Agenda

1. Introduction
   - Questionnaire
   - Syllabus, Objectives, Motivation
2. Warmup: categorical/quantitative, response/explanatory
3. Introduction to R, RStudio
4. Introduction to R Markdown
5. Statistical Modeling: Four-step process, C-F-A-U
6. Simple Linear Regression: Mathematical Form, Conditions

Course Objectives

- Modeling: construct an *appropriate* model based on the data
- Flexibility: modify your model to *adapt* to new data or knowledge
- Validity: understand and verify the *assumptions* upon which your model is based
- Inference: what types of *inferences* are justifiable?
- Communication: present your findings to a not-necessarily technical audience *clearly* in both written and oral forms
- Computational Skill: working with messy data, visualization, proficiency in R

Motivation Even though linear regression models are relatively simple, they are very widely used. Some examples:

- Switzer & Horton: multiple regression is the dominant statistical technique in medical research (CHANCE)
- Google Flu Trends: generates predictions for flu patients based on (unrelated) search terms using logistic regression
- Basically anything you read at FiveThirtyEight.com is based on a regression model

Warmup: 0.1 – Categorical or quantitative? Suppose that a statistics professor records the following for each student enrolled in her class:

- Gender
- Major
- Score on first exam
- Number of quizzes taken (a measure of class attendance)
- Time spent sleeping the previous night
- Handedness (left- or right-handed)

- Political inclination (liberal, moderate, conservative)
- Time spent on the final exam
- Score on the final exam

For the following questions, identify the response variable and the explanatory variables(s). Also classify each variable as quantitative or categorical. For the categorical variables, also indicate whether the variable is binary.

1. Do the various majors differ with regard to average sleeping time?
2. Is a student's score on the first exam useful for predicting his or her score on the final exam?
3. Do male and female students differ with regard to the average time they spend on the final exam?
4. Can we tell much about a student's handedness by knowing hos or her major, gender, and time spent on the final exam?

Introduction to R, RStudio, and R Markdown Let's walk through the introduction here:

http://www.math.smith.edu/ bbaumer/mth292-f14/lab-intro.html

- You will need to get comfortable with RStudio either on the server or on your local machine
- You will complete all of your homework assignments in R Markdown
- To do this you will need to upload the rendered HTML version of your document to Moodle
- Stat TAs in Burton 301 from 7-9 pm, Sunday through Thursday can help you!
- Familiarize yourself with the Resources page
- MinimalR handout

Statistical Modeling

Simple Linear Regression

Activity: regression and you Collect data from students?

```
require(mosaic)
url = "https://docs.google.com/spreadsheets/d/1wQ2HlKWImzi7fQlUwLXDkhpRp_P--4XGfpVSMkJmTEE/export?forma
ds = fetchGoogle(url)
```

```
xyplot(numFriends ~ distTravelled, data=ds, type = c("p", "r"))

## Error in eval(expr, envir, enclos):  could not find function "xyplot"
```

1 Instructor's Notes

Why Use Linear Models?

- Ubiquity of Data: all around us now, and much of it is poorly understood
- Simplicity: linear models are conceptually simple, but surprisingly powerful
- Versatility: linear models are surprisingly versatile. Non-linear models can be useful, but complicated non-linear models bring additional challenges
- Universality: anyone can build a model, but how many truly understand the information that is conveyed by this model?

Themes for the course

- Data $\neq$ information
- Correlation $\not\Rightarrow$ causation
- Just because you have a high R^2, that doesn't mean your model is valid
- There is no perfect model – we're looking for a delicate balance of simplicity and effectiveness
- Your model is worthless if you can't convince someone else of your findings