Exam 3 - Question 2

Adam McQuistan

Sunday, May 01, 2016

Problem 2 - 8.37

For this problem Y is total serious crime divided by total population (X9 / X4), X_1 will be population density (X4 / X3) and, X_2 will be unemployment rate (X13).

The dataset

Variable	Description
X1	County
X2	State
X3	Land area
X4	Population
X5	Percent of population 18-34
X6	Percent of population older than 65
X7	Number of physicians
X8	Number of hospital beds
X9	Total serious crimes
X10	Percent highschool graduates
X11	Percent bachelors degrees
X12	Percent below poverty level
X13	Percent unemployment
X14	Per capita income
X15	Total personal income
X16	Geographic region

```
df <- read.csv(file="data/8.37.csv")
df$Y <- df$X9 / df$X4; df$X_1 <- df$X4 / df$X3; df$X_2 <- df$X13
summary(df[,c("Y","X_1", "X_2")])
```

```
##        Y                 X_1                X_2        
##  Min.   :0.004601   Min.   :   13.26   Min.   : 2.200  
##  1st Qu.:0.038102   1st Qu.:  192.34   1st Qu.: 5.100  
##  Median :0.052429   Median :  335.91   Median : 6.200  
##  Mean   :0.057286   Mean   :  888.44   Mean   : 6.597  
##  3rd Qu.:0.072597   3rd Qu.:  756.55   3rd Qu.: 7.500  
##  Max.   :0.295987   Max.   :32403.72   Max.   :21.300
```

Part A

Fit a second-order regression model. Plot the residuals against the fitted values. How well does the model appear to fit the data? What is R squared?

The Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$$

First the inputs need to be centered to aleviate effects of multi-collinearity___

```
X_1_bar <- mean(df$X_1); X_2_bar <- mean(df$X_2)
df$X_1_cent <- df$X_1 - X_1_bar; df$X_2_cent <- df$X_2 - X_2_bar
result1 <- lm(Y ~ X_1_cent + X_2_cent + I(X_1_cent^2) + I(X_2_cent^2) + I(X_1_cent * X_2_cent),
              data=df)
result1_smry <- summary(result1); print(result1_smry)
```

```
##
## Call:
## lm(formula = Y ~ X_1_cent + X_2_cent + I(X_1_cent^2) + I(X_2_cent^2) +
##     I(X_1_cent * X_2_cent), data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055642 -0.016851 -0.002889  0.014810  0.085485
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.629e-02  1.260e-03  44.662  < 2e-16 ***
## X_1_cent                4.585e-06  9.841e-07   4.659 4.23e-06 ***
## X_2_cent               -8.800e-05  6.276e-04  -0.140   0.8886
## I(X_1_cent^2)           2.698e-12  5.932e-11   0.045   0.9637
## I(X_2_cent^2)           1.629e-04  9.541e-05   1.708   0.0884 .
## I(X_1_cent * X_2_cent)  8.334e-07  4.091e-07   2.037   0.0423 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02383 on 434 degrees of freedom
## Multiple R-squared:  0.2485, Adjusted R-squared:  0.2398
## F-statistic:  28.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

```
result1_aov <- fullRegressionAnova(anova(result1))
```

```
##          VariationSource  DF           SS           MS     F_stats
## 1             Regression   5 0.0814617240 0.0162923448  28.6984676
## 2               X_1_cent   1 0.0756708467 0.0756708467 133.2918848
## 3               X_2_cent   1 0.0003138473 0.0003138473   0.5528324
## 4          I(X_1_cent^2)   1 0.0022033375 0.0022033375   3.8811117
## 5          I(X_2_cent^2)   1 0.0009184249 0.0009184249   1.6177775
## 6 I(X_1_cent * X_2_cent)   1 0.0023552675 0.0023552675   4.1487318
## 7              Residuals 434 0.2463851985 0.0005677078          NA
## 8                  Total 439 0.3278469225           NA          NA
```

Correlation Among the Predictor Terms

```
library(knitr)
Terms = c("X_1 and X_1 Sqr","x_1_cent and x_1_cent Sqr",
          "X_2 and X_2 Sqr", "x_2_cent and x_2_cent Sqr")
Correlations = c(cor(df$X_1, df$X_1^2), cor(df$X_1_cent, df$X_1_cent^2),
                 cor(df$X_2, df$X_2^2), cor(df$X_2_cent, df$X_2_cent^2))
Correlations <- round(Correlations, 3)
cor_df <- data.frame(Terms, Correlations); kable(cor_df)
```

Terms	Correlations
X_1 and X_1 Sqr	0.868
x_1_cent and x_1_cent Sqr	0.848
X_2 and X_2 Sqr	0.957
x_2_cent and x_2_cent Sqr	0.623

Note that the correlations between the centered data is less than the uncentered data.

Test of Fit

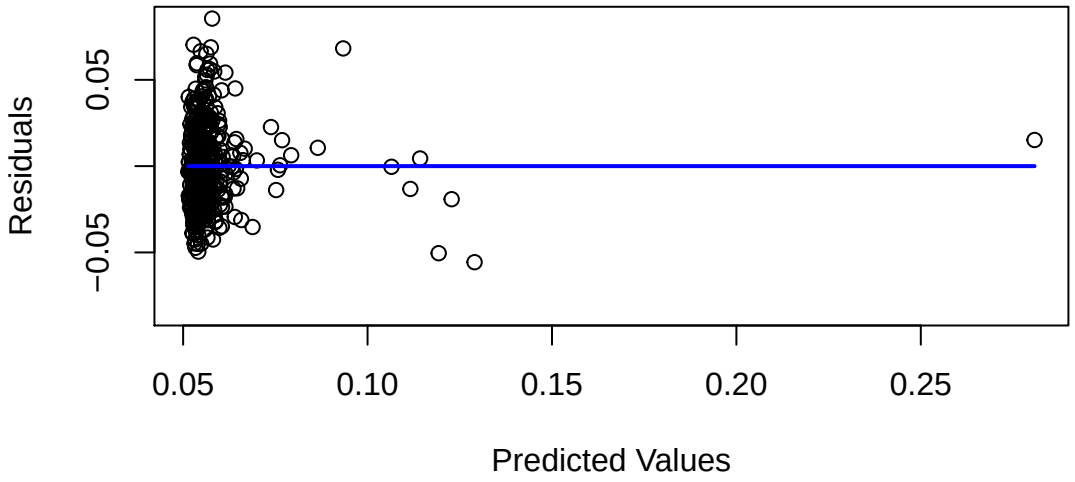
To determine if a lack of fit test can be used we'll search the data for replicates among the data.

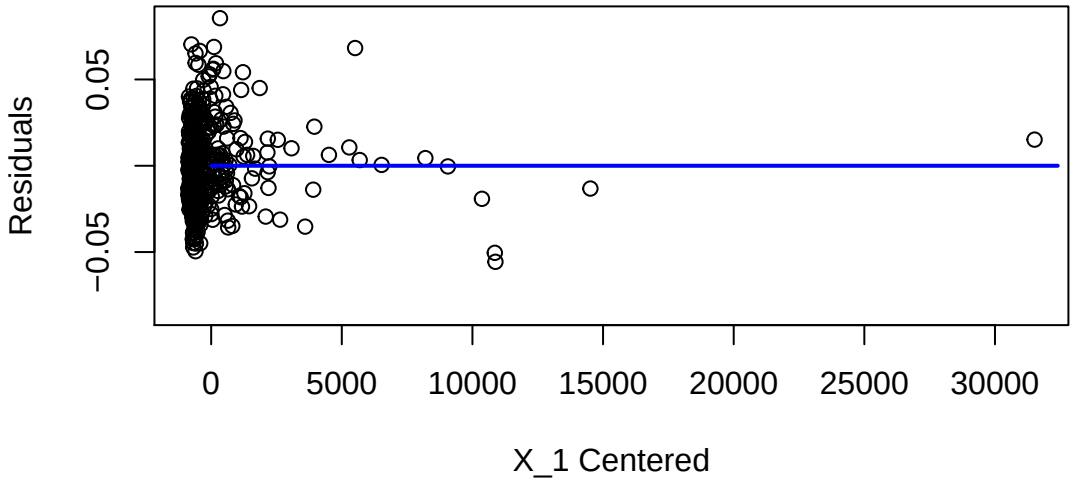
```
df$Terms <- paste(as.character(df$X_1_cent),as.character(df$X_2_cent),
                  as.character(df$X_1_cent^2),as.character(df$X_2_cent^2))
n <- dim(df)[1]; categories <- length(unique(df$Terms))
cat("Total Records ", n, "\nDistinct Categories ", categories, sep="")
```

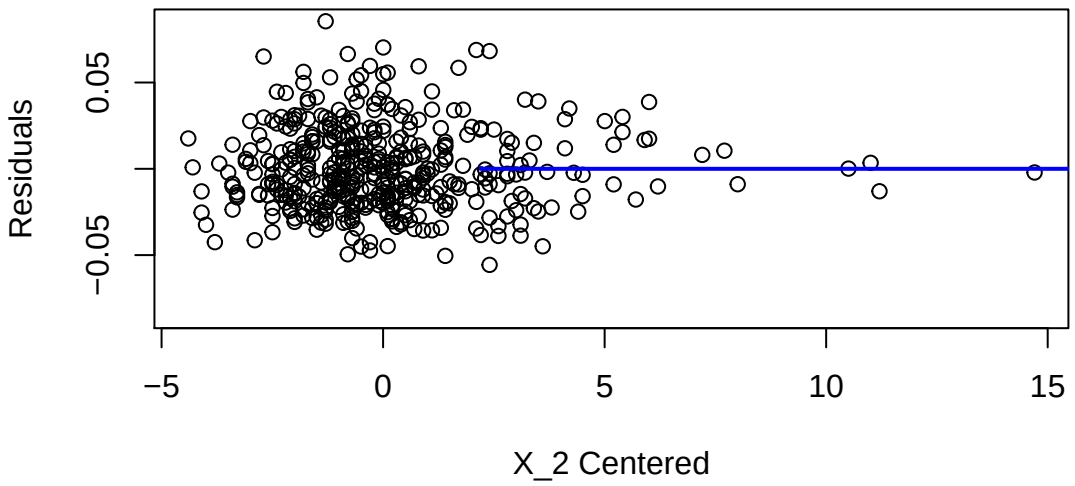
```
## Total Records 440
## Distinct Categories 440
```

Since there are not replicates a formal lack of fit test cannot be performed.

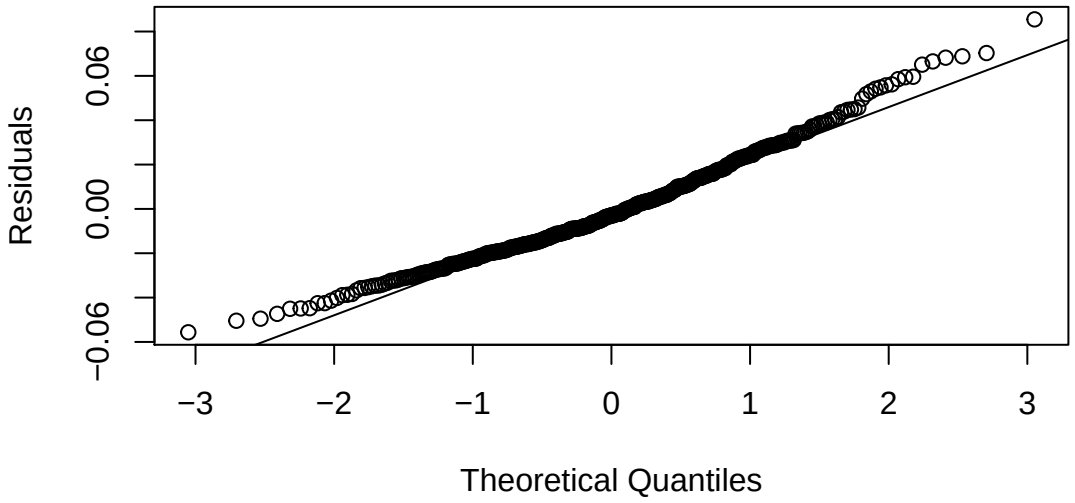
Residual Plots







Q-Q Normality Plot



Answer

The residual plots suggest that there is equal variance among the error terms and the model is appropriate for the data. The R squared value is 0.2485 and the adjusted R squared is 0.2398

Part B

Test whether or not all quadratics and interaction terms can be dropped from the regression model. Use $\alpha = 0.01$.

$H_o : \beta_{11} = \beta_{22} = \beta_{12} = 0$

H_a: Not all β_s in H_o equals 0

F stat = SSR($X_1^2, X_2^2, X_{12}|X_1, X_2$) / 3 / MSE

F crit = F(0.99, 3, 5)

If F Stat < F crit, conclude H_o

If F Stat >= F crit, conclude H_a

```
F_crit <- qf(0.99, 3, 5)
SSR <- sum(result1_aov$SS[4:6])
MSE <- result1_aov$MS[7]
F_stat <- SSR / 3 / MSE
msg = paste("F stat = ", F_stat, "\nF crit = ", F_crit)
result <- ifelse(F_stat < F_crit,
                 "\nConclude Ho, no curvature interaction effects are needed.",
                 "\nConclude Ha, curvature interaction effects are significant.")
cat(msg, result, sep="")
```

```
## F stat =  3.21587363957109
## F crit =  12.059953691652
## Conclude Ho, no curvature interaction effects are needed.
```

Answer

As the output suggests there H_o appears to be appropriate meaning that no curvature interactions are needed.

Part C

Fitting The Model with Population, Land Area, and Unemployment Rate

```
result2 <- lm(Y ~ X4 + I(X4^2) + X3 + X13, data = df)
summary(result2)
```

```
##
## Call:
## lm(formula = Y ~ X4 + I(X4^2) + X3 + X13, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05967 -0.01704 -0.00303  0.01410  0.19106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.250e-02  3.889e-03  10.927  < 2e-16 ***
```

```
## X4           3.206e-08  3.941e-09   8.133 4.41e-15 ***
## I(X4^2)     -3.356e-15  5.878e-16  -5.710 2.10e-08 ***
## X3          -5.576e-07  8.123e-07  -0.687    0.493
## X13          6.824e-04  5.302e-04   1.287    0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02539 on 435 degrees of freedom
## Multiple R-squared:  0.1444, Adjusted R-squared:  0.1365
## F-statistic: 18.35 on 4 and 435 DF,  p-value: 6.022e-14
```

The R squared value is 0.1444 and the adjusted R squared is 0.1365.

Since the above section proved that a quadratic effect is not required lets compare this to the first order model.

```
result3 <- lm(Y ~ X_1 + X_2, data=df)
summary(result3)
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053806 -0.016940 -0.003898  0.014680  0.084508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.959e-02  3.452e-03  14.368   <2e-16 ***
## X_1         5.973e-06  5.222e-07  11.439   <2e-16 ***
## X_2         3.618e-04  4.902e-04   0.738    0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02401 on 437 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2283
## F-statistic: 65.92 on 2 and 437 DF,  p-value: < 2.2e-16
```

Answer

For the model with population, land area, and unemployment rate the R squared value is 0.1444 and the adjusted R squared is 0.1365. Comparing that the first order model from part A-B we see that it actually explains less of the variation but both are pretty low R squared values.