Lab 6

Yunting Chiu

2021-02-12

R Lab 6

Exercise 1

1.1

- X is V2, which means the ACT test score.
- Y is V1, which means the GPA at the end of the freshman year.

```
GPA = read.table("./data/CH01PR19.txt")
attach(GPA)
X <- V2
Y <- V1
# X = ACT, Y = GPA
```

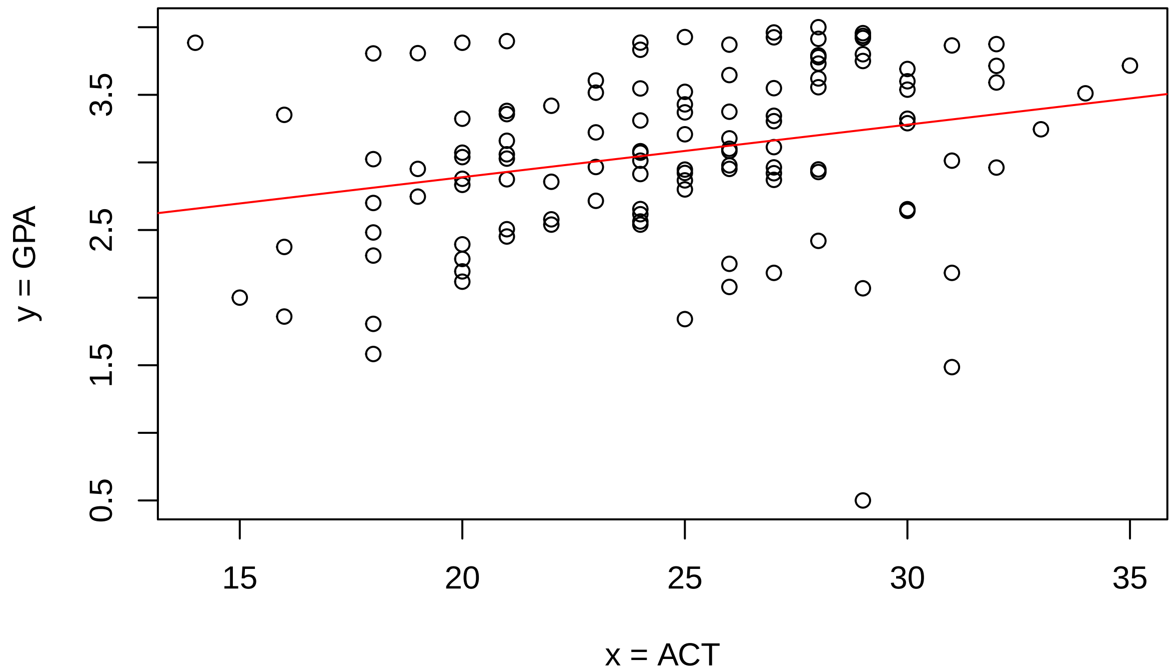
1.2

According to the plot below, we can see that many data points do not fit the linear regression line.

```
reg = lm(Y ~ X)
reg
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##     2.11405      0.03883
```

```
plot(X, Y, xlab = "x = ACT", ylab = "y = GPA")
abline(reg, col = "red")
```




```r
# b0 =  2.11405
# b1 = 0.03883
```

1.3

We use `predict` function to estimate it. If ACT score is 30, then the estimated mean of the freshman GPA would be 3.278863.

```r
predict(reg, data.frame(X = 30))
```

```
##        1
## 3.278863
```

1.4

The key is we should find b1. If the ACT score increases by one point, we estimate GPA would be changed 0.03882713 units.

```r
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## 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 ***
## X            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
```

```r
reg$coefficients[2]
```

```
##          X
## 0.03882713
```

1.5

e is the residuals. Another way is we can use `anova` table to find out the residuals.

```r
e = Y - fitted.values(reg)
e # ei each
```

```
##           1           2           3           4           5           6
##  0.96758105  1.22737094  0.57679116 -0.42824608  0.09858105  0.54730978
##           7           8           9          10          11          12
## -0.39451735  0.79861829 -2.74003597  0.05444541  0.26409967  0.25913691
##          13          14          15          16          17          18
##  0.03709967 -0.03290033 -0.15034448 -0.19938171  0.43727254 -0.30469022
##          19          20          21          22          23          24
## -0.13772746 -0.77259183 -0.48290033  0.42758105  0.52979116  0.76261829
##          25          26          27          28          29          30
##  0.35479116 -0.02255459 -0.78120884 -0.38924608  0.74744541  0.13058105
##          31          32          33          34          35          36
##  0.84227254 -0.36028332 -0.27220884  0.25144541 -0.11124608  0.02609967
##          37          38          39          40          41          42
##  0.45158105  0.01113691  0.38661829  0.52244541 -0.14555459 -0.62486309
##          43          44          45          46          47          48
## -0.50590033 -0.87355459 -1.17103597 -0.42890033 -1.13469022 -0.69645619
##          49          50          51          52          53          54
##  0.10023530  0.99306243 -0.29138171  0.61671668  0.14261829 -0.17155459
##          55          56          57          58          59          60
##  0.50109967  0.41213691  0.23058105 -0.69659183  0.04413691  0.69596403
##          61          62          63          64          65          66
## -0.16272746 -0.29107321  0.28527254  0.59892679 -0.63686309 -0.47741895
##          67          68          69          70          71          72
## -0.39090033  0.35748265 -1.00693757  0.50892679  0.14840817 -0.04107321
##          73          74          75          76          77          78
## -0.33093757 -0.11293757  0.67996403 -0.05659183  0.21492679 -0.03955459
##          79          80          81          82          83          84
##  0.79879116  0.07682840  0.43240817  0.18140817 -1.04455459  0.51848265
##          85          86          87          88          89          90
##  0.12327254 -0.24238171  0.18261829  0.71596403  0.95623530 -0.42341895
##          91          92          93          94          95          96
##  0.84009967 -0.97938171  0.34427254  0.21106243  0.50996403  0.78709967
##          97          98          99         100         101         102
## -0.04938171 -0.05441895 -0.10476470 -0.50193757 -1.24372746 -1.22993757
##         103         104         105         106         107         108
## -0.01159183  0.23448265 -0.13190033  0.24300127 -0.28472746  0.41979116
##         109         110         111         112         113         114
##  0.59079116 -0.21772746  0.45075392  0.32113691 -0.49659183 -0.60459183
##         115         116         117         118         119         120
```

```
## -1.83169022  0.99440817  0.55996403  0.71279116 -0.87528332 -0.25320884
```

```r
sum(e^2) # sum of the squared residuals
```

```
## [1] 45.81761
```

```r
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value   Pr(>F)
## 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
```

1.6

estimated variance is 0.3882848, and residual standard error: 0.6231 on 118 degrees of freedom

```r
n <- length(X)
var_est <- sum(e^2)/(n-2) # sum of the squared residuals - degrees of freedom (n -2, b0 and b1 so minus
var_est
```

```
## [1] 0.3882848
```

```r
sqrt(var_est)
```

```
## [1] 0.623125
```

```r
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## 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 ***
## X            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
```

Exercise 2

2.1

Zero is not included, because the b1 is 0.005385614 to 0.07226864.

```r
confint(reg, level = 0.99)
```

```
##                   0.5 %     99.5 %
## (Intercept) 1.273902675 2.95419590
## X           0.005385614 0.07226864
```

2.2

- Ho: $\beta 1 = 0$ (have NOT a linear association between ACT score (X) and GPA at the end of the freshman year (Y)).
- Ha: $\beta 1 \neq 0$ (have a linear association between ACT score (X) and GPA at the end of the freshman year (Y)).

According to tables of a linear regression model above, we have evidence to reject the null hypothesis in favor of the alternative hypothesis with the 99 % confidence interval. Thus, there is a linear association between student's ACT score (X) and GPA at the end of the freshman year (Y).

2.3

For the actual **population mean** response - confidence intervals

If ACT score is 28, we have 95% certain contains the population mean of freshman GPA is 3.061384 to 3.341033.

```r
predict(reg, data.frame(X = 28), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

2.4

For the individual response (actual response) - prediction intervals

If student obtained a 28 on the ACT, the 95% prediction interval is 1.959355 to 4.443063 of his freshman GPA.

```r
predict(reg, data.frame(X = 28), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

2.5

The majority of data points are in the range of upper band and lower band. We can conclude that the true regression relation has been precisely estimated.

```r
n = length(X) #sample sizes
e = reg$residuals # residuals
s = sqrt(sum(e^2)/(n-2)) # estimated standard deviation = root MSE
s
```

```
## [1] 0.623125
```

```r
W = sqrt(qf(0.95,2,n-2))  # quantity of F-distribution
W
```

```
## [1] 1.753023
```

```r
Yhat = fitted.values(reg) # Yhat = b0 + b1x = predict(reg)
Sxx = (n-1)*var(X)

margin = W*s*sqrt(1/n + (X - mean(X))^2/Sxx)
upper.band = Yhat + W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)
lower.band = Yhat - W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)

plot(X,Y,xlab="ACT", ylab="Y = GPA", xlim = c(20,30))
abline(reg,col="red")
lines(X,upper.band,col="blue")
lines(X,lower.band,col="blue")
```