Lesson: Frameworks and goals of statistical modeling

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

switches + dials

↓ ✓

Audio signal

Photo by Steve Harvey on Unsplash

$$Y = f(\mathbf{x}) + \varepsilon$$

Photo by Steve Harvey on Unsplash

temperature, humidity + fuel density + elevation, lightening storms + number of nearby campsites + …

↓ ?

Likelihood of large wildfire

Image from firelab.org: https://bit.ly/37q9EGA

Definition: A *statistical unit* is one member of the set of entities being studied.

Definition: A *population* is a collection of units about which research questions are asked.

Definition: A *sample* is a subset of the population. Typically, samples should be *representative.*

Photo by Max Adulyanukosol on Unsplash

Photo by ismail mohamed - SoviLe on Unsplash

Definition: *Inferential statistics and data science* is the process of learning about relationships in a sample in a way that is reliable enough to *generalize* from the sample to a population of interest.

Photo by Danny Nee on Unsplash

Lesson: The assumption of *concept validity*

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

Is cycling any less safe in the winter than in the summer?

Photo by Max Adulyanukosol on Unsplash

Definition: To *operationalize* a concept means to derive a set of steps to measure the concept.

Definition: The *validity* of a dataset or measurement tool is the extent to which the dataset or measurement tool measures what it claims to measure.

Photo by Max Adulyanukosol on Unsplash

Study on cycling safety:

http://bit.ly/3oZNd0x

Photo by Max Adulyanukosol on Unsplash

Lesson: The linear regression model

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

Motivation and Context

Basic statistical inference: estimate or test hypotheses about data generated from a distribution with a constant mean. For example:

But what if the mean depended on another variable? For example:

The Context of Linear Regression

Linear regression is used to explain or model the relationship between a single variable Y, and one or more variables $x_1, \ldots, x_p$.

Definition: Y is called the *response, outcome, output,* or *dependent* variable.

Definition: $x_1, \ldots, x_p$ are called *predictors*, *inputs*, *independent variables*, *explanatory variables,* or *features*. In some contexts, they are also called *covariates*.

We will assume, for now, that all variables are continuous (but will soon extend our methods to allow for discrete variables!).

The Context of Linear Regression

The simplest relationship between these variables is a linear relationship:

But note that, when we actually measure data, measurements aren't perfect—there is error. So, we might use the following to model our data:

Exact Linear Relationship

Noisy Linear Relationship

The Context of Linear Regression

Regression analysis has two main objectives:

1. *Prediction:* Predict an unmeasured/unseen Y using observed $x_1, \ldots, x_p$.

2. *Explanation*: To assess the effect of, or explain the relationship between, Y and $x_1, \ldots, x_p$.

Can we infer causality?

Lesson: Matrix representation of the linear regression model

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

The Linear Regression Model

Examples of linear models (i.e., models that linear methods can handle):

1.

2.

Examples of nonlinear models:

1.

Photo by Jason Leung on Unsplash

Lesson: Assumptions of the linear regression model

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

The Linear Regression Model

<u>Definition/Assumptions</u> of the linear regression model:

1. *Linearity*:

2. *Independence*:

3. *Homoskedasticity (constant variance):*

4. *Normality*:

Photo by Jason Leung on Unsplash

Lesson: Interpreting the linear regression model

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

Photo by Jason Leung on Unsplash

Interpreting the Regression Parameters

<u>Interpreting ***simple*** linear regression parameters</u>:

1. β_0: the intercept of the true regression line
2. β_1: the slope of the true regression line.

Interpreting the Regression Parameters

<u>Interpreting ***simple*** linear regression parameters</u>:

1. β_0 *is the average value of* Y *when x is zero.* Usually this is called the "baseline average".

2. β_1: *is the average change in* Y *associated with a 1-unit increase in the value of* x.

Interpreting the Regression Parameters

<u>Interpreting ***multiple*** linear regression parameters</u>:

1. β_0: the intercept of the true regression line. We interpret this as ***the average value of*** Y ***when all of the*** x***'s are zero.***

2. β_j: the slope of the true regression line, $j = 1, \ldots, p$.

Interpreting the Regression Parameters

<u>Interpreting ***multiple*** linear regression parameters</u>:

1. β_0 *is the average value of* Y *when all of the* x*'s are zero.*

2. β_j *is the average change in* Y *associated with a 1-unit increase in the value of* x_j *assuming all other predictors are held constant,* $j = 1, \ldots, p$.

Thus, these "slope" parameters are called ***partial*** or ***adjusted*** regression parameters/coefficients.

Lesson: The appropriateness of linear regression

Module: Introduction to statistical models

Photo by Robert Ritchie on Unsplash

How do we know linear regression is appropriate?

Photo by Jason Leung on Unsplash

Does a theory or “law” tell us how the variables relate?

Photo by Jason Leung on Unsplash

Does a theory or “law” tell us how the variables relate?

$$F = kx$$

Force

Spring Constant

Displacement

Photo by James Allen on Unsplash

What past studies related to this research question have been conducted?

Photo by Jason Leung on Unsplash

Can we explore the data to learn whether there are linear relationships between variables?

Photo by Jason Leung on Unsplash

Definition: A *circular analysis* or *double dipping* is the process of exploring a dataset in an attempt to discover what relationships exist, and then test hypotheses related to that exploration on the <u>same</u> dataset.

Photo by Jason Leung on Unsplash

Ways to avoid circular analyses:

1. Design the analysis and pre-specify research hypotheses before observing the data.

2. Subset the data

Photo by Jason Leung on Unsplash

Nonlinear relationships between the predictors and the response?