Chapter 2

Estimation, Inference, and Hypothesis Testing

Note: The primary reference for these notes is Ch. 7 and 8 of Casella and Berger (2001). This text may be challenging if new to this topic and Ch. 7 – 10 of Wackerly, Mendenhall, and Scheaffer (2001) may be useful as an introduction.

This chapter provides an overview of estimation, distribution theory, inference, and hypothesis testing. Testing an economic or financial theory is a multi-step process. First, any unknown parameters must be estimated. Next, the distribution of the estimator must be determined. Finally, formal hypothesis tests must be conducted to examine whether the data are consistent with the theory. This chapter is intentionally "generic" by design and focuses on the case where the data are independent and identically distributed. Properties of specific models will be studied in detail in the chapters on linear regression, time series, and univariate volatility modeling.

Three steps must be completed to test the implications of an economic theory:

- Estimate unknown parameters
- Determine the distributional of estimator
- Conduct hypothesis tests to examine whether the data are compatible with a theoretical model

This chapter covers each of these steps with a focus on the case where the data is independent and identically distributed (i.i.d.). The heterogeneous but independent case will be covered in the chapter on linear regression and the dependent case will be covered in the chapters on time series.

2.1 Estimation

Once a model has been specified and hypotheses postulated, the first step is to estimate the parameters of the model. Many methods are available to accomplish this task. These include

parametric, semi-parametric, semi-nonparametric and nonparametric estimators and a variety of estimation methods often classified as M-, R- and L-estimators.[1]

Parametric models are tightly parameterized and have desirable statistical properties when their specification is correct, such as providing consistent estimates with small variances. Nonparametric estimators are more flexible and avoid making strong assumptions about the relationship between variables. This allows nonparametric estimators to capture a wide range of relationships but comes at the cost of precision. In many cases, nonparametric estimators are said to have a *slower rate of convergence* than similar parametric estimators. The practical consequence of the rate is that nonparametric estimators are desirable when there is a proliferation of data and the relationships between variables may be difficult to postulate *a priori*. In situations where less data is available, or when an economic model proffers a relationship among variables, parametric estimators are generally preferable.

Semi-parametric and semi-nonparametric estimators bridge the gap between fully parametric estimators and nonparametric estimators. Their difference lies in "how parametric" the model and estimator are. Estimators which postulate parametric relationships between variables but estimate the underlying distribution of errors flexibly are known as semi-parametric. Estimators which take a stand on the distribution of the errors but allow for flexible relationships between variables are semi-nonparametric. This chapter focuses exclusively on parametric models and estimators. This choice is more reflective of the common practice than a critique of nonparametric methods.

Another important characterization of estimators is whether they are members of the M-, L- or R-estimator classes.[2] M-estimators (also known as extremum estimators) always involve maximizing or minimizing some objective function. M-estimators are the most commonly used class in financial econometrics and include maximum likelihood, regression, classical minimum distance and both the classical and the generalized method of moments. L-estimators, also known as linear estimators, are a class where the estimator can be expressed as a linear function of ordered data. Members of this family can always be written as

$$\sum_{i=1}^{n} w_i y_i$$

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 also includes the median by setting $w_i = 0$ for all i except $w_j = 1$ where $j = (n+1)/2$ (n is odd) or $w_j = w_{j+1} = 1/2$ where $j = n/2$ (n is even). R-estimators exploit the *rank* of the data. Common examples of R-estimators include the minimum, maximum and Spearman's rank correlation, which is the usual correlation estimator on the ranks of the data rather than on the data themselves. Rank statistics are often robust to outliers and non-linearities.

[1] There is another important dimension in the categorization of estimators: Bayesian or frequentist. Bayesian estimators make use of Bayes rule to perform inference about unknown quantities – parameters – conditioning on the observed data. Frequentist estimators rely on randomness averaging out across observations. Frequentist methods are dominant in financial econometrics although the use of Bayesian methods has been recently increasing.

[2] Many estimators are members of more than one class. For example, the median is a member of all three.

2.1.1 M-Estimators

The use of M-estimators is pervasive in financial econometrics. Three common types of M-estimators include the method of moments, both classical and generalized, maximum likelihood and classical minimum distance.

2.1.2 Maximum Likelihood

Maximum likelihood uses the distribution of the data to estimate any unknown parameters by finding the values which make the data as likely as possible to have been observed – in other words, by maximizing the likelihood. Maximum likelihood estimation begins by specifying the *joint* distribution, $f(\mathbf{y};\boldsymbol{\theta})$, of the observable data, $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, as a function of a k by 1 vector $\boldsymbol{\theta}$ which contains all parameters. Note that this is the joint density, and so it includes both the information in the marginal distributions of y_i and information relating the marginals to one another.[3] Maximum likelihood estimation "reverses" the likelihood to express the probability of $\boldsymbol{\theta}$ in terms of the observed $\mathbf{y}$, $L(\boldsymbol{\theta};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta})$.

The maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$, is defined as the solution to

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{y}) \tag{2.1}$$

where argmax is used in place of max to indicate that the maximum may not be unique – it could be set valued – and to indicate that the *global* maximum is required.[4] Since $L(\boldsymbol{\theta};\mathbf{y})$ is strictly positive, the log of the likelihood can be used to estimate $\boldsymbol{\theta}$.[5] The log-likelihood is defined as $l(\boldsymbol{\theta};\mathbf{y}) = \ln L(\boldsymbol{\theta};\mathbf{y})$. In most situations the maximum likelihood estimator (MLE) can be found by solving the k by 1 score vector,

$$\frac{\partial\, l(\boldsymbol{\theta};\mathbf{y})}{\partial\, \boldsymbol{\theta}} = \mathbf{0}$$

although a score-based solution does not work when $\boldsymbol{\theta}$ is constrained and $\hat{\boldsymbol{\theta}}$ lies on the boundary of the parameter space or when the permissible range of values for y_i depends on $\boldsymbol{\theta}$. The first problem is common enough that it is worth keeping in mind. It is particularly common when working with variances which must be (weakly) positive by construction. The second issue is fairly rare in financial econometrics.

[3] Formally the relationship between the marginal is known as the *copula*. Copulas and their use in financial econometrics will be explored in the second term.

[4] Many likelihoods have more than one maximum (i.e. local maxima). The maximum likelihood estimator is always defined as the global maximum.

[5] Note that the log transformation is strictly increasing and globally concave. If $z^\star$ is the maximum of $g(z)$, and thus

$$\left.\frac{\partial\, g(z)}{\partial\, z}\right|_{z=z^\star} = 0$$

then $z^\star$ must also be the maximum of $\ln(g(z))$ since

$$\left.\frac{\partial \ln(g(z))}{\partial\, z}\right|_{z=z^\star} = \left.\frac{g'(z)}{g(z)}\right|_{z=z^\star} = \frac{0}{g(z^\star)} = 0$$

which follows since $g(z) > 0$ for any value of z.

2.1.2.1 Maximum Likelihood Estimation of a Poisson Model

Realizations from a Poisson process are non-negative and discrete. The Poisson is common in ultra-high-frequency econometrics where the usual assumption that prices lie in a continuous space is implausible. For example, trade prices of US equities evolve on a grid of prices typically separated by \$0.01. Suppose $y_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. The pdf of a single observation is

$$f(y_i;\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \tag{2.2}$$

and since the data are independent and identically distributed (i.i.d.), the joint likelihood is simply the product of the n individual likelihoods,

$$f(\mathbf{y};\lambda) = L(\lambda;\mathbf{y}) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}.$$

The log-likelihood is

$$l(\lambda;\mathbf{y}) = \sum_{i=1}^{n} -\lambda + y_i \ln(\lambda) - \ln(y_i!) \tag{2.3}$$

which can be further simplified to

$$l(\lambda;\mathbf{y}) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} y_i - \sum_{j=1}^{n} \ln(y_j!)$$

The first derivative is

$$\frac{\partial l(\lambda;\mathbf{y})}{\partial \lambda} = -n + \lambda^{-1}\sum_{i=1}^{n} y_i. \tag{2.4}$$

The MLE is found by setting the derivative to 0 and solving,

$$\begin{aligned}
-n + \hat{\lambda}^{-1}\sum_{i=1}^{n} y_i &= 0 \\
\hat{\lambda}^{-1}\sum_{i=1}^{n} y_i &= n \\
\sum_{i=1}^{n} y_i &= n\hat{\lambda} \\
\hat{\lambda} &= n^{-1}\sum_{i=1}^{n} y_i
\end{aligned}$$

Thus the maximum likelihood estimator in a Poisson is the sample mean.

2.1.2.2 Maximum Likelihood Estimation of a Normal (Gaussian) Model

Suppose y_i is assumed to be i.i.d. normally distributed with mean μ and variance σ^2. The pdf of a normal is

$$f(y_i;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right). \tag{2.5}$$

where $\boldsymbol{\theta} = \begin{bmatrix}\mu & \sigma^2\end{bmatrix}'$. The joint likelihood is the product of the n individual likelihoods,

$$f(\mathbf{y};\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right).$$

Taking logs,

$$l(\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(y_i-\mu)^2}{2\sigma^2} \tag{2.6}$$

which can be simplified to

$$l(\boldsymbol{\theta};\mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

Taking the derivative with respect to the parameters $\boldsymbol{\theta} = \begin{pmatrix}\mu, \sigma^2\end{pmatrix}'$,

$$\frac{\partial\, l(\boldsymbol{\theta};\mathbf{y})}{\partial\mu} = \sum_{i=1}^{n}\frac{(y_i-\mu)}{\sigma^2} \tag{2.7}$$

$$\frac{\partial\, l(\boldsymbol{\theta};\mathbf{y})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^4}. \tag{2.8}$$

Setting these equal to zero, the first condition can be directly solved by multiplying both sides by $\hat{\sigma}^2$, assumed positive, and the estimator for μ is the sample average.

$$\begin{aligned}
\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})}{\hat{\sigma}^2} &= 0 \\
\hat{\sigma}^2\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})}{\hat{\sigma}^2} &= \hat{\sigma}^2 0 \\
\sum_{i=1}^{n} y_i - n\hat{\mu} &= 0 \\
n\hat{\mu} &= \sum_{i=1}^{n} y_i \\
\hat{\mu} &= n^{-1}\sum_{i=1}^{n} y_i
\end{aligned}$$

Plugging this value into the second score and setting equal to 0, the ML estimator of σ^2 is

$$\begin{aligned}
-\frac{n}{2\hat{\sigma}^2}+\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})^2}{\hat{\sigma}^4} &= 0 \\
2\hat{\sigma}^4\left(-\frac{n}{2\hat{\sigma}^2}+\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})^2}{\hat{\sigma}^4}\right) &= 2\hat{\sigma}^4 0 \\
-n\hat{\sigma}^2+\sum_{i=1}^{n}(y_i-\hat{\mu})^2 &= 0 \\
\hat{\sigma}^2 &= n^{-1}\sum_{i=1}^{n}(y_i-\hat{\mu})^2
\end{aligned}$$

2.1.3 Conditional Maximum Likelihood

Interest often lies in the distribution of a random variable conditional on one or more observed values, where the distribution of the observed values is not of interest. When this occurs, it is natural to use conditional maximum likelihood. Suppose interest lies in modeling a random variable Y conditional on one or more variables$\mathbf{X}$. The likelihood for a single observation is $f_i\left(y_i|\mathbf{x}_i\right)$, and when Y_i are conditionally i.i.d. , then

$$L(\boldsymbol{\theta};\mathbf{y}|\mathbf{X})=\prod_{i=1}^{n}f\left(y_i|\mathbf{x}_i\right),$$

and the log-likelihood is

$$l\left(\boldsymbol{\theta};\mathbf{y}|\mathbf{X}\right)=\sum_{i=1}^{n}\ln f\left(y_i|\mathbf{x}_i\right).$$

The conditional likelihood is not usually sufficient to estimate parameters since the relationship between Y and $\mathbf{X}$ has not been specified. Conditional maximum likelihood specifies the model parameters conditionally on $\mathbf{x}_i$. For example, in an conditional normal, $y|\mathbf{x}_i\sim N\left(\mu_i,\sigma^2\right)$ where $\mu_i=g\left(\boldsymbol{\beta},\mathbf{x}_i\right)$ is some function which links parameters and conditioning variables. In many applications a linear relationship is assumed so that

$$\begin{aligned}
y_i &= \boldsymbol{\beta}'\mathbf{x}_i+\epsilon_i \\
&= \sum_{j=1}^{k}\beta_i x_{i,j}+\epsilon_i \\
&= \mu_i+\epsilon_i.
\end{aligned}$$

Other relationships are possible, including functions $g\left(\boldsymbol{\beta}'\mathbf{x}_i\right)$ which limits to range of $\boldsymbol{\beta}'\mathbf{x}_i$ such as $\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)$ (positive numbers), the normal cdf ($\Phi\left(\boldsymbol{\beta}'\mathbf{x}\right)$) or the logistic function,

$$\Lambda\left(\boldsymbol{\beta}'\mathbf{x}_i\right)=\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)/\left(1+\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)\right),$$

since both limit the range to (0, 1).

2.1.3.1 Example: Conditional Bernoulli

Suppose Y_i and X_i are Bernoulli random variables where the conditional distribution of Y_i given X_i is

$$y_i | x_i \sim \text{Bernoulli}(\theta_0 + \theta_1 x_i)$$

so that the conditional probability of observing a success ($y_i = 1$) is $p_i = \theta_0 + \theta_1 x_i$. The conditional likelihood is

$$L\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right) = \prod_{i=1}^{n} \left(\theta_0 + \theta_1 x_i\right)^{y_i} \left(1 - \left(\theta_0 + \theta_1 x_i\right)\right)^{1-y_i},$$

the conditional log-likelihood is

$$l\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right) = \sum_{i=1}^{n} y_i \ln\left(\theta_0 + \theta_1 x_i\right) + \left(1 - y_i\right) \ln\left(1 - \left(\theta_0 + \theta_1 x_i\right)\right),$$

and the maximum likelihood estimator can be found by differentiation.

$$\begin{aligned}
\frac{\partial l\left(\hat{\theta} B; \mathbf{y}|\mathbf{x}\right)}{\partial \hat{\theta}_0} &= \sum_{i=1}^{n} \frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{1 - y_i}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} = 0 \\
\frac{\partial l\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right)}{\partial \theta_1} &= \sum_{i=1}^{n} \frac{x_i y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{x_i \left(1 - y_i\right)}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} = 0.
\end{aligned}$$

Using the fact that x_i is also Bernoulli, the second score can be solved

$$\begin{aligned}
0 = \sum_{i=1}^{n} x_i \left(\frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1} + \frac{\left(1 - y_i\right)}{\left(1 - \hat{\theta}_0 - \hat{\theta}_1\right)} \right) &= \sum_{i=1}^{n} \frac{n_{xy}}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{n_x - n_{xy}}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} \\
&= n_{xy}\left(1 - \left(\hat{\theta}_0 + \hat{\theta}_1\right)\right) - \left(n_x - n_{xy}\right)\left(\hat{\theta}_0 + \hat{\theta}_1\right) \\
&= n_{xy} - n_{xy}\left(\hat{\theta}_0 + \hat{\theta}_1\right) - n_x\left(\hat{\theta}_0 + \hat{\theta}_1\right) + n_{xy}\left(\hat{\theta}_0 + \hat{\theta}_1\right) \\
\hat{\theta}_0 + \hat{\theta}_1 &= \frac{n_{xy}}{n_x},
\end{aligned}$$

Define $n_x = \sum_{i=1}^{n} x_i$, $n_y = \sum_{i=1}^{n} y_i$ and $n_{xy} = \sum_{i=1}^{n} x_i y_i$. The first score than also be rewritten as

$$\begin{aligned}
0 = \sum_{i=1}^{n} \frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{1 - y_i}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} &= \sum_{i=1}^{n} \frac{y_i \left(1 - x_I\right)}{\hat{\theta}_0} + \frac{y_i x_i}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{1 - y_i \left(1 - x_I\right)}{1 - \hat{\theta}_0} - \frac{\left(1 - y_i\right) x_i}{1 - \hat{\theta}_0 - \hat{\theta}_1} \\
&= \sum_{i=1}^{n} \frac{y_i \left(1 - x_I\right)}{\hat{\theta}_0} - \frac{1 - y_i \left(1 - x_I\right)}{1 - \hat{\theta}_0} + \left\{ \frac{x_i y_i}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{x_i \left(1 - y_i\right)}{1 - \hat{\theta}_0 - \hat{\theta}_1} \right\} \\
&= \frac{n_y - n_{xy}}{\hat{\theta}_0} - \frac{n - n_y - n_x + n_{xy}}{1 - \hat{\theta}_0} + \{0\} \\
&= n_y - n_{xy} - \hat{\theta}_0 n_y + \hat{\theta}_0 n - \hat{\theta}_0 n + \hat{\theta}_0 n_y + \hat{\theta}_0 n_x - \hat{\theta}_0 n_{xy} \\
\hat{\theta}_0 &= \frac{n_y - n_{xy}}{n - n_x}
\end{aligned}$$

so that $\hat{\theta}_1 = \frac{n_{xy}}{n_x} - \frac{n_y - n_{xy}}{n - n_x}$. The "0" in the previous derivation follows from noting that the quantity in $\{\}$ is equivalent to the first score and so is 0 at the MLE. If X_i was not a Bernoulli random variable, then it would not be possible to analytically solve this problem. In these cases, numerical methods are needed.[6]

2.1.3.2 Example: Conditional Normal

Suppose $\mu_i = \beta x_i$ where Y_i given X_i is conditionally normal. Assuming that Y_i are conditionally i.i.d. , the likelihood and log-likelihood are

$$\begin{aligned} L\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right) \\ l\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right) &= \sum_{i=1}^{n} -\frac{1}{2}\left(\ln(2\pi) + \ln\left(\sigma^2\right) + \frac{(y_i - \beta x_i)^2}{\sigma^2}\right). \end{aligned}$$

The scores of the likelihood are

$$\begin{aligned} \frac{\partial l\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right)}{\partial \beta} &= \sum_{i=1}^{n} \frac{x_i\left(y_i - \hat{\beta} x_i\right)}{\hat{\sigma}^2} = 0 \\ \frac{\partial l\left(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}\right)}{\partial \sigma^2} &= -\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2} - \frac{\left(y_i - \hat{\beta} x_i\right)^2}{\left(\hat{\sigma}^2\right)^2} = 0 \end{aligned}$$

After multiplying both sides the first score by $\hat{\sigma}^2$, and both sides of the second score by $-2\hat{\sigma}^4$, solving the scores is straight forward, and so

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2} \\ \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^{n} (y_i - \beta x_i)^2. \end{aligned}$$

2.1.3.3 Example: Conditional Poisson

Suppose Y_i is conditional on X_1 i.i.d. distributed Poisson(λ_i) where $\lambda_i = \exp(\theta x_i)$. The likelihood and log-likelihood are

$$\begin{aligned} L\left(\theta; \mathbf{y}|\mathbf{x}\right) &= \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \\ l\left(\theta; \mathbf{y}|\mathbf{x}\right) &= \sum_{i=1}^{n} \exp(\theta x_i) + y_i(\theta x_i) - \ln(y_i!). \end{aligned}$$

[6] When X_i is not Bernoulli, it is also usually necessary to use a function to ensure p_i, the conditional probability, is in [0, 1]. Tow common choices are the normal cdf and the logistic function.

The score of the likelihood is

$$\frac{\partial l\left(\theta;\mathbf{y}|\mathbf{x}\right)}{\partial \theta} = \sum_{i=1}^{n} -x_i \exp\left(\hat{\theta} x_i\right) + x_i y_i = 0.$$

This score cannot be analytically solved and so a numerical optimizer must be used to dins the solution. It is possible, however, to show the score has conditional expectation 0 since $\mathrm{E}\left[Y_i|X_i\right] = \lambda_i$.

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial l\left(\theta;\mathbf{y}|\mathbf{x}\right)}{\partial \theta}|\mathbf{X}\right] &= \mathrm{E}\left[\sum_{i=1}^{n} -x_i \exp\left(\theta x_i\right) + x_i y_i|\mathbf{X}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[-x_i \exp\left(\theta x_i\right)|\mathbf{X}\right] + \mathrm{E}\left[x_i y_i|\mathbf{X}\right] \\
&= \sum_{i=1}^{n} -x_i\lambda_i + x_i \mathrm{E}\left[y_i|\mathbf{X}\right] \\
&= \sum_{i=1}^{n} -x_i\lambda_i + x_i\lambda_i = 0.
\end{aligned}$$

2.1.4 The Method of Moments

The Method of moments, often referred to as the classical method of moments to differentiate it from the *generalized* method of moments (GMM, chapter 6) uses the data to match *noncentral* moments.

Definition 2.1 (Noncentral Moment)**.** The r$^{\text{th}}$ noncentral moment is defined

$$\mu'_r \equiv \mathrm{E}\left[X^r\right] \tag{2.9}$$

for $r = 1, 2, \ldots$.

Central moments are similarly defined, only centered around the mean.

Definition 2.2 (Central Moment)**.** The r$^{\text{th}}$ central moment is defined

$$\mu_r \equiv \mathrm{E}\left[\left(X - \mu'_1\right)^r\right] \tag{2.10}$$

for $r = 2, 3, \ldots$ where the 1$^{\text{st}}$central moment is defined to be equal to the 1$^{\text{st}}$noncentral moment.

Since $\mathrm{E}\left[x_i^r\right]$ is not known any estimator based on it is *infeasible*. The obvious solution is to use the *sample analogue* to estimate its value, and the *feasible* method of moments estimator is

$$\hat{\mu}'_r = n^{-1}\sum_{i=1}^{n} x_i^r, \tag{2.11}$$

the sample average of the data raised to the r$^{\text{th}}$ power. While the classical method of moments was originally specified using noncentral moments, the central moments are usually the quantities of interest. The central moments can be directly estimated,

$$\hat{\mu}_r = n^{-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^r, \tag{2.12}$$

and so can be simply implemented by first estimating the mean ($\hat{\mu}_1$) and then estimating the remaining central moments. An alternative is to expand the noncentral moment in terms of central moments. For example, the second noncentral moment can be expanded in terms of the first two central moments,

$$\mu_2' = \mu_2 + \mu_1^2$$

which is the usual identity that states that expectation of a random variable squared, $\mathrm{E}[x_i^2]$, is equal to the variance, $\mu_2 = \sigma^2$, plus the mean squared, μ_1^2. Likewise, it is easy to show that

$$\mu_3' = \mu_3 + 3\mu_2\mu_1 + \mu_1^3$$

directly by expanding $\mathrm{E}\left[(X-\mu_1)^3\right]$ and solving for μ_3'. To understand that the method of moments is in the class of M-estimators, note that the expression in eq. (2.12) is the first order condition of a simple quadratic form,

$$\underset{\mu,\mu_2,\ldots,\mu_k}{\arg\min}\left(n^{-1}\sum_{i=1}^{n}x_i - \mu_1\right)^2 + \sum_{j=2}^{k}\left(n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j - \mu_j\right)^2, \tag{2.13}$$

and since the number of unknown parameters is identical to the number of equations, the solution is exact.[7]

2.1.4.1 Method of Moments Estimation of the Mean and Variance

The classical method of moments estimator for the mean and variance for a set of i.i.d. data $\{y_i\}_{i=1}^n$ where $\mathrm{E}[Y_i] = \mu$ and $\mathrm{E}\left[(Y_i-\mu)^2\right] = \sigma^2$ is given by estimating the first two noncentral moments and then solving for σ^2.

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} y_i$$
$$\hat{\sigma}^2 + \hat{\mu}^2 = n^{-1}\sum_{i=1}^{n} y_i^2$$

and thus the variance estimator is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n y_i^2 - \hat{\mu}^2$. Following some algebra, it is simple to show that the central moment estimator could be used equivalently, and so $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (y_i - \hat{\mu})^2$.

[7] Note that μ_1, the mean, is generally denoted with the subscript suppressed as μ.

2.1.4.2 Method of Moments Estimation of the Range of a Uniform

Consider a set of realization of a random variable with a uniform density over $[0, \theta]$, and so $y_i \overset{\text{i.i.d.}}{\sim} U(0, \theta)$. The expectation of y_i is $\mathrm{E}[Y_i] = \theta/2$, and so the method of moments estimator for the upper bound is

$$\hat{\theta} = 2n^{-1}\sum_{i=1}^{n} y_i.$$

2.1.5 Classical Minimum Distance

A third – and less frequently encountered – type of M-estimator is classical minimum distance (CMD) which is also known as minimum χ^2 in some circumstances. CMD differs from MLE and the method of moments in that it is an estimator that operates using initial parameter estimates produced by another estimator rather than on the data directly. CMD is most common when a simple MLE or moment-based estimator is available that can estimate a model without some economically motivated constraints on the parameters. This initial estimator, $\hat{\boldsymbol{\psi}}$ is then used to estimate the parameters of the model, $\boldsymbol{\theta}$, by minimizing a quadratic function of the form

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left(\hat{\boldsymbol{\psi}} - \mathbf{g}(\boldsymbol{\theta})\right)' \mathbf{W} \left(\hat{\boldsymbol{\psi}} - \mathbf{g}(\boldsymbol{\theta})\right) \tag{2.14}$$

where $\mathbf{W}$ is a positive definite weighting matrix. When $\mathbf{W}$ is chosen as the covariance of $\hat{\boldsymbol{\psi}}$, the CMD estimator becomes the minimum-χ^2 estimator since outer products of standardized normals are χ^2 random variables.

2.2 Convergence and Limits for Random Variables

Before turning to properties of estimators, it is useful to discuss some common measures of convergence for sequences. Before turning to the alternative definitions which are appropriate for random variables, recall the definition of a limit of a non-stochastic sequence.

Definition 2.3 (Limit)**.** Let $\{x_n\}$ be a non-stochastic sequence. If there exists N such that for ever $n > N$, $|x_n - x| < \epsilon \, \forall \epsilon > 0$, when x is called the limit of x_n. When this occurs, $x_n \to x$ or $\lim_{n\to\infty} x_n = x$.

A limit is a point where a sequence will approach, and eventually, always remain near. It isn't necessary that the limit is ever attained, only that for any choice of $\epsilon > 0$, x_n will eventually always be less than ϵ away from its limit.

Limits of random variables come in many forms. The first the type of convergence is both the weakest and most abstract.

Definition 2.4 (Convergence in Distribution)**.** Let $\{\mathbf{Y}_n\}$ be a sequence of random variables and let $\{F_n\}$ be the associated sequence of cdfs. If there exists a cdf F where $F_n(\mathbf{y}) \to F(\mathbf{y})$ for all $\mathbf{y}$ where F is continuous, then F is the limiting cdf of $\{\mathbf{Y}_n\}$. Let $\mathbf{Y}$ be a random variable with cdf F, then $\mathbf{Y}_n$ converges in distribution to $\mathbf{Y} \sim F$, $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y} \sim F$, or simply $\mathbf{Y}_n \xrightarrow{d} F$.

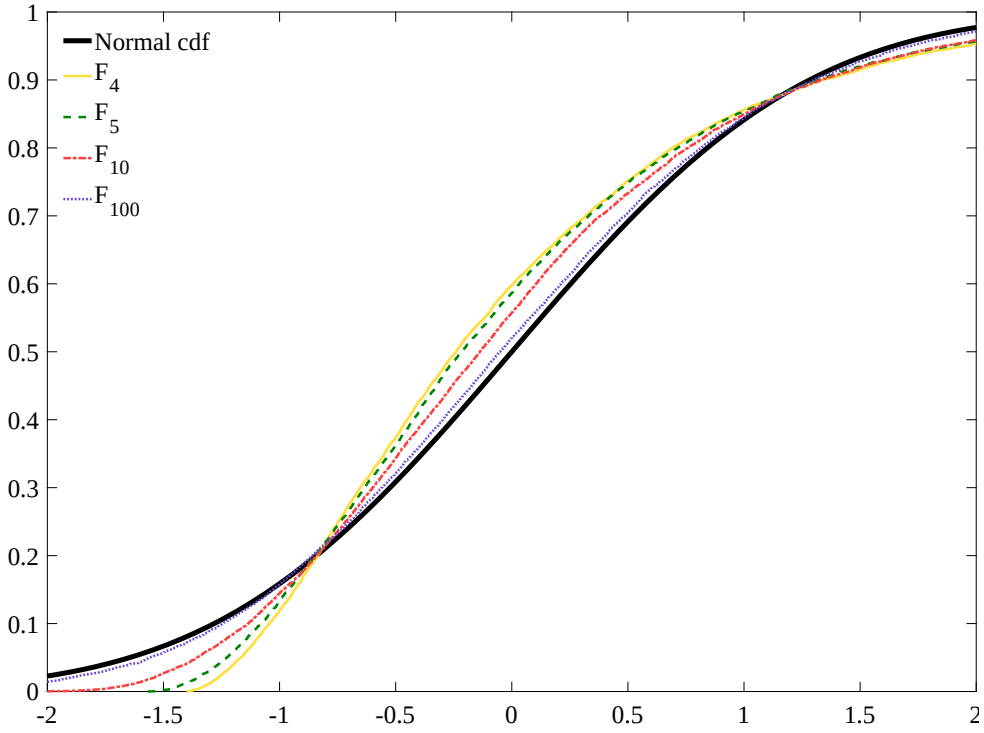

Figure 2.1: This figure shows a sequence of cdfs $\{F_i\}$ that converge to the cdf of a standard normal.

Convergence in distribution means that the limiting cdf of a sequence of random variables is the same as the convergent random variable. This is a very weak form of convergence since all it requires is that the distributions are the same. For example, suppose $\{X_n\}$ is an i.i.d. sequence of standard normal random variables, and Y is a standard normal random variable. X_n trivially converges to distribution to Y ($X_n \xrightarrow{d} Y$) even through Y is completely independent of $\{X_n\}$ – the limiting cdf of X_n is merely the same as the cdf of Y. Despite the weakness of convergence in distribution, it is an essential notion of convergence that is used to perform inference on estimated parameters.

Figure 2.1 shows an example of a sequence of random variables which converge in distribution. The sequence is

$$X_n = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i - 1}{\sqrt{2}}$$

where Y_i are i.i.d. χ^2_1 random variables. This is a studentized average since the variance of the average is $2/n$ and the mean is 1. By the time $n = 100$, F_{100} is nearly indistinguishable from the standard normal cdf.

Convergence in distribution is preserved through functions.

Theorem 2.1 (Continuous Mapping Theorem)**.** *Let* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and let the random variable* $g(\mathbf{X})$ *be defined by a function* $g(\mathbf{x})$ *that is continuous everywhere except possibly on a set with zero probability. Then* $g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X})$.

The continuous mapping theorem is useful since it facilitates the study of functions of sequences of random variables. For example, in hypothesis testing, it is common to use quadratic forms of normals, and when appropriately standardized, quadratic forms of normally distributed random variables follow a χ^2 distribution.

The next form of convergence is stronger than convergence in distribution since the limit is to a specific target, not just a cdf.

Definition 2.5 (Convergence in Probability)**.** The sequence of random variables $\{\mathbf{X}_n\}$ converges in probability to $\mathbf{X}$ if and only if

$$\lim_{n\to\infty} \Pr\left(|X_{i,n} - X_i| < \epsilon\right) = 1 \; \forall \epsilon > 0, \; \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ or equivalently $\operatorname{plim} \mathbf{X}_n = \mathbf{X}$ (or $\operatorname{plim} \mathbf{X}_n - \mathbf{X} = 0$) where plim is probability limit.

Note that $\mathbf{X}$ can be either a random variable or a constant (degenerate random variable). For example, if $X_n = n^{-1} + Z$ where Z is a normally distributed random variable, then $X_n \xrightarrow{p} Z$. Convergence in probability requires virtually all of the probability mass of $\mathbf{X}_n$ to lie near $\mathbf{X}$. This is a very weak form of convergence since it is possible that a small amount of probability can be arbitrarily far away from $\mathbf{X}$. Suppose a scalar random sequence $\{X_n\}$ takes the value 0 with probability $1 - 1/n$ and n with probability $1/n$. Then $\{X_n\} \xrightarrow{p} 0$ although $\mathrm{E}[X_n] = 1$ for all n.

Convergence in probability, however, is strong enough that it is useful work studying random variables and functions of random variables.

Theorem 2.2. *Let* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ *and let the random variable* $g(\mathbf{X})$ *be defined by a function* $g(x)$ *that is continuous everywhere except possibly on a set with zero probability. Then* $g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X})$ *(or equivalently* $\operatorname{plim} g(\mathbf{X}_n) = g(\mathbf{X})$*).*

This theorem has some, simple useful forms. Suppose the k-dimensional vector $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, the conformable vector $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y}$ and $\mathbf{C}$ is a conformable constant matrix, then

- $\operatorname{plim} \mathbf{C}\mathbf{X}_n = \mathbf{C}\mathbf{X}$
- $\operatorname{plim} \sum_{i=1}^{k} X_{i,n} = \sum_{i=1}^{k} \operatorname{plim} X_{i,n}$ – the plim of the sum is the sum of the plims
- $\operatorname{plim} \prod_{i=1}^{k} X_{i,n} = \prod_{i=1}^{k} \operatorname{plim} X_{i,n}$– the plim of the product is the product of the plims
- $\operatorname{plim} \mathbf{Y}_n\mathbf{X}_n = \mathbf{Y}\mathbf{X}$
- When $\mathbf{Y}_n$ is a square matrix and $\mathbf{Y}$ is nonsingular, then $\mathbf{Y}_n^{-1} \xrightarrow{p} \mathbf{Y}^{-1}$ – the inverse function is continuous and so plim of the inverse is the inverse of the plim
- When $\mathbf{Y}_n$ is a square matrix and $\mathbf{Y}$ is nonsingular, then $\mathbf{Y}_n^{-1}\mathbf{X}_n \xrightarrow{p} \mathbf{Y}^{-1}\mathbf{X}$.

These properties are very difference from the expectations operator. In particular, the plim operator passes through functions which allows for broad application. For example,

$$\mathrm{E}\left[\frac{1}{X}\right] \neq \frac{1}{\mathrm{E}[X]}$$

whenever X is a non-degenerate random variable. However, if $X_n \xrightarrow{p} X$, then

$$\begin{aligned}\operatorname{plim}\frac{1}{X_n} &= \frac{1}{\operatorname{plim} X_n} \\ &= \frac{1}{X}.\end{aligned}$$

Alternative definitions of convergence strengthen convergence in probability. In particular, convergence in mean square requires that the expected squared deviation must be zero. This requires that $\mathrm{E}[X_n] = X$ and $\mathrm{V}[X_n] = 0$.

Definition 2.6 (Convergence in Mean Square)**.** The sequence of random variables $\{\mathbf{X}_n\}$ converges in mean square to $\mathbf{X}$ if and only if

$$\lim_{n\to\infty} \mathrm{E}\left[(X_{i,n} - X_i)^2\right] = 0,\ \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$.

Mean square convergence is strong enough to ensure that, when the limit is random $\mathbf{X}$ than $\mathrm{E}[\mathbf{X}_n] = \mathrm{E}[\mathbf{X}]$ and $\mathrm{V}[\mathbf{X}_n] = \mathrm{V}[\mathbf{X}]$ – these relationships do not necessarily hold when only $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

Theorem 2.3 (Convergence in mean square implies consistency)**.** *If* $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$ *then* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

This result follows directly from Chebyshev's inequality. A final, and very strong, measure of convergence for random variables is known as almost sure convergence.

Definition 2.7 (Almost sure convergence)**.** The sequence of random variables $\{\mathbf{X}_n\}$ converges almost surely to $\mathbf{X}$ if and only if

$$\lim_{n\to\infty} \Pr(X_{i,n} - X_i = 0) = 1,\ \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$.

Almost sure convergence requires all probability to be on the limit point. This is a stronger condition than either convergence in probability or convergence in mean square, both of which allow for some probability to be (relatively) far from the limit point.

Theorem 2.4 (Almost sure convergence implications)**.** *If* $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ *then* $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$ *and* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

Random variables which converge almost surely to a limit are asymptotically degenerate on that limit.

The Slutsky theorem combines variables which converge in distribution with variables which converge in probability to show that the joint limit of functions behaves as expected.

Theorem 2.5 (Slutsky Theorem)**.** *Let* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and let* $\mathbf{Y} \xrightarrow{p} \mathbf{C}$*, a constant, then for conformable* $\mathbf{X}$ *and* $\mathbf{C}$*,*

1. $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + c$
2. $\mathbf{Y}_n\mathbf{X}_n \xrightarrow{d} \mathbf{C}\mathbf{X}$
3. $\mathbf{Y}_n^{-1}\mathbf{X}_n \xrightarrow{d} \mathbf{C}^{-1}\mathbf{X}$ *as long as* $\mathbf{C}$ *is non-singular.*

This theorem is at the core of hypothesis testing where estimated parameters are often asymptotically normal and an estimated parameter covariance, which converges in probability to the true covariance, is used to studentize the parameters.

2.3 Properties of Estimators

The first step in assessing the performance of an economic model is the estimation of the parameters. There are a number of desirable properties estimators may possess.

2.3.1 Bias and Consistency

A natural question to ask about an estimator is whether, on average, it will be equal to the population value of the parameter estimated. Any discrepancy between the expected value of an estimator and the population parameter is known as bias.

Definition 2.8 (Bias)**.** The bias of an estimator, $\hat{\boldsymbol{\theta}}$, is defined

$$\mathrm{B}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{E}\left[\hat{\boldsymbol{\theta}}\right] - \boldsymbol{\theta}_0 \tag{2.15}$$

where $\boldsymbol{\theta}_0$ is used to denote the population (or "true") value of the parameter.

When an estimator has a bias of $\mathbf{0}$ it is said to be unbiased. Unfortunately, many estimators are not unbiased. Consistency is a closely related concept that measures whether a parameter will be far from the population value in *large samples*.

Definition 2.9 (Consistency)**.** An estimator $\hat{\boldsymbol{\theta}}_n$ is said to be consistent if $\operatorname{plim}\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0$. The explicit dependence of the estimator on the sample size is used to clarify that these form a sequence, $\left\{\hat{\boldsymbol{\theta}}_n\right\}_{n=1}^{\infty}$.

Consistency requires an estimator to exhibit two features as the sample size becomes large. First, any bias must be shrinking. Second, the distribution of $\hat{\boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ must be shrinking in such a way that virtually all of the probability mass is arbitrarily close to $\boldsymbol{\theta}_0$. Behind consistency is a set of theorems known as *laws of large numbers*. Laws of large numbers provide conditions where an average will converge to its expectation. The simplest is the Kolmogorov Strong Law of Large numbers and is applicable to i.i.d. data.[8]

Theorem 2.6 (Kolmogorov Strong Law of Large Numbers)**.** *Let $\{y_i\}$ by a sequence of i.i.d. random variables with $\mu \equiv \mathrm{E}[y_i]$ and define $\bar{y}_n = n^{-1}\sum_{i=1}^{n} y_i$. Then*

$$\bar{y}_n \xrightarrow{a.s.} \mu \tag{2.16}$$

if and only if $\mathrm{E}\left[|y_i|\right] < \infty$.

In the case of i.i.d. data the only requirement for consistency is that the expectation exists, and so a law of large numbers will apply to an average of i.i.d. data whenever its expectation exists. For example, Monte Carlo integration uses i.i.d. draws and so the Kolmogorov LLN is sufficient to ensure that Monte Carlo integrals converge to their expected values.

The variance of an estimator is the same as any other variance, $\mathrm{V}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{E}\left[\left(\hat{\boldsymbol{\theta}} - \mathrm{E}[\hat{\boldsymbol{\theta}}]\right)^2\right]$ although it is worth noting that the variance is defined as the variation around its expectation, $\mathrm{E}[\hat{\boldsymbol{\theta}}]$, not the population value of the parameters, $\boldsymbol{\theta}_0$. Mean square error measures this alternative form of variation around the population value of the parameter.

[8] A law of large numbers is strong if the convergence is almost sure. It is weak if convergence is in probability.

Definition 2.10 (Mean Square Error)**.** The mean square error of an estimator $\hat{\boldsymbol{\theta}}$, denoted $\text{MSE}\left(\hat{\boldsymbol{\theta}}\right)$, is defined

$$\text{MSE}\left(\hat{\boldsymbol{\theta}}\right) = \text{E}\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^2\right]. \tag{2.17}$$

It can be equivalently expressed as the bias squared plus the variance, $\text{MSE}\left(\hat{\boldsymbol{\theta}}\right) = \text{B}\left[\hat{\boldsymbol{\theta}}\right]^2 + \text{V}\left[\hat{\boldsymbol{\theta}}\right]$.

When the bias and variance of an estimator both converge to zero, then $\hat{\boldsymbol{\theta}}_n \xrightarrow{m.s.} \boldsymbol{\theta}_0$.

2.3.1.1 Bias and Consistency of the Method of Moment Estimators

The method of moments estimators of the mean and variance are defined as

$$\begin{aligned}\hat{\mu} &= n^{-1}\sum_{i=1}^{n} y_i \\ \hat{\sigma}^2 &= n^{-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 .\end{aligned}$$

When the data are i.i.d. with finite mean μ and variance σ^2, the mean estimator is unbiased while the variance is biased by an amount that becomes small as the sample size increases. The mean is unbiased since

$$\begin{aligned}\text{E}[\hat{\mu}] &= \text{E}\left[n^{-1}\sum_{i=1}^{n} y_i\right] \\ &= n^{-1}\sum_{i=1}^{n}\text{E}[y_i] \\ &= n^{-1}\sum_{i=1}^{n}\mu \\ &= n^{-1}n\mu \\ &= \mu\end{aligned}$$

The variance estimator is biased since

$$\begin{aligned}\text{E}\left[\hat{\sigma}^2\right] &= \text{E}\left[n^{-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2\right] \\ &= \text{E}\left[n^{-1}\left(\sum_{i=1}^{n} y_i^2 - n\hat{\mu}^2\right)\right] \\ &= n^{-1}\left(\sum_{i=1}^{n}\text{E}\left[y_i^2\right] - n\text{E}\left[\hat{\mu}^2\right]\right)\end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\left(\sum_{i=1}^{n}\mu^2+\sigma^2-n\left(\mu^2+\frac{\sigma^2}{n}\right)\right)\\
&= n^{-1}\left(n\mu^2+n\sigma^2-n\mu^2-\sigma^2\right)\\
&= n^{-1}\left(n\sigma^2-n\frac{\sigma^2}{n}\right)\\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

where the sample mean is equal to the population mean plus an error that is decreasing in n,

$$
\begin{aligned}
\hat{\mu}^2 &= \left(\mu+n^{-1}\sum_{i=1}^{n}\epsilon_i\right)^2\\
&= \mu^2+2\mu n^{-1}\sum_{i=1}^{n}\epsilon_i+\left(n^{-1}\sum_{i=1}^{n}\epsilon_i\right)^2
\end{aligned}
$$

and so its square has the expected value

$$
\begin{aligned}
\mathrm{E}\left[\hat{\mu}^2\right] &= \mathrm{E}\left[\mu^2+2\mu n^{-1}\sum_{i=1}^{n}\epsilon_i+\left(n^{-1}\sum_{i=1}^{n}\epsilon_i\right)^2\right]\\
&= \mu^2+2\mu n^{-1}\mathrm{E}\left[\sum_{i=1}^{n}\epsilon_i\right]+n^{-2}\mathrm{E}\left[\left(\sum_{i=1}^{n}\epsilon_i\right)^2\right]\\
&= \mu^2+\frac{\sigma^2}{n}.
\end{aligned}
$$

2.3.2 Asymptotic Normality

While unbiasedness and consistency are highly desirable properties of any estimator, alone these do not provide a method to perform inference. The primary tool in econometrics for inference is the central limit theorem (CLT). CLTs exist for a wide range of possible data characteristics that include i.i.d., heterogeneous and dependent cases. The Lindberg-Lévy CLT, which is applicable to i.i.d. data, is the simplest.

Theorem 2.7 (Lindberg-Lévy)**.** *Let $\{y_i\}$ be a sequence of i.i.d. random scalars with $\mu \equiv \mathrm{E}[Y_i]$ and $\sigma^2 \equiv \mathrm{V}[Y_i] < \infty$. If $\sigma^2 > 0$, then*

$$
\frac{\bar{y}_n-\mu}{\bar{\sigma}_n} = \sqrt{n}\frac{\bar{y}_n-\mu}{\sigma} \xrightarrow{d} N(0,1) \tag{2.18}
$$

where $\bar{y}_n = n^{-1}\sum_{i=1}^{n} y_i$ and $\bar{\sigma}_n = \sqrt{\frac{\sigma^2}{n}}$.

Lindberg-Lévy states that as long as i.i.d. data have 2 moments – a mean and variance – the sample mean will be asymptotically normal. It can further be seen to show that other moments of i.i.d. random variables, such as the variance, will be asymptotically normal as long as two times the power of the moment exists. In other words, an estimator of the r$^{\text{th}}$ moment will be asymptotically normal as long as the 2r$^{\text{th}}$ moment exists – at least in i.i.d. data. Figure 2.2 contains density plots of the sample average of n independent χ^2_1 random variables for $n = 5$, 10, 50 and 100.[9] The top panel contains the density of the unscaled estimates. The bottom panel contains the density plot of the correctly scaled terms, $\sqrt{n}(\hat{\mu} - 1)/\sqrt{2}$ where $\hat{\mu}$ is the sample average. The densities are collapsing in the top panel. This is evidence of consistency since the asymptotic distribution of $\hat{\mu}$ is collapsing on 1. The bottom panel demonstrates the operation of a CLT since the appropriately standardized means all have similar dispersion and are increasingly normal.

Central limit theorems exist for a wide variety of other data generating process including processes which are independent but not identically distributed (i.n.i.d) or processes which are dependent, such as time-series data. As the data become more heterogeneous, whether through dependence or by having different variance or distributions, more restrictions are needed on certain characteristics of the data to ensure that averages will be asymptotically normal. The Lindberg-Feller CLT allows for heteroskedasticity (different variances) and/or different marginal distributions.

Theorem 2.8 (Lindberg-Feller)**.** *Let $\{y_i\}$ be a sequence of independent random scalars with $\mu_i \equiv \mathrm{E}[y_i]$ and $0 < \sigma_i^2 \equiv \mathrm{V}[y_i] < \infty$ where $y_i \sim F_i$, $i = 1, 2, \ldots$. Then*

$$\sqrt{n}\frac{\bar{y}_n - \bar{\mu}_n}{\bar{\sigma}_n} \xrightarrow{d} N(0,1) \tag{2.19}$$

and

$$\lim_{n\to\infty} \max_{1\leq i\leq n} n^{-1}\frac{\sigma_i^2}{\bar{\sigma}_n^2} = 0 \tag{2.20}$$

if and only if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \bar{\sigma}_n^2 n^{-1} \sum_{i=1}^{n} \int_{(z-\mu_n)^2 > \epsilon N \sigma_n^2} (z-\mu_n)^2 \, dF_i(z) = 0 \tag{2.21}$$

where $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$ and $\bar{\sigma}^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$.

The Lindberg-Feller CLT relaxes the requirement that the marginal distributions are identical in the Lindberg-Lévy CLT at the cost of a technical condition. The final condition, known as a Lindberg condition, essentially that no random variable is so heavy-tailed that it dominates the others when averaged. In practice, this can be a concern when the variables have a wide range of variances (σ_i^2). Many macroeconomic data series exhibit a large *decrease* in the variance of their shocks after 1984, a phenomenon is referred to as the *great moderation*. The statistical consequence of this decrease is that averages that use data both before and after 1984 not be well approximated by a CLT and caution is warranted when using asymptotic approximations. This phenomena is also present in equity returns where some periods – for example the technology "bubble" from 1997-2002 – have substantially higher volatility than periods before or after. These large persistent changes in the characteristics of the data have negative consequences on the quality of CLT approximations and large data samples are often needed.

[9] The mean and variance of a χ^2_ν are ν and 2ν, respectively.

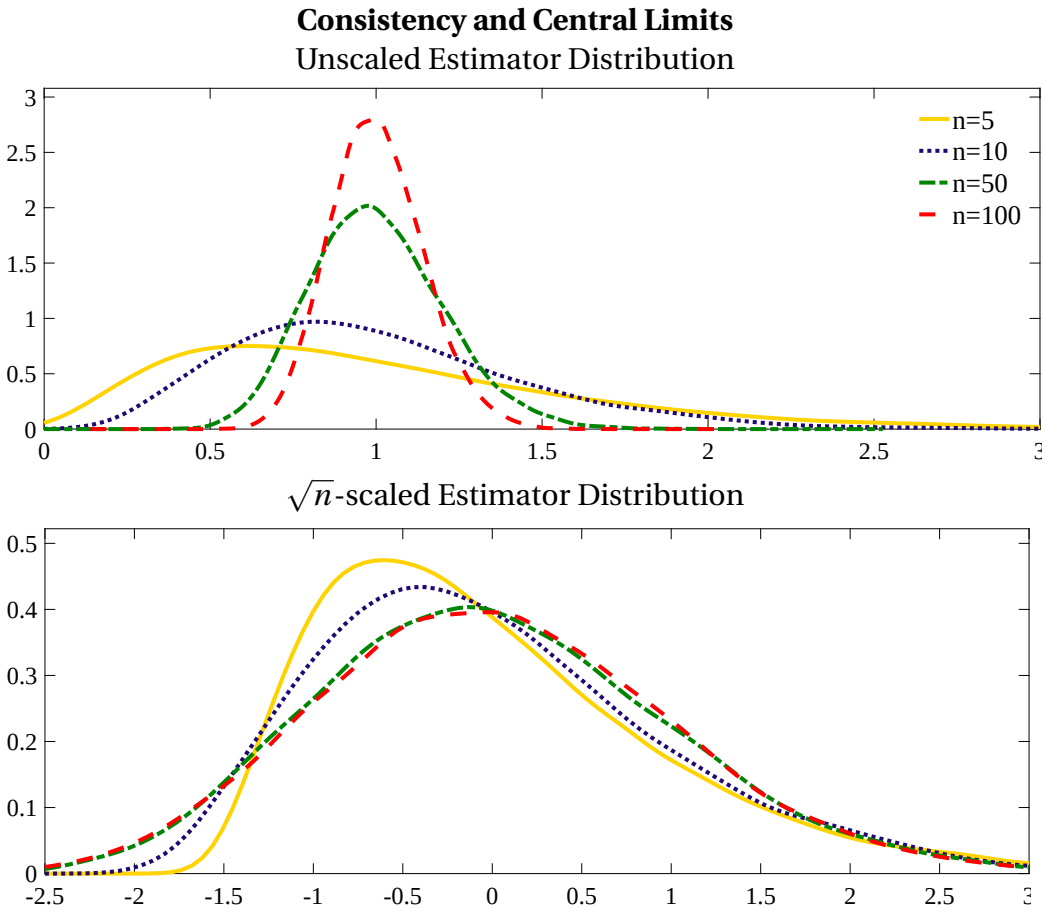

Figure 2.2: These two panels illustrate the difference between consistency and the correctly scaled estimators. The sample mean was computed 1,000 times using 5, 10, 50 and 100 i.i.d. χ^2 data points. The top panel contains a kernel density plot of the estimates of the mean. The density when $n = 100$ is much tighter than when $n = 5$ or $n = 10$ since the estimates are not scaled. The bottom panel plots $\sqrt{n}(\hat{\mu} - 1)/\sqrt{2}$, the standardized version for which a CLT applies. All scaled densities have similar dispersion although it is clear that the asymptotic approximation of the CLT is not particularly accurate when $n = 5$ or $n = 10$ due to the right skew in the χ^2_1 data.

2.3.2.1 What good is a CLT?

Central limit theorems are the basis of most inference in econometrics, although their formal justification is only asymptotic and hence only guaranteed to be valid for an arbitrarily large data set. Reconciling these two statements is an important step in the evolution of an econometrician.

Central limit theorems should be seen as approximations, and as an approximation, they can be accurate or arbitrarily poor. For example, when a series of random variables are i.i.d., thin-tailed and not skewed, the distribution of the sample mean computed using as few as 10 observations may be very well approximated using a central limit theorem. On the other hand, the approximation of a central limit theorem for the estimate of the autoregressive parameter, ρ, in

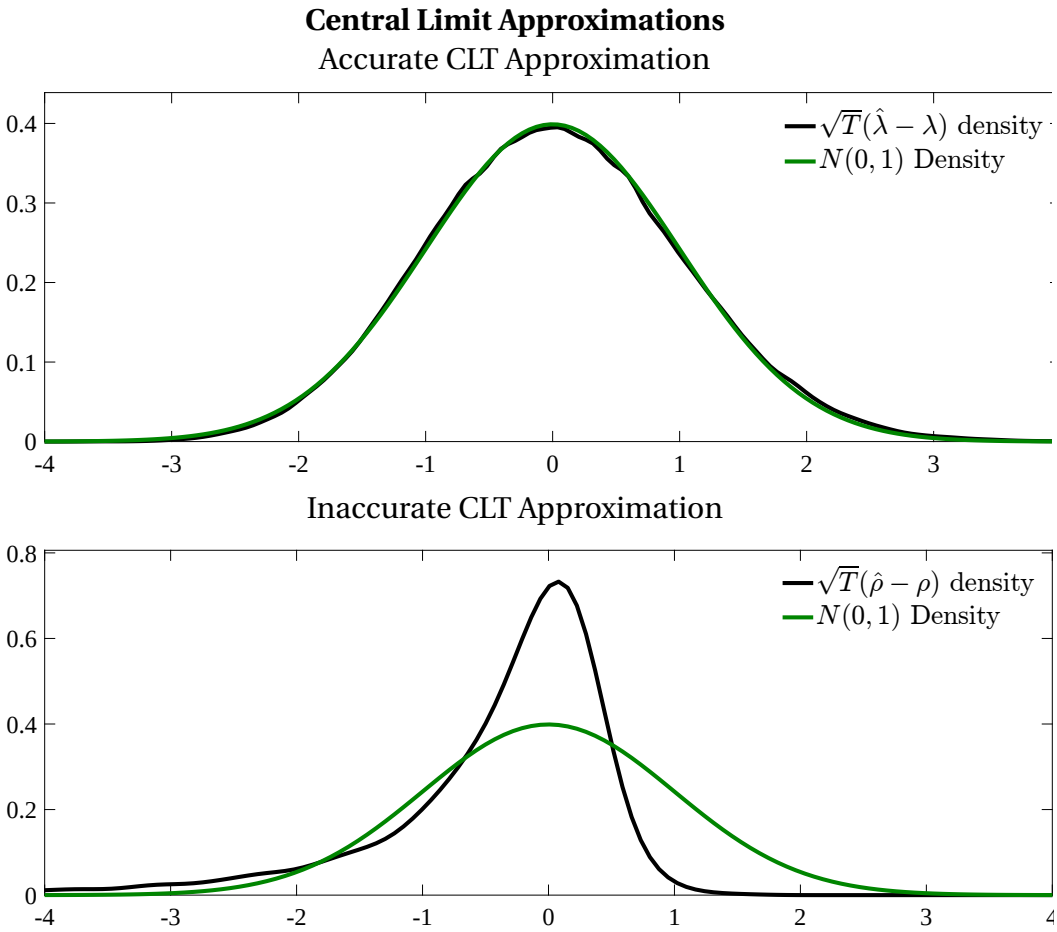

Figure 2.3: These two plots illustrate how a CLT can provide a good approximation, even in small samples (top panel), or a bad approximation even for moderately large samples (bottom panel). The top panel contains a kernel density plot of the standardized sample mean of $n = 10$ Poisson random variables ($\lambda = 5$) over 10,000 Monte Carlo simulations. Here the finite sample distribution and the asymptotic distribution overlay one another. The bottom panel contains the conditional ML estimates of ρ from the AR(1) $y_i = \rho y_{i-1} + \epsilon_i$ where ϵ_i is i.i.d. standard normal using 100 data points and 10,000 replications. While $\hat{\rho}$ is asymptotically normal, the quality of the approximation when $n = 100$ is poor.

$$y_i = \rho y_{i-1} + \epsilon_i \tag{2.22}$$

may be poor even for hundreds of data points when ρ is close to one (but smaller). Figure 2.3 contains kernel density plots of the sample means computed from a set of 10 i.i.d. draws from a Poisson distribution with $\lambda = 5$ in the top panel and the estimated autoregressive parameter from the autoregression in eq. (2.22) with $\rho = .995$ in the bottom. Each figure also contains the pdf of an appropriately scaled normal. The CLT for the sample means of the Poisson random variables is virtually indistinguishable from the actual distribution. On the other hand, the CLT approximation for $\hat{\rho}$ is very poor being based on 100 data points – $10\times$ more than in the i.i.d. uniform example. The difference arises because the data in the AR(1) example are not in-

dependent. With $\rho = 0.995$ data are highly dependent and more data is required for averages to be well behaved so that the CLT approximation is accurate.

There are no hard and fast rules as to when a CLT will be a good approximation. In general, the more dependent and the more heterogeneous a series, the worse the approximation for a fixed number of observations. Simulations (Monte Carlo) are a useful tool to investigate the validity of a CLT since they allow the finite sample distribution to be tabulated and compared to the asymptotic distribution.

2.3.3 Efficiency

A final concept, efficiency, is useful for ranking consistent asymptotically normal (CAN) estimators that have the same rate of convergence.[10]

Definition 2.11 (Relative Efficiency)**.** Let $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ be two $\sqrt{n}$-consistent asymptotically normal estimators for $\boldsymbol{\theta}_0$. If the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$, written avar $\left(\hat{\boldsymbol{\theta}}_n\right)$ is less than the asymptotic variance of $\tilde{\boldsymbol{\theta}}_n$, and so

$$\text{avar}\left(\hat{\boldsymbol{\theta}}_n\right) < \text{avar}\left(\tilde{\boldsymbol{\theta}}_n\right) \tag{2.23}$$

then $\hat{\boldsymbol{\theta}}_n$ is said to be relatively efficient to $\tilde{\boldsymbol{\theta}}_n$.[11]

Note that when $\boldsymbol{\theta}$ is a vector, avar $\left(\hat{\boldsymbol{\theta}}_n\right)$ will be a covariance matrix. Inequality for matrices $\mathbf{A}$ and $\mathbf{B}$ is interpreted to mean that if $\mathbf{A} < \mathbf{B}$ then $\mathbf{B} - \mathbf{A}$ is positive semi-definite, and so *all* of the variances of the inefficient estimator must be (weakly) larger than those of the efficient estimator.

Definition 2.12 (Asymptotically Efficient Estimator)**.** Let $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ be two $\sqrt{n}$-consistent asymptotically normal estimators for $\boldsymbol{\theta}_0$. If

$$\text{avar}\left(\hat{\boldsymbol{\theta}}_n\right) < \text{avar}\left(\tilde{\boldsymbol{\theta}}_n\right) \tag{2.24}$$

for any choice of $\tilde{\boldsymbol{\theta}}_n$ then $\hat{\boldsymbol{\theta}}_n$ is said to be the efficient estimator of $\boldsymbol{\theta}$.

One of the important features of efficiency comparisons is that they are only meaningful if both estimators are asymptotically normal, and hence consistent, at the same rate – $\sqrt{n}$ in the usual case. It is trivial to produce an estimator that has a smaller variance but is inconsistent. For example, if an estimator for a scalar unknown is $\hat{\theta} = 7$ then it has no variance: it will always be 7. However, unless $\theta_0 = 7$ it will also be biased. Mean square error is a more appropriate method to compare estimators where one or more may be biased since it accounts for the total variation, not just the variance.[12]

[10] In any consistent estimator the asymptotic distribution of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is degenerate. In order to perform inference on an unknown quantity, the difference between the estimate and the population parameters must be scaled by a function of the number of data points. For most estimators this rate is $\sqrt{n}$, and so $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ will have an asymptotically normal distribution. In the general case, the scaled difference can be written as $n^{\delta}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ where n^{δ} is known as the rate.

[11] The asymptotic variance of a $\sqrt{n}$-consistent estimator, written avar $\left(\hat{\boldsymbol{\theta}}_n\right)$ is defined as $\lim_{n\to\infty} \text{V}\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right]$.

[12] Some consistent asymptotically normal estimators have an asymptotic bias and so $\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(\mathbf{B}, \boldsymbol{\Sigma}\right)$. Asymptotic MSE defined as $\text{E}\left[n\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)'\right] = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}$ provides a method to compare estimators using their asymptotic properties.

2.4 Distribution Theory

Most distributional theory follows from a central limit theorem applied to the moment conditions or to the score of the log-likelihood. While the moment conditions or scores are not usually the objects of interest – $\boldsymbol{\theta}$ is – a simple expansion can be used to establish the asymptotic distribution of the estimated parameters.

2.4.1 Method of Moments

Distribution theory for the classical method of moments estimators is the most straightforward. Further, Maximum Likelihood can be considered a special case and so the method of moments is a natural starting point.[13] The method of moments estimator is defined as

$$\begin{aligned}
\hat{\mu} &= n^{-1}\sum_{i=1}^{n} x_i \\
\hat{\mu}_2 &= n^{-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\
&\vdots \\
\hat{\mu}_k &= n^{-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^k
\end{aligned}$$

To understand the distribution theory for the method of moments estimator, begin by reformulating the estimator as the solution of a set of k equations evaluated using the population values of $\mu, \mu_2, \ldots$

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n} x_i - \mu &= 0 \\
n^{-1}\sum_{i=1}^{n} (x_i - \mu)^2 - \mu_2 &= 0 \\
&\vdots \\
n^{-1}\sum_{i=1}^{n} (x_i - \mu)^k - \mu_k &= 0
\end{aligned}$$

Define $g_{1i} = x_i - \mu$ and $g_{ji} = (x_i - \mu)^j - \mu_j$, $j = 2, \ldots, k$, and the vector $\mathbf{g}_i$ as

[13] While the class of method of moments estimators and maximum likelihood estimators contains a substantial overlap, method of moments estimators exist that cannot be replicated as a score condition of any likelihood since the likelihood is required to integrate to 1.

$$\mathbf{g}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ki} \end{bmatrix}. \tag{2.25}$$

Using this definition, the method of moments estimator can be seen as the solution to

$$\mathbf{g}_n\left(\hat{\boldsymbol{\theta}}\right) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right) = 0.$$

Consistency of the method of moments estimator relies on a *law of large numbers* holding for $n^{-1}\sum_{i=1}^{n} x_i$ and $n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j$ for $j = 2,\ldots,k$. If x_i is an i.i.d. sequence and as long as $\mathrm{E}\left[\left|x_n-\mu\right|^j\right]$ exists, then $n^{-1}\sum_{i=1}^{n}(x_n-\mu)^j \xrightarrow{p} \mu_j$.[14] An alternative, and more restrictive approach is to assume that $\mathrm{E}\left[(x_n-\mu)^{2j}\right] = \mu_{2j}$ exists, and so

$$\mathrm{E}\left[n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j\right] = \mu_j \tag{2.26}$$

$$\begin{aligned}\mathrm{V}\left[n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j\right] &= n^{-1}\left(\mathrm{E}\left[(x_i-\mu)^{2j}\right] - \mathrm{E}\left[(x_i-\mu)^j\right]^2\right) \\ &= n^{-1}\left(\mu_{2j}-\mu_j^2\right),\end{aligned} \tag{2.27}$$

and so $n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j \xrightarrow{m.s.} \mu_j$ which implies consistency.

The asymptotic normality of parameters estimated using the method of moments follows from the asymptotic normality of

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\right) = n^{-1/2}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}_0\right), \tag{2.28}$$

an assumption. This requires the elements of $\mathbf{g}_n$ to be sufficiently well behaved so that averages are asymptotically normally distributed. For example, when x_i is i.i.d., the Lindberg-Lévy CLT would require x_i to have $2k$ moments when estimating k parameters. When estimating the mean, 2 moments are required (i.e. the variance is finite). To estimate the mean and the variance using i.i.d. data, 4 moments are required for the estimators to follow a CLT. As long as the moment conditions are differentiable in the actual parameters of interest $\boldsymbol{\theta}$ – for example, the mean and the variance – a *mean value expansion* can be used to establish the asymptotic normality of these parameters.[15]

[14] Technically, $n^{-1}\sum_{i=1}^{n}(x_i-\mu)^j \xrightarrow{a.s.} \mu_j$ by the Kolmogorov law of large numbers, but since a.s. convergence implies convergence in probability, the original statement is also true.

[15] The mean value expansion is defined in the following theorem.

Theorem 2.9 (Mean Value Theorem)**.** *Let $s:\mathbb{R}^k\to\mathbb{R}$ be defined on a convex set $\Theta\subset\mathbb{R}^k$. Further, let s be continuously differentiable on Θ with k by 1 gradient*

$$\nabla s\left(\hat{\boldsymbol{\theta}}\right) \equiv \left.\frac{\partial s\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.29}$$

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\hat{\boldsymbol{\theta}}) &= n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) + n^{-1}\sum_{i=1}^{n}\left.\frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \\ &= n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) + \mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \end{aligned} \tag{2.30}$$

where $\bar{\boldsymbol{\theta}}$ is a vector that lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, element-by-element. Note that $n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ by construction and so

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) + \mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= 0 \\ \mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= -n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) \\ \left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= -\mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)^{-1} n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) \\ \sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= -\mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)^{-1}\sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0) \\ \sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= -\mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)^{-1}\sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0) \end{aligned}$$

where $\mathbf{g}_n(\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta}_0)$ is the average of the moment conditions. Thus the normalized difference between the estimated and the population values of the parameters, $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)$ is equal to a scaled $\left(-\mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)^{-1}\right)$ random variable $\left(\sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0)\right)$ that has an asymptotic normal distribution. By assumption $\sqrt{n}\mathbf{g}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0,\boldsymbol{\Sigma})$ and so

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}'\right)^{-1}\right) \tag{2.31}$$

where $\mathbf{G}_n\left(\bar{\boldsymbol{\theta}}\right)$ has been replaced with its limit as $n\rightarrow\infty$, $\mathbf{G}$.

$$\begin{aligned} \mathbf{G} &= \text{plim}_{n\rightarrow\infty}\left.\frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \text{plim}_{n\rightarrow\infty} n^{-1}\sum_{i=1}^{n}\left.\frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \end{aligned} \tag{2.32}$$

Since $\hat{\boldsymbol{\theta}}$ is a consistent estimator, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and so $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ since it is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. This form of asymptotic covariance is known as a "sandwich" covariance estimator.

Then for any points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ there exists $\bar{\boldsymbol{\theta}}$ lying on the segment between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ such that $s(\boldsymbol{\theta}) = s(\boldsymbol{\theta}_0) + \nabla s\left(\bar{\boldsymbol{\theta}}\right)'(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$.

2.4.1.1 Inference on the Mean and Variance

To estimate the mean and variance by the method of moments, two moment conditions are needed,

$$n^{-1}\sum_{i=1}^{n} x_i = \hat{\mu}$$
$$n^{-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \hat{\sigma}^2$$

To derive the asymptotic distribution, begin by forming $\mathbf{g}_i$,

$$\mathbf{g}_i = \begin{bmatrix} x_i - \mu \\ (x_i-\mu)^2 - \sigma^2 \end{bmatrix}$$

Note that $\mathbf{g}_i$ is mean 0 and a function of a single x_i so that $\mathbf{g}_i$ is also i.i.d.. The covariance of $\mathbf{g}_i$ is given by

$$\begin{aligned}
\mathbf{\Sigma} = \mathrm{E}\left[\mathbf{g}_i\mathbf{g}_i'\right] &= \mathrm{E}\left[\begin{bmatrix} x_i - \mu \\ (x_i-\mu)^2 - \sigma^2 \end{bmatrix}\begin{bmatrix} x_i - \mu & (x_i-\mu)^2 - \sigma^2 \end{bmatrix}\right] \\
&= \mathrm{E}\begin{bmatrix} (x_i-\mu)^2 & (x_i-\mu)\left((x_i-\mu)^2-\sigma^2\right) \\ (x_i-\mu)\left((x_i-\mu)^2-\sigma^2\right) & \left((x_i-\mu)^2-\sigma^2\right)^2 \end{bmatrix} \\
&= \mathrm{E}\begin{bmatrix} (x_i-\mu)^2 & (x_i-\mu)^3 - \sigma^2(x_i-\mu) \\ (x_i-\mu)^3 - \sigma^2(x_i-\mu) & (x_i-\mu)^4 - 2\sigma^2(x_i-\mu)^2 + \sigma^4 \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}
\end{aligned} \tag{2.33}$$

and the Jacobian is

$$\begin{aligned}
\mathbf{G} &= \mathrm{plim}_{n\to\infty}\, n^{-1}\sum_{i=1}^{n} \left.\frac{\partial \mathbf{g}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \mathrm{plim}_{n\to\infty}\, n^{-1}\sum_{i=1}^{n} \begin{bmatrix} -1 & 0 \\ -2(x_i-\mu) & -1 \end{bmatrix}.
\end{aligned}$$

Since $\mathrm{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}(x_i-\mu) = \mathrm{plim}_{n\to\infty}\bar{x}_n - \mu = 0$,

$$\mathbf{G} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus, the asymptotic distribution of the method of moments estimator of $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)'$ is

$$\sqrt{n}\left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}\right)$$

since $\mathbf{G} = -\mathbf{I}_2$ and so $\mathbf{G}^{-1}\mathbf{\Sigma}\left(\mathbf{G}^{-1}\right)' = -\mathbf{I}_2\mathbf{\Sigma}(-\mathbf{I}_2) = \mathbf{\Sigma}$.

2.4.2 Maximum Likelihood

The steps to deriving the asymptotic distribution of a ML estimator are similar to those for a method of moments estimator where the score of the likelihood takes the place of the moment conditions. The maximum likelihood estimator is defined as the maximum of the log-likelihood of the data with respect to the parameters,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{y}). \tag{2.34}$$

When the data are i.i.d., the log-likelihood can be factored into n log-likelihoods, one for each observation[16],

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}; y_i). \tag{2.35}$$

It is useful to work with the average log-likelihood directly, and so define

$$\bar{l}_n(\boldsymbol{\theta}; \mathbf{y}) = n^{-1} \sum_{i=1}^{n} l_i(\boldsymbol{\theta}; y_i). \tag{2.36}$$

The intuition behind the asymptotic distribution follows from the use of the average. Under some regularity conditions, $\bar{l}_n(\boldsymbol{\theta}; \mathbf{y})$ converges uniformly in $\boldsymbol{\theta}$ to $\mathrm{E}[l(\boldsymbol{\theta}; y_i)]$. However, since the average log-likelihood is becoming a good approximation for the expectation of the log-likelihood, the value of $\boldsymbol{\theta}$ that maximizes the log-likelihood of the data and its expectation will be very close for n sufficiently large. As a result,whenever the log-likelihood is differentiable and the range of y_i does not depend on any of the parameters in $\boldsymbol{\theta}$,

$$\mathrm{E}\left[\left. \frac{\partial \bar{l}_n(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right] = \mathbf{0} \tag{2.37}$$

where $\boldsymbol{\theta}_0$ are the parameters of the data generating process. This follows since

$$\begin{aligned} \int_{\mathcal{S}_\mathbf{y}} \left. \frac{\partial \bar{l}_n(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{y}; \boldsymbol{\theta}_0)\, dy &= \int_{\mathcal{S}_\mathbf{y}} \frac{\left. \frac{\partial f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f(\mathbf{y}; \boldsymbol{\theta}_0)} f(\mathbf{y}; \boldsymbol{\theta}_0)\, dy \\ &= \int_{\mathcal{S}_\mathbf{y}} \left. \frac{\partial f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} dy \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left. \int_{\mathcal{S}_\mathbf{y}} f(\mathbf{y}; \boldsymbol{\theta}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} dy \end{aligned} \tag{2.38}$$

[16] Even when the data are not i.i.d., the log-likelihood can be factored into n log-likelihoods using conditional distributions for $y_2, \ldots, y_i$ and the marginal distribution of y_1,

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{n=2}^{N} l_i\left(\boldsymbol{\theta}; y_i \,|\, y_{i-1}, \ldots, y_1\right) + l_1(\boldsymbol{\theta}; y_1).$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \boldsymbol{\theta}} 1 \\
&= 0
\end{aligned}
$$

where $\mathcal{S}_{\mathbf{y}}$ denotes the support of $\mathbf{y}$. The scores of the average log-likelihood are

$$
\frac{\partial \bar{l}_n(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} = n^{-1} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \tag{2.39}
$$

and when y_i is i.i.d. the scores will be i.i.d., and so the average scores will follow a law of large numbers for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$. Thus

$$
n^{-1} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \xrightarrow{a.s.} \mathrm{E}\left[\frac{\partial l(\boldsymbol{\theta}; Y_i)}{\partial \boldsymbol{\theta}}\right] \tag{2.40}
$$

As a result, the population value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$, will also asymptotically solve the first order condition. The average scores are also the basis of the asymptotic normality of maximum likelihood estimators. Under some further regularity conditions, the average scores will follow a central limit theorem, and so

$$
\sqrt{n} \nabla_{\boldsymbol{\theta}} \bar{l}(\boldsymbol{\theta}_0) \equiv \sqrt{n} \left(n^{-1} \sum_{i=1}^{n} \frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{d} N(\mathbf{0}, \mathcal{J}). \tag{2.41}
$$

Taking a mean value expansion around $\boldsymbol{\theta}_0$,

$$
\begin{aligned}
\sqrt{n} \nabla_{\boldsymbol{\theta}} \bar{l}\left(\hat{\boldsymbol{\theta}}\right) &= \sqrt{n} \nabla_{\boldsymbol{\theta}} \bar{l}(\boldsymbol{\theta}_0) + \sqrt{n} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \bar{l}\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\
\mathbf{0} &= \sqrt{n} \nabla_{\boldsymbol{\theta}} \bar{l}(\boldsymbol{\theta}_0) + \sqrt{n} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \bar{l}\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\
-\sqrt{n} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \bar{l}\left(\bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \sqrt{n} \nabla_{\boldsymbol{\theta}} \bar{l}(\boldsymbol{\theta}_0) \\
\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \left[-\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \bar{l}\left(\bar{\boldsymbol{\theta}}\right)\right]^{-1} \sqrt{n} \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}_0)
\end{aligned}
$$

where

$$
\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} \bar{l}\left(\bar{\boldsymbol{\theta}}\right) \equiv n^{-1} \sum_{i=1}^{n} \frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Bigg|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} \tag{2.42}
$$

and where $\bar{\boldsymbol{\theta}}$ is a vector whose elements lie between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and so functions of $\bar{\boldsymbol{\theta}}$ will converge to their value at $\boldsymbol{\theta}_0$, and the asymptotic distribution of the maximum likelihood estimator is

$$
\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}\right) \tag{2.43}
$$

where

$$
\mathcal{I} = -\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \tag{2.44}
$$

and

$$\mathcal{J} = \mathrm{E}\left[\frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \tag{2.45}$$

The asymptotic covariance matrix can be further simplified using the information matrix equality which states that $\mathcal{I} - \mathcal{J} \xrightarrow{p} \mathbf{0}$ and so

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}^{-1}\right) \tag{2.46}$$

or equivalently

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{J}^{-1}\right). \tag{2.47}$$

The information matrix equality follows from taking the derivative of the expected score,

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} &= \frac{1}{f(\mathbf{y}; \boldsymbol{\theta})} \frac{\partial^2 f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \frac{1}{f(\mathbf{y}; \boldsymbol{\theta})^2} \frac{\partial f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'} &= \frac{1}{f(\mathbf{y}; \boldsymbol{\theta})} \frac{\partial^2 f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \end{aligned} \tag{2.48}$$

and so, when the model is correctly specified,

$$\begin{aligned} \mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right] &= \int_{\mathcal{S}_\mathbf{y}} \frac{1}{f(\mathbf{y}; \boldsymbol{\theta})} \frac{\partial^2 f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{y}; \boldsymbol{\theta}) d\,\mathbf{y} \\ &= \int_{\mathcal{S}_\mathbf{y}} \frac{\partial^2 f(\mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} d\,\mathbf{y} \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \int_{\mathcal{S}_\mathbf{y}} f(\mathbf{y}; \boldsymbol{\theta}_0) d\,\mathbf{y} \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} 1 \\ &= 0. \end{aligned}$$

and

$$\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = -\mathrm{E}\left[\frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right].$$

A related concept, and one which applies to ML estimators when the information matrix equality holds – at least asymptotically – is the Cramér-Rao lower bound.

Theorem 2.10 (Cramér-Rao Inequality)**.** *Let $f(\mathbf{y}; \boldsymbol{\theta})$ be the joint density of $\mathbf{y}$ where $\boldsymbol{\theta}$ is a k dimensional parameter vector. Let $\hat{\boldsymbol{\theta}}$ be a consistent estimator of $\boldsymbol{\theta}$ with finite covariance. Under some regularity condition on $f(\cdot)$*

$$\mathrm{avar}\left(\hat{\boldsymbol{\theta}}\right) \geq \mathcal{I}^{-1}(\boldsymbol{\theta}) \tag{2.49}$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2 \ln f(Y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]. \tag{2.50}$$

The important implication of the Cramér-Rao theorem is that maximum likelihood estimators, which are generally consistent, are asymptotically efficient.[17] This guarantee makes a strong case for using the maximum likelihood when available.

2.4.2.1 Inference in a Poisson MLE

Recall that the log-likelihood in a Poisson MLE is

$$l(\lambda;\mathbf{y}) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} y_i - \sum_{i=1}^{y_i}\ln(i)$$

and that the first order condition is

$$\frac{\partial l(\lambda;\mathbf{y})}{\partial \lambda} = -n + \lambda^{-1}\sum_{i=1}^{n} y_i.$$

The MLE was previously shown to be $\hat{\lambda} = n^{-1}\sum_{i=1}^{n} y_i$. To compute the variance, take the expectation of the negative of the second derivative,

$$\frac{\partial^2 l(\lambda; y_i)}{\partial \lambda^2} = -\lambda^{-2} y_i$$

and so

$$\begin{aligned}
\mathcal{I} = -\mathrm{E}\left[\frac{\partial^2 l(\lambda; y_i)}{\partial \lambda^2}\right] &= -\mathrm{E}\left[-\lambda^{-2} y_i\right] \\
&= \left[\lambda^{-2}\mathrm{E}\left[y_i\right]\right] \\
&= \left[\lambda^{-2}\lambda\right] \\
&= \left[\frac{\lambda}{\lambda^2}\right] \\
&= \lambda^{-1}
\end{aligned}$$

and so $\sqrt{n}\left(\hat{\lambda} - \lambda_0\right) \xrightarrow{d} N(0, \lambda)$ since $\mathcal{I}^{-1} = \lambda$.

Alternatively the covariance of the scores could be used to compute the parameter covariance,

$$\begin{aligned}
\mathcal{J} &= \mathrm{V}\left[\left(-1 + \frac{y_i}{\lambda}\right)^2\right] \\
&= \frac{1}{\lambda^2}\mathrm{V}\left[y_i\right] \\
&= \lambda^{-1}.
\end{aligned}$$

[17] The Cramér-Rao bound also applied in finite samples when $\hat{\boldsymbol{\theta}}$ is unbiased. While most maximum likelihood estimators are biased in finite samples, there are important cases where estimators are unbiased for any sample size and so the Cramér-Rao theorem will apply in finite samples. Linear regression is an important case where the Cramér-Rao theorem applies in finite samples (under some strong assumptions).

$\mathcal{I} = \mathcal{J}$ and so the IME holds when the data are Poisson distributed. If the data were not Poisson distributed, then it would not normally be the case that $\mathrm{E}[y_i] = \mathrm{V}[y_i] = \lambda$, and so $\mathcal{I}$ and $\mathcal{J}$ would not (generally) be equal.

2.4.2.2 Inference in the Normal (Gaussian) MLE

Recall that the MLE estimators of the mean and variance are

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} y_i$$
$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

and that the log-likelihood is

$$l(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

Taking the derivative with respect to the parameter vector, $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)'$,

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu} = \sum_{i=1}^{n}\frac{(y_i-\mu)}{\sigma^2}$$
$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^4}.$$

The second derivatives are

$$\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu \partial \mu} = -\sum_{i=1}^{n}\frac{1}{\sigma^2}$$
$$\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \mu \partial \sigma^2} = -\sum_{i=1}^{n}\frac{(y_i-\mu)}{\sigma^4}$$
$$\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \sigma^2 \partial \sigma^2} = \frac{n}{2\sigma^4} - \frac{2}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^6}.$$

The first does not depend on data and so no expectation is needed. The other two have expectations,

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \mu \partial \sigma^2}\right] &= \mathrm{E}\left[-\frac{(y_i - \mu)}{\sigma^4}\right] \\
&= -\frac{(\mathrm{E}[y_i] - \mu)}{\sigma^4} \\
&= -\frac{\mu - \mu}{\sigma^4} \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \sigma^2 \partial \sigma^2}\right] &= \mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{2}{2}\frac{(y_i - \mu)^2}{\sigma^6}\right] \\
&= \frac{1}{2\sigma^4} - \frac{\mathrm{E}\left[(y_i - \mu)^2\right]}{\sigma^6} \\
&= \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} \\
&= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \\
&= -\frac{1}{2\sigma^4}
\end{aligned}$$

Putting these together, the expected Hessian can be formed,

$$\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = \begin{bmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2\sigma^4} \end{bmatrix}$$

and so the asymptotic covariance is

$$\begin{aligned}
\mathcal{I}^{-1} = -\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} &= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}
\end{aligned}$$

The asymptotic distribution is then

$$\sqrt{n}\left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\right)$$

Note that this is different from the asymptotic variance for the method of moments estimator of the mean and the variance. This is because the data have been assumed to come from a normal distribution and so the MLE is correctly specified. As a result $\mu_3 = 0$ (the normal is symmetric) and the IME holds. In general the IME does not hold and so the asymptotic covariance may take a different form which depends on the moments of the data as in eq. (2.33).

2.4.3 Quasi Maximum Likelihood

While maximum likelihood is an appealing estimation approach, it has one important drawback: knowledge of $f(\mathbf{y};\boldsymbol{\theta})$. In practice the density assumed in maximum likelihood estimation, $f(\mathbf{y};\boldsymbol{\theta})$, is misspecified for the actual density of $\mathbf{y}$, $g(\mathbf{y})$. This case has been widely studied and estimators where the distribution is misspecified are known as quasi-maximum likelihood (QML) estimators. QML estimators generally lose all of the features that make maximum likelihood estimators so appealing: they are generally inconsistent for the parameters of interest, the information matrix equality does not hold and they do not achieve the Cramér-Rao lower bound.

First, consider the expected score from a QML estimator,

$$\begin{aligned}
\mathrm{E}_g\left[\frac{\partial l(\boldsymbol{\theta}_0;\mathbf{y})}{\partial \boldsymbol{\theta}}\right] &= \int_{\mathcal{S}_\mathbf{y}} \frac{\partial l(\boldsymbol{\theta}_0;\mathbf{y})}{\partial \boldsymbol{\theta}} g(\mathbf{y})\, dy \\
&= \int_{\mathcal{S}_\mathbf{y}} \frac{\partial l(\boldsymbol{\theta}_0;\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{f(\mathbf{y};\boldsymbol{\theta}_0)}{f(\mathbf{y};\boldsymbol{\theta}_0)} g(\mathbf{y})\, dy \\
&= \int_{\mathcal{S}_\mathbf{y}} \frac{\partial l(\boldsymbol{\theta}_0;\mathbf{y})}{\partial \boldsymbol{\theta}} \frac{g(\mathbf{y})}{f(\mathbf{y};\boldsymbol{\theta}_0)} f(\mathbf{y};\boldsymbol{\theta}_0)\, dy \\
&= \int_{\mathcal{S}_\mathbf{y}} h(\mathbf{y}) \frac{\partial l(\boldsymbol{\theta}_0;\mathbf{y})}{\partial \boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta}_0)\, dy
\end{aligned} \tag{2.51}$$

which shows that the QML estimator can be seen as a weighted average with respect to the density assumed. However these weights depend on the data, and so it will no longer be the case that the expectation of the score at $\boldsymbol{\theta}_0$ will necessarily be 0. Instead QML estimators generally converge to another value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$, that depends on both $f(\cdot)$ and $g(\cdot)$ and is known as the pseudo-true value of $\boldsymbol{\theta}$.

The other important consideration when using QML to estimate parameters is that the Information Matrix Equality (IME) no longer holds, and so "sandwich" covariance estimators must be used and likelihood ratio statistics will not have standard χ^2 distributions. An alternative interpretation of a QML estimator is that of a method of moments estimator where the scores of $l(\boldsymbol{\theta};\mathbf{y})$ are used to choose the moments. With this interpretation, the distribution theory of the method of moments estimator will apply as long as the scores, evaluated at the pseudo-true parameters, follow a CLT.

2.4.3.1 The Effect of the Data Distribution on Estimated Parameters

Figure 2.4 contains three distributions (left column) and the asymptotic covariance of the mean and the variance estimators, illustrated through joint confidence ellipses containing 80, 95 and 99% probability the true value is within their bounds (right column).[18] The ellipses were all derived from the asymptotic covariance of $\hat{\mu}$ and $\hat{\sigma}^2$ where the data are i.i.d. and distributed according to a *mixture of normals* distribution where

[18] The ellipses are centered at (0,0) since the population value of the parameters has been subtracted. Also note that even though the confidence ellipse for $\hat{\sigma}^2$ extended into the negative space, these must be divided by $\sqrt{n}$ and re-centered at the estimated value when used.

	p	μ_1	σ_1^2	μ_2	σ_2^2
Standard Normal	1	0	1	0	1
Contaminated Normal	.95	0	.8	0	4.8
Right Skewed Mixture	.05	2	.5	-.1	.8

Table 2.1: Parameter values used in the mixtures of normals illustrated in figure 2.4.

$$y_i = \begin{cases} \mu_1 + \sigma_1 z_i & \text{with probability } p \\ \mu_2 + \sigma_2 z_i & \text{with probability } 1-p \end{cases}$$

where z is a standard normal. A mixture of normals is constructed from mixing draws from a finite set of normals with possibly different means and/or variances and can take a wide variety of shapes. All of the variables were constructed so that $\mathrm{E}[y_i] = 0$ and $\mathrm{V}[y_i] = 1$. This requires

$$p\mu_1 + (1-p)\mu_2 = 0$$

and

$$p\left(\mu_1^2 + \sigma_1^2\right) + (1-p)\left(\mu_2^2 + \sigma_2^2\right) = 1.$$

The values used to produce the figures are listed in table 2.1. The first set is simply a standard normal since $p = 1$. The second is known as a contaminated normal and is composed of a frequently occurring (95% of the time) mean-zero normal with variance slightly smaller than 1 (.8), contaminated by a rare but high variance (4.8) mean-zero normal. This produces heavy tails but does not result in a skewed distribution. The final example uses different means and variance to produce a right (positively) skewed distribution.

The confidence ellipses illustrated in figure 2.4 are all derived from estimators produced assuming that the data are normal, but using the "sandwich" version of the covariance, $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$. The top panel illustrates the correctly specified maximum likelihood estimator. Here the confidence ellipse is symmetric about its center. This illustrates that the parameters are uncorrelated – and hence independent, since they are asymptotically normal – and that they have different variances. The middle panel has a similar shape but is elongated on the variance axis (x). This illustrates that the asymptotic variance of $\hat{\sigma}^2$ is affected by the heavy tails of the data (large 4^{th} moment) of the contaminated normal. The final confidence ellipse is rotated which reflects that the mean and variance estimators are no longer asymptotically independent. These final two cases are examples of QML; the estimator is derived assuming a normal distribution but the data are not. In these examples, the estimators are still consistent but have different covariances.[19]

2.4.4 The Delta Method

Some theories make predictions about *functions* of parameters rather than on the parameters directly. One common example in finance is the Sharpe ratio, S, defined

[19]While these examples are consistent, it is not generally the case that the parameters estimated using a misspecified likelihood (QML) are consistent for the quantities of interest.

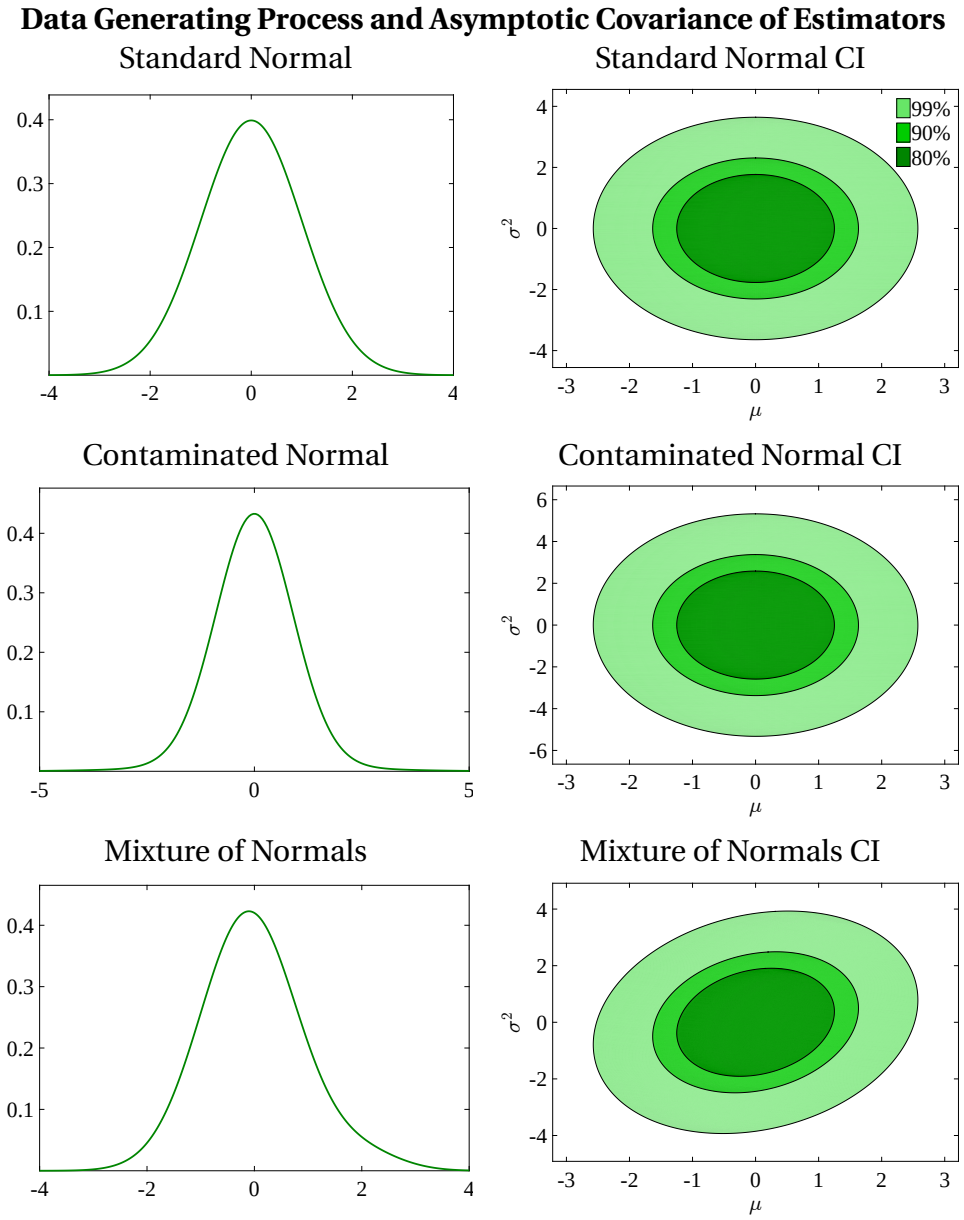

Figure 2.4: The six subplots illustrate how the data generating process, not the assumed model, determine the asymptotic covariance of parameter estimates. In each row of panels, the left shows the distribution of the data from a mixture of normals, $y_i = \mu_1 + \sigma_1 z_i$ with probability p and $y_i = \mu_2 + \sigma_2 z_i$ with probability $1-p$. The right shows the asymptotic distribution of $\hat{\mu}$ and $\hat{\sigma}^2$. The parameters were chosen so that $\mathrm{E}[y_i] = 0$ and $\mathrm{V}[y_i] = 1$. Different parameter configurations produce a standard normal (top), a heavy tailed distribution known as a contaminated normal (middle) and a skewed distribution (bottom).

$$S = \frac{\mathrm{E}\left[r - r_f\right]}{\sqrt{\mathrm{V}\left[r - r_f\right]}} \tag{2.52}$$

where r is the return on a risky asset and r_f is the risk-free rate – and so $r - r_f$ is the excess return on the risky asset. While the quantities in both the numerator and the denominator are standard statistics, the mean and the standard deviation, the ratio is not.

The delta method can be used to compute the covariance of functions of asymptotically normal parameter estimates.

Definition 2.13 (Delta method)**.** Let $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1}\boldsymbol{\Sigma}(\mathbf{G}')^{-1}\right)$ where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix. Further, suppose that $\mathbf{d}(\boldsymbol{\theta})$ is a m by 1 continuously differentiable vector function of $\boldsymbol{\theta}$ from $\mathbb{R}^k \to \mathbb{R}^m$. Then,

$$\sqrt{n}(\mathbf{d}(\hat{\boldsymbol{\theta}}) - \mathbf{d}(\boldsymbol{\theta}_0)) \xrightarrow{d} N\left(0, \mathbf{D}(\boldsymbol{\theta}_0)\left[\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}'\right)^{-1}\right]\mathbf{D}(\boldsymbol{\theta}_0)'\right)$$

where

$$\mathbf{D}(\boldsymbol{\theta}_0) = \left.\frac{\partial \mathbf{d}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{2.53}$$

2.4.4.1 Variance of the Sharpe Ratio

The Sharpe ratio is estimated by "plugging in" the usual estimators of the mean and the variance,

$$\hat{S} = \frac{\hat{\mu}}{\sqrt{\hat{\sigma}^2}}.$$

In this case $\mathbf{d}(\boldsymbol{\theta}_0)$ is a scalar function of two parameters, and so

$$\mathbf{d}(\boldsymbol{\theta}_0) = \frac{\mu}{\sqrt{\sigma}^2}$$

and

$$\mathbf{D}(\boldsymbol{\theta}_0) = \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix}$$

Recall that the asymptotic distribution of the estimated mean and variance is

$$\sqrt{n}\left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix}\right).$$

The asymptotic distribution of the Sharpe ratio can be constructed by combining the asymptotic distribution of $\hat{\boldsymbol{\theta}} = \left(\hat{\mu}, \hat{\sigma}^2\right)'$ with the $\mathbf{D}(\boldsymbol{\theta}_0)$, and so

$$\sqrt{n}\left(\hat{S} - S\right) \xrightarrow{d} N\left(0, \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix}'\right)$$

which can be simplified to

$$\sqrt{n}\left(\hat{S}-S\right) \xrightarrow{d} N\left(0, 1-\frac{\mu\mu_3}{\sigma^4}+\frac{\mu^2\left(\mu_4-\sigma^4\right)}{4\sigma^6}\right).$$

The asymptotic variance can be rearranged to provide some insight into the sources of uncertainty,

$$\sqrt{n}\left(\hat{S}-S\right) \xrightarrow{d} N\left(0, 1-S\times sk+\frac{1}{4}S^2\left(\kappa-1\right)\right),$$

where sk is the skewness and κ is the kurtosis. This shows that the variance of the Sharpe ratio will be higher when the data is negatively skewed or when the data has a large kurtosis (heavy tails), both empirical regularities of asset pricing data. If asset returns were normally distributed, and so $sk=0$ and $\kappa=3$, the expression of the asymptotic variance simplifies to

$$\mathrm{V}\left[\sqrt{n}\left(\hat{S}-S\right)\right] = 1+\frac{S^2}{2}, \tag{2.54}$$

which is expression commonly used as the variance of the Sharpe ratio. As this example illustrates the expression in eq. (2.54) is *only* correct if the skewness is 0 and returns have a kurtosis of 3 – something that would only be expected if returns are normal.

2.4.5 Estimating Covariances

The presentation of the asymptotic theory in this chapter does not provide a method to implement hypothesis tests since all of the distributions depend on the covariance of the scores and the expected second derivative or Jacobian in the method of moments. Feasible testing requires estimates of these. The usual method to estimate the covariance uses "plug-in" estimators. Recall that in the notation of the method of moments,

$$\mathbf{\Sigma} \equiv \mathrm{avar}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}_0\right)\right) \tag{2.55}$$

or in the notation of maximum likelihood,

$$\mathcal{J} \equiv \mathrm{E}\left[\frac{\partial\, l\left(\boldsymbol{\theta};Y_i\right)}{\partial\,\boldsymbol{\theta}}\frac{\partial\, l\left(\boldsymbol{\theta};Y_i\right)}{\partial\,\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]. \tag{2.56}$$

When the data are i.i.d., the scores or moment conditions should be i.i.d., and so the variance of the average is the average of the variance. The "plug-in" estimator for $\mathbf{\Sigma}$ uses the moment conditions evaluated at $\hat{\boldsymbol{\theta}}$, and so the covariance estimator for method of moments applications with i.i.d. data is

$$\hat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)' \tag{2.57}$$

which is simply the average outer-product of the moment condition. The estimator of $\mathbf{\Sigma}$ in the maximum likelihood is identical replacing $\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)$ with $\partial\, l\left(\boldsymbol{\theta};y_i\right)/\partial\,\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$,

$$\hat{\mathcal{J}} = n^{-1} \sum_{i=1}^{n} \frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta}'} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.58}$$

The "plug-in" estimator for the second derivative of the log-likelihood or the Jacobian of the moment conditions is similarly defined,

$$\hat{\mathbf{G}} = n^{-1} \sum_{i=1}^{n} \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{2.59}$$

or for maximum likelihood estimators

$$\hat{\mathcal{I}} = n^{-1} \sum_{i=1}^{n} -\frac{\partial^2 l(\boldsymbol{\theta}; y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.60}$$

2.4.6 Estimating Covariances with Dependent Data

The estimators in eq. (2.57) and eq. (2.58) are only appropriate when the moment conditions or scores are not correlated across i.[20] If the moment conditions or scores are correlated across observations the covariance estimator (but not the Jacobian estimator) must be changed to account for the dependence. Since $\boldsymbol{\Sigma}$ is defined as the variance of a sum it is necessary to account for both the sum of the variances *plus* all of the covariances.

$$\begin{aligned} \boldsymbol{\Sigma} &\equiv \text{avar}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}_0) \right) \\ &= \lim_{n\to\infty} n^{-1} \left(\sum_{i=1}^{n} \text{E}\left[\mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)' \right] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{E}\left[\mathbf{g}_j(\boldsymbol{\theta}_0) \mathbf{g}_{j-i}(\boldsymbol{\theta}_0)' + \mathbf{g}_{j-i}(\boldsymbol{\theta}_0) \mathbf{g}_j(\boldsymbol{\theta}_0) \right] \right) \end{aligned} \tag{2.61}$$

This expression depends on both the usual covariance of the moment conditions and on the covariance between the scores. When using i.i.d. data the second term vanishes since the moment conditions must be uncorrelated and so cross-products must have expectation 0.

If the moment conditions are correlated across i then covariance estimator must be adjusted to account for this. The obvious solution is to estimate the expectations of the cross terms in eq. (2.57) with their sample analogues, which would result in the covariance estimator

$$\hat{\boldsymbol{\Sigma}}_{\text{DEP}} = n^{-1} \left[\sum_{i=1}^{n} \mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right) \mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)' + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\mathbf{g}_j\left(\hat{\boldsymbol{\theta}}\right) \mathbf{g}_{j-i}\left(\hat{\boldsymbol{\theta}}\right)' + \mathbf{g}_{j-i}\left(\hat{\boldsymbol{\theta}}\right) \mathbf{g}_j\left(\hat{\boldsymbol{\theta}}\right)' \right) \right]. \tag{2.62}$$

This estimator is always zero since $\hat{\boldsymbol{\Sigma}}_{\text{DEP}} = n^{-1} \left(\sum_{i=1}^{n} \mathbf{g}_i \right) \left(\sum_{i=1}^{n} \mathbf{g}_i \right)'$ and $\sum_{i=1}^{n} \mathbf{g}_i = \mathbf{0}$, and so $\hat{\boldsymbol{\Sigma}}_{\text{DEP}}$

[20] Since i.i.d. implies no correlation, the i.i.d. case is trivially covered.

cannot be used in practice.[21] One solution is to truncate the maximum lag to be something less than $n-1$ (usually much less than $n-1$), although the truncated estimator is not guaranteed to be positive definite. A better solution is to combine truncation with a weighting function (known as a *kernel*) to construct an estimator which will consistently estimate the covariance and is guaranteed to be positive definite. The most common covariance estimator of this type is the Newey and West (1987) covariance estimator. Covariance estimators for dependent data will be examined in more detail in the chapters on time-series data.

2.5 Hypothesis Testing

Econometrics models are estimated in order to test hypotheses, for example, whether a financial theory is supported by data or to determine if a model with estimated parameters can outperform a naïveforecast. Formal hypothesis testing begins by specifying the null hypothesis.

Definition 2.14 (Null Hypothesis)**.** The null hypothesis, denoted H_0, is a statement about the population values of some parameters to be tested. The null hypothesis is also known as the maintained hypothesis.

The null defines the condition on the population parameters that is to be tested. A null can be either simple, for example, $H_0 : \mu = 0$, or complex, which allows for testing of multiple hypotheses. For example, it is common to test whether data exhibit any predictability using a regression model

$$y_i = \theta_1 + \theta_2 x_{2,i} + \theta_3 x_{3,i} + \epsilon_i, \tag{2.63}$$

and a composite null, $H_0 : \theta_2 = 0 \cap \theta_3 = 0$, often abbreviated $H_0 : \theta_2 = \theta_3 = 0$.[22]

Null hypotheses cannot be accepted; the data can either lead to *rejection of the null* or a *failure to reject the null*. Neither option is "accepting the null". The inability to accept the null arises since there are important cases where the data are not consistent with either the null or its testing complement, the alternative hypothesis.

Definition 2.15 (Alternative Hypothesis)**.** The alternative hypothesis, denoted H_1, is a complementary hypothesis to the null and determines the range of values of the population parameter that should lead to rejection of the null.

The alternative hypothesis specifies the population values of parameters for which the null should be rejected. In most situations, the alternative is the natural complement to the null in the sense

[21] The scalar version of $\hat{\Sigma}_{\text{DEP}}$ may be easier to understand. If g_i is a scalar, then

$$\hat{\sigma}^2_{\text{DEP}} = n^{-1}\left[\sum_{i=1}^{n} g_i^2\left(\hat{\theta}\right) + 2\sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} g_j\left(\hat{\theta}\right) g_{j-i}\left(\hat{\theta}\right)\right)\right].$$

The first term is the usual variance estimator and the second term is the sum of the $(n-1)$ covariance estimators. The more complicated expression in eq. (2.62) arises since order matters when multiplying vectors.

[22] $\cap$, the intersection operator, is used since the null requires both statements to be true.

that the null and alternative are exclusive of each other but inclusive of the range of the population parameter. For example, when testing whether a random variable has mean 0, the null is $H_0 : \mu = 0$ and the usual alternative is $H_1 : \mu \neq 0$.

In certain circumstances, usually motivated by theoretical considerations, one-sided alternatives are desirable. One-sided alternatives only reject for population parameter values on one side of zero and so test using one-sided alternatives may not reject even if both the null and alternative are false. Noting that a risk premium must be positive (if it exists), the null hypothesis of $H_0 : \mu = 0$ should be tested against the alternative $H_1 : \mu > 0$. This alternative indicates the null should only be rejected if there is compelling evidence that the mean is positive. These hypotheses further specify that data consistent with large negative values of μ should not lead to rejection. Focusing the alternative often leads to an increased probability to rejecting a false null. This occurs since the alternative is directed (positive values for μ), and less evidence is required to be convinced that the null is not valid.

Like null hypotheses, alternatives can be composite. The usual alternative to the null $H_0 : \theta_2 = 0 \cap \theta_3 = 0$ is $H_1 : \theta_2 \neq 0 \cup \theta_3 \neq 0$ and so the null should be rejected whenever any of the statements in the null are false – in other words if either or both $\theta_2 \neq 0$ or $\theta_3 \neq 0$. Alternatives can also be formulated as lists of exclusive outcomes.[23] When examining the relative precision of forecasting models, it is common to test the null that the forecast performance is equal against a composite alternative that the forecasting performance is superior for model A *or* that the forecasting performance is superior for model B. If δ is defined as the average forecast performance difference, then the null is $H_0 : \delta = 0$ and the composite alternatives are $H_1^A : \delta > 0$ and $H_1^B : \delta < 0$, which indicate superior performance of models A and B, respectively.

Once the null and the alternative have been formulated, a hypothesis test is used to determine whether the data support the alternative.

Definition 2.16 (Hypothesis Test)**.** A hypothesis test is a rule that specifies which values to reject H_0 in favor of H_1.

Hypothesis testing requires a test statistic, for example, an appropriately standardized mean, and a critical value. The null is rejected when the test statistic is larger than the critical value.

Definition 2.17 (Critical Value)**.** The critical value for a α-sized test, denoted C_α, is the value where a test statistic, T, indicates rejection of the null hypothesis when the null is true.

The region where the test statistic is outside of the critical value is known as the rejection region.

Definition 2.18 (Rejection Region)**.** The rejection region is the region where $T > C_\alpha$.

An important event occurs when the null is correct but the hypothesis is rejected. This is known as a Type I error.

Definition 2.19 (Type I Error)**.** A Type I error is the event that the null is rejected when the null is true.

A closely related concept is the size of the test. The size controls how often Type I errors should occur.

[23] The $\cup$ symbol indicates the union of the two alternatives.

		Decision	
		Do not reject H_0	Reject H_0
Truth	H_0	Correct	Type I Error (Size)
	H_1	Type II Error	Correct (Power)

Table 2.2: Outcome matrix for a hypothesis test. The diagonal elements are both correct decisions. The off diagonal elements represent Type I error, when the null is rejected but is valid, and Type II error, when the null is not rejected and the alternative is true.

Definition 2.20 (Size)**.** The size or level of a test, denoted α, is the probability of rejecting the null when the null is true. The size is also the probability of a Type I error.

Typical sizes include 1%, 5%, and 10%, although ideally, the selected size should reflect the decision makers preferences over incorrectly rejecting the null. When the opposite occurs, the null is *not* rejected when the alternative is true, a Type II error is made.

Definition 2.21 (Type II Error)**.** A Type II error is the event that the null is not rejected when the alternative is true.

Type II errors are closely related to the power of a test.

Definition 2.22 (Power)**.** The power of the test is the probability of rejecting the null when the alternative is true. The power is equivalently defined as 1 minus the probability of a Type II error.

The two error types, size and power are summarized in table 2.2.

A perfect test would have unit power against any alternative. In other words, whenever the alternative is true it would reject immediately. Practically the power of a test is a function of both the sample size and the distance between the population value of a parameter and its value under the null. A test is said to be consistent if the power of the test goes to 1 as $n \to \infty$ whenever the population value lies in the area defined by the alternative hypothesis. Consistency is an important characteristic of a test, but it is usually considered more important to have correct size rather than to have high power. Because power can always be increased by distorting the size, and it is useful to consider a related measure known as the *size-adjusted power*. The size-adjusted power examines the power of a test in excess of size. Since a test should reject at size even when the null is true, it is useful to examine the percentage of times it *will* reject in excess of the percentage it *should* reject.

One useful tool for presenting results of test statistics is the p-value, or simply the p-val.

Definition 2.23 (P-value)**.** The p-value is the probability of observing a value as large as the observed test statistic given the null is true. The p-value is also:

- The largest size (α) where the null hypothesis cannot be rejected.

- The smallest size where the null hypothesis can be rejected.

The primary advantage of a p-value is that it immediately demonstrates which test sizes would lead to rejection: anything above the p-value. It also improves on the common practice of reporting the test statistic alone since p-values can be interpreted without knowledge of the distribution of the test statistic. However, since it incorporates information about a specific test statistic and its associated distribution, the formula used to compute the p-value is problem specific.
A related representation is the confidence interval for a parameter.

Definition 2.24 (Confidence Interval)**.** A confidence interval for a scalar parameter is the range of values, $\theta_0 \in (\underline{C}_\alpha, \overline{C}_\alpha)$ where the null $H_0 : \theta = \theta_0$ cannot be rejected for a size of α.

The formal definition of a confidence interval is not usually sufficient to uniquely identify the confidence interval. Suppose that a $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$. The common 95% confidence interval is $(\hat{\theta} - 1.96\sigma^2, \hat{\theta} + 1.96\sigma^2)$. This set is known as the symmetric confidence interval and is formally defined as points $(\underline{C}_\alpha, \overline{C}_\alpha)$ where $\Pr(\theta_0) \in (\underline{C}_\alpha, \overline{C}_\alpha) = 1 - \alpha$ and $\underline{C}_\alpha - \theta = \theta - \overline{C}_\alpha)$. An alternative, but still valid, confidence interval can be defined as $(-\infty, \hat{\theta} + 1.645\sigma^2)$. This would also contain the true value with probability 95%. In general, symmetric confidence intervals should be used, especially for asymptotically normal parameter estimates. In rare cases where symmetric confidence intervals are not appropriate, other options for defining a confidence interval include shortest interval, so that the confidence interval is defined as values $(\underline{C}_\alpha, \overline{C}_\alpha)$ where $\Pr(\theta_0) \in (\underline{C}_\alpha, \overline{C}_\alpha) = 1 - \alpha$ subject to $\overline{C}_\alpha - \underline{C}_\alpha$ chosen to be as small as possible, or symmetric in probability, so that the confidence interval satisfies $\Pr(\theta_0) \in (\underline{C}_\alpha, \hat{\theta}) = \Pr(\theta_0) \in (\hat{\theta}, \overline{C}_\alpha) = 1/2 - \alpha/2$. When constructing confidence internals for parameters that are asymptotically normal, these three definitions coincide.

2.5.0.1 Size and Power of a Test of the Mean with Normal Data

Suppose n i.i.d. normal random variables have unknown mean μ but known variance σ^2 and so the sample mean, $\bar{y} = n^{-1}\sum_{i=1}^n y_i$, is then distributed $N(\mu, \sigma^2/N)$. When testing a null that $H_0 : \mu = \mu_0$ against an alternative $H_1 : \mu \neq \mu_0$, the size of the test is the probability that the null is rejected when it is true. Since the distribution under the null is $N(\mu_0, \sigma^2/N)$ and the size can be set to α by selecting points where $\Pr\left(\hat{\mu} \in \left(\underline{C}_\alpha, \overline{C}_\alpha\right) | \mu = \mu_0\right) = 1 - \alpha$. Since the distribution is normal, one natural choice is to select the points symmetrically so that $\underline{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(\alpha/2)$ and $\overline{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(1 - \alpha/2)$ where $\Phi(\cdot)$ is the cdf of a standard normal.

The power of the test is defined as the probability the null is rejected when the alternative is true. This probability will depend on the population mean, μ_1, the sample size, the test size and mean specified by the null hypothesis. When testing using an α-sized test, rejection will occur when $\hat{\mu} < \mu_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(\alpha/2)$ or $\hat{\mu} > \mu_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(1 - \alpha/2)$. Since under the alternative $\hat{\mu}$ is $N\left(\mu_1, \sigma^2\right)$, these probabilities will be

$$\Phi\left(\frac{\mu_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(\alpha/2) - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right) = \Phi\left(\frac{\underline{C}_\alpha - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right)$$

and

$$1-\Phi\left(\frac{\mu_0+\frac{\sigma}{\sqrt{N}}\Phi^{-1}(1-\alpha/2)-\mu_1}{\frac{\sigma}{\sqrt{N}}}\right)=1-\Phi\left(\frac{\overline{C}_\alpha-\mu_1}{\frac{\sigma}{\sqrt{N}}}\right).$$

The total probability that the null is rejected is known as the power function,

$$\text{Power}\left(\mu_0,\mu_1,\sigma,\alpha,N\right)=\Phi\left(\frac{\underline{C}_\alpha-\mu_1}{\frac{\sigma}{\sqrt{N}}}\right)+1-\Phi\left(\frac{\overline{C}_\alpha-\mu_1}{\frac{\sigma}{\sqrt{N}}}\right).$$

A graphical illustration of the power is presented in figure 2.5. The null hypothesis is $H_0:\mu=0$ and the alternative distribution was drawn at $\mu_1=.25$. The variance $\sigma^2=1$, $n=5$, and the size was set to 5%. The highlighted regions indicate the power: the area under the alternative distribution, and hence the probability, which is outside of the critical values. The bottom panel illustrates the power curve for the same parameters allowing n to range from 5 to 1,000. When n is small, the power is low even for alternatives far from the null. As n grows the power increases and when $n=1,000$, the power of the test is close to unity for alternatives greater than 0.1.

2.5.1 Statistical and Economic Significance

While testing can reject hypotheses and provide meaningful p-values, statistical significance is different from economic significance. Economic significance requires a more detailed look at the data than a simple hypothesis test. Establishing the statistical significance of a parameter is the first, and easy, step. The more difficult step is to determine whether the effect is economically important. Consider a simple regression model

$$y_i=\theta_1+\theta_2x_{2,i}+\theta_3x_{3,i}+\epsilon_i \tag{2.64}$$

and suppose that the estimates of both θ_2 and θ_3 are statistically different from zero. This can happen for a variety of reasons, including having an economically small impact accompanied with a very large sample. To assess the relative contributions other statistics such as the percentage of the variation that can be explained by either variable alone and/or the range and variability of the xs.

Another important aspect of economic significance is that rejection of a hypothesis, while formally as a "yes" or "no" question, should be treated in a more continuous manner. The p-value of a test statistic is a useful tool in this regard that can provide a deeper insight into the strength of the rejection. A p-val of .00001 is not the same as a p-value of .09999 even though a 10% test would reject for either.

2.5.2 Specifying Hypotheses

Formalized in terms of $\boldsymbol{\theta}$, a null hypothesis is

$$H_0:\mathbf{R}(\boldsymbol{\theta})=\mathbf{0} \tag{2.65}$$

where $\mathbf{R}(\cdot)$ is a function from $\mathbb{R}^k$ to $\mathbb{R}^m$, $m\leq k$, where m represents the number of hypotheses in a composite null. While this specification of hypotheses is very flexible, testing non-linear hypotheses raises some subtle but important technicalities and further discussion will be reserved

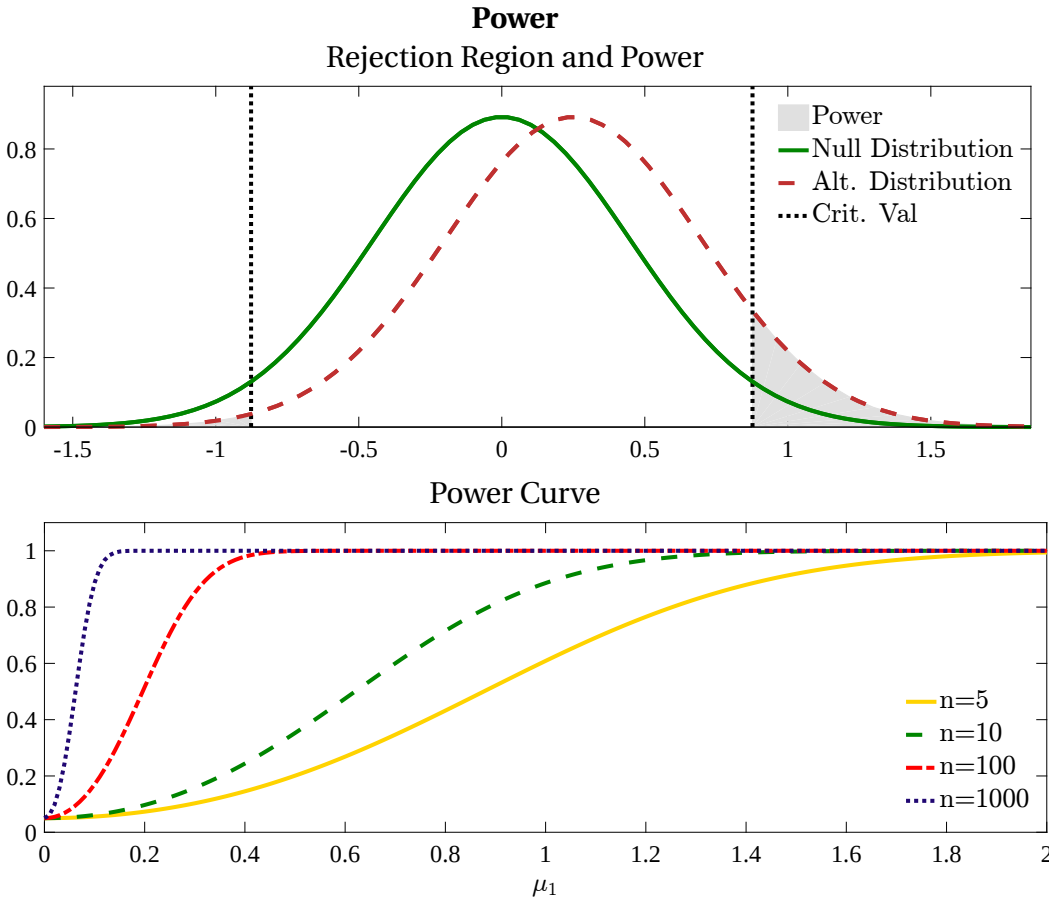

Figure 2.5: The top panel illustrates the power. The distribution of the mean under the null and alternative hypotheses were derived under that assumption that the data are i.i.d. normal with means $\mu_0 = 0$ and $\mu_1 = .25$, variance $\sigma^2 = 1$, $n = 5$ and $\alpha = .05$. The bottom panel illustrates the power function, in terms of the alternative mean, for the same parameters when $n = 5$, 10, 100 and 1,000.

---

for later. Initially, a subset of all hypotheses, those in the linear equality restriction (LER) class, which can be specified as

$$H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0} \tag{2.66}$$

will be examined where $\mathbf{R}$ is a m by k matrix and $\mathbf{r}$ is a m by 1 vector. All hypotheses in the LER class can be written as weighted sums of model parameters,

$$\begin{bmatrix} R_{11}\theta_1 + R_{12}\theta_2 \ldots + R_{1k}\theta_k &= r_1 \\ R_{21}\theta_1 + R_{22}\theta_2 \ldots + R_{2k}\theta_k &= r_2 \\ \vdots & \\ R_{m1}\theta_1 + R_{m2}\theta_2 \ldots + R_{mk}\theta_k &= r_i. \end{bmatrix} \tag{2.67}$$

Each linear hypothesis is represented as a row in the above set of equations. Linear equality constraints can be used to test parameter restrictions on $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)'$ such as

$$\begin{aligned} \theta_1 &= 0 \\ 3\theta_2 + \theta_3 &= 1 \\ \sum_{j=1}^{4} \theta_j &= 0 \\ \theta_1 = \theta_2 = \theta_3 &= 0. \end{aligned} \tag{2.68}$$

For example, the hypotheses in eq. (2.68) can be described in terms of $\mathbf{R}$ and $\mathbf{r}$ as

H_0	$\mathbf{R}$	$\mathbf{r}$
$\theta_1 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$	0
$3\theta_2 + \theta_3 = 1$	$\begin{bmatrix} 0 & 3 & 1 & 0 \end{bmatrix}$	1
$\sum_{j=1}^{k} \theta_j = 0$	$\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$	0
$\theta_1 = \theta_2 = \theta_3 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}'$

When using linear equality constraints, alternatives are generally formulated as $H_1 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} \neq 0$. Once both the null the alternative hypotheses have been postulated, it is necessary to determine whether the data are consistent with the null hypothesis using one of the many tests.

2.5.3 The Classical Tests

Three classes of statistics will be described to test hypotheses: Wald, Lagrange Multiplier, and Likelihood Ratio. Wald tests are perhaps the most intuitive: they directly test whether $\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}$, the value under the null, is close to zero by exploiting the asymptotic normality of the estimated parameters. Lagrange Multiplier tests incorporate the constraint into the estimation problem using a Lagrangian. If the constraint has a small effect on the value of objective function, the Lagrange multipliers, often described as the shadow price of a constraint in an economic application, should be close to zero. The magnitude of the scores 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 these restrictions are not statistically meaningful, this ratio should be close to one since the difference in the log-likelihoods should be small.

2.5.4 Wald Tests

Wald test statistics are possibly the most natural method to test a hypothesis and are often the simplest to compute since only the unrestricted model must be estimated. Wald tests directly

exploit the asymptotic normality of the estimated parameters to form test statistics with asymptotic χ^2_m distributions. Recall that a χ^2_ν random variable is defined to be the sum of ν independent standard normals squared, $\sum_{i=1}^{\nu} z_i^2$ where $z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. Recall that if $\mathbf{z}$ is a m-dimension normal vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.69}$$

then the standardized version of $\mathbf{z}$ can be constructed as

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{z} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I}). \tag{2.70}$$

Defining $\mathbf{w} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{z} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I})$, it is easy to see that $\mathbf{w}'\mathbf{w} = \sum_{m=1}^{M} w_m^2 \sim \chi^2_m$. In the usual case, the method of moments estimator, which nests ML and QML estimators as special cases, is asymptotically normal

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\right). \tag{2.71}$$

If null hypothesis, $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ is true, it follows directly that

$$\sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{R}\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{R}'\right). \tag{2.72}$$

This allows a test statistic to be formed

$$W = n\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right)'\left(\mathbf{R}\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{R}'\right)^{-1}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right) \tag{2.73}$$

which is the sum of the squares of m random variables, each asymptotically uncorrelated standard normal and so W is asymptotically χ^2_m distributed. A hypothesis test with size α can be conducted by comparing W against $C_\alpha = F^{-1}(1-\alpha)$ where $F(\cdot)$ is the cdf of a χ^2_m. If $W \geq C_\alpha$ then the null is rejected.

There is one problem with the definition of W in eq. (2.73): it is infeasible since it depends on $\mathbf{G}$ and $\boldsymbol{\Sigma}$ which are unknown. The usual practice is to replace the unknown elements of the covariance matrix with consistent estimates to compute a feasible Wald statistic,

$$W = n\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right)'\left(\mathbf{R}\hat{\mathbf{G}}^{-1}\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{G}}^{-1}\right)'\mathbf{R}'\right)^{-1}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right). \tag{2.74}$$

which has the same asymptotic distribution as the infeasible Wald test statistic.

2.5.4.1 t-tests

A t-test is a special case of a Wald and is applicable to tests involving a single hypothesis. Suppose the null is

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

where $\mathbf{R}$ is 1 by k, and so

$$\sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - r\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{R}').$$

The *studentized* version can be formed by subtracting the mean and dividing by the standard deviation,

$$t = \frac{\sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - r\right)}{\sqrt{\mathbf{R}\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{R}'}} \xrightarrow{d} N(0,1). \tag{2.75}$$

and the test statistic can be compared to the critical values from a standard normal to conduct a hypothesis test. t-tests have an important advantage over the broader class of Wald tests – they can be used to test one-sided null hypotheses. A one-sided hypothesis takes the form $H_0 : \mathbf{R}\boldsymbol{\theta} \geq r$ or $H_0 : \mathbf{R}\boldsymbol{\theta} \leq r$ which are contrasted with one-sided alternatives of $H_1 : \mathbf{R}\boldsymbol{\theta} < r$ or $H_1 : \mathbf{R}\boldsymbol{\theta} > r$, respectively. When using a one-sided test, rejection occurs when $\mathbf{R} - r$ is statistically different from zero and when $\mathbf{R}\boldsymbol{\theta} < r$ or $\mathbf{R}\boldsymbol{\theta} > r$ as *specified by the alternative*.

t-tests are also used in commonly encountered test statistic, the t-stat, a test of the null that a parameter is 0 against an alternative that it is not. The t-stat is popular because most models are written in such a way that if a parameter $\theta = 0$ then it will have no impact.

Definition 2.25 (t-stat)**.** The t-stat of a parameter θ_j is the t-test value of the null $H_0 : \theta_j = 0$ against a two-sided alternative $H_1 : \theta_j \neq 0$.

$$t\text{-stat} \equiv \frac{\hat{\theta}_j}{\sigma_{\hat{\theta}}} \tag{2.76}$$

where

$$\sigma_{\hat{\theta}} = \sqrt{\frac{\mathbf{e}_j\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{e}_j'}{n}} \tag{2.77}$$

and where $\mathbf{e}_j$ is a vector of 0s with 1 in the j$^{\text{th}}$ position.
Note that the t-stat is identical to the expression in eq. (2.75) when $\mathbf{R} = \mathbf{e}_j$ and $r = 0$. $\mathbf{R} = \mathbf{e}_j$ corresponds to a hypothesis test involving only element j of $\boldsymbol{\theta}$ and $r = 0$ indicates that the null is $\theta_j = 0$.

A closely related measure is the standard error of a parameter. Standard errors are essentially standard deviations – square-roots of variance – except that the expression "standard error" is applied when describing the estimation error of a parameter while "standard deviation" is used when describing the variation in the data or population.

Definition 2.26 (Standard Error)**.** The standard error of a parameter θ is the square root of the parameter's variance,

$$\text{s.e.}\left(\hat{\theta}\right) = \sqrt{\sigma_{\hat{\theta}}^2} \tag{2.78}$$

where

$$\sigma_{\hat{\theta}}^2 = \frac{\mathbf{e}_j\mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}^{-1}\right)'\mathbf{e}_j'}{n} \tag{2.79}$$

and where $\mathbf{e}_j$ is a vector of 0s with 1 in the j$^{\text{th}}$ position.

2.5.5 Likelihood Ratio Tests

Likelihood ratio tests examine how "likely" the data are under the null and the alternative. If the hypothesis is valid then the data should be (approximately) equally likely under each. The LR test statistic is defined as

$$LR = -2\left(l\left(\tilde{\boldsymbol{\theta}};\mathbf{y}\right) - l\left(\hat{\boldsymbol{\theta}};\mathbf{y}\right)\right) \tag{2.80}$$

where $\tilde{\boldsymbol{\theta}}$ is defined

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta};\mathbf{y}) \\ &\text{subject to } \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = 0\end{aligned} \tag{2.81}$$

and $\hat{\boldsymbol{\theta}}$ is the unconstrained estimator,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta};\mathbf{y}). \tag{2.82}$$

Under the null $H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = 0$, the $LR \xrightarrow{d} \chi^2_m$. The intuition behind the asymptotic distribution of the LR can be seen in a second order Taylor expansion around parameters estimated under the null, $\tilde{\boldsymbol{\theta}}$.

$$l(\mathbf{y};\tilde{\boldsymbol{\theta}}) = l(\mathbf{y};\hat{\boldsymbol{\theta}}) + \left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right)' \frac{\partial l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + \frac{1}{2}\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right)' \frac{1}{n}\frac{\partial^2 l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) + R^3 \tag{2.83}$$

where R^3 is a remainder term that is vanishing as $n \to \infty$. Since $\hat{\boldsymbol{\theta}}$ is an unconstrained estimator of $\boldsymbol{\theta}_0$,

$$\frac{\partial l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

and

$$-2\left(l(\mathbf{y};\tilde{\boldsymbol{\theta}}) - l(\mathbf{y};\hat{\boldsymbol{\theta}})\right) \approx \sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right)' \left(-\frac{1}{n}\frac{\partial^2 l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) \tag{2.84}$$

Under some mild regularity conditions, when the MLE is correctly specified

$$-\frac{1}{n}\frac{\partial^2 l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \xrightarrow{p} -\mathrm{E}\left[\frac{\partial^2 l(\mathbf{y};\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = \mathcal{I},$$

and

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}).$$

Thus,

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\right)'\frac{1}{n}\frac{\partial^2 l(\mathbf{y};\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\right)\xrightarrow{d}\chi^2_m \tag{2.85}$$

and so $2\left(l(\mathbf{y};\hat{\boldsymbol{\theta}})-l(\mathbf{y};\hat{\boldsymbol{\theta}})\right)\xrightarrow{d}\chi^2_m$. The only difficultly remaining is that the distribution of this quadratic form is a χ^2_m an not a χ^2_k since k is the dimension of the parameter vector. While formally establishing this is tedious, the intuition follows from the number of restrictions. If $\tilde{\boldsymbol{\theta}}$ were unrestricted then it must be the case that $\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}$ since $\hat{\boldsymbol{\theta}}$ is defined as the unrestricted estimators. Applying a single restriction leave $k-1$ free parameters in $\tilde{\boldsymbol{\theta}}$ and thus it should be close to $\hat{\boldsymbol{\theta}}$ except for this one restriction.

When models are correctly specified LR tests are very powerful against point alternatives (e.g. $H_0:\boldsymbol{\theta}=\boldsymbol{\theta}_0$ against $H_1:\boldsymbol{\theta}=\boldsymbol{\theta}_1$). Another important advantage of the LR is that the covariance of the parameters does not need to be estimated. In many problems, accurate parameter covariances may be difficult to estimate, and imprecise covariance estimators produce adverse consequence for test statistics, such as size distortions where a 5% test will reject substantially more than 5% of the time when the null is true.

It is also important to note that the likelihood ratio *does not* have an asymptotic χ^2_m when the assumed likelihood $f(\mathbf{y};\boldsymbol{\theta})$ is misspecified. When this occurs the information matrix equality fails to hold and the asymptotic distribution of the LR is known as a *mixture of* χ^2 *distribution*. In practice, the assumed error distribution is often misspecified and so it is important that the distributional assumptions used to estimate $\boldsymbol{\theta}$ are verified prior to using likelihood ratio tests.

Likelihood ratio tests are not available for method of moments estimators since no distribution function is assumed.[24]

2.5.6 Lagrange Multiplier, Score and Rao Tests

Lagrange Multiplier (LM), Score and Rao test are all the same statistic. While Lagrange Multiplier test may be the most appropriate description, describing the tests as score tests illustrates the simplicity of the test's construction. Score tests exploit the first order condition to test whether

[24] It is possible to construct a likelihood ratio-type statistic for method of moments estimators. Define

$$\mathbf{g}_n(\boldsymbol{\theta})=n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\theta})$$

to be the average moment conditions evaluated at a parameter $\boldsymbol{\theta}$. The likelihood ratio-type statistic for method of moments estimators is defined as

$$\begin{aligned}LM&=n\mathbf{g}_n'\left(\tilde{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{g}_n\left(\tilde{\boldsymbol{\theta}}\right)-n\mathbf{g}_n'\left(\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{g}_n\left(\hat{\boldsymbol{\theta}}\right)\\&=n\mathbf{g}_n'\left(\tilde{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{g}_n\left(\tilde{\boldsymbol{\theta}}\right)\end{aligned}$$

where the simplification is possible since $\mathbf{g}_n\left(\hat{\boldsymbol{\theta}}\right)=0$ and where

$$\hat{\boldsymbol{\Sigma}}=n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)'$$

is the sample covariance of the moment conditions evaluated at the *unrestricted* parameter estimates. This test statistic only differs from the LM test statistic in eq. (2.90) via the choice of the covariance estimator, and it should be similar in performance to the adjusted LM test statistic in eq. (2.92).

a null hypothesis is compatible with the data. Using the unconstrained estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, the scores must be zero,

$$\left.\frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{2.86}$$

The score test examines whether the scores are "close" to zero – in a statistically meaningful way – when evaluated using the parameters estimated subject to the null restriction, $\tilde{\boldsymbol{\theta}}$. Define

$$\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right) = \left.\frac{\partial l_i(\boldsymbol{\theta};y_i)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \tag{2.87}$$

as the i$^{\text{th}}$ score, evaluated at the restricted estimator. If the null hypothesis is true, then

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)\right) \xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma}). \tag{2.88}$$

This forms the basis of the score test, which is computed as

$$LM = n\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right)'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right) \tag{2.89}$$

where $\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right) = n^{-1}\sum_{i=1}^{n}\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)$. While this version is not feasible since it depends on $\boldsymbol{\Sigma}$, the standard practice is to replace $\boldsymbol{\Sigma}$ with a consistent estimator and to compute the feasible score test,

$$LM = n\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right) \tag{2.90}$$

where the estimator of $\boldsymbol{\Sigma}$ depends on the assumptions made about the scores. In the case where the scores are i.i.d. (usually because the data are i.i.d.),

$$\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)' \tag{2.91}$$

is a consistent estimator since $\mathrm{E}\left[\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)\right] = \mathbf{0}$ if the null is true. In practice a more powerful version of the LM test can be formed by subtracting the mean from the covariance estimator and using

$$\tilde{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)-\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right)\right)\left(\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right)-\bar{\mathbf{s}}\left(\tilde{\boldsymbol{\theta}}\right)\right)' \tag{2.92}$$

which must be smaller (in the matrix sense) than $\hat{\boldsymbol{\Sigma}}$, although asymptotically, if the null is true, these two estimators will converge to the same limit. Like the Wald and the LR, the LM follows an asymptotic χ^2_m distribution, and an LM test statistic will be rejected if $LM > C_\alpha$ where C_α is the $1-\alpha$ quantile of a χ^2_m distribution.

Scores test can be used with method of moments estimators by simply replacing the score of the likelihood with the moment conditions evaluated at the restricted parameter,

$$\mathbf{s}_i\left(\tilde{\boldsymbol{\theta}}\right) = \mathbf{g}_i\left(\tilde{\boldsymbol{\theta}}\right),$$

and then evaluating eq. (2.90) or (2.92).

2.5.7 Comparing and Choosing the Tests

All three of the classic tests, the Wald, likelihood ratio and Lagrange multiplier have the same limiting asymptotic distribution. In addition to all being asymptotically distributed as a χ^2_m, they are all *asymptotically equivalent* in the sense they all have an identical asymptotic distribution and if one test rejects, the others will also reject. As a result, there is no asymptotic argument that one should be favored over the other.

The simplest justifications for choosing one over the others are practical considerations. Wald requires estimation *under the alternative* – the unrestricted model – and require an estimate of the asymptotic covariance of the parameters. LM tests require estimation *under the null* – the restricted model – and require an estimate of the asymptotic covariance of the scores evaluated at the restricted parameters. LR tests require both forms to be estimated but do not require any covariance estimates. On the other hand, Wald and LM tests can easily be made robust to many forms of misspecification by using the "sandwich" covariance estimator, $\mathbf{G}^{-1}\mathbf{\Sigma}\left(\mathbf{G}^{-1}\right)'$ for moment-based estimators or $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$ for QML estimators. LR tests cannot be easily corrected and instead will have a non-standard distribution.

Models which are substantially easier to estimate under the null or alternative lead to a natural choice. If a model is easy to estimate in its restricted form, but not unrestricted LM tests are good choices. If estimation under the alternative is simpler then Wald tests are reasonable. If they are equally simple to estimate, and the distributional assumptions used in ML estimation are plausible, LR tests are likely the best choice. Empirically a relationship exists where $W \approx LR \geq LM$. LM is often smaller, and hence less likely to reject the null, since it estimates the covariance of the scores *under the null*. When the null may be restrictive, the scores will generally have higher variances when evaluated using the restricted parameters. The larger variances will lower the value of LM since the score covariance is inverted in the statistic. A simple method to correct this is to use the adjusted LM computed using the modified covariance estimator in eq. (2.92).

2.6 The Bootstrap and Monte Carlo

The bootstrap is an alternative technique for estimating parameter covariances and conducting inference. The name bootstrap is derived from the expression "to pick yourself up by your bootstraps" – a seemingly impossible task. The bootstrap, when initially proposed, was treated as an equally impossible feat, although it is now widely accepted as a valid, and in some cases, preferred method to plug-in type covariance estimation. The bootstrap is a simulation technique and is similar to Monte Carlo. However, unlike Monte Carlo, which requires a complete data-generating process, the bootstrap makes use of the observed data to simulate the data – hence the similarity to the original turn-of-phrase.

Monte Carlo is an integration technique that uses simulation to approximate the underlying distribution of the data. Suppose $Y_i \overset{\text{i.i.d.}}{\sim} F(\boldsymbol{\theta})$ where F is some distribution, and that interest is in the $\mathrm{E}[g(Y)]$. Further suppose it is possible to simulate from $F(\boldsymbol{\theta})$ so that a sample $\{y_i\}$ can be constructed. Then

$$n^{-1}\sum_{i=1}^{n} g(y_i) \xrightarrow{p} \mathrm{E}[g(Y)]$$

as long as this expectation exists since the simulated data are i.i.d. by construction.

The observed data can be used to compute the empirical cdf.

Definition 2.27 (Empirical cdf)**.** The empirical cdf is defined

$$\hat{F}(c) = n^{-1} \sum_{i=1}^{n} I_{[y_i<c]}.$$

As long as $\hat{F}$ is close to F, then the empirical cdf can be used to simulate random variables which should be approximately distributed F, and simulated data from the empirical cdf should have similar statistical properties (mean, variance, etc.) as data simulated from the true population cdf. The empirical cdf is a coarse step function and so *only* values which have been observed can be simulated, and so simulating from the empirical cdf of the data is identical to re-sampling the original data. In other words, the observed data can be directly used to simulate the from the underlying (unknown) cdf.

Figure 2.6 shows the population cdf for a standard normal and two empirical cdfs, one estimated using $n = 20$ observations and the other using $n = 1,000$. The coarse empirical cdf highlights the stair-like features of the empirical cdf estimate which restrict random numbers generated using the empirical cdf to coincide with the data used to compute the empirical cdf.

The bootstrap can be used for a variety of purposes. The most application of a bootstrap is to estimate the covariance matrix of some estimated parameters. This is an alternative to the usual plug-in type estimator and is simple to implement when the estimator is available in closed form.

Algorithm 2.1 (i.i.d. Nonparametric Bootstrap Covariance)**.**

1. *Generate a set of n uniform integers $\{j_i\}_{i=1}^{n}$ on $[1, 2, \ldots, n]$.*
2. *Construct a simulated sample $\{y_{j_i}\}$.*
3. *Estimate the parameters of interest using $\{y_{j_i}\}$, and denote the estimate $\tilde{\boldsymbol{\theta}}_b$.*
4. *Repeat steps 1 through 3 a total of B times.*
5. *Estimate the variance of $\hat{\boldsymbol{\theta}}$ using*

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1} \sum_{b=1}^{B} \left(\tilde{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}\right)'.$$

or alternatively

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1} \sum_{b=1}^{B} \left(\tilde{\boldsymbol{\theta}}_j - \overline{\tilde{\boldsymbol{\theta}}}\right)\left(\tilde{\boldsymbol{\theta}}_j - \overline{\tilde{\boldsymbol{\theta}}}\right)'.$$

The variance estimator that comes from this algorithm cannot be directly compared to the asymptotic covariance estimator since the bootstrap covariance is converging to 0. Normalizing the bootstrap covariance estimate by $\sqrt{n}$ will allow comparisons and direct application of the test statistics based on the asymptotic covariance. Note that when using a conditional model,

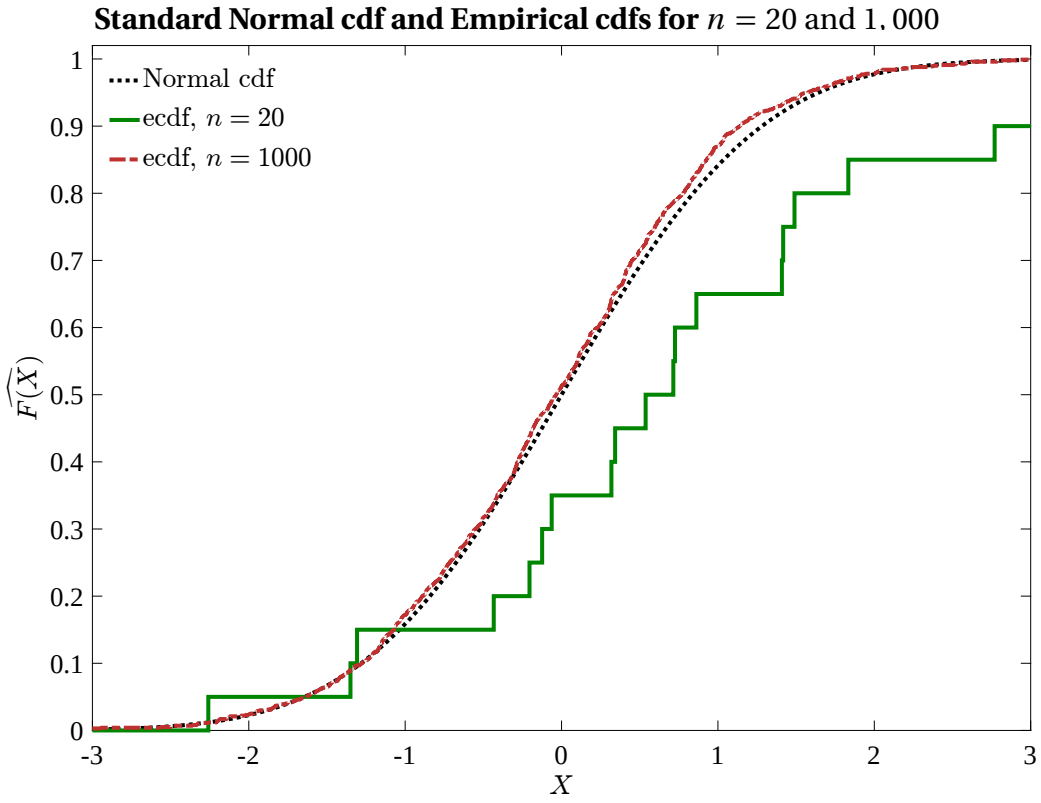

Figure 2.6: These three lines represent the population cdf of a standard normal, and two empirical cdfs constructed form simulated data. The very coarse empirical cdf is based on 20 observations and clearly highlights the step-nature of empirical cdfs. The other empirical cdf, which is based on 1,000 observations, appear smoother but is still a step function.

---

the vector $[y_i \ \mathbf{x}_i']'$ should be jointly bootstrapped. Aside from this small modification to step 2, the remainder of the procedure remains valid.

The nonparametric bootstrap is closely related to the residual bootstrap, at least when it is possible to appropriately define a residual. For example, when $Y_i|\mathbf{X}_i \sim N\left(\boldsymbol{\beta}'\mathbf{x}_i, \sigma^2\right)$, the residual can be defined $\hat{\epsilon}_i = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$. Alternatively if $Y_i|\mathbf{X}_i \sim \text{Scaled} - \chi^2_\nu\left(\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)\right)$, then $\hat{\epsilon}_i = y_i/\sqrt{\hat{\boldsymbol{\beta}}'\mathbf{x}}$. The residual bootstrap can be used whenever it is possible to express $y_i = g\left(\boldsymbol{\theta}, \epsilon_i, \mathbf{x}_i\right)$ for some known function g.

Algorithm 2.2 (i.i.d. Residual Bootstrap Covariance)**.**

1. *Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \ldots, n]$.*

2. *Construct a simulated sample $\{\hat{\epsilon}_{j_i}, \mathbf{x}_{j_i}\}$ and define $\tilde{y}_i = g\left(\hat{\boldsymbol{\theta}}, \tilde{\epsilon}_i, \tilde{\mathbf{x}}_i\right)$ where $\tilde{\epsilon}_i = \hat{\epsilon}_{j_i}$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_{j_i}$.*[25]

[25] In some models, it is possible to use independent indices on $\hat{\epsilon}$ and $\mathbf{x}$, such as in a linear regression when the data are conditionally homoskedastic (See chapter 3). In general it is not possible to explicitly break the link between ϵ_i and $\mathbf{x}_i$, and so these should usually be resampled using the same indices.

3. *Estimate the parameters of interest using* $\{\tilde{y}_i, \tilde{\mathbf{x}}_i\}$, *and denote the estimate* $\tilde{\boldsymbol{\theta}}_b$.
4. *Repeat steps 1 through 3 a total of* B *times.*
5. *Estimate the variance of* $\hat{\boldsymbol{\theta}}$ *using*

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)'.$$

or alternatively

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \bar{\tilde{\boldsymbol{\theta}}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \bar{\tilde{\boldsymbol{\theta}}}\right)'.$$

It is important to emphasize that the bootstrap is not, generally, a better estimator of parameter covariance than standard plug-in estimators.[26] Asymptotically both are consistent and can be used equivalently. Additionally, i.i.d. bootstraps can only be applied to (conditionally) i.i.d. data and using an inappropriate bootstrap will produce an inconsistent estimator. When data have dependence it is necessary to use an alternative bootstrap scheme.

When the interest lies in confidence intervals, an alternative procedure that directly uses the empirical quantiles of the bootstrap parameter estimates can be constructed (known as the percentile method).

Algorithm 2.3 (i.i.d. Nonparametric Bootstrap Confidence Interval)**.**

1. *Generate a set of* n *uniform integers* $\{j_i\}_{i=1}^{n}$ *on* $[1, 2, \ldots, n]$.
2. *Construct a simulated sample* $\{y_{j_i}\}$.
3. *Estimate the parameters of interest using* $\{y_{j_i}\}$, *and denote the estimate* $\tilde{\boldsymbol{\theta}}_b$.
4. *Repeat steps 1 through 3 a total of* B *times.*
5. *Estimate the* $1-\alpha$ *confidence interval of* $\hat{\theta}_k$ *using*

$$\left[q_{\alpha/2}\left(\left\{\tilde{\theta}_k\right\}\right), q_{1-\alpha/2}\left(\left\{\tilde{\theta}_k\right\}\right)\right]$$

where $q_\alpha\left(\left\{\tilde{\theta}_k\right\}\right)$ *is the empirical* α *quantile of the bootstrap estimates. 1-sided lower confidence intervals can be constructed as*

$$\left[\underline{R\left(\theta_k\right)}, q_{1-\alpha}\left(\left\{\tilde{\theta}_k\right\}\right)\right]$$

and 1-sided upper confidence intervals can be constructed as

$$\left[q_\alpha\left(\left\{\tilde{\theta}_k\right\}\right), \overline{R\left(\theta_k\right)}\right]$$

where $\underline{R\left(\theta_k\right)}$ *and* $\overline{R\left(\theta_k\right)}$ *are the lower and upper extremes of the range of* θ_k *(possibly* $\pm\infty$*).*

[26] There are some problem-dependent bootstraps that are more accurate than plug-in estimators in an asymptotic sense. These are rarely encountered in financial economic applications.

The percentile method can also be used directly to compute P-values of test statistics. This requires enforcing the null hypothesis on the data and so is somewhat more involved. For example, suppose the null hypothesis is $\mathrm{E}[y_i] = 0$. This can be enforced by replacing the original data with $\tilde{y}_i = y_i - \bar{y}$ in step 2 of the algorithm.

Algorithm 2.4 (i.i.d. Nonparametric Bootstrap P-value)**.**

1. *Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \ldots, n]$.*
2. *Construct a simulated sample using data where the null hypothesis is true, $\{\tilde{y}_{j_i}\}$.*
3. *Compute the test statistic of interest using $\{\tilde{y}_{j_i}\}$, and denote the statistic $T\left(\tilde{\boldsymbol{\theta}}_b\right)$.*
4. *Repeat steps 1 through 3 a total of B times.*
5. *Compute the bootstrap P-value using*

$$\widehat{P-val} = B^{-1}\sum_{b=1}^{B} I_{[T(\hat{\boldsymbol{\theta}})\leq T(\tilde{\boldsymbol{\theta}})]}$$

for 1-sided tests where the rejection region is for large values (e.g. a Wald test). When using 2-sided tests, compute the bootstrap P-value using

$$\widehat{P-val} = B^{-1}\sum_{b=1}^{B} I_{[|T(\hat{\boldsymbol{\theta}})|\leq|T(\tilde{\boldsymbol{\theta}})|]}$$

The test statistic may depend on a covariance matrix. When this is the case, the covariance matrix is usually estimated from the bootstrapped data using a plug-in method. Alternatively, it is possible to use any other consistent estimator (when the null is true) of the asymptotic covariance, such as one based on an initial (separate) bootstrap.

When models are maximum likelihood based, so that a complete model for the data is specified, it is possible to use a parametric form of the bootstrap to estimate covariance matrices. This procedure is virtually identical to standard Monte Carlo except that the initial estimate $\hat{\boldsymbol{\theta}}$ is used in the simulation.

Algorithm 2.5 (i.i.d. Parametric Bootstrap Covariance (Monte Carlo))**.**

1. *Simulate a set of n i.i.d. draws $\{\tilde{y}_i\}$ from $F\left(\hat{\boldsymbol{\theta}}\right)$.*
2. *Estimate the parameters of interest using $\{\tilde{y}_i\}$, and denote the estimates $\tilde{\boldsymbol{\theta}}_b$.*
3. *Repeat steps 1 through 4 a total of B times.*
4. *Estimate the variance of $\hat{\boldsymbol{\theta}}$ using*

$$\mathrm{V}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)'.$$

or alternatively

$$\mathrm{V}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \bar{\tilde{\boldsymbol{\theta}}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \bar{\tilde{\boldsymbol{\theta}}}\right)'.$$

When models use conditional maximum likelihood, it is possible to use parametric bootstrap as part of a two-step procedure. First, apply a nonparametric bootstrap to the conditioning data$\{\mathbf{x}_i\}$, and then, using the bootstrapped conditioning data, simulate $Y_i \sim F\left(\hat{\boldsymbol{\theta}}|\tilde{\mathbf{X}}_i\right)$. This is closely related to the residual bootstrap, only the assumed parametric distribution F is used in place of the data-derived residuals.

2.7 Inference on Financial Data

Inference will be covered in greater detail in conjunction with specific estimators and models, such as linear regression or ARCH models. These examples examine relatively simple hypotheses to illustrate the steps required in testing hypotheses.

2.7.1 Testing the Market Premium

Testing the market premium is a cottage industry. While current research is more interested in predicting the market premium, testing whether the market premium is significantly different from zero is a natural application of the tools introduced in this chapter. Let λ denote the market premium and let σ^2 be the variance of the return. Since the market is a traded asset it must be the case that the premium for holding market risk is the same as the mean of the market return. Monthly data for the Value Weighted Market (VWM) and the risk-free rate (RF) was available between January 1927 and June 2008. Data for the VWM was drawn from CRSP and data for the risk-free rate was available from Ken French's data library. Excess returns on the market are defined as the return to holding the market minus the risk-free rate, $VWM_i^e = VWM_i - RF_i$. The excess returns along with a kernel density plot are presented in figure 2.7. Excess returns are both negatively skewed and heavy-tailed – October 1987 is 5 standard deviations from the mean.

The mean and variance can be computed using the method of moments as detailed in section 2.1.4, and the covariance of the mean and the variance can be computed using the estimators described in section 2.4.1. The estimates were calculated according to

$$\begin{bmatrix} \hat{\lambda} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} n^{-1}\sum_{i=1}^n VWM_i^e \\ n^{-1}\sum_{i=1}^n \left(VWM_i^e - \hat{\lambda}\right)^2 \end{bmatrix}$$

and, defining $\hat{\epsilon}_i = VWM_i^e - \hat{\lambda}$, the covariance of the moment conditions was estimated by

$$\hat{\boldsymbol{\Sigma}} = n^{-1}\begin{bmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 & \sum_{i=1}^n \hat{\epsilon}_i\left(\hat{\epsilon}_i^2 - \hat{\sigma}^2\right) \\ \sum_{i=1}^n \hat{\epsilon}_i\left(\hat{\epsilon}_i^2 - \hat{\sigma}^2\right) & \sum_{i=1}^n \left(\hat{\epsilon}_i^2 - \hat{\sigma}^2\right)^2 \end{bmatrix}.$$

Since the plim of the Jacobian is $-\mathbf{I}_2$, the parameter covariance is also $\hat{\boldsymbol{\Sigma}}$. Combining these two results with a Central Limit Theorem (assumed to hold), the asymptotic distribution is

$$\sqrt{n}\left[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right] \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\theta} = \left(\lambda, \sigma^2\right)'$. These produce the results in the first two rows of table 2.3.

These estimates can also be used to make inference on the standard deviation, $\sigma = \sqrt{\sigma^2}$ and the Sharpe ratio, $S = \lambda/\sigma$. The derivation of the asymptotic distribution of the Sharpe ratio was

presented in 2.4.4.1 and the asymptotic distribution of the standard deviation can be determined in a similar manner where $\mathbf{d}(\boldsymbol{\theta}) = \sqrt{\sigma^2}$ and so

$$\mathbf{D}(\boldsymbol{\theta}) = \frac{\partial \mathbf{d}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} 0 & \frac{1}{2\sqrt{\sigma^2}} \end{bmatrix}.$$

Combining this expression with the asymptotic distribution for the estimated mean and variance, the asymptotic distribution of the standard deviation estimate is

$$\sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{d} N\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right).$$

which was computed by dividing the [2,2] element of the parameter covariance by $4\hat{\sigma}^2$.

2.7.1.1 Bootstrap Implementation

The bootstrap can be used to estimate parameter covariance, construct confidence intervals – either used the estimated covariance or the percentile method, and to tabulate the P-value of a test statistic. Estimating the parameter covariance is simple – the data is resampled to create a simulated sample with n observations and the mean and variance are estimated. This is repeated 10,000 times and the parameter covariance is estimated using

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= B^{-1}\sum_{b=1}^{B}\left(\begin{bmatrix}\tilde{\mu}_b \\ \tilde{\sigma}_b^2\end{bmatrix} - \begin{bmatrix}\hat{\mu}_b \\ \hat{\mu}_b^2\end{bmatrix}\right)\left(\begin{bmatrix}\tilde{\mu}_b \\ \tilde{\sigma}_b^2\end{bmatrix} - \begin{bmatrix}\hat{\mu}_b \\ \hat{\mu}_b^2\end{bmatrix}\right)' \\ &= B^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)'. \end{aligned}$$

The percentile method can be used to construct confidence intervals for the parameters as estimated and for functions of parameters such as the Sharpe ratio. Constructing the confidence intervals for a function of the parameters requires constructing the function of the estimated parameters using each simulated sample and then computing the confidence interval using the empirical quantile of these estimates. Finally, the test P-value for the statistic for the null H_0 : $\lambda = 0$ can be computed directly by transforming the returns so that they have mean 0 using $\tilde{r}_i = r_i - \bar{r}_i$. The P-value can be tabulated using

$$\widehat{\text{P}-\text{val}} = B^{-1}\sum_{b=1}^{B} I_{[\bar{r}\leq\bar{\bar{r}}_b]}$$

where $\bar{\bar{r}}_b$ is the average from bootstrap replication b. Table 2.4 contains the bootstrap standard errors, confidence intervals based on the percentile method and the bootstrap P-value for testing whether the mean return is 0. The standard errors are virtually identical to those estimated using the plug-in method, and the confidence intervals are similar to $\hat{\theta}_k \pm 1.96\text{s.e.}(\theta_k)$. The null that the average return is 0 is also strongly rejected.

Parameter	Estimate	Standard Error	t-stat
λ	0.627	0.173	3.613
σ^2	29.41	2.957	9.946
σ	5.423	0.545	9.946
$\frac{\lambda}{\sigma}$	0.116	0.032	3.600

Table 2.3: Parameter estimates and standard errors for the market premium (λ), the variance of the excess return (σ^2), the standard deviation of the excess return (σ) and the Sharpe ratio ($\frac{\lambda}{\sigma}$). Estimates and variances were computed using the method of moments. The standard errors for σ and $\frac{\lambda}{\sigma}$ were computed using the delta method.

		Bootstrap	Confidence Interval	
Parameter	Estimate	Standard Error	Lower	Upper
λ	0.627	0.174	0.284	0.961
σ^2	29.41	2.964	24.04	35.70
σ	5.423	0.547	4.903	5.975
$\frac{\lambda}{\sigma}$	0.116	0.032	0.052	0.179
$H_0 : \lambda = 0$				
P-value	3.00×10^{-4}			

Table 2.4: Parameter estimates, bootstrap standard errors and confidence intervals (based on the percentile method) for the market premium (λ), the variance of the excess return (σ^2), the standard deviation of the excess return (σ) and the Sharpe ratio ($\frac{\lambda}{\sigma}$). Estimates were computed using the method of moments. The standard errors for σ and $\frac{\lambda}{\sigma}$ were computed using the delta method using the bootstrap covariance estimator.

2.7.2 Is the NASDAQ Riskier than the S&P 100?

A second application examines the riskiness of the NASDAQ and the S&P 100. Both of these indices are value-weighted and contain 100 companies. The NASDAQ 100 contains only companies that trade on the NASDAQ while the S&P 100 contains large companies that trade on either the NYSE or the NASDAQ.

The null hypothesis is that the variances are the same, $H_0 : \sigma^2_{SP} = \sigma^2_{ND}$, and the alternative is that the variance of the NASDAQ is larger, $H_1 : \sigma^2_{ND} > \sigma^2_{SP}$.[27] The null and alternative can be reformulated as a test that $\delta = \sigma^2_{ND} - \sigma^2_{SP}$ is equal to zero against an alternative that it is greater than zero. The estimation of the parameters can be formulated as a method of moments problem,

[27] It may also be interesting to test against a two-sided alternative that the variances are unequal, $H_1 : \sigma^2_{ND} \neq \sigma^2_{SP}$.

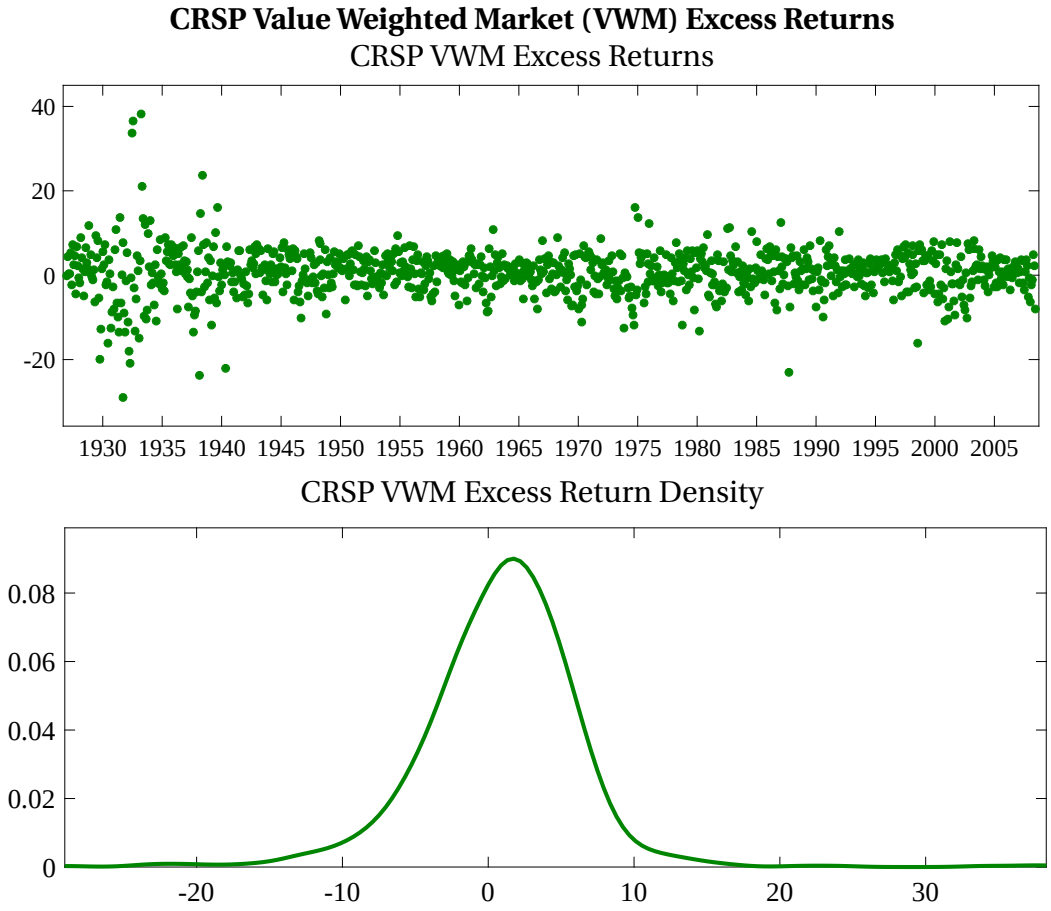

Figure 2.7: These two plots contain the returns on the VWM (top panel) in excess of the risk free rate and a kernel estimate of the density (bottom panel). While the mode of the density (highest peak) appears to be clearly positive, excess returns exhibit strong negative skew and are heavy tailed.

$$\begin{bmatrix} \hat{\mu}_{SP} \\ \hat{\sigma}^2_{SP} \\ \hat{\mu}_{ND} \\ \hat{\sigma}^2_{ND} \end{bmatrix} = n^{-1} \sum_{i=1}^{n} \begin{bmatrix} r_{SP,i} \\ (r_{SP,i} - \hat{\mu}_{SP})^2 \\ r_{ND,i} \\ (r_{ND,i} - \hat{\mu}_{ND})^2 \end{bmatrix}$$

Inference can be performed by forming the moment vector using the estimated parameters, $\mathbf{g}_i$,

$$\mathbf{g}_i = \begin{bmatrix} r_{SP,i} - \mu_{SP} \\ (r_{SP,i} - \mu_{SP})^2 - \sigma^2_{SP} \\ r_{ND,i} - \mu_{ND} \\ (r_{ND,i} - \mu_{ND})^2 - \sigma^2_{ND} \end{bmatrix}$$

and recalling that the asymptotic distribution is given by

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(0, \mathbf{G}^{-1}\boldsymbol{\Sigma}\left(\mathbf{G}'\right)^{-1}\right).$$

Daily Data

Parameter	Estimate	Std. Error/Correlation			
μ_{SP}	9.06	3.462	-0.274	0.767	-0.093
σ_{SP}	17.32	-0.274	0.709	-0.135	0.528
μ_{ND}	9.73	0.767	-0.135	4.246	-0.074
σ_{NS}	21.24	-0.093	0.528	-0.074	0.443

Test Statistics

δ	0.60	$\hat{\sigma}_\delta$	0.09	t-stat	6.98

Monthly Data

Parameter	Estimate	Std. Error/Correlation			
μ_{SP}	8.61	3.022	-0.387	0.825	-0.410
σ_{SP}	15.11	-0.387	1.029	-0.387	0.773
μ_{ND}	9.06	0.825	-0.387	4.608	-0.418
σ_{NS}	23.04	-0.410	0.773	-0.418	1.527

Test Statistics

δ	25.22	$\hat{\sigma}_\delta$	4.20	t-stat	6.01

Table 2.5: Estimates, standard errors and correlation matrices for the S&P 100 and NASDAQ 100. The top panel uses daily return data between January 3, 1983, and December 31, 2007 (6,307 days) to estimate the parameter values in the left-most column. The rightmost 4 columns contain the parameter standard errors (diagonal elements) and the parameter correlations (off-diagonal elements). The bottom panel contains estimates, standard errors, and correlations from monthly data between January 1983 and December 2007 (300 months). Parameter and covariance estimates have been annualized. The test statistics (and related quantities) were performed and reported on the original (non-annualized) values.

Using the set of moment conditions,

$$\begin{aligned}\mathbf{G} &= \text{plim}_{n\to\infty}\, n^{-1}\sum_{i=1}^{n}\begin{bmatrix} -1 & 0 & 0 & 0 \\ -2\left(r_{SP,i}-\mu_{SP}\right) & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -2\left(r_{ND,i}-\mu_{ND}\right) & -1 \end{bmatrix} \\ &= -\mathbf{I}_4.\end{aligned}$$

$\mathbf{\Sigma}$ can be estimated using the moment conditions evaluated at the estimated parameters, $\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)$,

$$\hat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\hat{\boldsymbol{\theta}}\right)\mathbf{g}_i'\left(\hat{\boldsymbol{\theta}}\right).$$

Noting that the (2,2) element of $\mathbf{\Sigma}$ is the variance of $\hat{\sigma}^2_{SP}$, the (4,4) element of $\mathbf{\Sigma}$ is the variance of $\hat{\sigma}^2_{ND}$ and the (2,4) element is the covariance of the two, the variance of $\hat{\delta} = \hat{\sigma}^2_{ND} - \hat{\sigma}^2_{SP}$ can be computed as the sum of the variances minus two times the covariance, $\Sigma_{[2,2]} + \Sigma_{[4,4]} - 2\Sigma_{[2,4]}$. Finally a *one-sided t*-test can be performed to test the null.

Data was taken from Yahoo! finance between January 1983 and December 2008 at both the daily and monthly frequencies. Parameter estimates are presented in table 2.5. The table also contains the parameter standard errors – the square-root of the asymptotic covariance divided by the number of observations ($\sqrt{\mathbf{\Sigma}_{[i,i]}/n}$) – along the diagonal and the parameter correlations – $\mathbf{\Sigma}_{[i,j]}/\sqrt{\mathbf{\Sigma}_{[i,i]}\mathbf{\Sigma}_{[j,j]}}$ – in the off-diagonal positions. The top panel contains results for daily data while the bottom contains results for monthly data. Returns scaled by 100 were used in both panels .

All parameter estimates are reported in annualized form, which requires multiplying daily (monthly) mean estimates by 252 (12), and daily (monthly) volatility estimated by $\sqrt{252}$ $\left(\sqrt{12}\right)$. Additionally, the delta method was used to adjust the standard errors on the volatility estimates since the actual parameter estimates were the means and variances. Thus, the reported parameter variance covariance matrix has the form

$$\mathbf{D}\left(\hat{\boldsymbol{\theta}}\right)\hat{\mathbf{\Sigma}}\mathbf{D}\left(\hat{\boldsymbol{\theta}}\right) = \begin{bmatrix} 252 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{252}}{2\sigma_{SP}} & 0 & 0 \\ 0 & 0 & 252 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{252}}{2\sigma_{ND}} \end{bmatrix} \hat{\mathbf{\Sigma}} \begin{bmatrix} 252 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{252}}{2\sigma_{SP}} & 0 & 0 \\ 0 & 0 & 252 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{252}}{2\sigma_{ND}} \end{bmatrix}.$$

In both cases δ is positive with a t-stat greater than 6, indicating a strong rejection of the null in favor of the alternative. Since this was a one-sided test, the 95% critical value would be 1.645 ($\Phi(.95)$).

This test could also have been implemented using an LM test, which requires estimating the two mean parameters but restricting the variances to be equal. One $\tilde{\boldsymbol{\theta}}$ is estimated, the LM test statistic is computed as

$$LM = n\mathbf{g}_n\left(\tilde{\boldsymbol{\theta}}\right)\hat{\mathbf{\Sigma}}^{-1}\mathbf{g}_n'\left(\tilde{\boldsymbol{\theta}}\right)$$

where

$$\mathbf{g}_n\left(\tilde{\boldsymbol{\theta}}\right) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i\left(\tilde{\boldsymbol{\theta}}\right)$$

and where $\tilde{\mu}_{SP} = \hat{\mu}_{SP}$, $\tilde{\mu}_{ND} = \hat{\mu}_{ND}$ (unchanged) and $\tilde{\sigma}^2_{SP} = \tilde{\sigma}^2_{ND} = \left(\hat{\sigma}^2_{SP} + \hat{\sigma}^2_{ND}\right)/2$.

2.7.2.1 Bootstrap Covariance Estimation

The bootstrap is an alternative to the plug-in covariance estimators. The bootstrap was implemented using 10,000 resamples where the data were assumed to be i.i.d.. In each bootstrap resample, the full 4 by 1 vector of parameters was computed. These were combined to estimate the parameter covariance using

$$\hat{\mathbf{\Sigma}} = B^{-1}\sum_{i=1}^{B}\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_b - \hat{\boldsymbol{\theta}}\right)'.$$

Daily Data					
Parameter	Estimate	BootStrap Std. Error/Correlation			
μ_{SP}	9.06	3.471	-0.276	0.767	-0.097
σ_{SP}	17.32	-0.276	0.705	-0.139	0.528
μ_{ND}	9.73	0.767	-0.139	4.244	-0.079
σ_{NS}	21.24	-0.097	0.528	-0.079	0.441
Monthly Data					
Parameter	Estimate	Bootstrap Std. Error/Correlation			
μ_{SP}	8.61	3.040	-0.386	0.833	-0.417
σ_{SP}	15.11	-0.386	1.024	-0.389	0.769
μ_{ND}	9.06	0.833	-0.389	4.604	-0.431
σ_{NS}	23.04	-0.417	0.769	-0.431	1.513

Table 2.6: Estimates and bootstrap standard errors and correlation matrices for the S&P 100 and NASDAQ 100. The top panel uses daily return data between January 3, 1983, and December 31, 2007 (6,307 days) to estimate the parameter values in the left-most column. The rightmost 4 columns contain the bootstrap standard errors (diagonal elements) and the correlations (off-diagonal elements). The bottom panel contains estimates, bootstrap standard errors and correlations from monthly data between January 1983 and December 2007 (300 months). All parameter and covariance estimates have been annualized.

Table 2.6 contains the bootstrap standard errors and correlations. Like the results in 2.5, the parameter estimates and covariance have been annualized, and volatility rather than variance is reported. The covariance estimates are virtually indistinguishable to those computed using the plug-in estimator. This highlights that the bootstrap is not (generally) a better estimator, but is merely an alternative.[28]

2.7.3 Testing Factor Exposure

Suppose excess returns were conditionally normal with mean $\mu_i = \boldsymbol{\beta}'\mathbf{x}_i$ and constant variance σ^2. This type of model is commonly used to explain cross-sectional variation in returns, and when the conditioning variables include only the market variable, the model is known as the Capital Asset Pricing Model (CAP-M, Sharpe (1964) and Lintner (1965)). Multi-factor models allow for additional conditioning variables such as the size and value factors (Ross, 1976; Fama and French, 1992; Fama and French, 1993). The size factor is the return on a portfolio which is long small cap stocks and short large cap stocks. The value factor is the return on a portfolio that is long high book-to-market stocks (value) and short low book-to-market stocks (growth).

This example estimates a 3 factor model where the conditional mean of excess returns on individual assets is modeled as a linear function of the excess return to the market, the size factor

[28] In this particular application, as the bootstrap and the plug-in estimators are identical as $B \to \infty$ for fixed n. This is not generally the case.

and the value factor. This leads to a model of the form

$$\begin{aligned} r_i - r_i^f &= \beta_0 + \beta_1 \left(r_{m,i} - r_i^f \right) + \beta_2 r_{s,i} + \beta_3 r_{v,i} + \epsilon_i \\ r_i^e &= \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i \end{aligned}$$

where r_i^f is the risk-free rate (short term government rate), $r_{m,i}$ is the return to the market portfolio, $r_{s,i}$ is the return to the size portfolio and $r_{v,i}$ is the return to the value portfolio. ϵ_i is a residual which is assumed to have a $N\left(0, \sigma^2\right)$ distribution.

Factor models can be formulated as a conditional maximum likelihood problem,

$$l\left(\mathbf{r}|\mathbf{X}; \boldsymbol{\theta}\right) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \ln\left(2\pi\right) + \ln\left(\sigma^2\right) + \frac{\left(r_i - \boldsymbol{\beta}'\mathbf{x}_i\right)^2}{\sigma^2} \right\}$$

where $\boldsymbol{\theta} = \left[\boldsymbol{\beta}' \, \sigma^2\right]'$. The MLE can be found using the first order conditions, which are

$$\begin{aligned} \frac{\partial l\left(r; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\beta}} &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} \mathbf{x}_i \left(r_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i \right) = 0 \\ \Rightarrow \hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{j=1}^{n} \mathbf{x}_i r_i \\ \frac{\partial l\left(r; \boldsymbol{\theta}\right)}{\partial \sigma^2} &= -\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2} - \frac{\left(r_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i \right)^2}{\hat{\sigma}^4} = 0 \\ \Rightarrow \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^{n} \left(r_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i \right)^2 \end{aligned}$$

The vector of scores is

$$\frac{\partial l\left(r_i|\mathbf{x}_i; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{x}_i \epsilon_i \\ -\frac{1}{2\sigma^2} + \frac{\epsilon_i^2}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \epsilon_i \\ \sigma^2 - \epsilon_i^2 \end{bmatrix} = \mathbf{S} \begin{bmatrix} \mathbf{x}_i \epsilon_i \\ \sigma^2 - \epsilon_i^2 \end{bmatrix}$$

where $\epsilon_i = r_i - \boldsymbol{\beta}'\mathbf{x}_i$. The second form will be used to simplify estimating the parameters covariance. The Hessian is

$$\frac{\partial^2 l\left(r_i|\mathbf{x}_i; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{x}_i \mathbf{x}_i' & -\frac{1}{\sigma^4}\mathbf{x}_i \epsilon_i \\ -\frac{1}{\sigma^4}\mathbf{x}_i \epsilon_i & \frac{1}{2\sigma^4} - \frac{\epsilon_i^2}{\sigma^6} \end{bmatrix},$$

and the information matrix is

$$\begin{aligned} \mathcal{I} &= -\mathrm{E}\begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{x}_i \mathbf{x}_i' & -\frac{1}{\sigma^4}\mathbf{x}_i \epsilon_i \\ -\frac{1}{\sigma^4}\mathbf{x}_i \epsilon_i & \frac{1}{2\sigma^4} - \frac{\epsilon_i^2}{\sigma^6} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2}\mathrm{E}\left[\mathbf{x}_i \mathbf{x}_i'\right] & -\frac{1}{\sigma^4}\mathrm{E}\left[\mathbf{x}_i \mathrm{E}\left[\epsilon_i|\mathbf{X}\right]\right] \\ -\frac{1}{\sigma^4}\mathrm{E}\left[\mathbf{x}_i \mathrm{E}\left[\epsilon_i|\mathbf{X}\right]\right] & \mathrm{E}\left[\frac{1}{2\sigma^4}\right] \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2}\mathrm{E}\left[\mathbf{x}_i \mathbf{x}_i'\right] & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}. \end{aligned}$$

The covariance of the scores is

$$\begin{aligned}
\mathcal{J} &= \mathrm{E}\left[\mathbf{S}\begin{bmatrix} \epsilon_i^2\mathbf{x}_i\mathbf{x}_i' & \sigma^2\mathbf{x}_i\epsilon_i - \mathbf{x}_i\epsilon_i^3 \\ \sigma^2\mathbf{x}_i'\epsilon_i - \mathbf{x}_i'\epsilon_i^3 & \left(\sigma^2-\epsilon_i^2\right)^2 \end{bmatrix}\mathbf{S}\right] \\
&= \mathbf{S}\begin{bmatrix} \mathrm{E}\left[\epsilon_i^2\mathbf{x}_i\mathbf{x}_i'\right] & \mathrm{E}\left[\sigma^2\mathbf{x}_i\epsilon_i - \mathbf{x}_i\epsilon_i^3\right] \\ \mathrm{E}\left[\sigma^2\mathbf{x}_i'\epsilon_i - \mathbf{x}_i'\epsilon_i^3\right] & \mathrm{E}\left[\left(\sigma^2-\epsilon_i^2\right)^2\right] \end{bmatrix}\mathbf{S} \\
&= \mathbf{S}\begin{bmatrix} \mathrm{E}\left[\mathrm{E}\left[\epsilon_i^2|\mathbf{X}\right]\mathbf{x}_i\mathbf{x}_i'\right] & \mathrm{E}\left[\sigma^2\mathbf{x}_i'\mathrm{E}\left[\epsilon_i|\mathbf{X}\right] - \mathbf{x}_i'\mathrm{E}\left[\epsilon_i^3|\mathbf{X}\right]\right] \\ \mathrm{E}\left[\mathrm{E}\left[\sigma^2\mathbf{x}_i'\epsilon_i - \mathbf{x}_i'\epsilon_i^3|\mathbf{X}\right]\right] & \mathrm{E}\left[\left(\sigma^2-\epsilon_i^2\right)^2\right] \end{bmatrix}\mathbf{S} \\
&= \mathbf{S}\begin{bmatrix} \sigma^2\mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i'\right] & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\mathbf{S} = \begin{bmatrix} \frac{1}{\sigma^2}\mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i'\right] & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}
\end{aligned}$$

The estimators of the covariance matrices are

$$\begin{aligned}
\hat{\mathcal{J}} &= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix}\begin{bmatrix} \mathbf{x}_i\hat\epsilon_i \\ \hat\sigma^2-\hat\epsilon_i^2 \end{bmatrix}\begin{bmatrix} \mathbf{x}_i'\hat\epsilon_i & \hat\sigma^2-\hat\epsilon_i^2 \end{bmatrix}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix} \\
&= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix}\begin{bmatrix} \hat\epsilon_i^2\mathbf{x}_i\mathbf{x}_i' & \hat\sigma^2\mathbf{x}_i\hat\epsilon_i - \mathbf{x}_i\hat\epsilon_i^3 \\ \hat\sigma^2\mathbf{x}_i'\hat\epsilon_i - \mathbf{x}_i'\hat\epsilon_i^3 & \left(\hat\sigma^2-\hat\epsilon_i^2\right)^2 \end{bmatrix}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mathcal{I}} &= -1\times n^{-1}\sum_{i=1}^{n}\begin{bmatrix} -\frac{1}{\hat\sigma^2}\mathbf{x}_i\mathbf{x}_i' & -\frac{1}{\hat\sigma^4}\mathbf{x}_i\epsilon_i \\ -\frac{1}{\hat\sigma^4}\mathbf{x}_i\epsilon_i & \frac{1}{2\hat\sigma^4} - \frac{\epsilon_i^2}{\hat\sigma^6} \end{bmatrix} \\
&= -1\times n^{-1}\sum_{i=1}^{n}\begin{bmatrix} -\frac{1}{\hat\sigma^2}\mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & \frac{1}{2\hat\sigma^4} - \frac{\hat\sigma^2}{\hat\sigma^6} \end{bmatrix} \\
&= -1\times n^{-1}\sum_{i=1}^{n}\begin{bmatrix} -\frac{1}{\hat\sigma^2}\mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & -\frac{1}{2\hat\sigma^4} \end{bmatrix} = n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix}\begin{bmatrix} \mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & 1 \end{bmatrix}
\end{aligned}$$

Note that the off-diagonal term in $\mathcal{J}$, $\hat\sigma^2\mathbf{x}_i'\hat\epsilon_i - \mathbf{x}_i'\hat\epsilon_i^3$, is not necessarily 0 when the data may be conditionally skewed. Combined, the QMLE parameter covariance estimator is then

$$\begin{aligned}
\hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}\hat{\mathcal{I}}^{-1} &= \left(n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1}\left[n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \hat\epsilon_i^2\mathbf{x}_i\mathbf{x}_i' & \hat\sigma^2\mathbf{x}_i\hat\epsilon_i - \mathbf{x}_i\hat\epsilon_i^3 \\ \hat\sigma^2\mathbf{x}_i'\hat\epsilon_i - \mathbf{x}_i'\hat\epsilon_i^3 & \left(\hat\sigma^2-\hat\epsilon_i^2\right)^2 \end{bmatrix}\right] \\
&\quad\times\left(n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1}
\end{aligned}$$

where the identical scaling terms have been canceled. Additionally, when returns are conditionally normal,

$$\begin{aligned}
\operatorname{plim}\hat{J} &= \operatorname{plim} n^{-1}\sum_{i=1}^{n}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix}\begin{bmatrix} \hat\epsilon_i^2\mathbf{x}_i\mathbf{x}_i' & \hat\sigma^2\mathbf{x}_i\hat\epsilon_i - \mathbf{x}_i\hat\epsilon_i^3 \\ \hat\sigma^2\mathbf{x}_i'\hat\epsilon_i - \mathbf{x}_i'\hat\epsilon_i^3 & \left(\hat\sigma^2-\hat\epsilon_i^2\right)^2 \end{bmatrix}\begin{bmatrix} \frac{1}{\hat\sigma^2} & 0 \\ 0 & \frac{1}{2\hat\sigma^4} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}\begin{bmatrix} \sigma^2\mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{x}_i\mathbf{x}_i' & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}\text{plim}\,\hat{\mathcal{I}} &= \text{plim}\, n^{-1}\sum_{i=1}^{n}\begin{bmatrix}\frac{1}{\hat{\sigma}^2}\mathbf{x}_i\mathbf{x}_i' & 0\\ 0 & \frac{1}{2\hat{\sigma}^4}\end{bmatrix}\\ &= \begin{bmatrix}\frac{1}{\sigma^2}\mathbf{x}_i\mathbf{x}_i' & 0\\ 0 & \frac{1}{2\sigma^4}\end{bmatrix},\end{aligned}$$

and so the IME, $\text{plim}\,\hat{\mathcal{J}} - \hat{\mathcal{I}} = 0$, will hold when returns are conditionally normal. Moreover, when returns are not normal, all of the terms in $\mathcal{J}$ will typically differ from the limits above and so the IME will not generally hold.

2.7.3.1 Data and Implementation

Three assets are used to illustrate hypothesis testing: ExxonMobil (XOM), Google (GOOG) and the SPDR Gold Trust ETF (GLD). The data used to construct the individual equity returns were downloaded from Yahoo! Finance and span the period September 2, 2002, until September 1, 2012.[29] The market portfolio is the CRSP value-weighted market, which is a composite based on all listed US equities. The size and value factors were constructed using portfolio sorts and are made available by Ken French. All returns were scaled by 100.

2.7.3.2 Wald tests

Wald tests make use of the parameters and estimated covariance to assess the evidence against the null. When testing whether the size and value factor are relevant for an asset, the null is $H_0 : \beta_2 = \beta_3 = 0$. This problem can be set up as a Wald test using

$$\mathbf{R} = \begin{bmatrix}0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\end{bmatrix}, \mathbf{r} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

and

$$W = n\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right)'\left[\mathbf{R}\hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}\hat{\mathcal{I}}^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right).$$

The Wald test has an asymptotic χ^2_2 distribution since the null imposes 2 restrictions.

t-stats can similarly be computed for individual parameters

$$t_j = \sqrt{n}\frac{\hat{\beta}_j}{\text{s.e.}\left(\hat{\beta}_j\right)}$$

where s.e. $\left(\hat{\beta}_j\right)$ is the square of the j$^{\text{th}}$ diagonal element of the parameter covariance matrix. Table 2.7 contains the parameter estimates from the models, t-stats for the coefficients and the Wald test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$. The t-stats and the Wald tests were implemented using both the sandwich covariance estimator (QMLE) and the maximum likelihood covariance estimator. The two sets of test statistics differ in magnitude since the assumption of normality is violated in the data, and so only the QMLE-based test statistics should be considered reliable.

[29] Google and the SPDR Gold Trust ETF both started trading after the initial sample date. In both cases, all available data was used.

2.7.3.3 Likelihood Ratio tests

Likelihood ratio tests are simple to implement when parameters are estimated using MLE. The likelihood ratio test statistic is

$$LR = -2\left(l\left(\mathbf{r}|\mathbf{X};\tilde{\boldsymbol{\theta}}\right) - l\left(\mathbf{r}|\mathbf{X};\hat{\boldsymbol{\theta}}\right)\right)$$

where $\tilde{\boldsymbol{\theta}}$ is the null-restricted estimator of the parameters. The likelihood ratio has an asymptotic χ^2_2 distribution since there are two restrictions. Table 2.7 contains the likelihood ratio test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$. Caution is needed when interpreting likelihood ratio test statistics since the asymptotic distribution is only valid when the model is correctly specified – in this case, when returns are conditionally normal, which is not plausible.

2.7.3.4 Lagrange Multiplier tests

Lagrange Multiplier tests are somewhat more involved in this problem. The key to computing the LM test statistic is to estimate the score using the restricted parameters,

$$\tilde{\mathbf{s}}_i = \begin{bmatrix} \frac{1}{\tilde{\sigma}^2}\mathbf{x}_i\tilde{\epsilon}_i \\ -\frac{1}{2\tilde{\sigma}^2} + \frac{\tilde{\epsilon}_i^2}{2\tilde{\sigma}^4} \end{bmatrix},$$

where $\tilde{\epsilon}_i = r_i - \tilde{\boldsymbol{\beta}}'\mathbf{x}_i$ and $\tilde{\boldsymbol{\theta}} = \left[\tilde{\boldsymbol{\beta}}' \; \tilde{\sigma}^2\right]'$ is the vector of parameters estimated when the null is imposed. The LM test statistic is then

$$LM = n\bar{\tilde{\mathbf{s}}}\tilde{\mathbf{S}}^{-1}\bar{\tilde{\mathbf{s}}}$$

where

$$\bar{\tilde{\mathbf{s}}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \text{ and } \tilde{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i'.$$

The improved version of the LM can be computed by replacing $\tilde{\mathbf{S}}$ with a covariance estimator based on the scores from the unrestricted estimates,

$$\hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'.$$

Table 2.7 contains the LM test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$ using the two covariance estimators. LM test statistics are naturally robust to violations of the assumed normality since $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$ are directly estimated from the scores and not based on properties of the assumed normal distribution.

2.7.3.5 Discussion of Test Statistics

Table 2.7 contains all test statistics for the three series. The test statistics based on the MLE and QMLE parameter covariances differ substantially in all three series, and importantly, the conclusions also differ for the SPDR Gold Trust ETF. The difference between the two sets of results from an implicit rejection of the the assumption that returns are conditionally normal with constant variance. The MLE-based Wald test and the LR test (which is implicitly MLE-based) have very similar magnitudes for all three series. The QMLE-based Wald test statistics are also always larger than the LM-based test statistics which reflects the difference of estimating the covariance under the null or under the alternative.

		ExxonMobil			
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.016	0.774 (0.439)	0.774 (0.439)	Wald (MLE)	251.21 (<0.001)
β_1	0.991	60.36 (<0.001)	33.07 (<0.001)	Wald (QMLE)	88.00 (<0.001)
β_2	-0.536	−15.13 (<0.001)	−9.24 (<0.001)	LR	239.82 (<0.001)
β_3	-0.231	−6.09 (<0.001)	−3.90 (<0.001)	LM ($\tilde{\mathbf{S}}$)	53.49 (<0.001)
				LM ($\hat{\mathbf{S}}$)	54.63 (<0.001)

		Google			
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.063	1.59 (0.112)	1.60 (0.111)	Wald (MLE)	18.80 (<0.001)
β_1	0.960	30.06 (<0.001)	23.74 (<0.001)	Wald (QMLE)	10.34 (0.006)
β_2	-0.034	−0.489 (0.625)	−0.433 (0.665)	LR	18.75 (<0.001)
β_3	-0.312	−4.34 (<0.001)	−3.21 (0.001)	LM ($\tilde{\mathbf{S}}$)	10.27 (0.006)
				LM ($\hat{\mathbf{S}}$)	10.32 (0.006)

		SPDR Gold Trust ETF			
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.057	1.93 (0.054)	1.93 (0.054)	Wald (MLE)	12.76 (0.002)
β_1	0.130	5.46 (<0.001)	2.84 (0.004)	Wald (QMLE)	5.16 (0.076)
β_2	-0.037	−0.733 (0.464)	−0.407 (0.684)	LR	12.74 (0.002)
β_3	-0.191	−3.56 (<0.001)	−2.26 (0.024)	LM ($\tilde{\mathbf{S}}$)	5.07 (0.079)
				LM ($\hat{\mathbf{S}}$)	5.08 (0.079)

Table 2.7: Parameter estimates, t-statistics (both MLE and QMLE-based), and tests of the exclusion restriction that the size and value factors have no effect ($H_0 : \beta_2 = \beta_3 = 0$) on the returns of the ExxonMobil, Google and SPDR Gold Trust ETF.

Shorter Problems

Problem 2.1. What influences the power of a hypothesis test?

Problem 2.2. Let Y_i be i.i.d. Exponential(λ) with pdf $f(y_i) = \lambda \exp(-\lambda y_i)$, $\lambda > 0$. Derive the MLE of λ where there are n observations.

Problem 2.3. If n observations of $Y_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) are observed, what is the MLE of p? The pdf of a single Bernoulli is

$$p^y (1-p)^{1-y}.$$

Problem 2.4. When performing a hypothesis test, what are Type I and Type II Errors?

Longer Exercises

Exercise 2.1. The distribution of a discrete random variable X depends on a discretely valued parameter $\theta \in \{1,2,3\}$ according to

x	$f(x\|\theta=1)$	$f(x\|\theta=2)$	$f(x\|\theta=3)$
1	$\frac{1}{2}$	$\frac{1}{3}$	0
2	$\frac{1}{3}$	$\frac{1}{4}$	0
3	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$
4	0	$\frac{1}{12}$	$\frac{1}{12}$
5	0	0	$\frac{3}{4}$

Find the MLE of θ if one value from X has been observed. *Note*: The MLE is a function that returns an estimate of θ given the data that has been observed. In the case where both the observed data and the parameter are discrete, a "function" will take the form of a table.

Exercise 2.2. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a gamma(α,β) distribution. The density of a gamma(α,β) is

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$$

where $\Gamma(z)$ is the gamma-function evaluated at z. Find the MLE of β assuming α is known.

Exercise 2.3. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the pdf

$$f(x|\theta) = \frac{\theta}{x^{\theta+1}}, \quad 1 \le x < \infty, \theta > 1$$

1. What is the MLE of θ?
2. What is $\mathrm{E}[X_j]$?
3. How can the previous answer be used to compute a method of moments estimator of θ?

Exercise 2.4. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta, \theta > 0$$

1. What is the MLE of θ? [This is tricky]
2. What is the method of moments Estimator of θ?
3. Compute the bias and variance of each estimator.

Exercise 2.5. Let $X_1, \ldots, X_n$ be an i.i.d. random sample from the pdf

$$f(x|\theta) = \theta x^{\theta-1}, \;\; 0 \leq x \leq 1, 0 < \theta < \infty$$

1. What is the MLE of θ?
2. What is the variance of the MLE?
3. Show that the MLE is consistent.

Exercise 2.6. Let $X_1, \ldots, X_i$ be an i.i.d. sample from a Bernoulli(p).

1. Show that $\bar{X}$ achieves the Cramér-Rao lower bound.
2. What do you conclude about using $\bar{X}$ to estimate p?

Exercise 2.7. Suppose you witness a coin being flipped 100 times with 56 heads and 44 tails. Is there evidence that this coin is unfair?

Exercise 2.8. Let $X_1, \ldots, X_i$ be an i.i.d. sample with mean μ and variance σ^2.

1. Show $\tilde{X} = \sum_{i=1}^{N} w_i X_i$ is unbiased if and only if $\sum_{i=1}^{N} w_i = 1$.
2. Show that the variance of $\tilde{X}$ is minimized if $w_i = \frac{1}{n}$ for $i = 1, 2, \ldots, n$.

Exercise 2.9. Suppose $\{X_i\}$ in i.i.d. sequence of normal variables with unknown mean μ and known variance σ^2.

1. Derive the power function of a 2-sided t-test of the null $H_0 : \mu = 0$ against an alternative $H_1 : \mu \neq 0$? The power function should have two arguments, the mean under the alternative, μ_1 and the number of observations n.
2. Sketch the power function for $n = 1, 4, 16, 64, 100$.
3. What does this tell you about the power as $n \to \infty$ for $\mu \neq 0$?

Exercise 2.10. Let X_1 and X_2 are independent and drawn from a Uniform($\theta, \theta + 1$) distribution with θ unknown. Consider two test statistics,

$$T_1 : \text{Reject if } X_1 > .95$$

and

$$T_2 : \text{Reject if } X_1 + X_2 > C$$

1. What is the size of T_1?

2. What value must C take so that the size of T_2 is equal to T_1

3. Sketch the power curves of the two tests as a function of θ. Which is more powerful?

Exercise 2.11. Suppose $\{y_i\}$ are a set of transaction counts (trade counts) over 5-minute intervals which are believed to be i.i.d. distributed from a Poisson with parameter λ. Recall the probability density function of a Poisson is

$$f(y_i;\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

1. What is the log-likelihood for this problem?

2. What is the MLE of λ?

3. What is the variance of the MLE?

4. Suppose that $\hat{\lambda} = 202.4$ and that the sample size was 200. Construct a 95% confidence interval for λ.

5. Use a t-test to test the null $H_0 : \lambda = 200$ against $H_1 : \lambda \neq 200$ with a size of 5%

6. Use a likelihood ratio to test the same null with a size of 5%.

7. What happens if the assumption of i.i.d. data is correct but that the data does not follow a Poisson distribution?

Upper tail probabilities for a standard normal z

Cut-off c	$\Pr(z > c)$
1.282	10%
1.645	5%
1.96	2.5%
2.32	1%

5% Upper tail cut-off for χ^2_q

Degree of Freedom q	Cut-Off
1	3.84
2	5.99
199	232.9
200	234.0

Exercise 2.12. Suppose $Y_i|X_i = x_i \overset{\text{i.i.d.}}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$

1. Write down the log-likelihood for this problem.

2. Find the MLE of the unknown parameters.

3. What is the asymptotic distribution of the parameters?

4. Describe two classes tests to test the null $H_0 : \beta_1 = 0$ against the alternative $H_0 : \beta_1 \neq 0$.

5. How would you test whether the errors in the model were conditionally heteroskedastic?

6. Suppose $X_i \overset{\text{i.i.d.}}{\sim} N\left(\mu_X, \sigma_X^2\right)$ and the X variables are independent of the shocks in the model. What are the values of:

   (a) $\mathrm{E}[Y_i]$

   (b) $\mathrm{E}\left[Y_i^2\right]$

   (c) $\mathrm{V}[Y_i]$

   (d) $\mathrm{Cov}[X_i, Y_i]$
   Note: If $Y \sim N\left(\mu, \sigma^2\right)$, then the pdf of Y is

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Exercise 2.13. Suppose $Y_i \overset{\text{i.i.d.}}{\sim}$ Exponential(λ), so that $\mathrm{E}[Y_i] = \lambda$.

1. Write down the log-likelihood for this problem.

2. Find the MLE of the unknown parameter.

3. What is the asymptotic distribution of the parameter estimate?

4. Suppose $n = 10$, $\sum y_i = 19$. Test the null $H_0 : \lambda = 1$ against a 2-sided alternative with a size of 5% test using a t-test.

5. Suppose $n = 10$, $\sum y_i = 19$. Test the null $H_0 : \lambda = 1$ against a 2-sided alternative with a size of 5% test using a likelihood-ratio.

6. When are sandwich covariance estimators needed in MLE problems?

7. Discuss the important considerations for building models using cross-sectional data?
   Notes:

   - If $Y \sim$ Exponential(λ), then the pdf of Y is

$$f(y; \lambda) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right)$$

   - The 5% critical value for a χ_1^2 is 3.8415, for a χ_2^2 is 5.9915 and for a χ_3^2 is 7.8147.

Exercise 2.14. Suppose $y_i | x_i \sim$ Exponential$(x_i\beta)$ where $x_i > 0$ and $\beta > 0$. This can be equivalently written $y_i \sim$ Exponential(λ_i) where $\lambda_i = x_i\beta$. The PDF of an exponential random variance with parameter λ is

$$f_Y(y) = \lambda \exp(-\lambda y).$$

Assume n pairs of observations on (y_i, x_i) are observed

1. What is the log-likelihood of the data?
2. Compute the maximum likelihood estimator $\hat{\beta}$.
3. What is the asymptotic distribution of $\sqrt{n}\left(\hat{\beta} - \beta\right)$?
4. Suppose the following quantities are observed

$$\begin{aligned} n &= 20 \\ \sum_{i=1}^{n} x_i &= 16.58 \\ \sum_{i=1}^{n} y_i &= 128.47 \\ \sum_{i=1}^{n} x_i y_i &= 11.23 \end{aligned}$$

   Perform a test for the null $H_0 : \beta = 1.5$ against the alternative $H_1 : \beta \neq 1.5$ using a t-test.
5. Explain how you would perform a likelihood-ratio test for the same null and alternative.