Tidy Linear Regression Analysis

Dr. Austin Brown

Kennesaw State University

11/29/2021

Introduction

- In the last session, we learned how we can use modern, tidy programming to perform traditional factorial ANOVA analyses, including assumption checking and data visualization.

- As you know, ANOVA types of analyses are not the only method available to a data analyst/scientist/statistician.

- When we think about ANOVA conceptually, the data structure we have is:
  a. A Continuous Outcome/Response
  b. One or more categorical predictors/explanatory variables

Introduction

- Regression analysis, on the other hand, we tend to think about somewhat differently (even though ANOVA and regression are the same):
  a. Used for prediction
  b. Continuous Outcome/Response
  c. Quantitative predictors/explanatory variables

- However, while the above is typically how regression analysis is thought of, it can also include categorical predictors via dummy coding (more on this later)

Introduction

- Let's consider a scenario where regression analysis would be appropriate:
- Suppose I want to build a predictive regression model for MLB Team wins per season using team-level batting statistics.
- Wins is a quantitative variable and batting statistics are also quantitative, so a regression methodology seems appropriate. An Excel spreadsheet with these data is contained in D2L.

Exploratory Correlational Analysis

- When you're in the beginning stages of model building, it's a good idea to do exploratory sorts of analyses to help get a better understanding of the data as well as the relationships therein that you may not have considered.

- One such visual method is a scatterplot matrix. Let's see how we can create and interpret one with these baseball data using traditional R methods and some more modern methods.

Model Building: The Traditional Approach

- While all regression-type of models are built in a similar way, the specific goal of the analysis dictates whether or not specific things should also be done.

- So for example, in traditional types of academic research, researchers are mostly interested in seeing whether or not some predictor variables are significantly related to some outcome variable.

- There's typically not a predictive component to this type of research, per se. But, we still have assumptions to assess in order to validate both the distributional assumptions as well as the overall fit (or effect size).

Model Building: The Traditional Approach

- One of the main reasons why someone would want to use R Markdown for compiling documents is that you can do everything in one file.
- Additionally, one of the main reasons why people keep developing new R packages is to make our lives easier and less code-heavy.
- With this in mind, what if we wanted to output our fitted model in R Markdown?

Model Building: The Traditional Approach

```
library(equatiomatic)
equatiomatic::extract_eq(mod2,use_coefs=FALSE,
                                 wrap=TRUE)
```

$$\begin{aligned} \text{Wins} = \alpha + \beta_1(\text{RBI}) + \beta_2(\text{SB}) + \beta_3(\text{CS}) + \\ \beta_4(\text{BB}) + \beta_5(\text{SO}) + \beta_6(\text{BA}) + \epsilon \end{aligned} \tag{1}$$

Model Building: The Traditional Approach

- And as we learned before, we can output our ANOVA table in a nice format using R Markdown.

```
knitr::kable(mod2 %>%
                 moderndive::get_regression_table() %>%
                 dplyr::select(-lower_ci,-upper_ci))
```

term	estimate	std_error	statistic	p_value
intercept	57.990	22.773	2.546	0.011
RBI	0.081	0.014	5.920	0.000
SB	0.008	0.022	0.370	0.712
CS	-0.079	0.069	-1.145	0.253
BB	0.034	0.011	3.161	0.002
SO	-0.020	0.005	-3.998	0.000
BA	-88.532	89.506	-0.989	0.323

Model Building: Dummy Coding

- As I mentioned before, regression analysis isn't restricted to having numeric predictors. We can also have categorical predictors, but we use them slightly differently through the use of dummy coding.

- Basically what this means is that for a categorical variable with k levels, we will have $k - 1$ dummy variables.

- Let's look at what this looks like in practice using the `penguins` dataframe.