STAT512 Division 1 HW2

Yi Yang

9/11/2018

1. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one of the team's products: Estimated regression equation: $\hat{Y} = 350.7 - .18X$. Two-sided P-value for estimated slope: 0.91. The student stated: "the message I get here is that the more we spend on advertising this product, the fewer units we sell!" Comment.

Solution:

The null hypothesis is $\beta_1 = 0$, P-value is 0.91, which means we fail to reject the null hypothesis. We do not have enough evidence to assure the linear relationship between X and Y, thus the statement is incorrect.

2. Refer to the problem 5 in the homework 1, use R to generate confidence band for

a) mean value prediction,
b) single value prediction,
c) mean of 3 new values prediction,
d) Working-hotelling confidence band. Comment on their difference.

Solution:

a) Mean value prediction is given as

```
airfreight_df <- read.csv('data/airfreight.csv', header = TRUE)
names(airfreight_df)
```

```
## [1] "X" "Y"
```

```
X <- airfreight_df$X
Y <- airfreight_df$Y
SSXY <- t(X - mean(X)) %*% (Y - mean(Y))
SSX <- t(X - mean(X)) %*% (X - mean(X))
slope <- SSXY / SSX
intercept <- mean(Y) - slope * mean(X)
Y_hat <- slope * X + intercept
Y_residual <- Y - Y_hat
SSE <- sum(Y_residual^2)
N <- length(X)
MSE <- SSE / (N - 2)
Xmean <- mean(X)
std_Y_hat <- rep(0,N)
CI_lbYhat <- rep(0,N)
CI_ubhat <- rep(0,N)
for (i in seq(N)){
  std_Y_hat[i] <- sqrt(MSE * (1 / N + (X[i] - Xmean)^2 / SSX))
  CI_lbYhat[i] <- Y_hat[i] - qt(1 - 0.01/2,N - 2) * std_Y_hat[i]
  CI_ubhat[i] <- Y_hat[i] + qt(1 - 0.01/2,N - 2) * std_Y_hat[i]
}
cat(paste('Mean value prediction at level (X =',X ,') confidence interval is\n
          [', CI_lbYhat,',', CI_ubhat,'].\n'))
```

```
## Mean value prediction at level (X = 1 ) confidence interval is
##
```

```
##           [ 12.6261838380802 , 15.7738161619198 ].
##  Mean value prediction at level (X = 0 ) confidence interval is
##
##           [ 7.97428783913104 , 12.425712160869 ].
##  Mean value prediction at level (X = 2 ) confidence interval is
##
##           [ 15.974287839131 , 20.425712160869 ].
##  Mean value prediction at level (X = 0 ) confidence interval is
##
##           [ 7.97428783913104 , 12.425712160869 ].
##  Mean value prediction at level (X = 3 ) confidence interval is
##
##           [ 18.6808400778595 , 25.7191599221405 ].
##  Mean value prediction at level (X = 1 ) confidence interval is
##
##           [ 12.6261838380802 , 15.7738161619198 ].
##  Mean value prediction at level (X = 0 ) confidence interval is
##
##           [ 7.97428783913104 , 12.425712160869 ].
##  Mean value prediction at level (X = 1 ) confidence interval is
##
##           [ 12.6261838380802 , 15.7738161619198 ].
##  Mean value prediction at level (X = 2 ) confidence interval is
##
##           [ 15.974287839131 , 20.425712160869 ].
##  Mean value prediction at level (X = 0 ) confidence interval is
##
##           [ 7.97428783913104 , 12.425712160869 ].
```

b) Single value prediction is given as

```
std_Y_hnew <- rep(0,N)
CI_lbYhnew <- rep(0,N)
CI_ubYhnew <- rep(0,N)
for (i2 in seq(N)){
  std_Y_hnew[i2] <- sqrt(MSE * (1 / N + (X[i2] - Xmean)^2 / SSX + 1))
  CI_lbYhnew[i2] <- Y_hat[i2] - qt(1 - 0.01 / 2, N - 2) * std_Y_hnew[i2]
  CI_ubYhnew[i2] <- Y_hat[i2] + qt(1 - 0.01 / 2, N - 2) * std_Y_hnew[i2]
}
cat(paste('Single value prediction at level (X =',X ,') confidence interval is\n
          [', CI_lbYhnew,',', CI_ubYhnew,'].\n'))
```

```
## Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 8.98024230191479 , 19.4197576980852 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 4.74814089156371 , 15.6518591084363 ].
##  Single value prediction at level (X = 2 ) confidence interval is
##
##           [ 12.7481408915637 , 23.6518591084363 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 4.74814089156371 , 15.6518591084363 ].
##  Single value prediction at level (X = 3 ) confidence interval is
```

```
##
##           [ 16.1046362148925 , 28.2953637851075 ].
##  Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 8.98024230191479 , 19.4197576980852 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 4.74814089156371 , 15.6518591084363 ].
##  Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 8.98024230191479 , 19.4197576980852 ].
##  Single value prediction at level (X = 2 ) confidence interval is
##
##           [ 12.7481408915637 , 23.6518591084363 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 4.74814089156371 , 15.6518591084363 ].
```

c) Mean of 3 new values prediction is given

```
std_Ybar_hnew <- rep(0,N)
CI_lbYbarhnew <- rep(0,N)
CI_ubYbarhnew <- rep(0,N)
for (i3 in seq(N)){
  std_Ybar_hnew[i3] <- sqrt(MSE * (1 / N + (X[i3] - Xmean)^2 / SSX + 1 / 3))
  CI_lbYbarhnew[i3] <- Y_hat[i3] - qt(1 - 0.01 / 2, N - 2) * std_Ybar_hnew[i3]
  CI_ubYbarhnew[i3] <- Y_hat[i3] + qt(1 - 0.01 / 2, N - 2) * std_Ybar_hnew[i3]
}
cat(paste('Single value prediction at level (X =',X ,') confidence interval is\n
          [', CI_lbYbarhnew,',', CI_ubYbarhnew,'].\n'))
```

```
## Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 10.9238404063213 , 17.4761595936787 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 6.56542726104247 , 13.8345727389575 ].
##  Single value prediction at level (X = 2 ) confidence interval is
##
##           [ 14.5654272610425 , 21.8345727389575 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 6.56542726104247 , 13.8345727389575 ].
##  Single value prediction at level (X = 3 ) confidence interval is
##
##           [ 17.6567840763031 , 26.7432159236969 ].
##  Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 10.9238404063213 , 17.4761595936787 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 6.56542726104247 , 13.8345727389575 ].
##  Single value prediction at level (X = 1 ) confidence interval is
##
##           [ 10.9238404063213 , 17.4761595936787 ].
```

```
##  Single value prediction at level (X = 2 ) confidence interval is
##
##           [ 14.5654272610425 , 21.8345727389575 ].
##  Single value prediction at level (X = 0 ) confidence interval is
##
##           [ 6.56542726104247 , 13.8345727389575 ].
```

d) Working-hoteling confidence band is given as

```
std_Y_hat <- rep(0,N)
CB_whlb <- rep(0,N)
CB_whub <- rep(0,N)
W <- sqrt(2 * qf(1 - 0.01, 2, N - 2))
for (i4 in seq(N)){
  std_Y_hat[i4] <- sqrt(MSE * (1 / N + (X[i4] - Xmean)^2 / SSX))
  CB_whlb[i4] <- Y_hat[i4] - W * std_Y_hat[i4]
  CB_whub[i4] <- Y_hat[i4] + W * std_Y_hat[i4]
}
cat(paste('Working hotelling band at level (X =',X ,') confidence interval is\n
          [', CB_whlb,',', CB_whub,'].\n'))
```

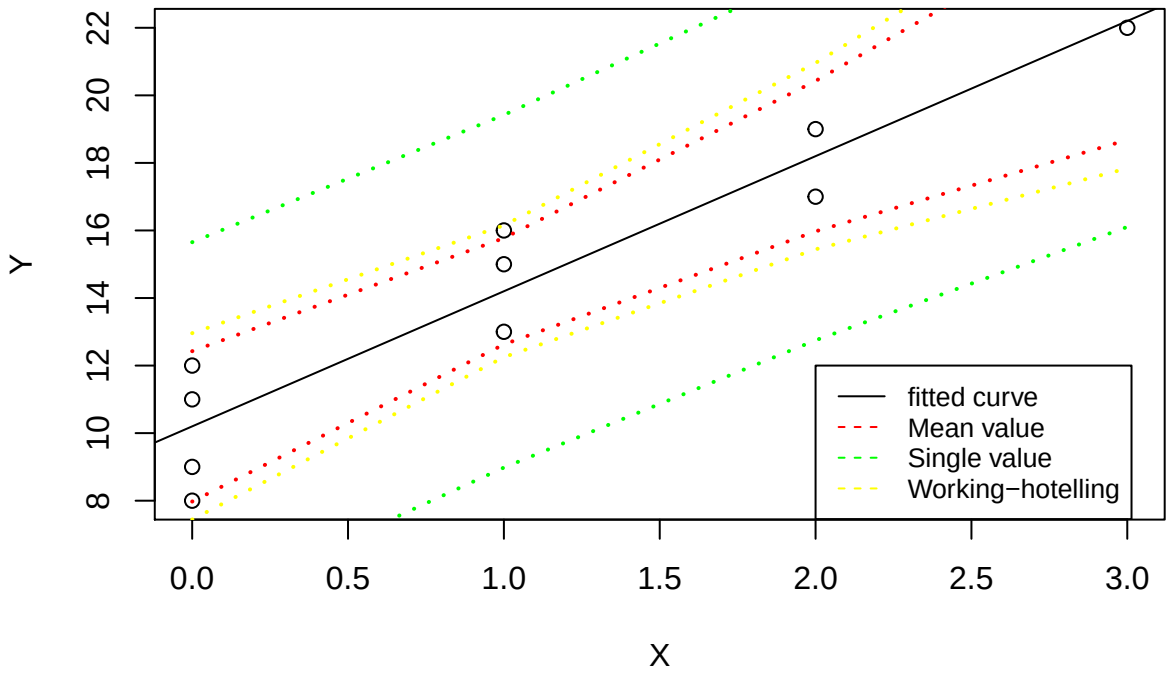
```
## Working hotelling band at level (X = 1 ) confidence interval is
##
##           [ 12.2492030649254 , 16.1507969350746 ].
##  Working hotelling band at level (X = 0 ) confidence interval is
##
##           [ 7.44115651698167 , 12.9588434830183 ].
##  Working hotelling band at level (X = 2 ) confidence interval is
##
##           [ 15.4411565169817 , 20.9588434830183 ].
##  Working hotelling band at level (X = 0 ) confidence interval is
##
##           [ 7.44115651698167 , 12.9588434830183 ].
##  Working hotelling band at level (X = 3 ) confidence interval is
##
##           [ 17.837885442875 , 26.562114557125 ].
##  Working hotelling band at level (X = 1 ) confidence interval is
##
##           [ 12.2492030649254 , 16.1507969350746 ].
##  Working hotelling band at level (X = 0 ) confidence interval is
##
##           [ 7.44115651698167 , 12.9588434830183 ].
##  Working hotelling band at level (X = 1 ) confidence interval is
##
##           [ 12.2492030649254 , 16.1507969350746 ].
##  Working hotelling band at level (X = 2 ) confidence interval is
##
##           [ 15.4411565169817 , 20.9588434830183 ].
##  Working hotelling band at level (X = 0 ) confidence interval is
##
##           [ 7.44115651698167 , 12.9588434830183 ].
```

```
plot(X,Y)
abline(a=intercept,b=slope,color='blue')
lines(X[order(X)], CI_lbYhat[order(X)],col="red", lwd=2, lty=3)
lines(X[order(X)], CI_ubhat[order(X)], col="red", lwd=2, lty=3)
```

```
lines(X[order(X)], CI_lbYhnew[order(X)],col="green", lwd=2, lty=3)
lines(X[order(X)], CI_ubYhnew[order(X)], col="green", lwd=2, lty=3)

lines(X[order(X)], CB_whlb[order(X)],col="yellow", lwd=2, lty=3)
lines(X[order(X)], CB_whub[order(X)], col="yellow", lwd=2, lty=3)

legend(2, 12, legend=c("fitted curve", "Mean value", "Single value", "Working-hotelling"),
       col=c("black", "red", "green", "yellow"), lty=c(1,2,2,2), cex=0.8)
```



It is revealed that the the working hotelling confidence band is a little wider than Mean value confidence band at any level of X. The prediction intervals for all levels of X will form a band with the largest width.

3. Tri-City office Equipment Corporation sells an imported copier on franchise basis and performs preventive maintainance and repair service on this copier. Data have been collected from 45 recent calls on users to perform routine preventive maintainance service; for each call, X is the numnber of copiers serviced and Y is the total number of minutes spent by the service person. Assume that simple linear regression model is appropriate. The following shows partial result.

```
Coefficients:
             Estimate Std. Error
(Intercept)  -0.5802     2.8039
X            15.0352     0.4831

Residual standard error: 8.914
```

a) Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. interpret your confidence interval.

b) Conduct a test to determine whether or not there is a linear association between X and Y here (i.e., $\beta_1 \neq 0$); At a significant level of 0.05, state the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

c) The manufacturer has suggested that the mean required time should not increase by more 14 minutes for each additional copier that is serviced on a service call. Conduct a test to test whether this standard is being satified by Tri-City. At a significant level of .05. State the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

d) Dose b_0 give any relevant information here about the "start-up" time on calls–i.e. about time required before service work is begun on the copiers at a customer location?

e) In order to perform the following hypothesis test ($\alpha = 0.05$),

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0$$

complete the following ANOVA table for the data. According to the F value and degree of freedom, use F tale to estimate the P-value of the test.

Source	degree of freedom	Sum of Squares	Mean Square	F-value
Model	1	76960	76960	968.78
Error	43	3416	79.44	
Corrected Total	44	80376		

f) Compare the F test statistic obtained here and demonstrate numerically its equivalence to the T test statisc in b).

Solution:

a) The change in the mean service time when the number of copiers serviced increased by 1 is 15.0352. The estimated standard deviation for b_1 is 0.4831.

```
cat(paste('The 90% confidence interval for b1 is\n
          [',15.0352 - qt(1 - 0.1/2, 45-2) * 0.4831, ',',
          15.0352 + qt(1 - 0.1/2, 45-2) * 0.4831, '].'))
```

```
## The 90% confidence interval for b1 is
##
##           [ 14.2230747432829 , 15.8473252567171 ].
```

We are 90% certain that the true slope β_1 will fall into this confidence interval.

b) The hypothesis test is stated as

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

Test statistic is given as $Ts = \frac{b_1 - 0}{s_{b_1}} = \frac{15.0352}{0.4831} = 31.1232$. We will reject the null hypothesis if $|t| > 2.0167$ or if p-value < 0.05. The p-value is given by p-value $= 2 \times P(t > Ts) \approx 0 < 0.05$. It is concluded that the null hypothesis that the linear association between X and Y does not exist can be rejected with high significance level of confidence.

c) The hypothesis test is given by

$$H_0 : \beta_1 \leq 14 \quad H_a : \beta_1 > 14$$

Test statistic is given as $Ts = \frac{b_1 - 14}{s_{b_1}} = \frac{1.0352}{0.4831} = 2.1428$. We will reject the null hypothesis if $t > 1.681$ or if p-value < 0.05. The p-value is given by p-value $= P(t > Ts) = 0.0189 < 0.05$. It means the null hypothesis can be rejected with sufficient evidence. The statement is hard to be achieved.

d) The intercept does not give relevant information on the "start-up" time on calls. Construct a hypothesis test $H_0 : \beta_0 = 0$, but it can not be rejected with high confidence since p-value $= 0.84$.

e) F value is computed using R

```
968.78 > qf(1-0.05,1,43)
```

```
## [1] TRUE
```

```
p.value <- 1 - pf(968.78,1,43)
print(paste('P-value is ', p.value))
```

```
## [1] "P-value is  0"
```

```
R2 <- 76960 / 80376
cat(paste('Coefficient of determination is ', R2))
```

```
## Coefficient of determination is  0.957499751169503
```

We reject the null hypothesis with the analysis of the variance.

f) Compare F test to the T test in b).

$$SSR = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y}) = b_1^2 \sum_{i=1}^{N}(X_i - \bar{X})^2$$

Then, we have

$$F^* = \frac{MSR}{MSE} = \frac{b_1^2 \sum_{i=1}^{N}(X_i - \bar{X})^2}{MSE} = (\frac{b_1}{s_{b_1}})^2 = (t^*)^2$$

The derivation implies the equivalence of the F test to the T test. Correspondingly, it can be verified numerically that

$$F^* = 968.78 \approx 31.1232^2 = T_s^2$$

$$(t(1-\alpha/2; N-2))^2 = F(1-\alpha; 1, N-2) = 2.0167^2 = 4.0670$$

4. Refer to the problem 2 in homework 1.

a) Complete the ANOVA table for the hypothesis test.

H_0 : ACT and GPA score are associated $\quad H_a$: ACT and GPA score are not associated

Or equivalently,

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$$

Estimate the p value and state the conclusion.

Source	degree of freedom	Sum of Squares	Mean Square	F-value
Model	1	0.009976081	0.009976081	0.1887
Error	8	0.4228315	0.05285394	
Corrected Total	9	0.4328076		

b) What is R^2? Perform a hypothesis test on the correlation, compute the test statistic and estimate p value, then state the conclusion. Use significant level of 0.05.

$$H_0 : \rho = 0, \quad H_a : \rho \neq 0$$

c) Compare 4a) to 3e), which model seems to be a better fit? Discuss the models based on MSE and R^2?

d) The GPA data used in this problem is actually the first 10 cases of a larger data set, and has a very small R^2, is it possible that for the complete set $n > 10$, R^2 will be zero? Could R^2 not be zero for the first 10 cases, yet equal to zero for all 30 cases? If applicable, sketch two scatter plots to demonstrate the two situations.

e) Use R to compute a 95% CI for the population coefficient.

Solution:

a) P-value is given by

$$\text{p-value} = P(F > F^*; 1, 8) = 0.394 > 0.05$$

Thus we fail to reject the null hypothesis.

b) R^2 is the coefficient of determination. The estimated correlation coefficient is 0.1518212.

```
r12 <- 0.1518212
n <- 10
Ts1 <- r12 * sqrt(n - 2) / sqrt(1 - r12^2)
cat(paste('Test statistic is ', Ts1, '\n'))
```

```
## Test statistic is  0.434451371752615
```

```
p.value1 <- (1 - pt(Ts1, n-2)) * 2
cat(paste('p-value is ', p.value1, '\n'))
```

```
## p-value is  0.675440841026002
```

```
cat(paste('R2 is estimated by', r12^2))
```

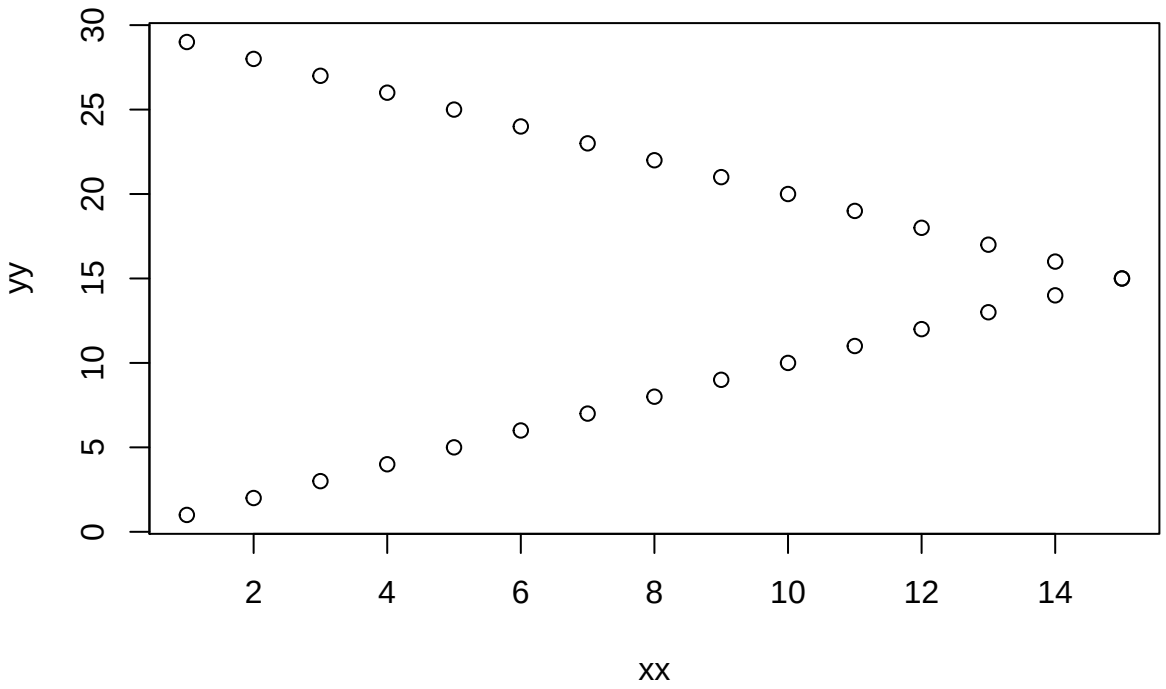
```
## R2 is estimated by 0.02304967676944
```

Since p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis, which means there is little linear association between GPA score and ACT score at high level of confidence significance.

c) R^2 in 3e) is 0.957499751169503, MSE in 3e) is 79.44. R^2 in 4a) is 0.02304967676944, MSE in 4a) is 0.05285394. Model 3e) seems to be a better fit, since the fitting curve in 3e) has a larger R^2 so that the variability of Y depends much more on X. However, we also like a smaller MSE since it indicates that the noise is smaller.

d) If the first 10 cases of a larger data set has a very small R^2, it is still possible that for the complete set $n > 10$, R^2 will not be zero. If R^2 is not zero for the first 10 cases, it can still be zero for all 30 cases. Let us first give a scatter plot for case II.

```
# plot the scatter point in case II
xx <- c(seq(1,15),seq(1,15))
yy <- c(xx[1:15],30-xx[16:30])
plot(xx,yy)
```




```
# Analyze the linear model
xx1 <- xx[1:10]
yy1 <- yy[1:10]
first10 <- data.frame(xx1,yy1)
first10.mod <- lm(yy1~xx1,first10)
summary(first10.mod)
```

```
##
## Call:
## lm(formula = yy1 ~ xx1, data = first10)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.661e-16 -1.157e-16  4.273e-17  2.153e-16  4.167e-16
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.123e-15  2.458e-16 4.571e+00  0.00182 **
## xx1         1.000e+00  3.961e-17 2.525e+16  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.598e-16 on 8 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 6.374e+32 on 1 and 8 DF,  p-value: < 2.2e-16
```

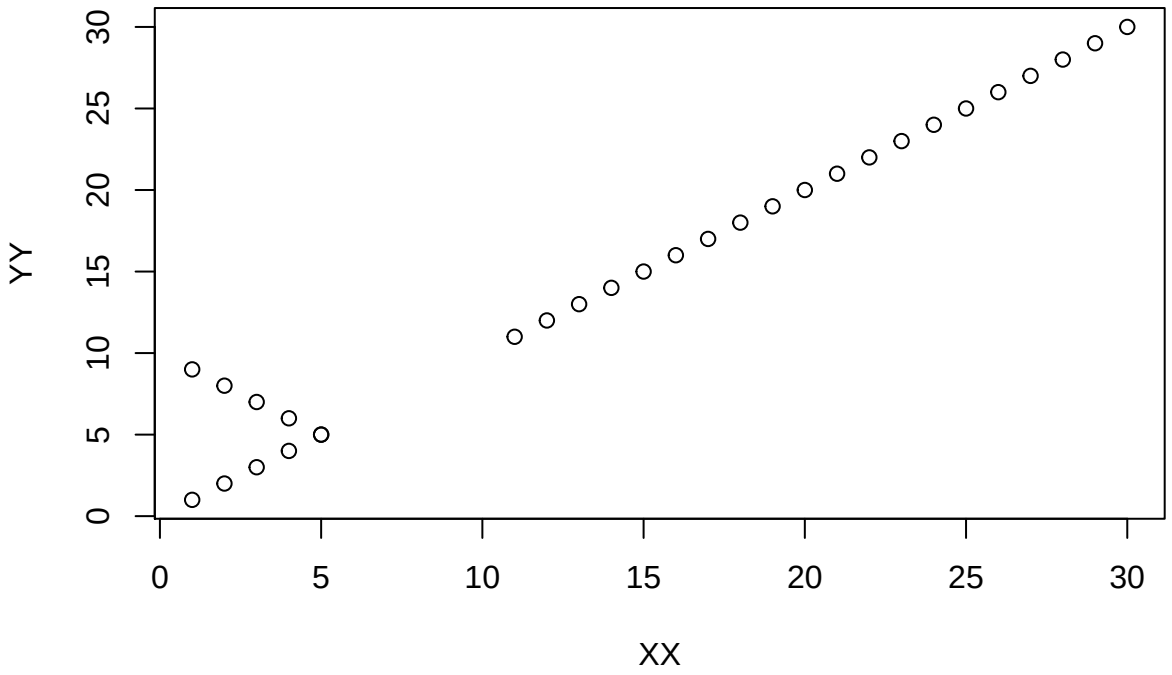
```
all30 <- data.frame(xx,yy)
all30.mod <- lm(yy~xx,all30)
summary(all30.mod)
```

```
##
## Call:
## lm(formula = yy ~ xx, data = all30)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.00  -6.75   0.00   6.75  14.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.500e+01  3.271e+00   4.585 8.62e-05 ***
## xx          1.501e-16  3.598e-01   0.000        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.515 on 28 degrees of freedom
## Multiple R-squared:  1.244e-32,  Adjusted R-squared:  -0.03571
## F-statistic: 3.482e-31 on 1 and 28 DF,  p-value: 1
```

For case I, let us see the following scatter plot.

```
# plot the scatter plot in case I
XX <- c(seq(1,5),seq(1,5),seq(11,30))
YY1 <- c(XX[1:5],10 - XX[6:10])
YY2 <- XX[11:30]
YY <- c(YY1,YY2)
plot(XX,YY)
```




```
# Analyze the linear model for case I
XX1 <- XX[1:10]
YY1 <- YY[1:10]
First10 <- data.frame(XX1,YY1)
First10.mod <- lm(YY1~XX1,First10)
summary(First10.mod)
```

```
##
## Call:
```

```
## lm(formula = YY1 ~ XX1, data = First10)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4.00  -1.75   0.00   1.75   4.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0000     2.0310   2.462   0.0392 *
## XX1           0.0000     0.6124   0.000   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.739 on 8 degrees of freedom
## Multiple R-squared:  5.259e-32,  Adjusted R-squared:  -0.125
## F-statistic: 4.207e-31 on 1 and 8 DF,  p-value: 1
```

```
All30 <- data.frame(XX,YY)
All30.mod <- lm(YY~XX,All30)
summary(All30.mod)
```

```
##
## Call:
## lm(formula = YY ~ XX, data = All30)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9364 -0.8912 -0.2176  0.3753  6.0636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0293     0.5772   3.516  0.00151 **
## XX            0.9071     0.0330  27.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.723 on 28 degrees of freedom
## Multiple R-squared:  0.9643, Adjusted R-squared:  0.963
## F-statistic: 755.7 on 1 and 28 DF,  p-value: < 2.2e-16
```

e) The 95% CI for the correlation coefficient is given as follows

```
library(DescTools)
zprime <- 1 / 2 * log((1 + r12) / (1 - r12))
sigma.zp <- sqrt(1 / (n - 3))
zeta.lb <- zprime - qnorm(1 - 0.05 / 2) * sigma.zp
zeta.ub <- zprime + qnorm(1 - 0.05 / 2) * sigma.zp
rho12.lb <- FisherZInv(zeta.lb)
rho12.ub <- FisherZInv(zeta.ub)
cat(paste('The CI is given by [',rho12.lb,',',rho12.ub,'].'))
```

```
## The CI is given by [ -0.528306219734338 , 0.713265964296451 ].
```

5. Use R to complete this problem. Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic, and the elapsed time since termination of the molding process. Sixteen batches of the plastic were made, and one test

item was molded from each batch. Each test item was randomly assigned to one of four predetermined time levels ($X = 16, 24, 32,$ or40 hours), and the hardness (Y) was measured after the assigned elapsed. Data is in plastic.csv

a) Obtain the estimated simple linear regression model (SLR or SLM). Plot the estimated regression function and the data. Does a linear regression function appear to be a good fit?
b) Plot the residuals against the fitted values to ascertain whether any departures from regression model are evident. State your findings.
c) Plot a normal probability plot of the residuals. Perform a Shapiro-Wilk normality test on the residuals. Does the normality assumption appear to be reasonable here?
d) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, $X \leq 24$, $X > 24$, use $\alpha = 0.05$. Does your conclusion support your preliminary findings in part b)?