LINEAR REGRESSION: Homework #7

Professor Jingchen Liu

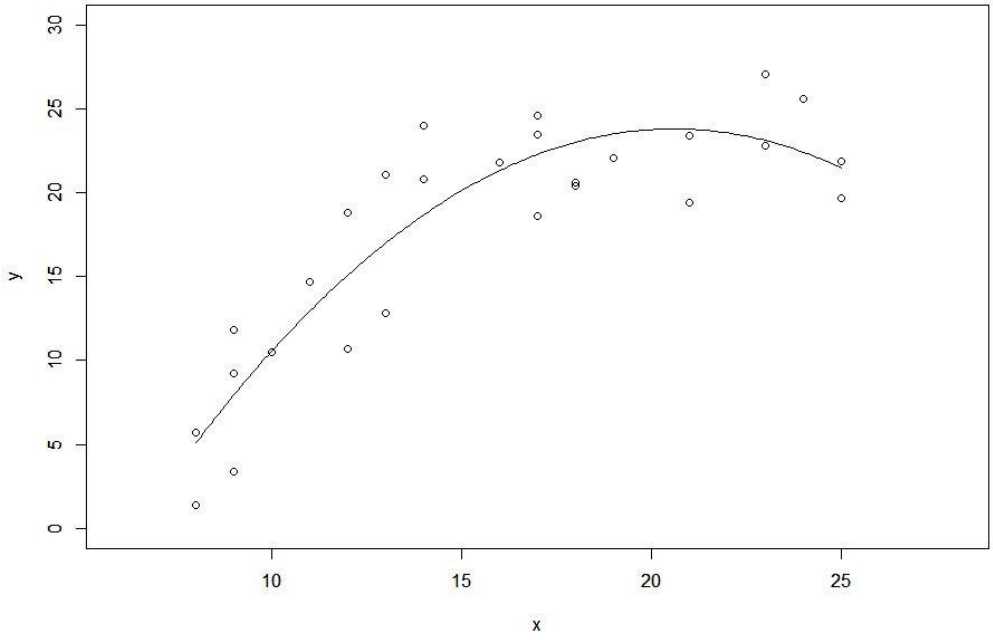
Fan Yang
UNI: fy2232

Problem 1 (8.6)

(a)

```
> x <- d6$V2
> x2 <- d6$V2^2
> model6a <- lm(y ~ x + x2)
> plot(x,y,xlim=c(6,28),ylim=c(0,30))
> coeff <- model6a$coefficients
> coeff
(Intercept)           x          x2
-26.3254125   4.8735744  -0.1184012
>
> f6a <- function(x) {coeff[1] + coeff[2]*x + coeff[3]*x^2}
> x <- c(80:250)/10
> par(new=T)
> plot(x, f6a(x),type = 'l',xlim=c(6,28),ylim=c(0,30),ylab="y")
```

The regression function is $y = -26.33 + 4.87x - 0.12x^2$



The quadratic regression function appears to be a good fit here.

```
> summary(model6a)$r.squared
[1] 0.8143372
```

The R^2 is 0.8143372

(b)

Hypothesis:

$$H_0: \ \beta_1 = \beta_{11} = 0$$
$$H_1: \ \text{not all } \beta = 0$$
$$F^* = \frac{SSR/2}{SSE/(n-3)}$$
$$\text{if } \ F^* \leq F(0.99, 2, n-3), \text{ then conclude } \ H_0$$
$$\text{if } \ F^* > F(0.99, 2, n-3), \text{ then conclude } \ H_1$$

```
> SSR <- sum(anova(model6a)$`Sum Sq`[1:2])
> SSE <- sum(anova(model6a)$`Sum Sq`[3])
> (SSR / 2) / (SSE / 24)
[1] 52.6333
> qf(0.99,2,24)
[1] 5.613591
> 1-pf((SSR/2)/(SSE/24),2,24)
[1] 1.67764e-09
```

Since $F^* = 52.6333 > F(0.99, 2, n-3) = 5.613591$, we can conclude H_1 that not all of the coefficients are 0, which means there is a regression relation. And the P-value is 1.67764e-09.

(c)

When $E\{Y_h\}$ is to be estimated for g levels X_h with family confidence coefficient $1-\alpha$, the Bonferroni confidence limits are:

$$\hat{Y}_h \pm Bs\{\hat{Y}_h\}$$

where

$$B = t(1-\alpha/2g; n-3)$$

and g is the number of confidence intervals in the family.
so $B = t(1-\alpha/2g; n-3) = t(0.9983333, 24) = 3.258373$

```
> Yh1 <- predict(model6a, data.frame(x=10,x2=100),level=0.99)
> Yh2 <- predict(model6a, data.frame(x=15,x2=225),level=0.99)
> Yh3 <- predict(model6a, data.frame(x=20,x2=400),level=0.99)
> X = cbind(rep(1,27),x,x2)
> MSE <- SSE / 24
> s2b <- MSE * solve(t(X)%*%X)
> Xh1 = c(1,10,10^2)
> Xh2 = c(1,15,15^2)
> Xh3 = c(1,20,20^2)
> s2Y1 = t(Xh1)%*%s2b%*%Xh1
> s2Y2 = t(Xh2)%*%s2b%*%Xh2
> s2Y3 = t(Xh3)%*%s2b%*%Xh3
> Yh1;Yh2;Yh3;s2Y1;s2Y2;s2Y3
       1
10.57021
       1
20.13792
       1
23.78558
          [,1]
[1,] 0.8534841
          [,1]
[1,] 0.7970226
          [,1]
[1,] 0.7353646
```

$$10.57021 - 3.258373 * \sqrt{0.8534841} \leq E\{Y_h\} \leq 10.57021 + 3.258373 * \sqrt{0.8534841}$$
$$20.13792 - 3.258373 * \sqrt{0.7970226} \leq E\{Y_h\} \leq 20.13792 + 3.258373 * \sqrt{0.7970226}$$
$$23.78558 - 3.258373 * \sqrt{0.7353646} \leq E\{Y_h\} \leq 23.78558 + 3.258373 * \sqrt{0.7353646}$$

which is

$$7.559985 \leq E\{Y_h\} \leq 13.58043$$
$$17.22897 \leq E\{Y_h\} \leq 23.04687$$
$$20.99141 \leq E\{Y_h\} \leq 26.57974$$

(d)

The 99 percent prediction interval for X_h is:

$$\hat{Y}_h \pm t(1-\alpha/2; n-3)s\{\text{pred}\}$$

where

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\}$$

In this situation, $\hat{Y}_h = 20.13792;\ t(1-\alpha/2; n-3) = t(0.995; 24) = 2.79694$

```
> Yh2 - qt(0.995,24)*sqrt(MSE+s2Y2)
         [,1]
[1,] 10.97342
> Yh2 + qt(0.995,24)*sqrt(MSE+s2Y2)
         [,1]
[1,] 29.30242
```

So the prediction interval is $[10.97342, 29.30242]$

(e)

Hypothesis:

$$H_0: \ \beta_2 = 0$$
$$H_1: \ \beta_2 \neq 0$$
$$\text{if } F^* \leq F(0.99, 1, n-3), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.99, 1, n-3), \text{ then conclude } H_1$$
$$F^* = \frac{SSE(X) - SSE(X, X^2)}{(n-2) - (n-3)} / \frac{SSE(X, X^2)}{n-3}$$

```
> model6e_f <- lm (y ~ x + x2)
> model6e_r <- lm (y ~ x)
> numerator <- sum(model6e_r$residuals^2) - sum(model6e_f$residuals^2)
> denominator <- sum(model6e_f$residuals^2) / model6e_f$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 25.45329
> qf(0.99, 1, model6e_f$df.residual)
[1] 7.822871
> 1 - pf(Fstatistic,1,model6e_f$df.residual)
[1] 3.70783e-05
```

Because $F^* = 25.45329 \geq F(0.99, 1, n-3) = 7.822871$, p-value 3.70783×10^{-5}, then we conclude $H_1: \ \beta_2 \neq 0$. The quadratic term cannot be dropped from the model.

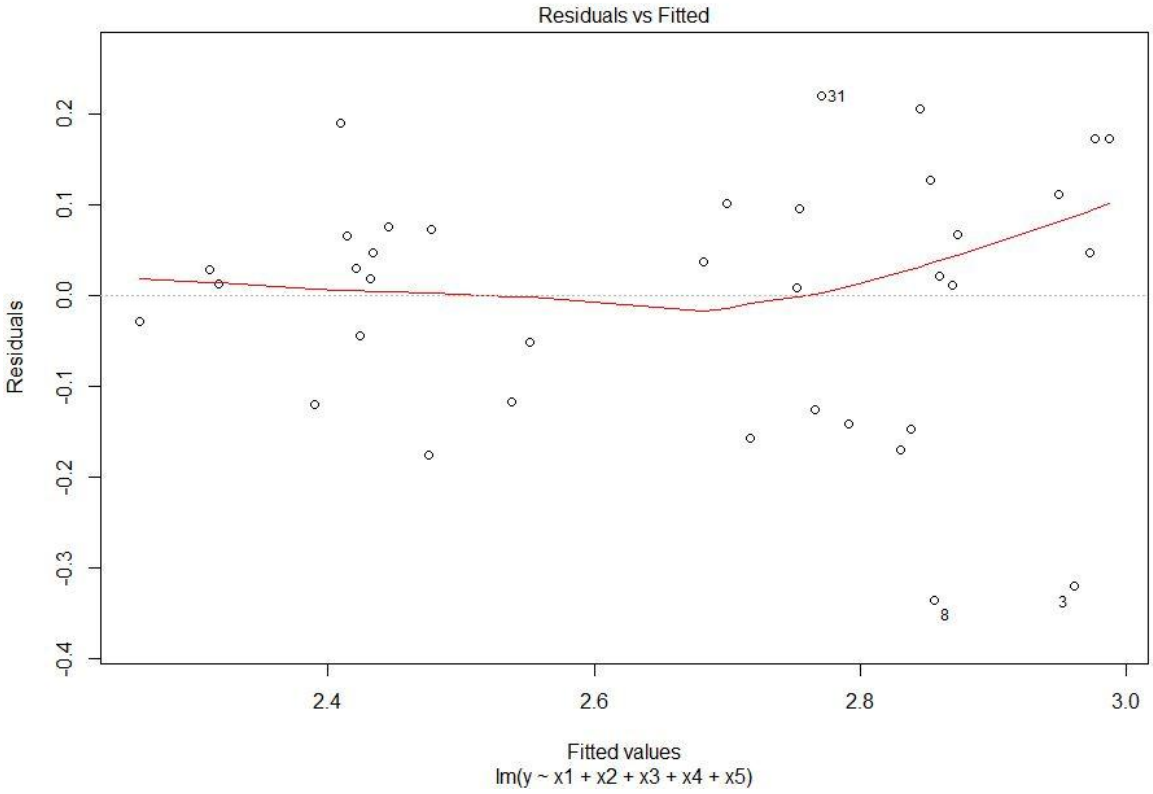
(f)

The regression function is $Y = -26.33 + 4.87X - 0.12X^2$

Problem 2 (8.42)

(a)

```
> y <- da3$‘Market share‘
> x1 <- da3$Price
> x2 <- da3$‘Gross Nielsen rating points‘
> x3 <- factor(da3$‘Discount price‘)
> x4 <- factor(da3$‘Package promotion‘)
> x5 <- factor(ifelse(da3$Year!=2000,da3$Year,0))
>
> model42a <- lm(y~x1 + x2 + x3 + x4 +x5)
> plot(model42a)
```



$$Y = 3.021 - 0.2470x_1 - 0.00009653x_2 + 0.4093x_3\{1\} + 0.124x_4\{1\} + 0.01324x_5\{1999\} - 0.1088x_5\{2001\} - 0.08306x_5\{2002\}$$

The residuals are some small along the fitted value. The first-order model appear to fit the data well.

(b)

Full model:

$$\begin{aligned} Y = \beta_0 &+ \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2 + \beta_4 X_2 \\ &+ \beta_5 X_2^2 + \beta_6 X_3 + \beta_7 X_4 \\ &+ \beta_8 X_5\{1999\} + \beta_9 X_5\{2001\} + \beta_10 X_5\{2002\} \end{aligned}$$

```
> x1_2 <- da3$Price^2
> x2_2 <- da3$`Gross Nielsen rating points`^2
> x12 <- da3$Price * da3$`Gross Nielsen rating points`
> model42b <- lm(y~x1 + x1_2 + x12 + x2 + x2_2 + x3 + x4 +x5)
```

Hypothesis:

$$H_0 : \beta_2 = \beta_3 = \beta_5 = 0$$
$$H_1 : \text{not all } \beta_2, \beta_3, \beta_5 = 0$$
$$\text{if } F^* \leq F(0.95, 3, n-11), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.95, 3, n-11), \text{ then conclude } H_1$$
$$F^* = \frac{SSE(\text{reduced}) - SSE(\text{full})}{(n-8)-(n-11)} \Big/ \frac{SSE(\text{full})}{n-11}$$

```
> numerator <- (sum(model42a$residuals^2) - sum(model42b$residuals^2))/3
> denominator <- sum(model42b$residuals^2) / model42b$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 0.3740023
> qf(0.95, 2, model42b$df.residual)
[1] 3.38519
> qf(0.95, 3, model42b$df.residual)
[1] 2.991241
> 1 - pf(Fstatistic,3,model42b$df.residual)
[1] 0.7724726
```

Because $F^* = 0.3740023 \leq F(0.95, 3, 25) = 2.991241$, with p-value 0.7724726, then we conclude H_0. All quadratic and interaction terms can be dropped from the regression model.

(c)

Full model:

$$\begin{aligned} Y = \beta_0 &+ \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \\ &+ \beta_5 X_5\{1999\} + \beta_6 X_5\{2001\} + \beta_7 X_5\{2002\} \end{aligned}$$

Reduced model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$$

Hypothesis:

$$H_0 : \beta_2 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_1 : \text{not all } \beta_2, \beta_5, \beta_6, \beta_7 = 0$$
$$\text{if } F^* \leq F(0.95, 4, n-8), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.95, 4, n-8), \text{ then conclude } H_1$$
$$F^* = \frac{SSE(\text{reduced}) - SSE(\text{full})}{(n-4)-(n-8)} \Big/ \frac{SSE(\text{full})}{n-8}$$

```
> model42c <- lm(y~x1 + x3 + x4)
> numerator <- (sum(model42c$residuals^2) - sum(model42a$residuals^2))/4
> denominator <- sum(model42a$residuals^2) / model42a$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 0.6817188
> qf(0.95, 4, model42a$df.residual)
[1] 2.714076
> 1 - pf(Fstatistic,4,model42a$df.residual)
[1] 0.6104856
```

Because $F^* = 0.6817188 \leq F(0.95, 4, n-8) = 2.714076$, with p-value 0.6104856, then we conclude H_0. Advertising index (X_2) and year (X_5) can be dropped from the model.

Problem 3 (8.43)

In order to make prediction, we separate the data to training set and test set. Randomly choose 100 sample as the test set and the rest data as the training set. Then our model will be build on training set and be predicted on test set.

```
> y <- da4$GPA
> x1 <- da4$‘High school class rank‘
> x2 <- da4$‘ACT score‘
> x3 <- factor(da4$‘Academic year‘)
> test_loc <- sample(1:nrow(da4),100, replace = F)
> test <- data.frame(y=y[test_loc],
+                    x1=x1[test_loc],x2=x2[test_loc],x3=x3[test_loc])
> train <- data.frame(y=y[-test_loc],
+                     x1=x1[-test_loc],x2=x2[-test_loc],x3=x3[-test_loc])
```

First use the first order model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\{1997\} + \beta_4 X_3\{1998\} + \beta_5 X_3\{1999\} + \beta_6 X_3\{2000\}$$

```
> model43 <- lm (y~ x1 + x2 + x3, data = train)
> summary(model43)
Call:
lm(formula = y ~ x1 + x2 + x3, data = train)
Residuals:
     Min       1Q   Median       3Q      Max
-1.86370 -0.28925  0.08984  0.39230  1.35648
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.261499   0.155838   8.095 3.22e-15 ***
x1          0.011141   0.001391   8.011 5.96e-15 ***
x2          0.033791   0.006357   5.316 1.50e-07 ***
x31997      0.030130   0.074066   0.407    0.684
x31998      0.077581   0.071327   1.088    0.277
x31999      0.016451   0.072789   0.226    0.821
x32000      0.034879   0.073254   0.476    0.634
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.5659 on 598 degrees of freedom
Multiple R-squared:  0.2096,Adjusted R-squared:  0.2016
F-statistic: 26.43 on 6 and 598 DF,  p-value: < 2.2e-16
```

As we can see the result summary, the p-value for X_3 (academic year) is large and not significant. So we decide to remove X_3. Let's test whether X_3 can be removed from the model.

Hypothesis:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_1 : \text{not all } \beta_3, \beta_4, \beta_5, \beta_6 = 0$$

$$\text{if } F^* \leq F(0.95, 4, n-7), \text{ then conclude } H_0$$

$$\text{if } F^* > F(0.95, 4, n-7), \text{ then conclude } H_1$$

$$F^* = \frac{SSE(\text{reduced}) - SSE(\text{full})}{(n-3)-(n-7)} \Big/ \frac{SSE(\text{full})}{n-7}$$

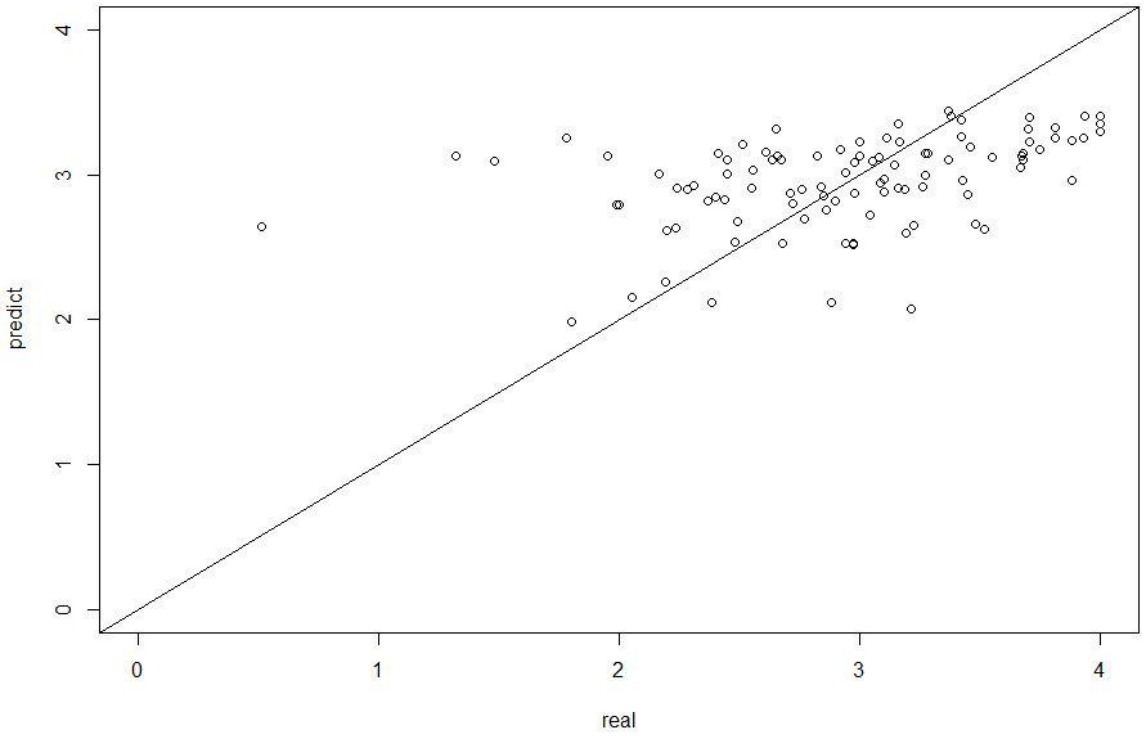
```
> model43_r <- lm (y~ x1 + x2, data = train)
> numerator <- (sum(model43_r$residuals^2) - sum(model43$residuals^2))/4
> denominator <- sum(model43$residuals^2) / model43$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 0.3327471
> qf(0.95, 4, model43$df.residual)
[1] 2.386834
> 1 - pf(Fstatistic,4,model43$df.residual)
[1] 0.8559695
```

Because $F^* = 0.3327471 \leq F(0.95, 4, n-7) = 2.386834$, with p-value 0.8559695, then we conclude H_0. $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$.

Therefore, we remove X_3 (academic year).

Now we make the prediction on test set:

```
> testpred <- predict(model43_r, test[,2:4], level=0.99)
> plot(test[,1],testpred,xlim=c(0,4),ylim=c(0,4),xlab="real",ylab="predict")
> abline(0,1)
```



This plot is predicting value against real value. We can see the model fit the data well to some degree. Now we try to find a more appropriate model. Let's introduce quadratic and interaction terms into the regression model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

Hypothesis:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_1 : \text{not all } \beta_3, \beta_4, \beta_5 = 0$$

```
> model43r2 <- lm (y~ x1 + I(x1*x2) +I(x1^2)+I(x2^2)+ x2, data = train)
> numerator <- (sum(model43_r$residuals^2) - sum(model43r2$residuals^2))/3
> denominator <- sum(model43r2$residuals^2) / model43r2$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 6.914549
> qf(0.95, 3, model43r2$df.residual)
[1] 2.619778
> 1 - pf(Fstatistic,3,model43r2$df.residual)
[1] 0.0001396
```

Because $F^* = 6.914549 \geq F(0.95, 3, n-6) = 2.619778$, with p-value 0.0001396, then we conclude H_1.

According to the summary of the model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.041e+00  6.740e-01   4.512 7.72e-06 ***
x1          -2.220e-02  8.557e-03  -2.594  0.00971 **
I(x1 * x2)   3.757e-04  3.724e-04   1.009  0.31351
I(x1^2)      1.832e-04  6.227e-05   2.941  0.00339 **
I(x2^2)      5.365e-04  1.239e-03   0.433  0.66531
x2          -2.249e-02  5.438e-02  -0.414  0.67931
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We notice that only X_1 and X_1^2 are significant, so we test the following reduced model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2$$

Hypothesis:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : \text{not both } \beta_4, \beta_5 = 0$$

```
> model43r3 <- lm (y~ x1 +x2+I(x1^2), data = train)
> numerator <- (sum(model43r3$residuals^2) - sum(model43r2$residuals^2))/3
> denominator <- sum(model43r2$residuals^2) / model43r2$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 0.6566811
> qf(0.95, 3, model43r2$df.residual)
[1] 2.619778
> 1 - pf(Fstatistic,3,model43r2$df.residual)
[1] 0.5789806
```

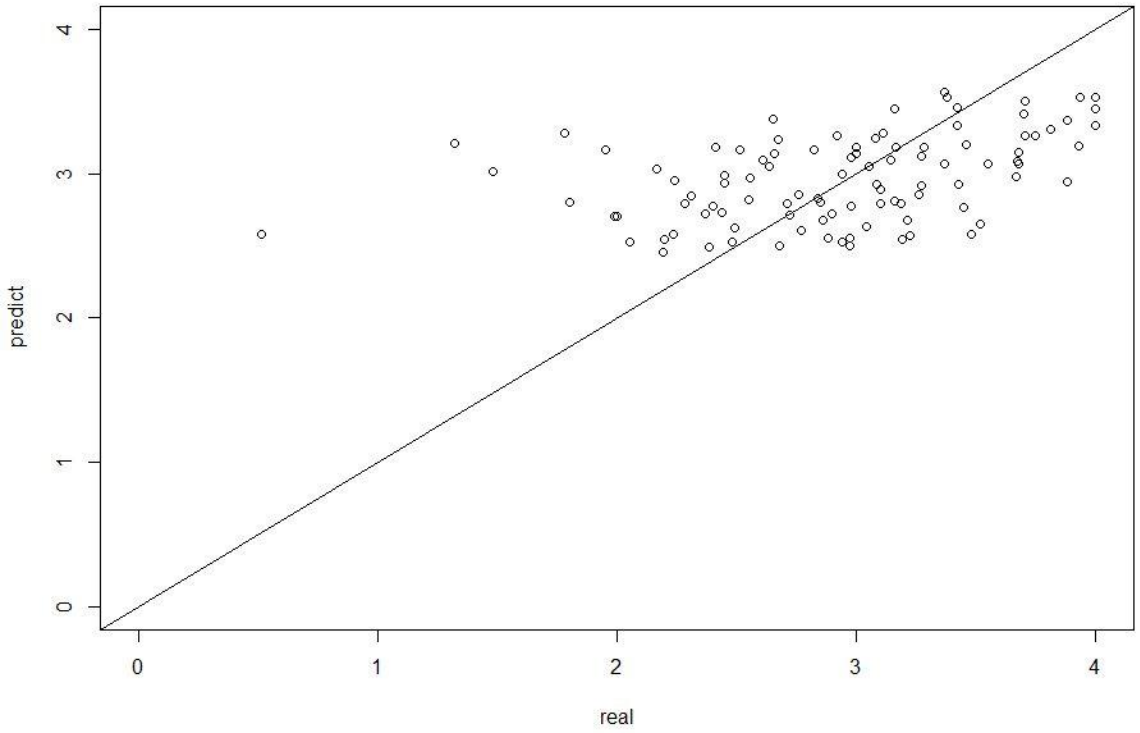
Because $F^* = 0.6566811 \geq F(0.95, 3, n-6) = 2.619778$, with p-value 0.5789806, then we conclude H_0.

Now it comes our final model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2$$

Let's make prediction on test set:

```
> testpred <- predict(model43r3, test[,2:4], level=0.95)
> plot(test[,1],testpred,xlim=c(0,4),ylim=c(0,4),xlab="real",ylab="predict")
> abline(0,1)
```



This is much better than the first order model fit. So we conclude that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2$$

is our final model and

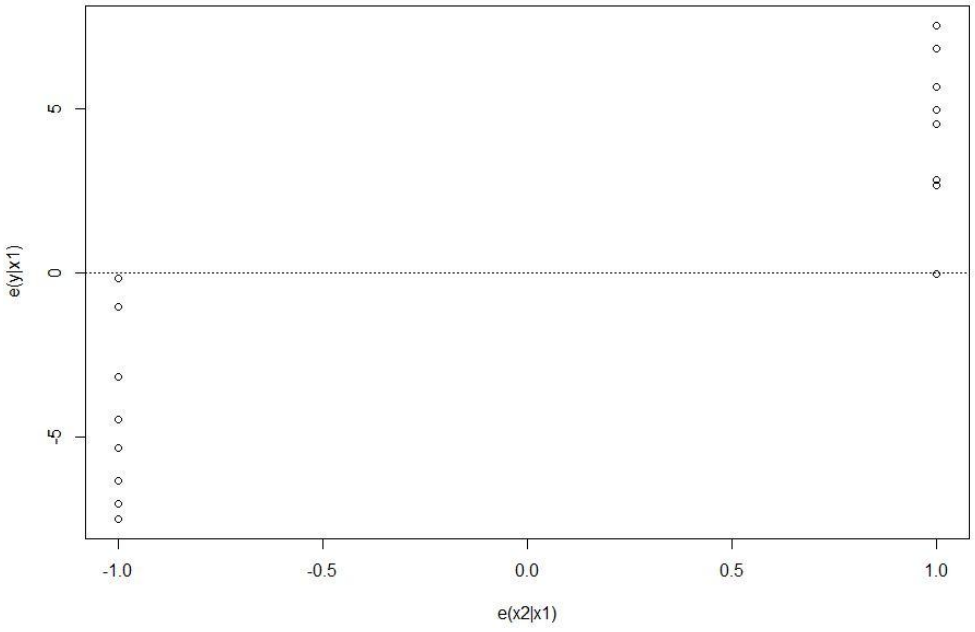
$$Y = 2.2621656555 - 0.0193636239X_1 + 0.0321995991X_2 + 0.0002266046X_2^2$$

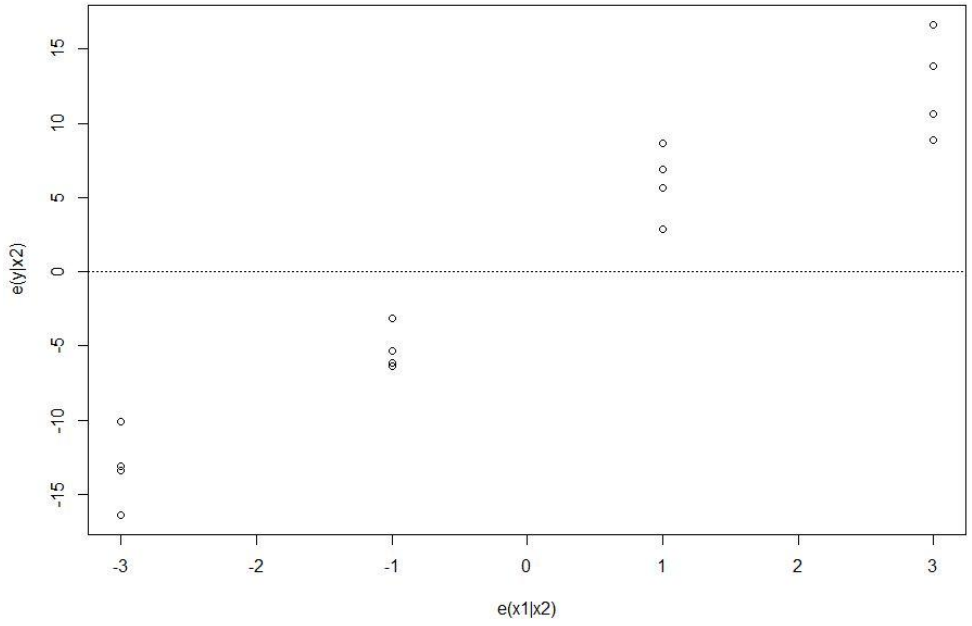
Although the model is much better, its use as a tool for admissions decisions is not much appropriate since admission should take more factors than these two into account.

Problem 4 (10.5)

(a)

```
> model5yx1 <- lm(y~x1)
> model5yx2 <- lm(y~x2)
> model5x1x2 <- lm(x1~x2)
> model5x2x1 <- lm(x2~x1)
>
> plot(model5x1x2$residuals, model5yx2$residuals,
+      xlab= "e(x1|x2)",ylab="e(y|x2)")
> abline(0,0)
> plot(model5x2x1$residuals, model5yx1$residuals,
+      xlab= "e(x2|x1)",ylab="e(y|x1)")
> abline(0,0)
```





(b)

According to the plots above, we can conclude the original regression function is appropriate for the predictor variables. Because the residuals point lay uniformly in the plot and each predictor variables does not show any significant influence.

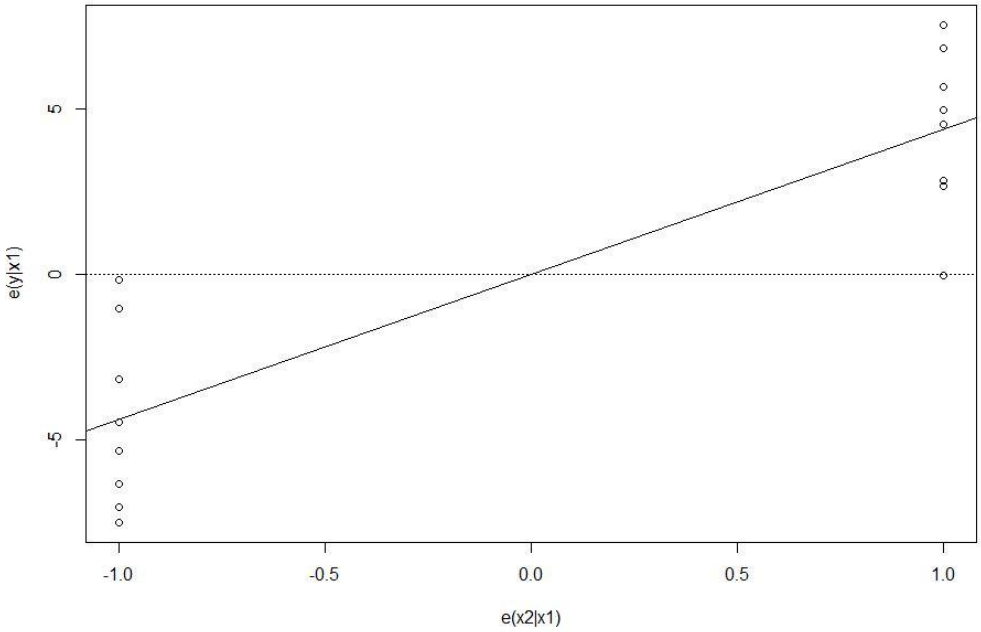
(c)

```
> model5yx1 <- lm(y~x1)
> model5x2x1 <- lm(x2~x1)
> model5yx1
Call:
lm(formula = y ~ x1)
Coefficients:
(Intercept)           x1
    1.93313      0.01357
> model5x2x1
Call:
lm(formula = x2 ~ x1)
Coefficients:
(Intercept)           x1
   17.20859      0.09531
```

Obtain:

$$\hat{Y}(X_1) = 1.93313 + 0.01357X_1$$
$$\hat{X_2}(X_1) = 17.20859 + 0.09531X_1$$

```
> yy <- model5yx1$residuals
> xx <- model5x2x1$residuals
> plot(model5x2x1$residuals, model5yx1$residuals,
+      xlab= "e(x2|x1)",ylab="e(y|x1)")
> abline(0,0, lty=3)
> modelyyxx <- lm(yy~xx)
> abline(modelyyxx$coefficients)
```



The plot shows the line through the origin with slope equal to the regression coefficient for the predictor variable if it were added to the fitted model. This plot provides some useful additional information. The plot follows the prototype in Figure 10.1 a, suggesting that X_1 is of little additional help in the model when X_2 is already present.

Problem 5 (10.9)

(a)

```
> X = cbind(rep(1,nrow(d5)),x1,x2)
> H = X%*%solve(t(X)%*%X)%*%t(X)
> Yhat = H%*%y
> e = (diag(rep(1,length(y)))-H)%*%y
>
> SSE = anova(lm(y~x1+x2))$`Sum Sq`[3]
> hii = diag(H)
>
> ti = e*sqrt((16-3-1)/(SSE*(1-hii)-e^2))
> data.frame(e,hii,ti)
       e    hii          ti
1  -0.10 0.2375 -0.04085498
2   0.15 0.2375  0.06128781
3  -3.10 0.2375 -1.36059879
4   3.15 0.2375  1.38602483
5  -0.95 0.1375 -0.36694571
6  -1.70 0.1375 -0.66490618
7  -1.95 0.1375 -0.76716157
8   1.30 0.1375  0.50461264
9   1.20 0.1375  0.46506694
10 -1.55 0.1375 -0.60436295
11  4.20 0.1375  1.82302030
12  2.45 0.1375  0.97784298
13 -2.65 0.2375 -1.13966417
14 -4.40 0.2375 -2.10272640
15  3.35 0.2375  1.48973208
16  0.60 0.2375  0.24572878
```

We shall use the Bonferroni simultaneous test procedure with a family significance level of $\alpha = 0.10$. We therefore require:

$$t(1 - \alpha/2n; n - p - 1) = t(0.996875; 12) = 3.307783$$

if $t_i \leq t(1 - \alpha/2n; n - p - 1)$, then conclude case i is not outlier;

if $t_i \geq t(1 - \alpha/2n; n - p - 1)$, then conclude case i not outlier;

Conclusion is that there is no outlier among the 16 cases.

(b)

The diagonal elements of the hat matrix are provided in part (a) as hii.

There are only two levels of the diagonal elements of the hat matrix which are 0.1375 and 0.2375.

(c)

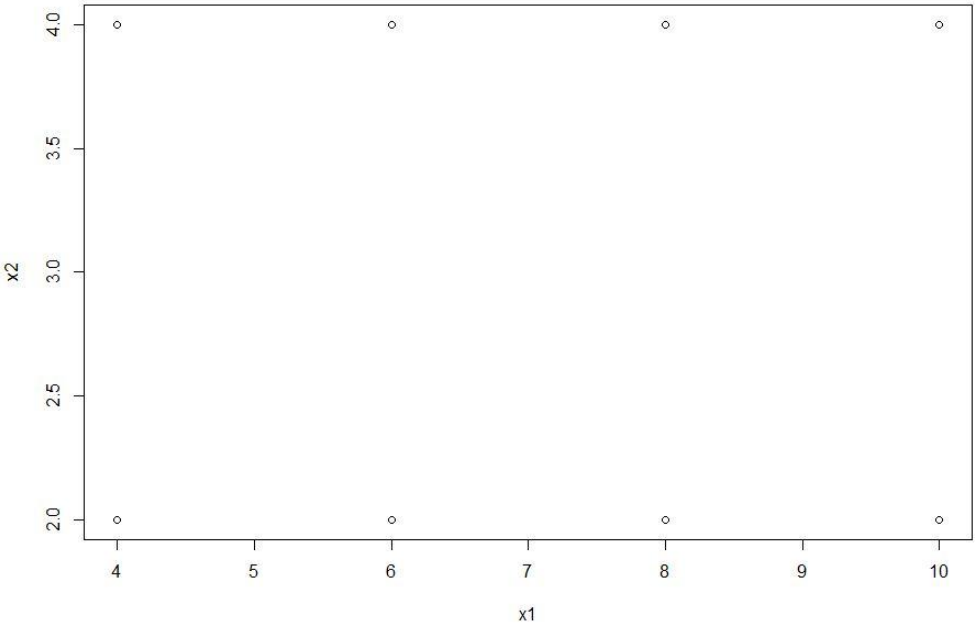
Leverage values greater than 2p/n are considered by this rule to indicate outlying cases with regard to their X values. Therefore,

$$\frac{2p}{n} = \frac{2 * 3}{16} = 0.375$$

```
> mean(hii)
[1] 0.1875
```

So there is not any of the observations outlying with regard to their X values according to the rule of thumb.

(d)



Visually this prediction does not involve an extrapolation beyond the range of the data.

```
> Xnew = c(1,10,3)
> t(Xnew)%*%solve(t(X)%*%X)%*%Xnew
      [,1]
[1,] 0.175
```

If $h_{\text{new.new}}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

Therefore, the small $h_{\text{new.new}}$ leads to the conclusion that no extrapolation is needed.

(g)

```
> MSE = anova(lm(y~x1+x2))$`Mean Sq`[3]
> e^2/(3*MSE)*(hii/(1-hii)^2)
               [,1]
 [1,] 0.0001877130
 [2,] 0.0004223542
 [3,] 0.1803921815
 [4,] 0.1862582123
 [5,] 0.0076655286
 [6,] 0.0245466787
 [7,] 0.0322971439
 [8,] 0.0143542862
 [9,] 0.0122308711
[10,] 0.0204060192
[11,] 0.1498281704
[12,] 0.0509831969
[13,] 0.1318214458
[14,] 0.3634123447
[15,] 0.2106609008
[16,] 0.0067576676
```

The larger either e_i or h_{ii} is, the larger D_i is.

The i^{th} case can be influential:

(1) by having a large residual e_i and only a moderate leverage value h_{ii}. or

(2) by having a large leverage value h_{ii} with only a moderately sized residual e_i, or

(3) by having both a large residual e_i and a large leverage value h_{ii} .

Therefore, none of the case is influential.