Introduction to Linear Regression and ANOVA with R

Laura Gutierrez Funderburk, Alice Roberts

Simon Fraser University

October 2019

SFU
SIMON FRASER UNIVERSITY

SFU
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

SFU
SIMON FRASER UNIVERSITY

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

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).
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 $\epsilon = \epsilon_1 + \epsilon_2 + ... + \epsilon_N$ note that ϵ_i is the error term for each variable X_i which together as ϵ 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)

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

SFU
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

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

SFU
SIMON FRASER UNIVERSITY

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

SFU
SIMON FRASER UNIVERSITY

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

SFU
SIMON FRASER UNIVERSITY

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.

SFU
SIMON FRASER UNIVERSITY

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.

SFU
SIMON FRASER UNIVERSITY

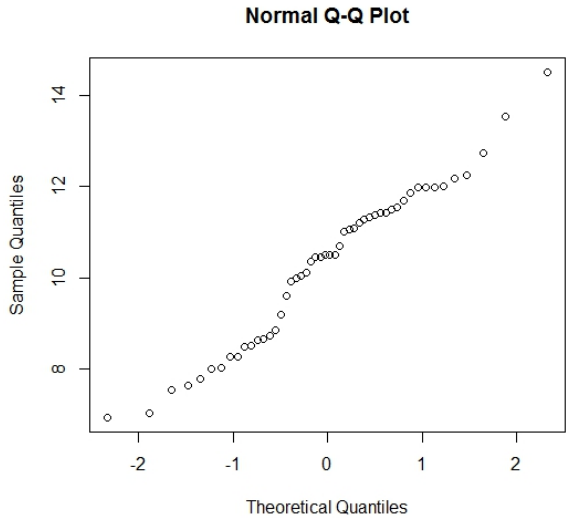
Background knowledge: Normal Q-Q Plot

The Q-Q plot, also known as the quantile-quantile plot, is a graphical tool whose purpose is to asses 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.

SFU
SIMON FRASER UNIVERSITY

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

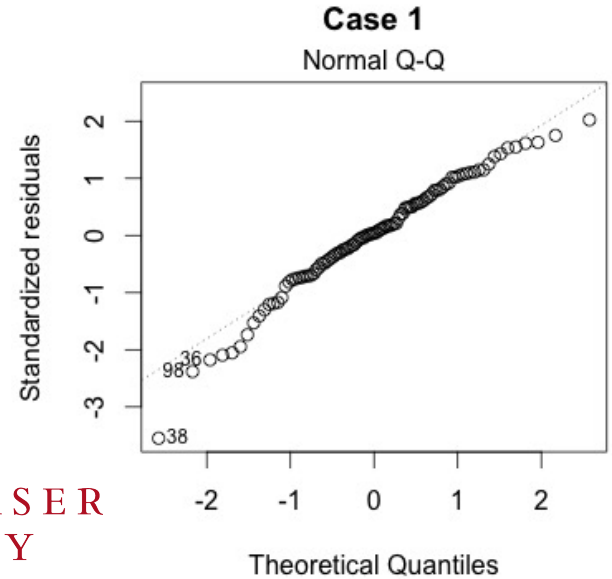
Data will be sorted in ascending order and plotted against quantiles from a theoretical distribution. A linear relationship indicates the data is more or less Normally distributed.

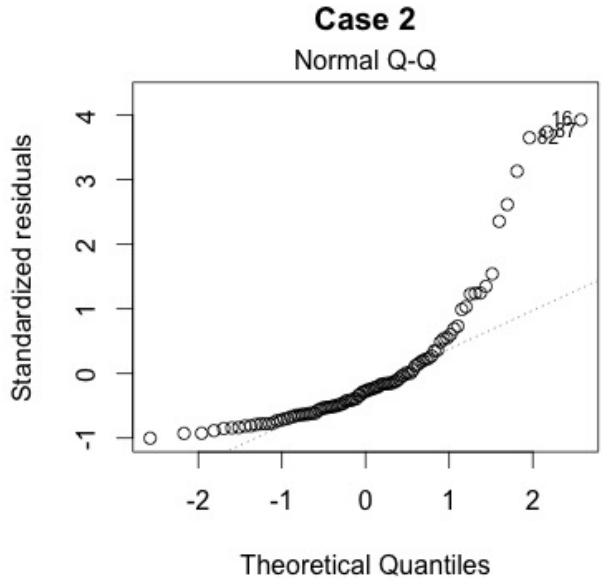


SFU
SIMON FRASER UNIVERSITY

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 extremeties indicates more extreme values than expected if the data was Normally distributed.

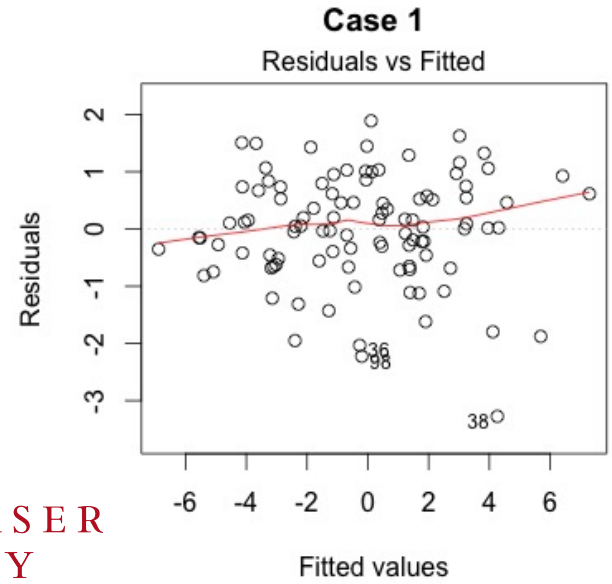


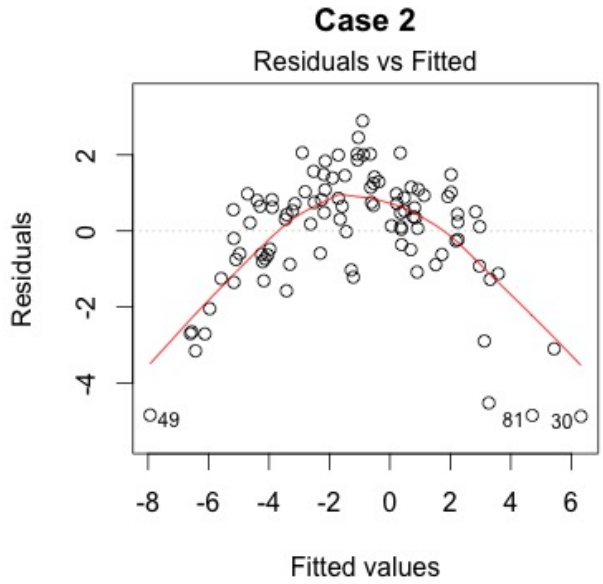


SFU
SIMON FRASER UNIVERSITY

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.

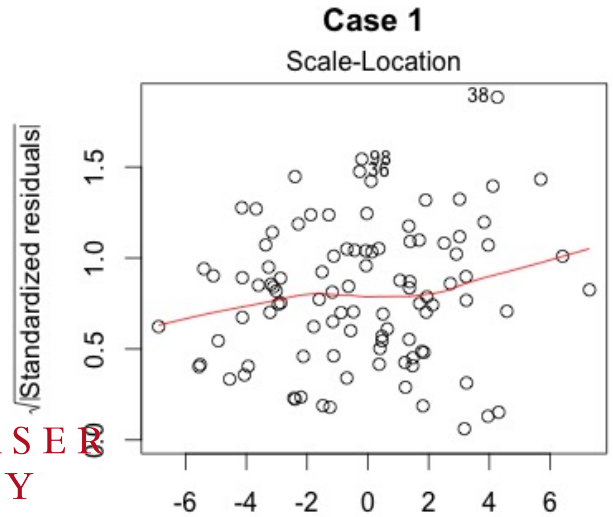


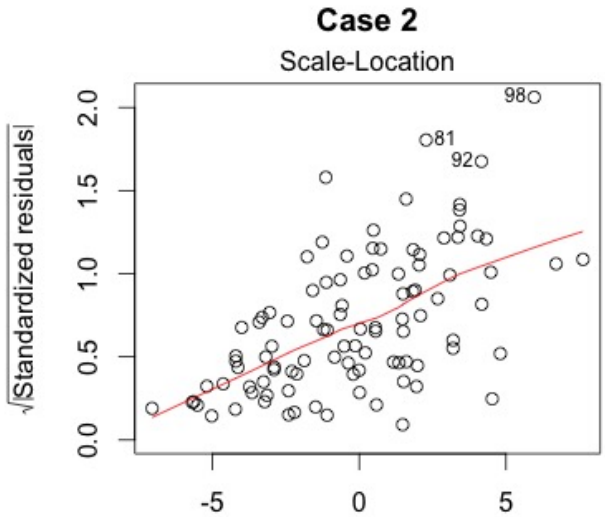


SFU
SIMON FRASER UNIVERSITY

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.

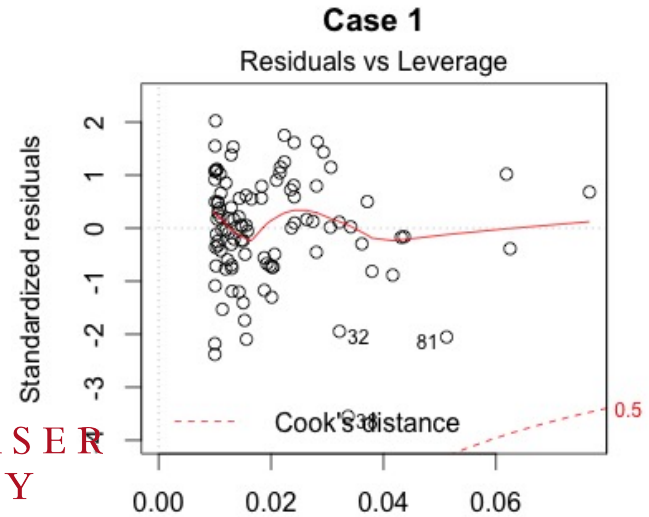


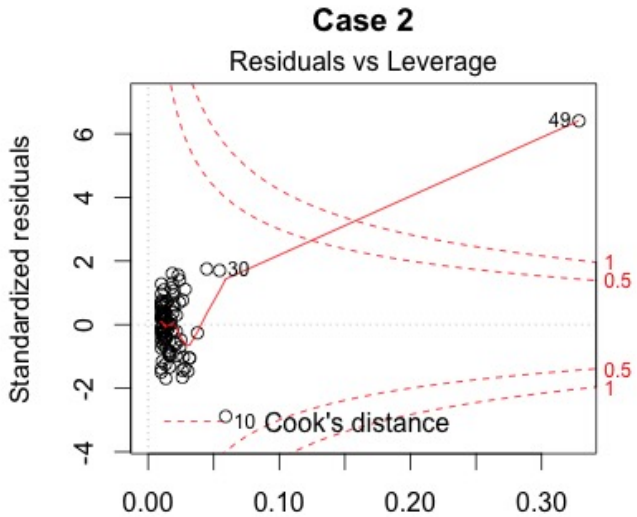


SFU
SIMON FRASER UNIVERSITY

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.





SFU
SIMON FRASER UNIVERSITY

What we are going to do today

- Perform statistical tests to determine of 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

SFU
SIMON FRASER UNIVERSITY

Time to practice

Open RStudio
Then open the files for Statistical Analysis

SFU
SIMON FRASER UNIVERSITY

Further Reading

More about Diagnostic Plots

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

https://medium.com/data-distilled/residual-plots-part-3-scale-location-plot-113e469b99c

https://medium.com/data-distilled/residual-plots-part-2-normal-qq-plots-c220ee9ed9fc

https://medium.com/data-distilled/residual-plots-part-1-residuals-vs-fitted-plot-f069849616b1

https:

//medium.com/data-distilled/residual-plots-part-4-residuals-vs-leverage-plot-14aeed009ef7

SFU
SIMON FRASER UNIVERSITY

Huge Thanks To

SciProg for hosting us and supporting us.