PSTAT 126 - Assignment 5

Fall 2022

Kayla Katakis

Note: ***Submit both your Rmd and generated pdf file to Canvas.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

1.

(a) In Lab 5 we showed that the OLS estimator for the Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

is given by

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n Y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i Y_i\right) \\ n\left(\sum_{i=1}^n x_i Y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right) \end{pmatrix}.$$

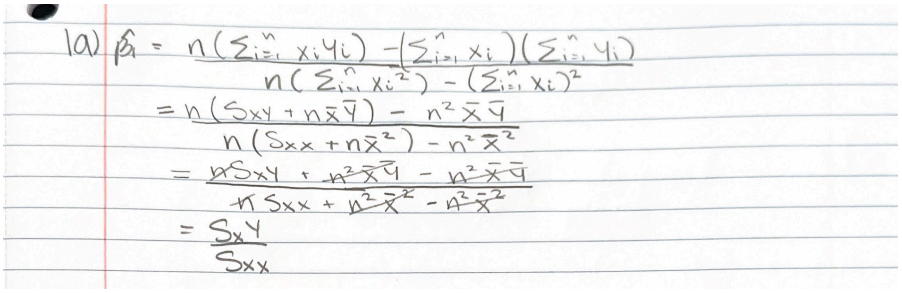
Show that this expression is equivalent to the familiar identity

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{x}\hat{\beta}_1 \\ S_{xY}/S_{xx} \end{pmatrix}.$$

Hint: Refer to Lab 1 for formulas for S_{xx} and S_{xY}.

Solution:

```
knitr::include_graphics("/Users/kaylakatakis/Desktop/beta1.pdf")
```



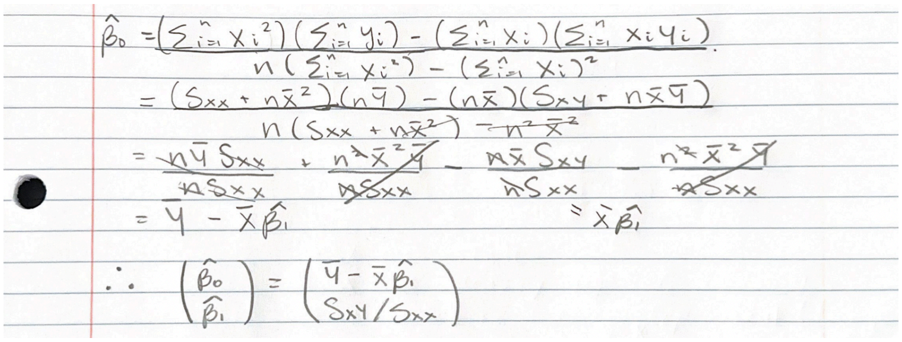
1a) $\hat{\beta}_1 = \dfrac{n(\sum_{i=1}^n x_i Y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$

$= \dfrac{n(S_{xY} + n\bar{x}\bar{Y}) - n^2\bar{x}\bar{Y}}{n(S_{xx} + n\bar{x}^2) - n^2\bar{x}^2}$

$= \dfrac{\cancel{n}S_{xY} + \cancel{n^2\bar{x}\bar{Y}} - \cancel{n^2\bar{x}\bar{Y}}}{\cancel{n}S_{xx} + \cancel{n^2\bar{x}^2} - \cancel{n^2\bar{x}^2}}$

$= \dfrac{S_{xY}}{S_{xx}}$

```r
knitr::include_graphics("/Users/kaylakatakis/Desktop/beta0.pdf")
```



$$\hat{\beta}_0 = \frac{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}.$$

$$= \frac{(S_{xx} + n\bar{x}^2)(n\bar{y}) - (n\bar{x})(S_{xy} + n\bar{x}\bar{y})}{n(S_{xx} + \cancel{n\bar{x}^2}) - \cancel{n^2\bar{x}^2}}$$

$$= \frac{\cancel{n}\bar{y}\cancel{S_{xx}}}{\cancel{n}\cancel{S_{xx}}} + \cancel{\frac{n^2\bar{x}^2\bar{y}}{nS_{xx}}} - \frac{\cancel{n}\bar{x}S_{xy}}{\cancel{n}S_{xx}} - \cancel{\frac{n^2\bar{x}^2\bar{y}}{nS_{xx}}}$$

$$= \bar{y} - \bar{x}\hat{\beta}_1 \qquad\qquad = \bar{x}\hat{\beta}_1$$

$$\therefore \begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}\bar{y} - \bar{x}\hat{\beta}_1 \\ S_{xy}/S_{xx}\end{pmatrix}$$

(b) An *intercept-only* model is an alternative way to express that univariate data form a random sample. $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ is equivalent to

$$Y_i = \mu + \epsilon_i \; , \quad i = 1, \ldots, n$$

with the standard model assumptions.

i. Write the intercept-only model in matrix form.

Solution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \beta_0 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

ii. Derive the least squares estimator of μ using the general OLS estimator $(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}$.

Solution:

2. For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors:

(a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary?

Solution: Using just the intervals, we could conclude that the p-value for age is significant at a 90% confidence level, but not at 95% because the interval spans over 0 in that case.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data('prostate', package = 'faraway')
View(prostate)
prostate_lm <- lm(lpsa~., data = prostate)
summary(prostate_lm)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
# 90% CI:
confint(prostate_lm, c('age'), level = 0.9)
```

```
##            5 %         95 %
## age -0.0382102 -0.001064151
```

```
#95% CI:
confint(prostate_lm, c('age'), level =0.95)
```

```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267
```

(b) Compute and display a 95% joint confidence region for the parameters associated with `age` and `lbph`. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

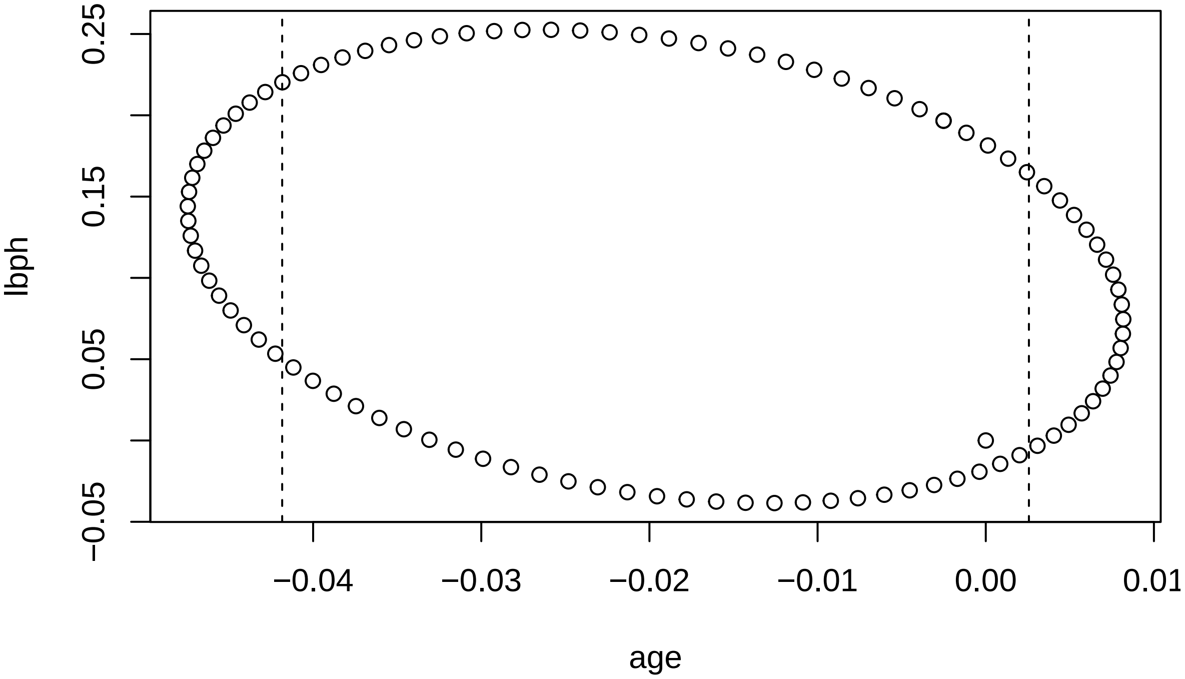
Solution: This display shows the hypothesis test with the null hypothesis age = lbph = 0. Here, we fail to reject the null hypothesis because the origin, (0,0) lies inside the confidence region.

```
library(ellipse)
```

```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
plot(ellipse(prostate_lm, c('age','lbph')))
points(0,0, pch =1)
abline(v=confint(prostate_lm)['age',], lty = 2)
abline(h=confint(prostate_lm)['lbph'], lty =2)
```



(c) In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for `age` in this model. (Hint: `summary(g)$coef[4,3]` gets you the t-statistic you need if the model is called `g`.)

Solution:

```
t_stat <- summary(prostate_lm)$coef[4,3]

x <- numeric(4000)
```

```
for (i in 1:4000){
  model = lm(lpsa~lcavol +lweight+sample(age)+lbph+svi+lcp+gleason+pgg45, data = prostate)
  x[i] = summary(model)$coef[4,3]

}

mean(abs(x) > abs(t_stat))
```

```
## [1] 0.084
```

(d) Remove all the predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

Solution: The new model is not significantly better than the original, so we would prefer the original.

```
prostate_lm_2 <- lm(lpsa~lcavol +lweight +svi, data = prostate)
summary(prostate_lm_2)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```