ddot{\beta}_2 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})Y_i}{\sum_{i=1}^{n}(Z_i - \bar{Z})X_i}$$

is used.

1. Is $\ddot{\beta}_2$ a reasonable estimator for β_2?
2. What is the variance of $\ddot{\beta}_2$, conditional on $\mathbf{x}$ and $\mathbf{z}$?
3. What does the variance limit to (i.e., not conditioning on $\mathbf{x}$ and $\mathbf{z}$)?
4. How is this estimator related to OLS, and what happens to its variance when OLS is used (Hint: What is $\mathrm{Corr}(X_i, X_i)$?)

Exercise 3.11. Let $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ conform to the small-sample assumptions and let $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. Define the estimator

$$\breve{\beta}_2 = \frac{\bar{Y}_H - \bar{Y}_L}{\bar{X}_H - \bar{X}_L}$$

where $\bar{X}_H$ is the average value of X_i given $X_i > \mathrm{median}(\mathbf{x})$, and $\bar{Y}_H$ is the average value of Y_i for i such that $X_i > \mathrm{median}(\mathbf{x})$. $\bar{X}_L$ is the average value of X_i given $X_i \leq \mathrm{median}(\mathbf{x})$, and $\bar{Y}_L$ is the average value of Y_i for i such that $X_i \leq \mathrm{median}(\mathbf{x})$ (both $\bar{X}$ and $\bar{Y}$ depend on the order of X_i, and not Y_i). For example, suppose the X_i were ordered such that $X_1 < X_2 < X_3 < \ldots < X_n$ and n is even. Then,

$$\bar{X}_L = \frac{2}{n}\sum_{i=1}^{n/2} X_i$$

and

$$\bar{X}_H = \frac{2}{n}\sum_{i=n/2+1}^{n} X_i$$

1. Is $\breve{\beta}_2$ unbiased, conditional on $\mathbf{X}$?

2. Is $\breve{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
3. What is the variance of $\breve{\beta}_2$, conditional on $\mathbf{X}$?

   Next consider the estimator
   $$\ddot{\beta}_2 = \frac{\bar{Y}}{\bar{X}}$$
   where $\bar{Y}$ and $\bar{X}$ are sample averages of $\{Y_i\}$ and $\{X_i\}$, respectively.
4. Is $\ddot{\beta}_2$ unbiased, conditional on $\mathbf{X}$?
5. Is $\ddot{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
6. What is the variance of $\ddot{\beta}_2$, conditional on $\mathbf{X}$?

Exercise 3.12. Suppose an unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i$$

1. Discuss which features of estimators each of the three major tests, Wald, Likelihood Ratio, and Lagrange Multiplier, utilize in testing.
2. Sketch the steps required to test a null $H_0 : \beta_2 = \beta_3 = 0$ in the large-sample framework using Wald, LM, and LR tests.
3. What are type I & II errors?
4. What is the size of a test?
5. What is the power of a test?
6. What influences the power of a test?
7. What is the most you can say about the relative power of a Wald, LM, and LR test of the same null?

Exercise 3.13. Consider the regression model

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the X_i. Also assume that X_i is i.i.d. with mean μ_x and variance σ_x^2, both finite.

1. Using scalar notation, derive the OLS estimators of β_1 and β_2.
2. Why are these estimators are consistent? Are any further assumptions needed?
3. Show that the matrix expression for the estimator of the regression parameters, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is identical to the estimators derived using scalar notation.

4. Suppose instead
$$Y_i = \gamma_1 + \gamma_2 (X_i - \bar{X}) + \varepsilon_i$$
was fit to the data. How are the estimates of the γs related to the βs?

5. What can you say about the relationship between the t-statistics of the γs and the βs?

6. How would you test for heteroskedasticity in the regression?

7. Since the errors are i.i.d. there is no need to use White's covariance estimator for this regression. What are the consequences of using White's covariance estimator if it is not needed?

Exercise 3.14. Suppose $Y_i = \alpha + \beta X_i + \varepsilon_i$ where $\mathrm{E}[\varepsilon_i | X] = 0$ and $\mathrm{V}[\varepsilon_i] = \sigma^2$ for all i .

1. Derive the OLS estimators of α and β.

2. Describe the trade-offs when deciding whether to use the classic parameter covariance estimator, $\hat{\sigma}^2 \Sigma_{XX}^{-1}$, and White's parameter covariance estimator, $\Sigma_{XX}^{-1} \mathbf{S} \Sigma_{XX}^{-1}$?

3. Describe a procedure to formally test whether White's covariance estimator is required.

4. Suppose the true model is as above, but instead the model $Y_i = \gamma + \varepsilon_i$ is fit. What is the most you can say about the the OLS estimate of $\hat{\gamma}$?

5. What is Windsorization in the context of a regression, and how is it useful?

Exercise 3.15. Consider the APT regression
$$R_t^e = \alpha + \beta_m R_{m,t}^e + \beta_s R_{smb,t} + \beta_v R_{hml,t} + \varepsilon_t$$
where $R_{m,t}^e$ is the excess return on the market, $R_{smb,t}$ is the return on the size factor, $R_{hml,t}$ is the return on value factor and R_t^e is an excess return on a portfolio of assets. Using the information provided in the tables below below, answer the following questions:

1. Is there evidence that this portfolio is market neutral?

2. Are the size and value factors needed in this portfolio to adequately capture the cross-sectional dynamics?

3. Is there evidence of conditional heteroskedasticity in this model?

4. What are the trade-offs for choosing a covariance estimator for making inference on this model?

5. Define the size and power of a statistical test.

6. What factors affect the power of a statistical test?

7. Outline the steps to implement the correct bootstrap covariance estimator for these parameters. Justify the method you chose using the information provided.

Notes: All models were estimated on $n = 100$ data points. Models 1 and 2 correspond to the specification above. In model 1 R_{smb} and R_{hml} have been excluded. Model 3, 4 and 5 are all version of

$$\begin{aligned}\hat{\varepsilon}_t^2 = \gamma_0 + \gamma_1 R_{m,t}^e + \gamma_2 R_{smb,t} + \gamma_3 R_{hml,t} + \gamma_4 \left(R_{m,t}^e\right)^2 + \gamma_5 R_{m,t}^e R_{smb,t} \\ + \gamma_6 R_{m,t}^e R_{hml,t} + \gamma_7 R_{smb,t}^2 + \gamma_8 R_{smb,t} R_{hml,t} + \gamma_9 R_{hml,t}^2 + \eta_t\end{aligned}$$

$\hat{\varepsilon}_t$ was computed using Model 1 for the results under Model 3, and using model 2 for the results under Models 4 and 5. R^2 is the R-squared and n is the number of observations.

Parameter Estimates

	Model 1	Model 2		Model 3	Model 4	Model 5
α	0.128	0.089	γ_0	0.984	0.957	0.931
β_m	1.123	0.852	γ_1	-0.779	-0.498	
β_{smb}		0.600	γ_2		-0.046	
β_{hml}		-0.224	γ_3		0.124	
			γ_4	0.497	0.042	0.295
			γ_5		0.049	
			γ_6		0.684	
			γ_7		0.036	-0.149
			γ_8		-0.362	
			γ_9		-0.005	0.128
R^2	0.406	0.527		0.134	0.126	0.037

Parameter Covariance Estimates

The estimated covariance matrices from the asymptotic distribution

$$\sqrt{n}\left(\hat{\beta}-\hat{\beta}_0\right) \xrightarrow{d} N(0,C)$$

are below where C is either $\hat{\sigma}^2\hat{\Sigma}_{XX}^{-1}$ or $\hat{\Sigma}_{XX}^{-1}\hat{S}\hat{\Sigma}_{XX}^{-1}$.

CAP-M

$\hat{\sigma}^2\hat{\Sigma}_{XX}^{-1}$

	α	β_m
α	1.365475	0.030483
β_m	0.030483	1.843262

$\hat{\Sigma}_{XX}^{-1}\hat{S}\hat{\Sigma}_{XX}^{-1}$

	α	β_m
α	1.341225	-0.695235
β_m	-0.695235	2.747142

Fama-French Model

$\hat{\sigma}^2\hat{\Sigma}_{XX}^{-1}$

	α	β_m	β_{smb}	β_{hml}
α	1.100680	0.103611	-0.088259	-0.063529
β_m	0.103611	1.982761	-0.619139	-0.341118
β_{smb}	-0.088259	-0.619139	1.417318	-0.578388
β_{hml}	-0.063529	-0.341118	-0.578388	1.686200

$\hat{\Sigma}_{XX}^{-1}\hat{S}\hat{\Sigma}_{XX}^{-1}$

	α	β_m	β_{smb}	β_{hml}
α	1.073227	-0.361618	-0.072784	0.045732
β_m	-0.361618	2.276080	-0.684809	0.187441
β_{smb}	-0.072784	-0.684809	1.544745	-1.074895
β_{hml}	0.045732	0.187441	-1.074895	1.947117

χ^2_m critical values

Critical value for a 5% test when the test statistic has a χ^2_m distribution.

m	1	2	3	4	8	9	10
Crit Val.	3.84	5.99	7.81	9.48	15.50	16.91	18.30

m	90	91	98	99	100
Crit Val.	113.14	114.26	122.10	123.22	124.34

Matrix Inverse

The inverse of a 2 by 2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Exercise 3.16. Suppose $Y_i = \alpha + \beta X_i + \varepsilon_i$ where $\mathrm{E}[\varepsilon_i|X] = 0$ and $\mathrm{V}[\varepsilon_i] = \sigma^2$ for all i.

1. Describe the trade-offs when deciding whether to use the classic parameter covariance estimator, $\hat{\sigma}^2\Sigma_{XX}^{-1}$, and White's parameter covariance estimator, $\Sigma_{XX}^{-1}\mathbf{S}\Sigma_{XX}^{-1}$?
2. Describe a procedure to formally test whether White's covariance estimator is required.
3. Suppose the true model is as above, but instead the model $Y_i = \gamma + \varepsilon_i$ is fit. What is the most you can say about the the OLS estimate of $\hat{\gamma}$?
4. Define the size and power of a statistical test.
5. What factors affect the power of a statistical test?
6. What is Windsorization in the context of a regression, and how is it useful?