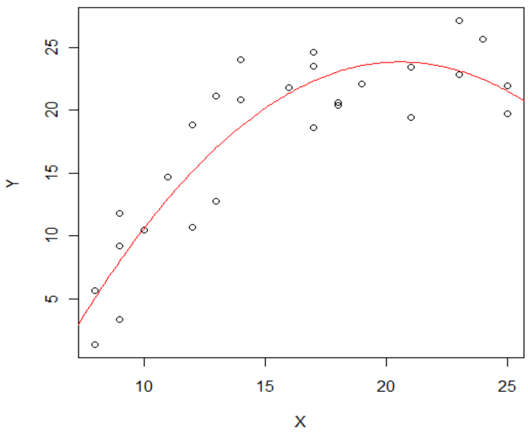
UNI: jh3561
Jiahong Hu
HW8
STAT 4315

8.6
Part a.
Fit regression model using quadratic function:

```
> summary(fit)

Call:
lm(formula = var2 ~ X1 + X11)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5463 -2.5369  0.3868  2.1973  5.3020 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.09416    0.91415  23.075  < 2e-16 ***
X1           1.13736    0.11546   9.851 6.59e-10 ***
X11         -0.11840    0.02347  -5.045 3.71e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.153 on 24 degrees of freedom
Multiple R-squared:  0.8143,	Adjusted R-squared:  0.7989 
F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

fitted regression function is shown below:
Y=21.09416+1.13736X-0.11840X^2.
Plot the fitted regression function and the data. The quadratic regression function appears to be a good fit.

$R^2 = 0.8143$



Part b.

```
> anova(fit)
Analysis of Variance Table

Response: Y
                      Df  Sum Sq Mean Sq F value    Pr(>F)    
poly(X, 2, raw = TRUE)  2 1046.27  523.13  52.633 1.678e-09 ***
Residuals             24  238.54    9.94                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ho: Both β_1 and β_{11} equal to 0

Ha: not both β_1 and β_{11} equal to 0.

Since $F^* = 52.633 > F^{-1}_{2,24}(0.99) = 5.613591$, we should conclude the H_α.

P-value is 1.678×10^{-9}.

Part c.

The 99% joint interval estimates for the mean steroid level of females aged 10, 15, 20 are [7.560736, 13.57968], [17.2297, 23.04615], [20.99211, 26.57905] respectively. With 99% confidence level, the average steroid level of female's aged 10 is between 7.560736 and 13.57968. Same interpretations applied for other two cases.

```
> steroid=read.table("CH08PR06.txt")
> attach(steroid)
> X=steroid[,2]
> sorted=steroid[order(X),]
> Y=sorted[,1]
> X=sorted[,2]
> fit=lm(Y~poly(X,2,raw=TRUE),data=sorted)
> B=qt(0.99833,24)
> ci=predict(fit,data.frame(X=10),se.fit=TRUE)
> ci$fit-B*ci$se.fit
       1 
7.560736 
> ci$fit+B*ci$se.fit
       1 
13.57968 
> ci1=predict(fit,data.frame(X=15),se.fit=TRUE)
> ci1$fit-B*ci1$se.fit
      1 
17.2297 
> ci1$fit+B*ci1$se.fit
       1 
23.04615 
> ci2=predict(fit,data.frame(X=20),se.fit=TRUE)
> ci2$fit-B*ci2$se.fit
       1 
20.99211 
> ci2$fit+B*ci2$se.fit
       1 
26.57905 
```

Part d.

```
> predict(fit,data.frame(X=15),interval="prediction",level=0.99)
       fit      lwr      upr
1 20.13792 10.97342 29.30242
```

So the prediction interval is [10.97342, 29.30242]. With 99 percent prediction level, the steroid levels of females aged 16 is between 10.97342 and 29.30242.

Part e.

```
> fit2=lm(Y~poly(X,1,raw=TRUE),data=sorted)
> anova(fit2,fit)
Analysis of Variance Table

Model 1: Y ~ poly(X, 1, raw = TRUE)
Model 2: Y ~ poly(X, 2, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     25 491.53
2     24 238.54  1    252.99 25.453 3.708e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(0.99,1,24)
[1] 7.822871
```

Ho: β_{11} =0

Ha: β_{11} is not equal to 0

$F^* = 25.453 > F_{1,24}^{-1}(0.99) = 7.822871$.

Therefore we should conclude H_α that the coefficient of quadratic term is not 0, and P-value is 3.71e-05.

Part f.

```
> lm(var2~age+age2)

Call:
lm(formula = var2 ~ age + age2)

Coefficients:
(Intercept)          age         age2
   -26.3254       4.8736      -0.1184
```

So the regression line is Y=-26.3254+4.8736X-0.1184X^2.

8.31

part a. (8.12)

$\hat{Y} = b_0 + b_1 x + b_{11} x^2,$

$x = X - \bar{X}$

$\Rightarrow \hat{Y} = \left(b_0 - b_1 \bar{X} + b_{11} \bar{X}^2\right) + \left(b_1 - 2 b_{11} \bar{X}\right) X + b_{11} X^2$

$\Rightarrow b_0^{'} = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2$

$b_1^{'} = b_1 - 2 b_{11} \bar{X},$

$b_{11}^{'} = b_{11}$

Part b.

$$\begin{bmatrix} b_0^{'} \\ b_1^{'} \\ b_{11}^{'} \end{bmatrix} = \begin{bmatrix} 1 & -\bar{X} & \bar{X}^2 \\ 0 & 1 & -2\bar{X} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_{11} \end{bmatrix}$$

$$\Rightarrow \sigma^2 \left\{ \begin{bmatrix} b_0^{'} \\ b_1^{'} \\ b_{11}^{'} \end{bmatrix} \right\} = \begin{bmatrix} 1 & -\bar{X} & \bar{X}^2 \\ 0 & 1 & -2\bar{X} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -\bar{X} & 1 & 0 \\ \bar{X}^2 & -2\bar{X} & 1 \end{bmatrix}$$

Where $\sigma_2^2 = \sigma^2 \left\{ b_{11} \right\}, \sigma_{02} = \sigma \left\{ b_0, b_{11} \right\}$ etc.

$$\Rightarrow \sigma^2 \left\{ b_0^{'} \right\} = \sigma_0^2 - 2\bar{X}\sigma_{01} + 2\bar{X}^2 \sigma_{02} + \bar{X}^2 \sigma_1^2 - 2\bar{X}^3 \sigma_{12} + \bar{X}^4 \sigma_2^2$$

$$\sigma^2 \left\{ b_1^{'} \right\} = \sigma_1^2 - 4\bar{X}\sigma_{12} + 4\bar{X}^2 \sigma_2^2$$

$$\sigma^2 \left\{ b_2^{'} \right\} = \sigma_2^2$$

$$\sigma \left\{ b_0^{'}, b_1^{'} \right\} = \sigma_{01} - 2\bar{X}\sigma_{02} + 3\bar{X}^2 \sigma_{12} - \bar{X}\sigma_1^2 - 2\bar{X}^3 \sigma_2^2$$

$$\sigma \left\{ b_0^{'}, b_2^{'} \right\} = \sigma_{02} - \bar{X}\sigma_{12} + \bar{X}^2 \sigma_2^2$$

$$\sigma \left\{ b_1^{'}, b_2^{'} \right\} = \sigma_{12} - 2\bar{X}\sigma_2^2$$

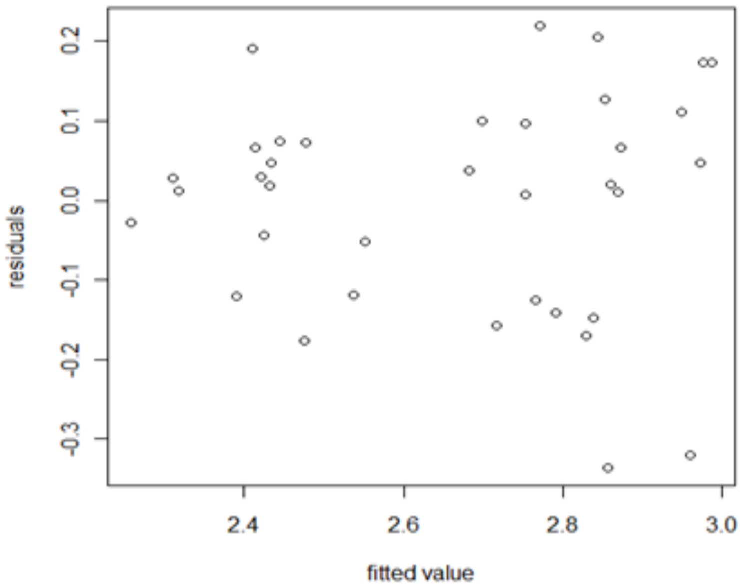
8.42

part a.

the fitted regression function is

$$\hat{Y} = 3.021 - 0.247X_1 - 0.000097X_2 + 0.4093X_3 + 0.124X_4 + 0.01324X_{51} - 0.1088X_{52} - 0.08306X_{53}$$

```
> data=read.csv("APPENC03.csv",header=FALSE)
> attach(data)
> Y=data[,2]
> X1=data[,3]
> X2=data[,4]
> X3=data[,5]
> X4=data[,6]
> X51=data[,8]
> X52=data[,9]
> X53=data[,10]
> fit1=lm(Y~X1+X2+X3+X4+X51+X52+X53)
> plot(fit1$fitted,fit1$resid,ylab="Residuals",xlab="Fitted Values")
> abline(0,0)
> summary(fit1)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X51 + X52 + X53)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.33558 -0.11872  0.02459  0.08020  0.21952 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.021e+00  4.705e-01   6.421 5.94e-07 ***
X1          -2.470e-01  1.982e-01  -1.246   0.2229    
X2          -9.653e-05  1.914e-04  -0.504   0.6181    
X3           4.093e-01  5.385e-02   7.601 2.80e-08 ***
X4           1.240e-01  5.484e-02   2.261   0.0317 *  
X51          1.324e-02  9.304e-02   0.142   0.8879    
X52         -1.088e-01  7.133e-02  -1.525   0.1385    
X53         -8.306e-02  8.657e-02  -0.959   0.3456    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1529 on 28 degrees of freedom
Multiple R-squared: 0.7326,     Adjusted R-squared: 0.6657 
F-statistic: 10.96 on 7 and 28 DF,  p-value: 1.382e-06
```



the residual plot doesn't show any apparent pattern. Thus the fitted function appears to fit the data well.

Part b.

According to the following R output, the fitted regression function involving quadratic interaction terms of quantitative variables is

$$\hat{Y} = 2.378 - 0.4527x_1 - 0.0001439x_2 + 0.0001629x_1x_2 + 0.9221x_1^2 + 0.0000005518x_2^2$$

$$+0.3941X_3 + 0.1149X_4 + 0.01236X_{51} - 0.1006X_{52} - 0.05807X_{53}.$$

```
> x1=X1-mean(X1)
> x2=X2-mean(X2)
> x1sq=x1^2
> x2sq=x2^2
> x12=x1*x2
> fit2=lm(Y~x1+x2+x12+x1sq+x2sq+X3+X4+X51+X52+X53)
> summary(fit2)

Call:
lm(formula = Y ~ x1 + x2 + x12 + x1sq + x2sq + X3 + X4 + X51 +
    X52 + X53)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33455 -0.08692  0.01892  0.07039  0.23931

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.378e+00  6.786e-02  35.039  < 2e-16 ***
x1          -4.527e-01  2.881e-01  -1.572   0.1286
x2          -1.439e-04  2.142e-04  -0.672   0.5080
x12          1.629e-04  1.393e-03   0.117   0.9078
x1sq         9.221e-01  1.069e+00   0.863   0.3965
x2sq         5.518e-07  7.375e-07   0.748   0.4613
X3           3.941e-01  6.098e-02   6.463 9.09e-07 ***
X4           1.149e-01  5.772e-02   1.991   0.0575 .
X51          1.236e-02  1.006e-01   0.123   0.9031
X52         -1.006e-01  7.476e-02  -1.345   0.1906
X53         -5.807e-02  9.541e-02  -0.609   0.5483
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1583 on 25 degrees of freedom
Multiple R-squared: 0.744,      Adjusted R-squared: 0.6417
F-statistic: 7.267 on 10 and 25 DF,  p-value: 2.837e-05
```

```
> fit3=lm(Y~x1+x2+X3+X4+X51+X52+X53)
> anova(fit3,fit2)
Analysis of Variance Table

Model 1: Y ~ x1 + x2 + X3 + X4 + X51 + X52 + X53
Model 2: Y ~ x1 + x2 + x12 + x1sq + x2sq + X3 + X4 + X51 + X52 + X53
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     28 0.65424
2     25 0.62614  3  0.028101 0.374 0.7725
> qf(0.95,3,25)
[1] 2.991241
```

Ho: all the coefficient of quadratic and interaction term equal to 0

H_α: not all the coefficients of quadratic and interaction terms equal to 0.

We now easily find $F^* = 0.374 < F_{3,25}^{-1}(0.95) = 2.991241.$

Hence we can conclude H_0.

Therefore we can drop them from the full model.

Part c.

```
> fit4=lm(Y~X1+X3+X4)
> anova(fit4,fit1)
Analysis of Variance Table

Model 1: Y ~ X1 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4 + X51 + X52 + X53
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     32 0.71795
2     28 0.65424  4  0.063715 0.6817 0.6105
```

Ho: both the coefficient of advertising index and year equal to 0

H_α: Not both the coefficients of advertising index and year equal to 0.

$F^* = 0.6817 < F_{4,28}^{-1}(0.95) = 2.714076$

Hence we conclude H_0.

8.43

1)fit the model interpolating quadratic and interaction terms:

```
> data=read.csv("APPENC04.csv",header=FALSE)
> attach(data)
> Y=data[,2]
> X1=data[,3]
> X2=data[,4]
> X3=data[,6]
> X4=data[,7]
> X5=data[,8]
> X6=data[,9]
> x1=X1-mean(X1)
> x2=X2-mean(X2)
> x1sq=x1^2
> x2sq=x2^2
> x12=x1*x2
> fit=lm(Y~x1+x2+x12+x1sq+x2sq+X3+X4+X5+X6)
> summary(fit)

Call:
lm(formula = Y ~ x1 + x2 + x12 + x1sq + x2sq + X3 + X4 + X5 +
    X6)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09218 -0.31352  0.07167  0.37913  1.32764

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.897e+00  5.190e-02  55.820  < 2e-16 ***
x1           1.434e-02  1.625e-03   8.825  < 2e-16 ***
x2           3.528e-02  5.947e-03   5.932 4.71e-09 ***
x12          5.652e-04  3.561e-04   1.587   0.1129
x1sq         1.476e-04  5.625e-05   2.623   0.0089 **
x2sq        -7.625e-05  1.148e-03  -0.066   0.9471
X3          -4.295e-02  6.723e-02  -0.639   0.5231
X4           1.698e-02  6.849e-02   0.248   0.8042
X5           5.689e-02  6.595e-02   0.863   0.3886
X6           2.238e-02  6.763e-02   0.331   0.7409
---
```

2) test whether the academic year variable can be dropped from the model:

```
> fit1=lm(Y~x1+x2+x12+x1sq+x2sq)
> anova(fit1,fit)
Analysis of Variance Table

Model 1: Y ~ x1 + x2 + x12 + x1sq + x2sq
Model 2: Y ~ x1 + x2 + x12 + x1sq + x2sq + X3 + X4 + X5 + X6
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    699 218.35
2    695 217.58  4   0.77103 0.6157 0.6514
>
```

The large P-value indicates that it is ok to drop the academic year variable.

3) test whether we can drop all the quadratic and interaction terms:

```
> fit2=lm(Y~x1+x2)
> anova(fit2,fit1)
Analysis of Variance Table

Model 1: Y ~ x1 + x2
Model 2: Y ~ x1 + x2 + x12 + x1sq + x2sq
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    702 225.81
2    699 218.35  3    7.4624 7.963 3.166e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

The very small P-value shows the quadratic and interaction terms may have some contributions to the response variable and we can't just drop them.

4) test if we can drop part of those quadratic and interaction terms:

```
> fit3=lm(Y~x1+x2+x1sq)
> anova(fit3,fit1)
Analysis of Variance Table

Model 1: Y ~ x1 + x2 + x1sq
Model 2: Y ~ x1 + x2 + x12 + x1sq + x2sq
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    701 219.30
2    699 218.35  2   0.94876 1.5186 0.2197
```

The P-value indicates that we can drop the interaction term X_1X_2 and the quadratic term X_2^2.

5) the final model is $Y = 2.905 + 0.01458x_1 + 0.0353x_2 + 0.0002075x_1^2$, where $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$. i.e. which R square is much larger than the previous ones.

$$Y = 2.1454 - 0.01736X_1 + 0.0353X_2 + 0.0002075X_1^2$$

The model may have relative good fit, but we still cannot guarantee its prediction accuracy owing to the value of R square. Therefore, admissions should find out some other information about these students and give a much fair judgment of them.