Pset 5

Jake Underland,
Ian Bamford,
Matthew Chen

Question 1

1) The OLS minimization problem is:

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

$$[\hat{\beta}_0] \quad \sum_i -2(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0$$

$$[\hat{\beta}_1] \quad \sum_i -2x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0$$

$$[\hat{\beta}_2] \quad \sum_i -2x_{i2}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0$$

From $[\hat{\beta}_0]$,

$$\sum_i y_i - n\beta_0 - \beta_1 \sum_i x_{i1} - \beta_2 \sum_i x_{i2} = 0$$

$[\hat{\beta}_1]$,

$$\sum_i x_{i1} y_i - \beta_0 \sum_i x_{i1} - \beta_1 \sum_i x_{i1}^2 - \beta_2 \sum_i x_{i1} x_{i2} = 0$$

$[\hat{\beta}_2]$,

$$\sum_i x_{i2} y_i - \beta_0 \sum_i x_{i1} - \beta_1 \sum_i x_{i1} x_{i2} - \beta_2 \sum_i x_{i2}^2$$

Let $\sum_i y_i = Q_y$ $\qquad \sum_i x_{i1} x_{i2} = Q_{x_1 x_2}$

$\sum_i x_{i1} y_i = Q_{x_1 y}$ $\qquad \sum_i x_{ik}^2 = Q_{x_k^2}$

$\sum_i x_{i2} y_i = Q_{x_2 y}$ $\qquad \sum_i x_{ik} = Q_{x_k}$

$$\hat{\beta}_0 = \frac{1}{n}(Q_y - \hat{\beta}_1 Q_{x_1} - \hat{\beta}_2 Q_{x_2})$$

$$\hat{\beta}_1 = \frac{1}{Q_{x_1^2}}(Q_{x_1 y} - \hat{\beta}_0 Q_{x_1} - \hat{\beta}_2 Q_{x_1 x_2})$$

$$\hat{\beta}_2 = \frac{1}{Q_{x_2^2}}(Q_{x_2 y} - \hat{\beta}_0 Q_{x_2} - \hat{\beta}_1 Q_{x_1 x_2})$$

1) a. Solving the above set of linear equations for $\hat{\beta}_1$, we get

$$\hat{\beta}_1 = \frac{Qx_2^2 \, Qx_1y - Qx_1x_2 \, Qx_2y}{Qx_1^2 \, Qx_2^2 - (Qx_1x_2)^2}$$

$$= \frac{\sum_i x_{i2}^2 \sum_i x_{i1}y_i - \sum_i x_{i1}x_{i2} \sum_i x_{i2}y_i}{\sum_i x_{i1}^2 \sum_i x_{i2}^2 - \left(\sum_i x_{i1}x_{i2}\right)^2}$$

b).

From class, we know that

$$\tilde{\beta}_1 = \frac{\sum_i (x_{i1} - \bar{x}_{n1}) y_i}{\sum_i (x_{i1} - \bar{x}_{n1})} \quad \text{where} \quad \bar{x}_{n1} = \frac{1}{n}\sum_i x_{i1}$$

c). The OLS estimate of δ_1 is

$$\hat{\delta}_1 = \frac{\sum_i (x_{i1} - \bar{x}_{n1}) x_{i2}}{\sum_i (x_{i1} - \bar{x}_{n1})^2}$$

$$\tilde{\beta}_1 = \frac{\sum_i (x_{i1} - \bar{x}_{n1}) y_i}{\sum_i (x_{i1} - \bar{x}_{n1})^2} = \frac{\sum_i (x_{i1} - \bar{x}_{n1})(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{u}_i)}{\sum_i (x_{i1} - \bar{x}_{n1})^2}$$

$$= \hat{\beta}_0 \frac{\sum_i (x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2} + \hat{\beta}_1 \frac{\sum_i x_{i1}(x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2} + \hat{\beta}_2 \frac{\sum_i x_{i2}(x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2} + \frac{\sum_i \hat{u}_i (x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2}$$

Since $\sum_i (x_{i1} - \bar{x}_{n1}) = n\bar{x}_{n1} - n\bar{x}_{n1} = 0$

$\sum_i x_{i1}(x_{i1} - \bar{x}_{n1}) = \sum_i x_{i1}(x_{i1} - \bar{x}_{n1}) - \bar{x}_{n1} \overset{0}{\sum_i (x_{i1} - \bar{x}_{n1})} = \sum_i (x_{i1} - \bar{x}_{n1})^2$

and from FOCs,

$$\sum_i \hat{u}_i x_i = 0, \quad \sum_i \hat{u}_i = 0,$$

1.

(c).

we can rewrite the equation as

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\sum_i x_{i2}(x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2}$$

$$= \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

$$E(\tilde{\beta}_1 | X) = E(\hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1 | X)$$

$$= E(\hat{\beta}_1 | X) + \hat{\delta}_1 E(\hat{\beta}_2 | X).$$

Now, we know that, if we let

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

then $(X'X)^{-1}X'Y$ yields the OLS estimate

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

and

$$E(\hat{\beta}|X) = \begin{pmatrix} E(\hat{\beta}_0|X) \\ E(\hat{\beta}_1|X) \\ E(\hat{\beta}_2|X) \end{pmatrix} = E((X'X)^{-1}(X'Y)|X)$$

$$= (X'X)^{-1}X' E(Y|X)$$

$$= (X'X)^{-1}X' E(X\beta|X)$$

$$= (X'X)^{-1}X'X\beta = \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Thus, $E(\hat{\beta}_0|X) = \beta_0$, $E(\hat{\beta}_1|X) = \beta_1$, $E(\hat{\beta}_2|X) = \beta_2$.

Therefore, $E(\tilde{\beta}_1|X) = \beta + \hat{\delta}_1\beta_2$ $\square$

d) From c),

$$\tilde{\beta}_1 = \hat{\beta}_1$$

$$\Longrightarrow \quad \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1 = \hat{\beta}_1$$

For this to hold, either $\hat{\beta}_2 = 0$, $\hat{\delta}_1 = 0$, or both must hold.

When $\hat{\beta}_2 = 0$, and in the sample, x_{i2} is irrelevant to y_i. When $\hat{\delta}_1 = 0$, then x_{1i} and x_{2i} are uncorrelated.

$$E[\tilde{\beta} | X] = \beta_1$$

$$\Longrightarrow \quad \beta_1 + \beta_2 \hat{\delta}_1 = \beta_1$$

In order for the above to hold, either

$\beta_2 = 0$, or x_2 is irrelevant to y in the true population, or $\hat{\delta}_1 = 0$, and x_{1i} and x_{2i} are uncorrelated.

1. c).

$$\tilde{\beta}_1 = \frac{\sum_i (x_{i1} - \bar{x}_{n1}) y_i}{\sum_i (x_{i1} - \bar{x}_{n1})} \qquad \text{Let } k_i = \frac{(x_{i1} - \bar{x}_{n1})}{\sum_i (x_{i1} - \bar{x}_{n1})^2}$$

$$Var(\tilde{\beta}_1 | X) = Var(\sum_i k_i y_i | X)$$

$$\overset{MLR2}{=} \sum_i k_i^2 Var(y_i | X)$$

$$\overset{MLR5}{=} \sigma_a^2 \sum_i k_i^2$$

$$= \frac{\sigma_a^2}{\sum_i (x_{i1} - \bar{x}_{n1})^2} = \frac{\sigma_a^2}{SST_1}$$

Now, let $\hat{r}_{ij}$ denote the ith residual of a regression of x_j on all other regressors.

$$x_{ij} = r_0 + r_1 x_{i1} + \cdots + r_{j-1} x_{i,j-1} + r_{j+1} x_{i,j+1} + r_k x_{ik} + r_{ij}$$

In this model, $x_{i1} = r_0 + r_2 x_{i2} + r_{i1}$

$$\Rightarrow \hat{r}_{i1} = x_{i1} - \hat{r}_0 - \hat{r}_2 x_{i2}$$

Then, we can write $\hat{\beta}_1 = \dfrac{\sum_i \hat{r}_{i1} y_i}{\sum_i \hat{r}_{i1}^2} = \dfrac{\sum_i \hat{r}_{i1} y_i}{SSR_1}$

$$Var(\hat{\beta}_1 | X) = Var\left(\frac{\sum_i \hat{r}_{i1} y_i}{SSR_1} \middle| X\right)$$

$$\overset{MLR.2}{=} \frac{\sum_i \hat{r}_{i1}^2 Var(y_i)}{SSR_1^2} \overset{MLR5}{=} \frac{\sigma_b^2 SSR_1}{SSR_1^2}$$

$$= \frac{\sigma_b^2}{SSR_1} = \frac{\sigma_b^2}{SST_1(1 - R_1^2)}$$

where $R_1^2 = 1 - \dfrac{SSR_1}{SST_1}$

$\longrightarrow$

1 (e). Since $SSR_1 = \sum_i \hat{r}_{i1}^2 \geq 0$, and $SST_1 \geq 0$

$$R_1^2 = 1 - \frac{SSR_1}{SST_1} \geq 0.$$

Furthermore, we note that since $\beta_2 = 0$, the two estimators are modeling the same relationship and that $\sigma_a^2 = \sigma_b^2 = \sigma^2$.

Thus,

$$Var(\tilde{\beta}_1|X) = \frac{\sigma^2}{SST_1} \leq \frac{\sigma^2}{SST_1(1-R_1^2)} = Var(\hat{\beta}_1|X)$$

Equality holds when $R_1^2 = 0 \Rightarrow SST_1 = SSR_1$

$$\Rightarrow x_{i1} - \hat{r}_0 - \hat{r}_2 x_{i2} = x_{i1} - \overline{x_{n1}}$$

This holds when, in the regression of x_1 on x_2, $\hat{r}_2 = 0$, and $\hat{r}_0 = \overline{x}_{n1}$. This is the case when x_1 and x_2 are uncorrelated. Thus, the equality holds when x_1 and x_2 are uncorrelated.

1.

(f).

Let

$$SSR_L = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$$

$$SSR_S = \sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1})^2$$

Then, since $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and $\tilde{\beta}_0, \tilde{\beta}_1$ are the minimizers of the above expressions,

$$SSR_L \leq SSR_S$$

This is because, due to MLR 3, the above function is strictly convex and the solutions for $\hat{\beta}_0, \hat{\beta}_1$ & $\hat{\beta}_2$ are unique and minimize the above function. If $SSR_S \leq SSR_L$, then $\hat{\beta}_2$ would just take the value of 0 and SSR_L would be equal to SSR_S. The respective R^2 values R_L^2 and R_S^2 would be:

$$R_L^2 = 1 - SSR_L$$

$$R_S^2 = 1 - SSR_S$$

and so

$$R_L^2 \geq R_S^2.$$

The equality would hold when $\hat{\beta}_2 = 0$, in which case each expression's minimization problem would be identical and yield the same unique results.

Problem Set 5

ECON 21020 Spring, 2021

Jake Underland, Groupmates: Ian Bamford, Matthew Chen

2021-05-19

Question 2.

The OLS problem is:

$$\min_{\beta} \frac{1}{n}\Sigma(y_i - \beta x_i)^2$$

FOCs:

$$\begin{aligned}
\{\beta\}: \ \frac{d}{d\beta}\frac{1}{n}\Sigma(y_i - \beta x_i)^2 = 0 \\
\implies -2\frac{1}{n}\Sigma x_i(y_i - \beta x_i) = 0 \\
\implies -2\frac{1}{n}\Sigma x_i y_i + 2\beta\frac{1}{n}\Sigma x_i^2 = 0 \\
\implies \hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}
\end{aligned}$$

Therefore, $\hat{\beta}$ is the OLS estimate of β. For bias,

$$\begin{aligned}
\hat{\beta} &= \frac{\Sigma x_i y_i}{\Sigma x_i^2} \\
E[\hat{\beta}|x_1, \cdots, x_n] &= E[\frac{\Sigma x_i y_i}{\Sigma x_i^2}|x_1, \cdots, x_n] \\
&= \frac{\Sigma x_i E[y_i|x_1, \cdots, x_n]}{\Sigma x_i^2} \\
&= \frac{\Sigma x_i E[\beta x_i + u_i|x_1, \cdots, x_n]}{\Sigma x_i^2} \\
&= \frac{\beta\Sigma x_i^2 + \Sigma x_i E[u_i|x_1, \cdots, x_n]}{\Sigma x_i^2} \\
&= \frac{\beta\Sigma x_i^2}{\Sigma x_i^2} \\
&= \beta \\
\implies E[\hat{\beta}] &= E[E[\hat{\beta}|x_1, \cdots, x_n]] = \beta
\end{aligned}$$

Thus, $\hat{\beta}$ is unbiased.

$$\begin{aligned}
\bar{\beta} &= \frac{\Sigma y_i}{\Sigma x_i} \\
E[\bar{\beta}|x_1, \cdots, x_n] &= E[\frac{\Sigma y_i}{\Sigma x_i}|x_1, \cdots, x_n] \\
&= \frac{\Sigma E[\beta x_i + u_i|x_1, \cdots, x_n]}{\Sigma x_i} \\
&= \frac{\beta \Sigma x_i + \Sigma E[u_i|x_1, \cdots, x_n]}{\Sigma x_i} \\
&= \frac{\beta \Sigma x_i}{\Sigma x_i} \\
&= \beta \\
\implies E[\bar{\beta}] &= E[E[\bar{\beta}|x_1, \cdots, x_n]] = \beta
\end{aligned}$$

Thus, the second estimator is also unbiased.
By the Gauss-Markov theorem, the OLS estimator would be the variance minimizing estimator under these assumptions. We can check as follows:

$$\begin{aligned}
Var(\hat{\beta}|X) &= Var(\frac{\Sigma x_i y_i}{\Sigma x_i^2}|X) \\
&\overset{MLR\,2}{=} \frac{\Sigma x_i^2 Var(y_i|X)}{(\Sigma x_i^2)^2} \\
&\overset{MLR\,5}{=} \frac{\sigma^2 \Sigma x_i^2}{(\Sigma x_i^2)^2} \\
&= \frac{\sigma^2}{\Sigma x_i^2}
\end{aligned}$$

$$\begin{aligned}
Var(\bar{\beta}|X) &= Var(\frac{\Sigma y_i}{\Sigma x_i}|X) \\
&\overset{MLR\,2}{=} \frac{\Sigma Var(y_i|X)}{(\Sigma x_i)^2} \\
&\overset{MLR\,5}{=} \frac{n\sigma^2}{(\Sigma x_i)^2} \\
&= \frac{n\sigma^2}{(\Sigma x_i)^2} \\
&= \frac{n\sigma^2}{(n\bar{x}_n)^2} = \frac{\sigma^2}{n\bar{x_n}^2}
\end{aligned}$$

Now, we know from Jensen's inequality that variance is always greater than or equal to 0:

$$\begin{aligned}
&\frac{1}{n}\Sigma x_i^2 - \bar{x_n}^2 \geq 0 \\
\implies &\Sigma x_i^2 - n\bar{x_n}^2 \geq 0 \ldots \text{ Since } n > 0 \\
\implies &\Sigma x_i^2 \geq n\bar{x_n}^2 \\
\implies &\frac{\sigma^2}{\Sigma x_i^2} \leq \frac{\sigma^2}{n\bar{x_n}^2}
\end{aligned}$$

Thus, we have that the variance of the OLS estimate conditioned on $X = x_1, \dots, x_n$ has a smaller variance than the other estimate, confirming the Gauss-Markov principle.

Question 3.

6.

(i) Let

$$\beta \equiv \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, Y \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X \equiv \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{pmatrix}, u \equiv \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

Then, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ can be written as

$$Y = X\beta + u$$

and, as we know from class, the OLS estimator of β is

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1}X'Y$$

$$\begin{aligned} E(\hat{\beta}|X) = \begin{pmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ E(\hat{\beta}_3) \end{pmatrix} &= E((X'X)^{-1}X'Y|X) \\ &= (X'X)^{-1}X'E(Y|X) \\ &= (X'X)^{-1}X'E(X\beta|X) \\ &= (X'X)^{-1}X'XE(\beta|X) \\ &= \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \end{aligned}$$

Thus, we have that $E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_2) = \beta_2$. Then,

$$E(\hat{\theta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2) = E(\hat{\beta}_1) + E(\hat{\beta}_2) = \beta_1 + \beta_2 = \theta_1$$

And thus $\hat{\theta}_1$ is an unbiased estimator of θ_1.

(ii)

$$\begin{aligned} Var(\hat{\theta}_1) &= Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2Cov(\hat{\beta}_1, \hat{\beta}_2) \\ &= Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2Corr(\hat{\beta}_1, \hat{\beta}_2)\sqrt{Var(\hat{\beta}_1)Var(\hat{\beta}_2)} \end{aligned}$$

11.

From class, we have

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\Sigma_i \hat{r}_{i1} y_i}{\Sigma_i \hat{r}_{i1}^2} \\
&= \frac{\Sigma_i \hat{r}_{i1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i)}{\Sigma_i \hat{r}_{i1}^2} \\
&= \frac{\beta_0 \Sigma_i \hat{r}_{i1} + \beta_1 \Sigma_i \hat{r}_{i1} x_{i1} + \beta_2 \Sigma_i \hat{r}_{i1} x_{i2} + \beta_3 \Sigma_i \hat{r}_{i1} x_{i3} + \Sigma_i \hat{r}_{i1} u_i}{\Sigma_i \hat{r}_{i1}^2} \\
&= \beta_1 + \beta_3 \frac{\Sigma_i \hat{r}_{i1} x_{i3}}{\Sigma_i \hat{r}_{i1}^2} + \frac{\Sigma_i \hat{r}_{i1} u_i}{\Sigma_i \hat{r}_{i1}^2}
\end{aligned}$$

Where the 4th equality follows from the properties of the residual of the regression x_1 on x_2:

$$\begin{aligned}
\Sigma_i \hat{r}_{i1} &= 0, \Sigma_i \hat{r}_{i1} x_{i2} = 0 \\
\Sigma_i \hat{r}_{i1} x_{i1} &= \Sigma_i \hat{r}_{i1}(\hat{r}_{i1} + \hat{x}_{i1}) \\
&= \Sigma_i \hat{r}_{i1}^2 + \Sigma_i \hat{r}_{i1}(\hat{\gamma}_0 + \hat{\gamma}_2 x_{i2}) \\
&= \Sigma_i \hat{r}_{i1}^2 + \hat{\gamma}_0 \underbrace{\Sigma_i \hat{r}_{i1}}_{=0} + \hat{\gamma}_2 \underbrace{\Sigma_i \hat{r}_{i1} x_{i2}}_{=0} \\
&= \Sigma_i \hat{r}_{i1}^2
\end{aligned}$$

Thus,

$$\begin{aligned}
E(\tilde{\beta}_1|X) &= E(\beta_1 + \beta_3 \frac{\Sigma_i \hat{r}_{i1} x_{i3}}{\Sigma_i \hat{r}_{i1}^2} + \frac{\Sigma_i \hat{r}_{i1} u_i}{\Sigma_i \hat{r}_{i1}^2}|X) \\
&= \beta_1 + \beta_3 \frac{\Sigma_i \hat{r}_{i1} x_{i3}}{\Sigma_i \hat{r}_{i1}^2} + \frac{\Sigma_i \hat{r}_{i1} E(u_i|X)}{\Sigma_i \hat{r}_{i1}^2} \\
&\overset{MLR\,4}{=} \beta_1 + \beta_3 \frac{\Sigma_i \hat{r}_{i1} x_{i3}}{\Sigma_i \hat{r}_{i1}^2}
\end{aligned}$$

C6

```r
library(wooldridge)
data(wage2)
```

(i)

```r
simple1 <- lm(IQ ~ educ, data = wage2)
delta_tilde <- simple1$coefficients
delta_tilde
```

```
## (Intercept)        educ
##   53.687154    3.533829
```

$\implies \tilde{\delta}_1 \approx 3.53$

(ii)

```r
simple2 <- lm(log(wage) ~ educ, data = wage2)
beta_tilde <- simple2$coefficients
beta_tilde
```

```
## (Intercept)        educ
##  5.97306245  0.05983921
```

$\implies \tilde{\beta}_1 \approx 0.06$

(iii)

```
multiple <- lm(log(wage) ~ educ + IQ, data = wage2)
beta_hats <- multiple$coefficients
beta_hats
```

```
## (Intercept)        educ          IQ
## 5.658287588 0.039119901 0.005863132
```

$\implies \hat{\beta}_1 \approx 0.04, \hat{\beta}_2 \approx 0.006$

(iv)

```
beta_hats[2] + beta_hats[3] * delta_tilde[2]
```

```
##       educ
## 0.05983921
```

```
beta_hats[2] + beta_hats[3] * delta_tilde[2] == beta_tilde[2]
```

```
## educ
## TRUE
```

Question 4.

8.

(i)

$$Var(\hat{\beta}_1 - 3\hat{\beta}_2) = Var(\hat{\beta}_1) + 9Var(\hat{\beta}_2) - 6Cov(\hat{\beta}_1, \hat{\beta}_2)$$

The standard error is

$$se(\hat{\beta}_1 - 3\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1) + 9Var(\hat{\beta}_2) - 6Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

(ii)

The t statistic is

$$T = \frac{\hat{\beta}_1 - 3\hat{\beta}_2 - 1}{\sqrt{Var(\hat{\beta}_1) + 9Var(\hat{\beta}_2) - 6Cov(\hat{\beta}_1, \hat{\beta}_2)}}$$

(iii)

$$\begin{aligned} \theta_1 &= \beta_1 - 3\beta_2 \\ \implies \beta_1 &= \theta_1 + 3\beta_2 \end{aligned}$$

Thus,

$$\begin{aligned} \theta_1 &= \beta_1 - 3\beta_2 \\ \implies \beta_1 &= \theta_1 + 3\beta_2 \\ y &= \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ &= \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_3 x_3 + u \end{aligned}$$

We can directly obtain $\hat{\theta}_1$ and its standard error by estimating the coefficient and standard error on x_1 in the above regression.

C3

(i)

```r
data("hprice1")
```

```r
log_price_basic <- lm(log(price) ~ sqrft + bdrms, data = hprice1)
log_price_basic$coefficients
```

```
## (Intercept)       sqrft       bdrms
## 4.766027213 0.000379446 0.028884467
```

```r
theta <- 150 * log_price_basic$coefficients[2] + log_price_basic$coefficients[3]
paste("Coefficient of theta is", theta)
```

```
## [1] "Coefficient of theta is 0.0858013664032184"
```

(ii)

Since $\beta_2 = \theta_1 - 150\beta_1$,

$$\ln(price) = \beta_0 + \beta_1(sqrft - 150bdrms) + \theta_1 bdrms + u$$

(iii)

```r
# create sqrft - 150 bdrms
hprice1$sqrft150bdrms = hprice1$sqrft - 150 * hprice1$bdrms
log_price_plugged <- lm(log(price) ~ sqrft150bdrms + bdrms, data = hprice1)
t_conf_intervals <- confint(log_price_plugged)
```

```r
# 95 % confidence intervals reported below coefficients,
# theta is coefficient of bdrms
stargazer(log_price_plugged, header = FALSE, type = "latex",
          ci.custom=list(t_conf_intervals))
```

Table 1:

	Dependent variable:
	log(price)
sqrft150bdrms	0.0004***
	(0.0003, 0.0005)
bdrms	0.086***
	(0.033, 0.139)
Constant	4.766***
	(4.573, 4.959)
Observations	88
R^2	0.588
Adjusted R^2	0.579
Residual Std. Error	0.197 (df = 85)
F Statistic	60.729*** (df = 2; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

C5.

```r
library(equatiomatic)
```

(i)

```r
data(mlb1)
log_salary_1 <- lm(log(salary) ~ years + gamesyr + bavg + hrunsyr, data=mlb1)
summary(log_salary_1)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr,
##     data = mlb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0642 -0.4614 -0.0271  0.4654  2.7216
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.020913   0.265719  41.476  < 2e-16 ***
## years        0.067732   0.012113   5.592 4.55e-08 ***
## gamesyr      0.015759   0.001564  10.079  < 2e-16 ***
## bavg         0.001419   0.001066   1.331    0.184
## hrunsyr      0.035943   0.007241   4.964 1.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7279 on 348 degrees of freedom
## Multiple R-squared:  0.6254, Adjusted R-squared:  0.6211
## F-statistic: 145.2 on 4 and 348 DF,  p-value: < 2.2e-16
```

```r
extract_eq(log_salary_1, use_coefs = TRUE, raw_tex = TRUE)
```

$$\log(\widehat{salary}) = 11.02 + 0.07(\operatorname{years}) + 0.02(\operatorname{gamesyr}) + 0(\operatorname{bavg}) + 0.04(\operatorname{hrunsyr})$$

Both statsitical significance and coefficient of *hrunsyr* increases.

(ii)

```r
log_salary_2 <- lm(log(salary) ~ years + gamesyr + bavg + hrunsyr
                   + runsyr + fldperc + sbasesyr, data=mlb1)
summary(log_salary_2)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
##     runsyr + fldperc + sbasesyr, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11554 -0.44557 -0.08808  0.48731  2.57872
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 10.4082678  2.0032546   5.196 3.50e-07 ***
## years        0.0699848  0.0119756   5.844 1.18e-08 ***
## gamesyr      0.0078995  0.0026775   2.950 0.003391 ** 
## bavg         0.0005296  0.0011038   0.480 0.631656    
## hrunsyr      0.0232106  0.0086392   2.687 0.007566 ** 
## runsyr       0.0173922  0.0050641   3.434 0.000666 ***
## fldperc      0.0010351  0.0020046   0.516 0.605936    
## sbasesyr    -0.0064191  0.0051842  -1.238 0.216479    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7176 on 345 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.6317 
## F-statistic: 87.25 on 7 and 345 DF,  p-value: < 2.2e-16
```

The factors with the stars next to them are individually statistically significant.

(iii)

```r
library(car)
```

```r
# Result of joint hypotheses F-test
linearHypothesis(log_salary_2, c("bavg = 0", "fldperc = 0", "sbasesyr = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## bavg = 0
## fldperc = 0
## sbasesyr = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + runsyr + fldperc + 
##     sbasesyr
##
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    348 178.72                          
## 2    345 177.66  3    1.0583 0.685 0.5617
```

Since the p-value is .56, we cannot reject the null hypothesis.

Question 5

C3

(i)

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u \\ &= \beta_0 + (\beta_1 + \beta_3 exper)educ + \beta_2 exper + u\end{aligned}$$

(ii)

$$\begin{aligned}H_0 &: \ \beta_3 = 0 \\ H_1 &: \ \beta_3 \neq 0\end{aligned}$$

We use a two-sided test because we cannot rule out the possibility that experience has a negative effect on the returns to education. For example, the more experienced a person is, the less significant their educational background may be in determining wages, as many companies use education as a signal for abiilty, but experience already works as a credible signal of one's ability and may lessen the importance of educational background.

(iii)

```r
data(wage2)
```

```r
return_educ <- lm(log(wage) ~ educ + exper + educ * exper, data = wage2)
res <- summary(return_educ)
res
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + educ * exper, data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88558 -0.24553  0.03558  0.26171  1.28836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.949455   0.240826  24.704   <2e-16 ***
## educ         0.044050   0.017391   2.533   0.0115 *
## exper       -0.021496   0.019978  -1.076   0.2822
## educ:exper   0.003203   0.001529   2.095   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 931 degrees of freedom
## Multiple R-squared:  0.1349, Adjusted R-squared:  0.1321
## F-statistic: 48.41 on 3 and 931 DF,  p-value: < 2.2e-16
```

Because the two-sided p-value of the interaction is 0.0365, at a 95% confidence level, we reject H_0.

(iv)

By following the hint we get the following equation:

$$= \beta_0 + \theta_1 educ + \beta_2 exper + \beta_3 educ(exper - 10) + u$$

```
# create exper-10
wage2$exper_minus_ten <- wage2$exper - 10
return_educ_2 <- lm(log(wage) ~ educ + exper + educ * (exper_minus_ten), data = wage2)
t_conf_intervals2 <- confint(return_educ_2)
```

```
# 95 % confidence intervals reported below coefficients,
# theta is coefficient on educ
stargazer(return_educ_2, header = FALSE, type = "latex",
          ci.custom=list(t_conf_intervals2))
```

Table 2:

	Dependent variable:
	log(wage)
educ	0.076^{***}
	(0.063, 0.089)
exper	-0.021
	(-0.061, 0.018)
exper_minus_ten	
educ:exper_minus_ten	0.003^{**}
	(0.0002, 0.006)
Constant	5.949^{***}
	(5.477, 6.422)
Observations	935
R^2	0.135
Adjusted R^2	0.132
Residual Std. Error	0.392 (df = 931)
F Statistic	48.407^{***} (df = 3; 931)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 6

Two analysts at a bank want to determine an appropriate credit limit for new customers with a given credit score using existing data on credit limit decisions. The first analyst studies customers with 'good' credit, and the second studies customers with 'excellent' credit. For the purpose of this exercise, suppose these are the only two categories. Suppose 'Good' credit is scored between 0-400 and 'Excellent' credit 400-800. The analysts separately report the following fitted equations:

$$\text{'Good'} : \hat{y}_i = 1000 + 0.5 score_i$$
$$\text{'Excellent'} : \hat{y}_i = 1500 + 0.7 score_i$$

a) Explain how you could combine the information from both of these univariate regressions by running a single (multivariate) linear regression. Deduce from the information provided what the parameter estimates would be in your multivariate regression and provide a derivation based on the OLS minimization problem.

Solution. We could combine the the two regressions by introducing a dummy variable. Denote

$$d_i = \begin{cases} 1 \text{ if } score_i \geq 400 \\ 0 \text{ if } score_i < 400 \end{cases}$$

Then, the regression equation would be

$$y_i = \beta_0 + \beta_1 score_i + d_i(\gamma + \delta score_i) + \epsilon_i$$

Estimating the parameters via OLS goes as follows:

$$\begin{aligned} &\min_{\beta_0,\beta_1,\gamma,\delta} \Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 score_i - d_i(\gamma + \delta score_i))^2 \\ =&\min_{\beta_0,\beta_1,\gamma,\delta} \Sigma_{i:d_i=0}(y_i - \beta_0 - \beta_1 score_i)^2 + \Sigma_{i:d_i=1}(y_i - \beta_0 - \beta_1 score_i - \gamma - \delta score_i)^2 \\ =&\min_{\beta_0,\beta_1,\gamma,\delta} \Sigma_{i:d_i=0}(y_i - \beta_0 - \beta_1 score_i)^2 + \Sigma_{i:d_i=1}(y_i - (\beta_0 + \gamma) - (\beta_1 + \delta) score_i)^2 \end{aligned}$$

Since the above functions under our assumptions are strictly convex, the minimizations of the first and second summation actually yield the same unique results as the 'Good' regression and 'Excellent' regression, respectively. Thus, the solution is

$$\begin{aligned} \hat{\beta}_0 &= 1000 \\ \hat{\gamma} &= 500 \\ \hat{\beta}_1 &= 0.5 \\ \hat{\delta} &= 0.2 \end{aligned}$$

b) What is the estimated size of the 'jump'? (The 'jump' does not occur at 0). What is the estimated 'kink'?

Solution. The 'jump' occurs at 400, since that is when the credit scores moves from "Good" to "Excellent". Therefore, the 'jump' is $\gamma + \delta * 400 = 500 + 0.2 * 400 = 580$. The 'kink' is the value of $\delta = 0.2$, which is the increase in the marginal impact of credit score on credit limit decisions.

c) Explain how to test the null hypothesis that the intercept and slope for the two credit categories are equal versus the alternative that they differ. What information would you need?

Solution.

$$\begin{aligned} H_0 &: \gamma = \delta = 0 \\ H_1 &: \gamma \neq 0 \text{ or } \delta \neq 0 \end{aligned}$$

The restricted and unrestricted regressions are as follows:

$$\begin{aligned} \text{Restricted: } & y_i \beta_0 + \beta_1 score_i + \epsilon_i \\ \text{Unrestricted: } & y_i = \beta_0 + \beta_1 score_i + d_i(\gamma + \delta score_i) + \epsilon_i \end{aligned}$$

The restricted regression gives us SSR_R, the unrestricted regression gives us SSR_U. The number of restrictions is 2, and we are estimating 4 parameters. From here we can compute the F-stat, which is

$$F = \frac{(SSR_R - SSR_U)/2}{SSR_U/(n-4-1)} = \frac{(SSR_R - SSR_U)/2}{SSR_U/(n-5)}$$

To compute this F-stat, we would need to run the above regressions and obtain information of n, SSR_R, SSR_U. We would reject the null hypothesis if $F > F_{2,n-5,1-\alpha}$.