Appendix

**Visualization**

* Create a scatter plot displaying the total cases versus new cases
* Add a line

library(ggplot2)

ggplot(DatasSet, aes(x='New Cases',y='Total Cases'))+

geom\_point()+

stat\_smooth(method =lm)

plot(`New Cases`,`Total Cases`)

abline(mod)

abline(mod,col=2,lwd=3)

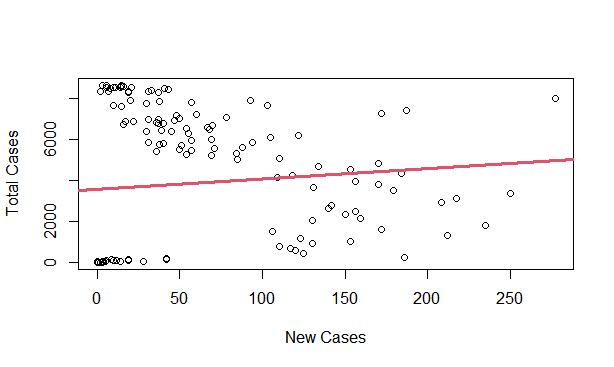


Figure 1 – Scatter graph of total cases versus new cases

The graph above suggests a linear increasing relationship between the total cases and the new cases variables. This is a good thing because one important assumption of the linear regression is that the relationship between the outcome and predictor variables is linear and additive.

It’s also possible to compute the correlation coefficient between the two variables using the R function cor():

cor(`New Cases`,`Total Cases`)

## [1] 0.09871877

The correlation coefficient measures the level of the association between two variables x and y. Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases).

A value closer to 0 suggests a weak relationship between the variables. A low correlation (-0.2 < x < 0.2) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x(=). In such case, we should probably look for better predictor variables.

In this case study, the correlation coefficient is large enough, so we can continue by building a linear model of y as a function of x.

**Computation**

The simple linear regression tries to find the best line to predict total cases on the new cases.

The linear model equation can be written as follow: total cases = b0 + b1 \* new cases

The R function 1m() can be used to determine the beta coefficients of the linear model:

mod <- lm(`Total Cases`~`New Cases`)

##

## Call:

## lm(formula = `Total Cases` ~ `New Cases`)

## Coefficients:

## (Intercept) `New Cases`

## 3561.612 5.080

The results show the intercept ad the beta coefficient for the new cases variable.

**Interpretation**

From the output above,

* the estimated regression line equation can be written as follow:

Total cases = 3561.612 + 5.080 \* new cases

* the intercept (b0) is 3561.612. It can be interpreted as the predicted total cases unit for a zero new cases.
* the regression beta coeffiecient for the variable new cases (b1), also known as the slope, is 5.080.

**Regression line**

To add the regression line onto the scatter plot, you can use the function stat\_smooth() [ggplot2]. By default, the fitted line is presented with confidence interval around it. The confidence bands reflect the uncertainty about the line. If you don’t want to display it, specify the option se = FALSE in the function stat\_smooth().

library(ggplot2)

ggplot(DatasSet, aes(x='New Cases',y='Total Cases'))+

geom\_point()+

stat\_smooth(method =lm)

plot(`New Cases`,`Total Cases`)

abline(mod)

abline(mod,col=2,lwd=3)

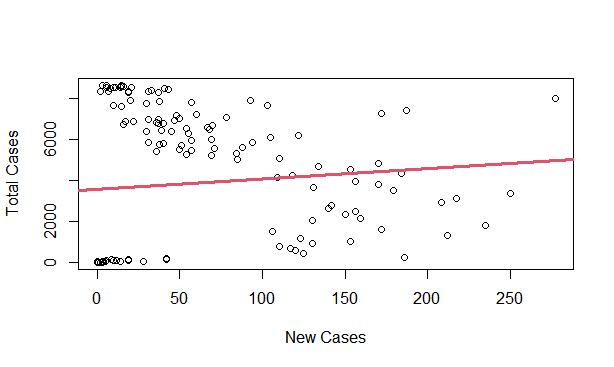


Figure 2 – Regression Line of the case study

**Model Assessment**

In the previous section, we built a linear model of sales as a function of new cases : total cases = b0 + b1 \* new cases

Before using this formula to predict future record, you should make sure that this model is statistically significant, that is:

* there is a statistically significant relationship between the predictor and the outcome variables
* the model that we built fits very well the data in our hand.

In this section, we’ll describe how to check the quality of a linear regression model.

Model summary

We start by displaying the statistical summary of the model using the R function summary():

summary(mod)

## Call:

## lm(formula = `Total Cases` ~ `New Cases`)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4268.5 -3544.7 3.6 3083.1 5010.1

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3561.612 358.920 9.923 <2e-16 \*\*\*

## `New Cases` 5.080 4.167 1.219 0.225

## ---

## Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

##

## Residual standard error: 3325 on 151 degrees of freedom

## Multiple R-squared: 0.009745, Adjusted R-squared: 0.003187

## F-statistic: 1.486 on 1 and 151 DF, p-value: 0.2247

The summary outputs shows 6 components, including:

* **Call.** Shows the function call used to compute the regression model.
* **Residuals.** Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero and the minimum and maximum should be roughly equal in absolute value.
* **Coefficients.** Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
* **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits our data.

Coefficients significance

The coefficients table, in the model statistical summary, shows:

* The estimates of the **beta coefficients**
* The **standard errors** (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
* The **t-statistic** and the associated **p-value**, which defines the statistical significance of the beta coefficients.

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3561.612 358.920 9.923 <2e-16

## `New Cases` 5.080 4.167 1.219 0.225

**t-statistic and p-value:**

For a given predictor, the t-statistic (and its associated p-value) tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable, that is whether or not the beta coefficient of the predictor is significantly different from zero.

The statistical hypotheses are as follow:

* Null hypothesis (H0): the coeffiecients are equal to zero (i.e., no relationship between x and y)
* Alternative hypothesis (Ha): the coefficients are not equal to zero (i.e., there is some relationship between x and y)

Mathematically, for a given beta coefficient (b), the t-test is computed as t = (b – 0)/SE(b), where SE(b) is the standard error of the coefficient b. The t-statistic will produce a small p-value. The higher the t-statistic (and the lower the p-value), the more significant the predictor. The symbols to the right visually specifies the level odf significance. The line below the table shows the definition of these symbols; one star means 0.01 < p > 0.05. The more the stars beside the variable’s p-value, the more significant the variable.

A statistically significant coefficient indicates that there is an association between the predictor (x) and the outcome (y) variable.

In our example, both the p-values for the intercept and the predictor variable are highly significant, so we can reject the null hypothesis and accept the alternative hypothesis, which means that there is a significant association between the predictor and the outcome variables.

The t-statistic is a very useful guide for whether or not to include a predictor in a model. High t-statistics (which go with low p-values near 0) indicate that a predictor should be retained in a model, while very low t-statistics indicate a predictor could be dropped (P. Bruce and Bruce 2017).

**Standard errors and confidence intervals:**

The standard error measures the variability/accuracy of the beta coefficients. It can be uses to compute the confidence intervals of the coefficients

For example, the 95% confidence interval for the coefficient b1 is defined as b1 +/- 2\*SE(b1), where:

* the lower limits of b1 = b1 - 2\*SE(b1) = 5.080 - 2\*4.167 = -3.254
* the upper limits of b1 = b1 + 2\*SE(b1) = 5.080 + 2\*4.167 = 13.414

That is, there is approximately a 95% chance that the interval [-3.254, 13.414] will contain the true of b1. Similarly the 95% confidence interval for b0 can be computed as b0 +/- 2\*SE(b0).

To get these information, simply type:

confint(mod)

## 2.5 % 97.5 %

## (Intercept) 2852.45924 4270.7649

## `New Cases` -3.15377 13.3142

Model accuracy

Once you identified that, at least one predictor variable is significantly associated the outcome, you should continue the diagnostic by checking how well the model fits the data. This process is also referred to as the *goodness-of-fit*

The overall quality if the linear regression fit can be assessed using the following three quantities, displayed in the model summary

1. The Residual Standard Error (RSE)

2. The R-squared (R2)

3. F-statistic

## rse r.squared f.statistic p.value

## 1 3325 0.009745 1.486 0.2247

1. **Residual standard error** (RSE).

The RSE (also known as the model sigma) is the residual variation, representing the average variation of the observations points around the fitted regression line. This is the standard deviation of residual errors.

RSE provides an absolute measure of patterns in the data that can’t be explained by the model. When comparing two models, the model with the small RSE is a good indication that this model fits the best the data.

Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible.

In this case study, RSE = 3325, meaning that the observed sales values deviate from the true regression line by approximately 33225 units in average.

Whether or not an RSE of 3325 units is an acceptable prediction error is subjective and depends on the prolem context. However, we can calculate the percentage error.

In our data set, the mean value of total cases is 55.81, and so the percentage error is 3325/55.81 = 60%.

sigma(mod)\*100/mean(DatasSet$`Total Cases`)

## [1] 86.33449

2. **R-squared and Adjusted R-squared:**

The R-squared (R2) ranges from 0 to 1 and represents the proportion of information (i.e. variation) in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom.

The R2 measures, how well the model fits the data. For a simple linear regression, R2 is the square of the Pearson correlation coefficient.

A high value of R2 is a good indication. However, as the value of R2 tends to increase when more predictors are added in the model, such as in multiple linear regression model, should mainly consider the adjusted R-squared, which is a penalized R2 for a higher number of predictors.

* An (adjusted) R2 that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
* A number near 0 indicates that the regression model did not explain much of the variability in the outcome.
* In our dataset, the R-squared is 0.009745. This shows that new cases explains 0.97% of the variation in total cases.

1. 3. **F-Statistic**:

The F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient.

In a simple linear regression, this test is not really interesting since it just duplicates the information in given by the t-test, available in the coefficient table. In fact, the F test is identical to the square of the t test: 1.486 = (1.219^2. This is true in any model with 1 degree of freedom.

The F-statistic becomes more important once we start using multiple predictors as in multiple linear regression.

A large F-statistic will corresponds to a statistically significant p-value (p<0.05).

In this case study, the F-statistic equal 1.486 producing a p-value of 0.2247, which is highly significant. As an overall, this shows that the factor we are using i.e. new cases are very relevant.

Summary

After computing a regression model, a first step is to check whether, at least, one predictor is significantly associated with outcome variables.

If one or more predictors are significant, the second step us to assess how well the model fits the data by inspecting the Residuals Standard Error (RSE), the R2 value and the F-statistics. These metrics give the overall quality of the model.

* RSE: Closer to zero the better
* R-Squared: Higher the better
* F-statistic: Higher the better

**Correlation**

Represent the scatterplot graph :

plot(`Total Cases`,`New Cases`, main="Scatterplot", las=1)

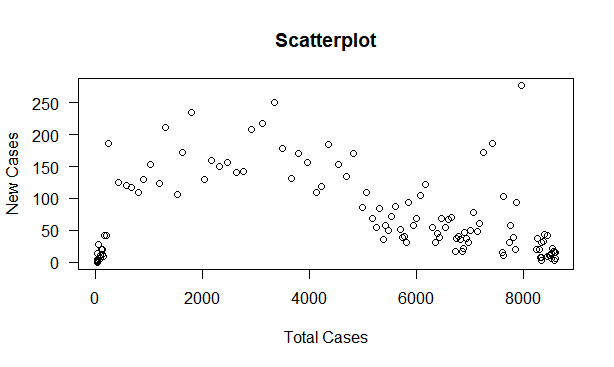


Figure 3 – Correlation of the case study

Calculate the value of correlation using **Pearson’s Product-Moment Correlation Coefficient**.

Assumes normality in both variables. There needs to be a linear relationship between the two variables. Two variables should be measured at the interval or ratio level.

It is sensitive to outliers, can have a large effect on the line of best fit and the Pearson correlation coeffient, leading to very difficult conclusions regarding the data.

cor(`Total Cases`,`New Cases`,method="pearson")

## [1] 0.09871877

cor.test(`Total Cases`,`New Cases`,method="pearson")

## Pearson's product-moment correlation

## data: Total Cases and New Cases

## t = 1.219, df = 151, p-value = 0.2247

## alternative hypothesis: true correlation is not equal to 0

## 95 percent confidence interval:

## -0.06091354 0.25342709

## sample estimates:

## cor

## 0.09871877

Suppose to be the result should be seen the total cases increases as the new cases increases. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the new cases and total cases.

**One Sample Test**

Represent the boxplot graph:

boxplot(Active, main="Boxplot")

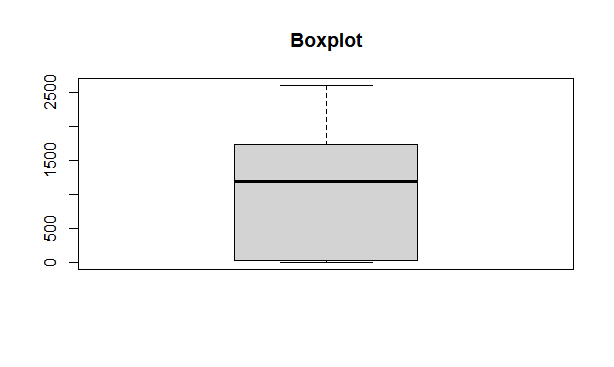


Figure 4 – One Sample Test of the case study

The output as follows:

t.test(Active, mu=0, alternative = "two.sided", level=0.95)

## One Sample t-test

##

## data: Active

## t = 14.69, df = 152, p-value = 1

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

## -Inf 1143.083

## sample estimates:

## mean of x

## 1027.346

So here we use two side alternative hypothesis. We found that the mean is not equal to 0. And for the confidence interval level we use 95%. For the result we got the mean of x is 1027.346.

**Chi-Square Test**

table(Active,Recoveries)

TAB = table(Active,Recoveries)

barplot(Active,Recoveries)

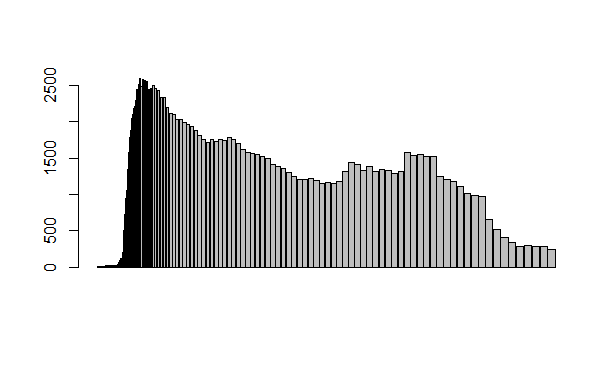


Figure 5 – Chi-Square Test of the case study

The output are as follows:

chisq.test(TAB, correct=T)

CHI = chisq.test(TAB, correct=T)

CHI

## Pearson's Chi-squared test

##

## data: TAB

## X-squared = 16673, df = 14274, p-value < 2.2e-16

attributes(CHI)

## $names

## [1] "statistic" "parameter" "p.value" "method" "data.name" "observed"

## [7] "expected" "residuals" "stdres"

##

## $class

## [1] "htest"