FinMath Academic Review
Regression

Mark Hendricks
Financial Mathematics, University of Chicago

September 2018

Contents

1 Projection

1.1 Environment

- We are interested in a random scalar variable, y.
- We observe a random $k \times 1$ column vector, $\mathbf{x}$.
- We do not put any restriction on the joint distribution of $(y, \mathbf{x})$.

1.2 Conditional expectation

Let's consider the conditional expectation in the context of an optimal approximation.

- Suppose we want to approximate y having observed $\mathbf{x}$.
- Let $f(\mathbf{x})$ denote an arbitrary function, resulting in the error $\varepsilon = y - f(\mathbf{x})$.
- If we specify a loss function that we wish to minimize, $\ell(\varepsilon)$, then we can consider the optimal approximating function, $f^*(\mathbf{x})$, that minimizes $\ell(\varepsilon)$.

For the **mean squared error** loss function,

$$\ell(\varepsilon) = \mathbb{E}\left[(y - f(\mathbf{x}))^2\right]$$

the **conditional expectation** is the optimal approximation,

$$f^*(\mathbf{x}) = \mathbb{E}\left[y \,|\, \mathbf{x}\right]$$

1.3 Decomposition

In general, the conditional expectation may,

- depend on the joint distribution of $(y, \mathbf{x})$
- be a nonlinear function of $\mathbf{x}$.

We can decompose y into two parts, the conditional expectation and the approximation error.

$$\begin{aligned} y &= \mathbb{E}\left[y \,|\, \mathbf{x}\right] + \varepsilon \\ 0 &= \mathbb{E}\left[\varepsilon \,|\, \mathbf{x}\right] \end{aligned}$$

Note that the second statement holds by construction. It is a significant restriction on ε, implying the following:

$$\mathbb{E}\left[\varepsilon \,|\, \mathbf{x}\right] = 0 \iff 0 = \mathbb{E}\left[f(\mathbf{x})\,\varepsilon\right], \forall f(\mathbf{x})$$

which implies,

$$\mathbb{E}\left[\mathbf{x}\,\varepsilon\right] = \mathbf{0}, \qquad 0 = \mathbb{E}\left[\varepsilon\right]$$

Furthermore, the last condition restricts the covariance and correlation:

$$\begin{aligned}\operatorname{cov}\left[f(\mathbf{x}),\, \varepsilon\right] &= \mathbf{0} = \operatorname{corr}\left[f(\mathbf{x}),\, \varepsilon\right],\ \forall f \\ \operatorname{cov}\left[\mathbf{x},\, \varepsilon\right] &= \mathbf{0} = \operatorname{corr}\left[\mathbf{x}\,\varepsilon\right]\end{aligned}$$

The approximation error cannot be improved through knowledge of $\mathbf{x}$: it is uncorrelated to any (nonlinear) function of $\mathbf{x}$, and certainly is uncorrelated to $\mathbf{x}$ itself.

1.4 Linear approximation

Suppose we restrict ourselves to linear approximation. If we again use the MSE loss function, we require a vector $\boldsymbol{\beta}^*$ that minimizes,

$$\ell(\varepsilon) = \mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\beta}\right)^2\right]$$

This is optimized for the choice, $\boldsymbol{\beta}^*$ which satisfies,[1]

$$\mathbb{E}\left[\mathbf{x}\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)\right] = \mathbf{0}$$

That, the optimal $\boldsymbol{\beta}^*$ ensures the approximation error is orthogonal to $\mathbf{x}$. Existence requires the following assumption,

Assumption 1 (Identified)**.** The $k \times k$ matrix, $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]$, is nonsingular, (and thus finite.)

Note,

- If this assumption does not hold, then the linear factors, $\mathbf{x}$ are not **identified**. That is, one of them is linearly a function of the other and thus its impact on y is not well defined in relation to the other factors.

[1] Consider the MSE for an arbitrary $\boldsymbol{\beta}$,

$$\begin{aligned}\ell\left(\epsilon\right) =& \mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\beta}\right)^2\right] \\ =& \mathbb{E}\left[\left[\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right) + \mathbf{x}'\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right)\right]^2\right] \\ =& \mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)^2\right] + 2\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right)' \underbrace{\mathbb{E}\left[\mathbf{x}\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)\right]}_{0 \text{ by orthogonality of } \boldsymbol{\beta}^*} + \mathbb{E}\left[\left(\mathbf{x}'\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right)\right)^2\right] \\ =& \mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)^2\right] + \mathbb{E}\left[\left(\mathbf{x}'\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right)\right)^2\right] \\ >& \mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)^2\right]\end{aligned}$$

Thus, any choice, β has MSE at least as large as $\boldsymbol{\beta}^*$.

- This assumption is not too restrictive; if one of the k elements of $\mathbf{x}$ is an exact linear function of the others, simply drop it from consideration. It additionally requires that $\mathbf{x}$ have a well defined (finite) mean.

Rearranging, we have

$$\boldsymbol{\beta}^* = \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1}\mathbb{E}\left[\mathbf{x}y\right] \tag{1}$$

Thus, $\mathbb{L}(y \mid \mathbf{x})$. is the optimal linear approximation of y given $\mathbf{x}$.

$$\mathbb{L}(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}^*$$

1.5 Example: Linear dependence

Consider an example where Assumption 1 is violated, and $\mathbf{x}\mathbf{x}'$ is not invertible.

- Suppose in analyzing mean excess returns for S&P 500 stocks, we wanted to also condition on whether the stock is in the Information Technology (IT) sector. We set two indicator variables in $\mathbf{x} = \begin{bmatrix} x^{[1]} \\ x^{[2]} \end{bmatrix}$.

$$x^{[1]} = \mathbb{I}_{\text{IT sector}},\ x^{[2]} = \mathbb{I}_{\text{not IT sector}}$$

- But clearly, $x^{[1]}$ and $x^{[2]}$ are not linearly independent:

$$x_i^{[2]} = 1 - x_i^{[1]},\ \forall i$$

- We need to drop either linearly dependent variable.

This example is contrived: it should have been clear we did not need to set separate complementary indicator variables. However, this issue does arise if we are setting indicator variables for many categories, and forget to exclude one category to make it the baseline.

For instance, we might look at how membership in each of the 11 GICS-defined sectors impacts a stock's mean excess returns. Then y would be mean excess return, and $\mathbf{x}$ should have indicator variables for just 10 of the 11 sectors. Whichever sector we exclude is seen as the baseline, for which each of the elements of $\boldsymbol{\beta}^*$ is measuring relative impact.

1.6 Including a constant

Suppose we allow an affine approximation by including a constant as one of the elements of $\mathbf{x}$.

$$\mathbf{x} = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}} \end{bmatrix}$$

where $\tilde{\mathbf{x}}$ are the non-constant elements of $\mathbf{x}$. The second moments are then equivalent to centered second moments:

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}\varepsilon\right] =& \operatorname{cov}\left[\mathbf{x}\varepsilon\right] \\ \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] =& \operatorname{var}\left[\mathbf{x}\right]\end{aligned}$$

It can be shown that the linear projection is simply,

$$\mathbb{L}(y \mid \mathbf{x}) = \mathbb{E}\left[y\right] + \left(\tilde{\mathbf{x}} - \mathbb{E}\left[\tilde{\mathbf{x}}\right]\right)' \left(\operatorname{var}\left[\tilde{\mathbf{x}}\right]\right)^{-1} \operatorname{cov}\left[\tilde{\mathbf{x}} y\right]$$

which in the case of scalar $\tilde{\mathbf{x}} = x$ reduces to

$$\mathbb{L}\left[y \mid \begin{bmatrix}1 \\ x\end{bmatrix}\right] = \mathbb{E}\left[y\right] + \left(x - \mathbb{E}\left[x\right]\right) \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}$$

This makes clear that including a constant is equivalent to de-meaning all the variables and then approximating.

1.7 Projection

$\mathbb{L}[y \mid \mathbf{x}]$ is decomposing y into a portion spanned by $\mathbf{x}$ and a portion orthogonal to $\mathbf{x}$. Thus, it is formally a projection of y onto $\mathbf{x}$.[2]

Note that

- It is a linear projection of y in that the operator is a linear function of $\mathbf{x}$.
- The coefficient vector of this linear function, $\boldsymbol{\beta}^*$, depends only on the second moments of $(y, \mathbf{x})$, not the entire joint distribution.

Thus we can decompose y into two parts, the linear projection and the orthogonal piece:

$$\begin{aligned}y =& \mathbb{L}\left[y \mid \mathbf{x}\right] + \epsilon \\ y =& \mathbf{x}'\boldsymbol{\beta}^* + \epsilon\end{aligned}$$

This is indeed a projection, and we can verify that

$$0 = \mathbb{L}\left[\epsilon \mid \mathbf{x}\right]$$

[2]Recall that a projection is a linear operator, P, such that $P^2 = P$.

$$\begin{aligned}\mathbb{L}[\mathbb{L}[y \mid \mathbf{x}] \mid \mathbf{x}] =& \mathbf{x}' \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x}\, \mathbb{L}[y \mid \mathbf{x}]\right] \\ =& \mathbf{x}' \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x}\, \underbrace{\mathbf{x}' \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x} y\right]}_{\mathbb{L}(y \mid \mathbf{x})}\right] \\ =& \mathbf{x}' \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x} y\right] \\ =& \mathbf{x}' \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}\left[\mathbf{x} y\right] \\ =& \mathbb{L}[y \mid \mathbf{x}]\end{aligned}$$

alternately stated as

$$0 = \mathbb{E}[\mathbf{x}\epsilon]$$

The last equality says ϵ is orthogonal to $\mathbf{x}$. Contrast this with the approximation error of the conditional expectation, ε, which is not just orthogonal to x but has zero conditional expectation.

Theorem 1. If $\mathbb{E}[\epsilon \mid \mathbf{x}] = 0$, then the linear projection is the conditional expectation.

$$\mathbb{E}[y \mid \mathbf{x}] = \mathbb{L}[y \mid \mathbf{x}]$$

Equivalently, if we assume that the conditional expectation is of the form,

$$\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}\boldsymbol{\theta} + \varepsilon$$

then the linear projection is the conditional expectation, with

$$\boldsymbol{\beta}^* = \boldsymbol{\theta}$$

1.8 Nonlinearity

Example 1 (Call Option)**.** Consider the value of a call option, denoted c, and the value of the underlying stock, denoted s. Suppose call options have the following relationship to underlying stock price:[3]

$$c = f(s) + \epsilon$$

where

- f is a nonlinear equation
- ϵ is the impact of transaction costs, has zero mean, and is independent of s.

Then by assumption, $\mathbb{E}[c \mid s] = f(s)$, with $\mathbb{E}[\varepsilon \mid s] = 0$. However, the linear projection of c onto $\mathbf{x} = \begin{bmatrix} 1 \\ s \end{bmatrix}$ is

$$\begin{aligned} \mathbb{L}[c \mid \mathbf{x}] &= \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} \mathbb{E}[\mathbf{x}c] \\ &= \mathbb{E}[c] + \frac{\operatorname{cov}(s,c)}{\operatorname{var}(s)}(s - \mathbb{E}[s]) \end{aligned}$$

Thus,

$$c = \mathbb{E}[c] + \frac{\operatorname{cov}(s,c)}{\operatorname{var}(s)}(s - \mathbb{E}[s]) + \epsilon$$

where the only restriction on ϵ is that

$$0 = \mathbb{E}[\mathbf{x}c] \iff 0 = \operatorname{corr}(s,c), 0 = \mathbb{E}[\epsilon]$$

So ϵ may predictably be small or large for certain values of s, since we do not have $\mathbb{E}[\epsilon \mid s] = 0$. However, the ways in which s predicts ϵ are not linear, given the zero correlation.

[3] This model is motivated by the Black-Scholes modeling you will see in FINM 33000, but it simplifies a number of things to focus on our point regarding approximation.

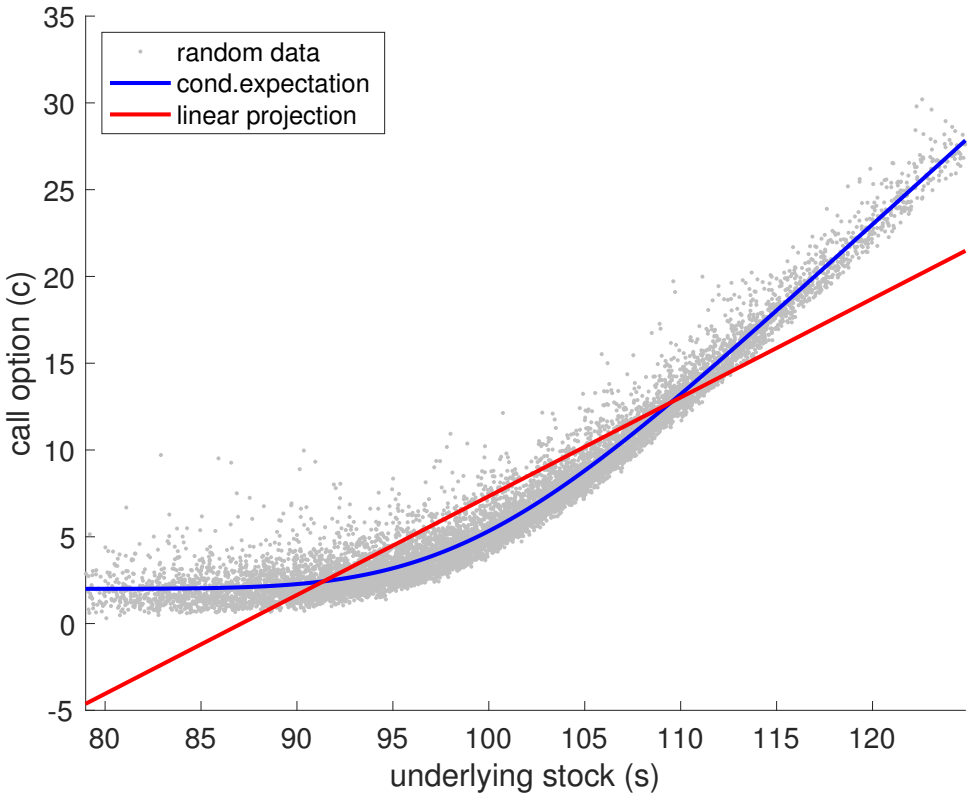

Figure 1: Illustration of Example 1

1.9 Partial derivative

The projection coefficient vector, $\boldsymbol{\beta}^*$ tells us that if we randomly observe an increase in $\mathbf{x}$, our optimal linear approximation is impacted by $\boldsymbol{\beta}^*$.

However, we are not inferring a causal impact from $\mathbf{x}$ to y.

- Suppose some other variable, z, causes change in y, but z is correlated with $\mathbf{x}$.
- The projection coefficients on x may be due to the implied change to z.
- For the conditional expectation, this indirect channel is wider: even if z is uncorrelated with $\mathbf{x}$, there may be nonlinear relationships between z from $\mathbf{x}$ which are useful in approximating y.

If we do not condition on z, yet it is correlated with both $\mathbf{x}$ and y, then it is known as a **confounding variable**. It causes the projection vector to identify not just causal impact, but inferred impact from z.

This question of **identification** is simple in theory: we should just condition on z in addition to $\mathbf{x}$. But in realized data, we'll see this may not be possible.

In many areas, researchers are focused on identifying causal impacts:

- y = salary, x = education, z = intelligence, responsibility, etc.
- y = crime rate, x = policing, z = lawfulness
- y = profitability, x = leverage, z = operating margin

However, there are many cases where this is not an issue, including many areas wehre machine-learning is increasingly used:

- Tax authorities predicting whether someone has unreported income.
- Netflix predicting your rating of a movie.
- Assessing portfolio risk to certain factors.

1.10 Confounding variables

Example 2 (Equity Premia)**.** Suppose we want to understand how a stock's characteristics, $\mathbf{x}$, impacts its premium, y.[4]

$$\mathbb{E}[y \mid x] = \mathbb{L}[y \mid x] = x\boldsymbol{\beta}^*$$

Let our population be the stocks of the S&P 500 from Jan 2005 to July 2018. Let $\mathbf{x}$ be a scalar variable, the stock's dividend-yield. Table 1 gives the projection coefficients.

	cond. on $\mathbf{x}$	cond. on ($\mathbf{x}$,z)
dividend-yield	0.0393	(-0.0009)
volatility		0.0796

Table 1: Projection of all S&P 500 mean excess returns on the stock's own dividend yield and return volatility.

Only using $\mathbf{x}$, the vector $\boldsymbol{\beta}^*$ is our best (linear) estimate of the impact of a stock return. However, it is not causal. It is based on the population's covariance between dividend-yield and other relevant factors. Should a firm take this as a prescription that if it increases the dividend yield by one point, the stock's mean excess return will change as indicated by $\boldsymbol{\beta}^*$?

Let z denote the stock's return volatility, and include it in the projection:

$$\mathbb{L}[y \mid \mathbf{x}, z] = \mathbf{x}'\boldsymbol{\beta}^* + z\gamma$$

See Table 1 for the projection coefficients.

[4] In FINM 36700, we discuss the equity premium as the mean excess return and will see that the Fundamental Theorem of Asset Pricing implies that the conditional expectation for y is linear in a certain set of variables, $\mathbf{x}$.

- Now, the estimated impact of dividend-yield on mean stock return almost disappears, while the impact through volatility is large.

- This happens because there is a positive covariance between dividend-yield and volatility. Thus, in the projection without volatility, dividend-yield is measuring its own impact as well as the associated impact via volatility.

- Specifically, for every point of dividend yield, we expect 0.5 points increase in standardized volatility.

From this, it seems that if a firm changes its dividend policy without making any more fundamental changes, it would lead to almost no impact on the mean excess return.

2 Stochastic processes

2.1 Stationarity

A stochastic process, $\{z_i\}$ is **stationary** if the joint distribution between z_i and z_{i+h} depends only on h, not on i. We often only require a weaker version, **covariance-stationary**, which says that the $\mathbb{E}[z_i]$ does not depend on i and that the covariances only depend on h.[5]

$$\gamma_h \equiv \text{ cov}[z_i, z_{i-h}] = \text{cov}[z_j, z_{j-h}]$$

Let $\mathbf{z}$ denote a set of n observations of the process $\{z_i\}$,

$$\mathbf{z} \equiv \begin{bmatrix} z_i & z_{i+1} & \dots & z_{i+n-1} \end{bmatrix}'$$

The $n \times n$ covariance matrix is then restricted to being a Toeplitz matrix,

$$\text{cov}[\mathbf{z}] = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-3} & \gamma_{n-2} & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-4} & \gamma_{n-3} & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-5} & \gamma_{n-4} & \gamma_{n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \gamma_{n-3} & \gamma_{n-4} & \gamma_{n-5} & \cdots & \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_{n-2} & \gamma_{n-3} & \gamma_{n-4} & \cdots & \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}$$

Thus for a set of n values from a (covariance) stationary process, the covariance matrix is restricted to n unique parameters rather than the usual $n(n+1)/2$.

2.2 Ergodicity

We give only a brief description of ergodicity and stationarity; the technical details require a separate treatment.

A stochastic process, $\{z_i\}$ is **ergodic** if it is asymptotically independent. Specifically, for any bounded, real-valued functions f and g,

$$\lim_{h \to \infty} |\mathbb{E}[f(z_i)g(z_{i+h})]| = |\mathbb{E}[f(z_i)]||\mathbb{E}[g(z_{i+h})]|$$

A few comments,

- Ergodicity allows for dependence among observations and for different observations to come from different distributions.
- Intuitively, ergodicity says that, given enough time, the process is drawn across the entire population—it does not get stuck in a cycle or become forever altered based on earlier realizations.

[5] Of course, this requires that the covariances are finite.

- We typically are interested in a slightly stronger version of ergodicity that guarantees that conditional expectations converge to unconditional expectations as the forecast horizon goes to infinity. We skip the technical details.[6]

- This easily generalizes to vector-valued stochastic processes, $\{\mathbf{z}_i\}$.

These two give us the following results,

- It implies that the population moments are well approximated by sample averages, for sufficiently large samples.

- Stationarity implies that the population distribution from which $(y_i, \mathbf{x}_i)$ is drawn is stable.

2.3 Consistency

Recall that a sequence of random variables, $\{z_n\}$ **converges in probability** to θ, if $\forall \epsilon > 0$,

$$\lim_{n \to \infty} \Pr\left(|\, z_n - \theta \,| > \epsilon\right) = 0$$

We denote this,

$$z_n \xrightarrow{P} \theta$$

Suppose we have an estimator, $A(\{x\}) = A_n$ which is a function of a random sample $\{x\}$ of size n and is intended to estimate some parameter of interest, α.

The estimator generates a sequence of random estimates, indexed by the sample size on which each estimate is based, $\{A_n\}$. We say this estimator is **consistent** if its probability limit equals the parameter being estimated, $\{A_n\} \xrightarrow{P} \alpha$.

2.4 Law of Large Numbers

Assumption 2 (Ergodic Stationarity)**.** We have a random sample that comes from a stochastic process which is **jointly stationary and ergodic.**

With this assumption, we can invoke perhaps the most important theorem in statistics: **The Law of Large Numbers (LLN).**

Theorem 2 (LLN)**.** Suppose the stochastic process $\{z\}_i$ satisfies Assumption 2, with $\mathbb{E}[z] = \mu$. Denote the sample average,

$$\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

[6]The extra conditions are often referred to as Gordin's condition, and the main restriction is that the conditional expectation of z becomes the unconditional expectation in the limit.

$\overline{z}$ is a consistent estimator of the population mean, μ.

$$\overline{z} \xrightarrow{P} \mu$$

- This is more general than the more familiar LLN which requires that the stochastic process is i.i.d.
- The LLN is important as it tells us that under Assumption 2, sample averages are consistent estimators for population means, including higher-order moments.

Example 3 (Estimating a variance)**.** Suppose we have observed n values of a scalar i.i.d.process $\{x_i\}$, with $\mathbb{E}[x_i] = \mu$ and

$$\text{var}[x_i] = \mathbb{E}\left[x_i^2\right] - (\mathbb{E}[x_i])^2 = \sigma^2$$

The LLN tells us that we can consistently estimate this second moment as

$$\begin{aligned} s^2 \equiv & \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \\ = & \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s^2 \xrightarrow{P} & \ \sigma^2 \end{aligned}$$

Simply taking the sample average of the targeted moment is a consistent estimator.[7]

Example 4 (Lazy estimator)**.** Suppose we estimated the mean by simply taking the average of the largest and smallest values of the sample,

$$\hat{x} = x_1$$

Not surprisingly, this is not consistent. As the sample size increases, $\hat{x}$ ignores the additional information and simply uses the first value seen.

$$\hat{x} \xrightarrow{P} \mu$$

[7] One might note s^2 usually adjusts for the degrees of freedom by dividing by $n-1$ instead of by n. However, the LLN tells us this is not required for a consistency which is concerned with the limiting behavior, and the usual degrees-of-freedom adjustment does not matter in the limit.

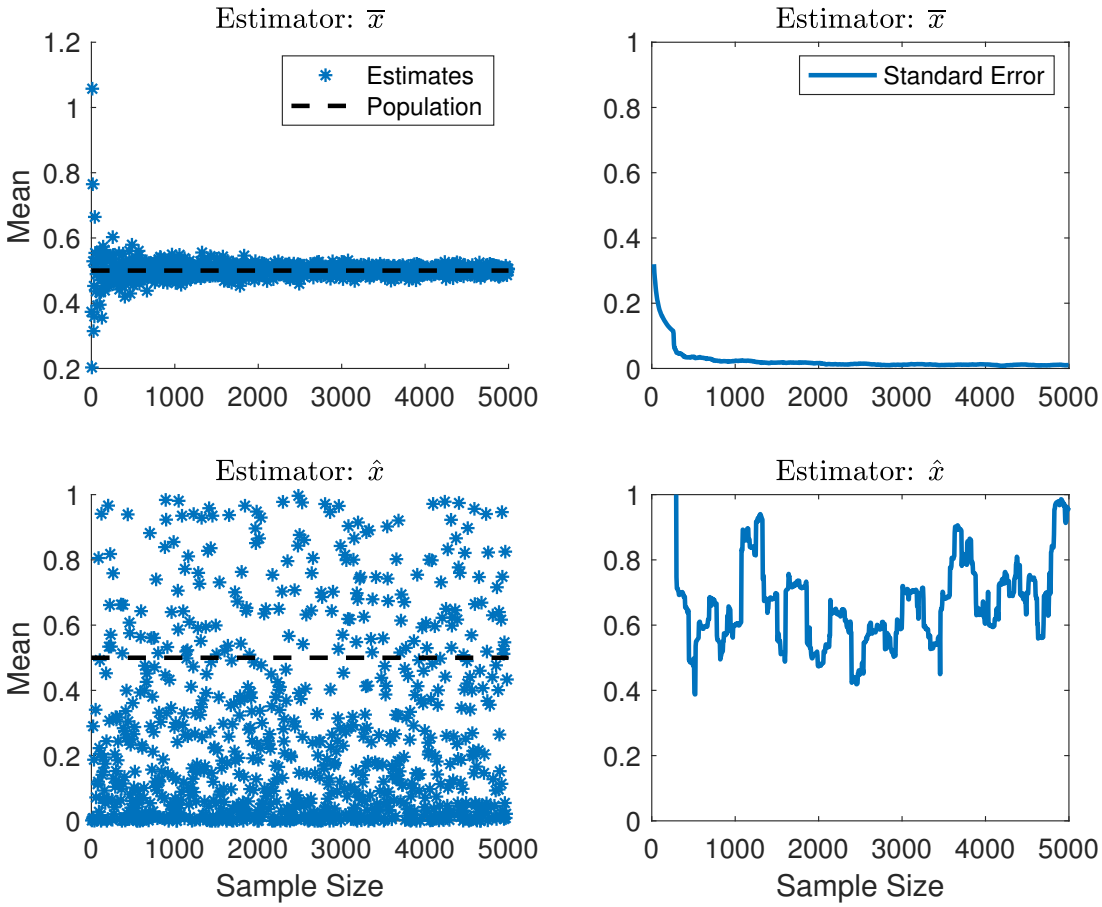

Figure 2: Data is simulated from a gamma distribution, with shape 0.5 and scale 1.0. The convergence is based on samples ranging from $n = 1$ to $n = 10,000$.

2.5 Variance of a sum

Suppose we have a set of scalar random variables, z_i, which do not satisfy Assumption 3 but do satisfy Assumption 2, meaning they are covariance stationary. By the arithmetic of covariances, it is easy to show that

$$\begin{aligned}
\text{var}\left(\bar{z}\right) =& \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right) \\
=& \frac{1}{n^2}\ \text{var}\left(\sum_{i=1}^{n} z_i\right) \\
=& \frac{1}{n^2}\sum_{h=-(n-1)}^{n-1} \left(n - |h|\right)\gamma_h \\
\gamma_h \equiv& \ \text{cov}\left[z_i, z_{i-h}\right]
\end{aligned}$$

Extending this for a process of random vectors, $\{\mathbf{z}_i\}$ leads to,

$$\begin{aligned} \text{cov}\left[\bar{\mathbf{z}}\right] &= \frac{1}{n^2}\sum_{h=-(n-1)}^{n-1}\left(n-|h|\right)\boldsymbol{\Gamma}_h \\ \boldsymbol{\Gamma}_h &\equiv \text{cov}\left[\mathbf{z}_i, \mathbf{z}_{i-h}\right] \end{aligned} \tag{2}$$

Note that we are taking advantage of the stationarity to rewrite second moments only as a function of the lag, h, not the indexing, (i, j),

$$\boldsymbol{\Gamma}_h = \mathbb{E}\left[\mathbf{z}_i\mathbf{z}_{i-h}'\right] = \mathbb{E}\left[\mathbf{z}_j\mathbf{z}_{j-h}'\right]$$

In fact, we can use the fact that $\boldsymbol{\Gamma}_{-h} = \boldsymbol{\Gamma}_h'$ to write,

$$\text{cov}\left[\bar{\mathbf{z}}\right] = \frac{1}{n}\boldsymbol{\Gamma}_0 + \frac{1}{n^2}\sum_{h=1}^{n-1}\left(n-h\right)\left(\boldsymbol{\Gamma}_h + \boldsymbol{\Gamma}_h'\right)$$

2.6 Asymptotic Distribution

Recall what is meant by a sequence of variables converging in distribution. For a sequence of random variables, $\{z_n\}$, we have a sequence of associated c.d.f.'s, denoted $\{F_n\}$. If

$$\{z_n\} \xrightarrow{D} z$$

implies that some variable, z, with c.d.f F, and $\lim_{n\to\infty} F_n = F$

The **asymptotic distribution** of an estimator refers to the limiting distribution of the sequence of estimators indexed by sample size. The **Central Limit Theorem (CLT)** gives the asymptotic distribution of a sample average under broad conditions.

Theorem 3 (CLT)**.** Let $\{\mathbf{z}_i\}$ be a stochastic process satisfying Assumption 2. Define the following notation, (and assume all the moments exist and are finite.)

$$\begin{aligned} \mathbb{E}\left[\mathbf{z}_i\right] &= \boldsymbol{\mu} \\ \text{cov}\left[\mathbf{z}_i\mathbf{z}_{i-h}\right] &= \boldsymbol{\Gamma}_h, \ \forall h \geq 0 \\ \bar{\mathbf{z}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i \end{aligned}$$

Then

$$\sqrt{n}\left(\bar{\mathbf{z}} - \boldsymbol{\mu}\right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\text{lim}}\right)$$

where the asymptotic covariance must take account of all potential covariances, given that the process is not i.i.d.

$$\boldsymbol{\Sigma}_{\text{lim}} \equiv \sum_{h=-\infty}^{\infty}\boldsymbol{\Gamma}_h$$

2.7 Special case: i.i.d.

ften, the CLT is applied with the additional assumption that the process is **Independently and Identically Distributed (i.i.d.)**

Assumption 3 (i.i.d.)**.** A process, $\{y_i, \mathbf{x}_i\}$ is i.i.d.conditional on $\{\mathbf{x}_i\}$ if the joint distribution of $(y_i, \mathbf{x}_i)$ is identical for all i and independent of the joint distribution for $(y_j, \mathbf{x}_j)$, for all $j \neq i$.

With Assumption 3, Theorem 3 simplifies as there is no longer any covariation between ϵ_i and ϵ_j. Thus, $\mathbf{\Gamma}_i = \mathbf{0}$ for all $i \neq 0$.

$$\mathbf{\Sigma}_{\text{lim}} = \mathbf{\Gamma}_0$$

Example 5. Gamma Distribution

Consider variables, z_i with an i.i.d.Gamma distribution. Figure 4 shows that the sample average as an estimator of $\mathbb{E}[z_i]$ converges to a normal distribution with smaller and smaller asymptotic variance.

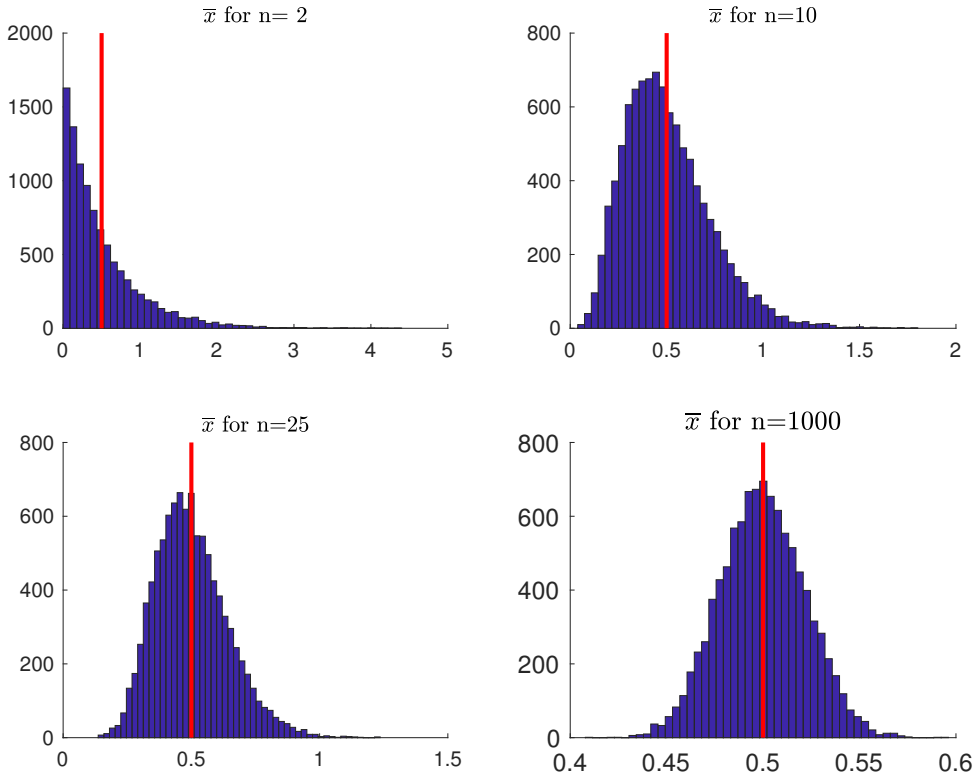

Figure 3: Illustration of the Central Limit Theorem. Sample average converges in distribution to a normal, even if the underlying data comes from a non-normal distribution. This simulation uses a Gamma distribution with shape 0.5 and scale 1.0. Estimated 10,000 times, with each sample mean based on $n = (2, 10, 25, 1000)$.

3 The Projection Estimate

3.1 Estimators

Estimators are functions of random data. Thus, the output of this function, the estimate, is itself a random variable. Its randomness comes from the randomness of the function's input.

For example, $\overline{x}$ is a function of sample data $\{x_i\}$, and the random variation in the different samples will cause random variation in the sample average, $\overline{x}$. See Figure 4 to see a histogram of $\overline{x}$ estimates, each based on samples of n data points. Note that in some samples, the resulting estimate is far from the true parameter.

Thus, we have several questions to consider when trying to estimate a parameter of interest.

- What estimator, (function,) is being used?
- Do the random estimates center around the true parameter?
- Do the random estimates tightly cluster around the true parameter?
- What is the distribution of the estimates? i.e. What is the probability that the estimate is a certain distance from the true parameter value?

We will be further analyzing each of these issues below, in the context of estimating the projection coefficient vector, $\boldsymbol{\beta}^*$.

Example 6 (Lazy estimator)**.** Consider again the lazy estimator of the mean of x_i:

$$\hat{x} = x_1$$

Figure 4 shows that this estimator does not center nor bunch tightly around μ.

3.2 Sample notation

- The linear projection vector in (1) depends on population moments.
- In applications, we rarely know the population; rather we must estimate based on a sample of data.

Suppose we observe n realizations of the stochastic process, $\{y_i, \mathbf{x}_i\}$.

- The observations are denoted $(y_i, \mathbf{x}_i)$ for $i = 1 \dots n$.

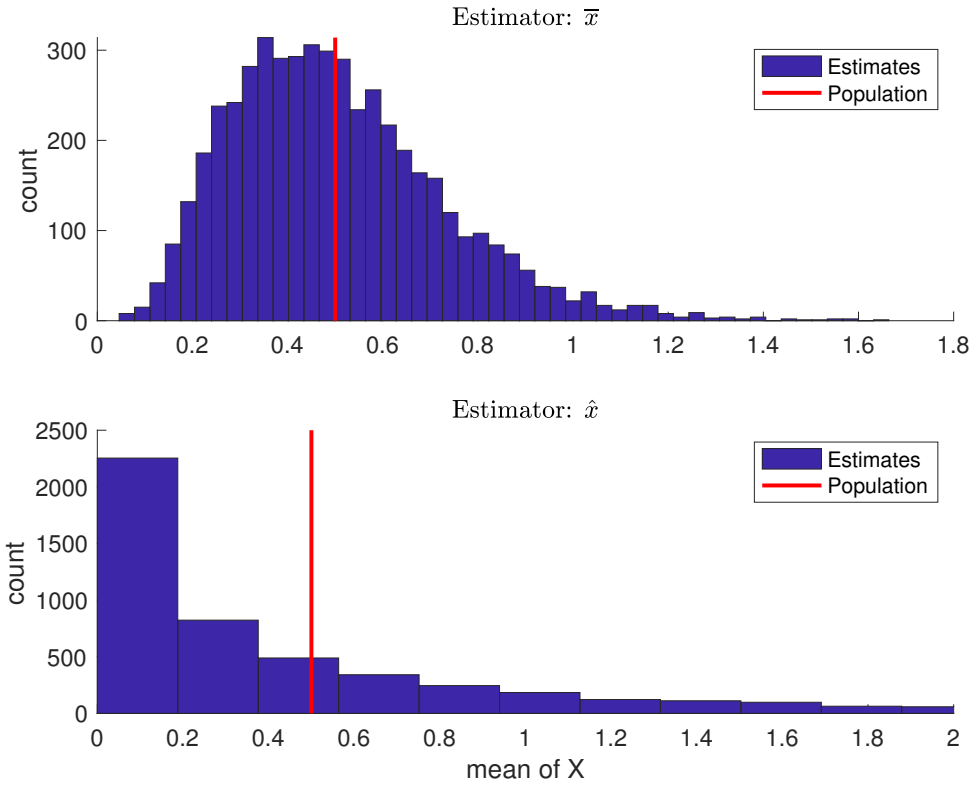

Figure 4: Data is simulated from a gamma distribution, with shape 0.5 and scale 1.0. The histogram is based on 5000 estimates of $\overline{x}$, which each are estimated from a sample size of 10.

- The total sample is denoted with the $n \times 1$ vector $\mathbf{y}$ and the $n \times k$ vector $\mathbf{X}$, where each row of $\mathbf{X}$ is an observation, $\mathbf{x}_i'$.

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times k}\underbrace{\mathbf{b}}_{k\times 1} + \underbrace{\mathbf{e}}_{n\times 1}$$

Consider estimating the projection from this sample:

- An estimate of $\boldsymbol{\beta}^*$ with some coefficient vector, $\mathbf{b}$, will decompose the sample, $\{y, \mathbf{x}\}_n$ into a sample of $\mathbf{x}'\mathbf{b}$ and $\mathbf{e}$.

- The projection errors, ϵ_i, associated with the sample values will be unobserved and unknown.

Consider the sample notation applied to Example 2:

$$\underbrace{\begin{bmatrix} \text{mean return}_1 \\ \text{mean return}_2 \\ \vdots \\ \text{mean return}_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \text{div}_1 & \text{vol}_1 \\ \text{div}_2 & \text{vol}_2 \\ \vdots & \vdots \\ \text{div}_n & \text{vol}_n \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} -.0009 \\ .0796 \end{bmatrix}}_{\boldsymbol{\beta}^*} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\text{unobserved error}}$$

Contrast this with the notation for a single observation:

$$\underbrace{\text{mean return}_i}_{y_i} = \underbrace{\begin{bmatrix}\text{div}_i & \text{vol}_i\end{bmatrix}}_{(\mathbf{x}_i)'} \underbrace{\begin{bmatrix}-.0009 \\ .0796\end{bmatrix}}_{\boldsymbol{\beta}^*} + \epsilon_i$$

3.3 Projection estimate

Consider estimating the projection coefficient vector, $\mathbf{b}^*$, using the same formula but simply replacing the population moments with their sample averages:[8]

$$\mathbf{b}^* = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Assumption 1 is required to imply that $\mathbf{X}'\mathbf{X}$ is nonsingular—that it has a non-zero determinant. This assumption will rarely be a problem mathematically; if it is violated, then we likely just need to drop the linearly dependent element of $\mathbf{X}$.[9] While this assumption will not be a problem mathematically, we will find that "nearly" violating it causes our biggest statistical issues.

This estimator ensures the resulting in-sample decomposition is indeed a projection,

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\mathbf{b}^* + \mathbf{e} \\ \mathbf{Y} &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} + \mathbf{e} \\ \mathbf{Y} &= \mathbf{P}\mathbf{Y} + \mathbf{e}\end{aligned}$$

where the projection matrix, $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

and it is easy to check that it indeed is a projection, maintaining its own map. Furthermore, one can easily show that the sample residual is uncorrelated to $\mathbf{X}$ and to $\mathbf{P}\mathbf{Y}$.

$$\mathbf{P}^2 = \mathbf{P} \qquad \mathbf{P}\mathbf{e} = 0 \qquad \mathbf{X}'\mathbf{e} = 0$$

Notice that the estimated errors, $\mathbf{e}$, will be different than the population errors, $\boldsymbol{\epsilon}$.

3.4 $\mathbf{b}^*$ as a random variable

The estimated linear projection vector is itself a random variable, which can be written as the sum of the population projection vector plus an error term: Rewrite the random variable, $\mathbf{b}^*$,

$$\begin{aligned}\mathbf{b}^* &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}\right) \\ &= \underbrace{\boldsymbol{\beta}^*}_{\text{constant}} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}\end{aligned} \tag{3}$$

[8] Below we motivate this approach by showing that under certain conditions it leads to consistent estimators. Generally, this approach is known as the method of moments.

[9] Technically, Assumption 1 only guarantees the condition in the population, not in any particular finite sample. Still, this means it will hold almost surely as the sample size goes to infinity.

Thus, the estimate is a constant plus a random variable estimation error:

$$\underbrace{\mathbf{b}^*}_{\text{random vector}} = \underbrace{\boldsymbol{\beta}^*}_{\text{unknown constant vector}} + \underbrace{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}}_{\text{estimation error}}$$

- $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of projection errors associated with the sample data, $\{y_i, \mathbf{x}_i\}$. They are unobserved.
- $\mathbf{e}$ is the $n \times 1$ vector of estimated projection errors associated with the sample data and the sample estimate, $\mathbf{b}^*$.
- By construction, $\boldsymbol{\epsilon}$ is orthogonal to $\mathbf{x}$ in population, while $\mathbf{e}$ is orthogonal to the sample $\mathbf{X}$.
- Even though ϵ is orthogonal to $\mathbf{x}$ in population, it almost surely will not be orthogonal to the sample $\mathbf{X}$, due to random variation.
- The estimation error, $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ is an unobserved random variable, given that it is the function of the unobserved $n \times 1$ vector, $\boldsymbol{\epsilon}$.

3.5 Consistent projection

Theorem 4. Given Assumptions 1 and 2, the estimator $\mathbf{b}^*$ is a consistent estimator of $\boldsymbol{\beta}^*$.

$$\mathbf{b}^* \xrightarrow{P} \boldsymbol{\beta}^*$$

The proof is direct.[10] In fact, we can view $\mathbf{b}^*$ as being based on consistent estimates for the two second

[10]

Proof. Recall from Equation (3),

$$\mathbf{b}^* = \boldsymbol{\beta}^* + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}$$

Thus, the consistency of $\mathbf{b}^*$ depends on two things. First, that $\lim_{n\to\infty}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is well defined (finite), which is guaranteed by Assumption 1. It also depends on

$$\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{P} \mathbf{0}$$

Given Assumption 2, the sample average converges to the population moment,

$$\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{P} \mathbb{E}\left[\mathbf{x}\epsilon\right]$$

By construction, linear projection ensures a decomposition with an ϵ such that,

$$\mathbb{E}\left[\mathbf{x}\epsilon\right] = \mathbf{0}$$

□

moments in a projection:

$$\begin{aligned}\frac{1}{n}\left(\mathbf{X}'\mathbf{X}\right) &\xrightarrow{P} \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] \\ \frac{1}{n}\left(\mathbf{X}'\mathbf{Y}\right) &\xrightarrow{P} \mathbb{E}\left[\mathbf{x}y\right] \\ \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} &\xrightarrow{P} \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1}\mathbb{E}\left[\mathbf{x}y\right] = \boldsymbol{\beta}^*\end{aligned}$$

3.6 Variability of the estimation error

Though $\mathbf{b}^*$ has arbitrary accuracy in estimating $\boldsymbol{\beta}^*$ as $n \to \infty$, we want to know how much variation there is in $\mathbf{b}^*$ for a fixed sample size n, and how fast this variation shrinks as n grows.

Consider the second moment of the estimation error.[11] Define the notation,

$$\boldsymbol{\upsilon}_i \equiv \mathbf{x}_i\epsilon_i$$

Define the sample averages estimating second moments,

$$\begin{aligned}\mathbf{S}_{\mathbf{x}} &\equiv \frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \\ \bar{\boldsymbol{\upsilon}} &\equiv \frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i\end{aligned}$$

We can write $\mathbf{b}^*$ as,

$$\mathbf{b}^* = \boldsymbol{\beta}^* + \mathbf{S}_{\mathbf{x}}^{-1}\bar{\boldsymbol{\upsilon}}$$

Then the variation in $\mathbf{b}^*$ is

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{b}} =& \mathbb{E}\left[\left(\mathbf{b}^* - \boldsymbol{\beta}^*\right)\left(\mathbf{b}^* - \boldsymbol{\beta}^*\right)' \mid \mathbf{X}\right] \\ =& \mathbb{E}\left[\mathbf{S}_{\mathbf{x}}^{-1}\bar{\boldsymbol{\upsilon}}\bar{\boldsymbol{\upsilon}}'\mathbf{S}_{\mathbf{x}}^{-1} \mid \mathbf{X}\right]\end{aligned}$$

Rewrite this as

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{b}} =& \mathbf{S}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}\mathbf{S}_{\mathbf{x}}^{-1} \\ \boldsymbol{\Sigma}_{\boldsymbol{\upsilon}} \equiv& \mathbb{E}\left[\bar{\boldsymbol{\upsilon}}\bar{\boldsymbol{\upsilon}}' \mid \mathbf{X}\right]\end{aligned} \tag{4}$$

3.7 Asymptotic distribution of $\mathbf{b}^*$

Theorem 5 (Limiting distribution of $\mathbf{b}^*$)**.** Under Assumptions 1 and 2,

$$\sqrt{n}\left(\mathbf{b}^* - \boldsymbol{\beta}^*\right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}^{\lim}\right) \tag{5}$$

[11] Throughout this section we discuss uncentered second moments rather than the covariances. This is because we want to avoid assuming that ϵ has zero mean conditional on x, which is a much stronger assumption that guarantees our linear projection is the conditional expectation.

where we define notation,

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{x}} &\equiv \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] \\ \boldsymbol{\Sigma}_{\mathbf{b}}^{\lim} &\equiv \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\upsilon}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\end{aligned}$$

Note that the asymptotic covariance is the same as the finite sample covariance in (4), but replacing the sample estimates $\mathbf{S}_{\mathbf{x}}$ with the population moment $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]$.

We refer to $\boldsymbol{\Sigma}_{\mathbf{b}}^{\lim}$ as the asymptotic covariance matrix of $\mathbf{b}^*$. In applications with large n, use $\boldsymbol{\Sigma}_{\mathbf{b}}$ as the covariance matrix of the estimation error.[12]

The theorem is important as it tells us how the random estimate, $\mathbf{b}^*$ is distributed around the value we are estimating, $\boldsymbol{\beta}^*$. This is necessary in order to test hypotheses on $\boldsymbol{\beta}^*$

[12] Careful readers will note that in finite samples we have not shown $\boldsymbol{\Sigma}_{\mathbf{b}}$ is a proper covariance matrix but only a second-moment. The consistency of $\mathbf{b}^*$ guarantees that in the limit, $\boldsymbol{\Sigma}_{\mathbf{b}}^{\lim}$ is a covariance.

4 Measuring variability of b*

This section considers how to estimate $\mathbf{\Sigma_b}$ under both an i.i.d.$\{\epsilon_i\}$ process as well as a non-i.i.d.process.

4.1 Errors that are i.i.d.

Let $\mathcal{I}$ denote the identity matrix. Under Assumption 3, we have

$$\begin{aligned}\mathbb{E}\left[\epsilon_i\epsilon_j \,|\, \mathbf{x}\right] =& 0,\ i \neq j\\ \mathbb{E}\left[\epsilon_i\epsilon_i \,|\, \mathbf{x}\right] =& \gamma_0^2,\ i = 1, \ldots, n\end{aligned}$$

This simplifies Equation (4),

$$\begin{aligned}\mathbf{\Sigma_\upsilon} &= \mathbb{E}\left[\mathbf{x}_i\epsilon_i\epsilon_i\mathbf{x}_i'\right]\\ &= \mathbf{\Sigma_x}\mathbf{\Sigma}_\epsilon\\ \mathbf{\Sigma}_\epsilon \equiv& \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \,|\, \mathbf{x}\right] = \gamma_0^2\mathcal{I}\end{aligned}$$

Thus,

$$\mathbf{\Sigma_b} = \gamma_0^2\ \mathbf{\Sigma_x}^{-1}$$

Estimate these moments with the sample averages,

$$\begin{aligned}s_0^2 \equiv& \frac{1}{n}\mathbf{e}'\mathbf{e}\\ \mathbf{S_b} =& s_0^2\ \left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned} \tag{6}$$

4.2 Simplifying to conditional homoskedasticity

Homoskedasticity refers to having constant second moments.[13] **Unconditional homoskedasticity** says that second moments are constant:

$$\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \sigma_{i,j}$$

Under the stationarity of Assumption 2, $\{\epsilon_i\}$ is unconditionally homoskedastic:

$$\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \gamma_{i-j}$$

For estimation of $\mathbf{\Sigma_\upsilon}$, we are interested in whether $\{\epsilon_i\}$ is **conditionally homoskedastic**, which is not guaranteed by Assumption 2.

[13] This is often stated in terms of covariances, but uncentered second moments allows for the case that the conditional mean is nonzero.

Assumption 4 (Conditional Homoskedasticity)**.** The process $\{\epsilon_i\}$ is homoskedastic conditional on $\{\mathbf{x}_i\}$.

$$\mathbb{E}\left[\epsilon_i \epsilon_j \,|\, \mathbf{x}_i, \mathbf{x}_j\right] = \sigma_{i,j}$$

Under Assumption 4, $\mathbf{\Sigma_v}$ simplifies to

$$\mathbf{\Sigma_v} = \mathbf{X}'\mathbf{\Sigma}_\epsilon\mathbf{X}$$

Estimating $\mathbf{\Sigma_x}$ is simple, but estimating $\mathbf{\Sigma}_\epsilon$ is harder given that we are assuming only conditional homoskedasticity, not i.i.d.. Thus, ϵ_i and ϵ_j can covary, but this covariance cannot depend on $\mathbf{x}$.

- To directly estimate $\mathbf{\Sigma}_\epsilon$, we have only n realizations with which to estimate the $n(n+1)/2$ unique elements.
- We could simply use the cross-products of the sample residuals to calculate each element, $\gamma_{i,j}$, as $e_i e_j$, but this is estimating making an estimate on each second moment based on a single sample cross product, $e_i e_j$.
- We need to add more structure in order to get any power to this estimate.

Under Assumption 2, the joint distribution of (ϵ_i, ϵ_j) depends only on $h = i - j$, not on i, j. Then for any pair of points in the sample,

$$\mathbb{E}\left[\epsilon_i \epsilon_j\right] = \mathbb{E}\left[\epsilon_{i+h} \epsilon_{j+h}\right]$$

This structure restricts the covariance matrix of n realizations of $\{\epsilon_i\}$ to

$$\mathbf{\Sigma}_\epsilon = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-3} & \gamma_{n-2} & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-4} & \gamma_{n-3} & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-5} & \gamma_{n-4} & \gamma_{n-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{n-3} & \gamma_{n-4} & \gamma_{n-5} & \cdots & \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_{n-2} & \gamma_{n-3} & \gamma_{n-4} & \cdots & \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}$$

Given a sample size of n, we can then estimate each γ_h with s_h:

$$s_h = \frac{1}{n-h} \sum_{i=h+1}^{n} e_i e_{i-h} \tag{7}$$

Let $\mathbf{S_e}$ denote the $n \times n$ matrix of these estimates, s_h, for $0 \leq h \leq n-1$.

Then use the estimators,

$$\begin{aligned} \mathbf{S_u} &= \mathbf{X}'\mathbf{S_e}\mathbf{X} \\ \mathbf{S_b} &= \mathbf{S_x}^{-1}\mathbf{S_u}\mathbf{S_x}^{-1} \end{aligned} \tag{8}$$

Estimating s_h for h near n will still have low power and possibly give poor estimates. Thus, an additional assumption is often added:

Assumption 5 (Orthogonality)**.** Error terms, ϵ_i and ϵ_{i-h} are conditionally orthogonal beyond for large h.

$$\mathbb{E}\left[\epsilon_i \epsilon_j\right] = 0, \quad \forall \ |i-j| \ > H$$

Suppose Assumption 5 holds for $H = 2$,

$$\mathbf{\Sigma}_\epsilon = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & & & \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & & \mathbf{0} & \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & \cdots & \gamma_0 & \gamma_1 & \gamma_2 \\ & \mathbf{0} & & \cdots & \gamma_1 & \gamma_0 & \gamma_1 \\ & & & \cdots & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}$$

- This does not increase our power for estimating γ_h for $h = 0, 1, 2$.
- However, it eliminates the estimation of the cross products for which we had the lowest power, (due to fewer high-lag cross products.)
- So this simplification makes sense when we have theoretical reasons to believe in orthogonality past a certain lag or simply a case where we do not trust the noisily estimated high-lag terms.

4.3 Estimation with conditional heteroskedasticity

Without Assumption 3 nor 4, we have the general Equation (4):

$$\begin{aligned} \mathbf{\Sigma_b} =& \mathbf{S_x^{-1}} \mathbf{\Sigma_v} \mathbf{S_x^{-1}} \\ \mathbf{\Sigma_v} \equiv& \mathbb{E}\left[\bar{\boldsymbol{v}} \bar{\boldsymbol{v}}' \,|\, \mathbf{X}\right] \end{aligned}$$

Note that $\mathbf{\Sigma_v}$ is the second moment matrix of a sample average, $\bar{\boldsymbol{v}}$. Thus, Equation (9) indicates that in general,

$$\begin{aligned} \mathbf{\Sigma_v} =& \frac{1}{n^2} \sum_{h=-(n-1)}^{n-1} (n - |h|) \, \mathbf{\Gamma}_h^v \\ =& \frac{1}{n^2} \left[n \, \mathbf{\Gamma}_0^v + 2 \sum_{h=1}^{n-1} (n-h) \left(\mathbf{\Gamma}_h^v + (\mathbf{\Gamma}_h^v)' \right) \right] \\ \mathbf{\Gamma}_h^v \equiv& \mathbb{E}\left[\boldsymbol{v}_i \boldsymbol{v}_{i-h}'\right] \end{aligned}$$

This suggests an estimator for $\mathbf{S_u}$, with elements that use sample averages to replace the covariance moments above.

$$\begin{aligned} \mathbf{S_u} =& \frac{1}{n^2} \sum_{h=-(n-1)}^{n-1} (n - |h|) \, \mathbf{\mathcal{S}}_h \\ =& \frac{1}{n^2} \left[n \, \mathbf{\mathcal{S}}_0 + \sum_{h=1}^{n-1} (n-h) \left(\mathbf{\mathcal{S}}_h + \mathbf{\mathcal{S}}_h' \right) \right] \end{aligned}$$

Define the estimator of the mixed second-moments,

$$\boldsymbol{\mathcal{S}}_h \equiv \frac{1}{n-h}\sum_{i=h+1}^{n} \boldsymbol{u}_i \boldsymbol{u}_{i-h}'$$

Substituting, we have the well-known spectral density estimate or $\boldsymbol{\Sigma}_{\boldsymbol{v}}$,

$$\mathbf{S}_{\boldsymbol{u}} = \frac{1}{n^2}\left[\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i' + \sum_{h=1}^{n-1}\sum_{i=h+1}^{n} \left(\boldsymbol{u}_i \boldsymbol{u}_{i-h}' + \boldsymbol{u}_{i-h}\boldsymbol{u}_i'\right)\right] \tag{9}$$

4.3.1 Restricting for large h

The **Hansen-Hodrick correction** implements Equation (9) with additionally making Assumption 5.

$$\boldsymbol{\mathcal{S}}_h = \begin{cases} \dfrac{1}{n-h}\displaystyle\sum_{i=h+1}^{n} \boldsymbol{u}_i \boldsymbol{u}_{i-h}' & \text{for } h < H \\ \mathbf{0} & \text{for } h \geq H \end{cases}$$

Thus the restriction limits the estimate to considering H total sample second-moments,

$$\mathbf{S}_{\boldsymbol{u}} = \frac{1}{n^2}\left[\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i' + \sum_{h=1}^{H-1}\sum_{i=h+1}^{n} \left(\boldsymbol{u}_i \boldsymbol{u}_{i-h}' + \boldsymbol{u}_{i-h}\boldsymbol{u}_i'\right)\right] \tag{10}$$

Of course, setting $H = n$ leaves us with the unrestricted Equation (9).

4.3.2 Ensuring the covariance is positive definite

Unfortunately, Equation (10) is not guaranteed to be positive definite, and often is not. The **Newey-West estimator** remedies this by putting less weight on $\boldsymbol{\mathcal{S}}$ for covariances with larger lags, h:

$$\mathbf{S}_{\boldsymbol{u}} = \frac{1}{n^2}\left[n\,\boldsymbol{\mathcal{S}}_0 + \sum_{h=1}^{H-1}(n-h) w_h \left(\boldsymbol{\mathcal{S}}_h + \boldsymbol{\mathcal{S}}_h'\right)\right]$$
$$w_h \equiv 1 - \frac{h}{H}$$

Thus,

$$\mathbf{S}_{\boldsymbol{u}} = \frac{1}{n^2}\left[\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i' + \sum_{h=1}^{H-1}\left(1 - \frac{h}{H}\right)\sum_{i=h+1}^{n} \left(\boldsymbol{u}_i \boldsymbol{u}_{i-h}' + \boldsymbol{u}_{i-h}\boldsymbol{u}_i'\right)\right] \tag{11}$$

The Newey West Estimator in (11) is widely implemented in computational statistics libraries.

4.3.3 Assuming no serial correlation

While the estimator in Equation (11) is very general, it requires trusting that the estimates for high-order serial correlation are better than a theoretical restriction that these equal zero.

- This is the usual tradeoff between assumptions and statistical power.
- In particular, financial return data often has very low serial correlation, and supported by basic financial theory.
- Thus, we might want an estimator that allows for heteroskedasticity without allowing for serial correlation.

White's Estimator is a popular way to do this, and is simply a restricted version of (10) and (11):

$$\mathbf{S}_{\boldsymbol{u}} = \frac{1}{n^2}\left[\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i'\right] \tag{12}$$

4.3.4 Putting it together to get $\boldsymbol{\Sigma}_{\mathbf{b}}$

Using the estimators in Equations (9), (10), or (11), we finally have a consistent estimator for $\boldsymbol{\Sigma}_{\mathbf{b}}$:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{b}} =& \mathbf{S}_{\mathbf{x}}^{-1}\mathbf{S}_{\boldsymbol{u}}\mathbf{S}_{\mathbf{x}}^{-1} \\ =& (\mathbf{X}\mathbf{X})^{-1}\left[\sum_{i=1}^{n} \boldsymbol{u}_i\boldsymbol{u}_i' + \sum_{h=1}^{n-1}\sum_{i=h+1}^{n}\left(\boldsymbol{u}_i\boldsymbol{u}_{i-h}' + \boldsymbol{u}_{i-h}\boldsymbol{u}_i'\right)\right](\mathbf{X}\mathbf{X})^{-1}\end{aligned}$$

where we note that the $\frac{1}{n^2}$ in $\mathbf{S}_{\boldsymbol{u}}$ cancels with the $\frac{1}{n}$ inside each $\mathbf{S}_{\mathbf{x}}$.

4.4 Estimating the asymptotic variance

Theorem 5 contains an expression for the asymptotic covariance of $\mathbf{b}^*$, denoted $\boldsymbol{\Sigma}_{\mathbf{b}}^{\lim}$. Estimating this asymptotic covariance given a sample size of n is done by estimating $\boldsymbol{\Sigma}_{\mathbf{b}}$ using Equation (11), (8), or (6) according to the assumptions used.

As a corollary, the consistency of $\mathbf{b}^*$ ensures the consistency of using sample averages to estimate $\boldsymbol{\Sigma}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}$, and $\boldsymbol{\Sigma}_{\mathbf{b}}$.

$$s_0^2 \xrightarrow{P} \gamma_0^2 \qquad \mathbf{S}_{\mathbf{e}} \xrightarrow{P} \boldsymbol{\Sigma}_{\epsilon} \qquad \mathbf{S}_{\boldsymbol{u}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{v}} \qquad \mathbf{S}_{\mathbf{b}} \xrightarrow{P} \boldsymbol{\Sigma}_{\mathbf{b}}$$

5 Inference

5.1 Notation

Define the i-coordinate vector,

$$\boldsymbol{\iota}_i \equiv \begin{bmatrix} 0 & 0 & \dots & \underbrace{1}_{\text{element } i} & \dots & 0 \end{bmatrix}'$$

For any vector, $\mathbf{z}$, let $\mathbf{z}_{[i]}$ denote element i of $\mathbf{z}$.

$$\mathbf{z}_{[i]} \equiv (\boldsymbol{\iota}_i)' \mathbf{z}$$

5.2 z-statistic

We wish to calculate probabilities of observing the estimate $\mathbf{b}^*_{[i]}$ conditional on $\boldsymbol{\beta}^*_{[i]}$ equaling some stated value, β_i.

$$\Pr\left(\text{estimate for } \boldsymbol{\beta}^* \leq \mathbf{b}^*_{[i]} \mid \boldsymbol{\beta}^*_{[i]} = \beta_i\right) \tag{13}$$

Recall that $\mathbf{b}^*$ has asymptotic distribution given by Equation (5),

$$\sqrt{n}\left(\mathbf{b}^* - \boldsymbol{\beta}^*\right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\text{lim}}_{\mathbf{b}}\right)$$

then for sample sizes with large n, we approximate,

$$\mathbf{b}^* \approx \mathcal{N}\left(\boldsymbol{\beta}^*, \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{b}}\right)$$

Define the *z*-statistic as

$$\mathsf{z}^* = \frac{\mathbf{b}^*_{[i]} - \beta_i}{\sigma^{\mathbf{b}}_{[i]}}$$

Then in terms of z, the expression in (13) can be rewritten as,

$$\Pr\left(\mathsf{z} \leq \mathsf{z}^*\right) = \Phi_{\mathcal{N}}\left(\mathsf{z}^*\right)$$

- $\left(\sigma^{\mathbf{b}}_{[i]}\right)^2$ is element i of diag $\left(\boldsymbol{\Sigma}^{\text{lim}}_{\mathbf{b}}\right)$.

$$\sigma^{\mathbf{b}}_{[i]} \equiv \sqrt{\boldsymbol{\iota}_i' \boldsymbol{\Sigma}_{\mathbf{b}} \boldsymbol{\iota}_i}$$

- $\Phi_{\mathcal{N}}$ denotes the cdf for a standard normal, and $\mathsf{t}^* \to \mathsf{z}^*$.

Define the p-statistic as the complementary probability, which gives the probability, conditional on the assumed value for $\boldsymbol{\beta}^*_{[i]}$, that an estimate would have been larger than the realized $\mathbf{b}^*_{[i]}$:

$$\mathsf{p} = 1 - \Phi_{\mathcal{N}}\left(\mathsf{z}^*\right)$$

5.3 t-statistic

The z-statistic above relies on $\mathbf{\Sigma_b}$, which depends on population moments which we do not know. Using the estimator, $\mathbf{S_b}$ changes the Normal distribution to a t-Distribution with $n-k$ degrees of freedom.

Thus estimating (13) can be done as,

$$\begin{aligned}
\mathbf{s}^{\mathbf{b}}_{[i]} &\equiv \sqrt{\boldsymbol{\iota}_i' \mathbf{S_b} \boldsymbol{\iota}_i} \\
\mathsf{t}^* &= \frac{\mathbf{b}^*_{[i]} - \beta_i}{\mathbf{s}^{\mathbf{b}}_{[i]}} \\
\mathsf{p} &= 1 - \Phi_{\mathsf{t}}\left(\mathsf{t}^*, n-k\right)
\end{aligned}$$

where Φ_{t} denotes the cdf for a t-distribution.

Notice that as $n \to \infty$, the t-Distribution converges to a standard normal distribution.

5.4 Hypothesis Testing

Being able to calculate the probability in (13) through the t-statistic gives us a way to test the hypothesis that a particular element of $\boldsymbol{\beta}^*$ equals a particular number, β.

The implementation depends on whether we are testing against the right-tail alternative, the left-tail alternative, or the two-sided alternative.

5.4.1 Right-tailed test

$$\begin{aligned}
\mathcal{H}_0 &: \ \boldsymbol{\beta}^*_{[i]} = \beta_i \\
\mathcal{H}_1 &: \ \boldsymbol{\beta}^*_{[i]} > \beta_i
\end{aligned}$$

To implement the test, we must choose a **significance level**, which is

$$\Pr\left(\text{reject } \mathcal{H}_0 \mid \mathcal{H}_0 = \text{true}\right) = \hat{\mathsf{p}}$$

The test result is simply,

$$\begin{aligned}
\mathsf{p} \geq \hat{\mathsf{p}} &\quad \text{Do not reject } \mathcal{H}_0 \\
\mathsf{p} < \hat{\mathsf{p}} &\quad \text{Reject } \mathcal{H}_0
\end{aligned}$$

Or, in terms of the t-statistic, it is

$$\begin{aligned}
\mathsf{t}^* \leq \mathsf{t}_{\hat{\mathsf{p}}} &\quad \text{Do not reject } \mathcal{H}_0 \\
\mathsf{t}^* > \mathsf{t}_{\hat{\mathsf{p}}} &\quad \text{Reject } \mathcal{H}_0 \\
\mathsf{t}_{\hat{\mathsf{p}}} &\equiv \Phi_{\mathsf{t}}^{-1}\left(1 - \hat{\mathsf{p}}, n-k\right)
\end{aligned}$$

5.4.2 Left-tailed test

Simplify modify to

$$\begin{aligned}\mathcal{H}_1: \ \boldsymbol{\beta}^*_{[i]} &< \beta_i \\ \mathsf{p} &\equiv \Phi_{\mathsf{t}}\left(\mathsf{t}^*, n-k\right) \\ \mathsf{t}_{\hat{\mathsf{p}}} &\equiv \ \Phi_{\mathsf{t}}^{-1}\left(\hat{\mathsf{p}}, n-k\right)\end{aligned}$$

5.4.3 Two-sided test

Suppose we test $\mathcal{H}_0$ against a two-sided alternative,

$$\begin{aligned}\mathcal{H}_0: \ \boldsymbol{\beta}^*_{[i]} &= \beta_i \\ \mathcal{H}_1: \ \boldsymbol{\beta}^*_{[i]} &\neq \beta_i\end{aligned}$$

Then we simply modify the test as follows,

$$\begin{aligned}|\mathsf{t}^*| \leq \mathsf{t}_{\hat{\mathsf{p}}} &\ \text{ Do not reject } \mathcal{H}_0 \\ |\mathsf{t}^*| > \mathsf{t}_{\hat{\mathsf{p}}} &\ \text{ Reject } \mathcal{H}_0 \\ \mathsf{t}_{\hat{\mathsf{p}}} \equiv &\ \Phi_{\mathsf{t}}^{-1}\left(1-\frac{\hat{\mathsf{p}}}{2}, n-k\right)\end{aligned}$$

5.5 Errors and Power

A **Type 1 error** is the probability that we reject $\mathcal{H}_0$ even though it is true.

- Thus, $\hat{\mathsf{p}}$ is the probability that using the test will lead to a Type 1 error.
- We can reduce this simply by choosing a lower significance level for the test.
- The significance level is commonly chosen to be 0.05.

A **Type 2 error** is the probability that we do not reject $\mathcal{H}_0$ even though it is false.

- The more cautious we are about Type 1 errors, the more we make Type 2 errors.

The **Power** of the test is the probability that we reject $\mathcal{H}_0$ when it is false.

- Power is the complement of the Type 2 error

$$\text{Power} = \ 1 - \text{Type 2 error}$$

- A consistent test has

$$\text{Power} \underset{n\to\infty}{\lim} 1$$

6 Descriptive statistics

6.1 R-squared

The **R-squared**, or coefficient of determination, in a regression is defined as

$$\begin{aligned} R^2_{y,x} =& \frac{\text{regression sum of squares}}{\text{total sum of squares}} \\ =& 1 - \frac{\text{error sum of squares}}{\text{total sum of squares}} \end{aligned}$$

Algebraically, this is

$$\begin{aligned} R^2_{y,x} =& \frac{\mathbf{b}^* \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ =& 1 - \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

Intuitively, the R-squared is the square of the correlation between y and the projection of y onto $\mathbf{x}$.

$$R^2_{y,\mathbf{x}} = [\text{corr}(\mathbf{Y}, \mathbf{PY})]^2$$

In a univariate regression of y on x,

$$R^2_{y,x} = [\text{corr}(y, x)]^2$$

6.1.1 Caveat: Regressing on a constant

The interpretation and formula for R-squared does not hold if there is no constant regressor.

- Without a constant, the R-squared will not necessarily be between 0 and 1.
- Without a constant, the R-squared will not necessarily be the square of the correlation between the sample $\mathbf{Y}$ and the projected Y values.
- Without a regressor, the fit can be improved simply by shifting the sample $\mathbf{Y}$ data by a constant.

6.2 Problems with multicollinearity

If regressor j is highly correlated with the other regressors, then the variance of the coefficient estimate can be written as

$$\text{var}\left[b_j \mid \mathbf{x}\right] = \left(\frac{1}{1 - R^2_{x_j, \mathbf{x}_{[-j]}}}\right) \frac{\sigma^2}{\sum_{i=1}^{n} \left(\mathbf{x}_{i,j} - \bar{\mathbf{x}}_j\right)^2}$$

where

$$R^2_{x_j, \mathbf{x}_{[-j]}}$$

denotes the R-squared from regressing x_j on the remaining regressors, (all columns except column j of $\mathbf{X}$.)

The term

$$\frac{1}{1 - R^2_{x_j, \mathbf{x}_{[-j]}}}$$

is known as the **variance inflation factor**.

- The VIF is closely related to the condition number of $\mathbf{X}'\mathbf{X}$.
- The condition number in linear algebra measures the sensitivity of inverting a matrix.
- It compares the largest and smallest eigenvalues.
- Most software packages will warn the user if the condition number (VIF) is too big.

6.3 Variation in regressors

The sample variance of x_j reduces the variance of the OLS estimate b_j.

- Again write

$$\text{var}\left[b_j \mid \mathbf{x}\right] = \left(\frac{1}{1 - R^2_{x_j, \mathbf{x}_{[-k]}}}\right) \frac{\sigma^2}{\sum_{i=1}^{n} \left(\mathbf{x}_{i,j} - \bar{\mathbf{x}}_j\right)^2}$$

- The denominator of the second term is the scaled sample variance.
- If there is not enough variation in the regressor data, then the OLS estimation can not precisely estimate $\boldsymbol{\beta}_j^*$.

The standard error of the OLS estimator depends on the variation in the regressors.

- The standard error of $\mathbf{b}^*$ decreases as the variation in $\mathbf{X}$ increases, holding other things equal.
- Recall that net interest margin refers to the spread in the lending and borrowing of banks.
- Consider using this as a regressor, for some financial or economic data.

References

[1] Fumio Hayashi, *Econometrics*. 2000.

[2] John Cochrane, *Asset Pricing*. 2005.

[3] Jeffrey Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. 2010.

[4] James Hamilton *Time Series Analysis*. 1994

[5] Ruey Tsay, *Analysis of Financial Time Series*. 2010