**Lab 8: Polynomial Regression, Residual Plots and Normality**

Use the below data on cafeteria to answer the following questions.

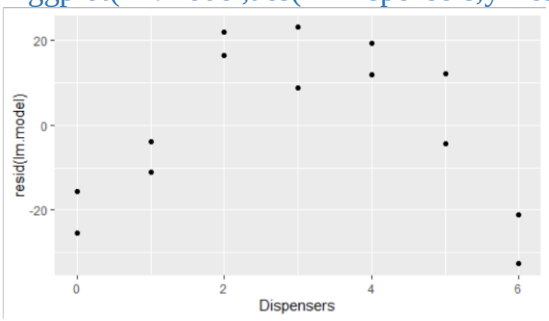
| Obs | Cafeteria | Dispensers | Sales |
|---|---|---|---|
| **1** | 1 | 0 | 507.9 |
| **2** | 2 | 0 | 498.0 |
| **3** | 3 | 1 | 568.3 |
| **4** | 4 | 1 | 575.5 |
| **5** | 5 | 2 | 651.6 |
| **6** | 6 | 2 | 657.1 |
| **7** | 7 | 3 | 713.8 |
| **8** | 8 | 3 | 699.6 |
| **9** | 9 | 4 | 758.4 |
| **10** | 10 | 4 | 765.7 |
| **11** | 11 | 5 | 797.7 |
| **12** | 12 | 5 | 814.3 |
| **13** | 13 | 6 | 836.8 |
| **14** | 14 | 6 | 825.1 |

The variables in the data set are:

**Cafeteria** - cafeteria identification number
**Dispensers** - number of dispensers at each cafeteria
**Sales** - coffee sales in hundreds of gallons

1. Perform a simple linear regression with Sales as the response and Dispensers as the predictor variable. What do you see in the residual versus predicted plot? What would you do to fix this problem?

```
> lm.model=lm(Sales~Dispensers,data=caf)
> ggplot(lm.model,aes(x=Dispensers,y=resid(lm.model)))+geom_point()
```



Looks like we are missing a polynomial term.

2. Perform a forward selection (by hand) using the AIC criteria (you will need to use the command AIC(model) to get the AIC values for each model). The "smallest" model should be the just the intercept. The "biggest" model should be Dispensers up to the power of 4 (be sure to follow model hierarchy). What was the best degree for the polynomial based on AIC?

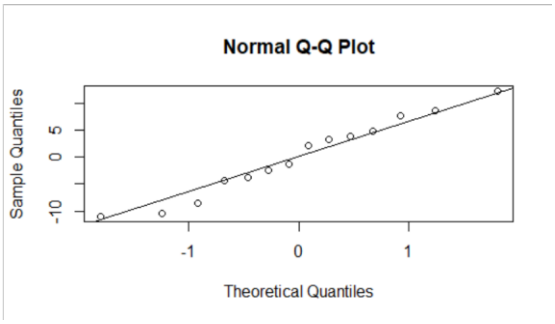
```
> lm0=lm(Sales~1,data=caf)

> AIC(lm0)

[1] 176.0696

> lm1=lm(Sales~Dispensers,data=caf)

> AIC(lm1)

[1] 126.8844

> lm2=lm(Sales~Dispensers + I(Dispensers^2),data=caf)

> AIC(lm2)

[1] 101.9835

> lm3=lm(Sales~Dispensers + I(Dispensers^2) + I(Dispensers^3),data=caf)

> AIC(lm3)

[1] 102.6002

> lm4=lm(Sales~Dispensers + I(Dispensers^2) + I(Dispensers^3)+I(Dispensers^4),data=caf)

> AIC(lm4)

[1] 104.5498
```

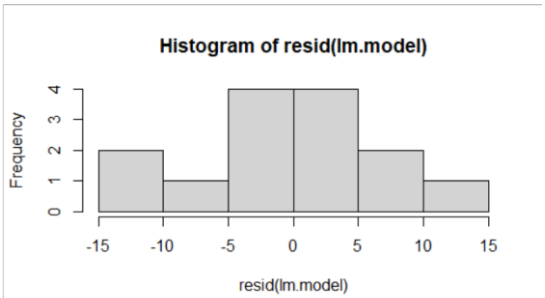
We would stop after a second degree polynomial. So, we have $\hat{y}_i = 499.4 + 84.7x_i - 4.8x_i^2$

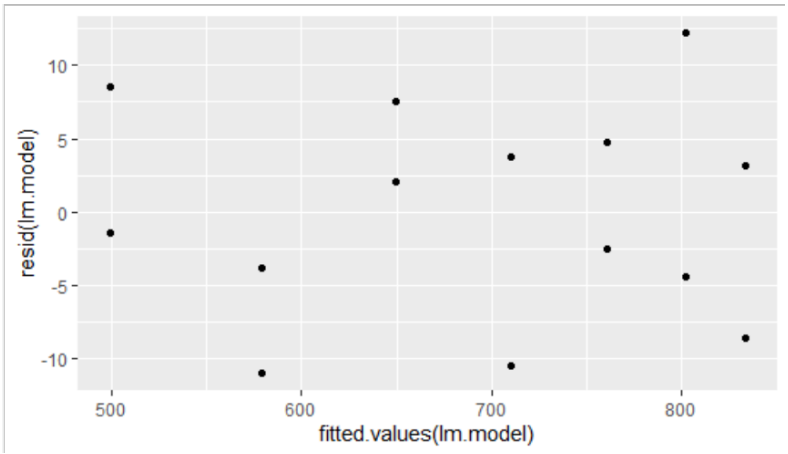
3. Run the model you selected in #2 and look at the residual versus predicted plot. What do you see?
    a. In this model, does the Q-Q plot provide evidence of normality?
    b. Does the histogram of residuals look normally distributed?

```
> lm.model=lm(Sales~Dispensers + I(Dispensers^2),data=caf)

> ggplot(lm.model,aes(x=fitted.values(lm.model),y=resid(lm.model)))+geom_point()

> qqnorm(resid(lm.model))

> qqline(resid(lm.model))

> hist(resid(lm.model))
```

Based on the histogram and QQplot, residuals appear to be normally distributed.


Normal Q-Q Plot
Sample Quantiles
-10
0
5
-1
0
1
Theoretical Quantiles


Histogram of resid(lm.model)
Frequency
0
1
2
3
4
-15
-10
-5
0
5
10
15
resid(lm.model)


resid(lm.model)
10
5
0
-5
-10
500
600
700
800
fitted.values(lm.model)