Mini-Project 2 - Regression

Timo Meiendresch

WS 20/21

1 Introduction

In the previous project you worked on a systematic way to explore a new dataset, called Exploratory Data Analysis (EDA). Here, you'll repeat what you have done before, in particular conducting the EDA process, but you will take it one step further by applying a multivariate regression model.

The outcome will again be a short presentation on a dataset. However, the focus will be on your understanding of regression modeling applied to your specific dataset. Structure of your analysis:

1) EDA (similar to mini project 1)
2) Regression Analysis (Focus!) based on your results from EDA (1) as well as based on

- Course content on Regression (part 1 and part 2)
- Chapter 3 of Introduction to Statistical Learning (ISL)

Before starting with the project, I advise to read Chapter 3: Linear Regression of ISL carefully.

2 EDA

As before, start with your EDA and look for possible variations and covariation of your data. This step is important to inform you about later model choices.

3 Regression Analysis

The following questions should guide your research on regression and your analysis. You may answer these in your presentation and may be asked some of them in the subsequent Q&A. Start with simple, bivariate regression analysis of your dependent variable y and move on to multiple regression. Based on your findings from your EDA you may already have some idea on relations in your data. Now, try to quantify your findings using regression analysis. The subsequent list of questions is rather exhaustive. Most of them can be easily answered and will become easier once you get used to regression analysis. They may help you to structure your analysis:

- What is your target variable y and why?
  - Inspect this variable visually and quantitatively
- Check Covariation in your data (you may have done this in your EDA)
  - Are there clear relationships between predictor and response variables?

3.1 Assess the accuracy of the coefficient estimates: Hypothesis testing and confidence intervals

- How "good" is your estimate of the coefficients? How to assess its quality? Which variables are significant? What does it mean?
- What is a t-statistic? Which estimates are necessary to get the t-statistic?

- What is a p-value? How are t-statistic and p-value related?
- Relation between t-statistic, p-value and standard error of coefficients? When to reject the null hypothesis? What does a rejection mean?
- What's the relation between confidence intervals and hypothesis testing in regression analysis?
- "If we cannot reject the null hypothesis of β_i, then the model reduces to ... "?

Relate these questions to your data and interpret what you think is important. It is not necessary to talk about everything and most questions are closely linked. Often it helps to look at the equations to realize how they are related.

3.2 Interpreting your regression coefficients

- Be able to interpret coefficients, standard error, t-statistic and p-value
- How strong is the relationship? Positive or negative?
- If you have used **logarithmic variables** (either as independent or dependent variables): How to interpret these? –> Advanced (this is more sophisticated and will be positively recognized if covered)!
- **Qualitative predictors** (independent variables): Did you use categorical predictors like `gender`, `status`, `ethnicity`?
  - How to interpret these? I.e. what happens if person i was female (male)?
  - How are these variables encoded in your regression analysis?
- What is the so-called "ceteris paribus" assumption in regression analysis? What does this mean in your own words?
- How to include and interpret **interactions** and **nonlinearity** in your regression? Advanced!
  - Interactions between variables? If you include interaction terms, pay attention to the **hierarchy principle**
  - Nonlinearity?
- Why is it important that our variables are assumed to be uncorrelated? What does it mean? In other words, what is the "multicollinearity problem?" –> Advanced!

Statements in regression analysis are generally conditional on all other variables in the model. Be careful of what is in the model but be particularly careful of variables that may be missing.

Side note on causality:

Mosteller, Tukey, Box, Pearl and other researchers of causal inference highlight the importance of actively checking causal relations. Causality is not a passive endeavor but testing hypotheses actively by doing (Experimental or quasi-experimental approaches). Beware of interpreting your regression findings as causal! Here, we deal with observational data so we won't be able to infer any "true" causality here.

> "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

3.3 Assessing the overall accuracy of the model

- How to assess the overall accuracy of a model?
- How to calculate the R^2 and what does this measure mean?

3.4 Model selection

It is generally not obvious which variables to include in the regression and which one to leave out. If you cannot motivate your variable choices theoretically there are some common approaches you may use, namely **Forward selection** and **Backward selection** for example

- What are these two concepts?
  - Did you use one of them? Explain your choice of variables!

Note that these two approaches are rather mechanical. Ideally, you would like to understand the domain that you are working on and motivate the inclusion of variables using theory and other research.

3.5 Other questions to consider

- Is at least one of the predictors useful in predicting the response?
  - F-statistic?
- Are all the variables helpful in explaining our y, or is only a subset of the predictors useful?
- Given a set of predictor values, what response value would we predict?
- Think about the underlying mechanisms: Does it make sense that a certain variable x may be associated with a change of y?
- What about possible missing variables? Do you think that there are other variables that are missing which may be associated with your dependent variable y?

4 Outcomes

Don't be afraid, the outcome is a short presentation. Total amount of time for each group is 5 minutes per group member (10 minutes for groups of two and 15 minutes for groups of three).

You should briefly introduce us to the general story of your data as in miniproject 1 and then cover your regression analysis.

Again, we expect to see nice, informative visualizations and a coherent and interesting story told with your data. Creativity and interesting presentations will be greatly appreciated! Make it fun and interesting for everyone. In addition, you should clearly motivate your regression analysis and point out interesting findings which may be worthwhile of further investigation.

After each presentation, we should have learned something that we did not know before on the topic that you have covered ("What did we learn that we did not know before?").

5 Groups and data

Group	Data	Source
#1 (Atique Nazar, Rishab Battacharyya, Sibora Domi)	Ames Housing dataset	Kaggle
#2 (Hammad Ahmed, Hantao Hui, Hatim Ali Asghar)	CASchools {AER}	library(AER)
#3 (Mehtanin Rashikh, Pruthvi Hegde, Steljana Lleshi)	Student Performance	UCI
#4 (Antonia Pütz, Pavel Raschetnov, Tauseef Ahmad Awan)	Wine Quality Data Set	UCI
#5 (Eric Li, Lin Henry, Md Razaul Haque)	Life Expectancy (WHO)	Kaggle

You may decide to look for a different dataset of your own choice but I need to check whether it is suitable. Please email me so I can check your suggested data and decide on it on monday. Possible sources:

- Kaggle
- tidytuesday
- UCI
- lionbridge overview article