Applied Regression - Exam 1

Adam McQuistan

Tuesday, March 08, 2016

1. Do problem 3.8 on page 148.

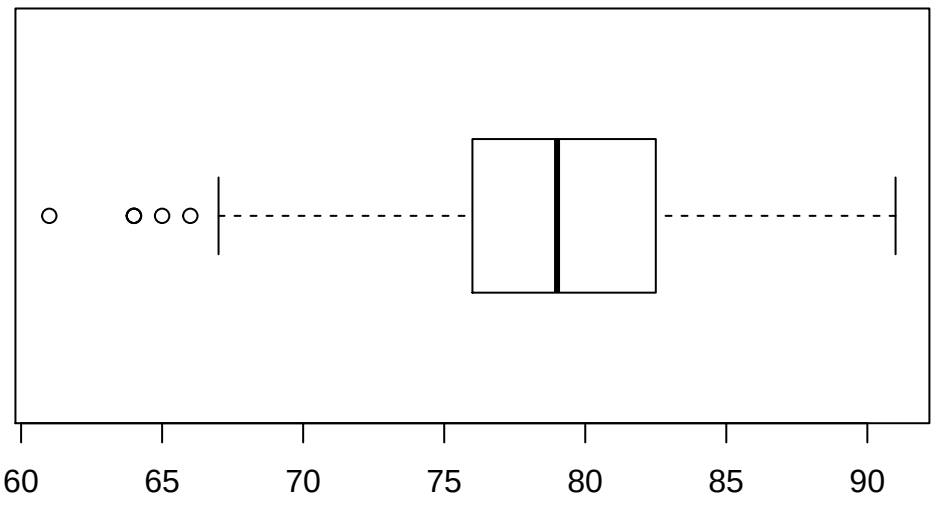
Problem 3.8: Refer to crime rate problem 1.28

A. Prepare a box-plot for the percentage of individuals in the county having at least a high-school diploma X_i. What information does your plot provide?

```
setwd("C:\\Users\\AdamMcQuistan\\Documents\\ISQA 8340\\Exam 1")
df <- read.csv("data/Ch. 1(PR28).csv")
names(df) = c("County","CrimeRate", "Percentage")
summary(df$Percentage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.00   76.00   79.00   78.60   82.25   91.00
```

```
boxplot(df$Percentage, main="",
        xlab="Percent of Individuals with atleast HS Diploma",
        horizontal=T)
```



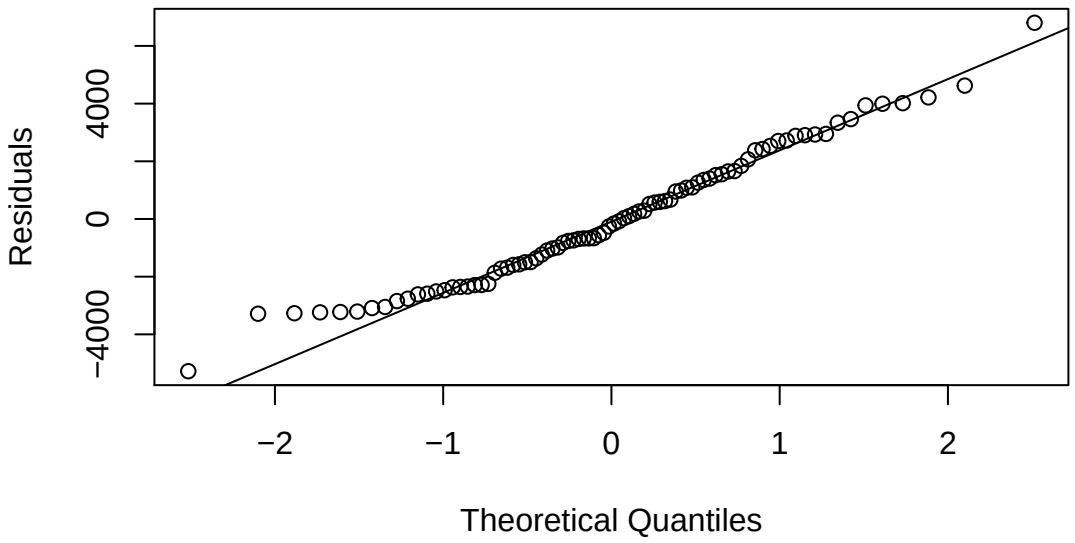
The boxplot shows that the variable for percentage of crimes with at least a highschool diploma is roughly normally distributed with slight left skew. If you use a determination for ourliers as being either 1st quartile -

3 x IQR or 3rd quartile + 3 x IQR then there are no outliers in the dataset. This roughly means that 50% of the data is within the box while

B. Obtain residuals e_i and prepare a normal probability plot of the residuals. Does the distribution of the residuals appear to be symmetrical?

```
result <- lm(CrimeRate ~ Percentage, data=df)

qqnorm(result$residuals, ylab="Residuals", main="")
qqline(result$residuals)
```

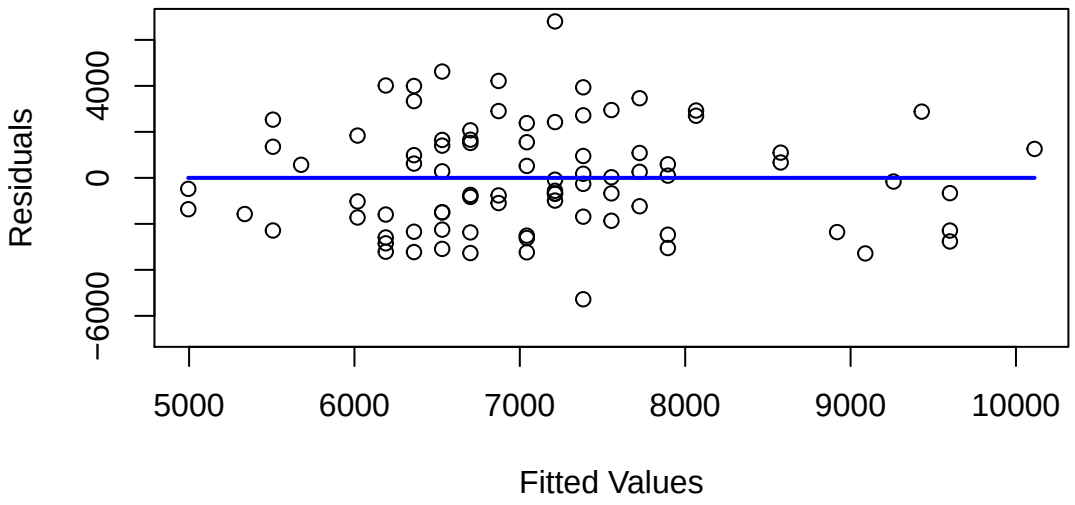


The Q-Q plot of the residuals shows the data has some lack of normality in the error terms below the first quartile.

C. Prepare a residual plot of e_i versus $\widehat{Y}$. What does the plot show?

```
with(result, {
     plot(x=fitted.values, y=residuals,
          ylim=c(-max(residuals), max(residuals)),
          xlab="Fitted Values", ylab="Residuals")

     points(c(min(fitted.values), max(fitted.values)),
            c(0,0), type="l", lwd="2", col="blue")
     })
```



The plot of e_i vs $\widehat{Y}$ shows no glaringly obvious pattern or outliers suggesting the model is appropriate.

D. Obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption using Table B.6 and $\alpha = 0.05$. What do you conclude?

```
df$Residuals <- result$residuals
df$Rank <- rank(df$Residuals)
df$Prob <- (df$Rank - 0.375) / (length(df$Rank) + 0.25)
df$Z <- qnorm(df$Prob, mean=0, sd=1)
result_smry <- summary(result)
df$EV <- result_smry$sigma * df$Z
cor_coef <-  with(df,cor(Residuals, EV))
cat("Correlation coefficient of ordered residuals vs expected values: ", cor_coef, "\n", sep="")
```

```
## Correlation coefficient of ordered residuals vs expected values: 0.9887589
```

If r > r critical (0.986) conclude the residuals are normally distributed.

Since the coefficient of correlation for ordered residuals versus expected values is 0.9888, which exceeds the extrapolated value from Table B.6 for n = 84 of 0.9855 then we have support for the distribution of error terms being normally distributed. The Q-Q normal probability plot also suggests this.

E. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups $X \leq 69$, $X > 69$, and use $\alpha = 0.05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c).

```
library(lawstat)
df$Group <- ifelse(df$Percentage > 69, "Gtr69", "Lte69")
with(df, levene.test(Residuals, as.factor(Group), location="median"))
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  Residuals
## Test Statistic = 0.126, p-value = 0.7235
```

```
t_crit <- qt(0.975, dim(df)[1])
```

If t* <= 1.9886097, conclude error variance is constant

If t* > 1.9886097, conclude error variance is not constant

Since the p-value for the Brown-Forsythe test is 0.7235 and T* < r t_crit‘ we can conclude that the error variance is constant.

F. Also perform the teset of normality on error after you save the residuals using Kolmogorov-Smirnov and Shapiro-Wilk test and state your hypothesis test.

```
ks.test(df$Residuals[df$Group == 'Gtr69'], df$Residuals[df$Group == 'Lte69'])
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  df$Residuals[df$Group == "Gtr69"] and df$Residuals[df$Group == "Lte69"]
## D = 0.3158, p-value = 0.3916
## alternative hypothesis: two-sided
```

```
shapiro.test(df$Residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Residuals
## W = 0.9776, p-value = 0.1515
```

Condition of Kolmogorov-Smirnov test: The null hypothesis is that the two groups are drawn from same distribution.

H_o: there is no difference in the samples if p-value > 0.05

H_o: there is a difference in the sample distributions if p-value <= 0.05.

Since the p-value of the Kolmogorov-Smirnov test is greater than 0.05 we can be confident that the two distributions are very similar.

Condition of Kolmogorov-Smirnov test: the null hypothesis is that the population is normally distributed.

H_o: If p-value > 0.05 the data is normally distributed

H_o: If p-value <= 0.05 the data is not normally distributed

Since the p-value is > o.o5 we can conclude that the data is normally distributed.

G. Conduct the Breusch-Pagen test with 0.05 level of significance.

```r
library(lmtest)
bptest(result)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  result
## BP = 0.0061, df = 1, p-value = 0.9378
```

H_0 : no heteroskedacity - the error variances are equal H_a : error variances are not equal

The pvalue of 0.9378 leads us to conclude the null hypthosis