Handout 1.1: Introduction to Modern Regression Methods

Dr. Bean - Stat 5100

1 About me

- Graduated from BYU-Idaho in 2014 with a Bachelors of Science in Applied Mathematics.
- Graduated from Utah State University in 2019 with a PhD in Mathematical Sciences.
- Current interests include basketball, cross country skiing, hiking and spending time with my wife and daughter.

(Groups) What is a creative, yet appropriate, question that you have about the life/career of the instructor?

2 Why Modern Regression Methods?

Statistics, in the words of Dr. Bin Yu, is the "science that solves data problems." This science becomes more and more relevant in a world inundated with data. From the late Leo Breiman:

> The uses of statistics pervade our society. They are used and terribly misused all through the social sciences and health fields. ... It is surprising how much the world around us depends on the use of statistics. ... It's odd that even though the articles

involving statistics in the newspapers far outnumber those involving say, physics or chemistry, people in general know very little about what we do.

In this class, we will learn several of the foundational approaches for using data to make **predictions**. Perhaps more importantly, we will discuss the **cautions** we must consider when using and interpreting model output.

(Groups) Why are YOU taking this course?

3 Functional vs Statistical Modeling

We learn about functions in Math 1050 (College Algebra), a functional model takes a set of inputs X and produces a (set of) outputs Y, i.e.

$$Y = f(X)$$

Example: You write a function to model the profits from your lemonade stand. You rent the stand for \$200 a month and sell each glass of lemonade for \$1.00. If it costs you \$0.25 to make the lemonade then your monthly profits Y could be modeled as a function of the number of lemonade glasses you sell x

$$Y = 0.75x - 200.$$

The key to a functional model is that each input x produces a **unique** output Y.

In a **statistical model** we assume that the values of Y can be modeled by a function *plus* some "random noise" ϵ. The presence of the ϵ term allows for many different values of Y for the same set in inputs x.

$$Y = f(X) + \epsilon$$

Example: The relationship between ground snow and elevation in Utah (see figure 1).

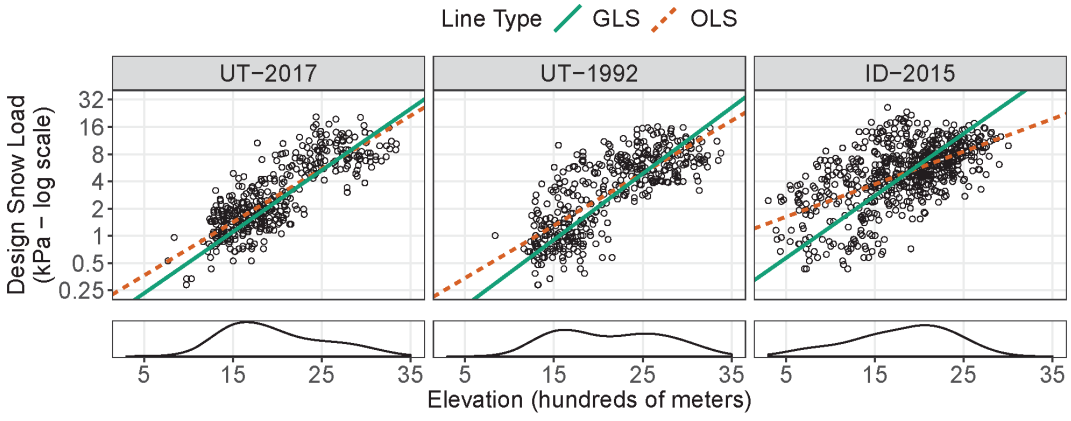

Figure 1: Plot of design ground snow loads (log-scale) vs elevation in and near Utah.

4 The Key Assumption (why this class exists)

The key assumption (and the foundation for this course) we make is that ϵ follows a probability distribution. Specifically we assume that

$$\epsilon \sim^{i.i.d} N(0, \sigma^2). \tag{1}$$

If this assumption is valid, then our **estimates** of the **model parameters** will come from well-defined probability distributions, which will allow us to determine if the linear relationships between our explanatory variables and our response variable are significant. This process is often called **statistical inference**.

(Groups) What do each of the symbols mean in (1) and why might they be important?

- **independence:** Knowing the value of one of the residuals should tell us nothing about the rest.
- **identically distributed:** Each of the residuals come from the same probability distribution.
- **zero mean:** The residuals have an average value of zero (i.e. the model is not biased).
- **constant variance:** The spread of the residuals is constant across all predictions.

Why they are important is something we will talk about for the next several weeks.

5 Why "Linear Regression"?

5.1 Why Linear?

Model are composed of:

- **coefficients:** These are *constant* values that are *estimated* to optimize the model fit.
- **variables:** These are the *observed* values, calculated from the data that we use to estimate parameters or make predictions.

A model is considered "linear" if it can be written as a sum of coefficients β multiplied by a set of variables x, i.e.

$$Y = \sum_i \beta_i X_i$$

This means that you can have nonlinear variables as long as the coefficients are linear.

(Individual) Which of the following models are linear and which are non-linear?

- $Y = \beta_0 + \beta_1 X_1 + \epsilon$
- $Y = \beta_0 + \beta_1 e^{X_1} + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 X_2 + \epsilon$
- $Y = \beta_0 + X_1^{\beta_1} + \epsilon$

First three are linear, last one is not.

5.2 Why Regression?

Based on concept that things tend to "regress" to the mean:
Example: heights of fathers vs sons:

- Tall fathers tend to have tall sons, but those sons will tend to be shorter than their fathers.
- Short fathers tend to have short sons, but those sons will tend to be taller than their fathers.
- Thus, a line comparing "standard deviations" of fathers and sons heights will have a slope approximately equal to one, while the **regression** line will have a slope that is less than one.
- Because things regress to the mean, the regression line will always be flatter than the standard deviation line.

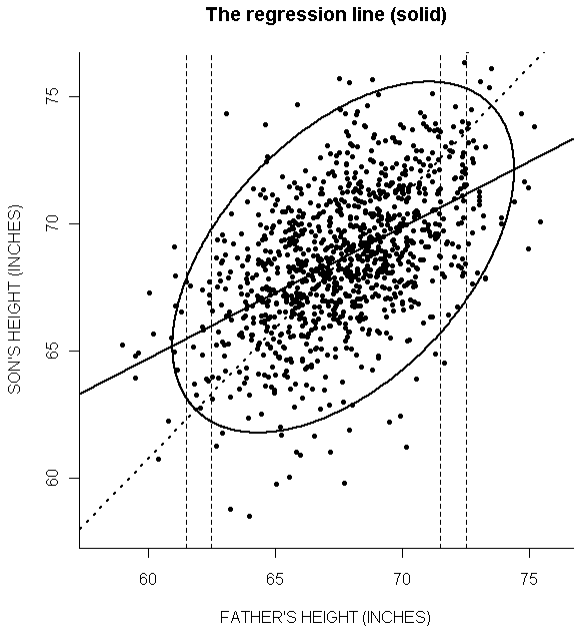

Figure 2: Plot of father vs sons' heights. The dotted line is the standard deviation line while the solid line is the regression line.