Lab

Bivariate Data and Linear Regression

Lab objectives:

- To become familiar with bivariate statistics
- To calculate and visualise linear regression functions

Getting started

First download the Lab Data from CampusMoodle. *DrillData.csv* contains data on the drilling time (days) and depth (m) for 20 wells. We seek to examine the relationship between these two variables.

Start R Studio. Clear your workspace (obtained from the Session menu). Set your working directory to the location in which you have placed *DrillData.csv*.

First Look at the Data

Read the data into a dataframe and examine it. There isn't much data, so it can all be shown on screen at once. With more data, commands like **head**, **tail**, **summary** and **str** are useful.

```
drill.data <- read.csv("DrillData.csv", header = T,
stringsAsFactors=T)
drill.data
```

The WellNo is just a label, so we have two variables of interest, Depth and DrillingTime. We could look at each of them as univariate distributions, using the commands from last week's lab. However, we are naturally interested in the relationship between these variables. First, calculate the covariance (this gives the *sample covariance*) and correlation between the two variables.

```
cov(drill.data$Depth,drill.data$DrillingTime)
```

or using the pipe (`|>`)notation

```
drill.data |> select(Depth, DrillingTime) |> cov()
```

Note that the output is presented in a slightly different way, but the value of interest is the same. The numbers in the the diagonal (every variable against itself) vill have positive covariance. Here, there is only one value of interest in the table, the covariance between the 2 different variables.

Calculating the correlation coefficient (2 options). The default method is persons.

```
cor(drill.data$Depth,drill.data$DrillingTime)
cor(drill.data[2],drill.data[3])
```

Alternatively, using pipes

```
drill.data |> select(Depth, DrillingTime) |> cor()
```

Note: If the scatterplot visualisation shows outliers or is indicative or a non-linear relationship between the 2 variables, then the value of Pearson’s correlation coefficient can be misleading. In such situations, or when one or both of the variables are on an ordinal scale, Spearman's rank correlation coefficient should be used by stating method=”spearman”. For example,

```
cor(drill.data$Depth,drill.data$DrillingTime, method =
“spearman”)
```

or

```
drill.data |> select(Depth, DrillingTime) |>
cor(method="spearman")
```

We can create a scatterplot as in previous labs (remember to load the appropriate library).

```
p <- ggplot(drill.data,aes(x=Depth,y=DrillingTime))
p <- p + geom_point()
p <- p +  labs(title = "Time vs Drilling Time",
               x="Depth (m)",
               y = "Drilling time (days)")
p
```

The variables have a strong linear correlation and the graph looks reasonably linear, so we shall proceed to a linear regression.

Linear Regression

We can obtain the coefficients of a linear regression using the `lm` command. More information can be obtained by assigning the output to a variable, as follows.

```
lm.output<-lm(formula=DrillingTime~Depth,data=drill.data)
```

Notice the use of ‘formula’ to define the model we are fitting. DrillingTime is to be regressed on Depth. The ‘data’ argument tells R that these variables are both columns of the dataframe drill.data.

Now we can examine the object lm.output piece by piece. For example, we can obtain the coefficients from lm.output$coefficients. Applying `names` to `lm.output` gives a full list of possibilities. `summary` can also be applied to this output, giving a lot of useful information.

Summary Explained

```
summary(lm.output)

Call:
lm(formula = DrillingTime ~ Depth, data = drill.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.9714  -0.8996   0.4538   4.9692   7.5699 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -15.398101   4.138643  -3.721  0.00157 ** 
Depth         0.024246   0.002144  11.307 1.31e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.967 on 18 degrees of freedom
Multiple R-squared:  0.8766,    Adjusted R-squared:  0.8697 
F-statistic: 127.9 on 1 and 18 DF,  p-value: 1.306e-09
```

Here, the *Intercept Estimate* (15.398101) is β_0 and the Depth Estimate (slope) (0.024246) is β_1. Hence, the equation (in the form $y = \beta_0 + \beta_1 x$, where y is the drilling time and x is the depth) is

```
      DrillingTime = -15.398101 + 0.024246 * Depth
```

The summary output shows the original call to `lm()` and a 5-number summary of the residuals (we already know their mean is zero). We could look at the residuals in more detail, but we can already see from the quartiles that the distribution does not look normal.

To look at the residuals:

```
lm.output$residuals
```

The coefficients table contains the coefficients (Estimate column), the corresponding standard errors (a measure of the uncertainty of the coefficient estimates) and some information on the statistical significance of the results. Note, in particular, the very small Pr(>|t|) for depth. This is the probability that the linear coefficient is zero, equivalent to drilling time being independent of depth. Had we chosen to test our model against that null hypothesis at the

0.001 significance level (same as 99.9% confidence), we could reject the null hypothesis. That is indicated by the '***' at the end of the depth row. The probability that the intercept is zero is somewhat larger (0.00157) and this is indicated by the '**'.

'Multiple R-squared' is the R^2 as defined in lectures; 'Adjusted R-squared' is a more conservative measure of the model's effectiveness.

The p-value corresponding to the F- statistic indicates the significance of the whole model. For a linear regression on one variable, this is the same as the `Pr(>|t|)` for the variable. Although most people look at the R^2 first, it is more sensible to check the `p-value` first. If the `p-value` (from an F-test) shows that our result is not significant, the rest of the values should be treated with suspicion.

Predictions and Residual Plots

We can make a prediction for one value or many at the same time. For a depth of 1700 metres, as before, we can obtain a prediction for drilling time as follows:

```
newdata  <- data.frame(Depth=1700)
predict(lm.output,newdata)
```

The first argument to `predict()` is our stored model, the second the data we want predictions for. If we omit the second argument, predict() returns fitted values for the original data (the twenty depth values we used to build the model).

We can plot the residuals against the depth, the drilling time and versus a theoretical normal distribution. First, we construct the dataframe.

First, we create a dataframe with the 2 variables and the residual.

```
lmData <-
data.frame(residuals = lm.output$residuals,
                Depth=drill.data$Depth,
               DrillingTime= drill.data$DrillingTime)
```

To obtain the residuals against the DrillingTime

```
p <- ggplot(lmData, aes(x=DrillingTime, y = residuals))
p <- p + geom_point(size=2, colour="red")
p <- p +  theme_classic()
p <- p + theme(text = element_text(size = 20))
p
```

The points should look random.

To obtain the residuals against the Depth

```
p <- ggplot(lmData, aes(x=Depth, y = residuals))
p <- p + geom_point(size=2, colour="red")
p <- p +  theme_classic()
p <- p + theme(text = element_text(size = 20))
p
```

The points should look random.

To obtain the Q-Q plot (residuals against a theoretical normal distribution)

```
p <- ggplot(lmData, aes(sample = residuals))
p <- p + stat_qq(size=2) +
      stat_qq_line( alpha = 0.9, color = "red",
                   linetype = "dashed")
p <- p +  theme_classic()
p <- p + theme(text = element_text(size = 20))
p
```

The points should be close to the theoretical line.

We can see that the residuals do not obey the model assumptions very well.

We can test whether the residuals distribution is normal using the Shapiro-Wilk test. A p-value > 0.05 would indicate that it is reasonable to assume that the distribution is normal.

```
shapiro.test(lmData$residuals)
```

It can be seen that it is not reasonable to assume that the distribution is normal.

Adding an LR line to a Scatterplot

We can add an LR line to the data, using the smoothing facility. The default smoothing is a 'loess' curve, so we have to alter that.

```
p <- ggplot(drill.data,aes(x=Depth,y=DrillingTime))
p <- p + geom_point() + stat_smooth(method=”lm”)
```

We can remove the grey confidence band by setting se to F:

```
p <- p + geom_point() + stat_smooth(method=”lm”,se=F)
p
```

The graph can be modified in all the usual ways (title, labels, colours, etc).

Exercises

1. Above you have built a linear regression model of drilling time on depth and predicted the drilling time for 1700 metres depth. Construct the linear regression model of depth on drilling time and estimate the depth that can be drilled in the time that you got from 1700 depth . Do you get 1700 metres?
2. The loginTimes dataset contains data collected over 25 days. The data includes the number of users, average time to login (s), average time to access email (s), average time between mail messages received (s) and room temperature (degrees centigrade). Investigate relationships between pairs of variables. If you obtain a linear regression equation, comment on the p-value.
    a. Is there a relationship between time to login (response) and time between messages (predictor)? If there is, what is it?
    b. Can you predict the time to login if the time between messages is 200?
    c. Can you predict the time to login if the time between messages is 20 seconds?
3. Anscombe's quartet is available in R – just type 'anscombe' to see the data. Try applying linear regression to each of the four data sets (x1 pairs with y1, etc.). Does any of the information returned distinguish the sets?
4. Explore the Advertising.csv dataset. Are there any linear relationships between the various advertising methods and sales?
5. Fourteen students were employed to test a new information retrieval tool which is intended to locate relevant passages in electronic books. The time (minutes) taken to perform an indexing task and the recall score (%) for each student is as follows

Subject	Time	Recall
1	9.2	61
2	13.3	53
3	8	57
4	11.9	50
5	12.1	53
6	8.9	54
7	10.8	54
8	7.2	65
9	14.9	46
10	7.3	87
11	6.3	88
12	10.4	52
13	13.8	45
14	7.6	72

Investigate whether there is a linear relationship between time and recall. Which correlation coefficient should you use?