CSCI E-106:Assignment 3

Due Date: September 28, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

Five observations on Y are to be taken when X = 4, 8, 12, 16, and 20, respectively. The true regression function is E{Y} = 20 + 4X, and the ϵ_i are independent N(O, 25). (40 points, 10 points each)

a-) Generate five normal random numbers (use set.seed(1023)), with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at X = 4,8, 12, 16, and 20 and calculate Y_1, Y_2, Y_3, Y_4 , and Y_5. Obtain the least squares estimates b_0 and b_1, when fitting a straight line to the five cases. Also calculate Y_h when $X_h = 10$ and obtain a 95 percent confidence interval for E(Y_h) when $X_h = 10$.

##Solution: b_0 is 22.38 and b_1 is 3.87. The predicted value is 61.09 for $X_h = 10$. The 95 percent confidence interval is $51.74121 \leq \hat{Y_h} \leq 70.44508$.

```
x=c(4,8, 12, 16,20)
#r function is rnorm(n, mean = 0, sd = 1)#
set.seed(1023)
ei<-rnorm(5,0,5)
yi<-20 + 4*x+ei
f<-lm(yi~x)
summary(f)
```

```
##
## Call:
## lm(formula = yi ~ x)
##
## Residuals:
##      1      2      3      4      5
## -5.251  6.857 -2.020  4.472 -4.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.3852     6.4975   3.445  0.04108 *
## x             3.8708     0.4898   7.903  0.00422 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.195 on 3 degrees of freedom
## Multiple R-squared:  0.9542, Adjusted R-squared:  0.9389
## F-statistic: 62.46 on 1 and 3 DF,  p-value: 0.004222
```

```
predict(f,data.frame(x=10),interval =  "confidence")
```

```
##        fit      lwr      upr
## 1 61.09314 51.74121 70.44508
```

b-) Repeat part (a) 200 times, generating new random numbers each time.

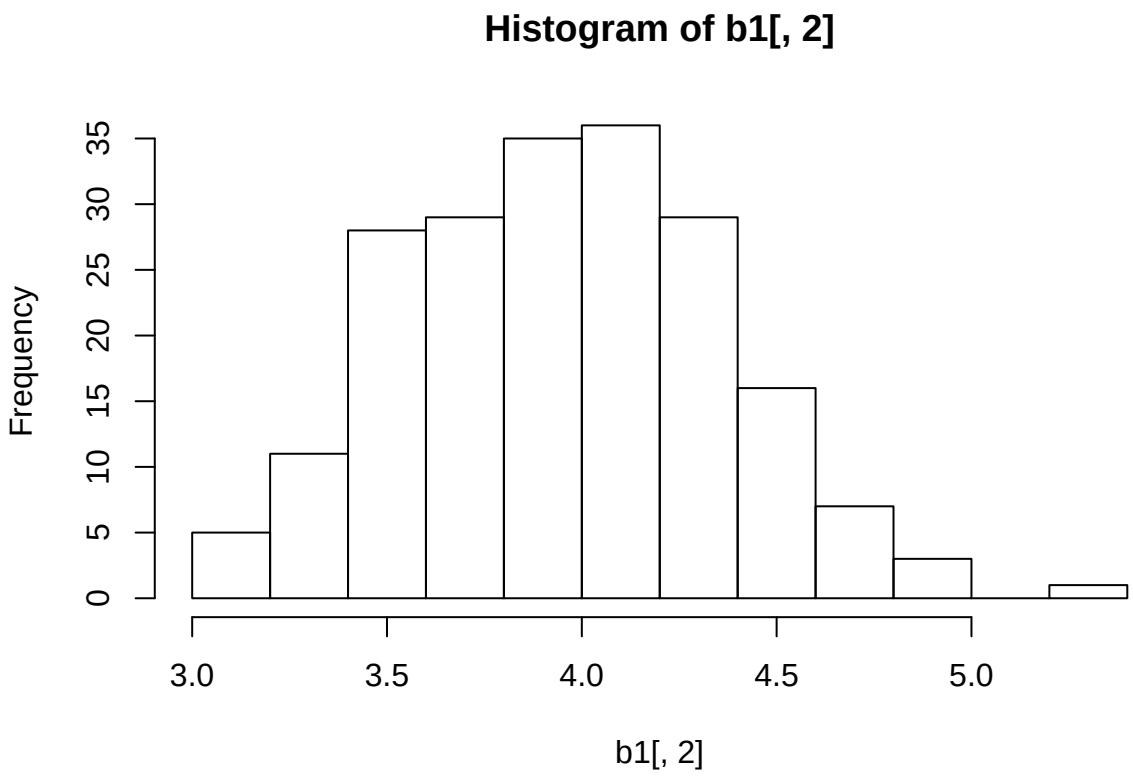
##Solution: see below

```
prg.hw2<-function(x,y,n){
  out<-matrix(0,nrow=n,ncol=5)
   for (i in 1:n){
     ei<-rnorm(5,0,5)
     yi<-20 + 4*x+ei
     f<-lm(yi~x)
     out[i,1:2]=f$coefficients
     out[i,3:5]=predict(f,data.frame(x=10),interval =  "confidence")
   }
  dimnames(out)[[2]]<-c("b0","b1","yhat","Lb","Ub")
  out
}
b1<-prg.hw2(x,y,200)
```

c-) Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations?

##Solution: The mean b_1 is 3.954 and standard deviation is 0.4. The results are in line with the theoretical expectations.

```
hist(b1[,2])
```




```
apply(b1,2,mean)
```

```
##        b0        b1      yhat        Lb        Ub
## 20.621705  3.953793 60.159630 53.690205 66.629055
```

```
sqrt(apply(b1,2,var))
```

```
##        b0        b1      yhat        Lb        Ub
## 5.1022883 0.4023784 2.4160362 3.7508009 3.5625701
```

d-) What proportion of the 200 confidence intervals for $E(Y_h)$ when $X_h = 10$ include $E(Y_h)$? Is this result consistent with theoretical expectations?

##Solutions: it is 92% of the confidence interval contain 60. It is not consistent wth theoretical expectations.

```
sum(I(b1[,4]<=60)*I(b1[,5]>=60))/200
```

```
## [1] 0.92
```

Problem 2

Refer to the CDI data set (used in homework 1). The number of active physicians in a CDI (Y) is expected to be related to total population, number of hospital beds, and total personal income. (30 points)Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians? (20 Point)

##Solution:

R^2 are for total population, number of hospital beds, and total personal income are 0.8840674, 0.9033826, 0.8989137. Number of hospital beds gives us the highest R^2.

```
CDI.Data <- read.csv("/cloud/project/Fall 2020/CDI Data.csv")
f1<-summary(lm(Number.of.active.physicians~Total.population,data=CDI.Data))
```

```
f2<-summary(lm(Number.of.active.physicians~Number.of.hospital.beds,data=CDI.Data))
f3<-summary(lm(Number.of.active.physicians~Total.personal.income,data=CDI.Data))
cbind(f1$r.squared,f2$r.squared,f3$r.squared)
```

```
##           [,1]      [,2]      [,3]
## [1,] 0.8840674 0.9033826 0.8989137
```

Problem 3

Refer to the CDI data set (use in previous homework). For each geographic region, regress per capita income in a CDI(Y) against the percentage of individuals in a county having at least a bachelor's degree (X). Obtain a separate interval estimate of β_1, for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes? (20 points)

##Solutions

For Region1: $448.5001 \leq b1 \leq 595.8176$ For Region2: $184.6841 \leq b1 \leq 292.6547$ For Region3: $277.0031 \leq b1 \leq 384.2203$ For Region4: $349.9378 \leq b1 \leq 530.6936$ The confidence intervals for region 2 and 3 are overlapping. The confidence intervals for region 1 and 4 are overlapping. The slope for regions 2 and 3 are different from region 1 and 4.

```
r1<-lm(Per.capita.income~Percent.bachelor.s.degrees,data=CDI.Data[CDI.Data$Geographic.region==1,])
r2<-lm(Per.capita.income~Percent.bachelor.s.degrees,data=CDI.Data[CDI.Data$Geographic.region==2,])
r3<-lm(Per.capita.income~Percent.bachelor.s.degrees,data=CDI.Data[CDI.Data$Geographic.region==3,])
r4<-lm(Per.capita.income~Percent.bachelor.s.degrees,data=CDI.Data[CDI.Data$Geographic.region==4,])

confint(r1)
```

```
##                                2.5 %     97.5 %
## (Intercept)                7534.1318 10913.4995
## Percent.bachelor.s.degrees  448.5001   595.8176
```

```
confint(r2)
```

```
##                                 2.5 %     97.5 %
## (Intercept)                12441.1260 14721.6844
## Percent.bachelor.s.degrees   184.6841   292.6547
```

```
confint(r3)
```

```
##                                2.5 %     97.5 %
## (Intercept)                9319.5741 11739.9961
## Percent.bachelor.s.degrees  277.0031   384.2203
```

```
confint(r4)
```

```
##                                2.5 %     97.5 %
## (Intercept)                6518.9683 10711.1370
## Percent.bachelor.s.degrees  349.9378   530.6936
```

Problem 4

In a small-scale regression study, five observatiol)s on Y were obtained corresponding to X = 1, 4,10, ll, and 14. Assume that $\sigma = .6$, $\beta_0 = 5$, and $\beta_1, = 3$. (20 points, 10 points each)

a-) What are the expected values MSR and MSE? ##Solutions: E(MSE)=0.36 and E(MSR)=1026.364

```
mse=0.6^2
x=c(1,4,10,11,14)
ssx=4*var(x)
msr=mse+9*4*var(x)
cbind(mse,msr)
```

```
##       mse     msr
## [1,] 0.36 1026.36
```

b-) For determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at X = 6,7, 8, 9, and 1O? Why? Would the same answer apply if the principal purpose were to estimate the mean response for X = 8? Discuss.

##Solutions: E(MSE)=0.36 and E(MSR)=90.36. No it is not better as E(MSR) is smaller.

```
mse=0.6^2
x=seq(6:10)
ssx=4*var(x)
msr=mse+9*4*var(x)
cbind(mse,msr)
```

```
##       mse   msr
## [1,] 0.36 90.36
```