R Intermediate Short Course
Session 7 (of 7) - Analysis of Variance (continued)

The University of the West Indies, St Augustine

Sunday 24th July 2022
(5:00pm - 7:00pm)
(online)

Completely Randomised Block Design

The completely randomised block design is an extension to the ANOVA or F test. It involves the introduction of blocking factors or blocks in order to deal with nuisance factors (or noise).

This additional source, blocks reduces the error (or residuals) produced in an ANOVA table. Block effects however, are not tested as it restricts the process of randomisation.

Therefore the hypotheses tested still involve difference in means among treatment groups:

H_0: $\mu_1 = \mu_2 = ... = \mu_a$ versus H_1: At least one μ_i is different

where a is the number of treatments or groups being compared.

We reject the null hypothesis H_0 if,

1. The test statistic $>$ critical value
2. The p-value $<$ level of significance, α

For the block effect:
H_0: There is no block effect.
H_1: There is a block effect,

where a is the number of treatments or groups being compared.
NB: If we have a block effect, then blocking made the analysis better.

The test statistic and the p-value are both calculated by means of an ANOVA table where the general form of the ANOVA in a completely randomised block design is as follows:

Source	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Treatments	a - 1	SSTr	MSTr	MSTr/MSE
Blocks	b - 1	SSB	MSB	
Residuals	(a-1)×(b-1)	SSE	MSE	
Total	N - 1	SST		

where b = the number of blocks

and N = a × b = the number of observations in the data set.

In R, the p-values are included as the last column in the table and the row of totals is excluded.

The test statistic and p-value for blocks is also generated, even though we do not test the significance of blocks.

The critical value F_{α, df_1, df_2} is computed by:

```
qf(1 - alpha, df1, df2)
```

where df1 = a - 1 (treatment degree of freedom)
and df2 = (a-1)×(b-1) (residual degree of freedom)

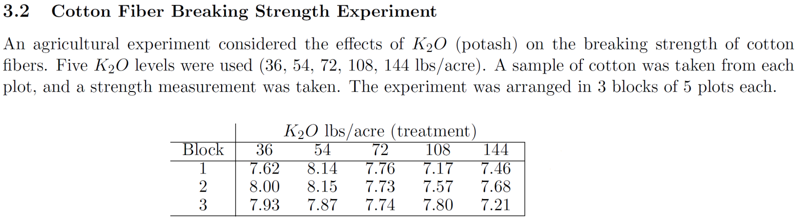
Performing CRBD in R

```
data = c(table)
T = c("Trt1", "Trt2",..., "Trta")
trt = gl(a, 1, N, factor(T))
blk = gl(b, a, N)
crbd = aov(data ~ trt + blk)
summary(crbd)
```

where,
table = the data set arranged in the format of a table or matrix
T = a vector containing the names of the treatments
trt and *blk* = the generalized linear models required for the CRBD.

Note that the `c()` command creates a vector of the table by columns. This vector should be in the order of the blocks. Therefore if a table is in the form where the blocks are rows then we must first transpose the table i.e. `data = c(t(table))`.

Example 1: Using the data provided below, construct an ANOVA table for the completely randomised block design (CRBD) and test the relevant hypotheses at $\alpha = 0.05$.



3.2 Cotton Fiber Breaking Strength Experiment

An agricultural experiment considered the effects of K_2O (potash) on the breaking strength of cotton fibers. Five K_2O levels were used (36, 54, 72, 108, 144 lbs/acre). A sample of cotton was taken from each plot, and a strength measurement was taken. The experiment was arranged in 3 blocks of 5 plots each.

	K_2O lbs/acre (treatment)				
Block	36	54	72	108	144
1	7.62	8.14	7.76	7.17	7.46
2	8.00	8.15	7.73	7.57	7.68
3	7.93	7.87	7.74	7.80	7.21

Required to test

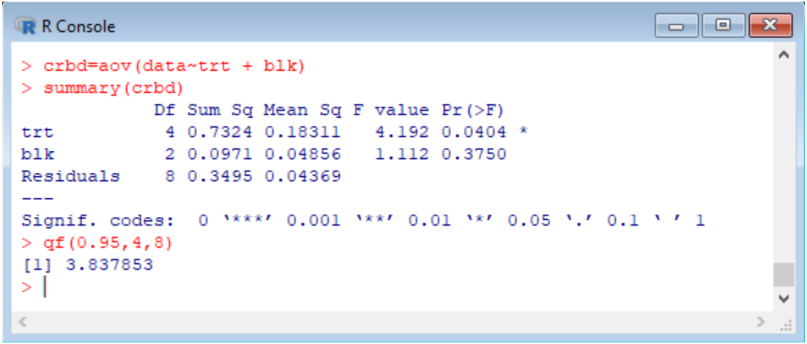
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs H_1: At least one μ_i is different

STEPS:

1. Enter the data as a table or matrix in R, this can be entered manually or in an external data file to be imported.
2. Perform the CRBD by first specifying T, a, b and N. Recall N = a×b.
3. Generate the ANOVA table and related critical value.
4. Interpret the R output to make a conclusion from the test.

```
> table = matrix(c(7.62,8.14,7.76,7.17,7.46,8.00,
    8.15,7.73,7.57,7.68,7.93,7.87,7.74,7.80,7.21),
    nrow=3, ncol=5, byrow=TRUE)
> data = c(t(table))
> T = c("36", "54", "72", "108", "144")
> a = 5
> b = 3
> N = a * b
> trt = gl(a, 1, N, factor(T))
> blk = gl(b, a, N)
> crbd = aov(data ~ trt + blk)
> summary(crbd)
```

Critical value: `> qf(0.95,4,8)`



```
> crbd=aov(data~trt + blk)
> summary(crbd)
            Df Sum Sq Mean Sq F value Pr(>F)
trt          4 0.7324 0.18311   4.192 0.0404 *
blk          2 0.0971 0.04856   1.112 0.3750
Residuals    8 0.3495 0.04369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> qf(0.95,4,8)
[1] 3.837853
> 
```

The treatment test statistic F $= 4.192 >$ the critical value $F_{0.05,4,8} = 3.838$ and the p-value $= 0.0404 < 0.05$ so we reject H_0 concluding that at least one level of potash has a significantly different mean.

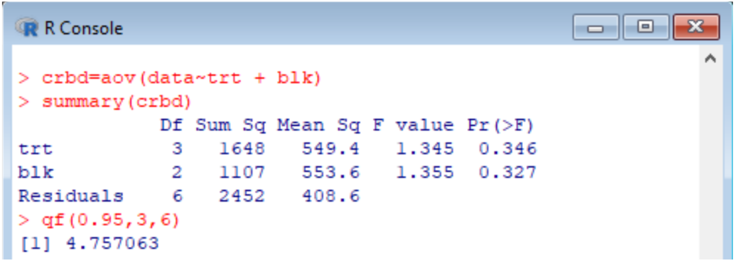
Example 2: A study was carried out to determine the lifelines of 4 premium brands of pens. It was thought that the writing surface might affect lifelines so 3 different surfaces were randomly selected. The table below shows the lifelines collected in minutes.

	Writing Surface		
Brand of Pen	1	2	3
1	709	713	660
2	668	722	692
3	659	666	678
4	698	704	686

(i) Construct an ANOVA table for the data.
(ii) Test at $\alpha = 0.05$ whether brand of pen has an effect on lifeline.

Required to test
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: At least one μ_i is different

```
> table = matrix(c(709,713,660,668,722,692,659,
    666,678,698,704,686), nrow=4, ncol=3,
    byrow=TRUE)
> data = c(table)
> T = c("BP1", "BP2", "BP3", "BP4")
> a = 4
> b = 3
> N = a * b
> trt = gl(a, 1, N, factor(T))
> blk = gl(b, a, N)
> crbd = aov(data ~ trt + blk)
> summary(crbd)
> qf(0.95,3,6)
```



```
> crbd=aov(data~trt + blk)
> summary(crbd)
            Df Sum Sq Mean Sq F value Pr(>F)
trt          3   1648   549.4   1.345  0.346
blk          2   1107   553.6   1.355  0.327
Residuals    6   2452   408.6
> qf(0.95,3,6)
[1] 4.757063
```

The treatment test statistic F $= 1.345 <$ the critical value $F_{0.05,3,6} = 4.757$ and the p-value $= 0.346 > 0.05$ so we do not reject H_0 concluding that the mean lifelines of all 4 brands of pen are equal.

Example 3: The built-in data set *VADeaths* gives the death rate in Virginia for the year 1940. By first setting the data set to a single vector,

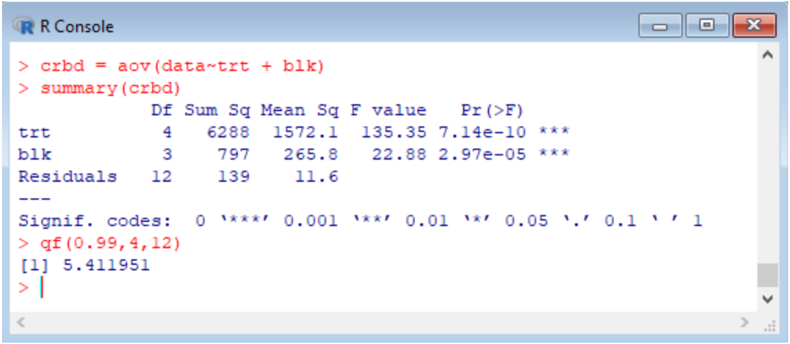
(i) Construct an ANOVA table for the block design.

(ii) Test the hypothesis that at least one age group has a significantly different death rate. Use a 1% level of significance.

There are 5 age groups in the *VADeaths* data set $\therefore$ a = 5. Also, the data is arranged as treatments (age) are rows and blocks are the columns

Required to test
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs H_1: At least one μ_i is different

```
> data = c(VADeaths)
> T = c("50-54","55-59","60-64","65-69","70-74")
> a = 5
> b = 4
> N = a * b
> trt = gl(a, 1, N, factor(T))
> blk = gl(b, a, N)
> crbd = aov(data ~ trt + blk)
> summary(crbd)
> qf(0.99,4,12)
```



```
> crbd = aov(data~trt + blk)
> summary(crbd)
            Df Sum Sq Mean Sq F value   Pr(>F)    
trt          4   6288  1572.1  135.35 7.14e-10 ***
blk          3    797   265.8   22.88 2.97e-05 ***
Residuals   12    139    11.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(0.99,4,12)
[1] 5.411951
> 
```

The treatment test statistic F $= 135.35 >$ the critical value $F_{0.01,4,12} =$ 5.412 and the p-value $= 7.14 \times 10^{-10} < 0.01$ so we reject H_0 concluding at least one age group has a significantly different mean death rate.

Session 5 - Regression Analysis

Simple Linear Regression

A **deterministic model** is a mathematical model used to predict or determine the outcomes of one variable using the known values of another variable and the relationship between both of them.

The **simple linear regression** model is the simplest deterministic model for the relationship between variables x and y.

It is defined by the equation: $y = \beta_0 + \beta_1 x + \varepsilon$

where y = response or dependent variable
x = predictor or independent variable
β_0 = population intercept
β_1 = population slope
and ε = random error or residual

Graphically this model is the best fit line in a scatter-plot of y_i values versus x_i values related to the regression.

The random error ε in a regression model is normally distributed with mean 0 and variance σ^2 i.e. $\varepsilon \sim N(0, \sigma^2)$

The assumptions are as follows:

1. Residuals are normally distributed
2. Residuals have a constant variance
3. Residuals are independent

When $\varepsilon = 0$ we obtain a fitted model which estimates the true linear regression model (given previously).
The fitted regression model is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Performing a regression in R

```
> reg = lm(y ∼ x)
> summary(reg)
```

where x and y are both vectors.
This summary gives the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ required for the fitted model.

The output above also gives the values R^2 and adjusted R^2.

The **coefficient of variation,** R^2 for a simple linear regression tells us how much variation is accounted for by the model.
Values of R^2 range from 0 to 1 i.e. $0 \leq r^2 \leq 1$. An R^2 value above 0.60 or 60% indicates that the regression model is fit and adequate.

When comparing two linear regression models we consider the **adjusted** R^2 values or $\bar{R}^2$ from both models.
The regression model with a higher $\bar{R}^2$ value is the better model.

Testing the significance of a regression model

We can test the significance of a linear regression model in two ways:

1. F test or one-way ANOVA
2. T distribution hypothesis test

1. F-test

The F-test for a regression, tests the hypotheses

H_0: regression is insignificant vs H_1: regression is significant

where the null hypothesis H_0 is rejected if

1. The test statistic $>$ critical value
2. The p-value $<$ level of significance, α.

The test statistic F and the p-value are both computed by means of an ANOVA table.

The general form of the ANOVA table for a regression is as follows:

Source	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	1	SSR	MSR	MSR/MSE
Residuals	n - 2	SSE	MSE	
Total	n - 1	SST		

The ANOVA generated in R includes a p-value and excludes the last row.

Performing F-test for regression in R

Generating ANOVA table:

```
anova.reg = aov(reg)
summary(anova.reg)
```

where `reg` was defined previously.

Generating the critical value F_{α,df_1,df_2}:

```
qf(1 - alpha, df1, df2)
```

where $df_1 = 1$ and $df_2 =$ n - 2 for a sample size n

2. T test

The T test for a regression model, determines whether the slope β_1 is significantly different from 0. This is because if the slope is 0 then the model will simply be the horizontal line $y = \beta_0$.

The hypotheses tested are

H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$

We reject H_0 if one of the 3 conditions are satisfied:

1. |test statistic| $>$ critical value

The test statistic t, is the t value for x given the summary of the regression performed.

The critical value $t_{\alpha/2,df}$ is found by

```
qt(alpha/2, df, lower.tail=F, log.p=F)
```

where df = n - 2

2. p-value $< \alpha$

The p-value is also given in the summary of the regression performed.

3. confidence interval of $\hat{\beta}_1$ does not contain 0

The confidence interval for both $\hat{\beta}_0$ and $\hat{\beta}_1$ is generated by

```
confint(reg, level = CL)
```

where CL = 1 - α and `reg` was defined previously.

We are interested in the second confidence interval i.e. for x.

Testing the assumptions of residuals

Recall the 3 assumptions of residuals given in slide 15.
After performing a regression, we can determine whether these assumptions hold true using R.

1. Residuals are normally distributed

For residuals to be normally distributed one of two conditions must hold:

(i) A histogram plot of residuals follows the shape of a normal distribution i.e. symmetric.

(ii) A Q-Q or quantile-quantile plot shows thats the majority of data points lie on the Q-Q line.

Histogram:

```
hist(reg$residuals, main="Histogram of Residuals")
```

Q-Q plot:

```
qqnorm(reg$residuals, pch = 20)
qqline(reg$residuals)
```

2. Residuals have constant or equal variance

For this assumption to hold, a plot of residuals versus fitted values should split the data into high and low values (band).

Plot of residuals versus fitted values:

```
plot(reg$fitted.values, reg$residuals, main =
    "Residuals versus Fitted", pch=20)
abline(h=0, lty=2)
```

If this assumption does not hold then all our outputs are worthless (heteroscedasticity).

3. Residuals are independent

This assumption is true if a plot of residuals versus time produces data points which oscillate about 0 (x-axis) with no apparent pattern.

Plot of residuals versus time:

```
plot(1:n,reg$residuals, main = "Residuals versus
    time order", pch = 20)
abline(h=0, lty=2)
```

where n = sample size

All four graphs can be constructed in one window in R (see R script).

Plotting a linear regression model

```
plot(x, y, main = "Regression line")
abline(reg,col="black")
```

Example 1: A chemist wishes to investigate how the pH of milk changes over time. The values for x and y are shown below where y represents the milk pH and x represents day.

x	1	2	3	4	5	6	7	8	9
y	6.8	6.6	6.6	6.4	6.1	5.7	5.5	5.2	4.9

(i) Determine a fitted regression model for the data and comment on the value of R^2.

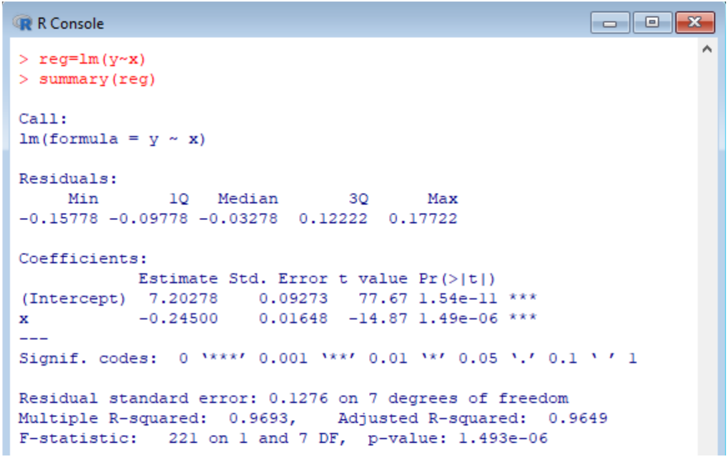
(ii) By first constructing an ANOVA table, test whether the regression model is significant at $\alpha = 0.10$.

(iii) Using an appropriate t-test, construct a confidence interval at $\alpha = 0.10$ to determine whether the regression is significant. Does your result match the decision made in (ii)?

(iv) Plot a regression line to represent the data given.

(v) Test whether the three assumptions of residuals hold true for the regression model.

(i)
```
> x = c(1,2,3,4,5,6,7,8,9)
> y = c(6.8,6.6,6.6,6.4,6.1,5.7,5.5,5.2,4.9)
```



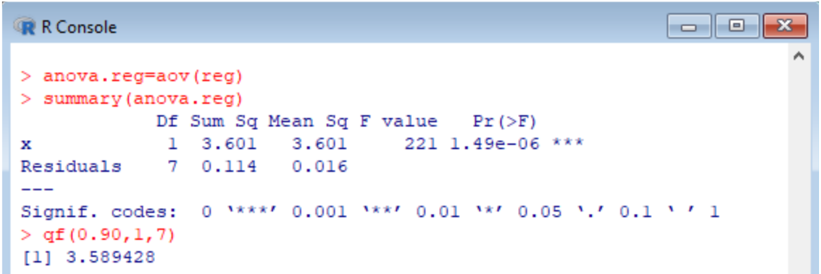
```
> reg=lm(y~x)
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.15778 -0.09778 -0.03278  0.12222  0.17722 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.20278    0.09273   77.67 1.54e-11 ***
x           -0.24500    0.01648  -14.87 1.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1276 on 7 degrees of freedom
Multiple R-squared:  0.9693,    Adjusted R-squared:  0.9649 
F-statistic:   221 on 1 and 7 DF,  p-value: 1.493e-06
```

The fitted regression model is $\hat{y} = 7.203 - 0.245x$

$R^2 = 0.9693$, this means that the regression model explains 96.93% of the total variation.

(ii) Required to test at $\alpha = 0.10$

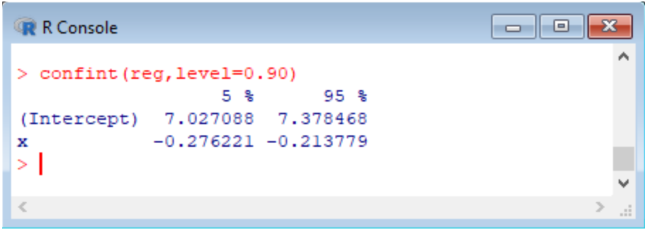
H_0: regression is insignificant vs H_1: regression is significant



```
> anova.reg=aov(reg)
> summary(anova.reg)
            Df Sum Sq Mean Sq F value   Pr(>F)
x            1  3.601   3.601     221 1.49e-06 ***
Residuals    7  0.114   0.016
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(0.90,1,7)
[1] 3.589428
```

The test statistic F $= 221 > F_{0.10,1,7} = 3.589$ and the p-value < 0.10 so we reject H_0 concluding that the regression model is significant.

(iii) Required to test at $\alpha = 0.10$ H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$



```
> confint(reg,level=0.90)
                    5 %      95 %
(Intercept)  7.027088  7.378468
x           -0.276221 -0.213779
> 
```

The 90% confidence interval for $\hat{\beta}_1$ is (-0.276, -0.214). This interval does not contain 0, therefore we can reject H_0 concluding that the slope and hence regression is significant.

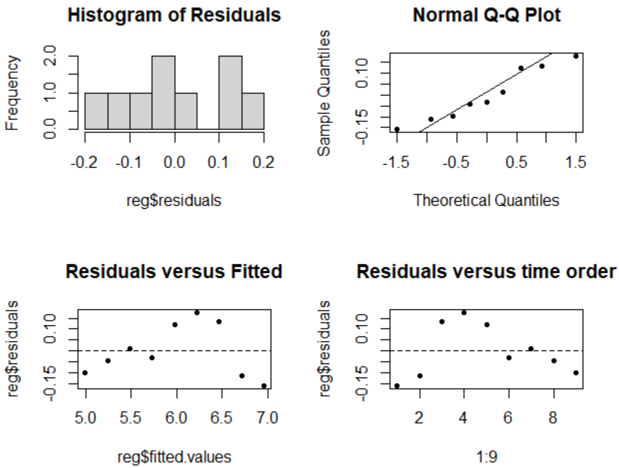
This is the same result obtained in (ii).

(iv) The regression line can be generated by

```
> plot(x,y,xlab="Day",ylab="pH",main="Relationship
    between milk pH and day")
> abline(reg,col="black")
```

See R script and output for regression line plot.

(v)



The histogram of residuals is not symmetric and only 2 of the 9 data points lie on the Q-Q line of the Q-Q normality plot. The residuals are therefore not normally distributed.

The plot of residuals versus fitted values show that the data points are split into high and low values in a band of (-0.15, 0.10). Residuals therefore have constant variance.

The residuals versus time plot shows data points oscillating about 0 with no apparent pattern. Thus the residuals are independent.

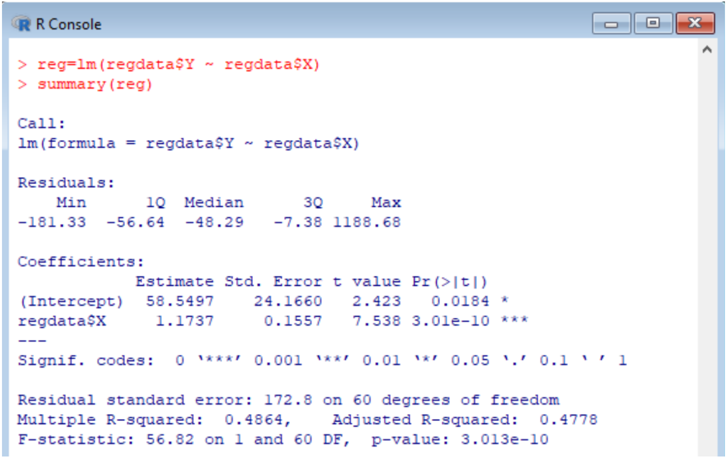
Example 2: The *slr.xlsx* data set provided shows data collected from an experiment.

(i) Obtain a fitted regression model to represent the data given.

(ii) By computing a suitable critical value at $\alpha = 0.05$, conduct a t-test to determine if the slope coefficient $\hat{\beta}_1$ is significant.

(iii) If a similar experiment conducted produced a regression model with an adjusted R^2 value of 0.45, which model is better?

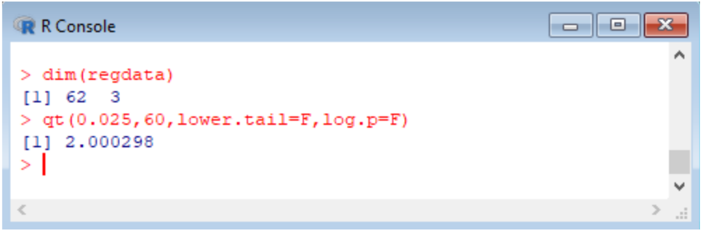
(i)
```
> library(readxl)
> regdata = read_excel("slr.xlsx")
```



```
> reg=lm(regdata$Y ~ regdata$X)
> summary(reg)

Call:
lm(formula = regdata$Y ~ regdata$X)

Residuals:
    Min      1Q  Median      3Q     Max 
-181.33  -56.64  -48.29   -7.38 1188.68 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.5497    24.1660   2.423   0.0184 *  
regdata$X     1.1737     0.1557   7.538 3.01e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 172.8 on 60 degrees of freedom
Multiple R-squared:  0.4864,	Adjusted R-squared:  0.4778 
F-statistic: 56.82 on 1 and 60 DF,  p-value: 3.013e-10
```

The fitted regression model is $\hat{y} = 58.550 + 1.174x$.

(ii) Required to test at $\alpha = 0.05$ H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$



```
> dim(regdata)
[1] 62  3
> qt(0.025,60,lower.tail=F,log.p=F)
[1] 2.000298
> 
```

The regression summary in (i) shows that the test statistic, t = 7.538. Since t > critical value $t_{0.025,60}$ (= 2.000) we can reject H_0 to conclude that the slope coefficient $\hat{\beta}_1$ is significant.

(iii) The adjusted R^2 value for the current model is 0.4778 which is larger than that of the new model with $\bar{R}^2 = 0.45$.
Therefore the current model is better.