**Question 1.**

(a)

> gpa=read.table("a1GPA.txt",header=T)

> gpa

y x

1 3.897 21

2 3.885 14

3 3.778 28

4 2.540 22

5 3.028 21

6 3.865 31

. . .

117 3.800 29

118 3.914 28

119 1.860 16

120 2.948 28

> n=length(gpa$y)

> n

[1] 120

> xbar=mean(gpa$x)

> xbar

[1] 24.725

> ybar=mean(gpa$y)

> ybar

[1] 3.07405

> ssx=sum(gpa$x^2)

> ssx

[1] 75739

> ssx = 0

> for(i in 1:n){

+ ssx = ssx + x[i]^2

+ }

> ssx

[1] 75739

> spxy = sum(gpa$x \*gpa$ y)

> spxy

[1] 9213.112

> spxy = 0

> for(i in 1:n){

+ spxy = spxy + gpa$x[i] \* gpa$y[i]

+ }

> spxy

[1] 9213.112

> b1 = (spxy - n \* xbar \* ybar)/(ssx - n \* xbar^2)

> b1

[1] 0.03882713

> b0 = ybar - b1 \* xbar

> b0

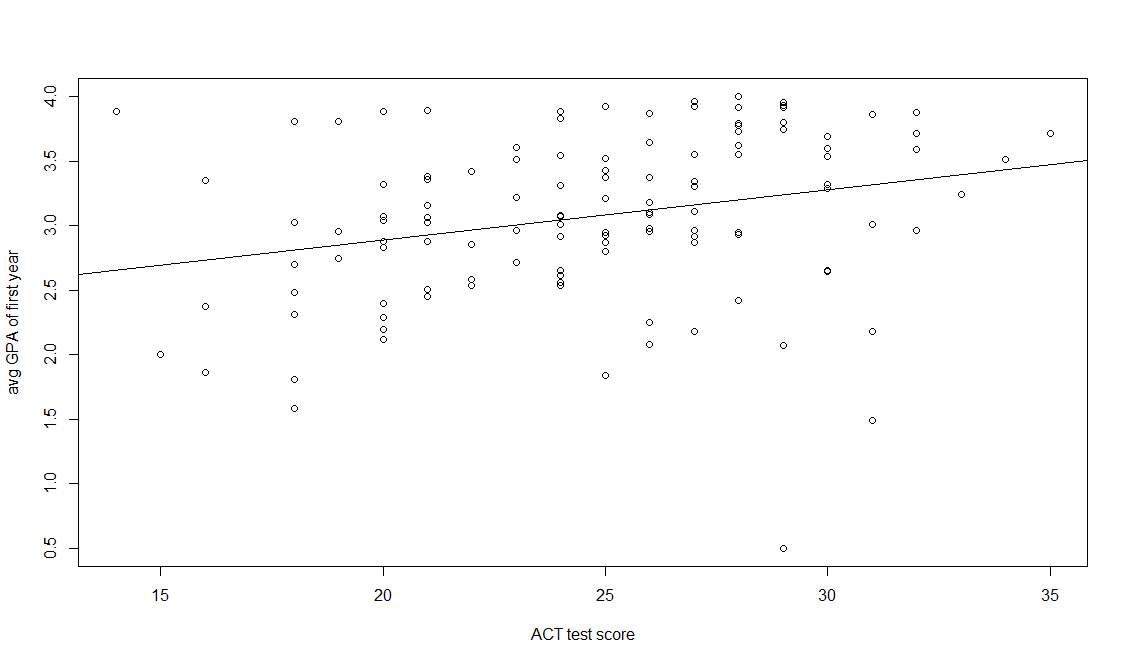
[1] 2.114049

Therefore, the estimated regression function is :

(b)

> plot(gpa$x,gpa$y,xlab="ACT test score",ylab="avg GPA of first year")

> abline(a=b0,b=b1)



Yes, the estimated regression function fits the date well.

(c)

when x=30, the point estimate of the mean freshman GPA is:

> yhat\_30=b0+b1\*30

> yhat\_30

[1] 3.278863

i.e,.

(d)

From the estimated regression function, we can see that the entrance test score increases by one point, the mean freshman GPA for students will increase 0.03882713 point.

**Question2.**

(a)

> yhat=b0+b1\*gpa$x

> e=gpa$y-yhat

> e

[1] 0.96758105 1.22737094 0.57679116 -0.42824608 0.09858105 0.54730978 -0.39451735

. . .

[113] -0.49659183 -0.60459183 -1.83169022 0.99440817 0.55996403 0.71279116 -0.87528332

[120] -0.25320884

> sse=sum(e)

> sse

[1] -3.996803e-15

Therefore, the sum of residuals e is zero in accord with



(b)

> SSE=sum(e^2)

> SSE

[1] 45.81761

> MSE=SSE/(n-2)

> MSE

[1] 0.3882848

> s=sqrt(MSE)

> s

[1] 0.623125

Hence, =MSE=0.3882848 and is the positive square root of MSE=0.623125.



is expressed in points as test score.



**Question3.**

(a)

(b)

(c)

**Question4.**

**Question5.**

> s2b1=MSE/(ssx-n\*xbar^2)

> s2b1

[1] 0.00016315

> sb1=sqrt(s2b1)

> sb1

[1] 0.01277302

> ttab=qt((1-0.01/2),n-2)

> ttab

[1] 2.618137

> c(b1-ttab\*sb1,b1+ttab\*sb1)

[1] 0.005385614 0.072268640

Therefore, the confidence interval of is: (0.005385614, 0.072268640)



Interpretation: If we draw many many samples of size n=120, there is 99% probability that would fall between 0.005385614 and 0.072268640.



The confidence interval does not include zero; The director of admissions can know if the regression model is significant (i.e. whether the ACT test score has significant explanatory power for the mean freshman GPA) through checking whether zero is contained in the confidence interval; In this question, the 99% confidence interval does not contain zero which implies that the regression model is significant (i.e. we reject the null hypothesis: =0), and we can conclude that the administrator can can strongly predict the mean freshman GPA by ACT test score.



**Question6.**

(a)

> MSE

[1] 0.3882848

> yhhat\_28=b0+b1\*28

> yhhat\_28

[1] 3.201209

> s2yhhat=MSE\*((1/n)+((28-xbar)^2/(ssx-n\*xbar^2)))

> s2yhhat

[1] 0.004985593

> syhhat=sqrt(s2yhhat)

> syhhat

[1] 0.07060873

i.e.,

> tsta=qt((1-0.05/2), n-2)

> tsta

[1] 1.980272

> c(yhhat\_28-tsta\*syhhat, yhhat\_28+tsta\*syhhat)

[1] 3.061384 3.341033

Hence, the 95% interval estimate is:

Interpretation: There is 95% probability that the mean freshman GPA is somewhere between 3.061384 and 3.341033 when ACT test score=28.

(b)

> s2sprd=MSE+s2yhhat

> s2sprd

[1] 0.3932704

> ssprd=sqrt(s2sprd)

> ssprd

[1] 0.6271128

i.e.,

> c(yhhat\_28-tsta\*ssprd, yhhat\_28+tsta\*ssprd)

[1] 1.959355 4.443063

Hence, the 95% prediction interval is:

Interpretation: We predict that there is 95% probability that Mary Jones's mean freshman GPA will be somewhere between 1.959355 and 4.443063 when she obtained a score of 28 on the entrance test.

(c)

Yes, the prediction interval in part (b) is wider than the confidence interval in part (a).

Yes, it should be. Because when predicting the mean freshman GPA, we encounter both variability in from sample to sample and the variation among ACT test scores within the probability distribution of Y as well.



**Question 7.**

(a)

> reg\_fit=lm(gpa$y~gpa$x)

> anova(reg\_fit)

Analysis of Variance Table

Response: gpa$y

Df Sum Sq Mean Sq F value Pr(>F)

gpa$x 1 3.588 3.5878 9.2402 0.002917 \*\*

Residuals 118 45.818 0.3883

---

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

(b)

MSR estimates and is estimated by MSE (i.e., ).



If there is no linear association (i.e.,=0 ), we would expect the ratio MSR/MSE=1, namely, MSR and MSE estimate the same quantity when =0.



(c)

H0: = 0



Ha: ≠0



since

> FTAB <- qf(0.99, 1, 118)

> FTAB

[1] 6.854641

i.e., f(0.99, 1 , 118) =6.854641

Decision rule: When=0.01, reject the null hypothesis when F > 6.854641



accept the null hypothesis when F < 6.854641

Conclusion: since our F = 9.2402 which is greater than 6.854641, we reject the null hypothesis at 0.01 level of significance, and we conclude that the regression model is significant.

(d)

The absolute magnitude of the reduction in variation of Y is SSR which equals to 3.588;

The relative reduction is SSR/SST = SSR/(SSR+SSE) = 3.588/(3.588 + 45.818) = 0.07262, which is called "the coefficient of determination" ().



(e)

r is the square root of , i.e,. sqrt(0.0726) = 0.26948, which is positive because the sign of b1 is positive.



(f)

Since shows the amount of variability in response variable that is explained by the fitted linear regression model, while r shows the strength of linear association between two variable. Therefore I would say that the has the more clear-cut operational interpretation.

