Michael Leibert
Math 651
Homework 7

7.7. Refer to Commercial properties Problem 6.18.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_4; with X_1, given X_4; with X_2, given X_1 and X_4; and with X_3, given X_1, X_2 and X_4.

$SSR(X_4) = 67.775$

```
options(stringsAsFactors = FALSE)
options(scipen=999)
cp<-read.table("CommercialProperties.txt")
head(cp)
```

```
##     V1 V2    V3   V4     V5
## 1 13.5  1  5.02 0.14 123000
## 2 12.0 14  8.19 0.27 104079
## 3 10.5 16  3.00 0.00  39998
## 4 15.0  4 10.70 0.05  57112
## 5 14.0 11  8.97 0.07  60000
## 6 10.5 15  9.45 0.24 101385
```

```
colnames(cp)<-c("Y","X1","X2","X3" ,"X4")


anova(  lm(cp$Y ~ cp$X4 )       )
```

```
## Analysis of Variance Table
##
## Response: cp$Y
##           Df  Sum Sq Mean Sq F value       Pr(>F)
## cp$X4      1  67.775  67.775  31.723 0.0000002628 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_1|X_4) = 42.275$

```
anova(  lm(cp$Y ~ cp$X4+cp$X1)  ) #£
```

```
## Analysis of Variance Table
##
## Response: cp$Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## cp$X4      1  67.775  67.775  41.788 8.076e-09 ***
## cp$X1      1  42.275  42.275  26.065 2.275e-06 ***
## Residuals 78 126.508   1.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_2|X_1, X_4) = 27.857$

```
anova(  lm(cp$Y ~ cp$X1+cp$X4+cp$X2)    )
```

```
## Analysis of Variance Table
##
## Response: cp$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## cp$X1      1 14.819  14.819  11.566  0.001067 **
## cp$X4      1 95.231  95.231  74.331 6.439e-13 ***
## cp$X2      1 27.857  27.857  21.744 1.287e-05 ***
## Residuals 77 98.650   1.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_3|X_1, X_2, X_4) = 0.420$

```
anova(  lm(cp$Y ~ cp$X1+cp$X4+cp$X2+cp$X3)       ) #£
## Analysis of Variance Table
##
## Response: cp$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cp$X1      1 14.819  14.819 11.4649  0.001125 **
## cp$X4      1 95.231  95.231 73.6794 8.379e-13 ***
## cp$X2      1 27.857  27.857 21.5531 1.412e-05 ***
## cp$X3      1  0.420   0.420  0.3248  0.570446
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Test whether X3 can be dropped from the regression model given that X_1, X_2 and X_4 are retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

```
anova( lm(cp$Y ~ cp$X1+cp$X4+cp$X2+cp$X3))
## Analysis of Variance Table
##
## Response: cp$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cp$X1      1 14.819  14.819 11.4649  0.001125 **
## cp$X4      1 95.231  95.231 73.6794 8.379e-13 ***
## cp$X2      1 27.857  27.857 21.5531 1.412e-05 ***
## cp$X3      1  0.420   0.420  0.3248  0.570446
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
cp.F<-anova( lm(cp$Y ~ cp$X1+cp$X4+cp$X2+cp$X3))[[4]][4] ; cp.F
## [1] 0.3247534
cp.n=nrow(cp)
cp.alpha<-.01
qf(1-cp.alpha,1,(cp.n-5))
## [1] 6.980578
if( cp.F < qf(1-cp.alpha,1,(cp.n-5))) {print("Fail to reject H0")} else {
print("Accept Ha")}
## [1] "Fail to reject H0"
pf(cp.F,1,cp.n-5) #P-Value
## [1] 0.4295543
```

Test the alternatives:

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

$F(.99; 1, 76) = 6.980578$

The decision rule:

$$\begin{aligned} &\text{If } F^* \leq 6.980578, \text{ conclude } H_0 \\ &\text{If } F^* > 6.980578, \text{ conclude } H_a \end{aligned}$$

Since $F^* = 0.3248 \leq 6.980578$, we fail to reject H_0.

P-Value: 0.4295543

7.15. Refer to Commercial properties Problems 6.18 and 7.7. Calculate R^2_{Y4}, R^2_{Y1}, $R^2_{Y1|4}$, R^2_{Y14}, $R^2_{Y2|14}$, $R^2_{Y3|124}$, and R^2 . Explain what each coefficient measures and interpret your results. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_4?

$R^2_{Y4} = 0.2865058$, measures the proportionate reduction of total variation associated with the use of the predictor variable X_4. When only X_4 is in the model the error sum of squares is reduced by 28.65 percent.

```
summary( lm(cp$Y ~ cp$X4)       )

##
## Call:
## lm(formula = cp$Y ~ cp$X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1390 -0.7930  0.2890  0.9653  3.4415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.378e+01  2.903e-01  47.482  < 2e-16 ***
## cp$X4       8.437e-06  1.498e-06   5.632 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 79 degrees of freedom
## Multiple R-squared:  0.2865,Adjusted R-squared:  0.2775
## F-statistic: 31.72 on 1 and 79 DF,  p-value: 2.628e-07

summary( lm(cp$Y ~ cp$X4)       )[[8]]

## [1] 0.2865058
```

$R^2_{Y1} = 0.06264236$, measures the proportionate reduction of total variation associated with the use of the predictor variable X_1. When only X_1 is in the model the error sum of squares is reduced by 6.264 percent.

```
summary( lm(cp$Y ~ cp$X1)       )

##
## Call:
```

```
## lm(formula = cp$Y ~ cp$X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1759 -0.9545  0.1705  0.9157  4.4444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.64918    0.28978  54.003   <2e-16 ***
## cp$X1       -0.06489    0.02824  -2.298   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.675 on 79 degrees of freedom
## Multiple R-squared:  0.06264,Adjusted R-squared:  0.05078
## F-statistic: 5.279 on 1 and 79 DF,  p-value: 0.02422
```

```
summary( lm(cp$Y ~ cp$X1)       )[[8]]
```

```
## [1] 0.06264236
```

$R^2_{Y1|4} = 0.2504679$, measures the proportionate reduction in the variation in Y remaining after X_4 is included in the model that is gained by also including X_1 in the model. When X_1 is added to the regression model containing X_4, the error sum of squares $SSE(X_4)$ is reduced by 25.05 percent.

```
anova( lm(cp$Y ~ cp$X4+cp$X1 ) )  [[2]][2]/anova( lm(cp$Y ~ cp$X4 ) )  [[2]][2] #£
```

```
## [1] 0.2504679
```

$R^2_{Y14} = 0.4652132$ measures the proportionate reduction of total variation associated with the use of the predictor variables X_1 and X_4. When both X_1 and X_4 are in the model the error sum of squares is reduced by 46.52 percent.

```
sum(anova(lm(cp$Y ~ cp$X4+cp$X1 )) [[2]][1:2])/sum(anova(lm(cp$Y ~ cp$X4+cp$X1 ))[[2]])
```

```
## [1] 0.4652132
```

$R^2_{Y2|14} = 0.2202037$, measures the proportionate reduction in the variation in Y remaining after X_1 and X_4 are included in the model that is gained by also including X_2 in the model. When X_2 is added to the regression model containing X_1 and X_4, the error sum of squares $SSE(X_1, X_4)$ is reduced by 22.02 percent.

```
anova(lm(cp$Y ~ cp$X4+cp$X1+cp$X2))[[2]][3] /anova(lm(cp$Y ~ cp$X4+cp$X1 ))[[2]][3] #£
```

```
## [1] 0.2202037
```

$R^2_{Y3|124} = 0.004254889$ measures the proportionate reduction in the variation in Y remaining after X_1, X_2 and X_4 are included in the model that is gained by also including X_3 in the model. When X_3 is added to the regression model containing X_1, X_2 and X_4, the error sum of squares $SSE(X_1, X_2, X_4)$ is reduced by .43 percent.

```
anova(lm(cp$Y ~ cp$X4+cp$X1+cp$X2+cp$X3))  [[2]][4] /anova(     lm(cp$Y ~ cp$X4+cp$X1 +cp$X2 ))  [
```

```
## [1] 0.004254889
```

$R^2 = 0.5847$ measures the proportionate reduction of total variation associated with the use of the predictor variables X_1, X_2 , X_3 and X_4. When X_1, X_2 , X_3 and X_4 are in the model the error sum of squares is reduced by 58.47 percent.

```
summary(        lm(cp$Y ~ cp$X4+cp$X1+cp$X2+cp$X3)      )  #£

##
## Call:
## lm(formula = cp$Y ~ cp$X4 + cp$X1 + cp$X2 + cp$X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## cp$X4        7.924e-06  1.385e-06   5.722 1.98e-07 ***
## cp$X1       -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## cp$X2        2.820e-01  6.317e-02   4.464 2.75e-05 ***
## cp$X3        6.193e-01  1.087e+00   0.570     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847,Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14

summary(        lm(cp$Y ~ cp$X4+cp$X1+cp$X2+cp$X3)      )  [[8]]#£

## [1] 0.5847496
```

7.28. **a.** Define each of the following extra sums of squares: (1) $SSR\,(X_5|X_1)$; (2) $SSR\,(X_3, X_4|X_1)$; (3) $SSR\,(X_4|X_1, X_2, X_3)$.

(1) $SSR\,(X_5|X_1) = SSE(X_1) - SSE(X_1, X_5)$

Considering the marginal effect of adding X_5 into the model when X_1 is already in the model.

(2) $SSR\,(X_3, X_4|X_1) = SSE(X_1) - SSE(X_1, X_3, X_4)$

Considering the marginal effect of adding X_3 and X_4 into the model when X_1 is already in the model.

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

Considering the marginal effect of adding X_4 into the regression model when X_1, X_2, and X_3 are already in the model.

b. For a multiple regression model with five X variables, what is the relevant extra sum of squares for testing whether or not $\beta_5 = 0$? whether or not $\beta_2 = \beta_4 = 0$?

$$SSE(X_1, X_2, X_3, X_4) - SSE(X_1, X_2, X_3, X_4, X_5) = SSR\,(X_5|X_1, X_2, X_3, X_4)$$

$$SSE(X_1, X_3, X_5) - SSE(X_1, X_2, X_3, X_4, X_5) = SSR\left(X_2, X_4 | X_1, X_3, X_5\right)$$

7.29. Show that:

a. $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3)$

$$\begin{aligned} SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3) =& \cancel{SSR(X_1)} + SSR(X_1, X_2, X_3) \cancel{- SSR(X_1)} - SSR(X_2|X_1) - \\ & SSR(X_1) - SSR(X_3|X_1, X_2) + SSR(X_1, X_2, X_3, X_4) \\ =& SSR(X_1, X_2, X_3) + SSR(X_1, X_2, X_3, X_4) - \\ & \left[SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)\right] \\ =& SSR(X_1, X_2, X_3, X_4) + \cancel{SSR(X_1, X_2, X_3)} - \cancel{\left[SSR(X_1, X_2, X_3)\right]} \\ =& SSR(X_1, X_2, X_3, X_4) \end{aligned}$$

b. $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3)$

$$\begin{aligned} SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3) =& SSR(X_2, X_3) + \cancel{SSR(X_1|X_2, X_3)} - \cancel{SSR(X_1|X_2, X_3)} - \\ & SSR(X_3|X_2) - SSR(X_2) + SSR(X_1, X_2, X_3, X_4) \\ =& SSR(X_2, X_3) - \left[SSR(X_3|X_2) + SSR(X_2)\right] + \\ & SSR(X_1, X_2, X_3, X_4) \\ =& \cancel{SSR(X_2, X_3)} - \cancel{\left[SSR(X_2, X_3)\right]} + SSR(X_1, X_2, X_3, X_4) \\ =& SSR(X_1, X_2, X_3, X_4) \end{aligned}$$

8.20. Refer to Grade point average Problem 1.19. 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. Assume that regression model (8.33) is appropriate, where X_1 is entrance test score and $X_2 = 1$ if student had indicated a major field of concentration at the time of application and 0 if the major field was undecided.

a. Explain how each regression coefficient in model (8.33) is interpreted here.

We see that the $E[Y]$, is a linear function of ACT score, X_1, with the same slope β_1 for both types of students. β_2 indicates how much higher (lower) the response function for declared students is than the one for undeclared students, for any given ACT score. Thus, β_2 measures the differential effect of type of student. In general, β_2 shows how much higher (lower) the mean response line is for the class coded 1 than the line for the class coded 0, for any given level of X_1.

b. Fit the regression model and state the estimated regression function.

$$Y = 2.19842 + 0.03789X_1 - 0.09430X_2$$

```
gpa<-read.table("GradePointAverage.txt")
colnames(gpa)<-c("GPA","ACT")
gpa$IND<-scan("GradePointAverageX2.txt")
(lm(gpa$GPA~gpa$ACT+gpa$IND))
```

```
##
## Call:
## lm(formula = gpa$GPA ~ gpa$ACT + gpa$IND)
##
## Coefficients:
## (Intercept)      gpa$ACT      gpa$IND
##     2.19842      0.03789     -0.09430
```

c. Test whether the X_2 variable can be dropped from the regression model; use $\alpha = .01$. State the alternatives. decision rule. and conclusion.

```
summary(lm(gpa$GPA~gpa$ACT+gpa$IND)) #£
```

```
##
## Call:
## lm(formula = gpa$GPA ~ gpa$ACT + gpa$IND)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70304 -0.35574  0.02541  0.45747  1.25037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19842    0.33886   6.488 2.18e-09 ***
## gpa$ACT      0.03789    0.01285   2.949  0.00385 **
## gpa$IND     -0.09430    0.11997  -0.786  0.43341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6241 on 117 degrees of freedom
## Multiple R-squared:  0.07749,Adjusted R-squared:  0.06172
## F-statistic: 4.914 on 2 and 117 DF,  p-value: 0.008928
```

```
qt(1-.01/2,nrow(gpa)-3)
```

```
## [1] 2.618504
```

Test the alternatives:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

$t(.99, 117) = 2.618504$

The decision rule:

$$\text{If } t^* \leq 2.618504, \text{ conclude } H_0$$
$$\text{If } t^* > 2.618504, \text{ conclude } H_a$$

Since $t^* = 2.618504 < 2.949$, reject H_0.

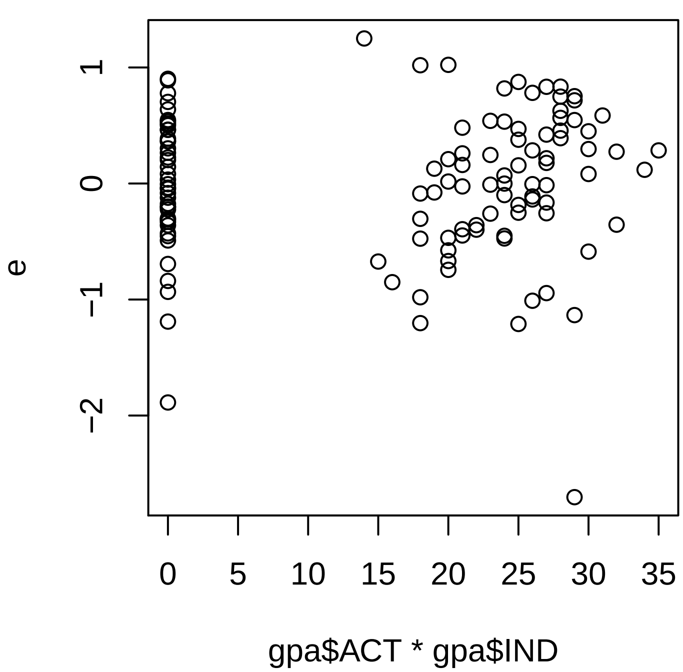
d. Obtain the residuals for regression model (8.33) and plot them against X_1X_2. Is there any evidence in your plot that it would be helpful to include an interaction term in the model?

```r
summary((lm(gpa$GPA~gpa$ACT*gpa$IND))) #£
```

```
##
## Call:
## lm(formula = gpa$GPA ~ gpa$ACT * gpa$IND)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80187 -0.31392  0.04451  0.44337  1.47544
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.226318   0.549428   5.872 4.18e-08 ***
## gpa$ACT         -0.002757   0.021405  -0.129   0.8977
## gpa$IND         -1.649577   0.672197  -2.454   0.0156 *
## gpa$ACT:gpa$IND  0.062245   0.026487   2.350   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 116 degrees of freedom
## Multiple R-squared:  0.1194,Adjusted R-squared:  0.09664
## F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

```r
X<-matrix(0,nrow(gpa),3)
X[,1]<-1
X[,2:3]<-data.matrix(gpa[,2:3])
H<-X%*%solve(t(X)%*%X)%*%t(X)
Y<-data.matrix(gpa[,1])
e<-((diag(nrow(gpa))-H)%*%Y)
```

```r
par(mar=c(5.1, 4.1, 2, 2.1))
plot(gpa$ACT*gpa$IND,e)
```



It appears there is a relationship when the indicator variable is 1, and when it is 0 the errors are located around 0.

8.20. Refer to Grade point average Problems 1.19 and 8.l6.

a. Fit regression model (8.49) and state the estimated regression function.

$3.226318 - 0.002757X_1 - 1.649577X_2 + 0.062245X_1X_2$

```
colnames(gpa)<-c("Y","X1","X2"  )
gpa.n<-nrow(gpa)

gpa$X1X2<-gpa$X2*gpa$X1
lm(gpa$Y~gpa$X1+gpa$X2+gpa$X1X2 )

##
## Call:
## lm(formula = gpa$Y ~ gpa$X1 + gpa$X2 + gpa$X1X2)
##
## Coefficients:
## (Intercept)       gpa$X1       gpa$X2     gpa$X1X2
##    3.226318    -0.002757    -1.649577     0.062245

(lm(gpa$Y~gpa$X1+gpa$X2+gpa$X1X2 )) #£

##
## Call:
## lm(formula = gpa$Y ~ gpa$X1 + gpa$X2 + gpa$X1X2)
##
## Coefficients:
## (Intercept)       gpa$X1       gpa$X2     gpa$X1X2
##    3.226318    -0.002757    -1.649577     0.062245
```

b. Test whether the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives. decision rule, and conclusion. If the interaction term cannot be dropped from the model. describe the nature of the interaction effect.

```
gpa.t<-as.numeric(lm(gpa$Y~gpa$X1+gpa$X2+gpa$X1X2 )[[1]][4] )  /
                as.numeric(summary(lm(gpa$Y~gpa$X1+gpa$X2+gpa$X1X2 ))[[4]][ 4,2])
gpa.t

## [1] 2.350029

gpa.alpha<-(1-.95)/2

qt(1-gpa.alpha,gpa.n-3)

## [1] 1.980448

abs(gpa.t) < qt(1-gpa.alpha,gpa.n-3)

## [1] FALSE

#reject H0
```

Test the alternatives:

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

$t(.95, 116) = 1.980448$

The decision rule:

$$\text{If } t^* \leq 1.980448, \text{ conclude } H_0$$
$$\text{If } t^* > 1.980448, \text{ conclude } H_a$$

Since $t^* = 2.350029 > 1.980448$, reject H_0.

Whether the student declared or was undeclared did have an effect.