Linear Regression Models
P8111

Lecture 01

Jeff Goldsmith

January 19, 2016

THE DEPARTMENT OF BIOSTATISTICS

Columbia University MAILMAN SCHOOL OF PUBLIC HEALTH

Today's Lecture

- Intro to regression
- Getting started with R

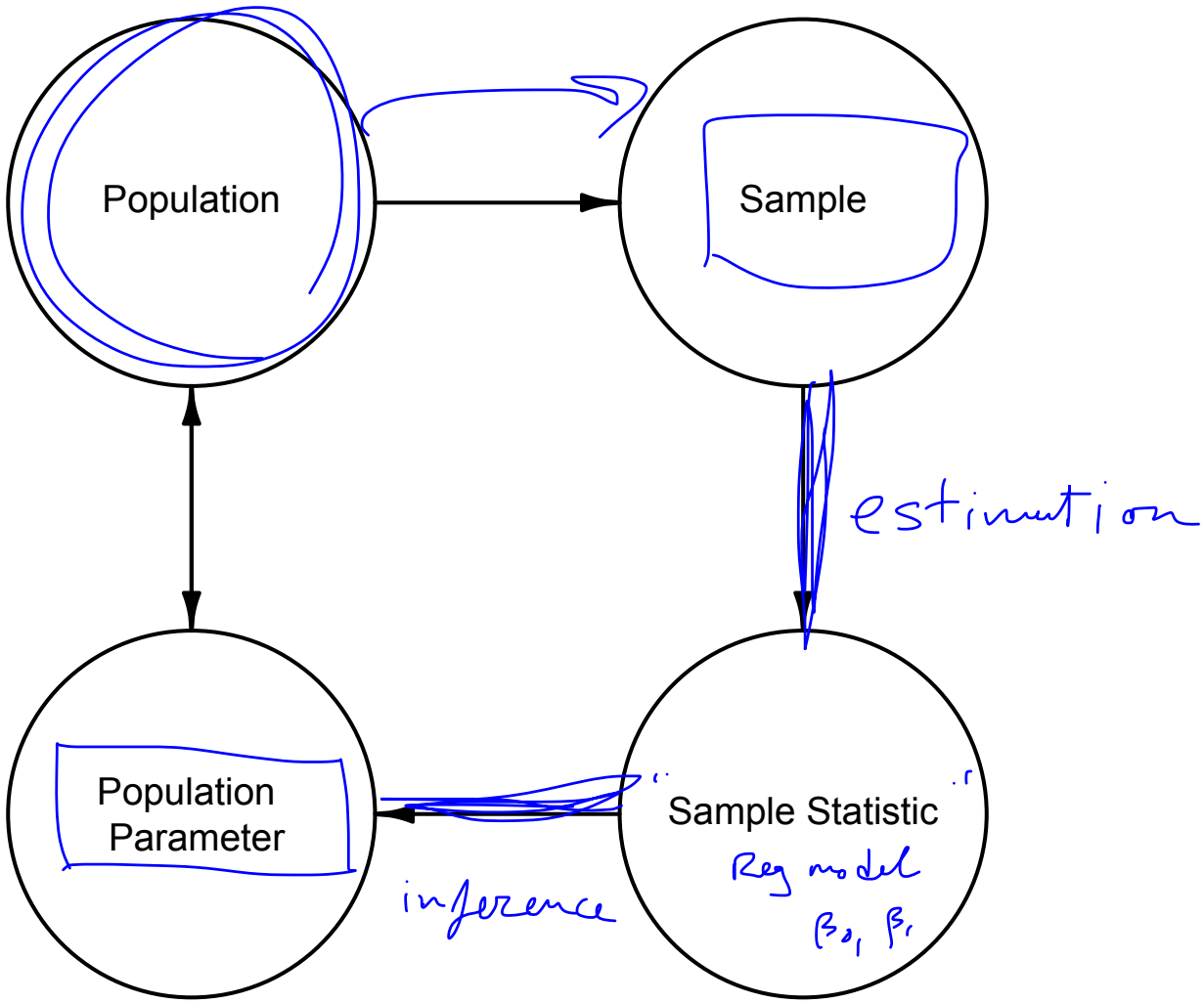
What is regression?

"...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)

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
- Linear regression modeling is often (though not always) focused on descriptive and inferential statistics

Circle of Life



What we want in regression

$[y|x]$

$E[y|x]$

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 $x \equiv x_1, x_2, \ldots x_p$.

- For this course, y is continuous. (Called outcome, response, "~~dependent variable~~").
- The x's can be continuous, binary, categorical. (Called predictor, covariate, "~~independent variable~~").
- We want $E(y|x) = f(x)$; we observe $y = f(x) + \epsilon$.

$E[y|x] + \epsilon$

Regression model

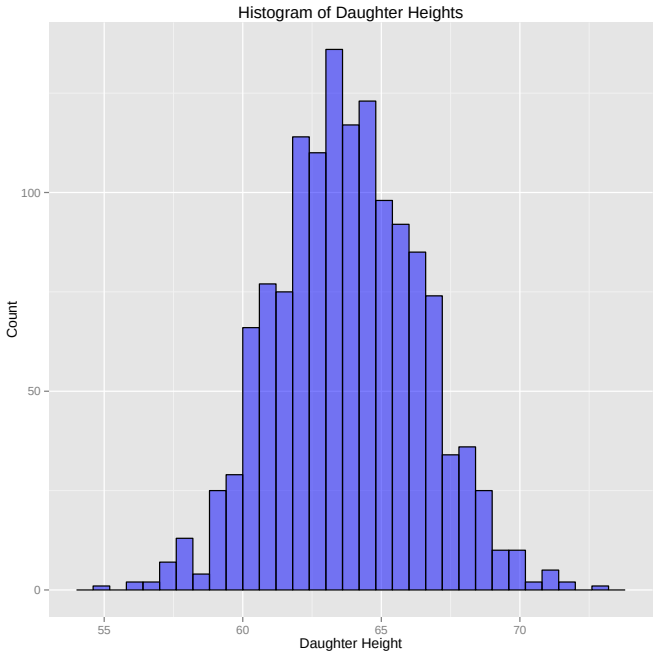
The process of using data to describe the relationship between outcomes and predictors is called modeling.

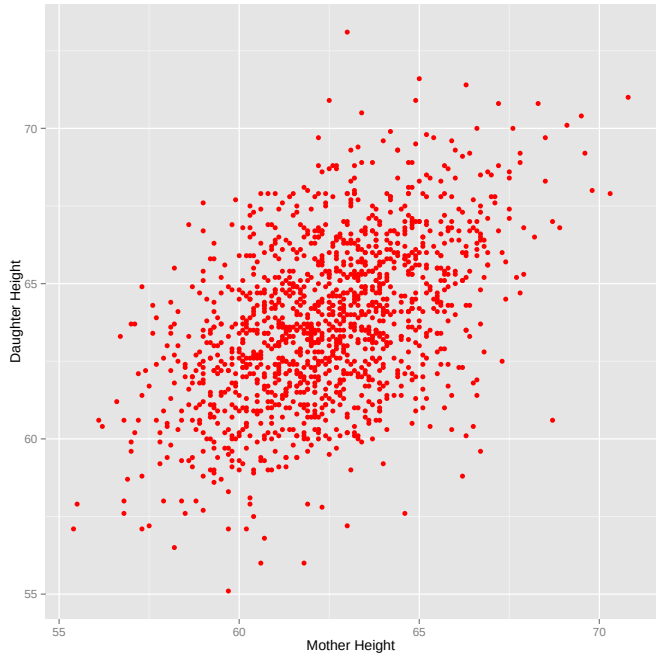
- Models are models, not reality.
- "All models are wrong, but some are useful."
- Introduce structure to $f(x)$ to make the problem of estimation easier (this also introduces elements not found in the data, including judgement calls about important features and assumptions about the world).
- We largely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

Example

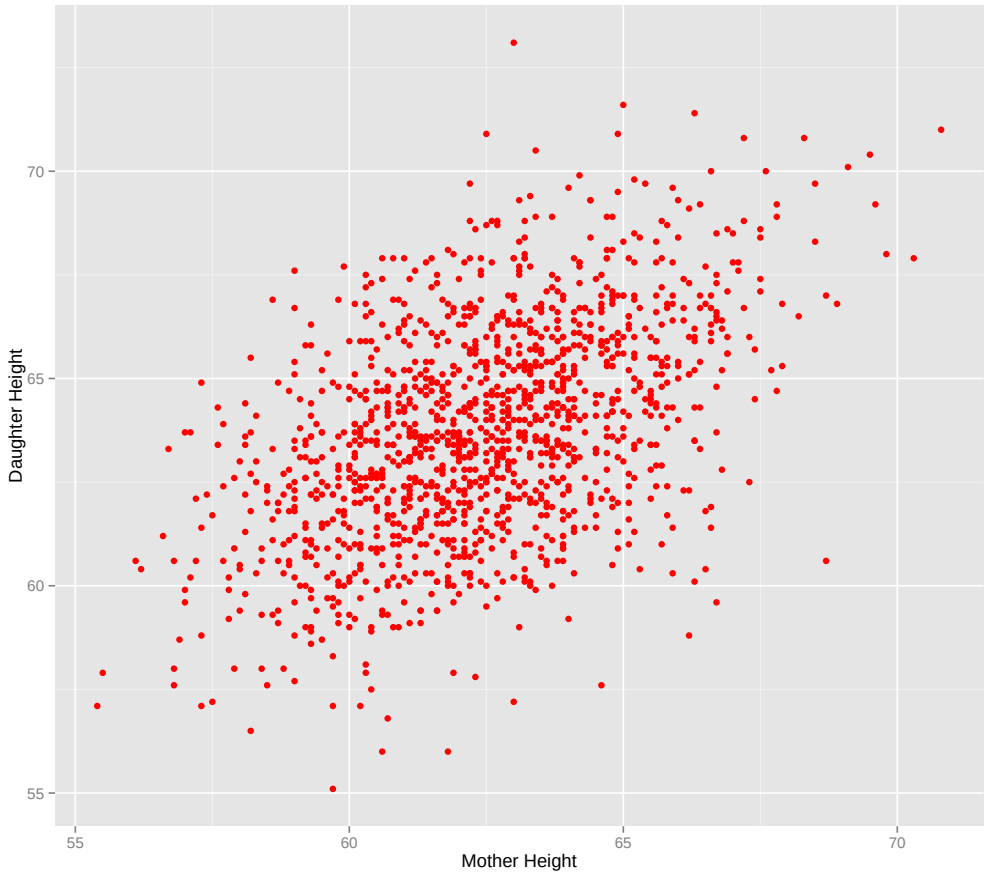
- Data on heights of $n = 1375$ mothers in the UK under the age of 65 and one of their adult daughters over the age of 18 (collected and organized during the period 1893–1898 by the famous statistician Karl Pearson)
- A historical use of regression to study inheritance of height from generation to generation.

Example

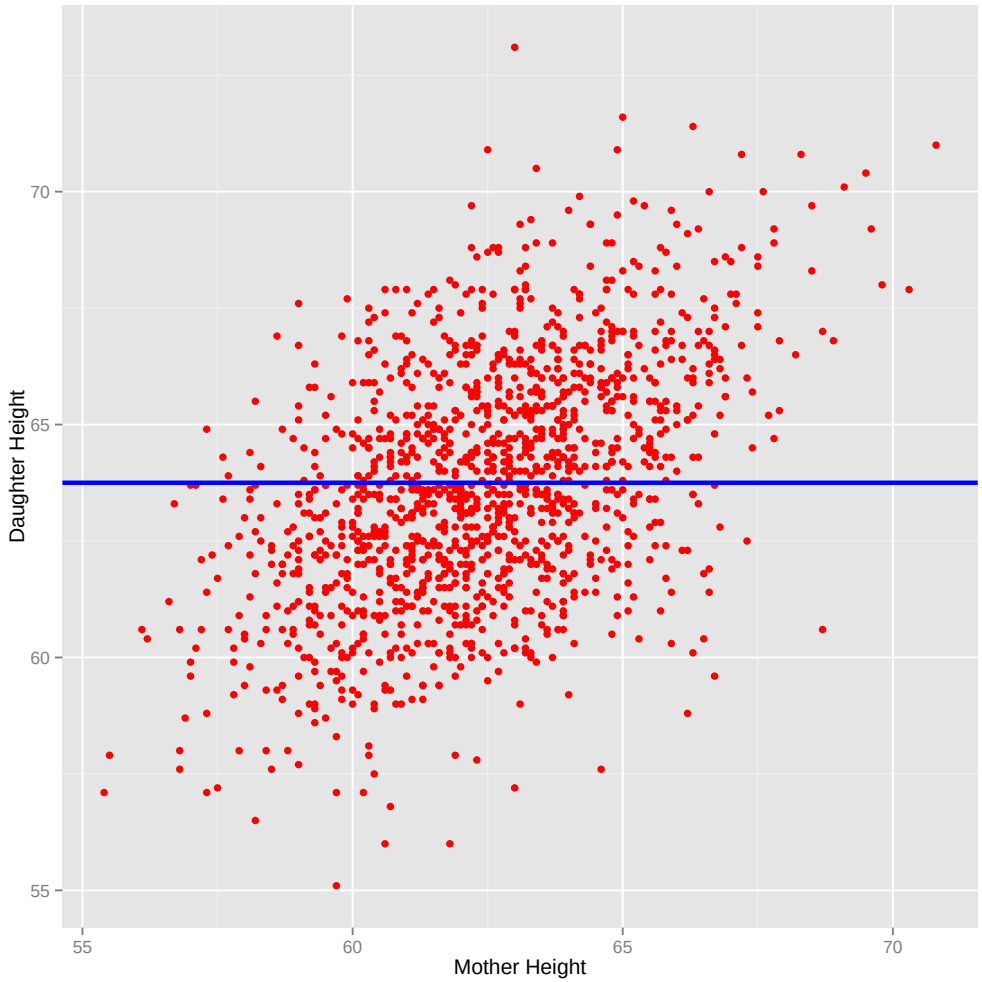

Histogram of Daughter Heights
Count
100
50
0
55
60
65
70
Daughter Height


Daughter Height
70
65
60
55
55
60
65
70
Mother Height

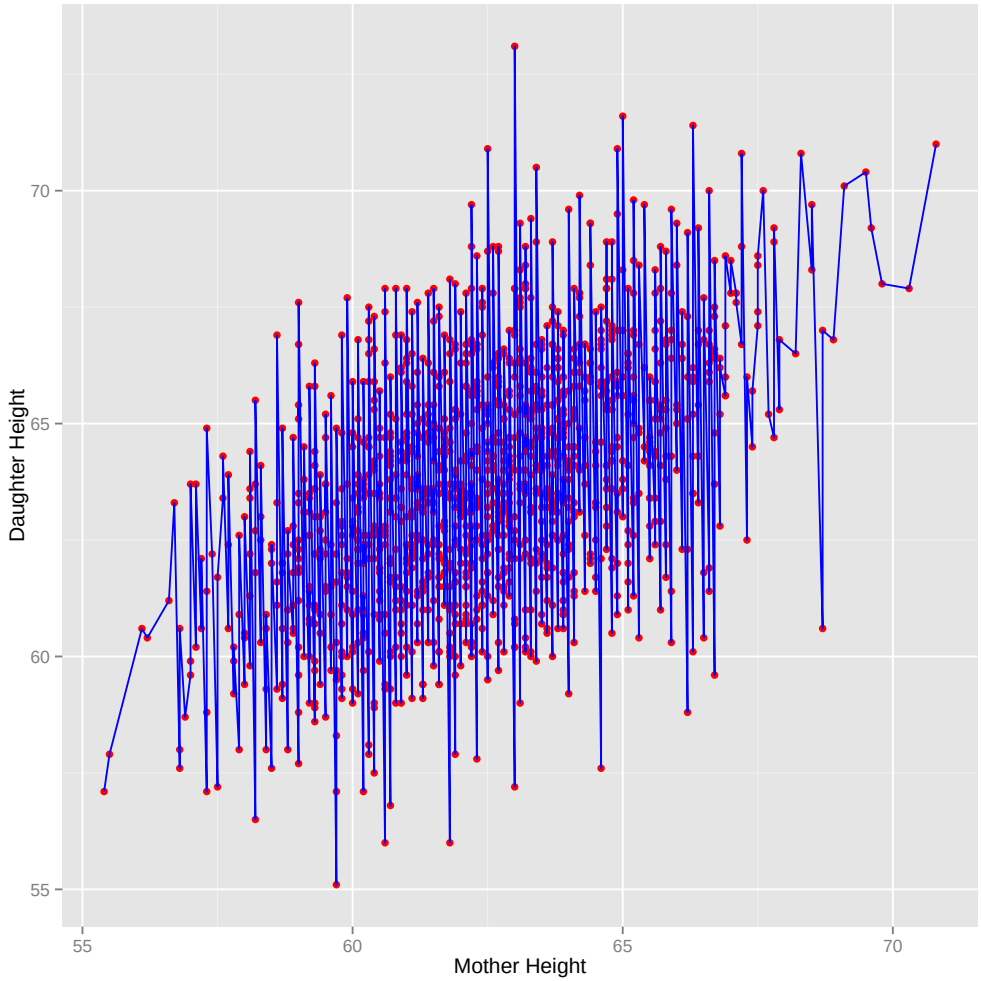
Scatterplot



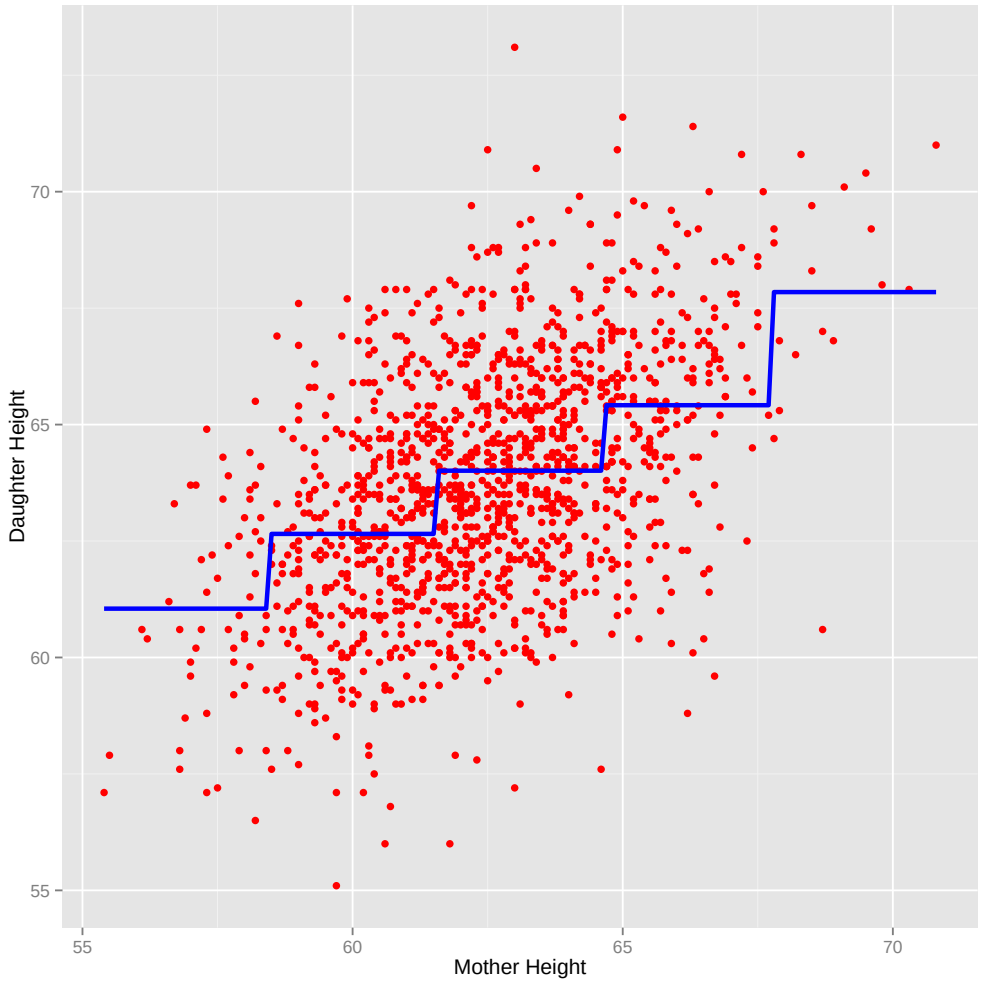
Regression: Mean



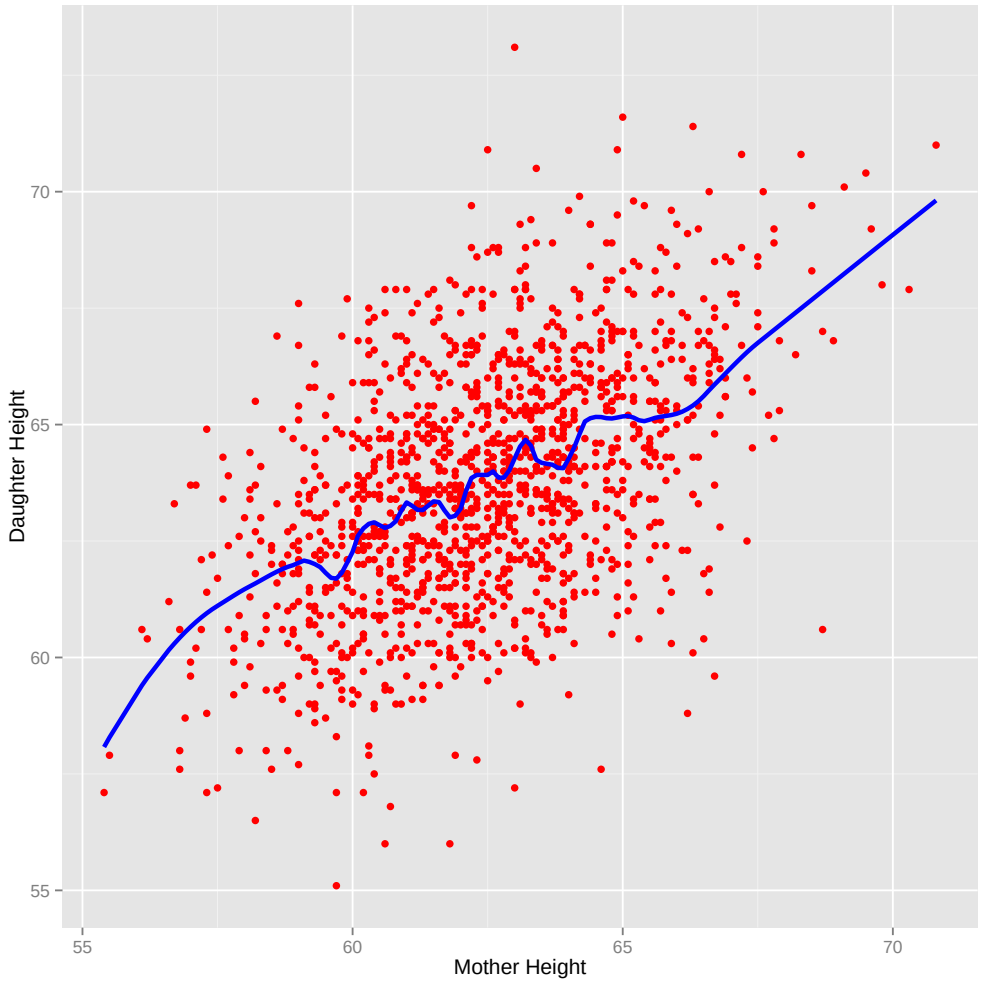
Regression: Interpolation



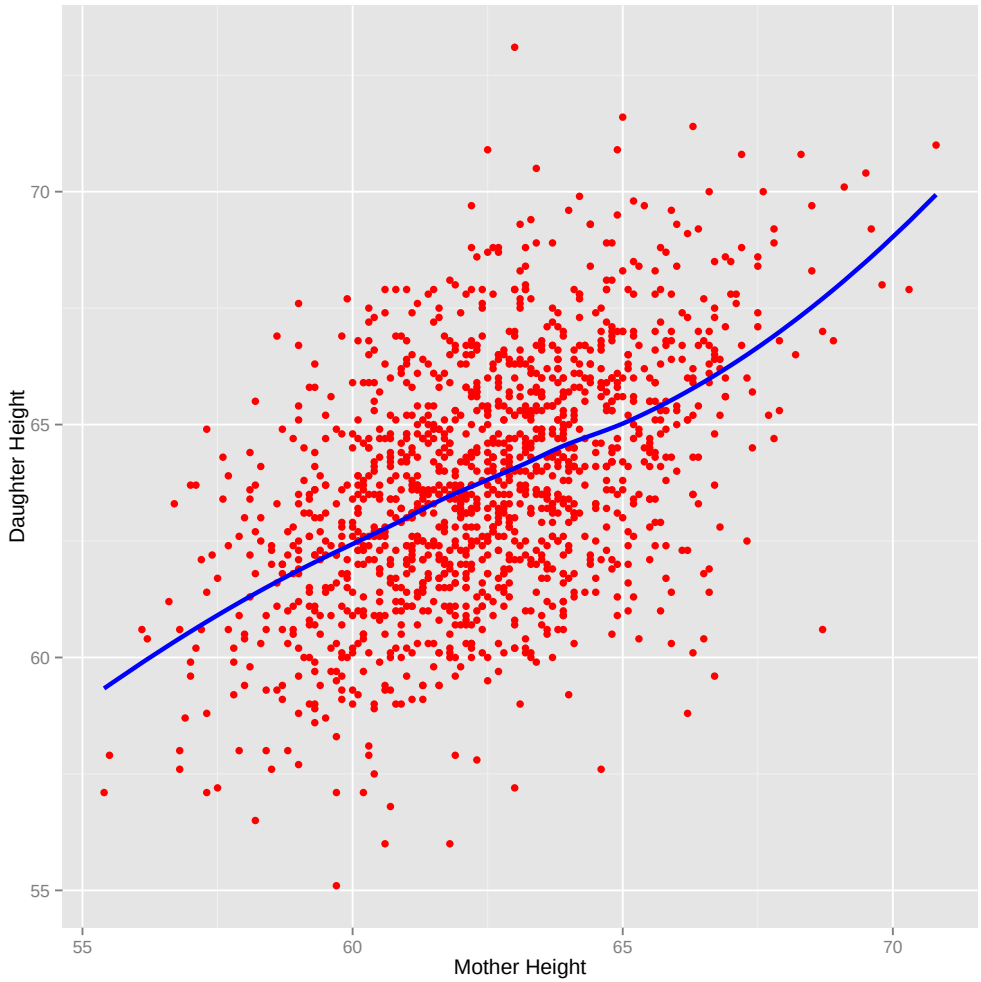
Regression: Bin Means



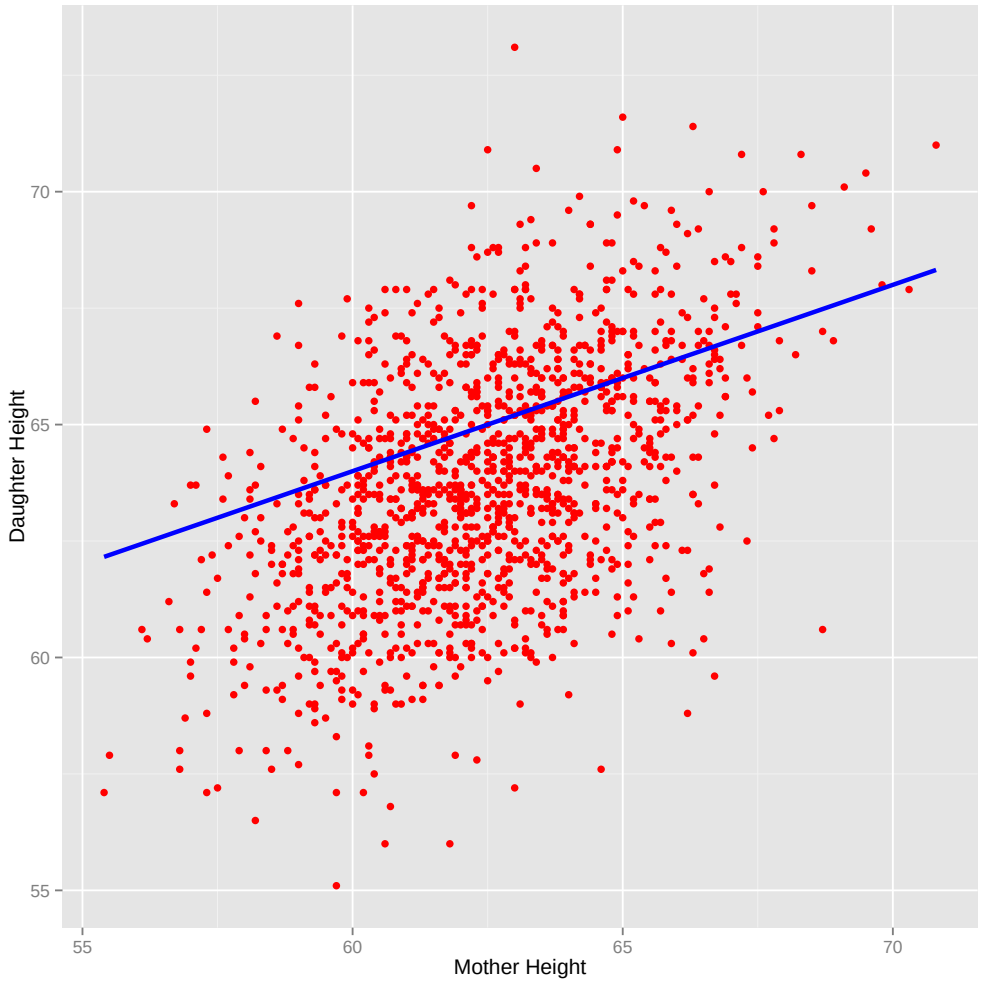
Regression: Curve



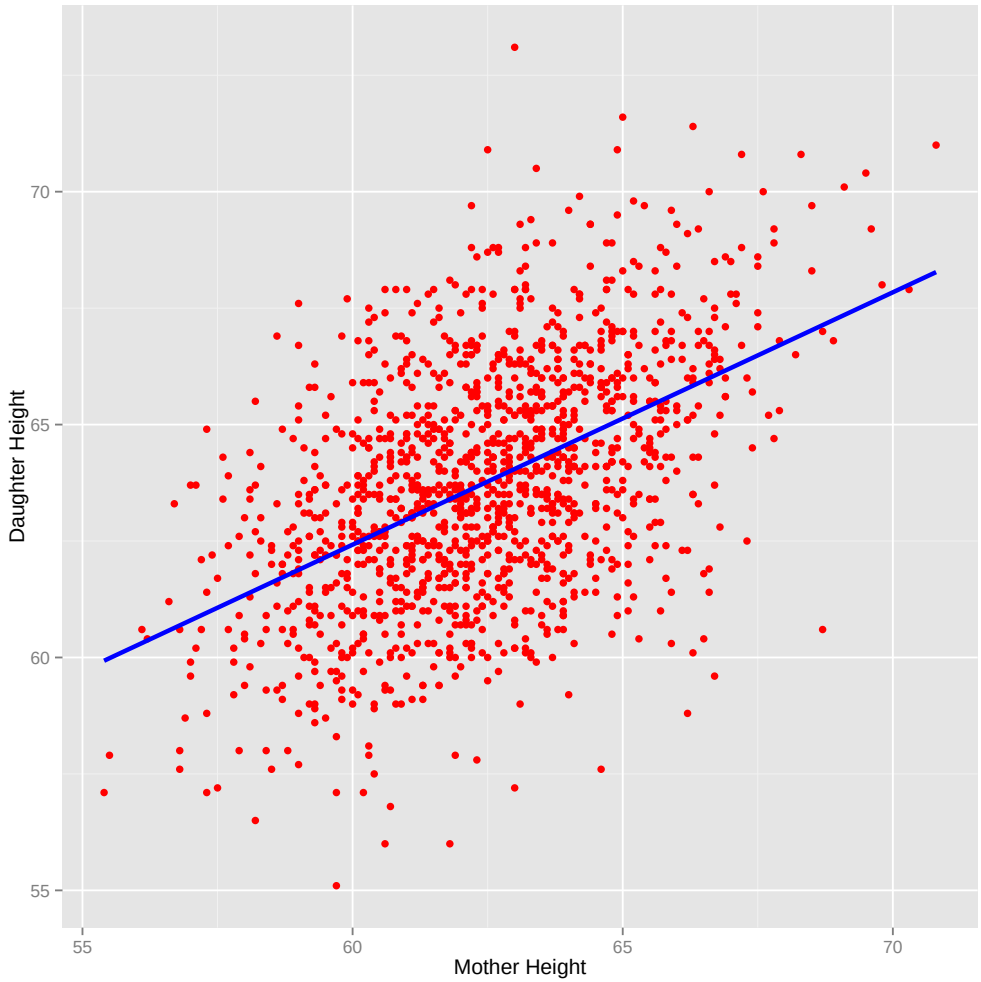
Regression: Curve 2



Regression: Line



Regression: Line 2



Lessons

- Lots of possible models
- Tradeoffs between flexibility and interpretability
- Concerns about over- and under-fitting
- Choices will often depend on context

Linear Regression Models

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

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

Why is linear regression so popular?

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

Height example

- Model is $DH_i = \beta_0 + \beta_1 MH_i + \epsilon_i$
- (Equivalently, $E(DH_i \mid MH_i) = \beta_0 + \beta_1 MH_i$)
- Estimate β_0, β_1

Height example

```
> library(alr3)
> library(broom)
> data(heights)
> linmod = lm(Dheight ~ Mheight, data = heights)

> summary(linmod)
Call:
lm(formula = Dheight ~ Mheight, data = heights)

Residuals:
   Min     1Q Median     3Q    Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.91744    1.62247   18.44   <2e-16 ***
Mheight      0.54175    0.02596   20.87   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> tidy(linmod)
         term  estimate  std.error statistic      p.value
1 (Intercept) 29.917437 1.62246940  18.43945 5.211879e-68
2     Mheight  0.541747 0.02596069  20.86797 3.216915e-84
```

Coding

- You can't *do* regression without coding
- More generally, much of the rest of your professional life will involve coding
- $\Rightarrow$ Your life will be easier the better you get at coding
- For this class, we'll use R exclusively

R

Before next class:

- Download R from the page

  `http://www.r-project.org/`

- Download and install R Studio
- Check out `swirl`'s R programming module (esp 1, 2, 3, 4, 6, 7, 8, 9)

swirl

```
> install.packages("swirl")
> library(swirl)
> swirl()
```

Then, follow instructions on screen ...

General coding tips

- Get better at using computers
- Get better at using google

General coding tips

- R will do *exactly* what you tell it to
  - Case matters; typos matter; order matters.
- Clear your workspace before you do anything
  - Re-run everything from scratch to make sure it still works

General coding tips

- Your most frequent collaborator is you from three months ago ... and you from three months ago never responds to email
  - Write code that you will understand in six months
  - Use comments liberally
  - Treat your code as something that has real value (because it does)

General coding tips

- Good organization will save you tons of headache (and heartache)
  - Put stuff in a reasonable place
  - Call stuff reasonable things
  - "HW 1.R" is bad; "20160120_Goldsmith_P8111_HW1.R" is good (and is better if it lives at " /Documents/P8111/Assignments/HW1/")
  - Variables `x1, x2, x2.1, x2_final` are bad; variables `age, gender, gender_as_num` are good (or at least better)

General coding tips

- If fixing a bug (or changing a decision) at the beginning of your code breaks everything that comes after it, you have bad code

Today's big ideas

- Intro to regression
- Intro to coding

- Faraway Ch 1, 2.1; ISLR 2.1, 2.3, 3.1
- `swirl`
- STAT 545 "Deep thoughts..."
- R Programming for Data Science (ch 4, 5)
- Elements of Data Analytic Style