Welcome to Math 261A - Regression Theory and Methods

Let me introduce myself: My name is Martina Bremer, my office is in McQuarrie Hall (318A) and the best way to reach me is usually by e-mail (martina.bremer@sjsu.edu). I will try my best to reply to you within 24 hours (36 hours on weekends).

For prerequisites, grading policies and exam information please refer to the course syllabus. The schedule as well as many other materials will be available throughout the semester on our course web site on Canvas:

`https://sjsu.instructure.com`

Canvas is also the place to go to find the lecture notes, homework assignments (and later their solutions), data sets, office hour schedule and more. Review material for exams and material for the projects later in the semester as well as your homework, project, and exam grades will also be posted on Canvas.

If you should ever find yourself falling behind in this course please come and see me as soon as possible so that we can discuss all the options you have of catching up. Statistics classes tend to be very sequential. If you miss a section you will likely have difficulties following the subsequent sections. This is a graduate course. It will require a significant time committment (usually at least 9 hours per course per week). Appropriate (self-) motivation and work ethics are expected.

The Goals of this Course

Regression is one of the most frequently used analysis techniques in statistical practice. In general, a regression model represents the dependence of one or more dependent variables (response) on one or more independent variables (predictor). The predictor variables can be categorical or quantitative. The response is always quantitative. In this course, we will investigate the theory behind statistical regression models and learn how to draw conclusions from the models for practical data applications.

The simplest case is a model in which one quantitative response is modeled though one quantitative predictor. This case is called SIMPLE LINEAR REGRESSION.

Examples:

- The height of a young child can be modeled as a linear function of its age.
- The score on a test might be modeled as a linear function of the time spent studying.

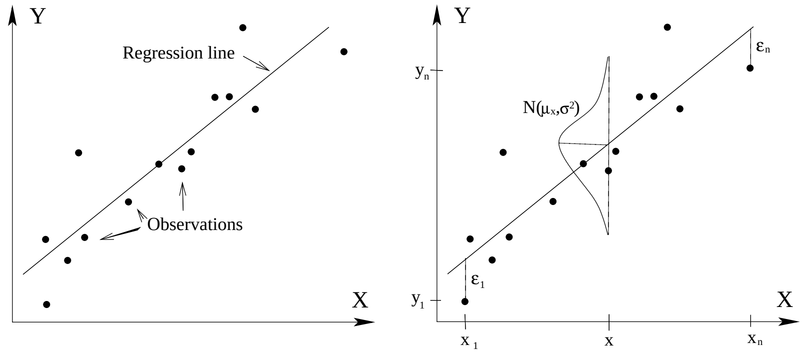
Question: Why are height and test score the natural responses in the examples above? Would it also make sense to make age or time the response? How do you know what to use as the response variable in an application?

The predictor variables in regression models are usually denoted with X and the response variables with Y. In general, capital letters (X, Y) are reserved for random variables and lower-case letters (x, y) for observations.

The simple linear regression model can be written as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 is the intercept and β_1 is the slope of the true regression model line shown in the scatter plot below.



Note, that not all the points that represent the observations (x, y) fall directly on the line. To account for this, the model includes a RESIDUAL term, denoted ϵ. The residuals are also sometimes called the ERROR TERMS of the regression model. A regression model is "statistical" because we make a probabilistic assumption on the ϵ terms. Usually, it is assumed that the residuals are independent and Normally distributed with mean zero and constant variance σ^2 (that does not depend on the value(s) of x or y).

There are several different methods with which the parameters (β_0, β_1, and σ) of a simple linear regression model can be estimated from the data (pairs of observations (x_k, y_k)). Not all regression models are "good". Sometimes it simply makes no sense to fit a line to the data. But a curve might make sense. We will investigate how to judge the quality or FIT of a regression model.

Note: Correlation between two variables can be established with regression models. If two variables are highly correlated, the observations on them will fall very close to a line and a linear regression will provide an excellent fit. However, correlation should not be confused with causation. Regression models can aid in establishing a cause and effect relationship between two variables, but they cannot be used as sole proof.

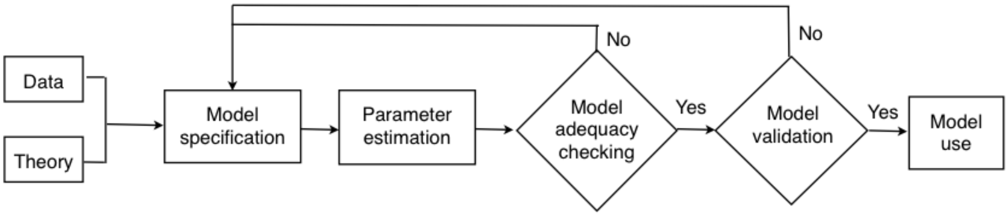
Uses of Regression

Regression models can be used for different purposes. Among the most common are

- Description of the data (graphical or with a linear equation)
- Parameter estimation (and interpretation in the context of the model)
- Prediction and estimation
- Control (if variables are related in a causal manner than one variable can be controlled by changing the value of another if their linear relationship is understood).

Building a Regression Model

For more complex regression models with more than one predictor and possibly also more than one response variable, finding the best fitting model is an iterative process.



Some steps of this process, such as parameter estimation are best conducted with a computer and software such as R. For other steps, human judgment is necessary. You will learn how to develop and apply such judgment in this course.

Regression in Data Science

Why is learning about regression modeling a useful tool in the arsenal of a statistician, data analyst, or a data scientist? In these occupations, your job will be to make sense of (usually large, complex, messy) datasets and to explain the relationships between variables. Sometimes (but not always) the relationships can be explained well through the use of linear models. The linear models you will encounter in this course include

- Simple linear regression
- Multiple linear regression (one response, many predictors, of which some may be categorical variables)
- Non-linear relationships between predictor(s) and response, e.g., polynomial regression or kernel regression

- Models for non-continuous response variables (e.g., logistic regression, Poisson regression etc).

In order to use any model to make sense of your data you need to know that the model exists, how to estimate its parameters, how to tell whether it is indeed a good fit for your specific data set (or how to make it fit better) and how to interpret the fitted model. This process is what you'll learn in this course. While each chapter of your textbook usually focusses on just one part of the process, in real life, you will have to be able to put all the parts together into one cohesive story to make sense of your data. You'll practice that skill during your projects.