CSE 635, Spring 2021, Homework 3
Sima Shafaei

Hello! Because I miss understood about "summary of statistics" in homework 2, I add this part to this homework. I would appreciate if you could accept it as a part of HW2

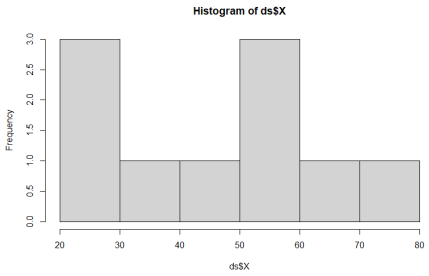
1. Entropy of X.

Code:

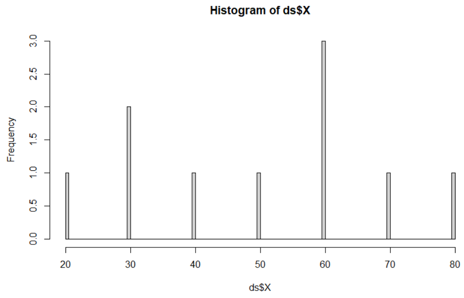
```
#==============================================================================
#                        Entropy of X
#==============================================================================
setwd("D:\\PhD+\\semester 4\\Datamining with linear model\\assignments\\HW1")
ds=read.table("myData.txt", header = TRUE)
head(ds)
str(ds)
# we can get some of statistical summary using "summary" command in R
summary(ds)
hx=hist(ds$X)
hx100=hist(ds$X,breaks = 100)
length(hx100$counts)
sum(hx100$counts)
px=hx100$counts/sum(hx100$counts)
plot(px)
qx=px[px>0]
length(qx) #shorter than p
entropy_x = sum(-qx*log2(qx))
entropy_x
max_entropy=log2(100)
max_entropy
```

Results:

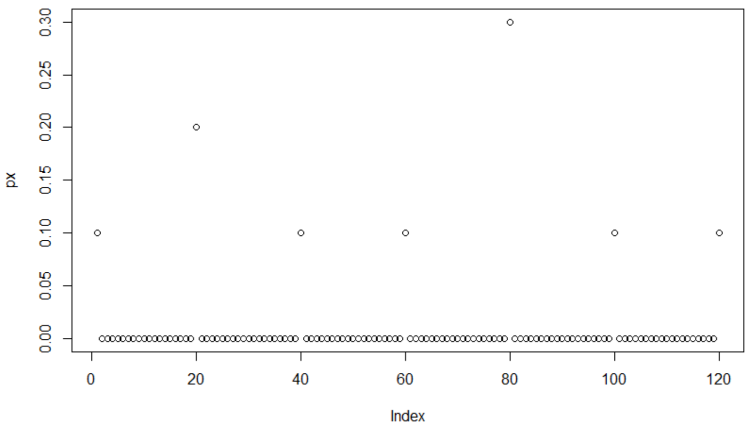
Histogram of X:



Histogram of X after breaking to 100 bins:



P(X):



Entropy of x: 2.646439

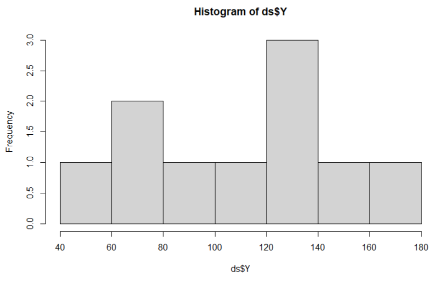
2. Entropy of Y.

Code:

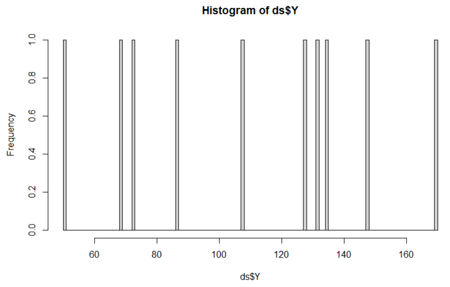
```
#==========================================================================
#                    Entropy of Y
#==========================================================================
hy=hist(ds$Y)
hy100=hist(ds$Y,breaks = 100)
length(hy100$counts)
sum(hy100$counts)
py=hy100$counts/sum(hy100$counts)
plot(py)
qy=py[py>0]
length(qy) #shorter than p
entropy_y = sum(-qy*log2(qy))
entropy_y
max_entropy_y=log2(100)
max_entropy_y
```

Results:

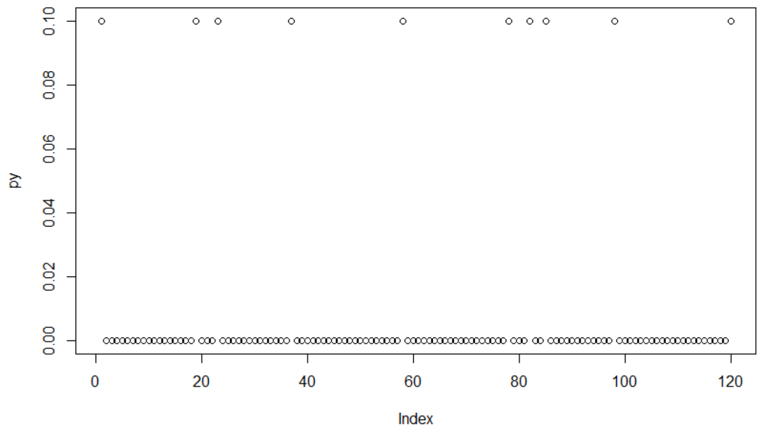
Histogram of Y:



Histogram of Y after breaking to 100 bins:



P(Y):



Entropy of x: 3.321928

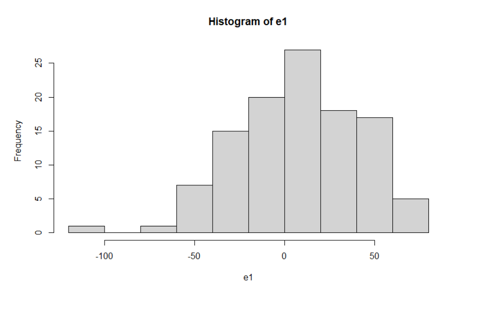
3. Summary of statistics and entropy of error in Eyeball model

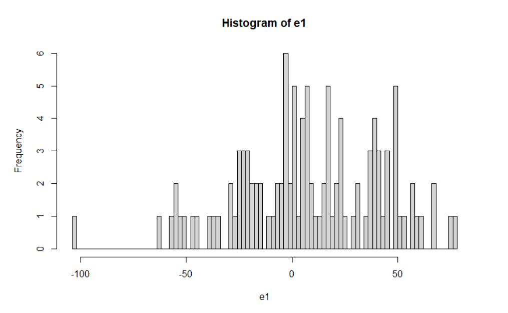
Code:

```
mean(e1)
sd(e1)
var(e1)
median(e1)
IQR(e1)
skewness(e1)
kurtosis(e1)
min(e1)
max(e1)
range(e1)
h_e1=hist(e1)


h_e1_100=hist(e1,breaks = 100)
length(h_e1_100$counts)
sum(h_e1_100$counts)
p_e1=h_e1_100$counts/sum(h_e1_100$counts)
plot(p_e1)
q_e1=p_e1[p_e1>0]
length(q_e1) #shorter than p
entropy_e1 = sum(-q_e1*log2(q_e1))
entropy_e1
max_entropy_e1=log2(100)
max_entropy_e1
```

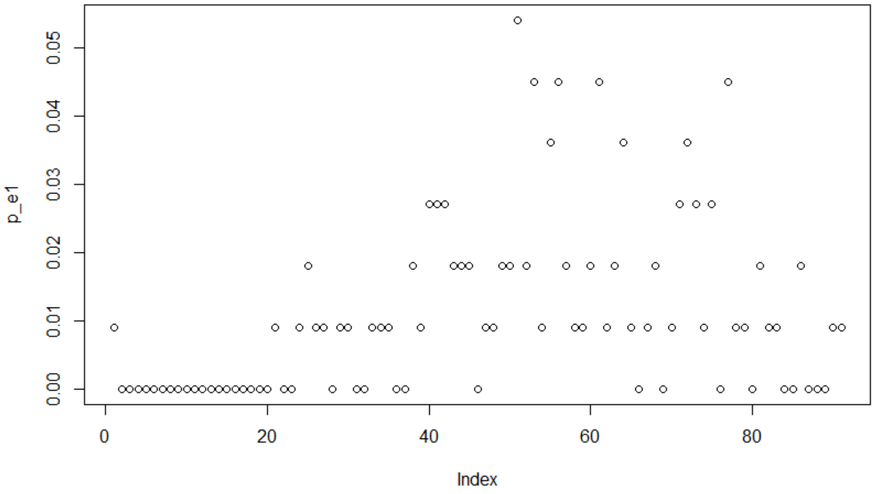
Results:

Histogram of eyeball Error:	Histogram of eyeball error after breaking to 100 bins:





P(eyeball_error):



Summary of statistics of eyeball error:

mean	**7.797387**
Standard deviation	**33.80003**
Variance	**1142.442**
Median	**7.3**
IQR	**52.1**
Skewness	**-0.324083**
Kurtosis	**3.049029**
Min	**-103.74**
Max	**76.18**
Range	**-103.74 .. 76.18**
Entropy	**5.510833**

Skewness is near zero and kurtosis is about 3 whish shows that the shape of distribution is almost near to normal distribution. However, mean is 7.7 which is very higher than zero and entropy is 5.5 which is near maximum entropy (6.64) therefore this error contains information and the model is missing this information (the model is not good)

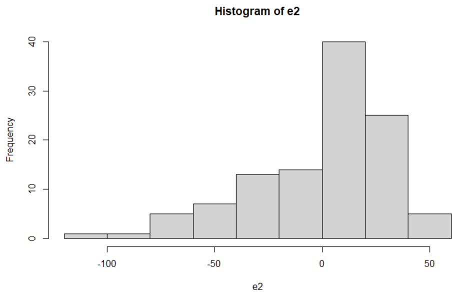
4. Summary of statistics and entropy of error in LM model

Code:

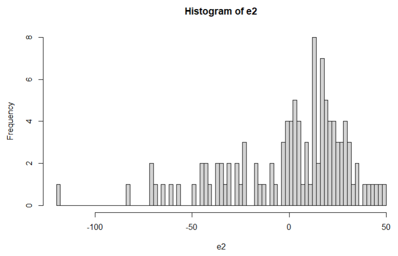
```
mean(e2)
sd(e2)
var(e2)
median(e2)
IQR(e2)
skewness(e2)
kurtosis(e2)
min(e2)
max(e2)
range(e2)
h_e2=hist(e2)


h_e2_100=hist(e2,breaks = 100)
length(h_e2_100$counts)
sum(h_e2_100$counts)
p_e2=h_e2_100$counts/sum(h_e2_100$counts)
plot(p_e2)
q_e2=p_e2[p_e2>0]
length(q_e2) #shorter than p
entropy_e2 = sum(-q_e2*log2(q_e2))
entropy_e2
max_entropy_e2=log2(100)
max_entropy_e2
```

Results:

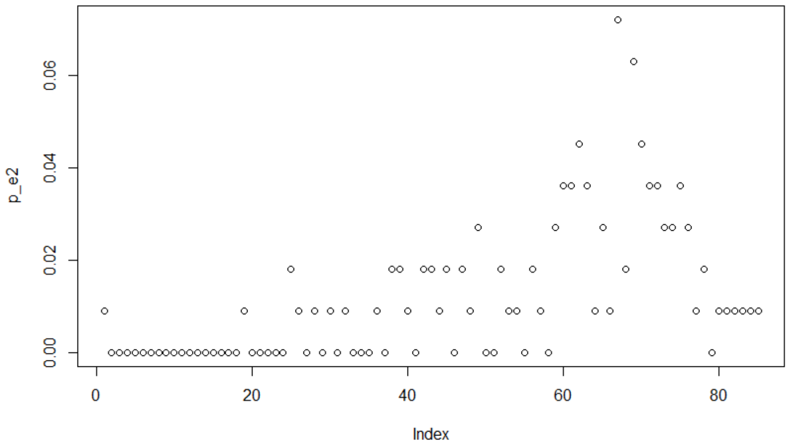
Histogram of LM Error:



Histogram of LM error after breaking to 100 bins:



P(LM_error):



Summary of statistics of LM error:

mean	**1.882125e-14**
Standard deviation	**31.19182**
Variance	**972.9294**
Median	**8.86447**
IQR	**37.73405**
Skewness	**-1.127785**
Kurtosis	**4.214179**
Min	**-119.1359**
Max	**48.29161**
Range	**-119.13594 .. 48.29161**
Entropy	**5.304325**

Mean is near zero but the skewness is negative so the distribution has longer tail in the left side and kurtosis is about 4.2>3 thus the distribution is pointier than normal distribution. These information show that the shape of distribution is not like to normal distribution. Entropy is 5.3 which is near maximum entropy (6.64) therefore this error contains information and the model is missing this information (the model is not good). However, this model is better than eyeball model

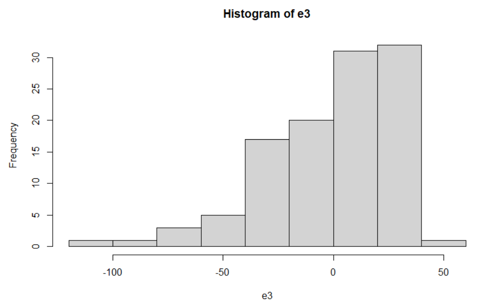
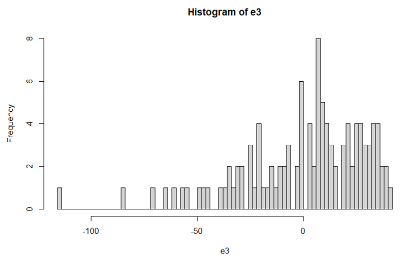
5. Summary of statistics and entropy of error in Polynomial model

Code:

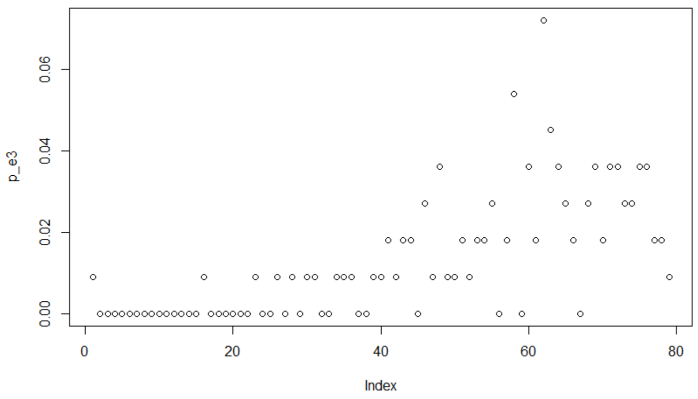
```
mean(e3)
sd(e3)
var(e3)
median(e3)
IQR(e3)
skewness(e3)
kurtosis(e3)
min(e3)
max(e3)
```

```
range(e3)
h_e3=hist(e3)


h_e3_100=hist(e3,breaks = 100)
length(h_e3_100$counts)
sum(h_e3_100$counts)
p_e3=h_e3_100$counts/sum(h_e3_100$counts)
plot(p_e3)
q_e3=p_e3[p_e3>0]
length(q_e3) #shorter than p
entropy_e3 = sum(-q_e3*log2(q_e3))
entropy_e3
max_entropy_e3=log2(100)
max_entropy_e3
```

Results:

Histogram of Polynomial Error:	Histogram of Polynomial error after breaking to 100 bins:
	

P(Polynomial_error):



Summary of statistics of Polynomial error:

mean	**1.589654e-14**
Standard deviation	**29.56798**
Variance	**874.2652**
Median	**6.437755**

IQR	**40.85343**
Skewness	**-1.089931**
Kurtosis	**4.39418**
Min	**-115.1173**
Max	**40.15503**
Range	**-115.11731 .. 40.15503**
Entropy	**5.284067**

Mean is almost 0 but skewness is negative so the distribution has longer tail in the left side and kurtosis is about 4.39>3 thus the distribution is pointier than normal distribution. These information show that the shape of distribution is not like to normal distribution. Entropy is 5.28 which is near maximum entropy (6.64) therefore this error contains information and the model is missing this information (the model is not good).

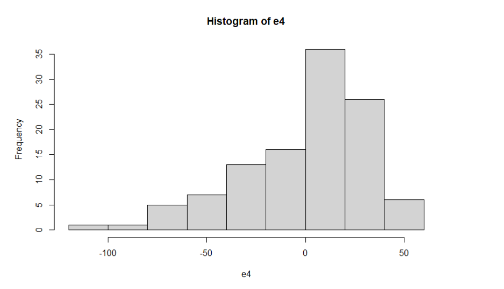
6. Summary of statistics and entropy of error in GLM model

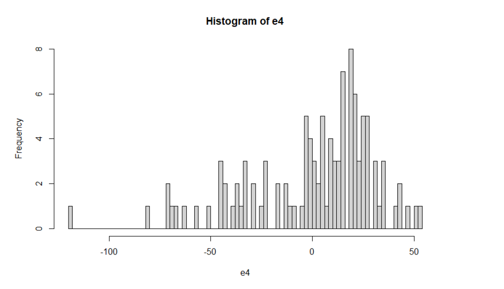
Code:

```
mean(e4)
sd(e4)
var(e4)
median(e4)
IQR(e4)
skewness(e4)
kurtosis(e4)
min(e4)
max(e4)
range(e4)
h_e4=hist(e4)


h_e4_100=hist(e4,breaks = 100)
length(h_e4_100$counts)
sum(h_e4_100$counts)
p_e4=h_e4_100$counts/sum(h_e4_100$counts)
plot(p_e4)
q_e4=p_e4[p_e4>0]
length(q_e4) #shorter than p
entropy_e4 = sum(-q_e4*log2(q_e4))
entropy_e4
max_entropy_e4=log2(100)
max_entropy_e4
```

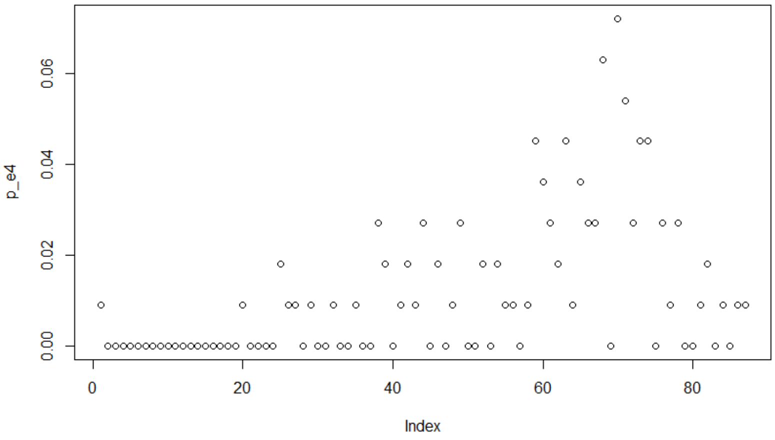
Results:

Histogram of GLM Error:	Histogram of GLM error after breaking to 100 bins:





P(GLM_error):



Summary of statistics of LM error:

mean	**-6.324124e-14**
Standard deviation	**31.76991**
Variance	**1009.327**
Median	**9.065036**
IQR	**38.83755**
Skewness	**-1.087975**
Kurtosis	**4.11407**
Min	**-119.9123**
Max	**52.0042**
Range	**-119.9123 .. 52.0042**
Entropy	**5.169246**

Mean is almost 0 but skewness is negative so the distribution has longer tail in the left side and kurtosis is about 4.11>3 thus the distribution is pointier than normal distribution. These information show that the shape of distribution is not like to normal distribution. Entropy is 5.16 which is near maximum entropy (6.64) therefore this error contains information and the model is missing this information (the model is not good).

Thank you