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

**Statistical inference** is the act of making a guess about a population using a sample.

# 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

We can use poly to fit a polygon of defined degree:

**lm.fit5 <- lm(medv ~ poly(lstat, 5))**

### 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

### Selecting all variables except one

lm.fit1 <- lm(medv ~ . - age, data = Boston)

### Changing a model

lm.fit1 <- update(lm.fit, ~ . - age)

## Extracting model info

**contrasts(DF$var):** Returns the coding that `R` uses for the dummy variables

**coef(model):** Gets model coefficients.

**confint(model**, **parm, level = 0.95):** Calculate model confident intervals.

**predict.lm(model, newdata = DF, interval = c("none", "confidence", "prediction")):** Can calculate prediction with confidence or prediction intervals.

**summary(model)$r.sq:** Gives us the R2.

**summary(model)$sigma:** Gives us the RSE.

**summary(model)$coef:** Gives coefficients table and p-values.

**car::vif(model):** can be used to compute *variance inflation factors*.

**par(mfrow = c(2, 2)); plot(model):** Plot 4 diagnostic plots

**plot(DF$lstat, DF$medv); abline(model):** Plot a simple linear model.

## 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

**anova(model, model2):** Check is there is a statistical significant difference between two or more models.

## 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 |

## Conditions for inference for regression (LINE)

1. Linearity of relationship between variables (residual analysis)

Chart, scatter chart

Description automatically generated

1. Independence of the residuals (how the data was collected)

For example, given that the same professor taught these first four courses, it is reasonable to expect that these four teaching scores are related to each other.

1. Normality of the residuals (residual analysis)

Chart, histogram

Description automatically generated

1. Equality of variance of the residuals (residual analysis)

Chart, scatter chart

Description automatically generated

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 | | |

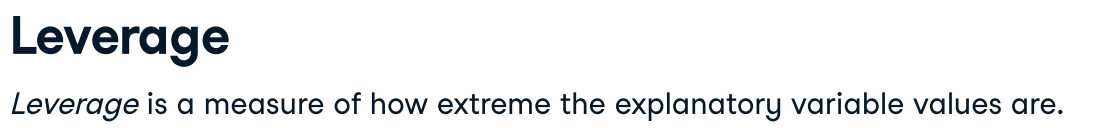
### Broom package

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

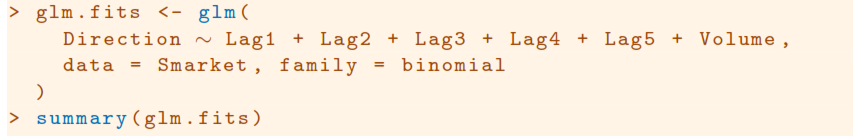
Description automatically generated

# Logistic or Poisson regressions

## Training

### Long Format

|  |  |
| --- | --- |
| Graphical user interface, application  Description automatically generated | Graphical user interface, application  Description automatically generated |



### Wider format

|  |  |
| --- | --- |
| Graphical user interface, application  Description automatically generated | Graphical user interface, application  Description automatically generated |

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

## Extracting model information

**predict.glm(model, type = “response”):** To output probabilities of the form P (Y =1|X), as opposed to other information such as the logit.

**table(predicted\_values, original\_values):** Creates a confusion matrix.

# Linear or Quadratic Discriminant Analysis (LDA or QDA)

## Model creation

A picture containing text

Description automatically generated

## Extracting information

**predict()** function returns a **list** with three elements.

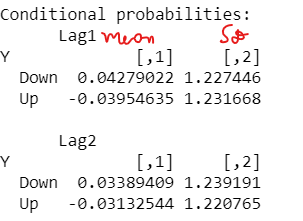
* **class**, contains LDA’s predictions about the movement of the market.
* **posterior**, is a matrix whose kth column contains the posterior probability that the corresponding observation belongs to the kth class
* **x,** contains the linear discriminants

# Naive Bayes

## Model Creation

This implementation of the naive Bayes classiﬁer models each quantitative feature using a **Gaussian distribution**.



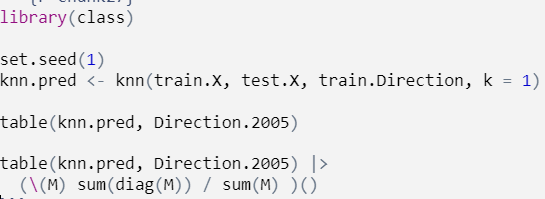


## Extracting information

**predict(model, new\_data, type = "class")**: Returns the predicted classes.

**predict(model, new\_data, type = "raw")**: Estimates of the probability that each observation belongs to a particular class.

# KNN



# Inference process

* *Any theory-based method is ultimately an approximation to the simulation-based method.*
* *The only assumption that needs to be met in the simulation-based method is that the sample is selected at random.*
* *We as authors much prefer the use of* ***confidence intervals*** *for statistical inference, since in our opinion they are much less prone to large misinterpretation*

## Possible scenarios

Graphical user interface, text, application, email

Description automatically generated

## Importance of 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 |

## Confidence interval

### Confidence level

To define the confident interval, we need to specify a ***confidence level***. Higher confidence levels tend to produce wider confidence intervals.

**Precise interpretation:** *If we repeated our sampling procedure many times, we expect about* ***95% of the resulting confidence intervals*** *to capture the value of the population parameter.*

|  |  |
| --- | --- |
| 95% Percentile | 80% SE |

### Sample size

|  |  |
| --- | --- |
|  |  |

## Bootstrap

If we take a sample and resample it with replacement, we can create ***Bootstrap Distribution of the Sample Mean*** which approximates the ***Sampling Distribution of the Sample Mean***. That will open the possibility to create ***confidence intervals*** of a sample, but it doesn’t affect the mean.

1. The ***bootstrap distribution*** will likely not have the same center as the ***sampling distribution***.
2. ***Bootstrapping*** will give you a good estimate of the standard error.

### Manual implementation

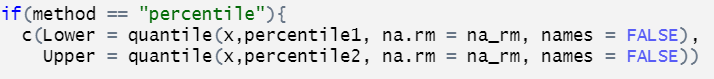
|  |  |
| --- | --- |
| Graphical user interface, text, application  Description automatically generated | Chart, histogram  Description automatically generated |

For example, we could be 95% “confident” that a 95% confidence interval captures the value of the population parameter.

A picture containing text

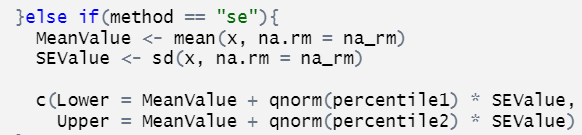
Description automatically generated

* Percentile method



* Standard error method

we can only this method when the bootstrap distribution is roughly normally shaped.



|  |  |
| --- | --- |
|  |  |

### Infer implementation

Diagram

Description automatically generated

|  |  |
| --- | --- |
|  |  |
|  | |

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

## Permutation testing

The term “permutation” is the mathematical term for “shuffling”: taking a series of values and reordering them randomly, as you did with the playing cards. While the bootstrap method involves resampling with replacement, permutation methods involve resampling without replacement.

* The ***null distribution*** is the sampling distribution of the test statistic assuming the null hypothesis H0 is true.
* The ***p-value*** is the probability of obtaining a *test statistic* just as extreme or more extreme than the observed test statistic assuming the null hypothesis H0 is true. For example, The p-value represents for the likelihood that the true mean for the promotion rates for males and females in the population is the same.

If the p-value falls below α, we would “reject the null hypothesis H0”. Alternatively, if the p-value does not fall below α, we would “fail to reject H0”. At the α = 0.05 significance level.

**Using non-statistical language:** we found enough evidence in this data to suggest that there was gender discrimination at play.

### Infer implementation

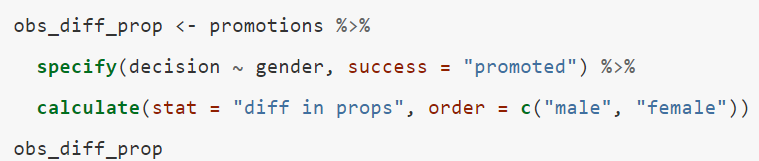
Diagram

Description automatically generated

* Calculating the Null Distribution



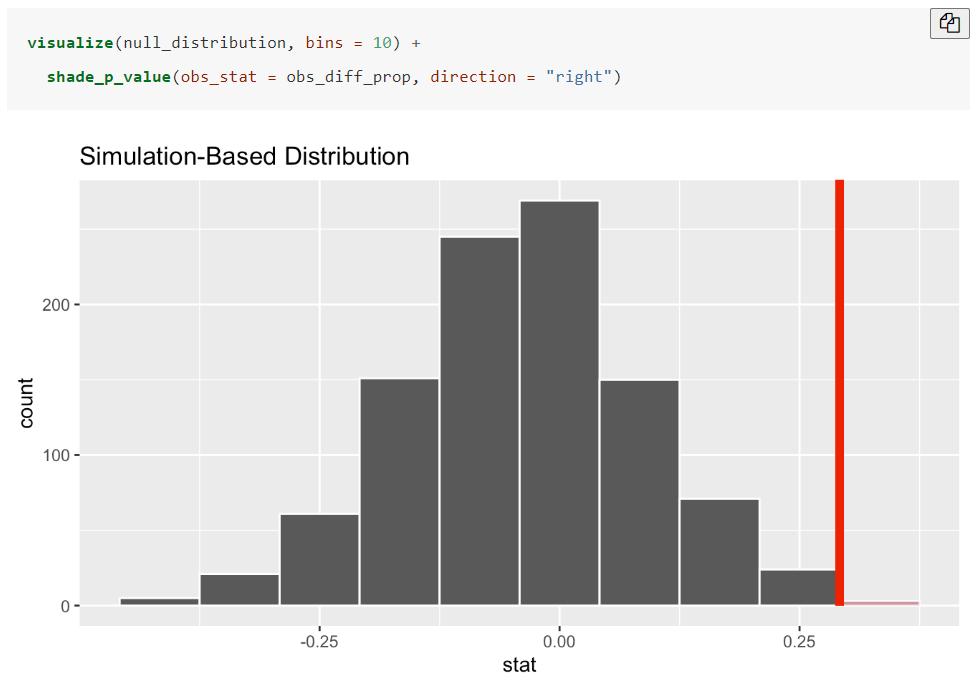
* Calculating the observed test statistic



* Calculating p-value



* Visualize results



## Types of errors

*There will always be the possibility of making either error when we use sample data.*

Table

Description automatically generated

* The probability of a **Type I Error** occurring is denoted by α. The value of α is called the significance level of the hypothesis test.
* The probability of a **Type II Error** is denoted by β. The value of 1−β is known as the power of the hypothesis test.