Statistics Bible Summary

# Modeling for exploration

Text

Description automatically generated

* Most crucially, looking at the raw data values.
* Computing summary statistics, such as means, medians, and interquartile ranges
* Creating data visualizations.

At the end, it should have:

* A sense of the distributions of the individual variables in your data
* Whether there are outliers and/or missing value
* Whether any potential relationships exist between variables

# Correlation calculation

The correlation coefficient is invariant to linear transformations.

Graphical user interface, text, application, email

Description automatically generated

A picture containing diagram

Description automatically generated

|  |  |
| --- | --- |
| Text  Description automatically generated | Text  Description automatically generated |

# Linear Regression

library(moderndive)

Minimum number of observations per coefficient

Graphical user interface, text, application, email

Description automatically generated

## Model Creation

### 1 numerical explanatory variable

|  |  |
| --- | --- |
| Graphical user interface, text, application  Description automatically generated | Chart, scatter chart  Description automatically generated |

We can use I() to elevate a variable.

Graphical user interface, text, application

Description automatically generated

### 1 categorical explanatory variable

We can also eliminate the Intercept by writing “+0” after the variable.

|  |  |
| --- | --- |
| Table  Description automatically generated with medium confidence | Table  Description automatically generated |

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

### 1 numerical and 1 categorical

**+** allows us to get one slope for both categories.

|  |  |
| --- | --- |
| Text  Description automatically generated with medium confidence | Chart, scatter chart  Description automatically generated |

**\*** allows us to get all model interactions. In this case we will have an interception and one slope for each categorical variable.

|  |  |
| --- | --- |
| Table  Description automatically generated | Chart, scatter chart  Description automatically generated |

We can get the same result by writing the interaction by hand.

Graphical user interface, text

Description automatically generated

We can limit interaction level using the next syntax

Graphical user interface, text, application

Description automatically generated

### 2 or more Numerical variables

*Taking into account all the other explanatory variables in our model*, for every increase of one unit in income (i.e., $1000), there is an associated decrease of on average $7.663 in debt.

|  |  |
| --- | --- |
| Text, table  Description automatically generated with medium confidence | Chart, scatter chart  Description automatically generated |
| A picture containing chart  Description automatically generated | |

### 2 Numerical and 1 categorical

Chart, scatter chart

Description automatically generated

## Model selection

We should try to keep the model as simple as possible. Here is an example where adding the interaction doesn’t affect the result too much.

|  |  |  |
| --- | --- | --- |
| Table  Description automatically generated | | Chart, scatter chart  Description automatically generated |
| A picture containing text  Description automatically generated | Text  Description automatically generated | |

Graphical user interface, text, application

Description automatically generated

## Math under regression hypothesis test

Diagram

Description automatically generated

Chart, diagram, box and whisker chart

Description automatically generated

|  |  |
| --- | --- |
| A picture containing chart  Description automatically generated | |
| A picture containing line chart  Description automatically generated | Line chart  Description automatically generated with medium confidence |

## Checking regression quality with ggfortify

Graphical user interface, text, application

Description automatically generated

## Making predictions with the model

Una variable predictiva

Graphical user interface, application

Description automatically generated

Varias variables

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, application

Description automatically generated

Una variable transformada

Text

Description automatically generated

Graficar las predicciones del modelo

Chart, scatter chart

Description automatically generated

## Simpson’s Paradox

Simpson’s Paradox occurs when trends that exist for the data in aggregate either disappear or reverse when the data are broken down into groups.

**To avoid problems**

* Articulate a question before you start modeling.
* Allow the question to select the model
* Try to plot the dataset y different ways
* Add more variables to the model and create charts

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Modeling the whole dataset suggests that playing more video games is related to a higher test score. If we reveal that each group represents the age of the child taking the test, it changes the interpretation. Now older children score more highly in the test, and playing lots of video games is related to a lower score.

## Extracting model components

### Base R

|  |  |  |
| --- | --- | --- |
| *Extraer los coeficientes* | *Extraer las predicciones con la data de entrenamiento* | *Extraer los residuales del modelo* |
| Graphical user interface, text, application, chat or text message  Description automatically generated | Graphical user interface, text, application  Description automatically generated | Graphical user interface, text, application, email  Description automatically generated |
| *Resumen del modelo con la data de entrenamiento* | | |
| Graphical user interface, text, application  Description automatically generated | | |

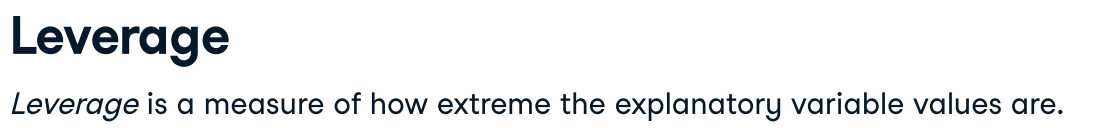
### Broom package

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated



A picture containing text

Description automatically generated

RSE == sigma

Leverage == .hat

Influence == .cooksd

Text

Description automatically generated with medium confidence

Text

Description automatically generated

# Sampling

|  |  |
| --- | --- |
| Table  Description automatically generated | * If the sampling of a sample of size *n* is done at ***random***, then * the ***sample is unbiased***and ***representative of the population*** of size N, thus * any result based on the sample can **generalize to the population**, thus * the ***point estimate is a “good guess”*** of the unknown population parameter, |
|  |  |

Shape

Description automatically generated with low confidence