PSTAT 126 - Assignment 5

Fall 2022

Due: Tuesday, November 8 at 11:59 pm on Gradescope

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:

Assignment 5

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

Recall:

$$S_{xx} = \sum_{i=1}^n (x_i-\bar{x})^2 = \left(\sum_{i=1}^n x_i^2\right) - n\bar{x}^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$$

$$S_{xy} = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$$

Therefore, from (1) we can write:

$$\frac{1}{n\left(\sum_{i=1}^n x_i^2\right)-n^2\bar{x}^2}\begin{pmatrix}n\bar{y}\left(\sum_{i=1}^n x_i^2\right) - n\bar{x}\left(\sum_{i=1}^n x_i y_i\right)\\ n\left(\sum_{i=1}^n x_i y_i\right) - n^2\bar{x}\bar{y}\end{pmatrix}$$

↑ using formulas for $\bar{x}$ and $\bar{y}$.

Now, write:

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

cancelling n terms,

$$\Rightarrow \begin{pmatrix}\frac{\bar{y}\left(\sum_{i=1}^n x_i^2\right) - \bar{x}\left(\sum_{i=1}^n x_i y_i\right)}{\left(\sum_{i=1}^n x_i^2\right)-n\bar{x}^2}\\ \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\left(\sum_{i=1}^n x_i^2\right)-n\bar{x}^2}\end{pmatrix} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$$

Now substitute S_{xx} and S_{xy} and added extra terms to top

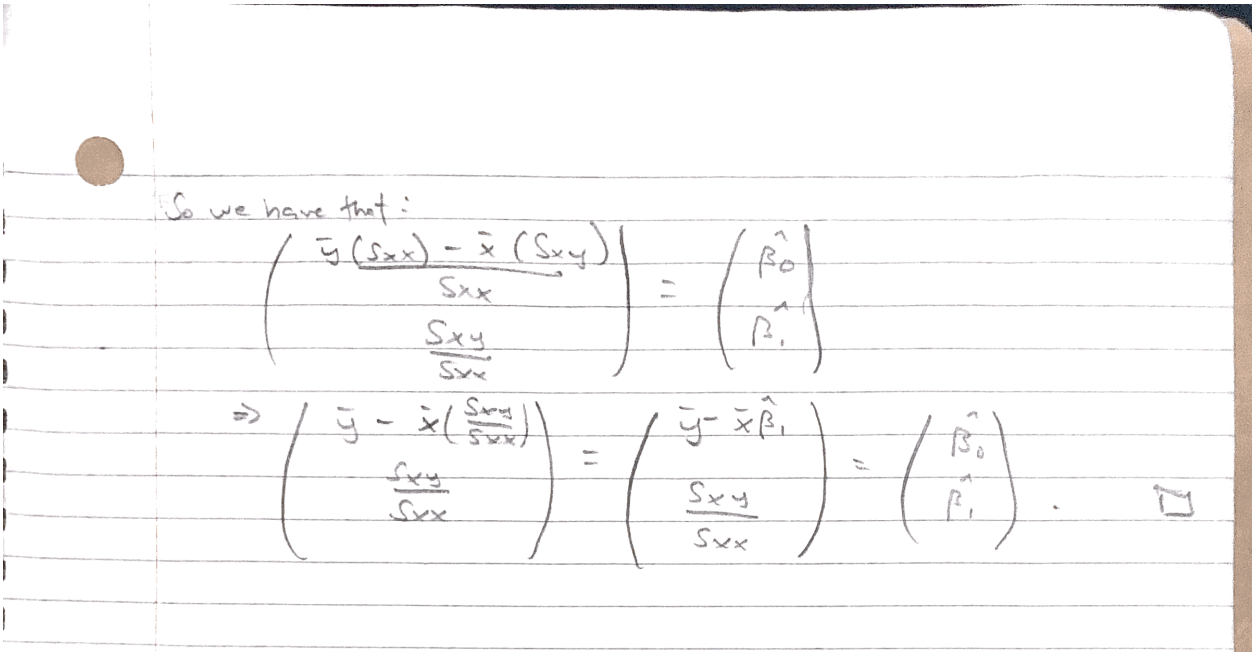
$$\Rightarrow \begin{pmatrix}\frac{\bar{y}\left(\sum_{i=1}^n x_i^2\right) - \bar{x}\left(\sum_{i=1}^n x_i y_i\right) + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y}}{S_{xx}}\\ \frac{S_{xy}}{S_{xx}}\end{pmatrix} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$$

Writing the top of $\hat{\beta}_0$, we can rearrange to:

$$\bar{y}\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) + \bar{x}\left(-\sum_{i=1}^n x_i y_i + n\bar{x}\bar{y}\right)$$

$$\Rightarrow \bar{y}(S_{xx}) - \bar{x}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

$$\Rightarrow \bar{y}(S_{xx}) - \bar{x}(S_{xy}) \quad \text{for top of } \hat{\beta}_0 \text{ term.}$$



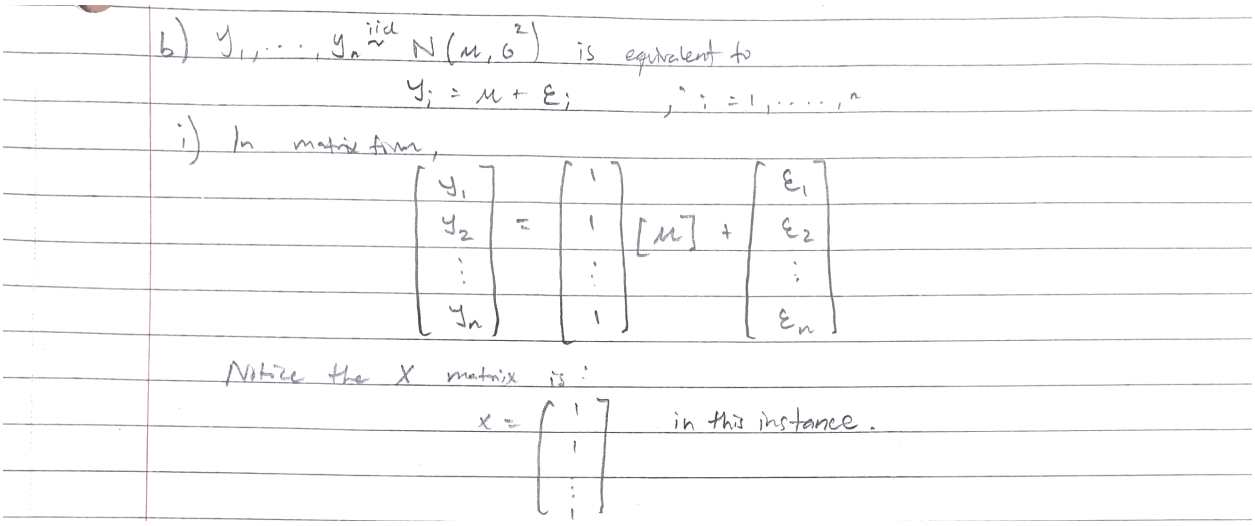
(b) An *intercept-only* model is an alternative way to express that univariate data form a random sample. $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ is equivalent to

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

with the standard model assumptions.

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

Solution:



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

Solution:

ii) Using the OLS estimator $(x^T x)^{-1} x^T y$,

we have that $x^T x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}$

$= [1 + 1 + 1 \dots + 1] = n \Rightarrow (x^T x)^{-1} = \frac{1}{n}$

Also, $x^T y =$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + \dots + y_n \end{bmatrix}$$

Therefore, the LSE of μ is

$$\hat{\mu} = \frac{1}{n}[y_1 + y_2 + \dots y_n] = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$= \bar{y}$$

Thus, $\boxed{\hat{\mu} = \bar{y}}$

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?

```
library(faraway)
```

Solution:

```
data('prostate')
fit <- lm(lpsa ~ .,prostate)
summary(fit)
```

```
##
## 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
```

```
confint(fit, c("age"), .90)
```

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

```
confint(fit, c("age"), .95)
```

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

We see that the 95% CI for age includes 0 but the 90% CI for age doesn't include 0, hence age is not significant on lpsa at 0.05, but it is at 0.1. So, we can expect the p-value to be between 0.05 and 0.1. Looking at the summary, the p-value for age is 0.08229, which lies in the interval as expected.

(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.

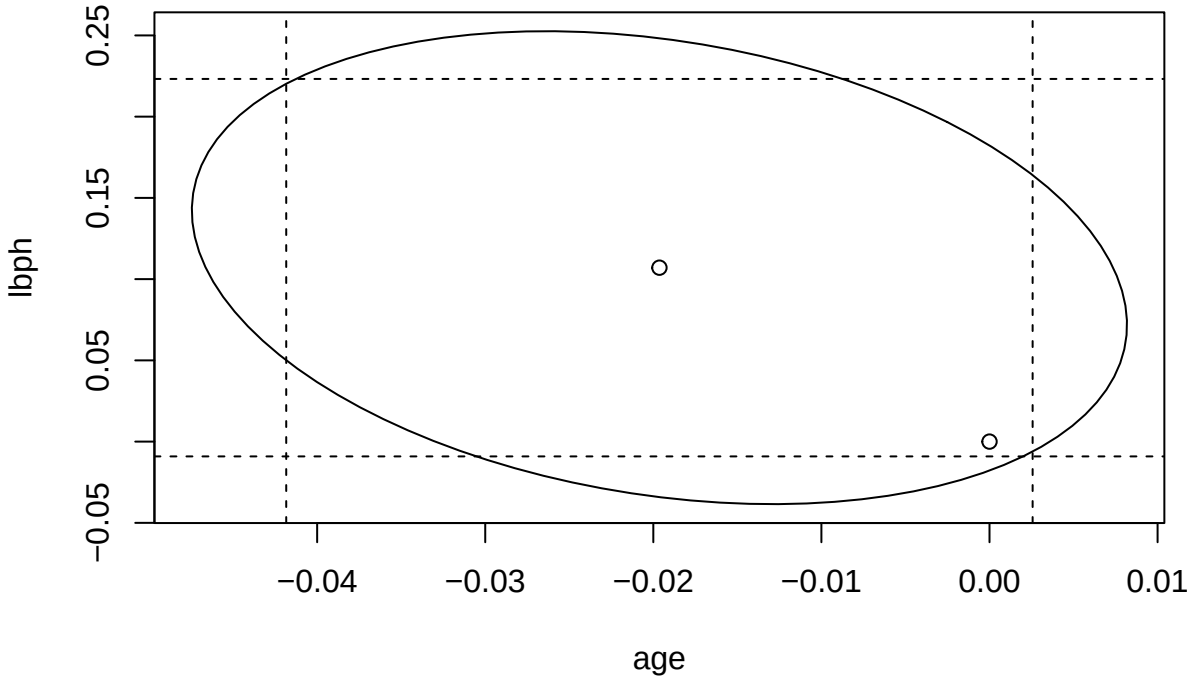
Solution:

```
library(ellipse)
```

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

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

```
plot(ellipse(fit, c('age', 'lbph')), type = "l") # 0.95 level default
points(0, 0)
points(coef(fit)['age'],coef(fit)['lbph'])
abline(v= confint(fit)['age',], lty = 2)
abline(h= confint(fit)['lbph',], lty = 2)
```



We see that the origin lies within the 95% confidence region, therefore we do not reject the null hypothesis, that age = lbph = 0.

(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:

```
tval <- summary(fit) %>% coef() %>% .['age', 't value']

# permutation test
permute <- function(nsims) {
    map_dbl(1:nsims,
            ~ lm(sample(lpsa) ~ ., data = prostate) %>%
            summary() %>%
            coef() %>%
            .['age', 't value'])
}
```

```
mean(abs(permute(10000)) > abs(tval))
```

```
## [1] 0.0842
```

We see that we obtain a p-value approximately equal to 0.08229 (calculated earlier) permutating n-times.

(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:

```
fit1 <- update(fit, . ~ lcavol + lweight + svi)
summary(fit1)
```

```
##
## 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
```

```
anova(fit,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
## Model 2: lpsa ~ lcavol + lweight + svi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     88 44.163
## 2     93 47.785 -5   -3.6218 1.4434 0.2167
```

We can see by looking at the Pr column (0.2167) that there is not a significant improvement in this model compared to the original model. Hence we choose the original model.