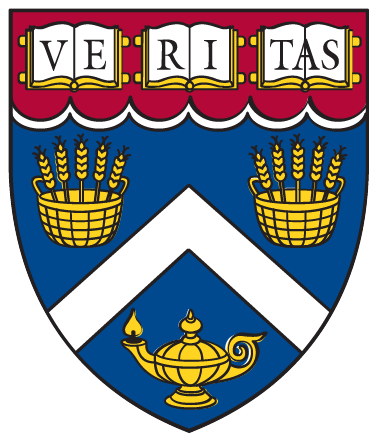
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 2

Due: Monday, 09/30/19 at 7pm 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 generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. The regression model we would like to study is:

and

a-) Write down the likelihood function (5pts)



b-) Find the MLE for and (10pts)



1. Refer to the Grade Point Average (GPA) date set attached below.
2. Obtain the least squares estimates of β0 and β1, and state the estimated regression function. (5pts)

1.48 ≤ β0 ≤ 2.75 and 0.013 ≤ β1 ≤ 0.064

f<-lm(GPA~ACT,data=GPA)

confint(f)

2.5 % 97.5 %

(Intercept) 1.47859015 2.74950842

ACT 0.01353307 0.06412118

1. Obtain a 99 percent confidence interval for β1.Interpret your confidence interval. (5pts)

0.005 ≤ β1 ≤ 0.07, it is wider than that of part a. The confidence interval does not contain 0. Hence, β1 is significant.

confint(f,level=0.99)

0.5 % 99.5 %

(Intercept) 1.273902675 2.95419590

ACT 0.005385614 0.07226864

1. Test, using the test statistic t\*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). (5pts)

Ho: β1 = 0

Ha: β1 ≠ 0

From the table below,

t = 0.03883 / 0.01277 = 3.040

Pvalue= 0.002 is less than 0.05. Reject Ho, ACT score is significant and the linear association exists.

summary(f)

Call:

lm(formula = GPA ~ ACT, data = GPA)

Residuals:

Min 1Q Median 3Q Max

-2.74004 -0.33827 0.04062 0.44064 1.22737

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*

ACT 0.03883 0.01277 3.040 0.00292 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6231 on 118 degrees of freedom

Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476

F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

1. Refer to the Grade Point Average (GPA) date set attached below.
2. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval. (5pts)

Yhat = 3.2

3.06 ≤ Yhat ≤ 3.34

Please see below for the Rcode and results:

f<-lm(GPA~ACT,data=GPA)

X<-data.frame(ACT=28)

predict(f,X,interval = "confidence")

fit lwr upr

3.201209 3.061384 3.341033

1. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA-using a %95 prediction interval. Interpret your prediction interval. (5pts)

3.02 ≤ Yhat ≤ 4.43

Please see below for the Rcode and results:

predict(f,X,interval = "prediction")

fit lwr upr

3.201209 1.959355 4.443063

1. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be? (5pts)

Yes it is wider and Yes it should be wider.

1. Calculate %95 percent confidence band for the regression line when Xh = 28. Is your-confidence band wider at this point than the confidence interval in

part (a)? Should it be? (5pts)

3.03 ≤ β0+β1X≤ 3.38

Please see below for the Rcode and results:

W <- sqrt( 2 \* qf(0.95,2,118))

CI<-predict(f,X,se.fit=TRUE,interval="confidence")

cbind(CI$fit[,1]-W\*CI$se.fit, CI$fit[,1] + W\*CI$se.fit )

[,1] [,2]

3.026159 3.376258

1. Repeat question 3, by building the models on the development sample (a random sample of 70% of GPA data), and calculating MSE’s on the hold out sample (remainder 30% of the GPA data).

a-) Yhat = 3.19

3.02≤ Yhat ≤ 3.36

b-) 1.92≤ Yhat ≤ 4.46

c-) yes and yes

d-) 2.98 ≤ β0+β1X≤ 3.4

e-) MSE =0.38,

set.seed(12345)

ind <- sample(1:nrow(GPA), size =nrow(GPA)\*0.70)

dev <- GPA[ind,]

holdout <- GPA[-ind,]

f11<-lm(GPA~ACT,data=dev)

predict(f11,X,interval = "confidence")

fit lwr upr

3.19435 3.023149 3.365551

predict(f11,X,interval = "prediction")

fit lwr upr

3.19435 1.922307 4.466392

W <- sqrt( 2 \* qf(0.95,2,82)

CI<-predict(f11,X,se.fit=TRUE,interval="confidence")

cbind(CI$fit[,1]-W\*CI$se.fit, CI$fit[,1] + W\*CI$se.fit )

[,1] [,2]

[1,] 2.980994 3.407706

pred<-predict(f11,holdout)

dim(holdout)

[1] 36 2

MSE<-sum((holdout[,1]-pred)^2)/34

MSE

[1] 0.380144

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 εi are independent N(0, 25).
2. Generate five normal random numbers, 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 Y1, Y2, Y3, Y4 , and Y5. Obtain the least squares estimates β0 and β1, when fitting a straight line to the five cases. Also calculate when Xh = 10 and obtain a %95 confidence interval for

E{Yh} when Xh = 10. (10 pts)

25.534 + 3.696 X and prediction for X=10 is 62.49

set.seed(1073)

e=rnorm(5, mean = 0, sd = 5)

x=c(4,8, 12, 16, 20)

y=20+(4\*x)+e

f5<-lm(y~x)

f5

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

19.73 3.83

X5=data.frame(x=10)

predict(f5,X5, interval = "confidence")

fit lwr upr

58.03327 44.795 71.27154

1. Repeat part (a) 200 times, generating new random numbers each time. (15 pts)

e1=matrix(rnorm(5\*200, mean = 0, sd = 5),nrow=5,ncol=200)

y1<- 20+(4\*x)+e1

prg<-function(x,e,y,xp){

out<-matrix(0,nrow=200,ncol=5)

for (i in 1:200){

f<-lm(y[,i]~x)

out[i,1:2]<- cbind(f$coefficients[1],f$coefficients[2])

out[i,3:5]<- predict(f,xp, interval = "confidence")

}

dimnames(out)[[2]]<-c("bo","b1","fit","lb","ub")

out

}

out<-prg(x,e1,y1,X5)

1. Make a frequency distribution of the 200 estimates β1. Calculate the mean and standard deviation of the 200 estimates β1. Are the results consistent with theoretical expectations? (10 pts)

mean(out[,2])

[1] 4.013783

sqrt(var(out[,2]))

[1] 0.4140188

1. What proportion of the 200 confidence intervals for E{Yh} when Xh = 10 include E{Yh}? Is this result consistent with theoretical expectations? (10 pts)

dim(out[(out[,5]>=58.03327 ) & (out[,4]<=58.03327 ),])[1]/200

[1] 0.92

Expected proportion is .95. However, we can 92% due the number of simulations. As the number of simulations increase, we should get 95% coverage.