Homework 5

Brady Lamson

Problem 7.1

State the degrees of freedom that are associated with each of the following extra sum of squares.

$$SSR(X_1|X_2) : df = 1$$
$$SSR(X_2|X_1, X_3) : df = 1$$
$$SSR(X_1, X_2|X_3, X_4) : df = 2$$
$$SSR(X_1, X_2X_3|X_4, X_5) : df = 3$$

Problem 7.2

$SSR(X_1)$ is an extra sum of squares in the context of a decomposition, such as with $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$.

Problem 7.3*

```r
brands <- read.table("../datasets/CH06PR05.txt")
colnames(brands) <- c("brand_like", "moisture", "sweetness")

brand_lm <- lm(brand_like ~ moisture + sweetness, data = brands)

anova(brand_lm)
```

```
Analysis of Variance Table

Response: brand_like
          Df  Sum Sq Mean Sq F value    Pr(>F)
moisture   1 1566.45 1566.45 215.947 1.778e-09 ***
sweetness  1  306.25  306.25  42.219 2.011e-05 ***
Residuals 13   94.30    7.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part B

$$\begin{aligned} H_0 &: B_2 = 0 \\ H_a &: B_2 \neq 0 \\ F &= 42 \\ p &\approx 2.01 \cdot 10^{-5} \\ \alpha &= .01 \end{aligned}$$

As $p < \alpha$ there is sufficient evidence to reject the null hypothesis that the model should not include X_2.

Problem 7.4**

Part A

```r
grocer <- read.table("../datasets/CH06PR09.txt")
colnames(grocer) <- c("hours", "cases_shipped", "costs", "holiday")

grocer_reg <- lm(hours ~ cases_shipped + holiday + costs, data = grocer)
anova(grocer_reg)
```

```
Analysis of Variance Table

Response: hours
              Df  Sum Sq Mean Sq F value    Pr(>F)
cases_shipped  1  136366  136366  6.6417   0.01309 *
holiday        1 2033565 2033565 99.0443 2.963e-13 ***
costs          1    6675    6675  0.3251   0.57123
Residuals     48  985530   20532
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part B

$$
\begin{aligned}
H_0 &: B_2 = 0 \\
H_a &: B_2 \neq 0 \\
F &= 0.3251 \\
p &\approx .57 \\
\alpha &= .05
\end{aligned}
$$

There is not significant evidence to reject the null hypothesis that the slope of $\beta_2 = 0$. As such costs can be left out of the model given that cases shipped and holidays are kept in.

Part C

We have the look at a different model here.

```r
lm(hours ~ cases_shipped + costs + holiday, data = grocer) |> anova()
```

```
Analysis of Variance Table

Response: hours
              Df  Sum Sq Mean Sq F value    Pr(>F)    
cases_shipped  1  136366  136366  6.6417   0.01309 *  
costs          1    5726    5726  0.2789   0.59987    
holiday        1 2034514 2034514 99.0905 2.941e-13 ***
Residuals     48  985530   20532                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lm(hours ~ costs + cases_shipped, data = grocer) |> anova()
```

```
Analysis of Variance Table

Response: hours
              Df  Sum Sq Mean Sq F value Pr(>F)
costs          1   11395   11395  0.1849 0.6691
```

```
cases_shipped  1  130697  130697  2.1206 0.1517
Residuals     49 3020044   61634
```

We need to pull a few values from these tables here.

$$\begin{aligned}
SSR(X_1) &= 136366 \\
SSR(X_2) &= 11395 \\
SSR(X_1|X_2) &= 130697 \\
SSR(X_2|X_1) &= 5726 \\
SSR(X_1) + SSR(X_2|X_1) &= SSR(X_2) + SSR(X_1|X_2) \\
142092 &= 142092
\end{aligned}$$

These are and must always be equal. Decomposition can be done in either order.

Problem 7.12

For the general R^2 and R^2_{12} as they are both the same in this case:

```r
r2 <- summary(brand_lm)$r.squared |> round(3)
glue::glue("R^2: {r2}")
```

```
R^2: 0.952
```

For R^2_{Y1},R^2_{Y2} I just fit models with one or the other predictor.

```r
x1_brand <- lm(brand_like ~ moisture, data = brands)
x2_brand <- lm(brand_like ~ sweetness, data = brands)

r2y1 <- summary(x1_brand)$r.squared |> round(3)
r2y2 <- summary(x2_brand)$r.squared |> round(3)

glue::glue("R^2_Y1: {r2y1}\nR^2_Y2: {r2y2}")
```

```
R^2_Y1: 0.796
R^2_Y2: 0.156
```

$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$

Of note that:

$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$

```
sse_x2 <- sum((x2_brand$fitted.values - brands$brand_like)^2)
sse_full <- sum((brand_lm$fitted.values - brands$brand_like)^2)

(sse_x2 - sse_full) / sse_x2
```

```
[1] 0.9432184
```

$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$

Of note that:

$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$

```
sse_x1 <- sum((x1_brand$fitted.values - brands$brand_like)^2)
sse_full <- sum((brand_lm$fitted.values - brands$brand_like)^2)

(sse_x1 - sse_full) / sse_x1
```

```
[1] 0.7645737
```

Problem 7.20

In an experimental setting this doesn't always have to be the case. Some experiment have the luxury of being able to very carefully select their predictors and don't have as many external variables to control for.

Problem 7.22

It is not uncommon for more complex models to predict better, even if certain predictors aren't statistically significant. What's important to examine is the relative performance increase of adding that many more predictors because doing so isn't free. A leaner model may be more computationally efficient, may be less likely to overfit the data, and also may have less of a "black box" issue more complex models have.

The problem doesn't specify *how much* better the predictions are. It also doesn't expand on the situations where it doesn't perform better (this better performance only happened in some initial trials).

Problem 7.24

Part A

```
x1_brand$coefficients
```

```
(Intercept)    moisture
     50.775       4.425
```

$$Y_i = 50.775 + 4.425X_1$$

Part B

```
brand_lm$coefficients
```

```
(Intercept)    moisture   sweetness
     37.650       4.425       4.375
```

The coefficients for moisture at the same in both models.

Part C

$1566.45 = 1566.45$. They are the exact same.

Part D

```
cor(brands)
```

```
           brand_like  moisture sweetness
brand_like  1.0000000 0.8923929 0.3945807
moisture    0.8923929 1.0000000 0.0000000
sweetness   0.3945807 0.0000000 1.0000000
```

Moisture and sweetness have a correlation of 0. If two predictors are uncorrelated, adding or removing the other will have 0 impact on their coefficients.

Problem 7.25

Part A

```
groc_x1_reg <- lm(hours ~ cases_shipped, data = grocer)
groc_x1_reg$coefficients
```

```
 (Intercept) cases_shipped
4.079870e+03  9.354971e-04
```

$$Y_i = 4.08 \cdot 10^3 + 9.35 \cdot 10^{-4} X_1$$

Part B

```
grocer_reg$coefficients
```

```
 (Intercept) cases_shipped       holiday         costs
4.149887e+03  7.870804e-04  6.235545e+02 -1.316602e+01
```

The cases shipped coefficient changes, slightly.

Part C

```r
blargh <- lm(hours ~ costs + cases_shipped, data = grocer)
blargh |> anova()
```

```
Analysis of Variance Table

Response: hours
              Df  Sum Sq Mean Sq F value Pr(>F)
costs          1   11395   11395  0.1849 0.6691
cases_shipped  1  130697  130697  2.1206 0.1517
Residuals     49 3020044   61634
```

$136366 \neq 130697$. The difference isnt very substantial though.

Part D

```r
cor(grocer)
```

```
                  hours cases_shipped      costs    holiday
hours         1.0000000    0.20766494 0.06002960 0.81057940
cases_shipped 0.2076649    1.00000000 0.08489639 0.04565698
costs         0.0600296    0.08489639 1.00000000 0.11337076
holiday       0.8105794    0.04565698 0.11337076 1.00000000
```

Cases shipped and costs only have a .08 correlation, which is tiny. This explains the very small shift in coefficient.