Chapter 7

Asymptotic Theory for Least Squares

7.1 Introduction

It turns out that the asymptotic theory of least squares estimation applies equally to the projection model and the linear CEF model. Therefore the results in this chapter will be stated for the broader projection model described in Section 2.18. Recall that the model is $Y = X'\beta + e$ with the linear projection coefficient $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right]$.

Maintained assumptions in this chapter will be random sampling (Assumption 1.2) and finite second moments (Assumption 2.1). We restate these here for clarity.

Assumption 7.1

1. The variables (Y_i, X_i), $i = 1, ..., n$, are i.i.d.
2. $\mathbb{E}\left[Y^2\right] < \infty$.
3. $\mathbb{E}\|X\|^2 < \infty$.
4. $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is positive definite.

The distributional results will require a strengthening of these assumptions to finite fourth moments. We discuss the specific conditions in Section 7.3.

7.2 Consistency of Least Squares Estimator

In this section we use the weak law of large numbers (WLLN, Theorem 6.1 and Theorem 6.2) and continuous mapping theorem (CMT, Theorem 6.6) to show that the least squares estimator $\widehat{\beta}$ is consistent for the projection coefficient β.

This derivation is based on three key components. First, the OLS estimator can be written as a continuous function of a set of sample moments. Second, the WLLN shows that sample moments converge in probability to population moments. And third, the CMT states that continuous functions preserve convergence in probability. We now explain each step in brief and then in greater detail.

First, observe that the OLS estimator

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY}$$

is a function of the sample moments $\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ and $\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$.

Second, by an application of the WLLN these sample moments converge in probability to their population expectations. Specifically, the fact that (Y_i, X_i) are mutually i.i.d. implies that any function of (Y_i, X_i) is i.i.d., including $X_i X_i'$ and $X_i Y_i$. These variables also have finite expectations under Assumption 7.1. Under these conditions, the WLLN implies that as $n \to \infty$,

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \underset{p}{\longrightarrow} \mathbb{E}\left[XX'\right] = \boldsymbol{Q}_{XX} \tag{7.1}$$

and

$$\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \underset{p}{\longrightarrow} \mathbb{E}\left[XY\right] = \boldsymbol{Q}_{XY}.$$

Third, the CMT allows us to combine these equations to show that $\widehat{\beta}$ converges in probability to β. Specifically, as $n \to \infty$,

$$\widehat{\beta} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} = \beta. \tag{7.2}$$

We have shown that $\widehat{\beta} \underset{p}{\longrightarrow} \beta$ as $n \to \infty$. In words, the OLS estimator converges in probability to the projection coefficient vector β as the sample size n gets large.

To fully understand the application of the CMT we walk through it in detail. We can write

$$\widehat{\beta} = g\left(\widehat{\boldsymbol{Q}}_{XX}, \widehat{\boldsymbol{Q}}_{XY}\right)$$

where $g(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{b}$ is a function of $\boldsymbol{A}$ and $\boldsymbol{b}$. The function $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is a continuous function of $\boldsymbol{A}$ and $\boldsymbol{b}$ at all values of the arguments such that $\boldsymbol{A}^{-1}$ exists. Assumption 7.1 specifies that $\boldsymbol{Q}_{XX}$ is positive definite, which means that $\boldsymbol{Q}_{XX}^{-1}$ exists. Thus $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is continuous at $\boldsymbol{A} = \boldsymbol{Q}_{XX}$. This justifies the application of the CMT in (7.2).

For a slightly different demonstration of (7.2) recall that (4.6) implies that

$$\widehat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe} \tag{7.3}$$

where

$$\widehat{\boldsymbol{Q}}_{Xe} = \frac{1}{n}\sum_{i=1}^{n} X_i e_i.$$

The WLLN and (2.25) imply

$$\widehat{\boldsymbol{Q}}_{Xe} \underset{p}{\longrightarrow} \mathbb{E}\left[Xe\right] = 0.$$

Therefore

$$\widehat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}0 = 0$$

which is the same as $\widehat{\beta} \underset{p}{\longrightarrow} \beta$.

Theorem 7.1 Consistency of Least Squares. Under Assumption 7.1, $\widehat{\boldsymbol{Q}}_{XX} \xrightarrow[p]{} \boldsymbol{Q}_{XX}$, $\widehat{\boldsymbol{Q}}_{XY} \xrightarrow[p]{} \boldsymbol{Q}_{XY}$, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1}$, $\widehat{\boldsymbol{Q}}_{Xe} \xrightarrow[p]{} 0$, and $\widehat{\beta} \xrightarrow[p]{} \beta$ as $n \to \infty$.

Theorem 7.1 states that the OLS estimator $\widehat{\beta}$ converges in probability to β as n increases and thus $\widehat{\beta}$ is consistent for β. In the stochastic order notation, Theorem 7.1 can be equivalently written as

$$\widehat{\beta} = \beta + o_p(1). \tag{7.4}$$

To illustrate the effect of sample size on the least squares estimator consider the least squares regression

$$\log(wage) = \beta_1 education + \beta_2 experience + \beta_3 experience^2 + \beta_4 + e.$$

We use the sample of 24,344 white men from the March 2009 CPS. We randomly sorted the observations and sequentially estimated the model by least squares starting with the first 5 observations and continuing until the full sample is used. The sequence of estimates are displayed in Figure 7.1. You can see how the least squares estimate changes with the sample size. As the number of observations increases it settles down to the full-sample estimate $\widehat{\beta}_1 = 0.114$.

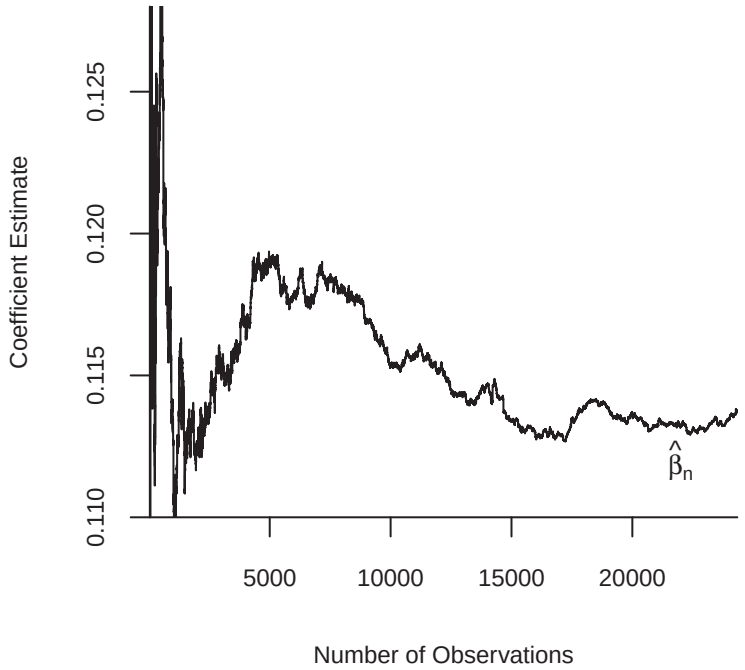


Figure 7.1: The Least-Squares Estimator as a Function of Sample Size

7.3 Asymptotic Normality

We started this chapter discussing the need for an approximation to the distribution of the OLS estimator $\widehat{\beta}$. In Section 7.2 we showed that $\widehat{\beta}$ converges in probability to β. Consistency is a good first step, but in itself does not describe the distribution of the estimator. In this section we derive an approximation typically called the **asymptotic distribution**.

The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied. The steps are as follows.

Take equation (7.3) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\right). \tag{7.5}$$

This shows that the normalized and centered estimator $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ is a function of the sample average $n^{-1}\sum_{i=1}^{n}X_iX_i'$ and the normalized sample average $n^{-1/2}\sum_{i=1}^{n}X_ie_i$.

The random pairs (Y_i,X_i) are i.i.d., meaning that they are independent across i and identically distributed. Any function of (Y_i,X_i) is also i.i.d. This includes $e_i=Y_i-X_i'\beta$ and the product X_ie_i. The latter is mean-zero ($\mathbb{E}[Xe]=0$) and has $k\times k$ covariance matrix

$$\Omega=\mathbb{E}\left[(Xe)(Xe)'\right]=\mathbb{E}\left[XX'e^2\right].$$

We show below that Ω has finite elements under a strengthening of Assumption 7.1. Since X_ie_i is i.i.d., mean zero, and finite variance, the central limit theorem (Theorem 6.3) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\underset{d}{\longrightarrow}\mathrm{N}(0,\Omega).$$

We state the required conditions here.

Assumption 7.2

1. The variables (Y_i,X_i), $i=1,...,n$, are i.i.d..
2. $\mathbb{E}\left[Y^4\right]<\infty$.
3. $\mathbb{E}\|X\|^4<\infty$.
4. $\boldsymbol{Q}_{XX}=\mathbb{E}\left[XX'\right]$ is positive definite.

Assumption 7.2 implies that $\Omega<\infty$. To see this, take its $j\ell^{th}$ element, $\mathbb{E}\left[X_jX_\ell e^2\right]$. Theorem 2.9.6 shows that $\mathbb{E}\left[e^4\right]<\infty$. By the expectation inequality (B.30), the $j\ell^{th}$ element of Ω is bounded by

$$\left|\mathbb{E}\left[X_jX_\ell e^2\right]\right|\leq\mathbb{E}\left|X_jX_\ell e^2\right|=\mathbb{E}\left[\left|X_j\right|\left|X_\ell\right|e^2\right].$$

By two applications of the Cauchy-Schwarz inequality (B.32), this is smaller than

$$\left(\mathbb{E}\left[X_j^2X_\ell^2\right]\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}\leq\left(\mathbb{E}\left[X_j^4\right]\right)^{1/4}\left(\mathbb{E}\left[X_\ell^4\right]\right)^{1/4}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}<\infty$$

where the finiteness holds under Assumption 7.2.2 and 7.2.3. Thus $\Omega < \infty$.

An alternative way to show that the elements of Ω are finite is by using a matrix norm $\|\cdot\|$ (See Appendix A.23). Then by the expectation inequality, the Cauchy-Schwarz inequality, Assumption 7.2.3, and $\mathbb{E}\left[e^4\right] < \infty$,

$$\|\Omega\| \leq \mathbb{E}\left\|XX'e^2\right\| = \mathbb{E}\left[\|X\|^2 e^2\right] \leq \left(\mathbb{E}\|X\|^4\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2} < \infty.$$

This is a more compact argument (often described as more *elegant*) but such manipulations should not be done without understanding the notation and the applicability of each step of the argument.

Regardless, the finiteness of the covariance matrix means that we can apply the multivariate CLT (Theorem 6.3).

Theorem 7.2 Assumption 7.2 implies that

$$\Omega < \infty \tag{7.6}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \underset{d}{\longrightarrow} \mathrm{N}(0, \Omega) \tag{7.7}$$

as $n \to \infty$.

Putting together (7.1), (7.5), and (7.7),

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}\mathrm{N}(0, \Omega) = \mathrm{N}\left(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\right)$$

as $n \to \infty$. The final equality follows from the property that linear combinations of normal vectors are also normal (Theorem 5.2).

We have derived the asymptotic normal approximation to the distribution of the least squares estimator.

Theorem 7.3 Asymptotic Normality of Least Squares Estimator
Under Assumption 7.2, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$$

where $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$, $\Omega = \mathbb{E}\left[XX'e^2\right]$, and

$$\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}. \tag{7.8}$$

In the stochastic order notation, Theorem 7.3 implies that $\widehat{\beta} = \beta + O_p(n^{-1/2})$ which is stronger than (7.4).

The matrix $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$. Consequently, $\boldsymbol{V}_\beta$ is often referred to as the **asymptotic covariance matrix** of $\widehat{\beta}$. The expression $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$ is called a **sandwich** form as the matrix Ω is sandwiched between two copies of $\boldsymbol{Q}_{XX}^{-1}$.

It is useful to compare the variance of the asymptotic distribution given in (7.8) and the finite-sample conditional variance in the CEF model as given in (4.10):

$$\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{7.9}$$

Notice that $\boldsymbol{V}_{\widehat{\beta}}$ is the exact conditional variance of $\widehat{\beta}$ and $\boldsymbol{V}_{\beta}$ is the asymptotic variance of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$. Thus $\boldsymbol{V}_{\beta}$ should be (roughly) n times as large as $\boldsymbol{V}_{\widehat{\beta}}$, or $\boldsymbol{V}_{\beta} \approx n\boldsymbol{V}_{\widehat{\beta}}$. Indeed, multiplying (7.9) by n and distributing we find

$$n\boldsymbol{V}_{\widehat{\beta}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

which looks like an estimator of $\boldsymbol{V}_{\beta}$. Indeed, as $n \to \infty$, $n\boldsymbol{V}_{\widehat{\beta}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$. The expression $\boldsymbol{V}_{\widehat{\beta}}$ is useful for practical inference (such as computation of standard errors and tests) as it is the variance of the estimator $\widehat{\beta}$, while $\boldsymbol{V}_{\beta}$ is useful for asymptotic theory as it is well defined in the limit as n goes to infinity. We will make use of both symbols and it will be advisable to adhere to this convention.

There is a special case where Ω and $\boldsymbol{V}_{\beta}$ simplify. Suppose that

$$\operatorname{cov}(XX', e^2) = 0. \tag{7.10}$$

Condition (7.10) holds in the homoskedastic linear regression model but is somewhat broader. Under (7.10) the asymptotic variance formulae simplify as

$$\begin{aligned}\Omega &= \mathbb{E}\left[XX'\right]\mathbb{E}\left[e^2\right] = \boldsymbol{Q}_{XX}\sigma^2 \\ \boldsymbol{V}_{\beta} &= \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1} = \boldsymbol{Q}_{XX}^{-1}\sigma^2 \equiv \boldsymbol{V}_{\beta}^{0}.\end{aligned} \tag{7.11}$$

In (7.11) we define $\boldsymbol{V}_{\beta}^{0} = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ whether (7.10) is true or false. When (7.10) is true then $\boldsymbol{V}_{\beta} = \boldsymbol{V}_{\beta}^{0}$, otherwise $\boldsymbol{V}_{\beta} \neq \boldsymbol{V}_{\beta}^{0}$. We call $\boldsymbol{V}_{\beta}^{0}$ the **homoskedastic asymptotic covariance matrix**.

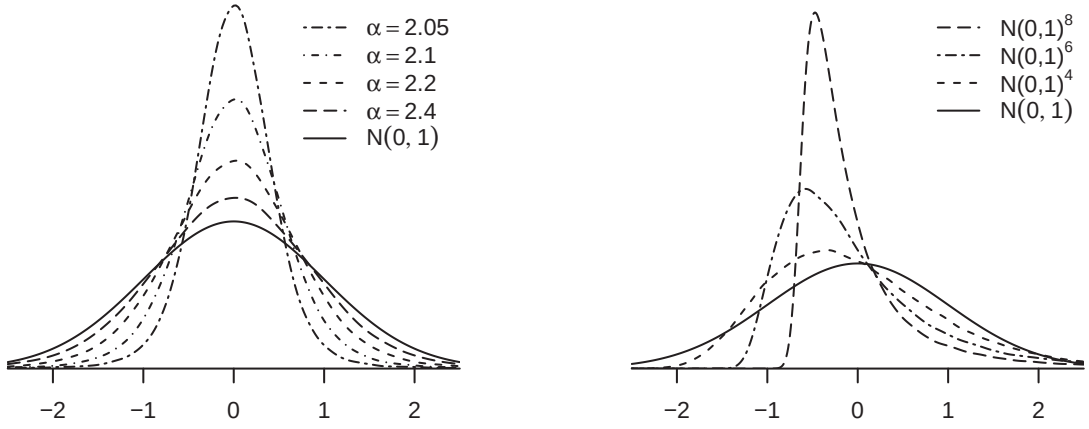
Theorem 7.3 states that the sampling distribution of the least squares estimator, after rescaling, is approximately normal when the sample size n is sufficiently large. This holds true for all joint distributions of (Y, X) which satisfy the conditions of Assumption 7.2. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$.

A difficulty is that for any fixed n the sampling distribution of $\widehat{\beta}$ can be arbitrarily far from the normal distribution. The normal approximation improves as n increases, but how large should n be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. The trouble is that no matter how large is the sample size, the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions. We illustrate this problem using a simulation. Let $Y = \beta_1 X + \beta_2 + e$ where X is $\mathrm{N}(0,1)$ and e is independent of X with the Double Pareto density $f(e) = \frac{\alpha}{2}|e|^{-\alpha-1}$, $|e| \geq 1$. If $\alpha > 2$ the error e has zero mean and variance $\alpha/(\alpha-2)$. As α approaches 2, however, its variance diverges to infinity. In this context the normalized least squares slope estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1-\beta_1\right)$ has the $\mathrm{N}(0,1)$ asymptotic distribution for any $\alpha > 2$. In Figure 7.2(a) we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1-\beta_1\right)$, setting $n = 100$ and varying the parameter α. For $\alpha = 3.0$ the density is very close to the $\mathrm{N}(0,1)$ density. As α diminishes the density changes significantly, concentrating most of the probability mass around zero.

Another example is shown in Figure 7.2(b). Here the model is $Y = \beta + e$ where

$$e = \frac{u^r - \mathbb{E}\left[u^r\right]}{\left(\mathbb{E}\left[u^{2r}\right] - \left(\mathbb{E}\left[u^r\right]\right)^2\right)^{1/2}} \tag{7.12}$$

and $u \sim \mathrm{N}(0,1)$. We show the sampling distribution of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ for $n = 100$, varying $r = 1$, 4, 6 and 8. As r increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figure 7.2 is that the $\mathrm{N}(0,1)$ asymptotic approximation is never guaranteed to be accurate.



(a) Double Pareto Error

(b) Error Process (7.12)

Figure 7.2: Density of Normalized OLS Estimator

7.4 Joint Distribution

Theorem 7.3 gives the joint asymptotic distribution of the coefficient estimators. We can use the result to study the covariance between the coefficient estimators. For simplicity, take the case of two regressors, no intercept, and homoskedastic error. Assume the regressors are mean zero, variance one, with correlation ρ. Then using the formula for inversion of a 2×2 matrix,

$$\boldsymbol{V}_{\beta}^{0} = \sigma^2 \boldsymbol{Q}_{XX}^{-1} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

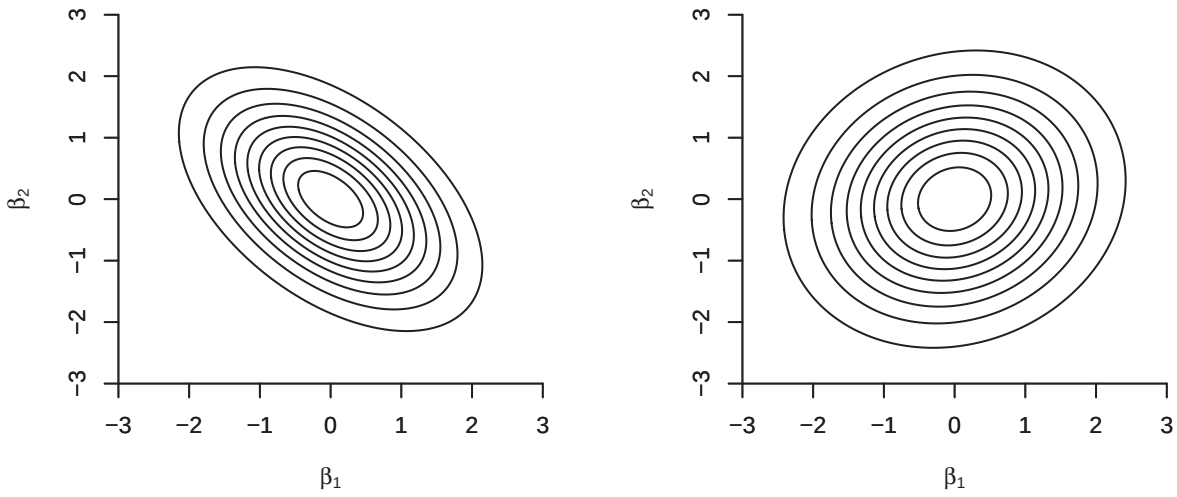
Thus if X_1 and X_2 are positively correlated ($\rho > 0$) then $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are negatively correlated (and vice-versa).

For illustration, Figure 7.3(a) displays the probability contours of the joint asymptotic distribution of $\widehat{\beta}_1 - \beta_1$ and $\widehat{\beta}_2 - \beta_2$ when $\beta_1 = \beta_2 = 0$ and $\rho = 0.5$. The coefficient estimators are negatively correlated because the regressors are positively correlated. This means that if $\widehat{\beta}_1$ is unusually negative, it is likely that $\widehat{\beta}_2$ is unusually positive, or conversely. It is also unlikely that we will observe both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ unusually large and of the same sign.

This finding that the correlation of the regressors is of opposite sign of the correlation of the coefficient estimates is sensitive to the assumption of homoskedasticity. If the errors are heteroskedastic then this relationship is not guaranteed.

This can be seen through a simple constructed example. Suppose that X_1 and X_2 only take the values $\{-1, +1\}$, symmetrically, with $\mathbb{P}[X_1 = X_2 = 1] = \mathbb{P}[X_1 = X_2 = -1] = 3/8$, and $\mathbb{P}[X_1 = 1, X_2 = -1] = \mathbb{P}[X_1 = -1, X_2 = 1] = 1/8$. You can check that the regressors are mean zero, unit variance and correlation 0.5, which is identical with the setting displayed in Figure 7.3(a).

Now suppose that the error is heteroskedastic. Specifically, suppose that $\mathbb{E}\left[e^2 \mid X_1 = X_2\right] = 5/4$ and $\mathbb{E}\left[e^2 \mid X_1 \neq X_2\right] = 1/4$. You can check that $\mathbb{E}\left[e^2\right] = 1$, $\mathbb{E}\left[X_1^2 e^2\right] = \mathbb{E}\left[X_2^2 e^2\right] = 1$ and $\mathbb{E}\left[X_1 X_2 e_i^2\right] = 7/8$. There-



(a) Homoskedastic Case

(b) Heteroskedastic Case

Figure 7.3: Contours of Joint Distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_2$

fore

$$\begin{aligned}\boldsymbol{V}_\beta &= \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1} \\ &= \frac{9}{16}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} 1 & \frac{7}{8} \\ \frac{7}{8} & 1 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \\ &= \frac{4}{3}\begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}.\end{aligned}$$

Thus the coefficient estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are positively correlated (their correlation is 1/4.) The joint probability contours of their asymptotic distribution is displayed in Figure 7.3(b). We can see how the two estimators are positively associated.

What we found through this example is that in the presence of heteroskedasticity there is no simple relationship between the correlation of the regressors and the correlation of the parameter estimators.

We can extend the above analysis to study the covariance between coefficient sub-vectors. For example, partitioning $X' = \left(X_1', X_2'\right)$ and $\beta' = \left(\beta_1', \beta_2'\right)$, we can write the general model as

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

and the coefficient estimates as $\widehat{\beta}' = \left(\widehat{\beta}_1', \widehat{\beta}_2'\right)$. Make the partitions

$$\boldsymbol{Q}_{XX} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}, \qquad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

From (2.43)

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

where $\boldsymbol{Q}_{11\cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot 1} = \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus when the error is homoskedastic

$$\operatorname{cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = -\sigma^2 \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}$$

which is a matrix generalization of the two-regressor case.

In general you can show that (Exercise 7.5)

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \tag{7.13}$$

where

$$\boldsymbol{V}_{11} = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\Omega_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{21} - \Omega_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{7.14}$$

$$\boldsymbol{V}_{21} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\Omega_{21} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{11} - \Omega_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{7.15}$$

$$\boldsymbol{V}_{22} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\Omega_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{12} - \Omega_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}_{22\cdot 1}^{-1}. \tag{7.16}$$

Unfortunately, these expressions are not easily interpretable.

7.5 Consistency of Error Variance Estimators

Using the methods of Section 7.2 we can show that the estimators $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{e}_i^2$ and $s^2 = (n-k)^{-1}\sum_{i=1}^n \widehat{e}_i^2$ are consistent for σ^2.

The trick is to write the residual $\widehat{e}_i$ as equal to the error e_i plus a deviation

$$\widehat{e}_i = Y_i - X_i'\widehat{\beta} = e_i - X_i'\left(\widehat{\beta} - \beta\right).$$

Thus the squared residual equals the squared error plus a deviation

$$\widehat{e}_i^2 = e_i^2 - 2e_i X_i'\left(\widehat{\beta} - \beta\right) + \left(\widehat{\beta} - \beta\right)' X_i X_i'\left(\widehat{\beta} - \beta\right). \tag{7.17}$$

So when we take the average of the squared residuals we obtain the average of the squared errors, plus two terms which are (hopefully) asymptotically negligible. This average is:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i X_i'\right)\left(\widehat{\beta} - \beta\right) + \left(\widehat{\beta} - \beta\right)'\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)\left(\widehat{\beta} - \beta\right). \tag{7.18}$$

The WLLN implies that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n e_i^2 &\longrightarrow_p \sigma^2 \\ \frac{1}{n}\sum_{i=1}^n e_i X_i' &\longrightarrow_p \mathbb{E}\left[eX'\right] = 0 \\ \frac{1}{n}\sum_{i=1}^n X_i X_i' &\longrightarrow_p \mathbb{E}\left[XX'\right] = \boldsymbol{Q}_{XX}. \end{aligned}$$

Theorem 7.1 shows that $\widehat{\beta} \longrightarrow_p \beta$. Hence (7.18) converges in probability to σ^2 as desired.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$ it follows that $s^2 = \left(\frac{n}{n-k}\right)\widehat{\sigma}^2 \longrightarrow_p \sigma^2$. Thus both estimators are consistent.

Theorem 7.4 Under Assumption 7.1, $\widehat{\sigma}^2 \xrightarrow[p]{} \sigma^2$ and $s^2 \xrightarrow[p]{} \sigma^2$ as $n \to \infty$.

7.6 Homoskedastic Covariance Matrix Estimation

Theorem 7.3 shows that $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ is asymptotically normal with asymptotic covariance matrix $\boldsymbol{V}_\beta$. For asymptotic inference (confidence intervals and tests) we need a consistent estimator of $\boldsymbol{V}_\beta$. Under homoskedasticity $\boldsymbol{V}_\beta$ simplifies to $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ and in this section we consider the simplified problem of estimating $\boldsymbol{V}_\beta^0$.

The standard moment estimator of $\boldsymbol{Q}_{XX}$ is $\widehat{\boldsymbol{Q}}_{XX}$ defined in (7.1) and thus an estimator for $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$. The standard estimator of σ^2 is the unbiased estimator s^2 defined in (4.31). Thus a natural plug-in estimator for $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ is $\widehat{\boldsymbol{V}}_\beta^0 = \widehat{\boldsymbol{Q}}_{XX}^{-1}s^2$.

Consistency of $\widehat{\boldsymbol{V}}_\beta^0$ for $\boldsymbol{V}_\beta^0$ follows from consistency of the moment estimators $\widehat{\boldsymbol{Q}}_{XX}$ and s^2 and an application of the continuous mapping theorem. Specifically, Theorem 7.1 established $\widehat{\boldsymbol{Q}}_{XX} \xrightarrow[p]{} \boldsymbol{Q}_{XX}$, and Theorem 7.4 established $s^2 \xrightarrow[p]{} \sigma^2$. The function $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ is a continuous function of $\boldsymbol{Q}_{XX}$ and σ^2 so long as $\boldsymbol{Q}_{XX} > 0$, which holds true under Assumption 7.1.4. It follows by the CMT that

$$\widehat{\boldsymbol{V}}_\beta^0 = \widehat{\boldsymbol{Q}}_{XX}^{-1}s^2 \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1}\sigma^2 = \boldsymbol{V}_\beta^0$$

so that $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for $\boldsymbol{V}_\beta^0$.

Theorem 7.5 Under Assumption 7.1, $\widehat{\boldsymbol{V}}_\beta^0 \xrightarrow[p]{} \boldsymbol{V}_\beta^0$ as $n \to \infty$.

It is instructive to notice that Theorem 7.5 does not require the assumption of homoskedasticity. That is, $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for $\boldsymbol{V}_\beta^0$ regardless if the regression is homoskedastic or heteroskedastic. However, $\boldsymbol{V}_\beta^0 = \boldsymbol{V}_\beta = \operatorname{avar}\left[\widehat{\beta}\right]$ only under homoskedasticity. Thus, in the general case $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for a well-defined but non-useful object.

7.7 Heteroskedastic Covariance Matrix Estimation

Theorems 7.3 established that the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ is $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$. We now consider estimation of this covariance matrix without imposing homoskedasticity. The standard approach is to use a plug-in estimator which replaces the unknowns with sample moments.

As described in the previous section a natural estimator for $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$ where $\widehat{\boldsymbol{Q}}_{XX}$ defined in (7.1).

The moment estimator for Ω is

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2,$$

leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_\beta^{\mathrm{HC0}} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}. \tag{7.19}$$

You can check that $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ where $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ is the HC0 covariance matrix estimator from (4.36).

As shown in Theorem 7.1, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}$, so we just need to verify the consistency of $\widehat{\Omega}$. The key is to replace the squared residual $\widehat{e}_i^2$ with the squared error e_i^2, and then show that the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned}\widehat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right).\end{aligned}$$

The first term is an average of the i.i.d. random variables $X_i X_i' e_i^2$, and therefore by the WLLN converges in probability to its expectation, namely,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \underset{p}{\longrightarrow} \mathbb{E}\left[XX' e^2\right] = \Omega.$$

Technically, this requires that Ω has finite elements, which was shown in (7.6).

To establish that $\widehat{\Omega}$ is consistent for Ω it remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \underset{p}{\longrightarrow} 0. \tag{7.20}$$

There are multiple ways to do this. A reasonably straightforward yet slightly tedious derivation is to start by applying the triangle inequality (B.16) using a matrix norm:

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\| X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 \left|\widehat{e}_i^2 - e_i^2\right|. \end{aligned} \tag{7.21}$$

Then recalling the expression for the squared residual (7.17), apply the triangle inequality (B.1) and then the Schwarz inequality (B.12) twice

$$\begin{aligned}\left|\widehat{e}_i^2 - e_i^2\right| &\le 2\left|e_i X_i' \left(\widehat{\beta} - \beta\right)\right| + \left(\widehat{\beta} - \beta\right)' X_i X_i' \left(\widehat{\beta} - \beta\right) \\ &= 2|e_i| \left|X_i' \left(\widehat{\beta} - \beta\right)\right| + \left|\left(\widehat{\beta} - \beta\right)' X_i\right|^2 \\ &\le 2|e_i| \|X_i\| \left\|\widehat{\beta} - \beta\right\| + \|X_i\|^2 \left\|\widehat{\beta} - \beta\right\|^2. \end{aligned} \tag{7.22}$$

Combining (7.21) and (7.22), we find

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le 2\left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^3 |e_i|\right) \left\|\widehat{\beta} - \beta\right\| + \left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^4\right) \left\|\widehat{\beta} - \beta\right\|^2 \\ &= o_p(1). \end{aligned} \tag{7.23}$$

The expression is $o_p(1)$ because $\left\|\widehat{\beta} - \beta\right\| \underset{p}{\longrightarrow} 0$ and both averages in parenthesis are averages of random variables with finite expectation under Assumption 7.2 (and are thus $O_p(1)$). Indeed, by Hölder's inequality (B.31)

$$\mathbb{E}\left[\|X\|^3 |e|\right] \le \left(\mathbb{E}\left[\left(\|X\|^3\right)^{4/3}\right]\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} = \left(\mathbb{E}\|X\|^4\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} < \infty.$$

We have established (7.20) as desired.

Theorem 7.6 Under Assumption 7.2, as $n \to \infty$, $\widehat{\Omega} \xrightarrow[p]{} \Omega$ and $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$.

For an alternative proof of this result, see Section 7.20.

7.8 Summary of Covariance Matrix Notation

The notation we have introduced may be somewhat confusing so it is helpful to write it down in one place.

The exact variance of $\widehat{\beta}$ (under the assumptions of the linear regression model) and the asymptotic variance of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ (under the more general assumptions of the linear projection model) are

$$\begin{aligned}\boldsymbol{V}_{\widehat{\beta}} &= \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \boldsymbol{V}_{\beta} &= \operatorname{avar}\left[\sqrt{n}\left(\widehat{\beta}-\beta\right)\right] = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}.\end{aligned}$$

The HC0 estimators of these two covariance matrices are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} &= \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}\end{aligned}$$

and satisfy the simple relationship $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$.

Similarly, under the assumption of homoskedasticity the exact and asymptotic variances simplify to

$$\begin{aligned}\boldsymbol{V}_{\widehat{\beta}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ \boldsymbol{V}_{\beta}^{0} &= \boldsymbol{Q}_{XX}^{-1}\sigma^2.\end{aligned}$$

Their standard estimators are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2 \\ \widehat{\boldsymbol{V}}_{\beta}^{0} &= \widehat{\boldsymbol{Q}}_{XX}^{-1} s^2\end{aligned}$$

which also satisfy the relationship $\widehat{\boldsymbol{V}}_{\beta}^{0} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$.

The exact formula and estimators are useful when constructing test statistics and standard errors. However, for theoretical purposes the asymptotic formula (variances and their estimates) are more useful as these retain non-generate limits as the sample sizes diverge. That is why both sets of notation are useful.

7.9 Alternative Covariance Matrix Estimators*

In Section 7.7 we introduced $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}}$ as an estimator of $\boldsymbol{V}_{\beta}$. $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}}$ is a scaled version of $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ from Section 4.14, where we also introduced the alternative HC1, HC2, and HC3 heteroskedasticity-robust covariance matrix estimators. We now discuss the consistency properties of these estimators.

To do so we introduce their scaled versions, e.g. $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC1}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC1}}$, $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC2}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC2}}$, and $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC3}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC3}}$. These are (alternative) estimators of the asymptotic covariance matrix $\boldsymbol{V}_{\beta}$.

First, consider $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}}$. Notice that $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC1}} = \frac{n}{n-k}\widehat{\boldsymbol{V}}_\beta^{\text{HC0}}$ where $\widehat{\boldsymbol{V}}_\beta^{\text{HC0}}$ was defined in (7.19) and shown consistent for $\boldsymbol{V}_\beta$ in Theorem 7.6. If k is fixed as $n \to \infty$, then $\frac{n}{n-k} \to 1$ and thus

$$\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} = (1 + o(1))\widehat{\boldsymbol{V}}_\beta^{\text{HC0}} \xrightarrow[p]{} \boldsymbol{V}_\beta.$$

Thus $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}}$ is consistent for $\boldsymbol{V}_\beta$.

The alternative estimators $\widehat{\boldsymbol{V}}_\beta^{\text{HC2}}$ and $\widehat{\boldsymbol{V}}_\beta^{\text{HC3}}$ take the form (7.19) but with $\widehat{\Omega}$ replaced by

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-2} X_i X_i' \widehat{e}_i^2$$

and

$$\overline{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1} X_i X_i' \widehat{e}_i^2,$$

respectively. To show that these estimators also consistent for $\boldsymbol{V}_\beta$ given $\widehat{\Omega} \xrightarrow[p]{} \Omega$ it is sufficient to show that the differences $\widetilde{\Omega} - \widehat{\Omega}$ and $\overline{\Omega} - \widehat{\Omega}$ converge in probability to zero as $n \to \infty$.

The trick is the fact that the leverage values are asymptotically negligible:

$$h_n^* = \max_{1\le i\le n} h_{ii} = o_p(1). \tag{7.24}$$

(See Theorem 7.17 in Section 7.21.) Then using the triangle inequality (B.16)

$$\begin{aligned}\left\|\overline{\Omega} - \widehat{\Omega}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|X_i X_i'\right\| \widehat{e}_i^2 \left|(1-h_{ii})^{-1} - 1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \widehat{e}_i^2\right)\left|\left(1-h_n^*\right)^{-1} - 1\right|.\end{aligned}$$

The sum in parenthesis can be shown to be $O_p(1)$ under Assumption 7.2 by the same argument as in in the proof of Theorem 7.6. (In fact, it can be shown to converge in probability to $\mathbb{E}\left[\|X\|^2 e^2\right]$.) The term in absolute values is $o_p(1)$ by (7.24). Thus the product is $o_p(1)$ which means that $\overline{\Omega} = \widehat{\Omega} + o_p(1) \xrightarrow[p]{} \Omega$.

Similarly,

$$\begin{aligned}\left\|\widetilde{\Omega} - \widehat{\Omega}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|X_i X_i'\right\| \widehat{e}_i^2 \left|(1-h_{ii})^{-2} - 1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \widehat{e}_i^2\right)\left|\left(1-h_n^*\right)^{-2} - 1\right| \\ &= o_p(1).\end{aligned}$$

Theorem 7.7 Under Assumption 7.2, as $n \to \infty$, $\widetilde{\Omega} \xrightarrow[p]{} \Omega$, $\overline{\Omega} \xrightarrow[p]{} \Omega$, $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} \xrightarrow[p]{} \boldsymbol{V}_\beta$, $\widehat{\boldsymbol{V}}_\beta^{\text{HC2}} \xrightarrow[p]{} \boldsymbol{V}_\beta$, and $\widehat{\boldsymbol{V}}_\beta^{\text{HC3}} \xrightarrow[p]{} \boldsymbol{V}_\beta$.

Theorem 7.7 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix.

To simplify notation, for the remainder of the chapter we will use the notation $\widehat{\boldsymbol{V}}_\beta$ and $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ to refer to any of the heteroskedasticity-consistent covariance matrix estimators HC0, HC1, HC2, and HC3, as they all have the same asymptotic limits.

7.10 Functions of Parameters

In most serious applications a researcher is actually interested in a specific transformation of the coefficient vector $\beta = (\beta_1, ..., \beta_k)$. For example, the researcher may be interested in a single coefficient β_j or a ratio β_j/β_l. More generally, interest may focus on a quantity such as consumer surplus which could be a complicated function of the coefficients. In any of these cases we can write the parameter of interest θ as a function of the coefficients, e.g. $\theta = r(\beta)$ for some function $r : \mathbb{R}^k \to \mathbb{R}^q$. The estimate of θ is

$$\widehat{\theta} = r(\widehat{\beta}).$$

By the continuous mapping theorem (Theorem 6.6) and the fact $\widehat{\beta} \xrightarrow[p]{} \beta$ we can deduce that $\widehat{\theta}$ is consistent for θ if the function $r(\cdot)$ is continuous.

Theorem 7.8 Under Assumption 7.1, if $r(\beta)$ is continuous at the true value of β then as $n \to \infty$, $\widehat{\theta} \xrightarrow[p]{} \theta$.

Furthermore, if the transformation is sufficiently smooth, by the Delta Method (Theorem 6.8) we can show that $\widehat{\theta}$ is asymptotically normal.

Assumption 7.3 $r(\beta) : \mathbb{R}^k \to \mathbb{R}^q$ is continuously differentiable at the true value of β and $\boldsymbol{R} = \frac{\partial}{\partial \beta} r(\beta)'$ has rank q.

Theorem 7.9 Asymptotic Distribution of Functions of Parameters
Under Assumptions 7.2 and 7.3, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow[d]{} \mathrm{N}(0, \boldsymbol{V}_\theta) \tag{7.25}$$

where $\boldsymbol{V}_\theta = \boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R}$.

In many cases the function $r(\beta)$ is linear:

$$r(\beta) = \boldsymbol{R}'\beta$$

for some $k \times q$ matrix $\boldsymbol{R}$. In particular if $\boldsymbol{R}$ is a "selector matrix"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix}$$

then we can partition $\beta = (\beta_1', \beta_2')'$ so that $\boldsymbol{R}'\beta = \beta_1$. Then

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & 0 \end{pmatrix} \boldsymbol{V}_\beta \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix} = \boldsymbol{V}_{11},$$

the upper-left sub-matrix of $\boldsymbol{V}_{11}$ given in (7.14). In this case (7.25) states that

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_{11}\right).$$

That is, subsets of $\widehat{\beta}$ are approximately normal with variances given by the conformable subcomponents of $\boldsymbol{V}$.

To illustrate the case of a nonlinear transformation take the example $\theta = \beta_j/\beta_l$ for $j \neq l$. Then

$$\boldsymbol{R} = \frac{\partial}{\partial \beta} r(\beta) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_j}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_k}\left(\beta_j/\beta_l\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \tag{7.26}$$

so

$$\boldsymbol{V}_\theta = \boldsymbol{V}_{jj}/\beta_l^2 + \boldsymbol{V}_{ll}\beta_j^2/\beta_l^4 - 2\boldsymbol{V}_{jl}\beta_j/\beta_l^3$$

where $\boldsymbol{V}_{ab}$ denotes the ab^{th} element of $\boldsymbol{V}_\beta$.

For inference we need an estimator of the asymptotic covariance matrix $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}$. For this it is typical to use the plug-in estimator

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \beta} r(\widehat{\beta})'. \tag{7.27}$$

The derivative in (7.27) may be calculated analytically or numerically. By analytically, we mean working out the formula for the derivative and replacing the unknowns by point estimates. For example, if $\theta = \beta_j/\beta_l$ then $\frac{\partial}{\partial \beta} r(\beta)$ is (7.26). However in some cases the function $r(\beta)$ may be extremely complicated and a formula for the analytic derivative may not be easily available. In this case numerical differentiation may be preferable. Let $\delta_l = (0 \cdots 1 \cdots 0)'$ be the unit vector with the "1" in the l^{th} place. The jl^{th} element of a numerical derivative $\widehat{\boldsymbol{R}}$ is

$$\widehat{\boldsymbol{R}}_{jl} = \frac{r_j(\widehat{\beta} + \delta_l \epsilon) - r_j(\widehat{\beta})}{\epsilon}$$

for some small ϵ.

The estimator of $\boldsymbol{V}_\theta$ is

$$\widehat{\boldsymbol{V}}_\theta = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}. \tag{7.28}$$

Alternatively, the homoskedastic covariance matrix estimator could be used leading to a homoskedastic covariance matrix estimator for θ.

$$\widehat{\boldsymbol{V}}_\theta^0 = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta^0\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{R}}s^2. \tag{7.29}$$

Given (7.27), (7.28) and (7.29) are simple to calculate using matrix operations.

As the primary justification for $\widehat{\boldsymbol{V}}_\theta$ is the asymptotic approximation (7.25), $\widehat{\boldsymbol{V}}_\theta$ is often called an **asymptotic covariance matrix estimator**.

The estimator $\widehat{\boldsymbol{V}}_\theta$ is consistent for $\boldsymbol{V}_\theta$ under the conditions of Theorem 7.9 because $\widehat{\boldsymbol{V}}_\beta \underset{p}{\longrightarrow} \boldsymbol{V}_\beta$ by Theorem 7.6 and

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \beta} r(\widehat{\beta})' \underset{p}{\longrightarrow} \frac{\partial}{\partial \beta} r(\beta)' = \boldsymbol{R}$$

because $\widehat{\beta} \underset{p}{\longrightarrow} \beta$ and the function $\frac{\partial}{\partial \beta} r(\beta)'$ is continuous in β.

Theorem 7.10 Under Assumptions 7.2 and 7.3, as $n \to \infty$, $\widehat{\boldsymbol{V}}_\theta \xrightarrow[p]{} \boldsymbol{V}_\theta$.

Theorem 7.10 shows that $\widehat{\boldsymbol{V}}_\theta$ is consistent for $\boldsymbol{V}_\theta$ and thus may be used for asymptotic inference. In practice we may set

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}} = n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}} \tag{7.30}$$

as an estimator of the variance of $\widehat{\theta}$.

7.11 Asymptotic Standard Errors

As described in Section 4.15, a standard error is an estimator of the standard deviation of the distribution of an estimator. Thus if $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ is an estimator of the covariance matrix of $\widehat{\beta}$ then standard errors are the square roots of the diagonal elements of this matrix. These take the form

$$s(\widehat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\beta}}\right]_{jj}}.$$

Standard errors for $\widehat{\theta}$ are constructed similarly. Supposing that $\theta = h(\beta)$ is real-valued then the standard error for $\widehat{\theta}$ is the square root of (7.30)

$$s(\widehat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}}} = \sqrt{n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}}.$$

When the justification is based on asymptotic theory we call $s(\widehat{\beta}_j)$ or $s(\widehat{\theta})$ an **asymptotic standard error** for $\widehat{\beta}_j$ or $\widehat{\theta}$. When reporting your results it is good practice to report standard errors for each reported estimate and this includes functions and transformations of your parameter estimates. This helps users of the work (including yourself) assess the estimation precision.

We illustrate using the log wage regression

$$\log(\mathit{wage}) = \beta_1 \, \mathit{education} + \beta_2 \, \mathit{experience} + \beta_3 \, \mathit{experience}^2/100 + \beta_4 + e.$$

Consider the following three parameters of interest.

1. Percentage return to education:

$$\theta_1 = 100\beta_1$$

   (100 times the partial derivative of the conditional expectation of log(wage) with respect to education.)

2. Percentage return to experience for individuals with 10 years of experience:

$$\theta_2 = 100\beta_2 + 20\beta_3$$

   (100 times the partial derivative of the conditional expectation of log wages with respect to experience, evaluated at experience= 10.)

3. Experience level which maximizes expected log wages:

$$\theta_3 = -50\beta_2/\beta_3$$

(The level of experience at which the partial derivative of the conditional expectation of log(wage) with respect to experience equals 0.)

The 4×1 vector $\boldsymbol{R}$ for these three parameters is

$$\boldsymbol{R} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 100 \\ 20 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -50/\beta_3 \\ 50\beta_2/\beta_3^2 \\ 0 \end{pmatrix},$$

respectively.

We use the subsample of married Black women (all experience levels) which has 982 observations. The point estimates and standard errors are

$$\widehat{\log(wage)} = \underset{(0.008)}{0.118}\ education + \underset{(0.006)}{0.016}\ experience - \underset{(0.012)}{0.022}\ experience^2/100 + \underset{(0.157)}{0.947}\ . \tag{7.31}$$

The standard errors are the square roots of the HC2 covariance matrix estimate

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \times 10^{-4}. \tag{7.32}$$

We calculate that

$$\widehat{\theta}_1 = 100\widehat{\beta}_1 = 100 \times 0.118 = 11.8$$

$$s(\widehat{\theta}_1) = \sqrt{100^2 \times 0.632 \times 10^{-4}} = 0.8$$

$$\widehat{\theta}_2 = 100\widehat{\beta}_2 + 20\widehat{\beta}_3 = 100 \times 0.016 - 20 \times 0.022 = 1.16$$

$$s(\widehat{\theta}_2) = \sqrt{\begin{pmatrix} 100 & 20 \end{pmatrix}\begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix}\begin{pmatrix} 100 \\ 20 \end{pmatrix} \times 10^{-4}} = 0.55$$

$$\widehat{\theta}_3 = -50\widehat{\beta}_2/\widehat{\beta}_3 = 50 \times 0.016/0.022 = 35.2$$

$$s(\widehat{\theta}_3) = \sqrt{\begin{pmatrix} -50/\widehat{\beta}_3 & 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix}\begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix}\begin{pmatrix} -50/\widehat{\beta}_3 \\ 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix} \times 10^{-4}} = 7.0.$$

The calculations show that the estimate of the percentage return to education is 12% per year with a standard error of 0.8. The estimate of the percentage return to experience for those with 10 years of experience is 1.2% per year with a standard error of 0.6. The estimate of the experience level which maximizes expected log wages is 35 years with a standard error of 7.

In Stata the `nlcom` command can be used after estimation to perform the same calculations. To illustrate, after estimation of (7.31) use the commands given below. In each case, Stata reports the coefficient estimate, asymptotic standard error, and 95% confidence interval.

Stata Commands

```
nlcom 100*_b[education]
nlcom 100*_b[experience]+20*_b[exp2]
nlcom -50*_b[experience]/_b[exp2]
```

7.12 t-statistic

Let $\theta = r(\beta) : \mathbb{R}^k \to \mathbb{R}$ be a parameter of interest, $\widehat{\theta}$ its estimator, and $s(\widehat{\theta})$ its asymptotic standard error. Consider the statistic

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})}. \tag{7.33}$$

Different writers call (7.33) a **t-statistic**, a **t-ratio**, a **z-statistic**, or a **studentized statistic**, sometimes using the different labels to distinguish between finite-sample and asymptotic inference. As the statistics themselves are always (7.33) we won't make this distinction, and will simply refer to $T(\theta)$ as a t-statistic or a t-ratio. We also often suppress the parameter dependence, writing it as T. The t-statistic is a function of the estimator, its standard error, and the parameter.

By Theorems 7.9 and 7.10, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \underset{d}{\longrightarrow} \mathrm{N}(0, V_\theta)$ and $\widehat{V}_\theta \underset{p}{\longrightarrow} V_\theta$. Thus

$$\begin{aligned} T(\theta) &= \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \\ &= \frac{\sqrt{n}\left(\widehat{\theta} - \theta\right)}{\sqrt{\widehat{V}_\theta}} \\ &\underset{d}{\longrightarrow} \frac{\mathrm{N}(0, V_\theta)}{\sqrt{V_\theta}} \\ &= Z \sim \mathrm{N}(0,1). \end{aligned}$$

The last equality is the property that affine functions of normal variables are normal (Theorem 5.2).

This calculation requires that $V_\theta > 0$, otherwise the continuous mapping theorem cannot be employed. In practice this is an innocuous requirement as it only excludes degenerate sampling distributions. Formally we add the following assumption.

Assumption 7.4 $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R} > 0$.

Assumption 7.4 states that $\boldsymbol{V}_\theta$ is positive definite. Since $\boldsymbol{R}$ is full rank under Assumption 7.3 a sufficient condition is that $\boldsymbol{V}_\beta > 0$. Since $\boldsymbol{Q}_{XX} > 0$ a sufficient condition is $\Omega > 0$. Thus Assumption 7.4 could be replaced by the assumption $\Omega > 0$. Assumption 7.4 is weaker so this is what we use.

Thus the asymptotic distribution of the t-ratio $T(\theta)$ is standard normal. Since this distribution does not depend on the parameters we say that $T(\theta)$ is **asymptotically pivotal**. In finite samples $T(\theta)$ is not necessarily pivotal but the property means that the dependence on unknowns diminishes as n increases.

It is also useful to consider the distribution of the **absolute t-ratio** $|T(\theta)|$. Since $T(\theta) \underset{d}{\longrightarrow} Z$ the continuous mapping theorem yields $|T(\theta)| \underset{d}{\longrightarrow} |Z|$. Letting $\Phi(u) = \mathbb{P}[Z \leq u]$ denote the standard normal distribution function we calculate that the distribution of $|Z|$ is

$$\begin{aligned}\mathbb{P}[|Z| \leq u] &= \mathbb{P}[-u \leq Z \leq u] \\ &= \mathbb{P}[Z \leq u] - \mathbb{P}[Z < -u] \\ &= \Phi(u) - \Phi(-u) \\ &= 2\Phi(u) - 1. \end{aligned} \tag{7.34}$$

Theorem 7.11 Under Assumptions 7.2, 7.3, and 7.4, $T(\theta) \underset{d}{\longrightarrow} Z \sim \mathrm{N}(0,1)$ and $|T(\theta)| \underset{d}{\longrightarrow} |Z|$.

The asymptotic normality of Theorem 7.11 is used to justify confidence intervals and tests for the parameters.

7.13 Confidence Intervals

The estimator $\widehat{\theta}$ is a **point estimator** for θ, meaning that $\widehat{\theta}$ is a single value in $\mathbb{R}^q$. A broader concept is a **set estimator** $\widehat{C}$ which is a collection of values in $\mathbb{R}^q$. When the parameter θ is real-valued then it is common to focus on sets of the form $\widehat{C} = [\widehat{L}, \widehat{U}]$ which is called an **interval estimator** for θ.

An interval estimator $\widehat{C}$ is a function of the data and hence is random. The **coverage probability** of the interval $\widehat{C} = [\widehat{L}, \widehat{U}]$ is $\mathbb{P}\left[\theta \in \widehat{C}\right]$. The randomness comes from $\widehat{C}$ as the parameter θ is treated as fixed. In Section 5.10 we introduced confidence intervals for the normal regression model which used the finite sample distribution of the t-statistic. When we are outside the normal regression model we cannot rely on the exact normal distribution theory but instead use asymptotic approximations. A benefit is that we can construct confidence intervals for general parameters of interest θ not just regression coefficients.

An interval estimator $\widehat{C}$ is called a **confidence interval** when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%. $\widehat{C}$ is called a $1-\alpha$ confidence interval if $\inf_\theta \mathbb{P}_\theta\left[\theta \in \widehat{C}\right] = 1-\alpha$.

When $\widehat{\theta}$ is asymptotically normal with standard error $s(\widehat{\theta})$ the conventional confidence interval for θ takes the form

$$\widehat{C} = \left[\widehat{\theta} - c \times s(\widehat{\theta}), \quad \widehat{\theta} + c \times s(\widehat{\theta})\right] \tag{7.35}$$

where c equals the $1-\alpha$ quantile of the distribution of $|Z|$. Using (7.34) we calculate that c is equivalently the $1-\alpha/2$ quantile of the standard normal distribution. Thus, c solves

$$2\Phi(c) - 1 = 1 - \alpha.$$

This can be computed by, for example, `norminv(1-α/2)` in MATLAB. The confidence interval (7.35) is symmetric about the point estimator $\widehat{\theta}$ and its length is proportional to the standard error $s(\widehat{\theta})$.

Equivalently, (7.35) is the set of parameter values for θ such that the t-statistic $T(\theta)$ is smaller (in absolute value) than c, that is

$$\widehat{C} = \{\theta : |T(\theta)| \leq c\} = \left\{\theta : -c \leq \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \leq c\right\}.$$

The coverage probability of this confidence interval is

$$\mathbb{P}\left[\theta \in \widehat{C}\right] = \mathbb{P}\left[|T(\theta)| \leq c\right] \to \mathbb{P}\left[|Z| \leq c\right] = 1 - \alpha$$

where the limit is taken as $n \to \infty$, and holds because $T(\theta)$ is asymptotically $|Z|$ by Theorem 7.11. We call the limit the **asymptotic coverage probability** and call $\widehat{C}$ an asymptotic $1-\alpha$% confidence interval for θ. Since the t-ratio is asymptotically pivotal the asymptotic coverage probability is independent of the parameter θ.

It is useful to contrast the confidence interval (7.35) with (5.8) for the normal regression model. They are similar but there are differences. The normal regression interval (5.8) only applies to regression coefficients β not to functions θ of the coefficients. The normal interval (5.8) also is constructed with the homoskedastic standard error, while (7.35) can be constructed with a heteroskedastic-robust standard error. Furthermore, the constants c in (5.8) are calculated using the student t distribution, while c in (7.35) are calculated using the normal distribution. The difference between the student t and normal values are typically small in practice (since sample sizes are large in typical economic applications). However, since the student t values are larger it results in slightly larger confidence intervals which is reasonable. (A practical rule of thumb is that if the sample sizes are sufficiently small that it makes a difference then neither (5.8) nor (7.35) should be trusted.) Despite these differences the coincidence of the intervals means that inference on regression coefficients is generally robust to using either the exact normal sampling assumption or the asymptotic large sample approximation, at least in large samples.

Stata by default reports 95% confidence intervals for each coefficient where the critical values c are calculated using the t_{n-k} distribution. This is done for all standard error methods even though it is only exact for homoskedastic standard errors and under normality.

The standard coverage probability for confidence intervals is 95%, leading to the choice $c = 1.96$ for the constant in (7.35). Rounding 1.96 to 2, we obtain the most commonly used confidence interval in applied econometric practice

$$\widehat{C} = \left[\widehat{\theta} - 2s(\widehat{\theta}),\quad \widehat{\theta} + 2s(\widehat{\theta})\right].$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval $\widehat{C}$ is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is an asymptotic 95.4% interval due to the substitution of 2.0 for 1.96 but this distinction is overly precise.)

Theorem 7.12 Under Assumptions 7.2, 7.3 and 7.4, for $\widehat{C}$ defined in (7.35) with $c = \Phi^{-1}(1-\alpha/2)$, $\mathbb{P}\left[\theta \in \widehat{C}\right] \to 1-\alpha$. For $c = 1.96$, $\mathbb{P}\left[\theta \in \widehat{C}\right] \to 0.95$.

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results look at the estimated coefficient estimates and the standard errors. For a parameter of interest compute the confidence interval $\widehat{C}$ and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about θ then do not jump to a conclusion about θ based on the point estimate alone.

For illustration, consider the three examples presented in Section 7.11 based on the log wage regression for married Black women.

Percentage return to education. A 95% asymptotic confidence interval is $11.8 \pm 1.96 \times 0.8 = [10.2, 13.3]$. This is reasonably tight.

Percentage return to experience (per year) for individuals with 10 years experience. A 90% asymptotic confidence interval is $1.1 \pm 1.645 \times 0.4 = [0.5, 1.8]$. The interval is positive but broad. This indicates that the return to experience is positive, but of uncertain magnitude.

Experience level which maximizes expected log wages. An 80% asymptotic confidence interval is $35 \pm 1.28 \times 7 = [26, 44]$. This is rather imprecise, indicating that the estimates are not very informative regarding this parameter.

7.14 Regression Intervals

In the linear regression model the conditional expectation of Y given $X = x$ is

$$m(x) = \mathbb{E}[Y \mid X = x] = x'\beta.$$

In some cases we want to estimate $m(x)$ at a particular point x. Notice that this is a linear function of β. Letting $r(\beta) = x'\beta$ and $\theta = r(\beta)$ we see that $\widehat{m}(x) = \widehat{\theta} = x'\widehat{\beta}$ and $\boldsymbol{R} = x$ so $s(\widehat{\theta}) = \sqrt{x'\widehat{\boldsymbol{V}}_{\widehat{\beta}}x}$. Thus an asymptotic 95% confidence interval for $m(x)$ is

$$\left[x'\widehat{\beta} \pm 1.96\sqrt{x'\widehat{\boldsymbol{V}}_{\widehat{\beta}}x}\right].$$

It is interesting to observe that if this is viewed as a function of x the width of the confidence interval is dependent on x.

To illustrate we return to the log wage regression (3.12) of Section 3.7. The estimated regression equation is

$$\widehat{\log(wage)} = x'\widehat{\beta} = 0.155x + 0.698$$

where $x =$*education*. The covariance matrix estimate from (4.43) is

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}.$$

Thus the 95% confidence interval for the regression is

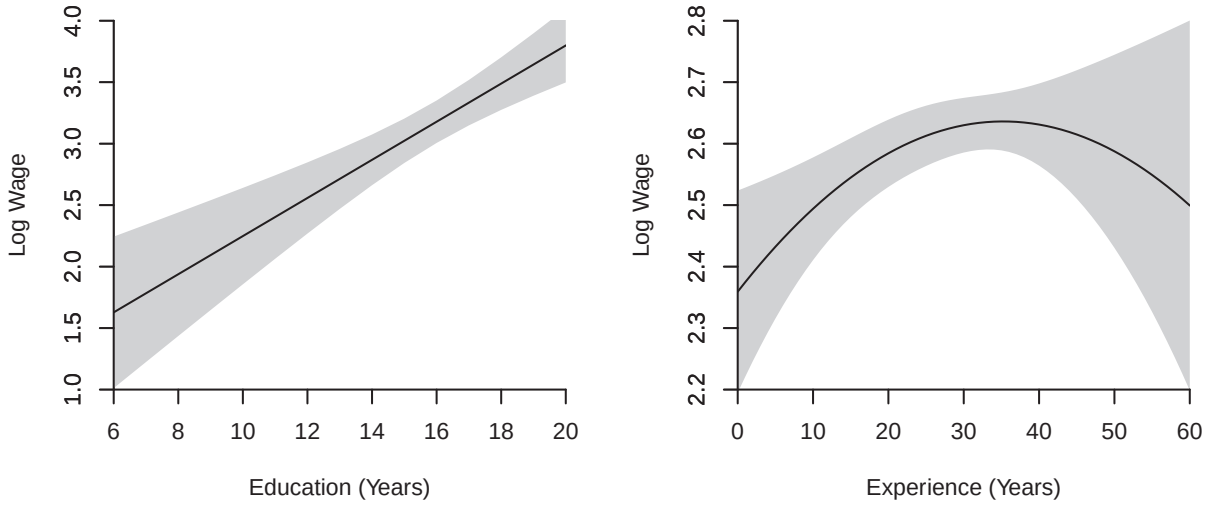
$$0.155x + 0.698 \pm 1.96\sqrt{0.001x^2 - 0.030x + 0.243}.$$

The estimated regression and 95% intervals are shown in Figure 7.4(a). Notice that the confidence bands take a hyperbolic shape. This means that the regression line is less precisely estimated for large and small values of *education*.

Plots of the estimated regression line and confidence intervals are especially useful when the regression includes nonlinear terms. To illustrate consider the log wage regression (7.31) which includes experience and its square and covariance matrix estimate (7.32). We are interested in plotting the regression estimate and regression intervals as a function of *experience*. Since the regression also includes *education*, to plot the estimates in a simple graph we fix *education* at a specific value. We select *education*=12. This only affects the level of the estimated regression since *education* enters without an interaction. Define the points of evaluation

$$z(x) = \begin{pmatrix} 12 \\ x \\ x^2/100 \\ 1 \end{pmatrix}$$

where $x =$*experience*.



(a) Wage on Education

(b) Wage on Experience

Figure 7.4: Regression Intervals

The 95% regression interval for *education*=12 as a function of x =*experience* is

$$
\begin{aligned}
&0.118 \times 12 \ + 0.016\, x - 0.022\, x^2/100 + 0.947 \\
&\pm 1.96 \sqrt{z(x)' \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} z(x) \times 10^{-4}} \\
&= 0.016\, x - .00022\, x^2 + 2.36 \\
&\pm 0.0196 \sqrt{70.608 - 9.356\, x + 0.54428\, x^2 - 0.01462\, x^3 + 0.000148\, x^4}.
\end{aligned}
$$

The estimated regression and 95% intervals are shown in Figure 7.4(b). The regression interval widens greatly for small and large values of experience indicating considerable uncertainty about the effect of experience on mean wages for this population. The confidence bands take a more complicated shape than in Figure 7.4(a) due to the nonlinear specification.

7.15 Forecast Intervals

Suppose we are given a value of the regressor vector X_{n+1} for an individual outside the sample and we want to forecast (guess) Y_{n+1} for this individual. This is equivalent to forecasting Y_{n+1} given $X_{n+1} = x$ which will generally be a function of x. A reasonable forecasting rule is the conditional expectation $m(x)$ as it is the mean-square minimizing forecast. A point forecast is the estimated conditional expectation $\widehat{m}(x) = x'\widehat{\beta}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\widehat{e}_{n+1} = Y_{n+1} - \widehat{m}(x) = e_{n+1} - x'\left(\widehat{\beta} - \beta\right)$. As the out-of-sample error e_{n+1} is inde-

pendent of the in-sample estimator $\widehat{\beta}$ this has conditional variance

$$\begin{aligned}\mathbb{E}\left[\widehat{e}_{n+1}^2 \mid X_{n+1}=x\right] &= \mathbb{E}\left[e_{n+1}^2 - 2x'\left(\widehat{\beta}-\beta\right)e_{n+1} + x'\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'x \mid X_{n+1}=x\right] \\ &= \mathbb{E}\left[e_{n+1}^2 \mid X_{n+1}=x\right] + x'\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'\right]x \\ &= \sigma^2(x) + x'\boldsymbol{V}_{\widehat{\beta}}x. \end{aligned} \tag{7.36}$$

Under homoskedasticity, $\mathbb{E}\left[e_{n+1}^2 \mid X_{n+1}\right] = \sigma^2$. In this case a simple estimator of (7.36) is $\widehat{\sigma}^2 + x'\boldsymbol{V}_{\widehat{\beta}}x$ so a standard error for the forecast is $\widehat{s}(x) = \sqrt{\widehat{\sigma}^2 + x'\boldsymbol{V}_{\widehat{\beta}}x}$. Notice that this is different from the standard error for the conditional expectation.

The conventional 95% forecast interval for Y_{n+1} uses a normal approximation and equals $\left[x'\widehat{\beta} \pm 2\widehat{s}(x)\right]$. It is difficult, however, to fully justify this choice. It would be correct if we have a normal approximation to the ratio

$$\frac{e_{n+1} - x'\left(\widehat{\beta}-\beta\right)}{\widehat{s}(x)}.$$

The difficulty is that the equation error e_{n+1} is generally non-normal and asymptotic theory cannot be applied to a single observation. The only special exception is the case where e_{n+1} has the exact distribution $\mathrm{N}(0,\sigma^2)$ which is generally invalid.

An accurate forecast interval would use the conditional distribution of e_{n+1} given $X_{n+1} = x$, which is more challenging to estimate. Due to this difficulty many applied forecasters use the simple approximate interval $\left[x'\widehat{\beta} \pm 2\widehat{s}(x)\right]$ despite the lack of a convincing justification.

7.16 Wald Statistic

Let $\theta = r(\beta) : \mathbb{R}^k \to \mathbb{R}^q$ be any parameter vector of interest, $\widehat{\theta}$ its estimator, and $\widehat{\boldsymbol{V}}_{\widehat{\theta}}$ its covariance matrix estimator. Consider the quadratic form

$$W(\theta) = \left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1}\left(\widehat{\theta}-\theta\right) = n\left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\theta}^{-1}\left(\widehat{\theta}-\theta\right). \tag{7.37}$$

where $\widehat{\boldsymbol{V}}_{\theta} = n\widehat{\boldsymbol{V}}_{\widehat{\theta}}$. When $q = 1$, then $W(\theta) = T(\theta)^2$ is the square of the t-ratio. When $q > 1$, $W(\theta)$ is typically called a **Wald statistic** as it was proposed by Wald (1943). We are interested in its sampling distribution.

The asymptotic distribution of $W(\theta)$ is simple to derive given Theorem 7.9 and Theorem 7.10. They show that $\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} Z \sim \mathrm{N}(0, \boldsymbol{V}_{\theta})$ and $\widehat{\boldsymbol{V}}_{\theta} \underset{p}{\longrightarrow} \boldsymbol{V}_{\theta}$. It follows that

$$W(\theta) = \sqrt{n}\left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\theta}^{-1}\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} Z'\boldsymbol{V}_{\theta}^{-1}Z$$

a quadratic in the normal random vector Z. As shown in Theorem 5.3.5 the distribution of this quadratic form is χ_q^2, a chi-square random variable with q degrees of freedom.

Theorem 7.13 Under Assumptions 7.2, 7.3 and 7.4, as $n \to \infty$, $W(\theta) \underset{d}{\longrightarrow} \chi_q^2$.

Theorem 7.13 is used to justify multivariate confidence regions and multivariate hypothesis tests.

7.17 Homoskedastic Wald Statistic

Under the conditional homoskedasticity assumption $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ we can construct the Wald statistic using the homoskedastic covariance matrix estimator $\widehat{\boldsymbol{V}}_\theta^0$ defined in (7.29). This yields a homoskedastic Wald statistic

$$W^0(\theta) = \left(\widehat{\theta} - \theta\right)' \left(\widehat{\boldsymbol{V}}_{\widehat{\theta}}^0\right)^{-1} \left(\widehat{\theta} - \theta\right) = n\left(\widehat{\theta} - \theta\right)' \left(\widehat{\boldsymbol{V}}_\theta^0\right)^{-1} \left(\widehat{\theta} - \theta\right). \tag{7.38}$$

Under the assumption of conditional homoskedasticity it has the same asymptotic distribution as $W(\theta)$.

Theorem 7.14 Under Assumptions 7.2, 7.3, and $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2 > 0$, as $n \to \infty$, $W^0(\theta) \xrightarrow[d]{} \chi_q^2$.

7.18 Confidence Regions

A confidence region $\widehat{C}$ is a set estimator for $\theta \in \mathbb{R}^q$ when $q > 1$. A confidence region $\widehat{C}$ is a set in $\mathbb{R}^q$ intended to cover the true parameter value with a pre-selected probability $1-\alpha$. Thus an ideal confidence region has the coverage probability $\mathbb{P}\left[\theta \in \widehat{C}\right] = 1 - \alpha$. In practice it is typically not possible to construct a region with exact coverage but we can calculate its asymptotic coverage.

When the parameter estimator satisfies the conditions of Theorem 7.13 a good choice for a confidence region is the ellipse

$$\widehat{C} = \{\theta : W(\theta) \leq c_{1-\alpha}\}$$

with $c_{1-\alpha}$ the $1-\alpha$ quantile of the χ_q^2 distribution. (Thus $F_q(c_{1-\alpha}) = 1-\alpha$.) It can be computed by, for example, `chi2inv(1-`α`,q)` in MATLAB.

Theorem 7.13 implies

$$\mathbb{P}\left[\theta \in \widehat{C}\right] \to \mathbb{P}\left[\chi_q^2 \leq c_{1-\alpha}\right] = 1 - \alpha$$

which shows that $\widehat{C}$ has asymptotic coverage $1-\alpha$.

To illustrate the construction of a confidence region, consider the estimated regression (7.31) of

$$\widehat{\log(\mathit{wage})} = \beta_1\, \mathit{education} + \beta_2\, \mathit{experience} + \beta_3\, \mathit{experience}^2/100 + \beta_4.$$

Suppose that the two parameters of interest are the percentage return to education $\theta_1 = 100\beta_1$ and the percentage return to experience for individuals with 10 years experience $\theta_2 = 100\beta_2 + 20\beta_3$. These two parameters are a linear transformation of the regression parameters with point estimates

$$\widehat{\theta} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 20 & 0 \end{pmatrix} \widehat{\beta} = \begin{pmatrix} 11.8 \\ 1.2 \end{pmatrix},$$

and have the covariance matrix estimate

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \begin{pmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 20 \end{pmatrix} \widehat{\boldsymbol{V}}_{\widehat{\beta}} \begin{pmatrix} 0 & 0 \\ 100 & 0 \\ 0 & 100 \\ 0 & 20 \end{pmatrix} = \begin{pmatrix} 0.632 & 0.103 \\ 0.103 & 0.157 \end{pmatrix}$$

with inverse

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1} = \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix}.$$

Thus the Wald statistic is

$$\begin{aligned} W(\theta) &= \left(\widehat{\theta} - \theta\right)' \widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1} \left(\widehat{\theta} - \theta\right) \\ &= \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix}' \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix} \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix} \\ &= 1.77 \left(11.8 - \theta_1\right)^2 - 2.32 \left(11.8 - \theta_1\right)\left(1.2 - \theta_2\right) + 7.13 \left(1.2 - \theta_2\right)^2. \end{aligned}$$

The 90% quantile of the χ_2^2 distribution is 4.605 (we use the χ_2^2 distribution as the dimension of θ is two) so an asymptotic 90% confidence region for the two parameters is the interior of the ellipse $W(\theta) = 4.605$ which is displayed in Figure 7.5. Since the estimated correlation of the two coefficient estimates is modest (about 0.3) the region is modestly elliptical.

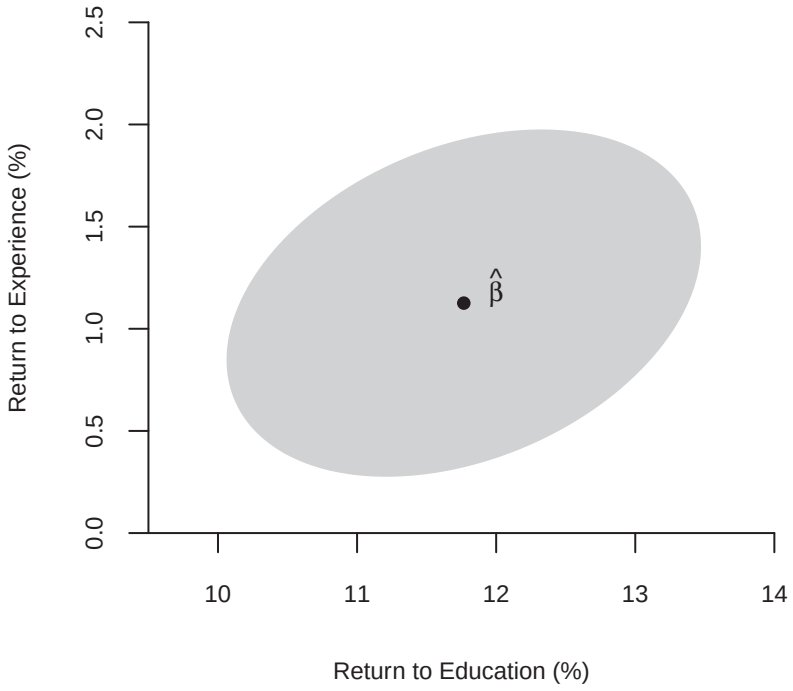


Figure 7.5: Confidence Region for Return to Experience and Return to Education

7.19 Edgeworth Expansion*

Theorem 7.11 showed that the t-ratio $T(\theta)$ is asymptotically normal. In practice this means that we use the normal distribution to approximate the finite sample distribution of T. How good is this approximation? Some insight into the accuracy of the normal approximation can be obtained by an

Edgeworth expansion which is a higher-order approximation to the distribution of T. The following result is an application of Theorem 9.11 of *Probability and Statistics for Economists*.

Theorem 7.15 Under Assumptions 7.2, 7.3, $\Omega > 0$, $\mathbb{E}\|e\|^{16} < \infty$, $\mathbb{E}\|X\|^{16} < \infty$, $g(\beta)$ has five continuous derivatives in a neighborhood of β, and $\mathbb{E}\left[\exp\left(t\left(\|e\|^4 + \|X\|^4\right)\right)\right] \leq B < 1$, as $n \to \infty$

$$\mathbb{P}[T(\theta) \leq x] = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + o\left(n^{-1}\right)$$

uniformly in x, where $p_1(x)$ is an even polynomial of order 2 and $p_2(x)$ is an odd polynomial of degree 5 with coefficients depending on the moments of e and X up to order 16.

Theorem 7.15 shows that the finite sample distribution of the t-ratio can be approximated up to $o(n^{-1})$ by the sum of three terms, the first being the standard normal distribution, the second a $O\left(n^{-1/2}\right)$ adjustment, and the third a $O\left(n^{-1}\right)$ adjustment.

Consider a one-sided confidence interval $\widehat{C} = \left[\widehat{\theta} - z_{1-\alpha} s(\widehat{\theta}), \infty\right)$ where $z_{1-\alpha}$ is the $1-\alpha^{th}$ quantile of $Z \sim \mathrm{N}(0,1)$, thus $\Phi(z_{1-\alpha}) - 1 - \alpha$. Then

$$\begin{aligned}\mathbb{P}\left[\theta \in \widehat{C}\right] &= \mathbb{P}[T(\theta) \leq z_{1-\alpha}] \\ &= \Phi(z_{1-\alpha}) + n^{-1/2} p_1(z_{1-\alpha})\phi(z_{1-\alpha}) + O\left(n^{-1}\right) \\ &= 1 - \alpha + O\left(n^{-1/2}\right).\end{aligned}$$

This means that the actual coverage is within $O\left(n^{-1/2}\right)$ of the desired $1-\alpha$ level.

Now consider a two-sided interval $\widehat{C} = \left[\widehat{\theta} - z_{1-\alpha/2} s(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2} s(\widehat{\theta})\right]$. It has coverage

$$\begin{aligned}\mathbb{P}\left[\theta \in \widehat{C}\right] &= \mathbb{P}[|T(\theta)| \leq z_{1-\alpha/2}] \\ &= 2\Phi(z_{1-\alpha/2}) - 1 + n^{-1} 2 p_2(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + o\left(n^{-1}\right) \\ &= 1 - \alpha + O\left(n^{-1}\right).\end{aligned}$$

This means that the actual coverage is within $O\left(n^{-1}\right)$ of the desired $1-\alpha$ level. The accuracy is better than the one-sided interval because the $O\left(n^{-1/2}\right)$ term in the Edgeworth expansion has offsetting effects in the two tails of the distribution.

7.20 Uniformly Consistent Residuals*

It seems natural to view the residuals $\widehat{e}_i$ as estimators of the unknown errors e_i. Are they consistent? In this section we develop a convergence result.

We can write the residual as

$$\widehat{e}_i = Y_i - X_i'\widehat{\beta} = e_i - X_i'\left(\widehat{\beta} - \beta\right). \tag{7.39}$$

Since $\widehat{\beta} - \beta \xrightarrow[p]{} 0$ it seems reasonable to guess that $\widehat{e}_i$ will be close to e_i if n is large.

We can bound the difference in (7.39) using the Schwarz inequality (B.12) to find

$$|\widehat{e}_i - e_i| = \left|X_i'\left(\widehat{\beta} - \beta\right)\right| \leq \|X_i\| \left\|\widehat{\beta} - \beta\right\|. \tag{7.40}$$

To bound (7.40) we can use $\left\|\widehat{\beta}-\beta\right\| = O_p\left(n^{-1/2}\right)$ from Theorem 7.3. We also need to bound the random variable $\|X_i\|$. If the regressor is bounded, that is, $\|X_i\| \le B < \infty$, then $|\widehat{e}_i - e_i| \le B\left\|\widehat{\beta}-\beta\right\| = O_p\left(n^{-1/2}\right)$. However if the regressor does not have bounded support then we have to be more careful.

The key is Theorem 6.15 which shows that $\mathbb{E}\|X\|^r < \infty$ implies $X_i = o_p\left(n^{1/r}\right)$ uniformly in i, or

$$n^{-1/r}\max_{1\le i\le n}\|X_i\| \xrightarrow[p]{} 0.$$

Applied to (7.40) we obtain

$$\max_{1\le i\le n}|\widehat{e}_i - e_i| \le \max_{1\le i\le n}\|X_i\|\left\|\widehat{\beta}-\beta\right\| = o_p\left(n^{-1/2+1/r}\right).$$

We have shown the following.

Theorem 7.16 Under Assumption 7.2 and $\mathbb{E}\|X\|^r < \infty$, then

$$\max_{1\le i\le n}|\widehat{e}_i - e_i| = o_p\left(n^{-1/2+1/r}\right). \tag{7.41}$$

The rate of convergence in (7.41) depends on r. Assumption 7.2 requires $r \ge 4$ so the rate of convergence is at least $o_p\left(n^{-1/4}\right)$. As r increases the rate improves.

We mentioned in Section 7.7 that there are multiple ways to prove the consistency of the covariance matrix estimator $\widehat{\Omega}$. We now show that Theorem 7.16 provides one simple method to establish (7.23) and thus Theorem 7.6. Let $q_n = \max_{1\le i\le n}|\widehat{e}_i - e_i| = o_p\left(n^{-1/4}\right)$. Since $\widehat{e}_i^2 - e_i^2 = 2e_i\left(\widehat{e}_i - e_i\right) + \left(\widehat{e}_i - e_i\right)^2$, then

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i'\left(\widehat{e}_i^2 - e_i^2\right)\right\| &\le \frac{1}{n}\sum_{i=1}^n \left\|X_i X_i'\right\|\left|\widehat{e}_i^2 - e_i^2\right| \\
&\le \frac{2}{n}\sum_{i=1}^n \|X_i\|^2 |e_i|\,|\widehat{e}_i - e_i| + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 |\widehat{e}_i - e_i|^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \|X_i\|^2 |e_i|\, q_n + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 q_n^2 \\
&\le o_p\left(n^{-1/4}\right).
\end{aligned}$$

7.21 Asymptotic Leverage*

Recall the definition of leverage from (3.40) $h_{ii} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X_i$. These are the diagonal elements of the projection matrix $\boldsymbol{P}$ and appear in the formula for leave-one-out prediction errors and HC2 and HC3 covariance matrix estimators. We can show that under i.i.d. sampling the leverage values are uniformly asymptotically small.

Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues of a symmetric square matrix $\boldsymbol{A}$ and note that $\lambda_{\max}(\boldsymbol{A}^{-1}) = (\lambda_{\min}(\boldsymbol{A}))^{-1}$.

Since $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX} > 0$, by the CMT $\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) \underset{p}{\longrightarrow} \lambda_{\min}\left(\boldsymbol{Q}_{XX}\right) > 0$. (The latter is positive since $\boldsymbol{Q}_{XX}$ is positive definite and thus all its eigenvalues are positive.) Then by the Quadratic Inequality (B.18)

$$\begin{aligned} h_{ii} &= X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \\ &\leq \lambda_{\max}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(X_i'X_i\right) \\ &= \left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\frac{1}{n}\left\|X_i\right\|^2 \\ &\leq \left(\lambda_{\min}\left(\boldsymbol{Q}_{XX}\right)+o_p(1)\right)^{-1}\frac{1}{n}\max_{1\leq i\leq n}\left\|X_i\right\|^2. \end{aligned} \tag{7.42}$$

Theorem 6.15 shows that $\mathbb{E}\left\|X\right\|^r < \infty$ implies $\max_{1\leq i\leq n}\left\|X_i\right\|^2 = \left(\max_{1\leq i\leq n}\left\|X_i\right\|\right)^2 = o_p\left(n^{2/r}\right)$ and thus (7.42) is $o_p\left(n^{2/r-1}\right)$.

Theorem 7.17 If X_i is i.i.d., $\boldsymbol{Q}_{XX} > 0$, and $\mathbb{E}\left\|X\right\|^r < \infty$ for some $r \geq 2$, then $\max_{1\leq i\leq n} h_{ii} = o_p\left(n^{2/r-1}\right)$.

For any $r \geq 2$ then $h_{ii} = o_p(1)$ (uniformly in $i \leq n$). Larger r implies a faster rate of convergence. For example $r = 4$ implies $h_{ii} = o_p\left(n^{-1/2}\right)$.

Theorem (7.17) implies that under random sampling with finite variances and large samples no individual observation should have a large leverage value. Consequently, individual observations should not be influential unless one of these conditions is violated.

7.22 Exercises

Exercise 7.1 Take the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$ with $\mathbb{E}\left[Xe\right] = 0$. Suppose that β_1 is estimated by regressing Y on X_1 only. Find the probability limit of this estimator. In general, is it consistent for β_1? If not, under what conditions is this estimator consistent for β_1?

Exercise 7.2 Take the model $Y = X'\beta + e$ with $\mathbb{E}\left[Xe\right] = 0$. Define the **ridge regression** estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' + \lambda \boldsymbol{I}_k\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right) \tag{7.43}$$

here $\lambda > 0$ is a fixed constant. Find the probability limit of $\widehat{\beta}$ as $n \to \infty$. Is $\widehat{\beta}$ consistent for β?

Exercise 7.3 For the ridge regression estimator (7.43), set $\lambda = cn$ where $c > 0$ is fixed as $n \to \infty$. Find the probability limit of $\widehat{\beta}$ as $n \to \infty$.

Exercise 7.4 Verify some of the calculations reported in Section 7.4. Specifically, suppose that X_1 and X_2 only take the values $\{-1,+1\}$, symmetrically, with

$$\begin{aligned}\mathbb{P}[X_1 = X_2 = 1] = \mathbb{P}[X_1 = X_2 = -1] &= 3/8 \\ \mathbb{P}[X_1 = 1, X_2 = -1] = \mathbb{P}[X_1 = -1, X_{2i} = 1] &= 1/8 \\ \mathbb{E}\left[e_i^2 \mid X_1 = X_2\right] &= \frac{5}{4} \\ \mathbb{E}\left[e_i^2 \mid X_1 \neq X_2\right] &= \frac{1}{4}.\end{aligned}$$

Verify the following:

(a) $\mathbb{E}[X_1] = 0$

(b) $\mathbb{E}\left[X_1^2\right] = 1$

(c) $\mathbb{E}[X_1 X_2] = \frac{1}{2}$

(d) $\mathbb{E}\left[e^2\right] = 1$

(e) $\mathbb{E}\left[X_1^2 e^2\right] = 1$

(f) $\mathbb{E}\left[X_1 X_2 e^2\right] = \frac{7}{8}$.

Exercise 7.5 Show (7.13)-(7.16).

Exercise 7.6 The model is

$$\begin{aligned}Y &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \\ \Omega &= \mathbb{E}\left[XX'e^2\right].\end{aligned}$$

Find the method of moments estimators $(\widehat{\beta}, \widehat{\Omega})$ for (β, Ω).

Exercise 7.7 Of the variables (Y^*, Y, X) only the pair (Y, X) are observed. In this case we say that Y^* is a **latent variable**. Suppose

$$\begin{aligned}Y^* &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \\ Y &= Y^* + u\end{aligned}$$

where u is a measurement error satisfying

$$\begin{aligned}\mathbb{E}[Xu] &= 0 \\ \mathbb{E}\left[Y^* u\right] &= 0.\end{aligned}$$

Let $\widehat{\beta}$ denote the OLS coefficient from the regression of Y on X.

(a) Is β the coefficient from the linear projection of Y on X?

(b) Is $\widehat{\beta}$ consistent for β as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.8 Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right)$ as $n \to \infty$.

Exercise 7.9 The model is $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the two estimators

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$
$$\widetilde{\beta} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i}{X_i}.$$

(a) Under the stated assumptions are both estimators consistent for β?

(b) Are there conditions under which either estimator is efficient?

Exercise 7.10 In the homoskedastic regression model $Y = X'\beta + e$ with $\mathbb{E}[e \mid x] = 0$ and $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ suppose $\widehat{\beta}$ is the OLS estimator of β with covariance matrix estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ based on a sample of size n. Let $\widehat{\sigma}^2$ be the estimator of σ^2. You wish to forecast an out-of-sample value of Y_{n+1} given that $X_{n+1} = x$. Thus the available information is the sample, the estimates $(\widehat{\beta}, \widehat{\boldsymbol{V}}_{\widehat{\beta}}, \widehat{\sigma}^2)$, the residuals $\widehat{e}_i$, and the out-of-sample value of the regressors X_{n+1}.

(a) Find a point forecast of Y_{n+1}.

(b) Find an estimator of the variance of this forecast.

Exercise 7.11 Take a regression model with i.i.d. observations (Y_i, X_i) with $X \in \mathbb{R}$

$$Y = X\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$
$$\Omega = \mathbb{E}\left[X^2 e^2\right].$$

Let $\widehat{\beta}$ be the OLS estimator of β with residuals $\widehat{e}_i = Y_i - X_i\widehat{\beta}$. Consider the estimators of Ω

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^n X_i^2 e_i^2$$
$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^n X_i^2 \widehat{e}_i^2.$$

(a) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\Omega} - \Omega\right)$ as $n \to \infty$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\Omega} - \Omega\right)$ as $n \to \infty$.

(c) How do you use the regression assumption $\mathbb{E}[e_i \mid X_i] = 0$ in your answer to (b)?

Exercise 7.12 Consider the model

$$Y = \alpha + \beta X + e$$
$$\mathbb{E}[e] = 0$$
$$\mathbb{E}[Xe] = 0$$

with both Y and X scalar. Assuming $\alpha > 0$ and $\beta < 0$ suppose the parameter of interest is the area under the regression curve (e.g. consumer surplus), which is $A = -\alpha^2/2\beta$.

Let $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})'$ be the least squares estimators of $\theta = (\alpha, \beta)'$ so that $\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d N(0, \boldsymbol{V}_\theta)$ and let $\widehat{\boldsymbol{V}}_\theta$ be a standard estimator for $\boldsymbol{V}_\theta$.

(a) Given the above, describe an estimator of A.

(b) Construct an asymptotic $1-\eta$ confidence interval for A.

Exercise 7.13 Consider an i.i.d. sample $\{Y_i, X_i\}$ $i = 1, ..., n$ where Y and X are scalar. Consider the reverse projection model $X = Y\gamma + u$ with $\mathbb{E}[Yu] = 0$ and define the parameter of interest as $\theta = 1/\gamma$.

(a) Propose an estimator $\widehat{\gamma}$ of γ.

(b) Propose an estimator $\widehat{\theta}$ of θ.

(c) Find the asymptotic distribution of $\widehat{\theta}$.

(d) Find an asymptotic standard error for $\widehat{\theta}$.

Exercise 7.14 Take the model

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + e \\ \mathbb{E}[Xe] &= 0 \end{aligned}$$

with both $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$, and define the parameter $\theta = \beta_1\beta_2$.

(a) What is the appropriate estimator $\widehat{\theta}$ for θ?

(b) Find the asymptotic distribution of $\widehat{\theta}$ under standard regularity conditions.

(c) Show how to calculate an asymptotic 95% confidence interval for θ.

Exercise 7.15 Take the linear model $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i^3 Y_i}{\sum_{i=1}^n X_i^4}.$$

Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.16 From an i.i.d. sample (Y_i, X_i) of size n you randomly take half the observations. You estimate a least squares regression of Y on X using only this sub-sample. Is the estimated slope coefficient $\widehat{\beta}$ consistent for the population projection coefficient? Explain your reasoning.

Exercise 7.17 An economist reports a set of parameter estimates, including the coefficient estimates $\widehat{\beta}_1 = 1.0$, $\widehat{\beta}_2 = 0.8$, and standard errors $s(\widehat{\beta}_1) = 0.07$ and $s(\widehat{\beta}_2) = 0.07$. The author writes "The estimates show that β_1 is larger than β_2."

(a) Write down the formula for an asymptotic 95% confidence interval for $\theta = \beta_1 - \beta_2$, expressed as a function of $\widehat{\beta}_1$, $\widehat{\beta}_2$, $s(\widehat{\beta}_1)$, $s(\widehat{\beta}_2)$ and $\widehat{\rho}$, where $\widehat{\rho}$ is the estimated correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(b) Can $\widehat{\rho}$ be calculated from the reported information?

(c) Is the author correct? Does the reported information support the author's claim?

Exercise 7.18 Suppose an economic model suggests

$$m(x) = \mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

where $X \in \mathbb{R}$. You have a random sample (Y_i, X_i), $i = 1, ..., n$.

(a) Describe how to estimate $m(x)$ at a given value x.

(b) Describe (be specific) an appropriate confidence interval for $m(x)$.

Exercise 7.19 Take the model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$ and suppose you have observations $i = 1, ..., 2n$. (The number of observations is $2n$.) You randomly split the sample in half, (each has n observations), calculate $\widehat{\beta}_1$ by least squares on the first sample, and $\widehat{\beta}_2$ by least squares on the second sample. What is the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)$?

Exercise 7.20 The variables $\{Y_i, X_i, W_i\}$ are a random sample. The parameter β is estimated by minimizing the criterion function

$$S(\beta) = \sum_{i=1}^{n} W_i \left(Y_i - X_i'\beta\right)^2$$

That is $\widehat{\beta} = \operatorname{argmin}_\beta S(\beta)$.

(a) Find an explicit expression for $\widehat{\beta}$.

(b) What population parameter β is $\widehat{\beta}$ estimating? Be explicit about any assumptions you need to impose. Do not make more assumptions than necessary.

(c) Find the probability limit for $\widehat{\beta}$ as $n \to \infty$.

(d) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.21 Take the model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ \mathbb{E}\left[e^2 \mid X\right] &= Z'\gamma \end{aligned}$$

where Z is a (vector) function of X. The sample is $i = 1, ..., n$ with i.i.d. observations. Assume that $Z'\gamma > 0$ for all Z. Suppose you want to forecast Y_{n+1} given $X_{n+1} = x$ and $Z_{n+1} = z$ for an out-of-sample observation $n+1$. Describe how you would construct a point forecast and a forecast interval for Y_{n+1}.

Exercise 7.22 Take the model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ Z &= X'\beta\gamma + u \\ \mathbb{E}[u \mid X] &= 0 \end{aligned}$$

where X is a k vector and Z is scalar. Your goal is to estimate the scalar parameter γ. You use a two-step estimator:

- Estimate $\widehat{\beta}$ by least squares of Y on X.
- Estimate $\widehat{\gamma}$ by least squares of Z on $X'\widehat{\beta}$.

(a) Show that $\widehat{\gamma}$ is consistent for γ.

(b) Find the asymptotic distribution of $\widehat{\gamma}$ when $\gamma = 0$.

Exercise 7.23 The model is $Y = X + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the estimator

$$\widetilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{X_i}.$$

Find conditions under which $\widetilde{\beta}$ is consistent for β as $n \to \infty$.

Exercise 7.24 The parameter β is defined in the model $Y = X^*\beta + e$ where e is independent of $X^* \geq 0$, $\mathbb{E}[e] = 0$, $\mathbb{E}\left[e^2\right] = \sigma^2$. The observables are (Y, X) where $X = X^* v$ and $v > 0$ is random scale measurement error, independent of X^* and e. Consider the least squares estimator $\widehat{\beta}$ for β.

(a) Find the plim of $\widehat{\beta}$ expressed in terms of β and moments of (X, v, e).

(b) Can you find a non-trivial condition under which $\widehat{\beta}$ is consistent for β? (By non-trivial we mean something other than $v = 1$.)

Exercise 7.25 Take the projection model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$. For a positive function $w(x)$ let $W_i = w(X_i)$. Consider the estimator

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} W_i X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} W_i X_i Y_i\right).$$

Find the probability limit (as $n \to \infty$) of $\widetilde{\beta}$. Do you need to add an assumption? Is $\widetilde{\beta}$ consistent for $\widetilde{\beta}$? If not, under what assumption is $\widetilde{\beta}$ consistent for β?

Exercise 7.26 Take the regression model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ \mathbb{E}\left[e^2 \mid X = x\right] &= \sigma^2(x) \end{aligned}$$

with $X \in \mathbb{R}^k$. Assume that $\mathbb{P}[e = 0] = 0$. Consider the infeasible estimator

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} e_i^{-2} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} e_i^{-2} X_i Y_i\right).$$

This is a WLS estimator using the weights e_i^{-2}.

(a) Find the asymptotic distribution of $\widetilde{\beta}$.

(b) Contrast your result with the asymptotic distribution of infeasible GLS.

Exercise 7.27 The model is $Y = X'\beta + e$ with $\mathbb{E}[e \mid X] = 0$. An econometrician is worried about the impact of some unusually large values of the regressors. The model is thus estimated on the subsample for which $|X_i| \leq c$ for some fixed c. Let $\widetilde{\beta}$ denote the OLS estimator on this subsample. It equals

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \mathbb{1}\{|X_i| \leq c\}\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \mathbb{1}\{|X_i| \leq c\}\right).$$

(a) Show that $\widetilde{\beta} \underset{p}{\longrightarrow} \beta$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\beta} - \beta\right)$.

Exercise 7.28 As in Exercise 3.26, use the `cps09mar` dataset and the subsample of white male Hispanics. Estimate the regression

$$\widehat{\log(wage)} = \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4.$$

(a) Report the coefficient estimates and robust standard errors.

(b) Let θ be the ratio of the return to one year of education to the return to one year of experience for *experience*= 10. Write θ as a function of the regression coefficients and variables. Compute $\widehat{\theta}$ from the estimated model.

(c) Write out the formula for the asymptotic standard error for $\widehat{\theta}$ as a function of the covariance matrix for $\widehat{\beta}$. Compute $s(\widehat{\theta})$ from the estimated model.

(d) Construct a 90% asymptotic confidence interval for θ from the estimated model.

(e) Compute the regression function at *education*= 12 and *experience*= 20. Compute a 95% confidence interval for the regression function at this point.

(f) Consider an out-of-sample individual with 16 years of education and 5 years experience. Construct an 80% forecast interval for their log wage and wage. [To obtain the forecast interval for the wage, apply the exponential function to both endpoints.]