Yichen Dong Module 8 HW

Yichen Dong

October 25, 2018

Problem 1

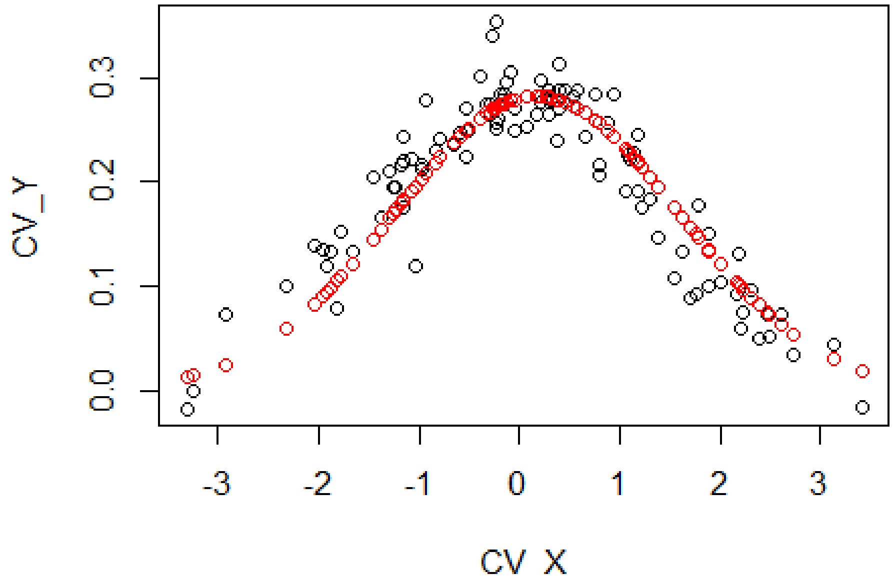
Part A

```
CV_X = scan("CV_X.txt")
CV_Y = scan("CV_Y.txt")
mean_CV = mean(CV_X)
mean_CV

## [1] 0.1734736
```

Our MLE for mu is the mean of the dataset, or .1734. This means that we have a normal pdf 1/sqrt(4*pi)*exp$^{(-(x-.1734)}$2/4)

```
CV = as.data.frame(cbind(CV_X,CV_Y))
CV = CV %>%
  mutate(normal = dnorm(CV_X,mean_CV,sqrt(2)))%>%
  mutate(error = abs(CV_Y-normal))
plot(CV_X,CV_Y)
points(CV_X,CV$normal,col = "red")
```



Here we have the points of the data along with the predicted values based on a N(.1734,2) distribution.

```
sum(CV$error)/length(CV$CV_X)

## [1] 0.02518254
```

This is our apparent error.

Part B and C

```
CV = CV %>%
  mutate(group = floor((as.numeric(rownames(CV))-1)/50))
mean_CV0 = mean(CV$CV_X[CV$group == 0])
mean_CV1 = mean(CV$CV_X[CV$group == 1])
CV = CV %>%
  mutate(norm_group_0 = dnorm(CV_X,mean_CV1,sqrt(2))
         ,norm_group_1 = dnorm(CV_X,mean_CV0,sqrt(2)))
CV = CV%>%
  mutate(error_0 = ifelse(group == 0, abs(CV_Y - norm_group_0),0)
         ,error_1 = ifelse(group == 1, abs(CV_Y - norm_group_1),0))
(sum(CV$error_0)+sum(CV$error_1))/length(CV$CV_X)

## [1] 0.03103595
```

This is our cross validated error using two halves of the dataset. As we can see, our second value was larger than the first. This means that our apparent error likely underpredicted the actual error that would appear. By using balnced half sampling, we are able to get closer to the true error that would occur with fitting the data.

Problem 2

Problem 2

$$\widehat{V(T)}_j = \frac{\sum_{i=1}^{r} (T_i^* - J(T))^2}{r(r-1)}$$

$T_j^* = rT - (r-1)T_{(-j)}$

$J(T) = \frac{1}{r}\sum_{i=1}^{r} T_j^* = rT - (r-1)\bar{T}_{(\cdot)}$

$r = n$

$\bar{T}_{(\cdot)} = \frac{1}{r}\sum_{j=1}^{r} T_{(-j)}$

$$= \frac{\sum_{j=1}^{r} (nT - (n-1)T_{(-j)} - (nT - (n-1)\bar{T}_{(\cdot)}))^2}{n(n-1)}$$

$$= \frac{\sum_{j=1}^{n} (T_{(-j)} - nT_{(-j)} - \bar{T}_{(\cdot)} + n\bar{T}_{(\cdot)})^2}{n(n-1)}$$

$$= \frac{\sum_{j=1}^{n} (T_{(-j)}(1-n) - (\bar{T}_{(\cdot)})(1-n))^2}{n(n-1)}$$

$$= \frac{\sum_{j=1}^{n} (-1)^2 (n-1)^2 (T_{(-j)} - \bar{T}_{(\cdot)})^2}{n(n-1)}$$

$$= \frac{(n-1)^2 \sum_{j=1}^{n} (T_{(-j)} - \bar{T}_{(\cdot)})^2}{n(n-1)}$$

Since $a^2 \cdot b^2 = (a \cdot b)^2$, and $(-1)(1-n) = (n-1)$, we can factor those out and cancel and get our final answer

$$\frac{n-1}{n} \sum_{j=1}^{n} (T_{(-j)} - \bar{T}_{(\cdot)})$$

Part b.

From part a: $\frac{n-1}{n} \sum_{j=1}^{n} (T_{(-j)} - \bar{T}_{(\cdot)})^2$

$T_j^{*} = nT - (n-1) T_{(-j)} \rightarrow T_{(-j)} = \frac{nT - T_j^{*}}{n-1}$

$$\frac{n-1}{n} \sum_{j=1}^{n} \left(T_{(-j)} - \frac{1}{n} \cdot \sum_{j=1}^{n} T_{(-j)} \right)^2$$

$$= \frac{n-1}{n} \sum_{j=1}^{n} \left(\left(\frac{1}{n-1} \right) \left(nT - T_j^{*} - \frac{1}{n} \cdot \sum_{j=1}^{n} (nT - T_j^{*}) \right) \right)^2$$

$$\sum_{j=1}^{n} nT = n^2 T \checkmark$$

$$= \frac{n-1}{n} \sum_{j=1}^{n} \left(\left(\frac{1}{n-1} \right) \left(nT - \frac{n^2}{n} T - T_j^{*} + \frac{1}{n} \cdot \sum_{j=1}^{n} T_j^{*} \right) \right)^2$$

$$= \frac{n-1}{n} \cdot \frac{1}{(n-1)^2} \cdot \sum_{j=1}^{n} (-T_j^{*} + J(T))^2$$

$$= \frac{\sum_{j=1}^{n} (-T_j^{*} + J(T))^2}{n(n-1)} = \frac{\sum_{j=1}^{n} (T_j - J(T))^2}{n(n-1)}$$

Since $(T_j - J(T))^2 \leq (T_j^{*} - T)^2$, then

$$\hat{V}(T)_j \leq \frac{\sum_{j=1}^{n} (T_j^{*} - T)^2}{n(n-1)}$$

Problem 3

Part a

Problem 3

part a.

$$b_{2(-j)} = \frac{\sum_{i=1}^{j-1}(y_i - \bar{y}(-j))^4 + \sum_{i=j+1}^{n}(y_i - \bar{y}(-j))^4}{\left(\sum_{i=1}^{j-1}(y_i - \bar{y}(-j))^2 + \sum_{i=j+1}^{n}(y_i - \bar{y}(-j))\right)^2}$$

$$\widehat{V(T)}_J = \frac{n-1}{n} \sum_{j=1} \left(b_{2(-j)} - \bar{b}_{2(\cdot)} \right)$$

where

$$\bar{b}_{2(\cdot)} = \frac{1}{r} \sum_{j=1}^{r} b_{2(-j)}$$

and

$$\bar{y}(-j) = \frac{\sum_{i=1}^{j-1} y_i + \sum_{i=j+1}^{n} y_i}{n-1}$$

Part b

```
jackknife = scan("Jackknife.txt")
jackknife = as.data.frame(jackknife)
jack_mean = mean(jackknife$jackknife)
b2 = sum((jackknife$jackknife -jack_mean)^4)/sum((jackknife$jackknife -jack_m
ean)^2)^2

k=5
groups = length(jackknife$jackknife)/k
jack_group = rep(1:groups,each =k)
jackknife= cbind.data.frame(jackknife,jack_group)
T_minus_j = NULL

for(i in 1:groups){
  jk_minus_j = jackknife$jackknife[jackknife$jack_group != i]
  jk_minus_j_mean = mean(jk_minus_j)
  T_minus_j[i] = sum((jk_minus_j -jk_minus_j_mean)^4)/sum((jk_minus_j -jk_min
us_j_mean)^2)^2
}

T_bar_dot = mean(T_minus_j)
J_T = groups*b2 - (groups-1)*T_bar_dot
```

```r
jk_var = (groups-1)/groups * sum((T_minus_j - T_bar_dot)^2)
paste("Values for k=",k,"; b2:",round(b2,4),", J_T:",round(J_T,6),", T_bar_do
t:",round(T_bar_dot,4),"SD", round(sqrt(jk_var),4))
```

```
## [1] "Values for k= 5 ; b2: 0.0267 , J_T: 1e-04 , T_bar_dot: 0.0281 SD 0.00
37"
```

```r
k=1
groups = length(jackknife$jackknife)/k
jack_group = rep(1:groups,each =k)
jackknife= cbind.data.frame(jackknife,jack_group)
T_minus_j = NULL

for(i in 1:groups){
  jk_minus_j = jackknife$jackknife[jackknife$jack_group != i]
  jk_minus_j_mean = mean(jk_minus_j)
  T_minus_j[i] = sum((jk_minus_j -jk_minus_j_mean)^4)/sum((jk_minus_j -jk_min
us_j_mean)^2)^2
}

T_bar_dot = mean(T_minus_j)
J_T = groups*b2 - (groups-1)*T_bar_dot

jk_var = (groups-1)/groups * sum((T_minus_j - T_bar_dot)^2)
paste("Values for k=",k,"; b2:",round(b2,4),", J_T:",round(J_T,6),", T_bar_do
t:",round(T_bar_dot,4),"SD", round(sqrt(jk_var),4))
```

```
## [1] "Values for k= 1 ; b2: 0.0267 , J_T: -0.00102 , T_bar_dot: 0.027 SD 0.
0067"
```

Part C

```r
for(iter in 1:10){
  k=1
  norm_rand_1 = as.data.frame(rnorm(100,1,sqrt(2)))
  colnames(norm_rand_1) = c("jackknife")
  groups = length(norm_rand_1$jackknife)/k
  jack_group = rep(1:groups,each =k)
  jackknife= cbind.data.frame(norm_rand_1,jack_group)
  jack_mean = mean(norm_rand_1$jackknife)
  b2 = sum((norm_rand_1$jackknife -jack_mean)^4)/sum((norm_rand_1$jackknife -
jack_mean)^2)^2

  T_minus_j = NULL

  for(i in 1:groups){
    jk_minus_j = jackknife$jackknife[jackknife$jack_group != i]
    jk_minus_j_mean = mean(jk_minus_j)
    T_minus_j[i] = sum((jk_minus_j -jk_minus_j_mean)^4)/sum((jk_minus_j -jk_m
inus_j_mean)^2)^2
  }
```

```r
  T_bar_dot = mean(T_minus_j)
  J_T = groups*b2 - (groups-1)*T_bar_dot

  jk_var = (groups-1)/groups * sum((T_minus_j - T_bar_dot)^2)

  print(paste("Values for k=",k,"; b2:",round(b2,4),", J_T:",round(J_T,6),",
T_bar_dot:",round(T_bar_dot,4),"SD", round(sqrt(jk_var),4)))
}
```

```
## [1] "Values for k= 1 ; b2: 0.0325 , J_T: 0.000518 , T_bar_dot: 0.0328 SD 0
.0031"
## [1] "Values for k= 1 ; b2: 0.029 , J_T: 0.000864 , T_bar_dot: 0.0293 SD 0.
0053"
## [1] "Values for k= 1 ; b2: 0.0242 , J_T: 0.000204 , T_bar_dot: 0.0244 SD 0
.0028"
## [1] "Values for k= 1 ; b2: 0.0255 , J_T: -8.5e-05 , T_bar_dot: 0.0257 SD 0
.0026"
## [1] "Values for k= 1 ; b2: 0.0273 , J_T: 5.7e-05 , T_bar_dot: 0.0275 SD 0.
0027"
## [1] "Values for k= 1 ; b2: 0.031 , J_T: 0.000543 , T_bar_dot: 0.0313 SD 0.
0041"
## [1] "Values for k= 1 ; b2: 0.0232 , J_T: 3.8e-05 , T_bar_dot: 0.0235 SD 0.
0022"
## [1] "Values for k= 1 ; b2: 0.032 , J_T: 0.000469 , T_bar_dot: 0.0323 SD 0.
0038"
## [1] "Values for k= 1 ; b2: 0.0265 , J_T: 0.000228 , T_bar_dot: 0.0268 SD 0
.0028"
## [1] "Values for k= 1 ; b2: 0.0262 , J_T: 0.000147 , T_bar_dot: 0.0265 SD 0
.0024"
```

It seems that T_bar_dot is always close to b2, but always slightly higher. J_T is usually close to 0, as well as the standard deviation.