Homework 9 Solutions

Due: Thursday 4/23/20 by 8:30am

Rubric:

- Maximum of 2 points each for 1., determined as follows:
  - 0 points for no solutions whatsoever or incomplete solutions;
  - 1 point for solutions provided for each part, but at least one incorrect solution;
  - 2 points for correct solutions to each part;
- Maximum of 3 points for 2.-3., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.

1. Problem 7.3 from the `.pdf` version of the textbook. Requires use of the `brand_preference` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/brand_preference.RData")
load(link)
close(link)
```

The analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1 and X_2, given X_1, is given below:

Source of Variation	SS	df	MS
Regression	1872.70	2	936.35
X_1	1566.45	1	1566.45
$X_2 \vert X_1$	306.25	1	306.25
Error	94.30	13	7.25
Total	1967.00	15	131.13

```
anova(lm(Y~X1 + X2, data = data))
```

(b)

```
n <- nrow(data)
an <- anova(lm(Y ~ X1 + X2, data = data))
F.star <- an$`Mean Sq`[2]/an$`Mean Sq`[3]
F.quantile.01 <- qf(0.99, 1, n  - 3)
p.val <- pf(F.star, 1, n - 3, lower.tail = FALSE)
```

We can test whether X_2 can be dropped from the regression model given that X_1 is retained using an F-test of the null hypothesis $H_0 : \beta_2 = 0$ versus the alternative $H_a : \beta_2 \neq 0$. The decision rule for a level-$\alpha = 0.01$ test based on the test statistic $F^* = MSR(X_2|X_1)/MSE(X_1, X_2)$ would be:

- If $F^* \leq F(0.99; 1; 13)$, conclude H_0
- If $F^* > F(0.99; 1; 13)$, conclude H_a.

Because $F^* = 42.22$ and $F(0.99; 1; 13) = 9.074$, we would reject H_0 and conclude H_a.

The p-value of the test is $P(F > F^*) = 2.0110474 \times 10^{-5}$, the probability that a F random variable with 1 and 13 degrees of freedom is greater than F^*.

2. Problem 7.6 from the `.pdf` version of the textbook. Requires use of the `patient_satisfaction` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/grocery_retailer.RData")
load(link)
close(link)
```

The analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1 and X_2, given X_1, is given below:

Source of Variation	SS	df	MS
Regression	2176606	3	725535.30
X_1	136366	1	136366.00
$X_2 \vert X_1$	5726	1	5726.00
$X_3 \vert X_2, X_1$	2034514	1	2034514.00
Error	985530	48	20531.88
Total	3162136	51	62002.67

```
anova(lm(Y~X1 + X2 + X3, data = data))
```

(b)

```
n <- nrow(data)
an <- anova(lm(Y ~ X1 + X3 + X2, data = data))
F.star <- an$`Mean Sq`[3]/an$`Mean Sq`[4]
F.quantile.025 <- qf(0.975, 1, n  - 4)
p.val <- pf(F.star, 1, n - 4, lower.tail = FALSE)
```

We can test whether X_2 can be dropped from the regression model given that X_1 and X_3 are retained using an F-test of the null hypothesis $H_0 : \beta_2 = 0$ versus the alternative $H_a : \beta_2 \neq 0$. The decision rule for a level-$\alpha = 0.025$ test based on the test statistic $F^* = MSR(X_2|X_1, X_3)/MSE(X_1, X_2, X_3)$ would be:

- If $F^* \leq F(0.975; 1; 48)$, conclude H_0
- If $F^* > F(0.975; 1; 48)$, conclude H_a.

Because $F^* = 0.33$ and $F(0.975; 1; 48) = 5.354$, we would conclude H_0.

The p-value of the test is $P(F > F^*) = 0.5712274$, the probability that a F random variable with 1 and 48 degrees of freedom is greater than F^*.

(c)

```
an12 <- anova(lm(Y ~ X1 + X2, data = data))
an21 <- anova(lm(Y ~ X1 + X2, data = data))

ssr1 <- an12$`Sum Sq`[1]
ssr2 <- an21$`Sum Sq`[1]
ssr21 <- an12$`Sum Sq`[2]
ssr12 <- an21$`Sum Sq`[2]
```

Yes. Because of how we have defined $SSR(X_1)$, $SSR(X_2)$, $SSR(X_1|X_2)$, and $SSR(X_2|X_1)$, it must always be the case that $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2) = SSR(X_1, X_2)$.

3. Problem 7.16 from the `.pdf` version of the textbook. Requires use of the `brand_preference` data that has been posted on the Homework page.

(a)

```
link <- url("http://maryclare.github.io/stat525/content/homework/brand_preference.RData")
load(link)
close(link)

Y.mean <- mean(data$Y)
X1.mean <- mean(data$X1)
X2.mean <- mean(data$X2)

Y.sd <- sd(data$Y)
X1.sd <- sd(data$X1)
X2.sd <- sd(data$X2)

data$Y.std <- (data$Y - Y.mean)/Y.sd
data$X1.std <- (data$X1 - X1.mean)/X1.sd
data$X2.std <- (data$X2 - X2.mean)/X2.sd

linmod.std <- lm(Y.std ~ X1.std + X2.std, data = data)
b.star.0 <- linmod.std$coef[1]
b.star.1 <- linmod.std$coef[2]
b.star.2 <- linmod.std$coef[3]
```

We obtain estimated regression coefficents $b_0^* = 0$, $b_1^* = 0.89$, and $b_2^* = 0.39$.

(b)

We obtain $b_1^* = 0.89$, which is the average increase in how much a brand is liked in sample standard deviations when in moisture content X_1 increases by one sample standard deviation.

(c)

```
b.star <- linmod.std$coef
b1 <- b.star[2]*Y.sd/X1.sd
b2 <- b.star[3]*Y.sd/X2.sd
b0 <- Y.mean - b1*X1.mean - b2*X2.mean
```

We obtain the same estimated regression coefficient values that we obtained on Homework 7, $b_0 = 37.650$, $b_1 = 4.425$, $b_2 = 4.375$.