Introduction to Statistical Models

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

Topics

Review

t-distribution

- Recall from confidence intervals for a mean value:
  - If the population standard deviation is known we can use the normal distribution.
  - In practice, the population standard deviation is not known and we use the t-distribution.
- The t-distribution *approaches* the normal distribution as the degrees of freedom (n) increases.

One-Sample t-Tests

- Used to test significance of regression terms.
- Test statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{1}$$

- Rejection regions:
  - Lower one-tailed test: $t < t_{-\alpha,n-1}$
  - Upper one-tailed test: $t > t_{\alpha,n-1}$
  - Two-tailed test: $|t| > |t_{\alpha/2,n-1}|$

Rejection Regions in R

- $t_{-\alpha,n-1}$ (Lower one-tailed)
  - `qt(p=`α`, df=n-1, lower.tail=TRUE)`
- $t_{\alpha,n-1}$ (Upper one-tailed)
  - `qt(p=`α`, df=n-1, lower.tail=FALSE)`
- $t_{\alpha/2,n-1}$ (Two-tailed)
  - `qt(p=`α`/2, df=n-1, lower.tail=FALSE)`

p-values in R

- We may also use the `t.test()` function in R.
- This function takes at least one **vector** of values as the first argument.
- Produces the test statistic (t), confidence intervals, p-value, and sample mean.
- Usage:
    - Lower one-tailed: `t.test(x, mu =` μ_0`, alternative = "less", conf.level = 0.95)`
    - Upper one-tailed: `t.test(x, mu =` μ_0`, alternative = "greater", conf.level = 0.95)`
    - Two-tailed: `t.test(x, mu =` μ_0`, alternative = "two.sided", conf.level = 0.95)`

Review Examples A

- Use the `t.test()` function to test assumptions about the simulated data.

Normality Assumptions

- Some statistical tests are only valid when assumptions about the distribution hold.
- Data being drawn from a Normal distribution is common assumption for many tests.
- Some methods to determine if our data is normally distributed or not:
    - Quantile-Quantile plots (not a formal test).
    - ~~Kolmogorov–Smirnov test.~~
    - Shapiro-Wilk test of normality.

Quantile-Quantile (Q-Q) plots

- **(Q-Q) plot** is a visualisation method to determine if two distributions are the same.
- Use the theoretical quantiles from a normal distribution and plot them against the quantiles from the sample.
- If the two distributions are identical the Q-Q plot will follow the 45^o line.
- Otherwise we can conclude that the distributions are different (not normally distributed).

Shapiro–Wilk Test

- **Shapiro–Wilk test** is based on the ordered sample.
- The test statistic is complicated to calculate so we will use R.
- The null hypothesis is that the population is normally distributed.
- If we obtain a small p-value we can reject the null hypothesis of normality.

Review Examples B

- Examine the normality of the example data.

Nature of the Relationship

- One variable (Y) is the dependent (response) variable and other variables play the role of independent (explanatory) variables (X_1, X_2, ..., X_k)
- The relationship is not deterministic (functional) but is **statistical** (stochastic).
- There is a (conditional) distribution of the dependent variable associated with various combinations of independent (explanatory) variables.
- Initially we will focus on **linear** relationships.

Sample Correlation

- Often we do not have all of the observations in populations and we need to estimate parameters.

- **Sample correlation** estimates the actual correlation between two random variables:

$$\rho = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}}. \tag{2}$$

- Where S_{XY}, S_X, and S_Y are the sample covariance and standard deviations.

Sample Correlation in R

- R function: `cor(x, y, method = "pearson")`
  - `x`: a numeric vector, matrix, or data frame.
  - If `x` is a vector we must give a vector `y`.
  - `method = "pearson"` is the default method.

- `method = "kendall"` to measure the ordinal association between two measured quantities.
- `method = "spearman"` (rank correlation) to assess monotonic relationships between two measured quantities.

Correlation Matrix

Variable	X	Y	Z
X	$\rho_{XX} = 1$	ρ_{XY}	ρ_{XZ}
Y	ρ_{XY}	$\rho_{YY} = 1$	ρ_{YZ}
Z	ρ_{XZ}	ρ_{YZ}	$\rho_{ZZ} = 1$

Review Examples C

- Examine the relationships found in the example data.

Introduction to Statistical Models

Introduction

- The world we live in is extremely complex.
- "*All models are wrong but some are useful*" - Box, 1979
- Being in this room is a result of the outcomes of a complicated series of events and decisions.
- **Statistics is the explanation of variation in the context of what remains unexplained.**
- It is very important to pay close attention to the descriptive accuracy of statistical models.

Observational Studies & Experiments

- In an **observational study**, variables are observed without any attempt to change/control independent factors.
  - Background or lurking variables *may* be the cause of any changes in the dependent/response variable.
- In an **experiment** the independent variables are purposely controlled and varied and runs are executed in a way to minimize the impact of any lurking variables.

Experiments

- In general, causal inferences are *more certain* in experiments than observational studies.
- The explanatory variables are under the direct control of the researcher(s).
- Randomization provides a strong platform for inferences to be drawn during experimentation.

Models as Functions

- We are going to use the concept of a **function** throughout the regression course.
- A **function** is a mathematical concept that represents the relationship between an **output** and one or more **inputs**.
- In general we will use a *formula* to represent such relationships.

Response Variable

- The **response/dependent variable** is the variable whose variation/behavior the modeller is trying to understand.
- In graphical form, the response variable is located on the vertical axis.
- Often denoted by Y.

Explanatory Variables

- The **explanatory/independent variables** are the other variables that the modeller wants to use to explain the variation of the response variable.
- In graphical form, the explanatory variable is located on the horizontal axis.
- Often denoted by X_i.

Conditioning on Explanatory Variables

- In regression, we will take the value of the explanatory variables into account when looking at the response variables.
- For example: salary conditioned on years of experience.

Model Value

- The **model/fitted value** is the output of a function.
- The estimated function called the **model function** gives the typical value of the response variable conditioning on the explanatory variables.
- **The estimated simple linear regression equation is**:

$$\hat{Y} = b_0 + Xb_1.$$

  - b_0 and b_1 are estimates of β_0 and β_1.

Residuals

- The **residuals** show how far each observation is from its *model value*.
- Residuals are always measured: *actual value minus fitted value*.
- In linear regression the *residuals* (observed errors) are defined as follows:

$$e_i = Y_i - \hat{Y}_i.$$

Example 1

- Run the example data code provided.
- Estimate four separate simple linear regression models.
  - Which of your models fit well?
  - How large do the residuals appear to be?

- The idea of a function is fundamental to regression and statistical modelling in general.

- We use the model function to describe a relationship, our description will **NOT** be perfect.

- The residuals give us information as to how close each observation is to our model function.
  - Random portion or unexplained variation in the response variable.

Multiple Explanatory Variables

- Statistical models may contain more than one explanatory variable.
- As mentioned, real-life processes and relationships are often complex.
- Models may be able to capture relationships that are difficult to see using visualizations alone.

Example 2

- Using the *df5* data in the example code, create a linear regression model with multiple explanatory variables.
- How does the model perform?

Reading Models

- Understanding your results is extremely important.
- Visualizations can be useful to examine the relationships.
- Quantitative ways to read a model:
  1. Read out the *model value*.
  2. Characterize the relationship *described* by the model.

Read out the Model Value

- Plug in specific values for the explanatory variables and read out the resulting model value.
- Essentially examining the fitted values for specific combinations of explanatory variables.
- Specific *point*, not a general description of the relationship.

Characterize the Relationship

- Interested in the overall relationship.
- Essentially examining the fitted values for specific combinations of explanatory variables.
- Not focused on specific combinations of values.

Slope

- Generally, a slope can be used to characterize the relationship *described* by the model.
- The numerical size of the slope is a measure of the strength of the relationship (rise over run).
- The units of a slope: units of the response variable divided by the units of the explanatory variable.
  - Distinct slope associated with each independent variable.
- For categorical variables, *differences* are used instead of slopes.

Describing a Model

- The way we describe our model can carry implications including *causation*.
- Some examples:
  - "*The difference between typical wages*": No causation
  - "*Typical wages go up by 20 cents per hour for every year of age*": Implies causation
- It is important that the data are collected in an appropriate way to draw conclusions about causation.

Model Design Selection

- The model selection depends on its intended purpose.
- Also depends on the information available:
  1. The data
  2. The response variable
  3. The explanatory variables

Data

- How were the data collected?
  - Part of an experiment or are they observational?
  - Random sample or from a sampling frame?
  - Are the relevant variables being measured?
- What conclusions will you really be able to draw from the data?
- *Garbage in garbage out*

Response Variable

- *The choice of response variable is often obvious.*
    - The thing that you wish to predict or whose variability you would like to understand.
- Most of the methods we will cover require a numeric dependent variable.

Explanatory Variables

- Most of the consideration will be given to the selection of explanatory variables.
- We will cover situations where the inclusion of an explanatory variable can actually hurt the model.
- Variable selection and being able to explain why you made the selection you did is a very important combination of skills to learn.

Example 3

- Examine the variables in *df6* and decide which variables should be considered for a linear regression model.

- Estimate some models and see if you were on the right track.

Model Terms

- Explanatory variables can be included in a model in more than one way.
- The shape of the model is determined by the shape of the **model terms**.
- Describing models using model terms is useful for communicating with the computer and for dealing with multiple explanatory variables.
  - It is very difficult to *see* the relationships.
- We will evaluate the contributions of each model term and decide their relevance in the overall model.

Some Model Terms

- **Intercept term**: A baseline/overall average that is included in almost every model.
- **Main terms**: The direct effects of explanatory variables.
- **Interaction terms**: How the relationships between different explanatory variables influence the response variable.
- **Transformation terms**: Simple modifications to explanatory variables.

Notation in R

- We can include interaction terms and transformations in our models in R by using the `*` symbol *or* the `:` symbol.

- `lm1 <- lm(Y ~ X1 + X2 + X1*X2 + log(X3), data = data.frame)`

- `lm2 <- lm(Y~ X1 + X2 + X1:X2 + I(X3^2) -1, data = data.frame)`

- Use `I()` to perform transformations, and use −1 to remove the intercept.

Example 4

- As we know exactly how the variable Y was constructed in *df6*, construct two models, one with and one without an intercept.
- Which model seems to be better?

Exercise 1

- Consider the models that were estimated in Example 4.
  - Take some time to describe the models in detail.
  - In this case are we able to assume causation?

Exercise 2

- Think back to a regression model you have constructed in the past. How would you describe your model to others?

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/
2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm
- https://cran.r-project.org/web/packages/car/index.html