Agenda

1. Simple Linear Regression: Mathematical Form
2. Parameters and statistics
3. Data collection
4. R and RStudio intro

Model notation

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

- Y: response variable
- f: a function that makes up the model – there are *infinitely* many possible functions!
- X: explanatory variables (i.e. the data)
- ϵ: error or noise term

Simple Linear Regression Simple linear regression is a special case of the model above, wherein:

- Y is a quantitative variable
- f makes a *line*!
- X is a single quantitative variable
- $\epsilon \sim N(0, \sigma_\epsilon)$, where σ_ϵ is fixed

Notation for this may vary. Note that:

- Y, X: the (idea of the) quantities (e.g. Age, Price)
- y, x: specific values of those quantities (e.g. 22, $35,748)
- y_i, x_i: the ith specific value of a collection
- $\bar{y}, \bar{x}$: average (mean) value of a collection of values

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- $\hat{y}$: an *estimate* of an unknown value based on the model

We can write the model as follows:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon, \; \epsilon \sim N(0, \sigma_\epsilon)$$

and we estimate it using the following notation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

Parameters and statistics If we have time, we'll go through this fun explanation of polling, `http://rocknpoll.graphics/`. It's focused on election polling (quite relevant today!) but also helps show the difference between population parameters and sample statistics.

Data collection We are going to choose some variables we want to study about our class as a whole. This isn't going to be a sampling activity (hopefully, we will capture data about the entire class) but we will use linear modeling as a descriptive method. First, we need to brainstorm some variables to collect. Last year, the class collected variables about students credits for the semester, number of siblings, number of pairs of shoes, number of facebook friends, and distance travelled to Smith after winter break. But, we don't have to use those!
Go to this spreadsheet and enter your information.

R and RStudio intro We're going go through the first lab in the course, and we'll see how far we get. Depending on time, we'll do some analysis of your data!

```
install.packages(c("mosaic", "dplyr", "ggplot2", "Stat2Data", "car", "googlesheets"))
require(mosaic)
require(car)

data(Salaries)
str(Salaries)

histogram(~salary, data=Salaries)
densityplot(~salary, data=Salaries)
barchart(tally(~rank, data=Salaries))

require(ggplot2)
ggplot(Salaries) + geom_histogram(aes(x=salary))
ggplot(Salaries) + geom_density(aes(x=salary))
ggplot(Salaries) + geom_bar(aes(x=rank))

xyplot(salary~yrs.since.phd, data=Salaries)
bwplot(salary~sex, data=Salaries)
mosaicplot(sex~rank, data=Salaries, las=1)

ggplot(Salaries) + geom_point(aes(x=yrs.since.phd, y=salary))
ggplot(Salaries) + geom_boxplot(aes(x=sex, y=salary))

help(xyplot)
?geom_histogram

mean(~salary, data=Salaries)
sd(~salary, data=Salaries)
favstats(~salary, data=Salaries)
tally(~rank, data=Salaries)

require(googlesheets)
url = gs_url("https://docs.google.com/spreadsheets/d/111z767D_pbHg58QCfndbmB22wki86fk5v4DGPTb7Nkc/pub?g
ds = gs_read_csv(url)

xyplot(distance_home ~ num_countries, data=ds, type = c("p", "r"))
mod = lm(distance_home ~ num_countries, data=ds)
summary(mod)
```