Quantitative and Qualitative Analysis

Devan Becker

Introduction

Land Acknowledgement

UWO exists on the traditional territories of:

- Anishinaabek
- Haudenosaunee
- Lūnaapéewak
- Chonnonton

These lands are connected with the London Township and Sombra Treaties of 1796 and the Dish with One Spoon Covenant Wampum.

I acknowledge that Western University is located on the traditional territories of the Anishinaabek, Haudenosaunee, Lūnaapéewak, and Chonnonton Nations, on lands connected with the London Township and Sombra Treaties of 1796 and the Dish with One Spoon Covenant Wampum. The Dish with One Spoon is a metaphor for the responsibilities that we have when sharing the land; each of us eats from the same bowl and we must always leave enough for the others. It's easy for people like me to think of history books when talking about indigenous peoples, but I recognize that they are contemporary stewards of the land and vital contributors of our society. I also recognize the great diversity among indigenous peoples, both within and between nations. For instance, the Dish With One Spoon Covenant Wampum is not an "Indigenous Thing", it's specific to the Haudenosaunee peoples.

About Me

- B.Sc. Math (Laurier)
- M.Sc. Stats/Biostats
- Ph.D. Stats
- Postdoc - bioinformatics and data science
- Music, reading, outdoorsy stuff

A little about me before we begin. Did an undergrad thesis project with Doug Woolford. He introduced me to some people at Western for my master's, then followed me and became my PhD supervisor I defended my PhD thesis exactly

one week before the pandemic hit Canada, so I am extremely lucky that I got hired as a Postdoc in the Schulich school. Outside of academia, I listen to and play music (poorly), go camping, hiking, showshoeing, kayaking, and rock climbing, and I like to read a variety of things.

Outline

- **Quantitative:** Dealing with numbers
    - Any number in a range
    - Only 0's and 1's (maybe a 2)
    - Things we can turn into numbers

- **Qualitative:** Dealing with descriptions
    - Using your brain
    - Using your computer

- **Meditative:** Dealing with everything
    - How to get started
    - Accessing resources
    - Not being afraid of coding

I'm a statistician by training, so most of this will be spend on quantitative analysis. I'm going to focus on linear regression, which is an analysis technique that can be used when you're modelling something that could take any value in a range. I'll briefly go over a few other techniques that work for numbers in a range. I'll also talk about classification problems, which is when the thing we're modelling can only be one of a handful of numbers. I'll talk about qualitative analysis, but mostly from a data scientist's perspective. Finally, I'll do my best to get you started on learning your own analysis. There are lots of resources on campus and across the internet, but there are some caveats with all of those.

Before we begin

- Interrupt at any time

- All notes/links/resources are on GitHub

- I have allowed myself **ONE** equation.

```
The GitHub version also has my (approximate) script inside
:::notes::: tags, which show up as text in the pdf.
```

I'll add the link to the chat now, but I'll also have Mihaela send it out after the meeting.

Quantitative

Regression

So let's start with the fun one!

- The **Target** could be any number in a range.
    - A.k.a. dependent variable or response.
- The **Features** could be any data type
    - A.k.a. explantory or independent variables

illustration

Most things are linear models

illustration: y v. x1 and y v. x2, mostly linear-looking, with slopes and intercepts labelled

This is the blah

Linear models: intercepts

- Exists to make the model fit better

- Not always interpretable

Mean-centering

illustration: not mean-centered -> mean centered

Linear models: slopes

illustration: one unit increase in x1

For every one unit increase in x1, y goes up by blah.

Binary Features

Suppose x1 is labelled 0 or 1.

What does the slope represent?

The story so far

- Intercept is a mathematical necessity
- Slopes answer our questions

But how good is our model?

The most important part!

illustration: error

Residual plots

illustration: the subtle sine curve that I sometimes show

Residual plots

illustration: some interesting relationship between features

Putting it all together

1. Get data
   - Data cleaning is the hardest part.
2. Plot data
   - If you haven't plotted it, you're doing it wrong.
3. Fit model
4. Check model
   - If you haven't plotted it, you're doing it wrong.

Non-linear models?

illustration: sine, polynomial, spline smoothing

It's all just linear!

Feature Selection

This is hard

Machine Learning

Lasso Regression

- It's like linear regression, but it automatically removes features.

xgBoost

Remember the residual plots?

What if we fit a regression to those residuals?

Neural Nets