

# Introduction to Linear Regression and ANOVA with R

Laura Gutierrez Funderburk

Simon Fraser University

August 2019



SIMON FRASER  
UNIVERSITY

- 1 Introductions
- 2 Linear Regression (In a Nutshell)
- 3 ANOVA (In a Nutshell)
- 4 The Heart of This Workshop
- 5 Background Knowledge
- 6 Hands on activities

# About Me

- Mathematics Major, SFU
- DevOps Engineer for iReceptor, SFU
- My hope is to motivate people to explore programming in R
- Fun Fact: My favourite place on Earth is a Chiapas, Mexico



SIMON FRASER  
UNIVERSITY

# Tell me about you

- Name
- What you do
- What you hope to learn from this workshop
- Fun Fact about you



SIMON FRASER  
UNIVERSITY

# When to use Linear Regression

- To help determine whether there is a relationship between two or more variables
- If there is a relationship, determine model that will predict new values from existing data
- Help determine how much one can trust (or not) that model



# When to use ANOVA

- To determine if there are statistically significant differences between the means of two or more groups
- To determine if a linear regression is statistically significant



# 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).

This mathematical equation can be generalized as follows:

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

where,  $\beta_i$  are the coefficients associated to each  $X_i$  and  $\epsilon = \epsilon_1 + \epsilon_2 + \dots + \epsilon_N$  note that  $\epsilon_i$  is the error term for each variable  $X_i$  which together as  $\epsilon$  account for 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)

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

The logo for Simon Fraser University, consisting of the letters "SFU" in white on a red rectangular background.

SIMON FRASER  
UNIVERSITY



# 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



## 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:

1. The experimental errors of the data follow a Normal distribution
2. Equal variances between treatments (i.e. Homogeneity of variances and Homoscedasticity)
3. Independence of samples: Each sample is randomly selected and independent



# The Heart of This Workshop

We can think of ANOVA and Linear Regression as equivalent. Indeed, ANOVA can help us determine whether we should reject the hypothesis that one variable does not depend on other(s) in our linear model.



SIMON FRASER  
UNIVERSITY

# The Heart of This Workshop

Null hypothesis  $H_0$ : there is no statistically significant relationship between the variables in the linear model

Alternative hypothesis  $H_A$ : there is a statistically significant relationship between the variables in the linear model



## Background knowledge: P-value

The probability of finding the observed, or more extreme, results when the null hypothesis ( $H_0$ ) of a study question is true.

P is also described in terms of rejecting  $H_0$  when it is actually true, however, it is not a direct probability of this state.



## Background knowledge: P-value

The null hypothesis is usually an hypothesis of "no difference" e.g. no difference between blood pressures in group A and group B.

**Define a null hypothesis for each study question clearly before the start of your study.**



## Background knowledge: Normal Q-Q Plot

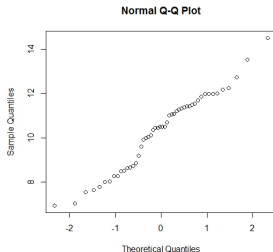
The Q-Q plot, also known as the quantile-quantile plot, is a graphical tool whose purpose is to assess if a set of data follows a Normal or an exponential distribution. This can help us check the Normality assumption of a dataset when we perform ANOVA on a model.





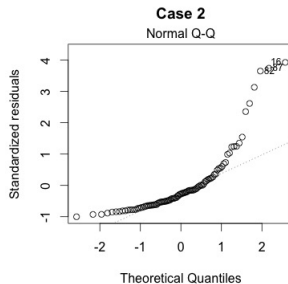
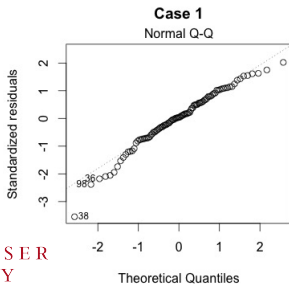
# Background knowledge: How to interpret the Normal Q-Q Plot?

Data will be sorted in ascending order and plotted against quantiles from a theoretical distribution. A linear relationship indicates the data is more or less Normally distributed.



# Background knowledge: How to interpret the Normal Q-Q Plot?

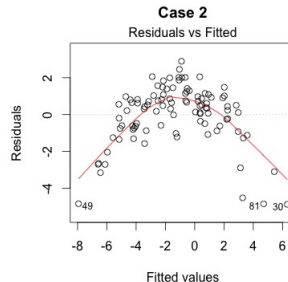
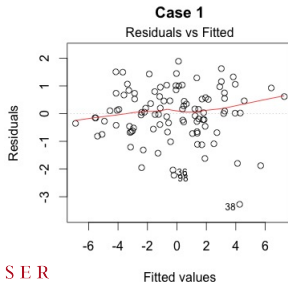
A curved relationship indicates skewed data, while a straight line which curves off in the extremities indicates more extreme values than expected if the data was Normally distributed.



SIMON FRASER  
UNIVERSITY

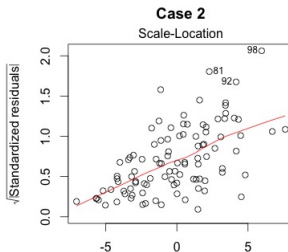
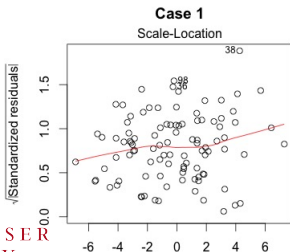
# Background knowledge: Residual vs Fitted Plot

This is a graphical tool that can help us see if residuals have a non-linear pattern. If we find equally spread residuals around a horizontal line with no distinctive patterns, this is a good indication that we are dealing with non-linear relationships.



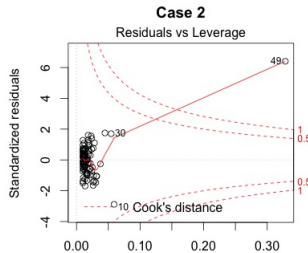
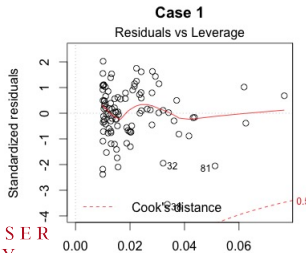
# Background knowledge: Scale-Location Plot

This plot is a graphical tool to help us see if residuals are spread equally along ranges of predictors. With this we can check the assumption for equal variance (also known as homoscedasticity). We are looking for a horizontal line with equally and randomly spread points.



# Background knowledge: Residuals vs Leverage Plot

This is a graphical tool which can help us determine influential cases, i.e. cases which if removed would change our model. Not all extreme values are influential cases. Outlying values can be found at the upper or lower right corners outside the dashed line. Such cases are influential to the regression results.



# What we are going to do today

- Perform statistical tests to determine if our data satisfies ANOVA assumptions
- Visualize data sets and explore linear models
- Use R to determine the model that fits the data best
- Use ANOVA to determine if the model is statistically significant



## Time to practice

Open RStudio

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



SIMON FRASER  
UNIVERSITY

## Further Reading

`https:  
//data.library.virginia.edu/diagnostic-plots/`



SIMON FRASER  
UNIVERSITY