MATH 564 - Assignment2

————————Mohammed Wasim R D(A20497053)——————

Problem1

Reading data and renaming columns

```
data<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR27.txt")
colnames(data)[1] ="Muscle"
colnames(data)[2]="Age"

#Linear regression model
mm_model = lm(Muscle ~ Age, data=data)
summary(mm_model)
```

```
##
## Call:
## lm(formula = Muscle ~ Age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## Age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

---

a) Hypothesis Testing : H0 : β1 = 0, H1 :β1 <0,α=0.05 as from the above T*-value=-13.19 and p−value=2∗10^−16 are for two-sided P(t< -13.19) for a t-distribution with degree of freedom = 58 which is still 0 and less than 0.05 so, we can reject the null hypothesis H0 There is sufficient evidence that there is negative linear association between amount of muscle mass and age

---

b) No as test of beta_0 not equal to zero is significant and wont provide related information on amount of muscle mass.

```
confint(mm_model)
```

```
##                  2.5 %     97.5 %
## (Intercept) 145.312572 167.380556
## Age          -1.370545  -1.009446
```

c) From the above solution we can see that confidence interval of beta_1 is (-1.37,-1.009). Here it is not necessary to know the specific age and this doesnt change as x changes and we assume slope remains same throughout the range of x

---

Problem 2

Reading the table with assigning column names

```
df<- read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR19.txt", head
er = FALSE, sep = " ")
GPA <- df[,2]
ACT <- df[,6]
lm <- lm(GPA~ACT)
lm
```

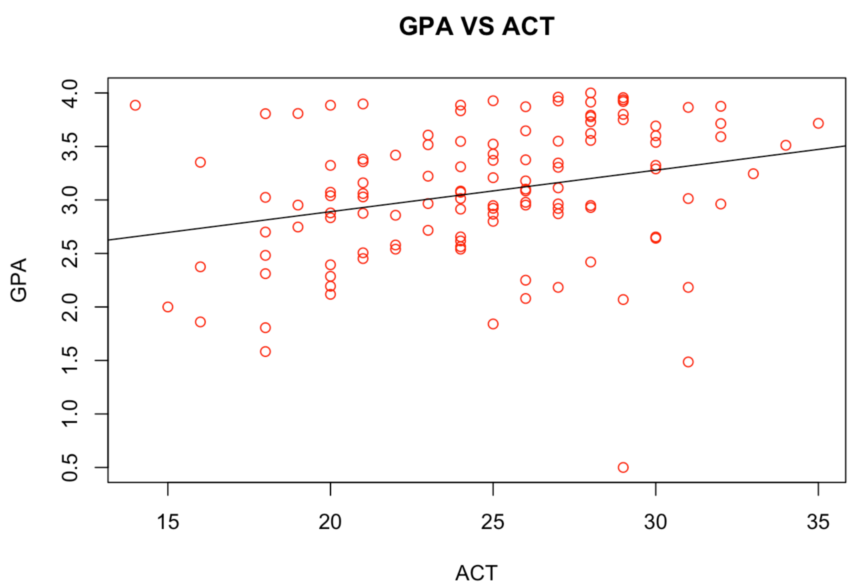
```
##
## Call:
## lm(formula = GPA ~ ACT)
##
## Coefficients:
## (Intercept)          ACT
##     2.11405      0.03883
```

a) The estimated Regression function is beta_0=2.11405 and beta_1=0.03883

Yi=2.11405+1.0.03883Xi+e

Plotting the estimated regression function

```
plot(GPA~ACT, main = "GPA VS ACT", xlab = "ACT", ylab = "GPA",
    col = "red")
abline(lm, col = "black")
```



b)According to the above plot we can see that data is too spread and might create problems as there is lot of variance in the data

---

c)Obtain a point estimate of the mean freshman GPA for students with ACT test score X=30

```
Y = lm$coefficients[[2]] * 30 + lm$coefficients[[1]]
Y
```

```
## [1] 3.278863
```

---

d)What is the point estimate of the change in the mean response when the entrance test score increases by one point?

Here beta_1 represents the slope of estimated regression line and therefore it indicates the change in the mean response when X increases by one measurement which is beta_1=0.03883

Problem 3

a)Obtain a 95% interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval

```
freshman.gpa <- data.frame(ACT=28)
gpa.confidence <- predict(lm, freshman.gpa, interval = "confidence", level = 0.95, se.fit = TRUE)
gpa.confidence
```

```
## $fit
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
##
## $se.fit
## [1] 0.07060873
##
## $df
## [1] 118
##
## $residual.scale
## [1] 0.623125
```

From the above results we can see that the students with ACT score 28 with confidence interval 95% will have GPA between 3.061384 and 3.341033

---

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

```
gpa.prediction <- predict(lm, freshman.gpa, interval = "prediction", level = 0.95, se.fit = TRUE)
gpa.prediction
```

```
## $fit
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
##
## $se.fit
## [1] 0.07060873
##
## $df
## [1] 118
##
## $residual.scale
## [1] 0.623125
```

From the above results we can see that Mary Jones with score of 28 using 95% prediction interval will have GPA between 1.95 and 4.44

---

c)Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

Ans: Yes the prediction interval is much wider than the confidence interval due to conceptual difference from the confidence interval. Here the prediction interval describes value for random variable and therefore it should have wider interval to allow for non parameterized variables to impact the predicted value

---

d)Determine the boundary values of the 95% 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?

```
W <- sqrt(2*qf(0.95,2,length(GPA)-2))
conf.band.upper <- gpa.confidence$fit[,1]+W*gpa.confidence$se.fit
conf.band.lower <- gpa.confidence$fit[,1]-W*gpa.confidence$se.fit
conf.band.upper
```

```
## [1] 3.376258
```

```
conf.band.lower
```

```
## [1] 3.026159
```

From the above results the confidence band for Xh=28 is 3.026159 <= beta_0+beta_1Xh<=3.376258.It is little wider than the confidence interval at Xh=28 as it is representing the confidence intervals for entire regression line

---

Probelm 4

a)Set up the ANOVA table

```
analysis<-anova(lm)
analysis
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value   Pr(>F)
## ACT         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)What is estimated by MSR in your ANOVA table?by MSE?Under what conditions do MSR and MSE estimate the same quantity?

Ans: Here MSR is the Sum of Squares due to regression by degree of freedom in the model and MSE is Mean Square Error due to error.When beta_1=0 MSR and MSE estimate the same quantity

c)Conduct an F test of whether or not β1=0 Control the alpha risk at 0.01. State the alternatives, decision rule, and conclusion

```
alpha <- 0.01
n <- length(GPA)
F.test.gpa <- qf((1-alpha),1,n-2)
F.test.gpa
```

```
## [1] 6.854641
```

F value from anova is 9.2402 and F value from ftest is 6.854 where we can say that it rejects null hypothesis and accepts alternative hypothesis when β1=0

d)What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model?What is the name of the latter meas

```
r summary(lm)
```

```
## ## Call: ## lm(formula = GPA ~ ACT) ## ## Residuals: ##       Min        1Q   Median        3Q       Max ## -2.74004 -0.33827  0.040
```

```
r r <- sqrt(0.07262) r
```

```
## [1] 0.269481
```

The relative reduction in the variation of Y when X is introduced into the regression model is the R^2 value 0.07262 also the latter measure is r^2 value

e)Obtain r and attach the appropriate sign

```
r <- sqrt(0.07262)
r
```

```
## [1] 0.269481
```

r value for linear model of gpa is 0.269481.The sign is positive as there is positive correlation between two sets of data

f)Which measure, R2 or r, has the more clear-cut operational interpretation? Explain

Ans: R^2 is more clear cut operational interpretation because it is the value between 0 and 1 which describes the percent od variable y explained by x which is more frequently used to describe the relationship of variables

PROBLEM 5 :

$$F_{a-1,N-a} = \frac{MST}{MSE} = \frac{\frac{SST}{a-1}}{\frac{SSE}{N-a}} \rightarrow (1)$$

$$t_K^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_p^2(1/n_1 + 1/n_2)} \rightarrow (2)$$

Denominator of equation (1) when $a=2$

$$MSE = \frac{SSE}{N-2} = \frac{\sum_{j=1}^{n}(y_{1j}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y})^2}{N-2} \rightarrow (3)$$

Formula for sample variance estimator is

$$S_i^2 = \sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2 / n_i - 1$$

Multiply and divide terms in numerator in eq^n (3) by $(n_i - 1)$ and get eq^n (4)

$$\frac{SSE}{N-2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = S_p^2 \rightarrow (4)$$

Numerator of equation (1) when $a = 2$.

$$\frac{SST}{2-1} = SST.$$

$$SST = \sum_{1}^{2} n_i(\bar{y}_i - \bar{y})^2 = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 \longrightarrow (5)$$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{N} \longrightarrow (6)$$

Replace (6) in (5)

$$SST = n_1\left[\bar{y}_1 - \left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{N}\right)\right]^2 + n_2\left[\bar{y}_2 - \left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{N}\right)\right]^2 \longrightarrow (7)$$

Substituting (7) equations seperately we get.

$$SST = \frac{n_1 n_2^2}{N^2}(\bar{y}_1 - \bar{y}_2)^2 + \frac{n_2 n_1^2}{N^2}(\bar{y}_2 - \bar{y}_1)^2 \longrightarrow (8)$$

$$SST = \left[\frac{n_1 n_2^2}{N^2} + \frac{n_2 n_1^2}{N^2}\right](\bar{y}_1 - \bar{y}_2)^2 \longrightarrow (9)$$

$$= \frac{n_1 n_2^2}{N^2} + \frac{n_2 n_1^2}{N^2}$$

N^2 is common denominator.

$$\frac{n_1 n_2(n_1 + n_2)}{N^2}$$

Replace $N = n_1 + n_2$

$$= \frac{n_1 n_2 N}{N^2}$$

$$= \frac{n_1 n_2}{N} = \frac{1}{\frac{N}{n_1 n_2}}$$

Replace $N = n_1 + n_2$

$$\frac{1}{\frac{n_1 + n_2}{n_1 n_2}} = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Replace above eq^n in (a).

$$SST = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}} (\bar{y}_1 - \bar{y}_2)^2$$

With above steps and $a = 2$ we have

$$\frac{SST}{2-1} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}.$$

$$\frac{SSE}{N-2} = S_p^2$$

Ratio of these expressions namely F statistic is,

$$F_{1,K} = \frac{\frac{SST}{2-1}}{\frac{SSE}{N-2}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = t_K^2$$

Therefore $t^2 = F$ //