Introduction to Linear Regression and ANOVA with R

Laura Gutierrez Funderburk

Simon Fraser University

August 2019

SFU

SIMON FRASER
UNIVERSITY

SFU

SIMON FRASER
UNIVERSITY

About Me

- Mathematics Major, SFU

- DevOps Engineer for iReceptor, SFU

SFU

SIMON FRASER
UNIVERSITY

How do I use Linear Regression

- When I want to determine whether there is a relationship between two or more varialbles
- If there is a relationship, I want a model that will help me predict new values from existing data
- Help determine how much I can trust my model

How do I use ANOVA

- To determine if there are statistically significant differences between two or more groups
- Can come in handy when your data involves control and treatment groups (health sciences, biology for example)

SFU

SIMON FRASER
UNIVERSITY

Tell me about you

- Name

- What you do

- One challenge you face in your work

- What you hope to learn from this workshop

SFU
SIMON FRASER
UNIVERSITY

What is Linear Regression

The purpose of linear regression is to model a continuous variable Y as a mathematical function of one or more X variable(s).

We can use this regression model to predict the Y when only the X is known.

This mathematical equation can be generalized as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N + \epsilon$$

where, β_i are the coefficients associated to each X_i and ϵ is the error term - the part that the model cannot explain.

What kinds of models are we going to explore today

- One example of linear (single variable)

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- One example of linear (multivariate)

$$Y = \beta_0 + \beta_1 X_1 + + \beta_2 X_2 + ... + \beta_N X_N + \epsilon$$

SFU

SIMON FRASER
UNIVERSITY

What kinds of models are we going to explore today

- One example of polynomial (quadratic)

$$Y = \beta_0 + \beta_1 X_1^2 + \epsilon$$

What is a one-way ANOVA test

The one-way analysis of variance (ANOVA) is an extension of independent two-samples t-test for comparing means in two or more groups

Data is organized into several groups based on a factor variable

SFU

SIMON FRASER
UNIVERSITY

ANOVA test hypotheses:

Null hypothesis H_0: the means of the different groups are the same

Alternative hypothesis H_A: At least one sample mean is not equal to the others

ANOVA assumptions:

• The observations are obtained independently and randomly from the population defined by the factor levels
• The data of each factor level are normally distributed
• These normal populations have a common variance

SFU
SIMON FRASER
UNIVERSITY

What we are going to do today

• Visualize data sets and explore linear, multivariate and polynomial models

• Use R to determine the model that fits the data best

• Use ANOVA to determine differences between different groups within a dataset

Time to practice

Open RStudio
If you do not have RStudio, go to https://sfu.syzygy.ca/

SFU

SIMON FRASER
UNIVERSITY