DATA 605 : Week 11 - Linear Regression Model

Ramnivas Singh

11/07/2021

Lets use the MPG dataset. I am going to use a dataset I worked with in the past since I feel it matches the topic of interest for this week's discussion. https://archive.ics.uci.edu/ml/datasets/Auto+MPG

We need to tidy the data up a bit before we do anything.

```
auto <- read.table(url("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data
```

```
names(auto) <- c("mpg", "cylinders",
                 "displacement",
                 " horsepower",
                 "weight",
                 "acceleration",
                 "model year",
                 "origin",
                 "car name")
```

```
auto.df<-data.frame(auto)
auto.df$X.horsepower <- as.numeric(as.character(auto.df$X.horsepower))
```

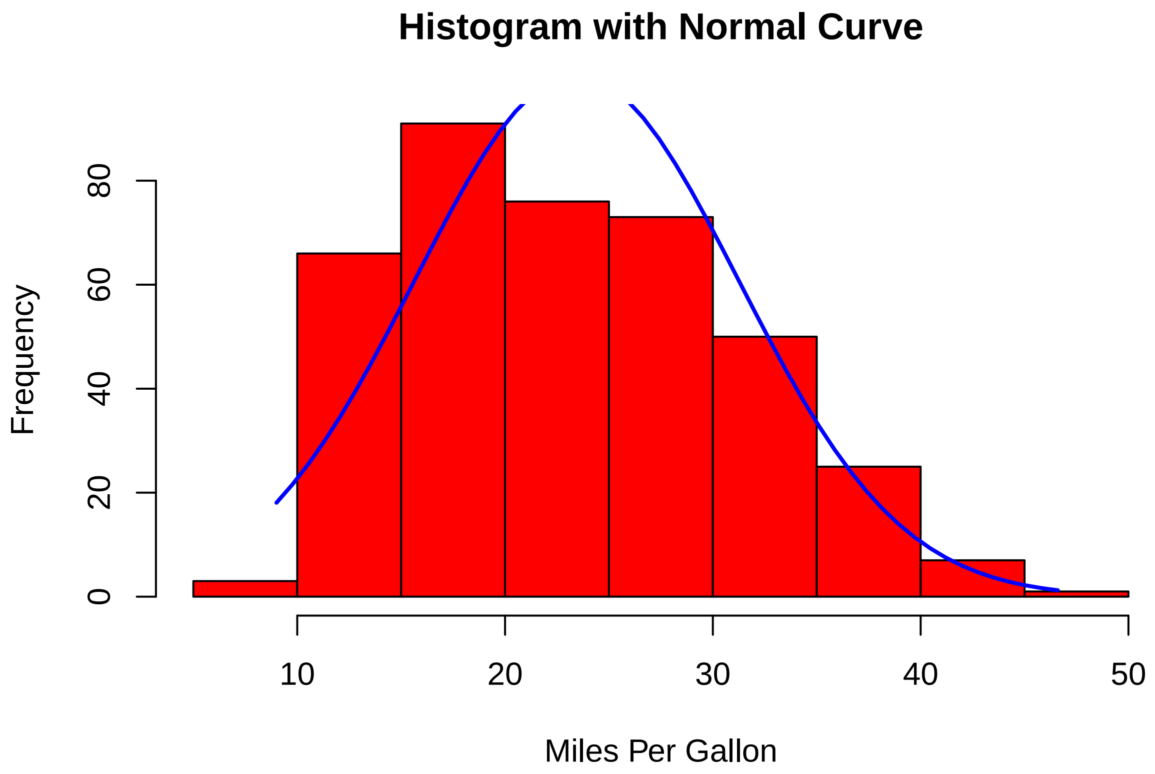
```
## Warning: NAs introduced by coercion
```

```
auto.df<-na.omit(auto.df)
summary(auto.df)
```

```
##       mpg          cylinders      displacement    X.horsepower       weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration     model.year        origin         car.name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

Check distribution of response variable

```
mpg <- auto.df$mpg
h<-hist(mpg, breaks=10, col="red", xlab="Miles Per Gallon",
    main="Histogram with Normal Curve")
xfit<-seq(min(mpg),max(mpg),length=50)
yfit<-dnorm(xfit,mean=mean(mpg),sd=sd(mpg))
yfit <- yfit*diff(h$mids[1:2])*length(mpg)
lines(xfit, yfit, col="blue", lwd=2)
```



Build model on variables that do not include levels

```
auto.df2 <- subset(auto.df, select = c(mpg, cylinders, displacement, X.horsepower, weight,acceleration)
```

Build a model

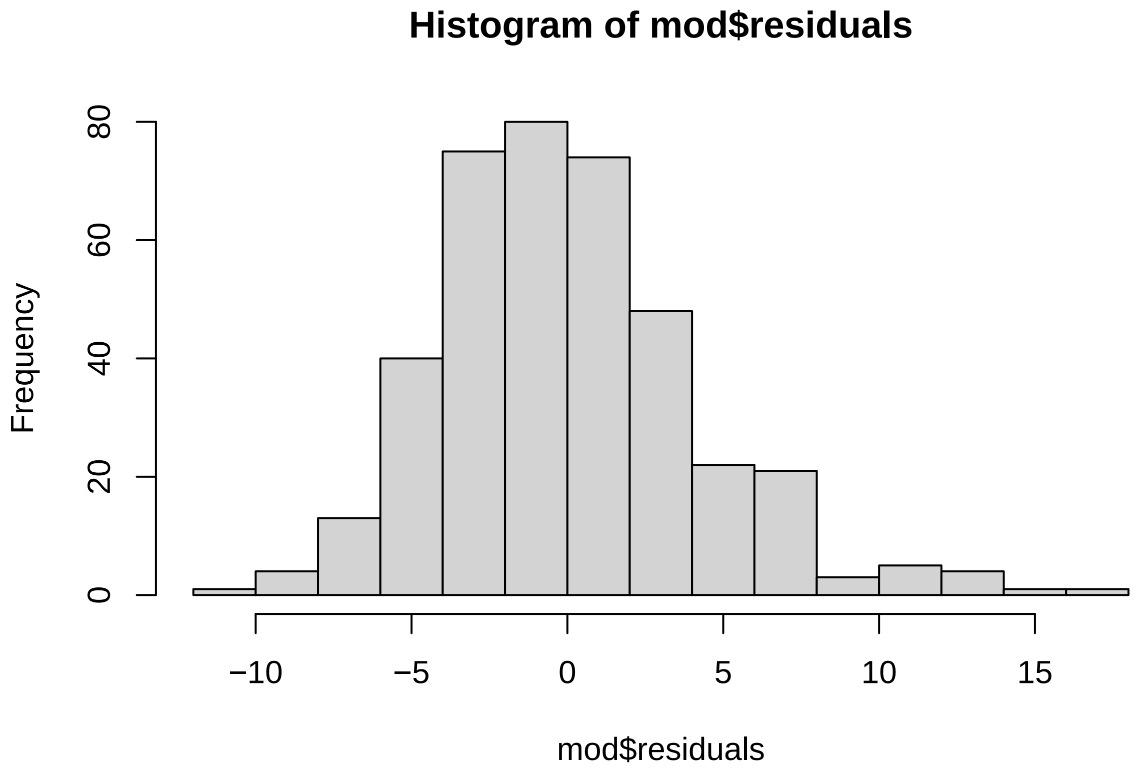
```
mod <- lm(mpg ~ ., data=auto.df2)
summary(mod)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto.df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5816  -2.8618  -0.3404   2.2438  16.3416
```

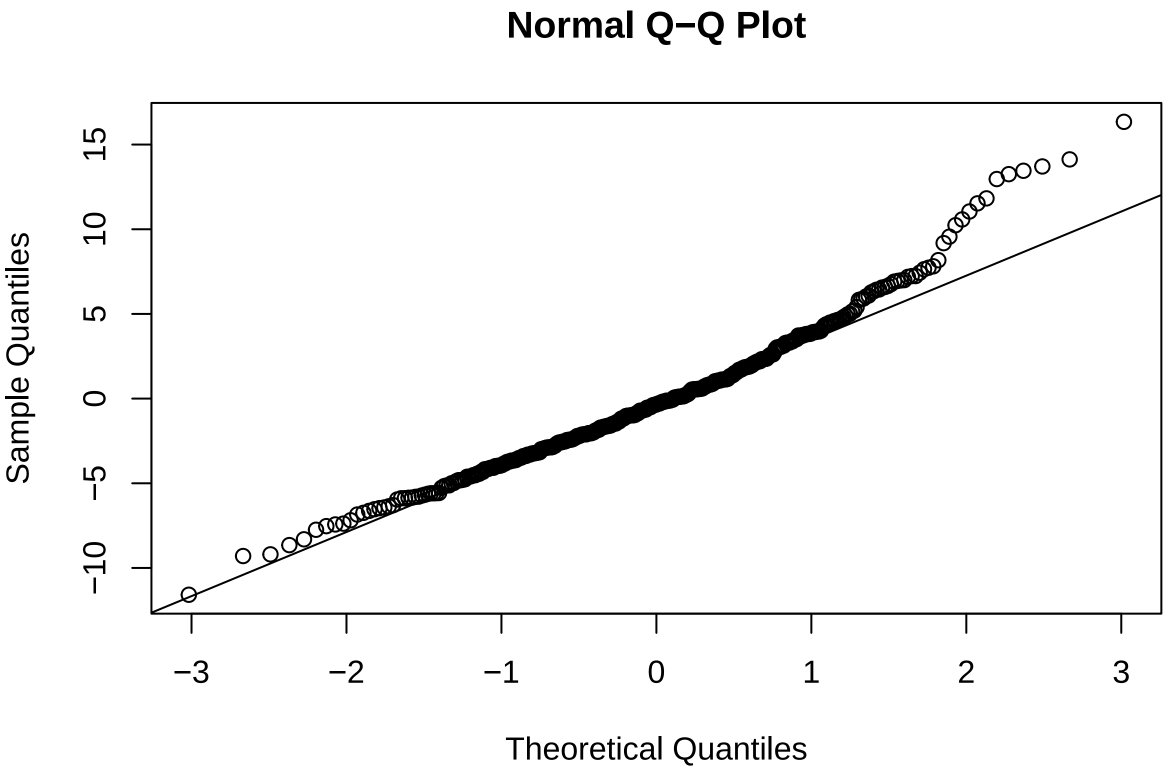
```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
## cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
## displacement -8.313e-05  9.072e-03  -0.009   0.9927
## X.horsepower -4.526e-02  1.666e-02  -2.716   0.0069 **
## weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
## acceleration -2.910e-02  1.258e-01  -0.231   0.8171
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.247 on 386 degrees of freedom
## Multiple R-squared:  0.7077, Adjusted R-squared:  0.7039
## F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

Model Validation Residuals

```
hist(mod$residuals);
```



```
qqnorm(mod$residuals);
qqline(mod$residuals)
```



Constant Variance

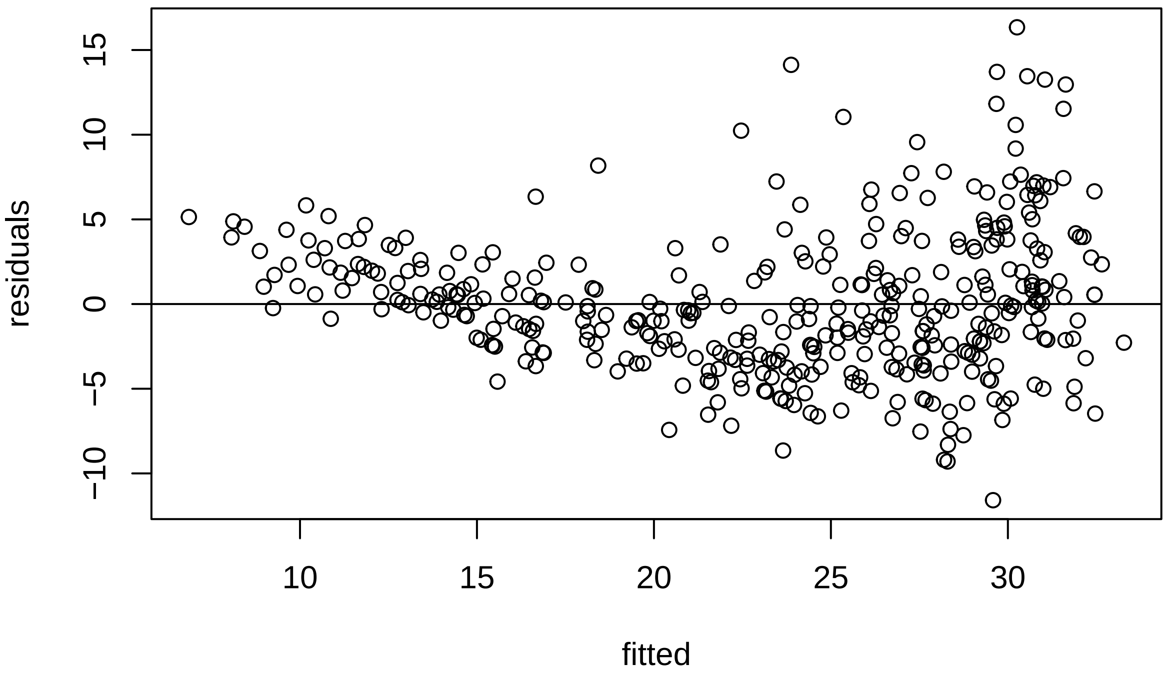
```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked _by_ '.GlobalEnv':
##
##     auto

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_test_breusch_pagan(mod);
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  -------------------------------
##  Response : mpg
##  Variables: fitted values of mpg
##
##          Test Summary
##  -------------------------------
##  DF              =    1
##  Chi2            =    39.77868
##  Prob > Chi2     =    2.844324e-10
```

```
plot(fitted(mod), residuals(mod), xlab="fitted", ylab="residuals")
abline(h=0)
```



Based on these residuals, a linear model is not a good fit. The residuals demonstrate the need for some transformation on the response variable. The constant variance check is also a strong indicator that a linear model is not the way to go.