## Loading required R packages

The following R packages are required for this chapter:

* tidyverse for data manipulation and visualization

**library**(tidyverse)

## Examples of data

We’ll use the marketing data set [datarium package], which contains the impact of the amount of money spent on three advertising medias (youtube, facebook and newspaper) on sales.

First install the datarium package using devtools::install\_github("kassmbara/datarium"), then load and inspect the marketing data as follow:

data("marketing", package = "datarium")

**head**(marketing, 4)

## youtube facebook newspaper sales

## 1 276.1 45.4 83.0 26.5

## 2 53.4 47.2 54.1 12.5

## 3 20.6 55.1 83.2 11.2

## 4 181.8 49.6 70.2 22.2

## Building model

We want to build a model for estimating sales based on the advertising budget invested in youtube, facebook and newspaper, as follow:

sales = b0 + b1\*youtube + b2\*facebook + b3\*newspaper

You can compute the model coefficients in R as follow:

model <- lm(sales ~ youtube + facebook + newspaper, data = marketing)

summary(model)

##

## Call:

## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

##

## Residuals:

## Min 1Q Median 3Q Max

## -10.59 -1.07 0.29 1.43 3.40

##

## Coefficients:

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

## (Intercept) 3.52667 0.37429 9.42 <2e-16 \*\*\*

## youtube 0.04576 0.00139 32.81 <2e-16 \*\*\*

## facebook 0.18853 0.00861 21.89 <2e-16 \*\*\*

## newspaper -0.00104 0.00587 -0.18 0.86

## ---

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

##

## Residual standard error: 2.02 on 196 degrees of freedom

## Multiple R-squared: 0.897, Adjusted R-squared: 0.896

## F-statistic: 570 on 3 and 196 DF, p-value: <2e-16

## Interpretation

The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary.

In our example, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.

To see which predictor variables are significant, you can examine the coefficients table, which shows the estimate of regression beta coefficients and the associated t-statitic p-values:

summary(model)$coefficient

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

## (Intercept) 3.52667 0.37429 9.422 1.27e-17

## youtube 0.04576 0.00139 32.809 1.51e-81

## facebook 0.18853 0.00861 21.893 1.51e-54

## newspaper -0.00104 0.00587 -0.177 8.60e-01

For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero.

It can be seen that, changing in youtube and facebook advertising budget are significantly associated to changes in sales while changes in newspaper budget is not significantly associated with sales.

For a given predictor variable, the coefficient (b) can be interpreted as the average effect on y of a one unit increase in predictor, holding all other predictors fixed.

For example, for a fixed amount of youtube and newspaper advertising budget, spending an additional 1 000 dollars on facebook advertising leads to an increase in sales by approximately 0.1885\*1000 = 189 sale units, on average.

The youtube coefficient suggests that for every 1 000 dollars increase in youtube advertising budget, holding all other predictors constant, we can expect an increase of 0.045\*1000 = 45 sales units, on average.

We found that newspaper is not significant in the multiple regression model. This means that, for a fixed amount of youtube and newspaper advertising budget, changes in the newspaper advertising budget will not significantly affect sales units.

As the newspaper variable is not significant, it is possible to remove it from the model:

model <- lm(sales ~ youtube + facebook, data = marketing)

summary(model)

##

## Call:

## lm(formula = sales ~ youtube + facebook, data = marketing)

##

## Residuals:

## Min 1Q Median 3Q Max

## -10.557 -1.050 0.291 1.405 3.399

##

## Coefficients:

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

## (Intercept) 3.50532 0.35339 9.92 <2e-16 \*\*\*

## youtube 0.04575 0.00139 32.91 <2e-16 \*\*\*

## facebook 0.18799 0.00804 23.38 <2e-16 \*\*\*

## ---

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

##

## Residual standard error: 2.02 on 197 degrees of freedom

## Multiple R-squared: 0.897, Adjusted R-squared: 0.896

## F-statistic: 860 on 2 and 197 DF, p-value: <2e-16

Finally, our model equation can be written as follow: sales = 3.5 + 0.045\*youtube + 0.187\*facebook.

The confidence interval of the model coefficient can be extracted as follow:

confint(model)

## 2.5 % 97.5 %

## (Intercept) 2.808 4.2022

## youtube 0.043 0.0485

## facebook 0.172 0.2038

## Model accuracy assessment

As we have seen in simple linear regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

**R-squared**:

In multiple linear regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response (James et al. 2014). A solution is to adjust the R2 by taking into account the number of predictor variables.

The adjustment in the “Adjusted R Square” value in the summary output is a correction for the number of x variables included in the prediction model.

In our example, with youtube and facebook predictor variables, the adjusted R2 = 0.89, meaning that “89% of the variance in the measure of sales can be predicted by youtube and facebook advertising budgets.

Thi model is better than the simple linear model with only youtube (Chapter simple-linear-regression), which had an adjusted R2 of 0.61.

**Residual Standard Error** (RSE), or sigma:

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

The error rate can be estimated by dividing the RSE by the mean outcome variable:

sigma(model)/mean(marketing$sales)

## [1] 0.12

In our multiple regression example, the RSE is 2.023 corresponding to 12% error rate.

Again, this is better than the simple model, with only youtube variable, where the RSE was 3.9 (~23% error rate) (Chapter simple-linear-regression).

## Discussion

This chapter describes multiple linear regression model.

Note that, if you have many predictors variable in your data, you don’t necessarily need to type their name when computing the model.

To compute multiple regression using all of the predictors in the data set, simply type this:

model <- lm(sales ~., data = marketing)

If you want to perform the regression using all of the variables except one, say newspaper, type this:

model <- lm(sales ~. -newspaper, data = marketing)

Alternatively, you can use the update function:

model1 <- update(model, ~. -newspaper)