STM4PSD – Workshop 12

Linear regression in R

For the questions in this lab we will work through some analyses of data collected from the La Trobe University childcare centre solar generation system. The data consists of several variables. The variables that we will be focussing on is total yield (kWh) as the response and two explanatory variables. The first, which we consider below, is the number of hours of bright sunshine for the day. The second, to be considered later, is the maximum temperature for the day in Celsius. This data set can be found on the LMS in a file called `Solar.csv`. Save the data to your computer and then load it in R.

1. Execute the following command to obtain least squares estimates and associated output in R.

```
lm.model <- lm(Yield ˜ Sun, data = Solar)
summary(lm.model)
```

(a) The least squares estimates are in the column called `Estimate`. The first is the estimate for β_0 and the second, for β_1, is for the coefficient to explanatory variable Sun. Write down the estimated regression model.

(b) What is the estimate of the coefficient for the explanatory variable? Interpret this estimated coefficient.

(c) The test statistics for the test of the hypotheses

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0$$

are in the column `t value`. The corresponding p-values are in the column `Pr(>|t|)`. Use the p-value to carry out the test of

$$H_1 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

Do you reject H_0 at the $\alpha = 0.05$ level of significance? Explain.

(d) Now, what does your findings from the above test tell you? Can you come up with a simple statement that summarises the findings of the test?

(e) The coefficient of determination (R^2), is labeled as `Multiple R-squared` in the output. Does this R^2 suggest that the regression model fits the data well? Explain.

(f) Observations with missing values have been removed from the analysis leaving us with a sample size of $n = 322$. The standard errors for the estimates are in the column `Std. Error`. Using the formula from Readings 10.4, calculate a 95% confidence interval for β_1.
You can use the command `qt(0.975, df = 320)` to obtain the value for $t_{320,0.975}$.

Multiple linear regression

In many cases we may have more than one explanatory variable. Fortunately, the method of least squares generalises to this setting easily where the **multiple linear regression model** is defined to be

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

based on the p explanatory variables $x_1, \ldots, x_p$. The same conditions hold for the error and interpretation of the estimated coefficients and the coefficient of determination remains the same. Additionally, we check for model violations using residuals versus fits plots and Q-Q plots in the same way too. A 95% confidence interval for β_j is

$$\widehat{\beta}_j \pm t_{n-p-1,0.975} \times \text{SE} \tag{1}$$

where $\widehat{\beta}_j$ is the least squares estimate of β_j and SE is the corresponding standard error.

2. We will now consider an additional explanatory variable denoted the maximum temperature reached (in Celsius). For this question we will also inspect the residual versus fits plot and Q-Q plot. Execute the following command to obtain least squares estimates and associated output in R.

```
lm.model.2 <- lm(Yield ~ Sun + MaxTemp, data = Solar)
summary(lm.model.2)
windows(width = 8, height = 4)
par(mfrow = c(1, 2))
plot(lm.model.2, which = 1:2) # R plots four diagnostic plots.  We focus on just two.
```

(a) Do you think the residuals versus fits plot or the Q-Q plot of the residuals suggest that there are any linear regression model violations that we need to be concerned with? Justify your answer with reference to both of the plots.

(b) Does the R output suggest that the regression model fits the data well?

(c) What is the estimate of the coefficient for the maximum temperature explanatory variable? Interpret this estimated coefficient.

(d) Let β_2 denote the true coefficient for the maximum temperature explanatory variable and consider the hypotheses

$$H_0 : \beta_2 = 0 \text{ versus } H_1 : \beta_2 \neq 0.$$

Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? Explain.

(e) After removing observations with missing values, the sample size used is now $n = 320$. Using the formula in (1), construct a 95% confidence interval for β_2. You can use the command `qt(0.975, df = 317)` to obtain the value for $t_{317,0.975}$.

Confidence intervals and prediction intervals for the response

Suppose that we have an estimated regression model denoted

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p$$

which is an estimate of $E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. Suppose we want to estimate the response for set vales of the explanatory variables given as

$$x_1 = x_{01}, x_2 = x_{02}, \ldots, x_p = x_{0p}.$$

Then an estimate for the response is

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \ldots + \widehat{\beta}_p x_{0p}.$$

There are two types of intervals that we may be interested in. They are:

- An interval estimate for $E(Y) = \beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p}$. That is, an interval estimate for the *mean response associated*. In this context we refer to these intervals as **confidence intervals** as we usually do for interval estimators for model parameters.
- An interval estimate for $Y = \beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p} + \epsilon$. This interval is for a single future response, not the mean, and therefore needs to account for error variability as well. We therefore refer to these intervals as **prediction intervals**.

3. Continuing the previous question, we will now obtain confidence and prediction intervals for the response for a set value of the explanatory variables.

(a) Run the below commands in R. This will create a new data frame with values of the explanatory variables maximum temperature and hours of sunshine set to 23 and 1 respectively.

```
new.data <- data.frame(MaxTemp = 23, Sun = 1)
new.data
```

(b) Run the command below to obtain a 95% confidence interval for the mean yield using these explanatory variables values.

```
predict(lm.model.2, new.data, interval = "confidence")
```

Displayed are the fit followed by the 95% confidence interval. Try to come up with as simple statement about your findings from this output.

(c) Run the command below to obtain a 95% prediction interval for a future yield using these explanatory variables values.

```
predict(lm.model.2, new.data, interval = "prediction")
```

Displayed are the fit followed by the 95% prediction interval. Try to come up with as simple statement about your findings from this output. The statement provided in the Overleaf solutions gives a good example.

(d) In the context of the problem, provide a justification as to why these intervals are different.

(e) Now repeat the above but this time set the sun explanatory variable to 10. Have the intervals changed much? Do your intervals suggest that hours of sun plays a big role in solar yield?