The OLS Estimator and its Properties

1 The OLS estimator in a simple linear regression model

These questions relate to the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

We assume that rank($\mathbf{X}$) = k and that observations are independent across $i = 1, 2, ..., n$.

1.1

Consider the special case with $k = 2$ and $x_{1i} = 1$ for all $i = 1, 2, ..., n$, so that

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i.$$

1.1.1

What property of the explanatory variable x_{2i} is required in this case to guarantee that rank($\mathbf{X}$) = 2?

Since $\mathbf{X}_1$ is a unit vector, the only property we place on the $\mathbf{X}_2$ vector to ensure full rank of $\mathbf{X}$ is that it is non-uniform – that there must be no linear dependence between the $\mathbf{X}_1$ and $\mathbf{X}_2$ vectors. In other words,

$$\exists X_{2i}, X_{2j} : X_{2i} \neq X_{2j}.$$

1.1.2

Show that the 2 normal equations obtained by minimising

$$\phi(\boldsymbol{\beta}) = \sum_{i=1}^{n} u_i^2(\boldsymbol{\beta}) = \mathbf{u}^\top \mathbf{u},$$

with respect to $\boldsymbol{\beta}$ gives

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta},$$

and implies that the 2 elements of the OLS estimator, $\hat{\boldsymbol{\beta}}_{OLS}$, can be expressed as

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{X}_2,$$
$$\hat{\beta}_2 = \frac{\mathbf{S_{XY}}}{\mathbf{S_{XX}}},$$

where

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\
\bar{X}_2 &= \frac{1}{n}\sum_{i=1}^{n} X_{2i}, \\
\mathbf{S_{Xy}} &= \frac{1}{n}\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(y_i - \bar{y}), \\
\mathbf{S_{XX}} &= \frac{1}{n}\sum_{i=1}^{n}(X_{2i} - \bar{X})^2.
\end{aligned}$$

Apply the chain rule of matrix calculus to find our first order condition:

$$\frac{\partial \phi(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{u}^\top \mathbf{u}}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{u}}{\partial \boldsymbol{\beta}} \frac{\partial \mathbf{u}^\top \mathbf{u}}{\partial \mathbf{u}} = -2\mathbf{X}^\top \mathbf{u} = \mathbf{0}.$$

Substituting our value for $\mathbf{u}$ gives:

$$\begin{aligned}
\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \mathbf{0} \\
\mathbf{X}^\top \mathbf{y} &= \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\
\implies \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \qquad (1)
\end{aligned}$$

Alternatively, to attain the expressions for $\hat{\beta}$, start by using the method of moments where we assume that the [population] expectation of $\mathbf{u}$ is equal to zero:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 X_i) = 0.$$

But this is one equation with two unknown parameters. However, we can attain another equation by assuming that the sample mean is zero and using the fact that the expectations operator is a linear operator:

$$\mathbb{E}(X_i u_i) = \mathbb{E}(\mathbb{E}(X_i u_i | X_i)) = \mathbb{E}(X_i \mathbb{E}(u_i | X_i)) = 0,$$

which gives us the the corresponding sample mean condition:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(y_i - \beta_1 - \beta_2 X_i) = 0.$$

We now have two unknowns and two equations, which can be written as:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \beta_1 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\beta_2 \qquad (2)$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i y_i = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\beta_1 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)\beta_2 \qquad (3)$$

Multiplying both sides of (2) and (3) by n and rearranging gives:

$$\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n X_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

which can be expressed as the simple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Therefore, with the following:

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n y_i \\ \frac{1}{n}\sum_{i=1}^n X_i y_i \end{bmatrix},$$
$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i & \frac{1}{n}\sum_{i=1}^n X_i^2 \end{bmatrix},$$

we can rewrite (1) as:

$$\frac{1}{n}\sum_{i=1}^n X_i y_i = \left(\frac{1}{n}\sum_{i=1}^n X_i^\top X_i\right)\left(\frac{1}{n}\sum_{i=1}^n X_i^\top X_i\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_i y_i.$$

Thus, (2) and (3) also give us expressions for $\hat{\boldsymbol{\beta}}$:

$$\hat{\beta}_1 = \bar{y} - \bar{X}\hat{\beta}_2,$$

and substituting this into (3) gives:

$$\hat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i y_i) - \bar{X}\bar{y}}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2},$$

and since the following summation rules hold:

$$\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y}) = \sum_{i=1}^n X_i y_i - n\bar{X}\bar{y},$$
$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2,$$

we have

$$\hat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\mathbf{S_{Xy}}}{\mathbf{S_{XX}}}.$$

2 Matrix properties

2.1

Let $k = 2$ so we have

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \implies \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix},$$

and so

$$(\mathbf{X}^\top \mathbf{X})^\top = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & (\mathbf{X}_2^\top \mathbf{X}_1)^\top \\ (\mathbf{X}_1^\top \mathbf{X}_2) & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix} = \mathbf{X}^\top \mathbf{X}.$$

$$\mathbf{A}\mathbf{X} = \mathbf{I}_k \implies \mathbf{A} = \mathbf{X}^{-1}$$

so long as $\mathbf{X}$ is a $k \times k$ square matrix.

If $\mathbf{A}\mathbf{A}^\top = \mathbf{Q}^{-1} \implies \mathbf{A}\mathbf{A}^\top = \mathbf{X}^{-1}(\mathbf{X}^{-1})^\top$ since

$$(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1} \implies \mathbf{Q}^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{A}\mathbf{A}^\top.$$

The transpose of the projection matrix is the projection matrix itself:

$$\mathbf{P}_\mathbf{X}^\top = \left(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\top\right)^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = \mathbf{P}_\mathbf{X}.$$

The product of the projection matrix is the projection matrix itself:

$$\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = \mathbf{P}_\mathbf{X}.$$

The projection matrix which projects onto the X subspace multiplied by X is X itself:

$$\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \mathbf{X} = \mathbf{X}.$$

The transpose of the elimination matrix is the elimination matrix itself:

$$\mathbf{M}_\mathbf{X}^\top = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top\right)^\top = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = \mathbf{M}_\mathbf{X}.$$

The product of the elimination matrix is the elimination matrix itself:

$$\begin{aligned} \mathbf{M}_\mathbf{X}\mathbf{M}_\mathbf{X} &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top)(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top) \\ &= \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \\ &= \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top = \mathbf{M}_\mathbf{X}. \end{aligned}$$

The elimination matrix which projects onto the subspace orthogonal to $\mathbf{X}$ multiplied by $\mathbf{X}$ is the zero matrix:

$$\mathbf{M}_\mathbf{X}\mathbf{X} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top)\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{O}.$$

2.2

Given

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix},$$

$$|\mathbf{W}| = w_{11}w_{22} - w_{12}w_{21}$$

$$\mathbf{W}^{-1} = \frac{1}{w_{11}w_{22} - w_{12}w_{21}} \begin{bmatrix} w_{22} & -w_{12} \\ -w_{21} & w_{11} \end{bmatrix}$$

$$n\mathbf{W}^{-1} = \frac{n}{w_{11}w_{22} - w_{12}w_{21}} \begin{bmatrix} w_{22} & -w_{12} \\ -w_{21} & w_{11} \end{bmatrix}$$

$$\mathbf{U} = \frac{\mathbf{W}}{n},$$
$$|\mathbf{U}| = \frac{|\mathbf{W}|}{n}$$

$$\mathbf{U}^{-1} = \left(\frac{\mathbf{W}}{n}\right)^{-1} = \frac{n}{w_{11}w_{22} - w_{12}w_{21}} \begin{bmatrix} \frac{w_{22}}{n} & \frac{-w_{12}}{n} \\ \frac{-w_{21}}{n} & \frac{w_{11}}{n} \end{bmatrix}$$

$$\mathbf{W}^{-1} = \mathbf{U}^{-1} = \left(\frac{\mathbf{W}}{n}\right)^{-1} = \frac{n}{|\mathbf{W}|} \begin{bmatrix} \frac{w_{22}}{n} & \frac{-w_{12}}{n} \\ \frac{-w_{21}}{n} & \frac{w_{11}}{n} \end{bmatrix}.$$

3 Confidence intervals and simple hypothesis tests

The OLS estimator $\hat{\boldsymbol{\beta}}$ *has the conditional distribution*

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \overset{d}{\sim} N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

We estimate the unknown scalar parameter σ^2 *using the OLS estimator*

$$s^2 = \frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{n-k},$$

where $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$*. This gives an estimator o the conditional variance of* $\hat{\boldsymbol{\beta}}$ *as*

$$\hat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = s^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

Let $\hat{v}_{ii}$ *denote the elements of the main diagonal of* $\hat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ *for* $i = 1, 2, ..., k$*, and let* $SE_i = \sqrt{\hat{v}_{ii}}$*. The scalar* SE_i *is the standard error for* $\hat{\beta}_i$*, where* $\hat{\beta}_i$ *denotes the* i*th element of the column vector* $\hat{\boldsymbol{\beta}}$*. i.e. the OLS estimator of the scalar parameter* β_i*.*

The statistic

$$t_{\beta_i} = \frac{\hat{\beta}_i - \beta_i}{SE_i},$$

can be shown to have a Student t-distribution with $n-k$ *degrees of freedom, i.e.* $t_{\beta_i} \overset{d}{\sim} t(n-k)$*.*

3.1

Use this property to construct a 95% confidence interval for the true value of the parameter β_i*. i.e. an interval that contains the true value with probability 0.95 under the maintained assumption.*

To construct a 95% confidence interval begin with

$$\Pr\left(t_{\alpha/2} \leq t_{\beta_i} \leq t_{1-\alpha/2}\right) = 1 - \alpha,$$

then substitute our test statistic, and our significance level, α, and rearrange:

$$\begin{aligned}
\Pr\left(t_{0.025} \leq \frac{\hat{\beta}_i - \beta_i}{SE_i} \leq t_{0.975}\right) &= 0.95 \\
\Pr\left(t_{0.025}SE_i \leq \hat{\beta}_i - \beta_i \leq t_{0.975}SE_i\right) &= 0.95 \\
\Pr\left(t_{0.025}SE_i - \hat{\beta}_i \leq -\beta_i \leq t_{0.975}SE_i - \hat{\beta}_i\right) &= 0.95 \\
\Pr\left(\hat{\beta}_i - t_{0.025}SE_i \leq \beta_i \leq \hat{\beta}_i - t_{0.975}SE_i\right) &= 0.95.
\end{aligned}$$

3.2

Outline a test of the null hypothesis that the true value of $\beta_i = \beta_i^0$ *against a two-sided alternative that* $\beta_i \neq \beta_i^0$ *at the 5% significance level.*

The hypothesis would to test for the true value of β_i is:

$$H_0 : \beta_i = \beta_i^0$$
$$H_a : \beta_i \neq \beta_i^0.$$

Our test statistic to test this hypothesis was given above:

$$t_{\beta_i} = \frac{\hat{\beta}_i - \beta_i}{SE_i} \overset{d}{\sim} t(n-k),$$

where we can use an exact distribution to test the hypothesis if we assume that $\mathbf{u}$ is normally distributed. If we do not assume this, then we would have to use an asymptotic distribution for our test statistic. It is worth noting remember when to use exact and asymptotic tests via the following table:

Table 1: **Exact and Asymptotic Tests**

Errors	Small Sample	Large Sample
$\mathbf{u}\|\mathbf{X} \sim N$	$t \overset{d}{\sim} t(n-K)$ and $F \overset{d}{\sim} F(p, n-K)$	$t \overset{d}{\sim} t(n-K) \overset{d}{\sim} N(0,1)$ and $F \overset{d}{\sim} F(p, n-K) \overset{d}{\sim} \frac{\chi_p^2}{p}$
$\mathbf{u}\|\mathbf{X} \sim$IID	N/A	Using LLN/CLT: $t \overset{a}{\sim} N(0,1)$ and $F \overset{a}{\sim} \frac{\chi_p^2}{p}$

Assuming that our errors are normally distributed, then $t_{\beta_i} \overset{d}{\sim} t(n-k)$, and our decision rule will be to reject the null hypothesis if $|t_{\beta_i}| > t_{0.025}(n-k)$. That is, if our test statistic lies in the rejection region of the $t(n-k)$ distribution. Since this is a two-tailed test, we have two rejection regions – one for each tail.

3.3

Show that the hypothesis test you considered in part 2 will reject H_0 against the two-sided alternative at the 5% significance level if and only if the candidate value β_i^0 lies outside the 95% confidence interval for β_i that you obtained in part 1.

From part i), we know that the 95% confidence interval for this hypothesis test is:

$$\Pr\left(\hat{\beta}_i - t_{0.025}SE_i \leq \beta_i \leq \hat{\beta}_i - t_{0.975}SE_i\right) = 0.95.$$

So if β_i^0 lies outside the 95% confidence interval, we have

$$\beta_i < \hat{\beta}_i - t_{0.025}SE_i, \text{ or}$$
$$\beta_i > \hat{\beta}_i - t_{0.975}SE_i \leftrightarrow \beta_i > \hat{\beta}_i + t_{0.025}SE_i.$$

Multiplying both inequalities by -1 gives

$$-\beta_i > -\hat{\beta}_i + t_{0.025}SE_i,$$
$$-\beta_i < -\hat{\beta}_i + t_{0.975}SE_i,$$

then add $\hat{\beta}_i$

$$\hat{\beta}_i - \beta_i > t_{0.025} SE_i,$$
$$\hat{\beta}_i - \beta_i < t_{0.975} SE_i,$$

then divide by SE_i

$$\frac{\hat{\beta}_i - \beta_i}{SE_i} > t_{0.025},$$
$$\frac{\hat{\beta}_i - \beta_i}{SE_i} < t_{0.975},$$

and since $t_{\beta_i} = (\hat{\beta}_i - \beta_i) SE_i^{-1}$, we have

$$t_{\beta_i} > t_{0.025},$$
$$t_{\beta_i} < t_{0.975.}.$$

For completeness, we can make clear that $t_{0.025} = t_{\alpha/2}$ and $t_{0.975} = t_{1-\alpha/2}$.

3.4

(for full question, see problem set)

Given the data, our test statistic is:

$$t_{\beta_i} = \frac{\hat{\beta}_i - \beta^0}{SE_k} = \frac{-4.30 - \beta_3}{1.04},$$
$$\implies \Pr(-4.30 - 2.042 \times 1.04 \leq \beta_3 \leq -4.30 + 2.042 \times 1.04) = 0.95.$$

3.4.1

$$t_{\beta_i} = \frac{-4.30}{1.04} = 4.135,$$

which lies in our rejection region, and so we reject H_0.

3.4.2

$$t_{\beta_i} = \frac{5 - 4.30}{1.04} = 0.673,$$

which does not lie in our rejection region, and so we are unable to reject H_0.

3.5

A simple giveaway would be that we cannot have negative standard errors (or variances).

4 Testing multiple linear restrictions

The F-statistic can be expressed as the following:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{ps^2} \overset{d}{\sim} F_{p,n-k}, \tag{4}$$

where p are the number of restrictions, $\mathbf{R}$ is the $p \times k$ restriction matrix, $\mathbf{r}$is a $p \times 1$ vector with p linear combinations of the elements of $\boldsymbol{\beta}$, and s^2 is the sample variance estimator. To be clear

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta},$$

and $\text{rank}(\mathbf{R}) = p$.

For example, suppose have the following simple linear regression model

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \mathbf{u},$$

and some theory predicts that

$$\begin{aligned} \beta_2 - \beta_3 &= 0 \\ \beta_1 &= 1, \end{aligned}$$

then our restriction matrices would be

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

so if the restriction and our hypothesis is true

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta}.$$

It's worth noting that if

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},\ \ \mathbf{u} \overset{d}{\sim} N(0, \sigma^2\mathbf{I}),$$

then

$$\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\sim} N(0, \sigma^2\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^{-1})$$

and

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{R}^\top\left[\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right]^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \overset{d}{\sim} \chi^2_p.$$

Since we know

$$\frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{\sigma^2} \sim \chi^2_{n-k},$$

and that if $y_1 \sim \chi^2(m_1)$ and $y_2 \sim \chi^2(m_2)$, and $y_1 \perp y_2$

$$F = \frac{\frac{y_1}{m_1}}{\frac{y_2}{m_2}} \sim F(m_1, m_2),$$

we find that

$$\frac{1}{p}\left(\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \mathbf{R}^\top \left[\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right]^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2}\right) \div \frac{1}{n-k}\left(\frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{\sigma^2}\right)$$
$$= \frac{\chi^2_p}{p} \div \frac{\chi^2_{n-k}}{n-k} \sim F(p, n-k).$$

Under the null assumption that $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, the above test statistics becomes

$$\frac{\frac{(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})^\top\left[\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})}{p}}{\frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{n-k}} \sim F(p, n-k)$$

which is what we had in (4).

4.1

Outline a test of the joint null hypothesis that the true value of $\beta_i = \beta_i^0$ and the true value of $\beta_j = \beta_j^0$ (where $i \neq j$) against the two sided alternative that $\beta_i \neq \beta_i^0$ and $\beta_j \neq \beta_j^0$ at the 5% significance level.

The null hypothesis is then:

$$H_0 : \begin{cases} \beta_i = \beta_i^0 & \text{and} \\ \beta_j = \beta_j^0, \end{cases}$$

$$H_a : \begin{cases} \beta_i \neq \beta_i^0 & \text{or} \\ \beta_j \neq \beta_j^0 \end{cases}$$

which implies that $p = 2$. Suppose if $i = 2$ and $j = 3$, then our restriction matrices would be

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta}$$
$$\begin{bmatrix}\beta_2^0 \\ \beta_3^0\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 \\ 0 & 0 & 1\end{bmatrix}\boldsymbol{\beta}$$

4.2

See problem set for full information.

The F-statistic for this hypothesis is:

$$\begin{aligned} F &= \frac{1}{2}\begin{bmatrix}\hat{\beta}_k - \beta_k & \hat{\beta}_l - \beta_l\end{bmatrix}\begin{bmatrix}\hat{v}_{kk} & \hat{v}_{kl} \\ \hat{v}_{lk} & \hat{v}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\hat{\beta}_k - \beta_k \\ \hat{\beta}_l - \beta_l\end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix}2.37 - 2 & 5 - 4.30\end{bmatrix}\begin{bmatrix}3.3124 & 1.6071 \\ 1.6071 & 1.0816\end{bmatrix}^{-1}\begin{bmatrix}2.36 - 2 \\ 5 - 4.30\end{bmatrix} \\ &= 0.4693, \end{aligned}$$

which does not lie within our rejection region, and so we cannot reject the null hypothesis.

5 Concentrated likelihood

Write the (conditional) log-likelihood function for the parameter vector ϕ in a parametric model as $L(\phi)$. Partition the parameter vector ϕ into two sub-vectors ϕ_1 and ϕ_2, such that $\phi = (\phi_1, \phi_2)$. Let $\hat{\phi}_{1,ML}$ denote the (conditional) maximum likelihood (ML) estimator of the parameter vector ϕ_1, such that $\hat{\phi}_{1,ML}$ maximises $L(\phi_1, \phi_2)$ with respect to ϕ_1.

Substituting the ML estimator $\hat{\phi}_{1,ML}$ for the true value ϕ_1 in the (conditional) log-likelihood function is referred to as concentrating the (conditional) log-likelihood with respect to ϕ_1. This gives the concentrated (conditional) log-likelihood function $L(\hat{\phi}_{1,ML}, \phi_2)$. The (conditional) ML estimator of the parameter vector ϕ_2 can be obtained by maximising $L(\hat{\phi}_{1,ML}, \phi_2)$ with respect to ϕ_2.

Apply this approach to derive an expression for the (conditional) ML estimator of the variance parameter in the classic linear regression model with normally distributed errors. Compare this ML estimator of σ^2 with the OLS estimator $\hat{\sigma}^2 = \hat{\mathbf{u}}^\top \hat{\mathbf{u}}(n-k)^{-1}$. For a fixed sample size n, is the ML estimator biased? How do the two estimators compare as the sample size n increases?

The likelihood function for the simple linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} \sim \text{IID } N(0, \sigma^2\mathbf{I})$$

is given as

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right),$$

and so the log-likelihood function is:

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{5}$$

Differentiating l wrt $\boldsymbol{\beta}$ and solving for $\boldsymbol{\beta}$ with the first order condition gives:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{0} &= \frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\beta}}\left[\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}^\top\right] \\
&= -\frac{1}{\sigma^2}\left[-\mathbf{X}^\top\mathbf{y} + \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right] \\
\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{y} \\
\therefore \hat{\boldsymbol{\beta}}_{ML} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},
\end{aligned}$$

and substituting this estimator back into (5) gives:

$$l^C(\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}),$$

and differentiating this concentrated log-likelihood function wrt σ^2 and solving for σ^2

gives:

$$\frac{\partial l^C(\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\hat{\mathbf{u}}_{ML}^\top \hat{\mathbf{u}}_{ML}}{2(\sigma^2)^2} = 0$$
$$\implies \hat{\sigma}^2_{ML} = \frac{\hat{\mathbf{u}}_{ML}^\top \hat{\mathbf{u}}_{ML}}{n}.$$

But, we know that $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{ML}$, and so $\hat{\mathbf{u}}_{ML} = \hat{\mathbf{u}}_{OLS}$. However, we can see that the MLE parameter for σ^2 is in fact biased. It has not been adjusted for its degrees of freedom, and hence it has not been adjusted for the fact that we overestimate $\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$. Asymptotically, however, the bias shrinks and the MLE parameter estimate converges to the OLS parameter estimate.