Homework 7

Yunting Chiu

2021-04-09

1. (**7.1**) State the number of degrees of freedom that are associated with each of the following extra sums of squares: SSReg(X1 | X2), SSReg(X2 | X1, X3), SSReg(X1, X2 | X3, X4), SSReg(X1, X2, X3 | X4, X5).

A note about the notation. SSReg(A | B) is the extra sum of squares that appeared as aresult of including variables A into the regression model that already had variables B in it. Thus, it is used to compare the full model with both A and B in it against the reduced model with only B.

Ans: We can calculate degrees of freedom by counting the number of variables to the left of the "|". - SSReg(X1 | X2) = 1 - SSReg(X2 | X1, X3) = 1 - SSReg(X1, X2 | X3, X4) = 2 - SSReg(X1, X2, X3 | X4, X5) = 3

2. (**7.2**) Explain in what sense the regression sum of squares SSReg(X1) is an extra sum of squares.

- Extra sum of squares uses extra sums of squares in tests for regression coefficients. For example, there is a response variable Y and 2 predictor variables X1 and X2:
- The reduce model is Y = $\beta 0$ + $\beta 1$X1 + ei and compute SSE(X1)
- The full model is Y = $\beta 0$ + $\beta 1$X1 + $\beta 2$X2 + ei and compute SSE(X1, X2)
- So the equation can be denoted as SSE(X1) = SSE(X1, X2) + SS? How can we define SS? As the extra sum of squares and denote it by SSR(X2|X1) so we can write as

$$SSR(X2|X1) = SSE(X1) - SSE(X1, X2)$$

- SSR(X2|X1) calculates the decrease in SSE when X2 is added to the regression model, given X1 is already present.

Reference: - https://365datascience.com/tutorials/statistics-tutorials/sum-squares/ - https://www.stat.colostate.edu/~riczw/teach/STAT540_F15/Lecture/lec09.pdf

3. (**7.28b**) For a multiple regression model with five X variables, what is the relevant extra sum of squares for testing

The equation might be:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + ei$$

(a) whether or not $\beta 5 = 0$?
- SSR(X5 | X2, X3, X4, X5)
(b) whether or not $\beta 2 = \beta 4 = 0$?
- SSR(X2, X4 | X1, X3, X5)

4. (**7.28b, Stat-615 only**) Show that SSReg(X1, X2, X3, X4) = SSReg(X2, X3)+SSReg(X1|X2, X3)+SSReg(X4 | X1, X2, X3)

Reference: - https://www.stat.colostate.edu/~riczw/teach/STAT540_F15/Lecture/lec09.pdf - https://www.math.arizona.edu/~piegorsch/571A/STAT571A.Ch07.pdf

4. $SSReg(X_1, X_2, X_3, X_4) = SSReg(X_2, X_3) + SSReg(X_1 | X_2, X_3) + SSReg(X_4 | X_1, X_2, X_3)$ Prove it!

$$SSReg(X_1 | X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3) = SSR(X_1, X_2, X_3) - SSR(X_2, X_3)$$

$$SSReg(X_4 | X_1, X_2, X_3) = SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4) = SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3)$$

$$SSReg(X_1, X_2, X_3, X_4) = \cancel{SSReg(X_2, X_3)} + \cancel{SSR(X_1, X_2, X_3)} - \cancel{SSR(X_2, X_3)} + SSR(X_1, X_2, X_3, X_4) - \cancel{SSR(X_1, X_2, X_3)}$$

$$= SSR(X_1, X_2, X_3, X_4) \ \#$$

5. (**7.3, 7.24, 7.30**) Continue working with the Brand Preference data, which are available on our Blackboard, on http://statweb.lsu.edu/EXSTWeb/StatLab/DataSets/NKNWData/CH06PR05.txt, and in the previous homework.

Recall the variables: It was collected to study the relation between degree of brand liking (Y) and moisture content (X1) and sweetness (X2) of the product.

(a) Obtain the ANOVA table that decomposes the regression sum of squares into extra sum of squares associated *with X1* and *with X2, given X1*.

- SSR(X1) = 1566.45
- SSR(X2|X1) = 306.25

```r
brand <- read.table("./data/CH06PR05.txt")
brand %>%
  rename(Y = V1, X1 = V2, X2 = V3) -> brand

# SSR(X1)
X1 <- lm(Y ~ X1, data = brand)

# SSR(X2|X1)
X2givenX1 <- lm(Y ~ X1 + X2, data = brand)

anova(X1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45  54.751 3.356e-06 ***
## Residuals 14  400.55   28.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(X2givenX1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## X2         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
```

(b) Test whether X2 can be dropped from the model while X1 is retained.

Consider dropping X2, the hypothesis is H0: $\beta 2 = 0$ vs $\beta 2 \neq 0$. According to the analysis of variance table above, the p-value of X2 is 2.011e-05, indicating that there is evidence that $\beta 2 \neq 0$, so X2 cannot be removed from the model.

(c) Fit first-order simple linear regression for relating brand liking (Y) to moisture content (X1).

```r
summary(X1)$coefficients[, 1]
```

```
## (Intercept)          X1
##      50.775       4.425
```

$$\hat{Y} = 50.775 + 4.425X_1$$

(d) Compare the estimated regression coefficient for X1 with the corresponding coefficient obtained in (a).

- In the `X2givenX1` model, the estimated regression coefficient for X1 is 4.425.
- In the `X1` model, the estimated regression coefficient for X1 is 4.425, too.

```r
summary(X2givenX1)$coefficients[2,1]
```

```
## [1] 4.425
```

```r
summary(X1)$coefficients[2,1]
```

```
## [1] 4.425
```

(e) Does SSreg(X1) equal SSreg(X1|X2) here? Is the difference substantial?

- There are no different between sum of squares of X1. The first model SSReg(X1) is 1566.45, and the second model SSReg(X1|X2) is 1566.45.

```r
# SSReg(X1)
anova(X1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45  54.751 3.356e-06 ***
## Residuals 14  400.55   28.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# SSReg(X1|X2)
X1givenX2 <- lm(Y ~ X2 + X1, data = brand)
anova(X1givenX2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f)

- Regress Y on X2 and obtain the residuals.

```
lm.fit.5fa <- lm(Y ~ X2 , data = brand)
lm.fit.5fa$residuals -> a
a
```

```
##       1       2       3       4       5       6       7       8       9      10
## -13.375 -13.125 -16.375 -10.125  -5.375  -6.125  -6.375  -3.125   5.625   2.875
##      11      12      13      14      15      16
##   8.625   6.875  10.625   8.875  16.625  13.875
```

- Regress X1 on X2 and obtain the residuals.

```
lm.fit.5fb <- (lm(X1 ~ X2, data = brand))
lm.fit.5fb$residuals -> b
b
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
## -3 -3 -3 -3 -1 -1 -1 -1  1  1  1  1  3  3  3  3
```

- Regress residuals from the model "Y on X2" on residuals from the model "X1 on X2"; compare the estimated slope, error sum of squares with #1. What about R^2?
- The regression of `Y` on `X2`: estimated slope is 4.375, SSE is 1660.75, R^2 is 0.1557.
- The regression of `Y` and `X1` on `X2`: estimated slope is 4.425, SSE is 94.3, R^2 is 0.9432.
- Because these two model are not regressing the same size so these two R-squares are completely different.

```
lm.fit.5fc <- lm(a ~ b)
summary(lm.fit.5fa) # lm(Y ~ X2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = brand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.375  -7.312  -0.125   8.688  16.625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68.625      8.610   7.970 1.43e-06 ***
## X2             4.375      2.723   1.607     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 14 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.09539
## F-statistic: 2.582 on 1 and 14 DF,  p-value: 0.1304
```

```
anova(lm.fit.5fa)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## X2         1  306.25  306.25  2.5817 0.1304
```

```
## Residuals 14 1660.75  118.62
```

```r
summary(lm.fit.5fc)
```

```
##
## Call:
## lm(formula = a ~ b)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.718e-17  6.488e-01    0.00        1
## b            4.425e+00  2.902e-01   15.25 4.09e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.595 on 14 degrees of freedom
## Multiple R-squared:  0.9432, Adjusted R-squared:  0.9392
## F-statistic: 232.6 on 1 and 14 DF,  p-value: 4.089e-10
```

```r
anova(lm.fit.5fc)
```

```
## Analysis of Variance Table
##
## Response: a
##           Df Sum Sq Mean Sq F value    Pr(>F)
## b          1 1566.5 1566.45  232.56 4.089e-10 ***
## Residuals 14   94.3    6.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. (**8.13**) Consider a regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, where X1 is a numerical variable, and X2 is a dummy variable. Sketch the response curves (the graphs of E(Y) as a function of X1 for different values of X2), if $\beta 0 = 25$, $\beta 1 = 0.2$, and $\beta 2 =$ -12.

- The blue line indicates the association between E(Y) and X1 when X2 = 0
- The green line indicates the association between E(Y) and X1 when X2 = 1

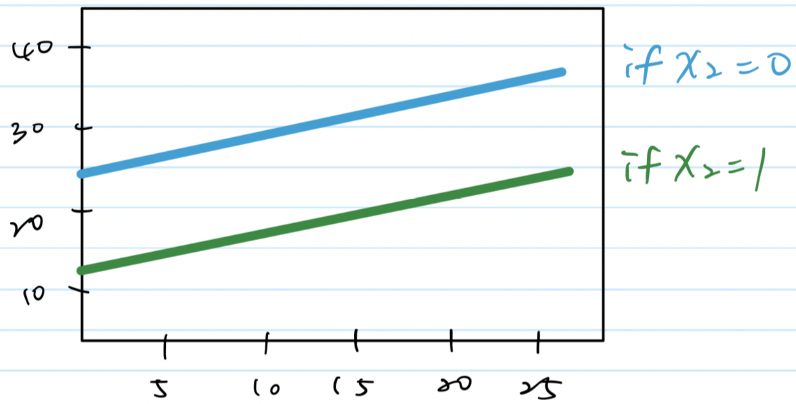
6. $Y = 25 + 0.2X_1 - 12X_2 + \varepsilon$

$E\{Y\} = 25 + 0.2X_1 + (-12)X_2$

As X_2 is a dummy variable, so the equation can be denoted as:

if $X_2 = 0$ $\quad E\{Y\} = 25 + 0.2X_1$

if $X_2 = 1$ $\quad E\{Y\} = 25 + 0.2X_1 - 12 = 13 + 0.2X_1$



7. Continue the previous exercise. Sketch the response curves for the model with interaction, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$, given that β_3 = -0.2

- The red line indicates the association between E(Y) and X1 when X2 = 0
- The green line indicates the association between E(Y) and X1 when X2 = 1

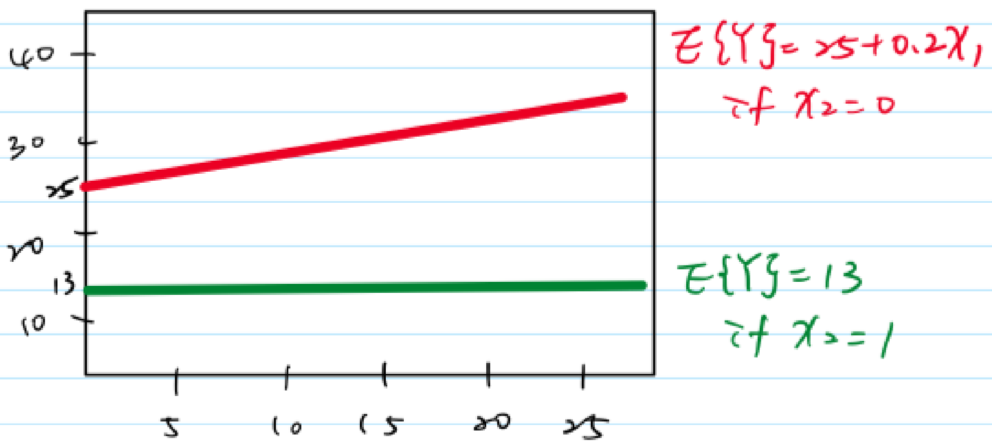
7. $Y = 25 + 0.2X_1 + (-12)X_2 + (-0.2)X_1X_2 + \varepsilon$

$E\{Y\} = 25 + 0.2X_1 + (-12)X_2 + (-0.2)X_1X_2$

if $X_2 = 0 \Rightarrow E\{Y\} = 25 + 0.2X_1$

if $X_2 = 1 \Rightarrow E\{Y\} = 25 + 0.2X_1 - 12 - 0.2X_1$

$= 25 - 12 = 13$



8. (**8.34**) In a regression study, three types of banks were involved, namely, (1) commercial, (2) mutual savings, and (3) savings and loan. Consider the following dummy variables for the type of bank:

Type of Bank	X_2	X_3
Commerical	1	0
Mutual Saving	0	1
Saving and loan	0	0

(a) Develop the first-order linear regression model (no interactions) for relating last year's profit or loss (Y) to size of bank (X1) and type of bank (X2, X3).

$$Yi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ei$$

(b) State the response function for the three types of banks.

- In this data, we can see the `X2` and `X3` are dummy variables. Also, `Y` represents profit or loss, `X1` represents the size of bank.

(b) Method: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Commercial $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2$

Mutual saving $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_3$

Saving and Loan $E\{Y\} = \beta_0 + \beta_1 X_1$

(c) Interpret each of the following quantities: (1) β_2, (2) β_3, (3) $\beta_2 - beta_3$.

1. β_2: The difference between the commercial bank's and the savings and loan bank's expected profit or loss.

2. β_3: The difference between the mutual saving bank's and the savings and loan bank's expected profit or loss.
3. $\beta_2 - \beta_3$: The difference between the mutual saving bank's and the commercial bank's expected profit or loss.
4. (**8.16, 8.20**) Refer to our old GPA data

An assistant to the director of admissions conjectured that the predictive power of the model could be improved by adding information on whether the student had chosen a major field of concentration at the time the application was submitted. Suppose that the first 10 students chose their major when they applied.

```r
GPA <- read.table("./data/CH01PR19.txt")

GPA %>%
  rename(Y = "V1", X1 = "V2") -> GPA
# Suppose that the first 10 students chose their major when they applied.
GPA %>%
  mutate(X2 = 0) -> GPA
GPA$X2[1:10] = 1
head(GPA)
```

```
##       Y X1 X2
## 1 3.897 21  1
## 2 3.885 14  1
## 3 3.778 28  1
## 4 2.540 22  1
## 5 3.028 21  1
## 6 3.865 31  1
```

(a) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, where X1 is the entrance test score and X2 = 1 if a student has indicated a major at the time of application, otherwise X2 = 0. State the estimated regression function.

```r
lm.fit <- lm(Y ~ X1 + X2, data = GPA)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81035 -0.33271  0.02987  0.44702  1.15523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11062    0.32220   6.551  1.6e-09 ***
## X1           0.03871    0.01282   3.018  0.00312 **
## X2           0.07728    0.20663   0.374  0.70910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6254 on 117 degrees of freedom
## Multiple R-squared:  0.07373,    Adjusted R-squared:  0.05789
## F-statistic: 4.656 on 2 and 117 DF,  p-value: 0.01133
```

State the Estimated Regression Function

$$\hat{Y} = 2.11062 + 2.11062X_1 + 0.07728X_2$$

(b) Test whether X2 can be dropped from the model, using $\alpha = 0.05$.

Significance of the whole model is tested by H0 : $\beta_2 = 0$ vs H1 : $\beta_2 \neq 0$. With a large p-value 0.7091 and a small test statistic F = 0.1399, we fail to reject the null hypothesis, meaning that we have no evidence to conclude that `X2` is significant so `X2` may be remove from the model.

```
lm.fit.dropedX2 <- lm(Y ~ X1, data = GPA)
anova(lm.fit, lm.fit.dropedX2)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    117 45.763
## 2    118 45.818 -1 -0.054703 0.1399 0.7091
```

(c) Fit the regression model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + e$ and state the estimated regression function. Interpret β_3. Test significance of the interaction term.

```
# interaction term
lm.fit.interaction <- lm(Y ~ X1 * X2, data = GPA)
summary(lm.fit.interaction)
```

```
##
## Call:
## lm(formula = Y ~ X1 * X2, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47832 -0.31337  0.04355  0.45001  1.07374
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.83364    0.33492   5.475 2.57e-07 ***
## X1           0.04992    0.01336   3.738  0.00029 ***
## X2           2.49114    1.00135   2.488  0.01428 *
## X1:X2       -0.09635    0.03915  -2.461  0.01531 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6123 on 116 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.09694
## F-statistic: 5.258 on 3 and 116 DF,  p-value: 0.001947
```

State the Estimated Regression Function

$$\hat{Y} = 1.83364 + 0.04992X_1 + 2.49114X_2 - 0.09635X_1X_2$$

- If X2 = 0: $\hat{Y} = 1.83364 + 0.04992X_1$
- If X2 = 1: $\hat{Y} = 1.83364 + 0.04992X_1 + 2.49114 - 0.09635X_1 = 4.32478 - 0.04645X_1$

- As previous stated, `X2` $= 1$ is the student has indicated a major at the time of application, otherwise `X2` is 0. The estimated value of $\beta 3$ is -0.09635, indicating that there is a expected difference value on GPA between the students in these two groups (whether they have indicated a major or not).