Chapter 1 Simple Linear Regression (Part 2)

1 Software R and regression analysis

Downloadable from http://www.r-project.org/; some useful commands

- setwd('path..') ... to change the directory for data loading and saving
- read.table .... for reading/loading data
- data$variable .... variable in the data
- plot(X, Y) ... plotting Y against X (starting a new plot);
- lines(X, Y)... to add lines on an existing plot.
- object = lm(y $\sim$ x)... to call "lm" to estimate a model and stored the calculation results in "object"
- Exporting the plotted figure (save as .pfd, .ps or other files)

Example 1.1 Suppose we have 10 observations for (X, Y): (1.2, 1.91), (2.3, 4.50), (3.5, 2.13), (4.9, 5.77), (5.9, 7.40), (7.1, 6.56), (8.3, 8.79), (9.2, 6.56), (10.5, 11.14), (11.5, 9.88). They are stored in file **(data010201.dat)**. We hope to fit a linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$$

code of R (the words after # are comments only)

```
mydata = read.table('data010201.dat')     # read the data from the file
```

```
X = mydata$V1     # select X
Y = mydata$V2     # select Y
plot(X, Y)     # plot the observations (data)
myreg = lm(Y ~ X)     # do the linear regression
summary(myreg)     # output the estimation
```

Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	1.3931	0.9726	1.432	0.189932	
X	0.7874	0.1343	5.862	0.000378	***

—

Sign. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.406 on 8 degrees of freedom
Multiple R-squared: 0.8111, Adjusted R-squared: 0.7875
F-statistic: 34.36 on 1 and 8 DF, p-value: 0.0003778

```
lines(X, myreg$fitted)     # plot the fitted
title("Scatter of (X,Y) and fitted linear regression model")     # add title
# Please get to know how to make a figure file for latter use
```

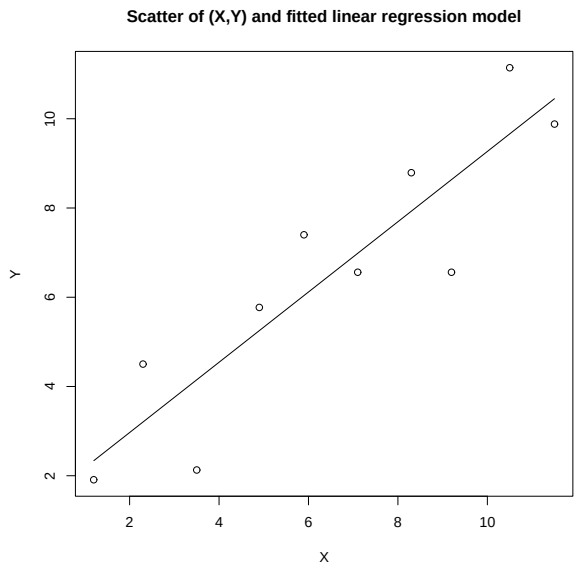

Figure 1: **(R code)**

The fitted regression line/model is

$$\hat{Y} = 1.3931 + 0.7874X$$

For any new subject/individual with X', its prediction of $E(Y)$ is

$$\hat{Y} = b_0 + b_1 X'.$$

For the above data,

- If $X' = -3$, then we predict $\hat{Y} = -0.9690$
- If $X' = 3$, then we predict $\hat{Y} = 3.7553$
- If $X' = 0.5$, then we predict $\hat{Y} = 1.7868$

2 Properties of Least squares estimators

Statistical properties in theory

- LSE is unbiased: $E\{b_1\} = \beta_1$, $E\{b_0\} = \beta_0$.

  Proof: By the model, we have

  $$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}$$

  and

  $$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1 \bar{X} - \bar{\varepsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

  recall that $E\varepsilon_i = 0$. It follows that

  $$Eb_1 = \beta_1.$$

For b_0,

$$\begin{aligned} E(b_0) &= E(\bar{Y} - b_1\bar{X}) = \beta_0 + \beta_1\bar{X} - E(b_1)\bar{X} = \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} \\ &= \beta_0 \end{aligned}$$

- Variance of the estimators

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad Var(b_0) = \frac{1}{n}\sigma^2 + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\sigma^2$$

[Proof:

$$\begin{aligned} Var(b_1) &= Var(\frac{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^n (X_i - \bar{X})^2}) \\ &= \{\sum_{i=1}^n (X_i - \bar{X})^2\}^{-2} Var\{\sum_{i=1}^n (X_i - \bar{X})\varepsilon_i\} \\ &= \{\sum_{i=1}^n (X_i - \bar{X})^2\}^{-2} \sum_{i=1}^n (X_i - \bar{X})^2\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \end{aligned}$$

We shall prove the second equation later.]

- Estimated (fitted) regression function $\hat{Y}_i = b_0 + b_1 X_i$. We also call $\hat{Y}_i = b_0 + b_1 X_i$ the fitted value.

$$E\{\hat{Y}_i\} = EY_i$$

[Proof:

$$E(\hat{Y}_i) = E(b_0 + b_1 X_i) = E(b_0) + E(b_1)X_i = \beta_0 + \beta_1 X_i = EY_i$$

]

Numerical properties of fitted regression line

Recall the normal equations

$$\begin{aligned} -2\sum_{i=1}^n (Y_i - b_0 - b_1 X_i) = 0 \\ -2\sum_{i=1}^n X_i(Y_i - b_0 - b_1 X_i) = 0 \end{aligned}$$

and $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$. It follows

$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} X_i e_i = 0$$

The following properties follows

- $\sum_{i=1}^{n} e_i = 0$
- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$
- $\sum_{i=1}^{n} e_i^2 = \min_{b_0,b_1}\{Q\}$
- $\sum_{i=1}^{n} X_i e_i = 0$
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
- Regression line always goes to $(\bar{X}, \bar{Y})$
- $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + \epsilon_i$, where $\epsilon_i = \varepsilon_i - \bar{\varepsilon}$.
- The coefficient and the correlation coefficient

$$b_1 = r_{X,Y}\frac{s_Y}{s_X}$$

where

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}, \quad s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$
$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

2.1 Estimation of Error Terms Variance σ^2

- **Sum of squares of residuals** or **error sum of squares** (SSE)

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Estimate σ^2 by

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

called mean squared error (MSE), i.e.

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

or denoted by $\hat{\sigma}^2$.

Why is it divided by $n-2$? because there are TWO constraints on $e_i, i = 1, ..., n$, i.e. the normal equations.

- **s^2 is unbiased estimator of σ^2, i.e.** $E(s^2) = \sigma^2$

[Proof: For any $\xi_1, ..., \xi_n$ IID with mean μ and variance σ^2, we have

$$\begin{aligned}
&E\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 = E\sum_{i=1}^{n}[(\xi_i - \mu) - (\bar{\xi} - \mu)]^2 \\
&= E\{\sum_{i=1}^{n}(\xi_i - \mu)^2 - n(\bar{\xi} - \mu)^2\} \\
&= \sum_{i=1}^{n} Var(\xi) - nVar(\bar{\xi}) \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

This is why we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2}{n-1}.$$

Consider

$$\begin{aligned}
Var(\xi_1 - \bar{\xi}) &= Var\{(1 - \frac{1}{n})\xi_1 - \overbrace{\frac{1}{n}\xi_2 - ... - \frac{1}{n}\xi_n}^{\text{n-1 terms}}\} \\
&= (1 - \frac{1}{n})^2\sigma^2 + \frac{1}{n^2}\sigma^2 + ... + \frac{1}{n^2}\sigma^2 \\
&= (1 - \frac{2}{n} + \frac{1}{n^2})\sigma^2 + \frac{n-1}{n^2}\sigma^2 \\
&= (1 - \frac{1}{n})\sigma^2.
\end{aligned}$$

similarly, for any i,

$$Var(\xi_i - \bar{\xi}) \quad = \quad (1 - \frac{1}{n})\sigma^2.$$

Now turn to the estimator s^2. Consider

$$\begin{aligned}
&E\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\} = \sum_{i=1}^{n} E(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Var(Y_i - \hat{Y}_i) + \{E(Y_i - \hat{Y}_i)\}^2 \\
&= \sum_{i=1}^{n} Var\{(Y_i - \bar{Y} - b_1(X_i - \bar{X}))^2\} \\
&= \sum_{i=1}^{n}\{Var(Y_i - \bar{Y}) - 2Cov(Y_i - \bar{Y}, b_1(X_i - \bar{X})) + Var(b_1)(X_i - \bar{X})^2\} \\
&= \sum_{i=1}^{n}\{Var(Y_i - \bar{Y}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\} \\
&= \sum_{i=1}^{n}\{Var(\varepsilon_i - \bar{\varepsilon}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\} \\
&= (n-1)\sigma^2 - 2Cov(\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= (n-1)\sigma^2 - 2Cov(b_1\sum_{i=1}^{n}(X_i - \bar{X})^2, b_1) + Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= (n-1)\sigma^2 - Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-2)\sigma^2.
\end{aligned}$$

Thus

$$E(s^2) = \sigma^2$$

Example For the above example, the MSE (estimator of $\sigma^2 = Var(\varepsilon_i)$) is

$$MSE = \sum_{i=1}^{n} e_i^2/(n-2) = 1.975997.$$

or

$$\hat{\sigma} = \sqrt{MSE} = 1.405702$$

which is also called **Residual standard error**.

How to find the value in the output of R?

3 Regression Without Predictors

At first glance, it doesn't seem that studying regression without predictors would be very useful. Certainly, we are not suggesting that using regression without predictors is a major data analysis tool. We do think that it is worthwhile to look at regression models without predictors to see what they can tell us about the nature of the constant. Understanding the regression constant in these simpler models will help us to understand both the constant and the other regression coefficients in later more complex models.

Model

$$Y_i = \beta_0 + \varepsilon_i, \qquad i = 1, 2, ..., n.$$

where as before, we assume

$$\varepsilon_i, i = 1, 2, ..., n \text{ are IID with } E(\varepsilon_i) = 0 \text{ and } Var(\varepsilon_i) = \sigma^2$$

(We shall call this model Regression Without Predictors)

The least square estimator b_0 is to minimizer of

$$Q = \sum_{i=1}^{n} \{Y_i - b_0\}^2$$

Note that

$$\frac{dQ}{db_0} = -2 \sum_{i=1}^{n} \{Y_i - b_0\}$$

Letting it equal 0, we have the **normal equation**

$$\sum_{i=1}^{n} \{Y_i - b_0\} = 0$$

which leads to the (ordinary) least square estimator

$$b_0 = \bar{Y}.$$

The fitted model is

$$\hat{Y}_i = b_0.$$

The fitted residuals are

$$e_i = Y_i - \hat{Y}_i = Y_i - \bar{Y}_i$$

- Can you prove the estimator is unbiased, i.e $Eb_0 = \beta_0$?
- How to estimate σ^2?

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

Why it is divided by $n-1$?

4 Inference in regression

Next, we consider the simple linear regression model

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned} \tag{1}$$

under assumptions of normal random errors.

- X_i is a known, observed, and nonrandom
- $\varepsilon_1, ..., \varepsilon_n$ are independent $N(0, \sigma^2)$, Thus Y_i is random
- β_0, β_1 and σ^2 are parameters.

By the assumption, we have

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

and

$$Var(Y_i) = \sigma^2$$

4.1 Inference of β_1

We need to check whether $\beta_1 = 0$ (or any other specified value, say -1.5), why

- To check whether X and Y has linear relationship

- To see whether the model can be simplified (if $\beta_1 = 0$, the model becomes $Y_i = \beta_0 + \varepsilon_i$, a regression model without predictors.) For example, **Hypotheses** $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 \neq 0$

Sample distribution of b_1 recall

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Theorem 4.1 For model (1) with normal assumption of ε_i then

$$b_1 \sim N\Big(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big)$$

Proof Recall the fact that *any linear combination of independent normal distributed random variables is still normal. To find its distribution, we only need to find its mean and variance.*

Since Y_i are normal and independent, thus b_1 is **normal**, and

$$Eb_1 = \beta_1$$

and (we have proved that)

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

The theorem follows.

Question: what is the distribution of $b_1/\sqrt{Var(b_1)}$ under H_0? Can we use this Theorem to test the hypothesis H_0? why

Estimated Variance of b_1. (Estimating σ^2 by MSE)

$$s^2(b_1) = \frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n e_i^2/(n-2)}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$s(b_1)$ is the **Standard Error (or S.E.)** of b_1, (or called Standard deviation)

sample distribution of $(b_1 - \beta_1)/s(b_1)$

$$\frac{b_1 - \beta_1}{s(b_1)} \text{ follows } t(n-2) \text{ for model (1)}$$

Confidence interval for β_1. Let $t_{1-\alpha/2}(n-2)$ or $t(1-\alpha/2, n-2)$ the $1-\alpha/2-$quantile of $t(n-2)$.

$$P(t(\alpha/2, n-2) \leq (b_1 - \beta_1)/s(b_1) \leq t(1-\alpha/2, n-2)) = 1 - \alpha$$

By symmetry of the distribution, we have

$$t(1-\alpha/2, n-2) = -t(\alpha/2, n-2)$$

Thus, with confidence $1-\alpha$, we have the confidence interval for β_1 is

$$-t(1-\alpha/2, n-2) \leq (b_1 - \beta_1)/s(b_1) \leq t(1-\alpha/2, n-2)$$

i.e.

$$b_1 - t(1-\alpha/2, n-2) * s(b_1) \leq \beta_1 \leq b_1 + t(1-\alpha/2, n-2) * s(b_1)$$

Example 4.2 For the example above, find the 95% confidence interval for β_1?

solution: since $n = 10$, we have $t(1-0.05/2, 8) = 2.306$; the SE for b_1 is $s(b_1) = 0.1343$. Thus the confidence interval is

$$b_1 \pm t(1-0.05/2, 8) * s(b_1) = 0.7874 \pm 2.306 * 0.1343 = [0.4777, 1.0971]$$

Test of β_1

- Two-sided Test: to check whether β_1 is 0

  $$H_0 : \beta_0 = 0, \quad H_a : \beta_1 \neq 0$$

  Under H_0, we have random variable

  $$t = \frac{b_1}{s(b_1)} \sim t(n-2)$$

  Suppose the **significance level** is α (usually, 0.05, 0.01). Calculate t, say t^*

  - If $|t^*| \leq t(1-\alpha/2; n-2)$, then accept H_0.
  - If $|t^*| > t(1-\alpha/2; n-2)$, then reject H_0.

  The test can also be done based on the **p-value**, defined as $p = P(|t| > |t^*|)$. It is easy to see that

  $$\text{p-value} < \alpha \Longleftrightarrow |t^*| > t(1-\alpha/2; n-2)$$

  Thus

– If p-value $\geq \alpha$, then accept H_0.

– If p-value $< \alpha$, then reject H_0.

- One-sided test: for example to check whether β_1 is positive (or negative)

$$H_0 : \beta_1 \geq 0, \quad H_a : \beta_1 < 0$$

Under H_0, we have

$$t = \frac{b_1}{s(b_1)} = \frac{b_1 - \beta_1}{s(b_1)} + \frac{\beta_1}{s(b_1)} \sim t(n-2) + \text{a positive term}$$

Suppose the **significance level** is α (usually, 0.05, 0.01). Calculate t, say t^*

– If $t^* \geq t(\alpha; n-2)$, then accept H_0.

– If $t^* < t(\alpha; n-2)$, then reject H_0.

4.2 Inference about β_0

Sample distribution of b_0

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Theorem 4.3 For model (1) with normal assumption of ε_i then

$$b_0 \sim N\Big(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}]\Big)$$

[**Proof** The expectation is

$$Eb_0 = E\{\bar{Y}\} - E(b_1)\bar{X} = (\beta_0 + \beta_1 \bar{X}) - \beta_1 \bar{X} = \beta_0$$

Let $k_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, then (see the proof at the beginning of this part)

$$b_1 = \beta_1 + \sum_{i=1}^n k_i \varepsilon_i.$$

Thus

$$b_0 = \beta_0 + \frac{1}{n}\sum_{i=1}^n \varepsilon_i - \sum_{i=1}^n k_i \varepsilon_i = \beta_0 + \sum_{i=1}^n [\frac{1}{n} - k_i \bar{X}]\varepsilon_i$$

The **variance** is

$$Var(b_0) = \sum_{i=1}^n [\frac{1}{n} - k_i \bar{X}]^2 \sigma^2 = [\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}]\sigma^2$$

Therefore the Theorem follows.]

Estimated Variance of b_0 (by replacing σ^2 with MSE).

$$s^2(b_0) = MSE\Big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big]$$

$s(b_0)$ is the **Standard Error (or S.E.)** of b_0, (or called Standard deviation)

Sample distribution of $(b_0 - \beta_0)/s(b_0)$

$$\frac{b_0 - \beta_0}{s(b_0)} \text{ follows } t(n-2) \text{ for model (1)}$$

Confidence interval for β_0: with confidence $1 - \alpha$, we have the confidence interval

$$b_0 - t(1 - \alpha/2, n-2) * s(b_0) \leq \beta_1 \leq b_0 + t(1 - \alpha/2, n-2) * s(b_0)$$

Test of β_0

- Two-sided Test: to check whether β_1 is 0

$$H_0 : \beta_0 = 0, \quad H_a : \beta_0 \neq 0$$

Under H_0, we have

$$t = \frac{b_0}{s(b_0)} \sim t(n-2)$$

Suppose the **significance level** is α (usually, 0.05, 0.01). If the calculated t, say t^*

  - If $|t^*| \leq t(1 - \alpha/2; n-2)$, then accept H_0.
  - If $|t^*| > t(1 - \alpha/2; n-2)$, then reject H_0.

Similarly, the test can also be done based on the **p-value**, defined as $p = P(|t| > |t^*|)$.

It is easy to see that

$$\text{p-value} < \alpha \Longleftrightarrow |t^*| > t(1 - \alpha/2; n-2)$$

Thus

  - If p-value $\geq \alpha$, then accept H_0.

– If p-value $< \alpha$, then reject H_0.

- One-sided test:to check whether β_1 is positive (or negative)

$$H_0 : \beta_0 \leq 0, \quad H_a : \beta_0 > 0$$

Example 4.4 For the example above, with significance level 0.05,

1. Test $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$
2. Test $H_0' : \beta_1 = 0$ versus $H_1' : \beta_1 \neq 0$
3. Test $H_0'' : \beta_0 \geq 0$ versus $H_1'' : \beta_0 < 0$

Answer:

1. since $n = 10, t(0.975, 8) = 2.306$. $|t^*| = 1.432 < 2.306$. Thus, we accept H_0

   (another approach: p-value $= 0.1899 > 0.05$, we accept H_0)

2. The t-value is $|t^*| = 5.862 > 2.306$, thus we reject H_0', i.e. b_1 is significantly different from 0.

   (another approach: p-value $= 0.000378 < 0.05$, we reject H_0')

3. $t(0.05, 8) = -1.86$, since $t^* = 1.3931 > -1.86$ we accept H_0''

How to find these test from the output of the R code?