**CSE 635, Spring 2021, Homework 2**
**Sima Shafaei**

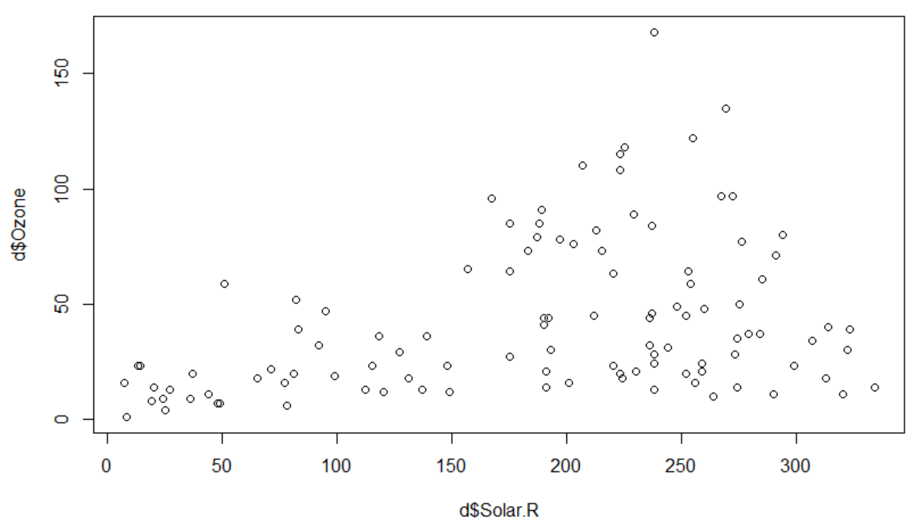
Use the airquality data set to do the following in R

1. Remove all records with NA entries. Find the number of the available records or rows.

| Code: |
|---|
| nrow(airquality)<br>d=na.omit(airquality) #remove null variables<br>head(d)<br>nrow(d) |
| **Answer:** |
| Number of available records before removing Null entries:153<br>Number of available records after removing Null entries:111 |

2. Plot the dependent variable Ozone as a function of the independent variable Solar.R

| Code: |
|---|
| plot(d$Ozone~d$Solar.R) |
| **Answer:** |

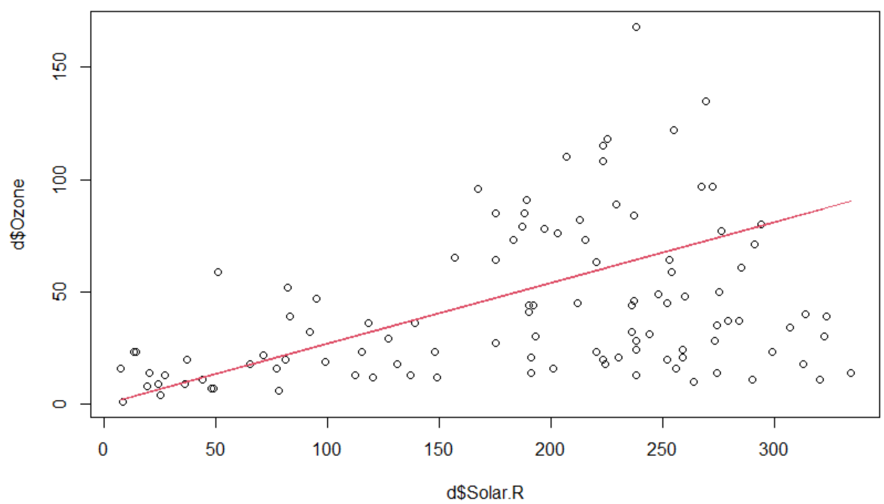



3. Evaluate the predictions using the following models for Ozone ~ Solar.R:
   a. Eyeball linear equation
   b. Linear model; lm

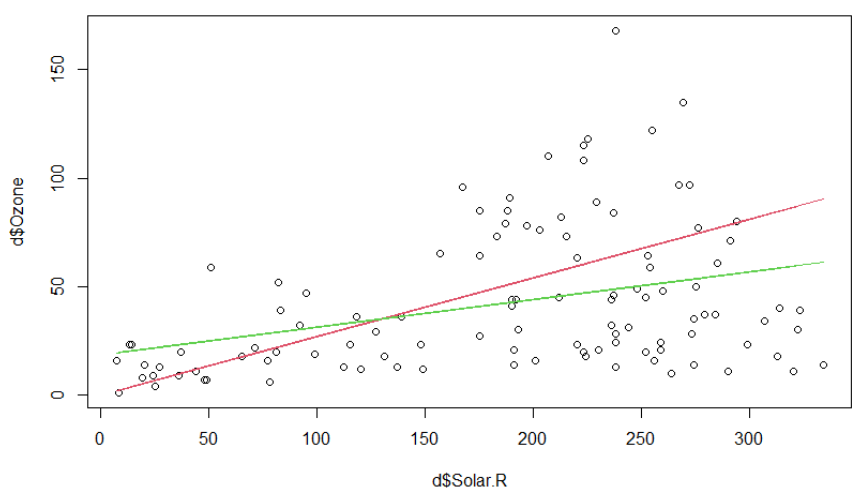
c. Second order polynomial

d. Generalized linear model; glm

| Code: |
| --- |
| # a.==========Eyeball linear equation =====================<br># by considering two point (10, 1) (300,75) az average point in the diagram the<br># slope would be=(75-1)/(280-10)=0.27<br># and my estimation for interception is 0<br># estimation model is: Ozon = 0.27Solar.R<br>p1=0.27*d$Solar.R<br>lines(d$Solar.R,p1,col=2)<br># b.==========Linear model; lm =====================<br>m2=lm(d$Ozone~d$Solar.R)<br>summary(m2)<br>c2=coef(m2)<br>p2=c2[1] + c2[2]*d$Solar.R<br>lines(d$Solar.R,p2,col=3)<br><br># c.========== Second order polynomial =====================<br>x2=d$Solar.R * d$Solar.R<br>m3=lm(d$Ozone~d$Solar.R+x2)<br>summary(m3)<br>c3=coef(m3)<br>p3= c3[1]+c3[2]*d$Solar.R+c3[3]*x2<br>lines(d$Solar.R,p3,col=4)<br><br># d.========== Generalized linear model; glm =====================<br>m4 = glm(d$Ozone~d$Solar.R, family = "poisson")<br>summary(m4)<br>c4=coef(m4)<br>p4 = exp(c4[1]+c4[2]*d$Solar.R)<br>lines(d$Solar.R,p4,col=5) |
| Answer: |
| Plot m1 on data: |




Plot m2 on data:



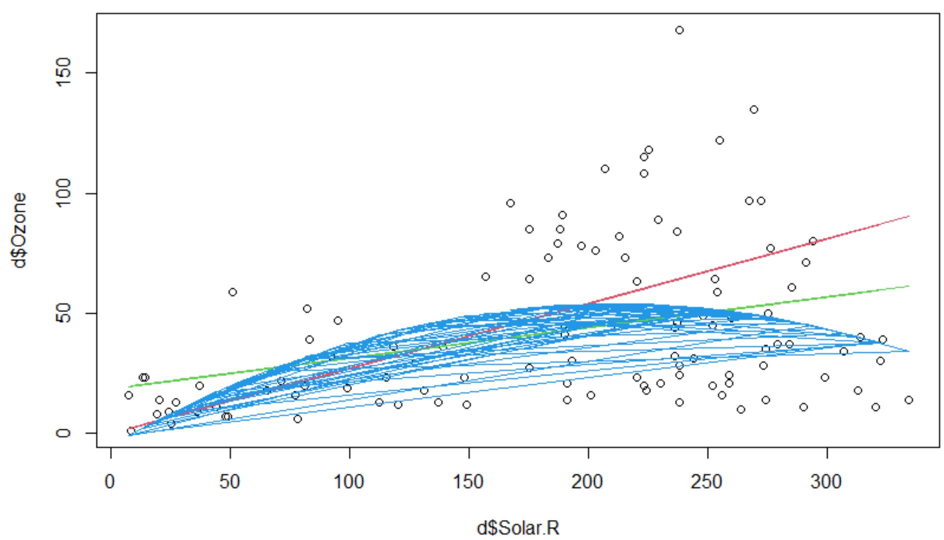

Summary of m2:

```
Call:
lm(formula = d$Ozone ~ d$Solar.R)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.292 -21.361  -8.864  16.373 119.136 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.59873    6.74790   2.756 0.006856 ** 
d$Solar.R    0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.33 on 109 degrees of freedom
Multiple R-squared:  0.1213,    Adjusted R-squared:  0.1133 
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

The small p-value for the intercept and the slope indicates that we can reject the null hypothesis. Here 3 star for slope shows that slop is more significant than intercept with 2 stars.

Multiple R-Squared = 0.1213 shows that the variation that the model carries from the variations given in the data is 0.1213 which shows it is a very poor model in terms of the variation

Plot m3 on data:



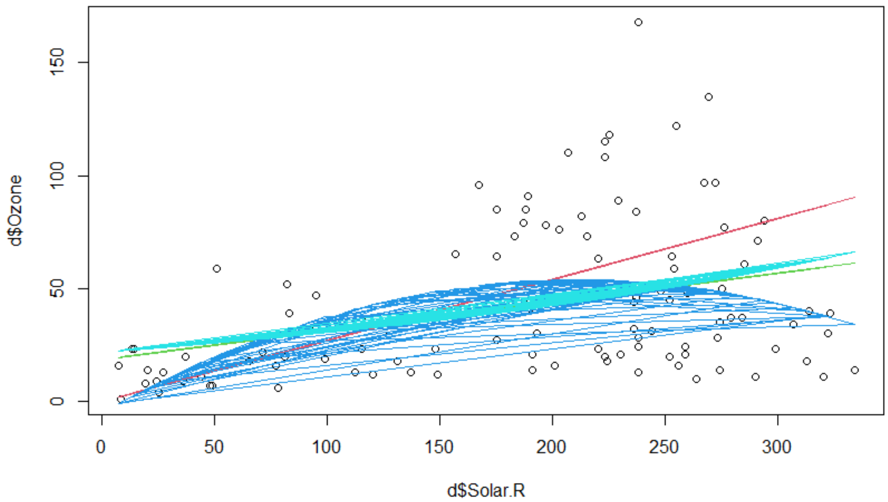

Summary of m3:

```
Call:
lm(formula = d$Ozone ~ d$Solar.R + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.155 -22.793  -6.438  18.061 115.117 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.7561171  9.2761865  -0.513 0.609192    
d$Solar.R    0.5550868  0.1264847   4.389 2.67e-05 ***
x2          -0.0013147  0.0003766  -3.491 0.000698 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.84 on 108 degrees of freedom
Multiple R-squared:  0.2104,	Adjusted R-squared:  0.1958 
F-statistic: 14.39 on 2 and 108 DF,  p-value: 2.875e-06
```

The small p-value for the intercept and the slope indicates that we can reject the null hypothesis. Here we can conclude that intercept is not a significant variable Solar.R and Solar.$R^2$ are significant variables.

Multiple R-Squared = 0.2104 shows that the variation that the model carries from the variations given in the data is 0.2104 which shows it is a poor model in terms of the variation but it is better than LM

Plot m4 on data:




Summary of m4:

```
Call:
glm(formula = d$Ozone ~ d$Solar.R, family = "poisson")

Deviance Residuals:
   Min       1Q  Median       3Q      Max
-8.100  -3.873  -1.713    2.616   13.435

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.0888635  0.0399598   77.30   <2e-16 ***
d$Solar.R   0.0032948  0.0001774   18.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2627.1  on 110  degrees of freedom
Residual deviance: 2256.4  on 109  degrees of freedom
AIC: 2844.5

Number of Fisher Scoring iterations: 5
```

The small p-value for the intercept and the slope indicates that we can reject the null hypothesis. Here we can conclude that intercept and Solar.R are both significant variables and reject null hypothesis

4. In each case present the following: coefficient, summary statistics of the error vector, and SSE. Also include a plot that shows the response of these models.

Code:

```
# a.==========Eyeball linear equation =====================
c1<- c(0,0.27)
c1
e1 = p1 - d$Ozon
summary(e1)
hist(e1)
# the histogram is not a normal distribution and mean is not 0 so it is not a good estimation
SSE1=sum(t(e1)*e1)
SSE1
# b.==========Linear model; lm ===========================
c2
e2=p2-d$Ozone
summary(e2)
hist(e2)
SSE2=sum(t(e2)*e2)
SSE2
# c.========== Second order polynomial =====================
c3
e3=p3-d$Ozone
summary(e3)
hist(e3)
SSE3=sum(t(e3)*e3)
SSE3
# d.========== Generalized linear model; glm =====================
c4
e4=p4-d$Ozone
summary(e4)
hist(e4)
SSE4=sum(t(e4)*e4)
SSE4
```

Answer:

Coefficients:

| | Intercept | Slope | $X^2$ |
|---|---|---|---|
| eyeball | 0 | 0.27 | --- |
| LM | 18.5987278 | 0.1271653 | --- |
| Polynomial | -4.756117103 | 0.555086789 | -0.001314735 |
| GLM | 3.088863547 | 0.003294806 | --- |

Summary statistics of the error vector:

| | Min | 1st Qu | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| eyeball | -103.740 | -15.770 | 7.300 | 7.797 | 36.330 | 76.180 |

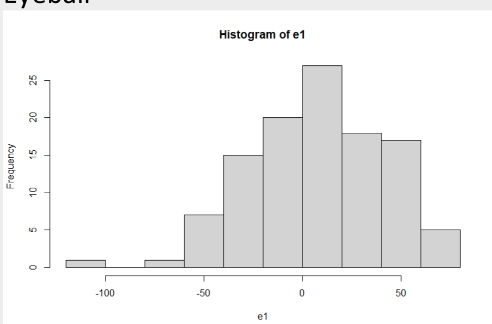
| | | | | | | |
|---|---|---|---|---|---|---|
| LM | -119.136 | -16.373 | 8.864 | 0.000 | 21.361 | 48.292 |
| Polynomial | -115.117 | -18.061 | 6.438 | 0.000 | 22.793 | 40.155 |
| GLM | -119.912 | -17.331 | 9.065 | 0.000 | 21.507 | 52.004 |

SSE:

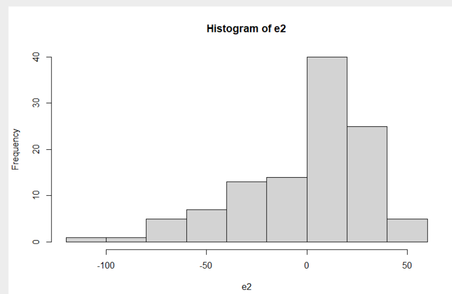
| | SSE |
|---|---|
| eyeball | 132417.4 |
| LM | 107022.2 |
| Polynomial | 96169.17 |
| GLM | 111026 |

Histograms of errors:

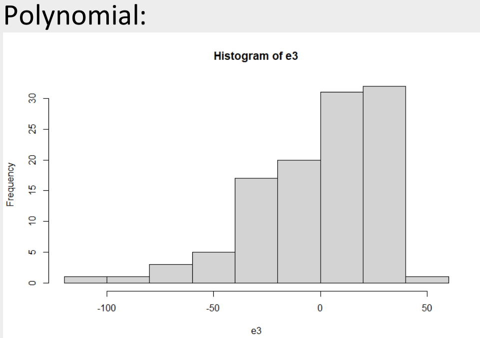
Eyeball



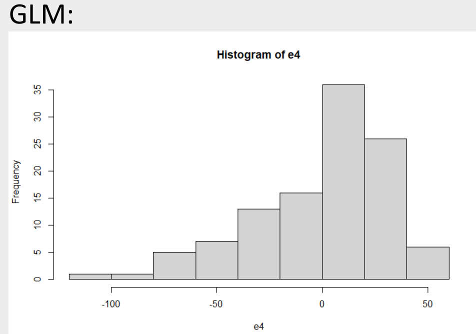

LM:




Polynomial:




GLM:




Which model is the best? And why?

1) None of the error histogram have a normal distribution shape (the distribution of GLM and LM are similar)
2) Considering SSE the best model is Polynomial and the worst model is eyeball. After eyeball GLM is the weakest model

3) Considering Multiple R-Squared that shows the goodness of fit of a model
   Polynomial is better than LM
   Based on above observation we conclude that for this problem Polynomial method is better than other models

**Thank you**