STAT 530 HW2

Md Muhtasim Billah

2/7/2020

Question

Consider the data set website names Hw1_2020.xls or .csv. Answer the questions based on this data set:

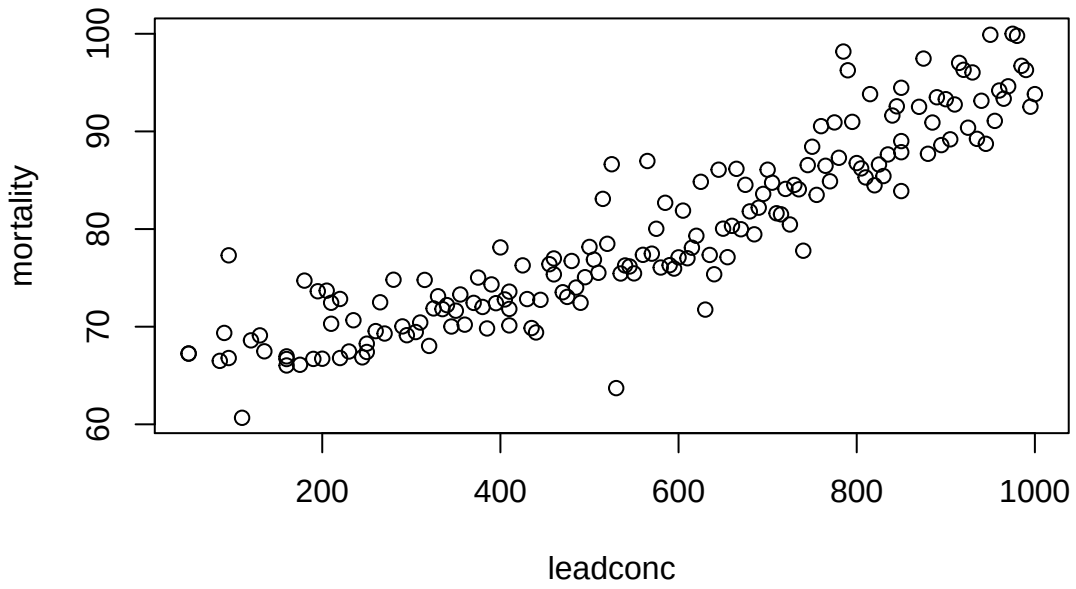
1. Plot the scatter plot and mention any possible violation of assumptions you see in this plot.
2. Plot the residual versus predicted plot and the normal probability plots. What do you see?
3. Do the correlation test for normality. For n>100, for alpha=.05, the critical value is: .987.
4. Do the Brown Forsyth Test or the Breusch Pagan test (whichever you think is appropriate). Discuss results.
5. Look at the Box-Cox transform to see what transformation if any is appropriate for the Y's.
6. What is the value of the Durbin Watson statistic?
7. Give an example from your experience where you have faced multiple regression.

Answer

1. Plot the scatter plot and mention any possible violation of assumptions you see in this plot.

Scatter plot of the explanatory (lead concentration) and response variable (mortality).

```
mydata=read.csv("/Users/muhtasim/Desktop/STAT530/HWs/HW2/HW1_2020.csv")
leadconc=mydata[,1]
mortality=mydata[,2]
plot(leadconc,mortality)
```



From a scatter plot, we can roughly estimate whether there are violations of the following two assumptions.

i) Non-linearity of the regression function.
ii) Non-constant variance.

But looking only at the scatter plot, it seems that there exist a positive linear pattern between the explanatory (x) and response (y) variables. So, the assumption of linearity holds to be true. Also, the variance seems to be constant. So, it can be said that these two assuptions were not violated.

2. Plot the residual versus predicted plot and the normal probability plots. What do you see?

Residuals vs predicted (fitted) plot (similar to the residuals against the explanatory variables plot).

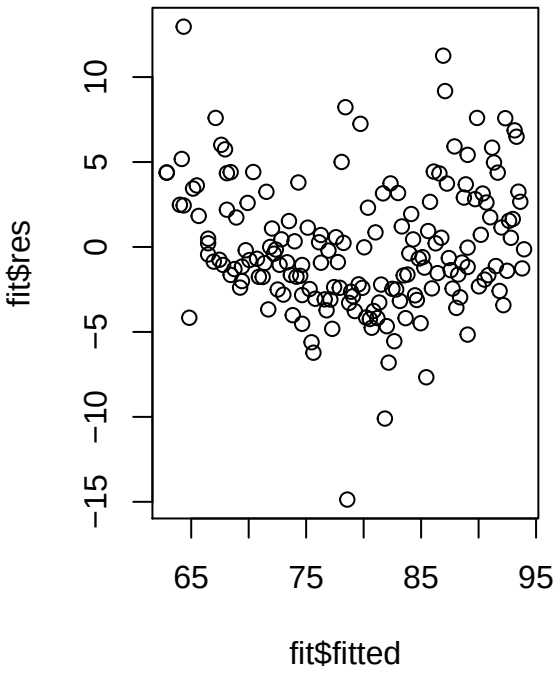
```
#linear regression
fit=lm(mortality~leadconc, data=mydata)
summary(fit)
```

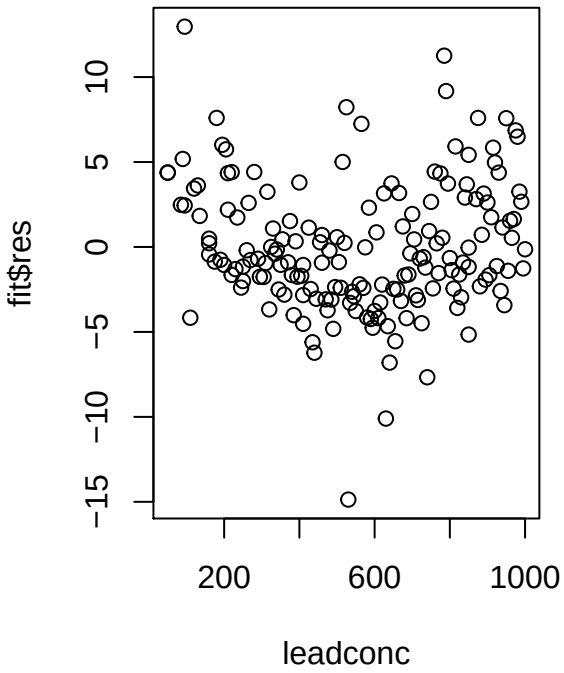
```
##
## Call:
## lm(formula = mortality ~ leadconc, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.867  -2.456  -0.684   2.538  12.966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.23565    0.69579   88.01   <2e-16 ***
## leadconc     0.03272    0.00112   29.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 169 degrees of freedom
## Multiple R-squared:  0.8346, Adjusted R-squared:  0.8336
## F-statistic: 852.8 on 1 and 169 DF,  p-value: < 2.2e-16
```

```
coefficients(fit)
```

```
## (Intercept)    leadconc
## 61.23564557  0.03271988
```

```
#plot of residual vs x and fitted values
par(mfrow=c(1,2))
plot(fit$fitted, fit$res)
plot(leadconc, fit$res)
```

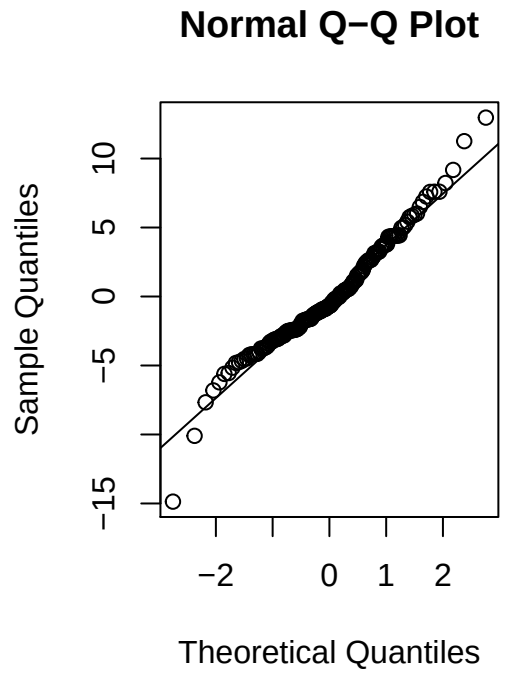


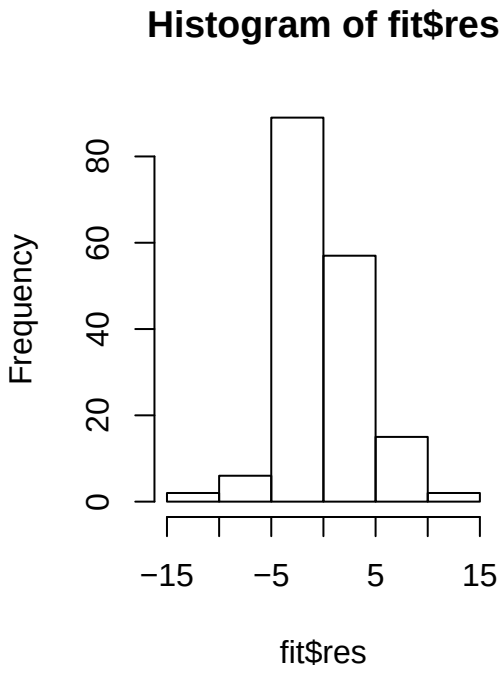


We don't see any pattern (like a funnel shape) in the above plots, rather we see a random scatter, which is desired. So, the plot indicates a possibility for constant variance.

Normal probability plots (QQ) of residuals.

```
# Normal probability plot of Residuals
par(mfrow=c(1,2))
qqnorm(fit$res)
qqline(fit$res)
hist(fit$res)
```





From the normal probability plot of the residuals (left), we see the tails of the plot at the lower and upper ends are deviated from the linear line which hints the non-normality of the residuals. The same is also seen from the histogram (right) of the errors. Though the higher frequencies are in the middle, the probability curve would be right skewed rather than being a perfect Bell curve.

3. Do the correlation test for normality. For n>100, for alpha=.05, the critical value is: .987.

Due to limited availabity of a correlation test for normality in R, the Shapiro-Wilk test for normality is performed.

```
#test for normality
shapiro.test(fit$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$res
## W = 0.97126, p-value = 0.001294
```

Here, the p-value is smaller than $\alpha = 0.05$ which indicates the rejection of null which means the errors are not normally distributed.

4. Do the Brown Forsyth Test or the Breusch Pagan test (whichever you think is appropriate). Discuss results.

We do the Breusch Pagan test because it has better power.

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

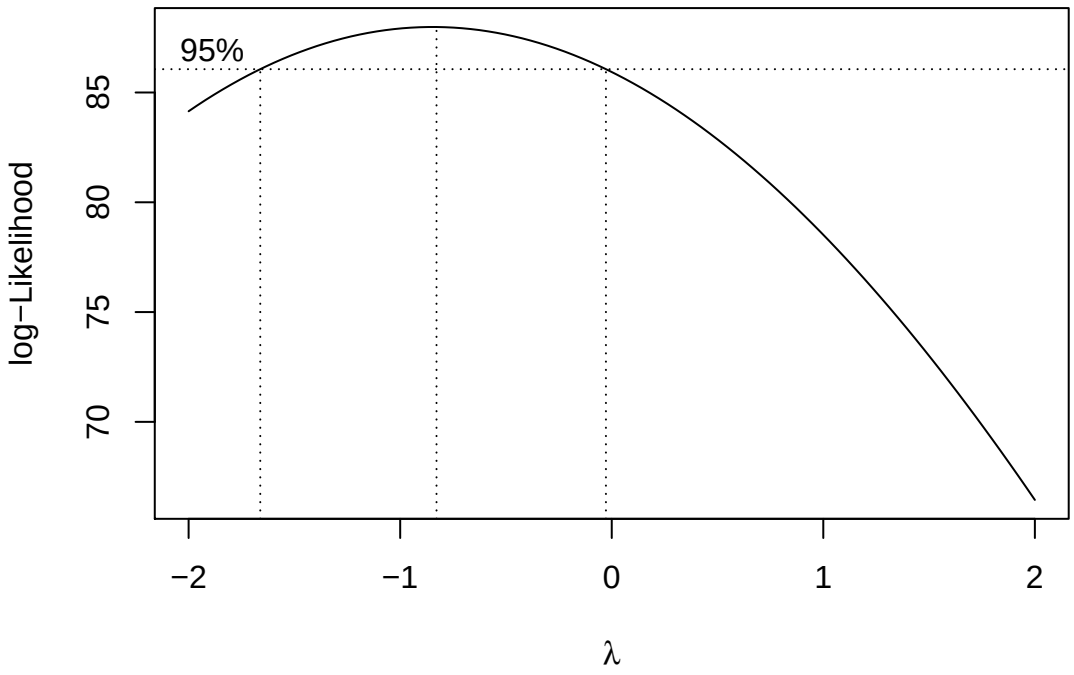
```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 0.036695, df = 1, p-value = 0.8481
```

Since, the $p - value > \alpha$, the null is retained. Which means that there is a constancy in variance.

5. Look at the Box-Cox transform to see what transformation, if any, is appropriate for the Y's.

```r
library(MASS)
boxcox(mortality ~ leadconc, data = mydata,
       lambda = seq(-2.0, 2.0, length = 10))
```



Since, the value of λ is around -1, so, an inverse transformation would be appropriate.

6. What is the value of the Durbin Watson statistic?

```r
#durbin-watson test
dwtest(fit)
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.7537, p-value = 0.04476
## alternative hypothesis: true autocorrelation is greater than 0
```

The value for Durbin Watson statistic is, $D = 1.7537$. This value indicates that there is a positive autocorrelation among the data.

7. Give an example from your experience where you have faced multiple regression.

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. A common example for such regression, that I have experienced, is the grading system for a class. The final grade (response variable) is directly related to several explanatory variables such as homeworks, project, mid term and final exam etc.

So, the explanatory variables are:

i) Homework, HW.
ii) Quiz, Q.
iii) Project, P.
iv) Midterm, Mid.
v) Final, Fin.

Where, the final grade, $Grade$ is the response variable.

So, the multiple linear regression model with the stochastic error becomes,

$$Grade = \beta_0 + \beta_1 HW + \beta_2 Q + \beta_3 P + \beta_4 Mid + \beta_5 Fin + \varepsilon_i$$