Overview of regression

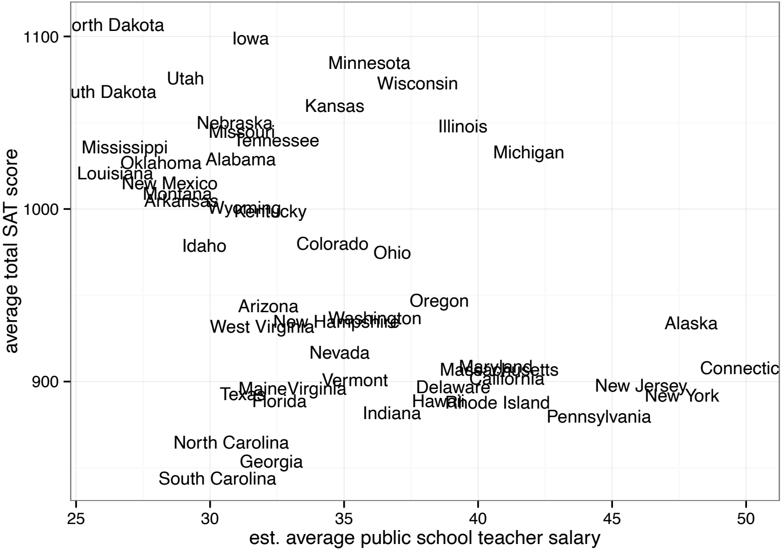
Author: Nicholas G Reich, Jeff Goldsmith

This material is part of the ***statsTeachR*** *project*

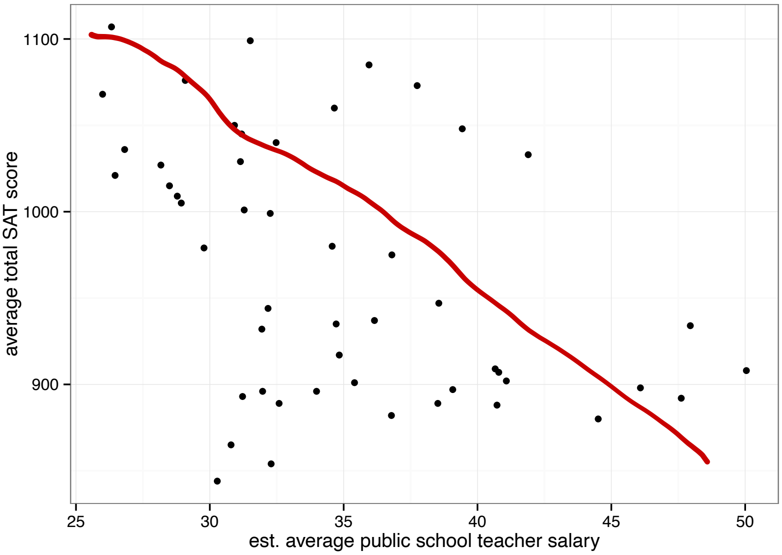
What is regression?

An informal introduction...

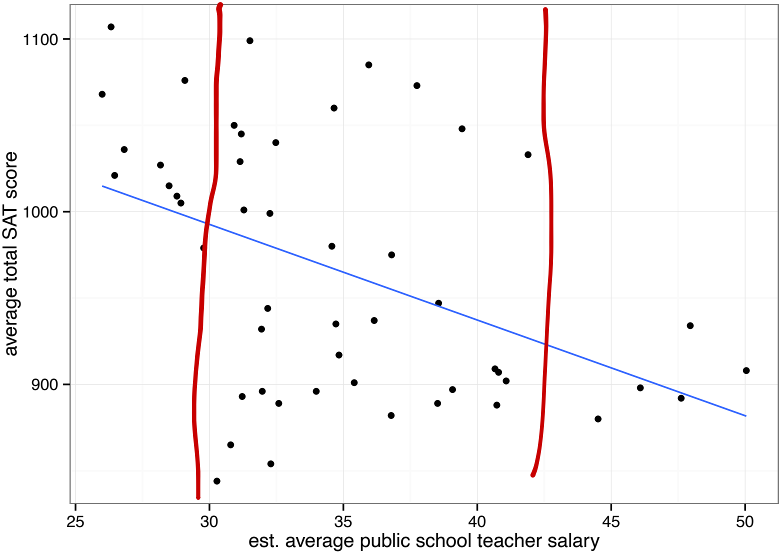
State-level SAT score data (1994-95)



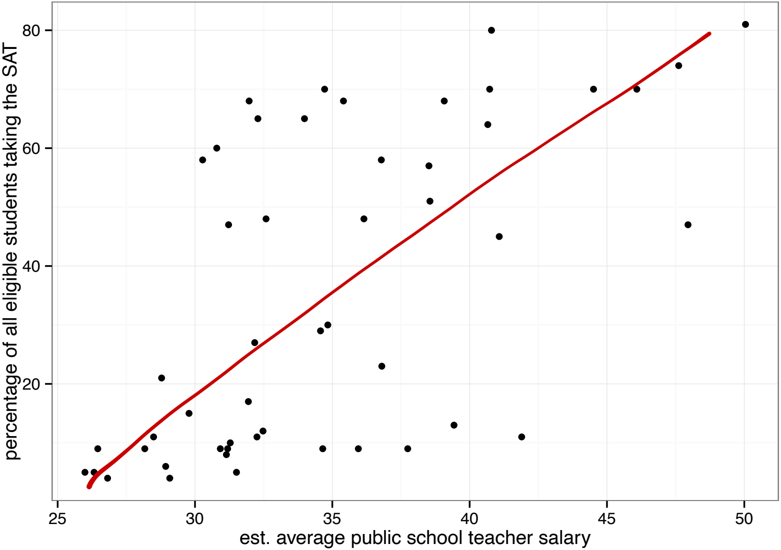
State-level SAT score data (1994-95)



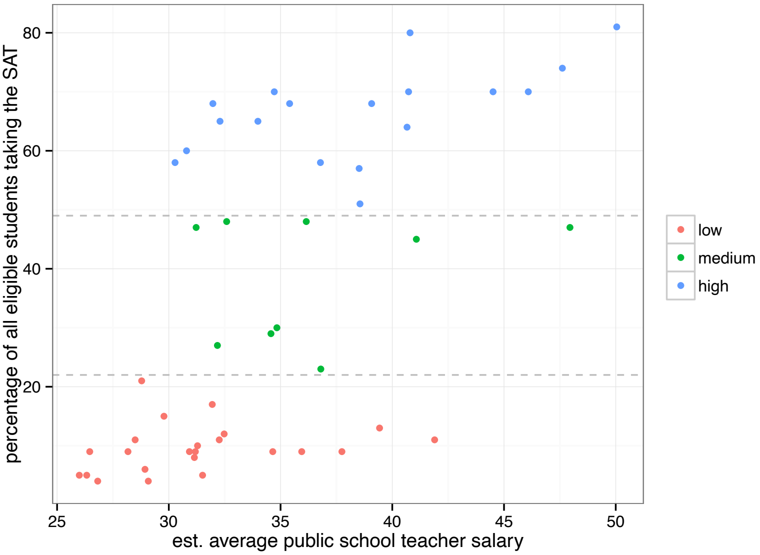
State-level SAT score data (1994-95)



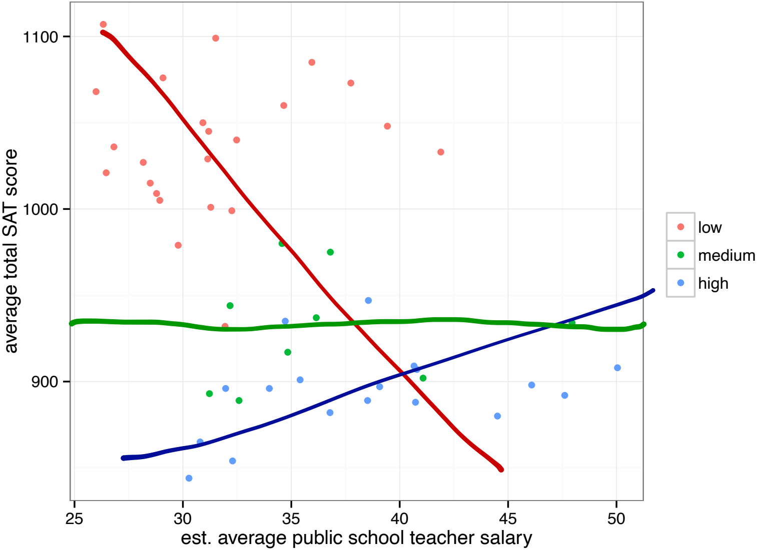
State-level SAT score data (1994-95)



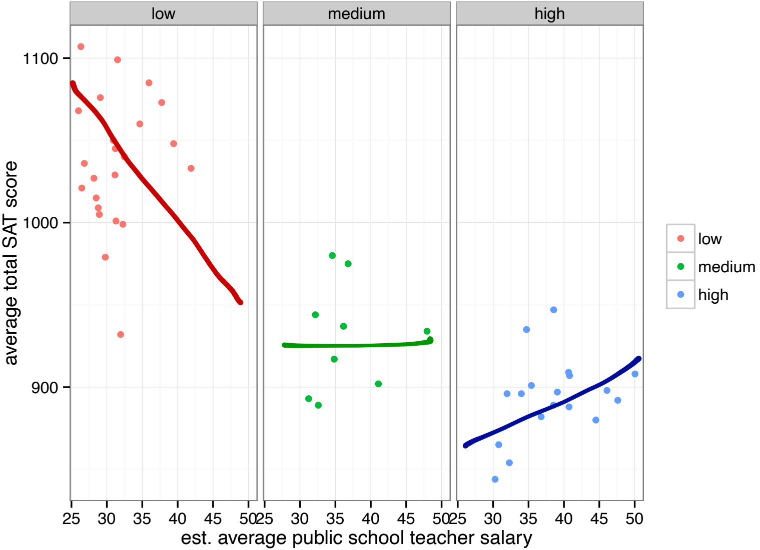
State-level SAT score data (1994-95)



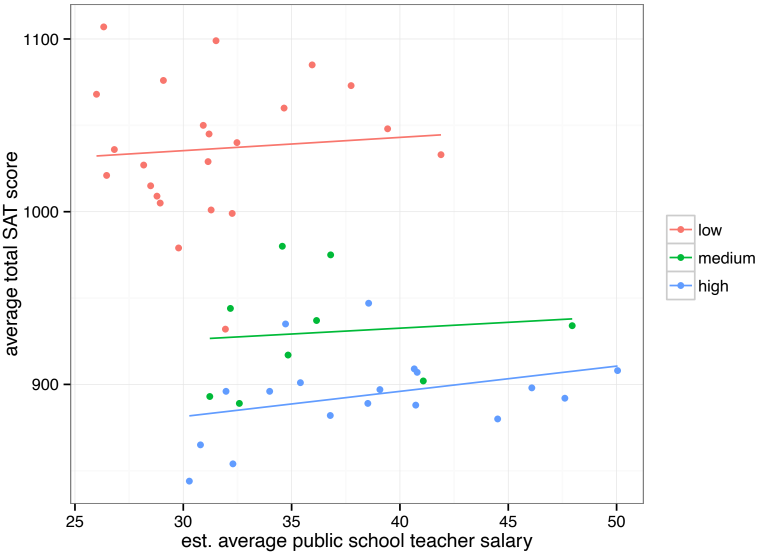
State-level SAT score data (1994-95)



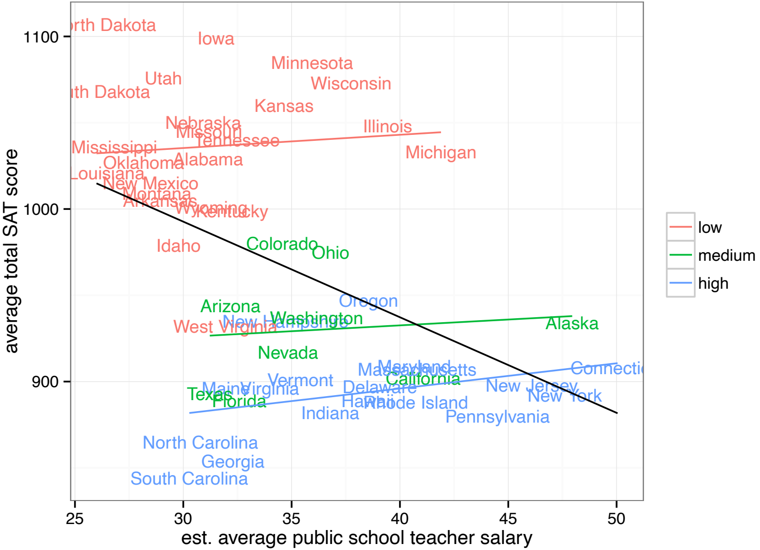
State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



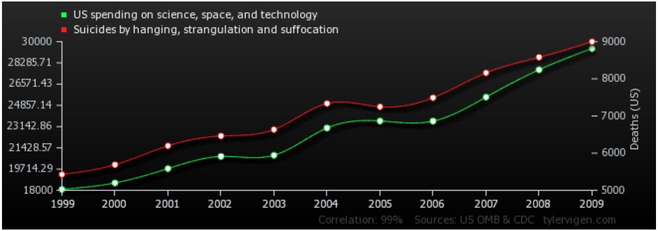
State-level SAT score data (1994-95)



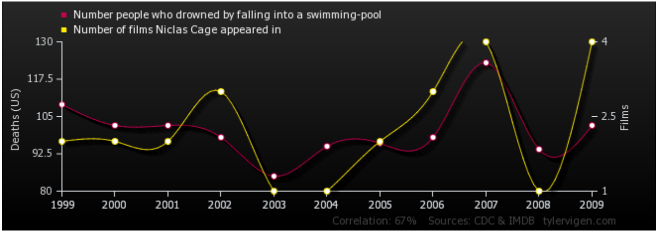
State-level SAT score data (1994-95)

What can we conclude from all of this? (BTW, this is an example of "Simpson's Paradox".)

Beware of correlation!



Beware of correlation!



1

[1] Hat tip to www.tylervigen.com

What is regression?

[Now, more formally...]

"...to understand as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors." – Cook and Weisberg (1999)

Good overview on Wikipedia.

What is regression?

- The goal is to learn about the relationship between a covariate (predictor) of interest and an outcome of interest.
  - Focus on prediction
  - Focus on description
- Regression is an exercise in inferential statistics: we are drawing evidence and conclusions from data about "noisy" systems.

Example data: heights of mothers and daughters

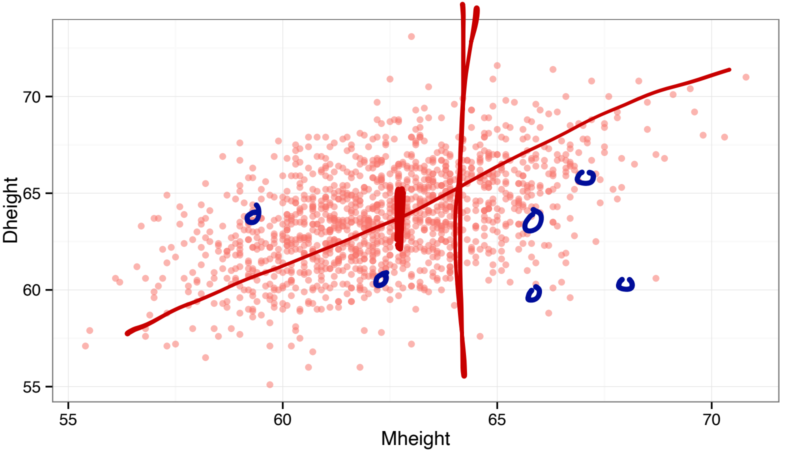
Heights of $n = 1375$ mothers in the UK under the age of 65 and one of their adult daughters over the age of 18 (collected and organized during the period 1893–1898 by the famous statistician Karl Pearson)

```
require(alr3)
data(heights)
head(heights)

##   Mheight Dheight
## 1    59.7    55.1
## 2    58.2    56.5
## 3    60.6    56.0
## 4    60.7    56.8
## 5    61.8    56.0
## 6    55.5    57.9
```

Example data: heights of mothers and daughters

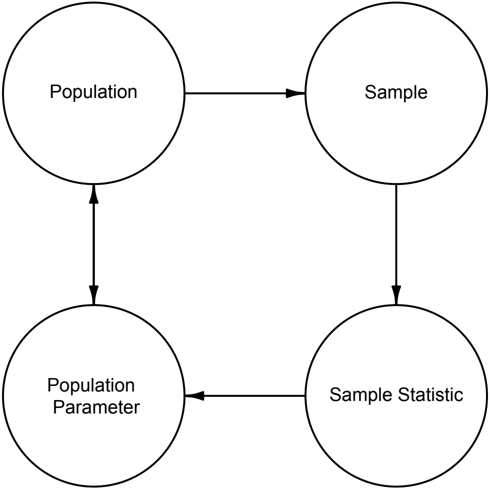
```
require(ggplot2)
qplot(Mheight, Dheight, data=heights, col="red",
      alpha=.5) + theme(legend.position="none")
```



Circle of Life

μ

β



$\bar{x}$

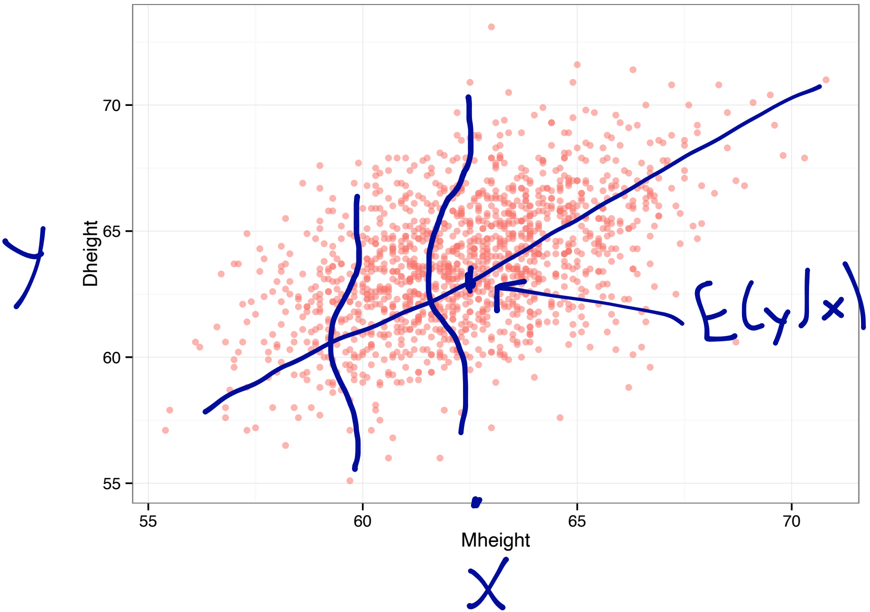
$\hat{\beta}$

What we want in regression

Given some data $y, x_1, x_2, \ldots x_p$, we are interesting finding a likely value for y given the value of predictors $x \equiv x_1, x_2, \ldots x_p$.

- Often, but not always, y is continuous. (Called outcome, response, "dependent variable").
- The x's can be continuous, binary, categorical. (Called predictor, covariate, "independent variable").
- We want $\mathbb{E}(y|x) = f(x)$; we observe $y = f(x) + \epsilon$.

Example data: heights of mothers and daughters



Regression model

The process of using data to describe the relationship between outcomes and predictors is called modeling.

- Models are models, not reality.
- "All models are wrong, but some are useful."
- Introduce structure to $f(x)$ to make the problem of estimation easier (this also introduces elements not found in the data, including judgement calls about important features and assumptions about the world).
- We largely focus on *parametric models* $f(x) = f(x; \beta)$ and worry about estimating β.

Linear Regression Models

intercept

A linear regression model is a particular type of parametric regression.

- Assume $f(x;\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$

  slope
- Focus is on $\beta_0, \beta_1, \ldots$
- "Linear" refers to the β's, not the x's:
  - $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model
  - $f(x) = \beta_0 + x^{\beta_1}$ is not

linear

not linear

Why is linear regression so popular?

- Easy to implement
- Lots of theory
- Straightforward interpretations
- Surprisingly flexible
- Good approximation in many cases

What do we need to assume?

Typical assumptions for a SLR model

- A1: The model: e.g. $y_i = f(x_i; \beta) + \epsilon_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$
- A2: Unbiased errors: $\mathbb{E}[\epsilon_i | x_i] = \mathbb{E}[\epsilon_i] = 0$
- A3: Uncorrelated errors: $cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
- A4: Constant variance: $Var[y_i | x_i] = \sigma^2$
- A5: Probability distribution: e.g. $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- A6: Representative sampling: generalize to population.

Things to come

- Where do estimates $\hat{\beta}_0, \hat{\beta}_1$ come from?
- How do we draw inference about these estimates?
- What about more complex models?