STA238 Tutorial 8

Luis Ledesma

2023-03-22

1 Announcements

- You can upload your work on Crowdmark from the end of the tutorial session to 5pm Friday of that week.
- All questions must be solved using RStudio.

2 Recall: Last tutorial

Last tutorial: We fitted a linear regression model to a dataset, and we interpreted some fitted values and derived statistics from the model, along with visualizations of the data.

Main takeaways:

1. The function `lm()` will be used to fit linear regression models, and with `summary()` one can extract the parameter estimates, along with their associated standard errors.
2. We obtained parameter estimates from the fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and the coefficient of determination R^2.
3. We visualized the data and fitted a least squares line on top of it.

3 Tutorial activity

For this tutorial, we want to:

1. Carry out more regression model fitting.
2. Understand the hypothesis tests associated with regression models, along with their associated test statistics and rejection regions.
3. Compute confidence intervals for the parameter fits.
4. Carry out linear regression by hand for simpler models.

3.1 Predicting runoff volume using rainfall volume

3.1.1 Visualizing the dataset

We will code the dataset as follows:

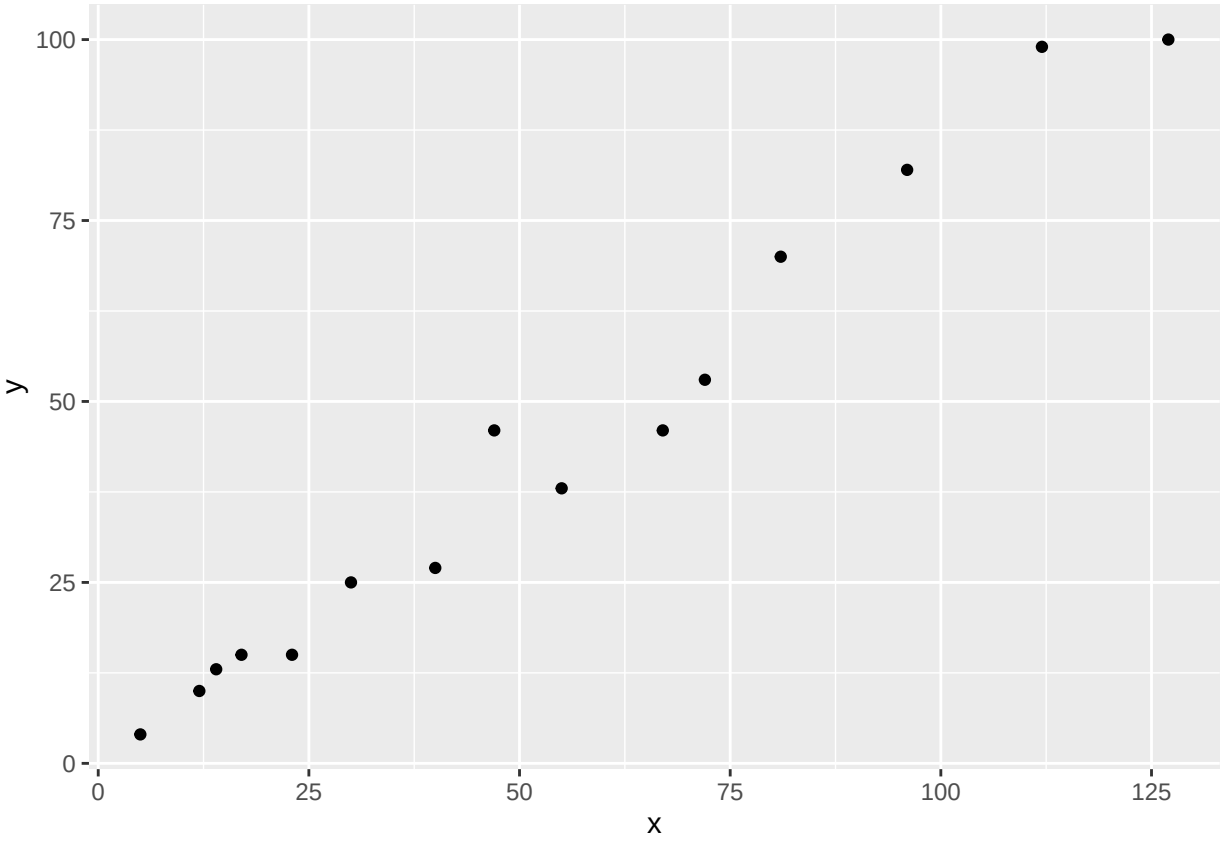
```
x <- c(5, 12, 14, 17, 23, 30, 40, 47, 55, 67, 72, 81, 96, 112, 127)
y <- c(4, 10, 13, 15, 15, 25, 27, 46, 38, 46, 53, 70, 82, 99, 100)
dataQ1 = data.frame(x = x, y=y)
```

The independent variable will be rainfall volume (in m^3), and we want to determine whether it is a predictor of runoff volume (in m^3) for a particular location. Our proposed model will be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Using `ggplot2`, one can visualize the data:

```
ggplot(dataQ1, aes(x, y)) +
  geom_point()
```



It does seem reasonable that linear regression can be used to fit the data.

3.1.2 Model fits, point estimates and coefficients of determination

Now, to look at point estimates of the data:

```
model1 <- lm(y ~ x, data = dataQ1)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = dataQ1)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.279 -4.424  1.205  3.145  8.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.12830    2.36778  -0.477    0.642
## x            0.82697    0.03652  22.642  7.9e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.24 on 13 degrees of freedom
## Multiple R-squared:  0.9753, Adjusted R-squared:  0.9734
## F-statistic: 512.7 on 1 and 13 DF,  p-value: 7.896e-12
```

```
model1
```

```
##
## Call:
## lm(formula = y ~ x, data = dataQ1)
##
## Coefficients:
## (Intercept)            x
##      -1.128        0.827
```

The estimate of the intercept will be $\hat{\beta}_0 = -1.128$ and the estimate of the slope term will be $\hat{\beta}_1 = 0.827$. To determine the point estimate of the true average of runoff volume when rainfall volume is 50:

$$\hat{y} = -1.128 + 0.827 \cdot 50$$

Alternatively, one could use `predict(model1,newdata=c(50))`.

The **point estimate of the standard deviation will be the residual standard error, 5.24** (from the model output). Similarly, the **coefficient of determination R^2 explains the proportion of variability in y due to x, $R^2 = 0.9753$.**

3.2 Computing linear regression estimates and confidence intervals manually

Note: While all of the following calculations can be done in R, you have to solve this problem by hand. You can use R to verify your answers.

The method of least squares will give the following estimates for $\hat{\beta}_0, \hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

We need to compute:

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 111 - \frac{23^2}{7} = 35.428$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 81 - \frac{23 \cdot 18}{7} = 21.857$$

$$\overline{y} = \frac{\sum y_i}{7} = \frac{18}{7} = 2.571$$

$$\overline{x} = \frac{\sum x_i}{7} = \frac{23}{7} = 3.286$$

Then, we can plug these into:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{21.857}{35.428} = 0.617$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 2.571 - (0.617 \cdot 3.286) = 0.544$$

Then, the least squares regression line will be:

$$\hat{y} = 0.544 + 0.617x$$

3.2.1 Hypothesis tests for regression coefficients

Suppose that we want to test whether the data provides sufficient evidence that x is a significant predictor of y. Using our model, the hypothesis test will be:

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

And, the test statistic will be:

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

Where $S_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$. **Under the null hypothesis, the above test statistic will be a t-distribution with $n - 2 = 7 - 2 = 5$ degrees of freedom**.

3.2.2 Computing the test statistic

For the above test statistic:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \quad SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

We need to compute:

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 62 - \frac{18^2}{7} = 15.714$$

Plugging in the values:

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 15.714 - \frac{21.857^2}{35.428} = 2.23$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{2.23}{5} = 0.446 \Rightarrow \hat{\sigma} = \sqrt{0.446} = 0.6678$$

$$S_{\hat{\beta}_1} = \frac{0.6678}{\sqrt{35.428}} = 0.1122$$

Then, the computed test statistic will be:

$$t_c = \frac{0.617}{0.1122} = 5.5$$

Recall that under the null hypothesis, the test statistic follows a t-distribution with 5 degrees of freedom.

3.2.3 Carrying out the hypothesis test

As the hypothesis test is two-sided, the rejection region will be $RR = \{|t_c| > t_{0.025,df=5}\}$. Since $t_{0.025,5} = 2.571$:

$$t_c = 5.5 > 2.571$$

Then, there is **enough evidence to reject H_0, which implies that there is a linear relationship between x and y.**

3.2.4 Computing confidence intervals

The 95% confidence interval for $\hat{\beta}_1$ will be:

$$\hat{\beta}_1 \pm t_{0.025,5} \cdot S_{\hat{\beta}_1} = 0.617 \pm 0.288 = (0.329, 0.905)$$