Stat 615/415 (Regression) HW # 7 – Solutions

Multivariate Regression – inference, tests, model comparison, categorical predictors
chap. 7–8 (7.1–7.3, 8.2–8.5)

1. (**7.1**) State the number of degrees of freedom that are associated with each of the following extra sums of squares: $\text{SSReg}(X_1 \mid X_2)$, $\text{SSReg}(X_2 \mid X_1, X_3)$, $\text{SSReg}(X_1, X_2 \mid X_3, X_4)$, $\text{SSReg}(X_1, X_2, X_3 \mid X_4, X_5)$.

   A note about the notation. SSReg($A \mid B$) is the extra sum of squares that appeared as a result 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.

   SOLUTION.
   For SSReg($X_1 \mid X_2$), $DF_{ex} =$ 1
   For SSReg($X_2 \mid X_1, X_3$), $DF_{ex} =$ 1
   For SSReg($X_1, X_2 \mid X_3, X_4$), $DF_{ex} =$ 2
   For SSReg($X_1, X_2, X_3 \mid X_4, X_5$). $DF_{ex} =$ 3

2. (**7.2**) Explain in what sense the regression sum of squares $\text{SSReg}(X_1)$ is an extra sum of squares.

   SOLUTION. *When we compare the full model $Y = \beta_0 + \beta_1 X + \varepsilon$ against the reduced model $Y = \beta_0 + \varepsilon$, the extra sum of squares is SSReg(X_1). It is the difference between SSErr(Reduced) $= \sum(Y_i - \bar{Y})^2 =$ SSTot and SSErr(Full) $= \sum(Y_i - \hat{Y}_i)^2 =$ SSErr, and SSTot $-$ SSErr $=$ SSReg $=$ SSReg(X_1).*

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

   (a) whether or not $\beta_5 = 0$?

   (b) whether or not $\beta_2 = \beta_4 = 0$?

   SOLUTION.

   (a) SSReg($X_5 \mid X_2, X_3, X_4, X_5$)

   (b) SSReg($X_2, X_4 \mid X_1, X_3, X_5$)

4. (**7.28b, Stat-615 only**)
   Show that $\text{SSReg}(X_1, X_2, X_3, X_4) = \text{SSReg}(X_2, X_3) + \text{SSReg}(X_1 \mid X_2, X_3) + \text{SSReg}(X_4 \mid X_1, X_2, X_3)$.

   SOLUTION. *SSReg($X_1 \mid X_2, X_3$) $=$ SSReg(X_1, X_2, X_3) $-$ SSReg(X_2, X_3)*
   and SSReg($X_4 \mid X_1, X_2, X_3$) $=$ SSReg(X_1, X_2, X_3, X_4) $-$ SSReg(X_1, X_2, X_3).

   Adding these SS to SSReg(X_2, X_3), we obtain

   $$\begin{aligned} & \text{SSReg}(X_2, X_3) + \text{SSReg}(X_1 \mid X_2, X_3) + \text{SSReg}(X_4 \mid X_1, X_2, X_3) \\ = \; & \text{SSReg}(X_2, X_3) + \text{SSReg}(X_1, X_2, X_3) - \text{SSReg}(X_2, X_3) + \text{SSReg}(X_1, X_2, X_3, X_4) - \text{SSReg}(X_1, X_2, X_3) \\ = \; & \text{SSReg}(X_1, X_2, X_3, X_4). \end{aligned}$$

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.

   (a) Obtain the ANOVA table that decomposes the regression sum of squares into extra sum of squares associated with X_1 and with X_2, given X_1.

   (b) Test whether X_2 can be dropped from the model while X_1 is retained.

   (c) Fit first-order simple linear regression for relating brand liking (Y) to moisture content (X_1).

(d) Compare the estimated regression coefficient for X_1 with the corresponding coefficient obtained in (a).

(e) Does $SS_{reg}(X_1)$ equal $SS_{reg}(X_1|X_2)$ here? Is the difference substantial?

(f) Regress Y on X_2 and obtain the residuals.

Regress X_1 on X_2 and obtain the residuals.

Regress residuals from the model "Y on X_2" on residuals from the model "X_1 on X_2"; compare the estimated slope, error sum of squares with the regression of Y on X_1, X_2. What about R^2?

SOLUTION. *Results are based on the R code and output below.*

(a)

	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			

From this table, $SSReg(X_1) =$ 1566.45 , $SSReg(X_2 \mid X_1) =$ 306.25 , and together they form $SSReg = SSReg(X_1, X_2) =$ 1872.70 .

(b) *$H_0 : \beta_2 = 0$ is rejected in favor of $H_1 : \beta_2 \neq 0$, with a low p-value $p = 0.00002011$. Thus, X_2 cannot be dropped from the model.*

(c) $\widehat{Y} = 50.775 + 4.425X_1$.

(d) *In the reduced model, $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, the estimated slope is $b_1 =$ 4.425 . In the full model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, $b_1 =$ 4.425 too.*

(e) *In the reduced model, $SS_{reg}(X_1) =$ 1566.45 . The extra sum of squares between the reduced and the full models is $SS_{reg}(X_1|X_2) =$ 1566.45 too.*

This is not always the case. Here, the slopes and the sums of squares happen to coincide because predictors X_1 and X_2 are orthogonal , i.e., $r_{X_1,X_2} = 0$. It means that they do not share any common information, and the contribution of one variable is exactly the same whether the other variable is in the model or not.

(f) *In the regression of Y on X_1, X_2, we have $b_1 = 4.425$, $SSErr = 1566.5$, and $R^2 = 0.9521$.*
In the regression of residuals, we have $b_1 = 4.425$, $SSErr = 1566.5$, and $R^2 = 0.0.9432$.

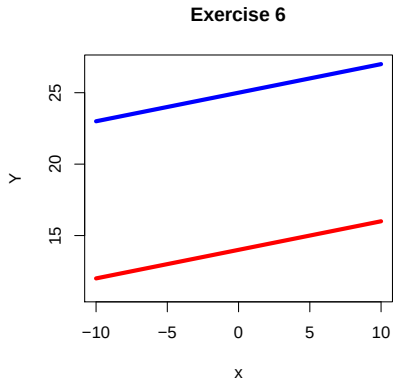
We knew about the equality of slopes and SSErr. But R^2 are different because their denominators SSTot are different. For the full model, $SSTot = \sum(Y_i - \bar{Y}_i)^2$. For the regression of residuals, $SSTot = \sum e_i^2 = SSErr(X_2)$.

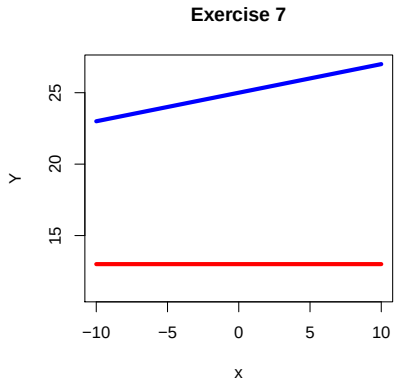
6. **(8.13)** Consider a regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_1 is a numerical variable, and X_2 is a dummy variable. Sketch the response curves (the graphs of $\mathbf{E}(Y)$ as a function of X_1 for different values of X_2), if $\beta_0 = 25$, $\beta_1 =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

SOLUTION. *See the picture on the left. No interaction, lines are parallel.*

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

SOLUTION. *See the picture on the right. Interaction results in different slopes.*





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
Commercial	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 (X_1) and type of bank (X_2, X_3).

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

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

SOLUTION.

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

(b) $$\textit{Profit or loss} = \begin{cases} (\beta_0 + \beta_2) + \beta_1(Size) & \textit{for commercial banks} \\ (\beta_0 + \beta_3) + \beta_1(Size) & \textit{for mutual saving banks} \\ \beta_0 + \beta_1(Size) & \textit{for saving and loan banks} \end{cases}$$

(c) *β_2 is the expected difference in profit or loss between the commercial banks and the saving and loan banks.*

β_3 is the expected difference in profit or loss between the mutual saving banks and the saving and loan banks.

$(\beta_2 - \beta_3)$ is the expected difference in profit or loss between the commercial banks and the mutual saving banks.

9. **(8.16, 8.20)**

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.

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

(b) Test whether X_2 can be dropped from the model, using $\alpha = 0.05$.

(c) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ and state the estimated regression function. Interpret β_3. Test significance of the interaction term.

SOLUTION. *Results are based on the R code and output below.*

(a) $\hat{Y} = 2.1106 + 0.0387X_1 + 0.0773X_2$

(b) *The P-value of the partial F-test of* $H_0 : \beta_2 = 0$ *vs* $\beta_2 \neq 0$ *is* $p = 0.70910$. *Hence,* H_0 *is not rejected, and there is no evidence that the dummy variable* X_2 *is significant, in addition to* X_1 *(ACT score) that is already in the model.*

(c) $\hat{Y} = 1.8336 + 0.0499X_1 + 2.4911X_2 - 0.0964X_1X_2$.

The slope β_3, *estimated here by* $b_3 = -0.0964$ *is the difference in slopes between the GPA of students who indicated their major at the time of their applications and students who did not. For the first group of students, the estimated regression equation is* $\hat{Y} = 4.3248 - 0.0464X_1$. *For the second group,* $\hat{Y} = 1.8336 + 0.0499X_1$.

R Code and Output for Problem #5

```
 # Enter the data. For example, from the internet...
> A = read.table(url("http://statweb.lsu.edu/EXSTWeb/StatLab/DataSets/NKNWData/CH06PR05.txt"))
> names(A)
[1] "V1" "V2" "V3"
> attach(A)
> Y=V1; X1=V2; X2=V3;

 # Full model with both X1 and X2
> full = lm(Y ~ X1 + X2)
> anova(full)               # Solving (a,b)
          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

 # Reduced model with X1 only (c)
> reducedX1 = lm(Y ~ X1)
> reducedX1
(Intercept)           X1
     50.775        4.425

 # Compare estimated slopes (d)
> full
(Intercept)           X1           X2
     37.650        4.425        4.375

 # Compare SSR (e)
> anova(reducedX1)
          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
```

```
 # Explanation: X1 and X2 are orthogonal
> cor(X1,X2)
[1] 0

 # To find SSR(X1|X2) instead of SSR(X2|X1), just switch the order of X1 and X2
> anova(lm(Y ~ X2 + X1))
          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

 # Regression of residuals (f)
> ey = residuals(lm( Y ~ X2 ))
> ex = residuals(lm( Y ~ X2 ))
> ex = residuals(lm( X1 ~ X2 ))
> reg = lm(ey ~ ex)
> anova(reg)
          Df Sum Sq Mean Sq F value    Pr(>F)
ex         1 1566.5 1566.45  232.56 4.089e-10 ***
Residuals 14   94.3    6.74

> summary(reg)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.718e-17  6.488e-01    0.00        1
ex           4.425e+00  2.902e-01   15.25 4.09e-10 ***

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

> summary(full)
Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9447
F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

R Code and Output for Problem #9

```
 # Enter the data, look at the first rows, rename variables, and introduce the dummy X2
> A = read.table(url("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/Kut
> attach(A)
> head(A)
     V1  V2
1 3.897 21
2 3.885 14
3 3.778 28
4 2.540 22
5 3.028 21
6 3.865 31
> n = length(V1)
> Y = V1; X1 = V2; X2 = V2; X2[1:10]=1; X2[11:n]=0;
> X2
  [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [62] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

 # Fit the regression model with the dummy but without the interaction
> reg = lm( Y ~ X1 + X2 )
> summary(reg)
            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

 # Add interaction into the model
> reg_interact = lm( Y ~ X1*X2 )     # Another way is lm( Y ~ X1 + X2 + X1:X2 )
> summary(reg_interact)
            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 *
```