ISTA 116: Statistical Foundations for the Information Age

Simple Linear Regression

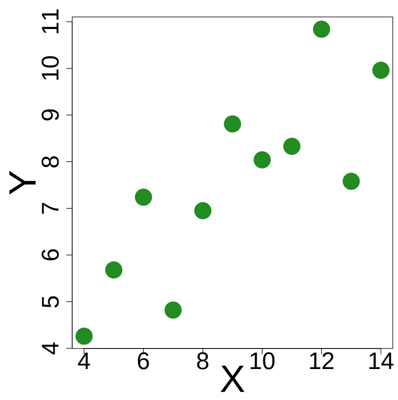
3, 5 and 7 October 2011

Outline

Prediction

- Correlations give us a **description** of the relationship between two numeric variables.
- However, when two variables are related, we can go further and use knowledge of one to make **predictions** about the other.
- Examples:
  - Use SAT scores to predict college GPA
  - Use economic indicators to predict stock prices
  - Use credit score to predict probability of default on a loan
  - Use biomarkers to predict disease progression
  - What else?

What's a Good Prediction?

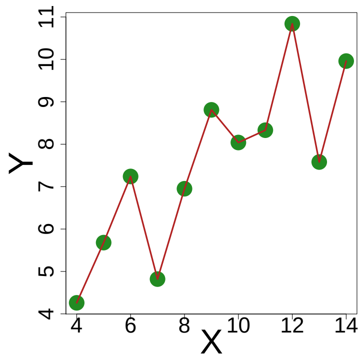


- Suppose I have this data.
- What would be a good prediction if I get a new X value of 12?
- What about an X value of 5.5?
- Why?

Modeling relationships with a function

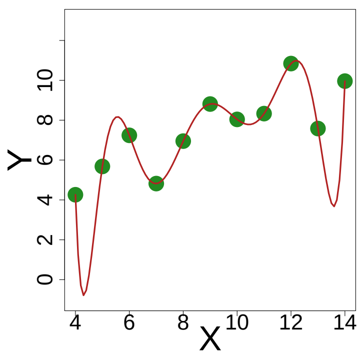
- We can capture all of our predictions by writing the y variable as a **function** of the x variable
- What's a function again?
- For every possible x value we put in, we get a single y value out.
- Examples:
  - $f(x) = x^2$
  - $f(x) = 1.6x + 20$
  - $f(x) = 5\cos(2\pi x)$

What's a Good Prediction?



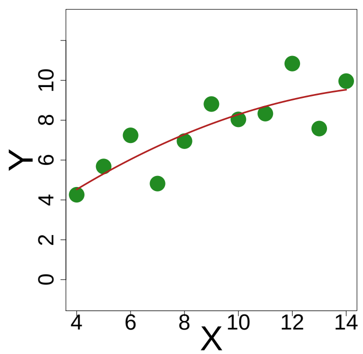
- How about this function?

What's a Good Prediction?



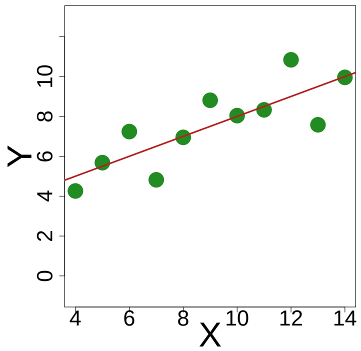
- Or this?

What's a Good Prediction?



- What about this?
- There's a tradeoff between how well we can fit the data and how simple our **model** (i.e., prediction function) is.

What's a Good Prediction?



- Pretty much the simplest model we can have is a straight line.
- Two things determine what line we have:
  - The intercept
  - The slope

Intercept Slope Form

- The intercept and slope are the **parameters** of our regression model.
- We denote them using β_0 for the intercept and β_1 for the slope.
- The general equation for a line is:

$$f(x) = a + bx$$

- In statistics notation, we write $\hat{y}$ ("y hat") to represent our *predicted* value of y.
- Given a value x_i, we predict using:

$$\hat{y} = \beta_0 + \beta_1 x_i$$

Hat Notation

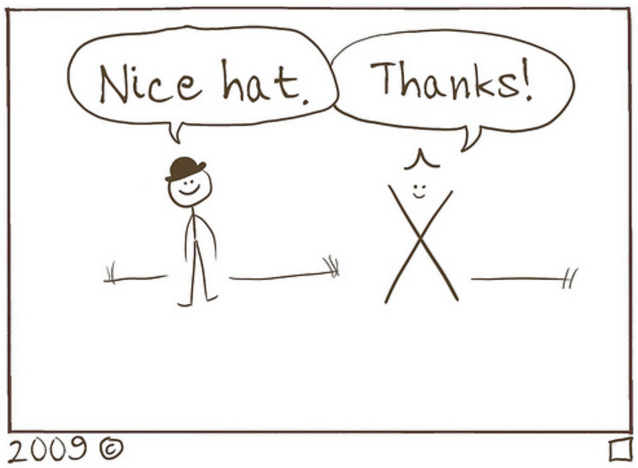


Figure: Source: `brownsharpie.com`

Systematic vs. Random

- We can split up each y value into two parts: a systematic (predictable) part and a "random" part.
- That is, we can write, for the y coordinate of the i^{th} data point:

$$y_i = f(x_i) + \varepsilon_i$$

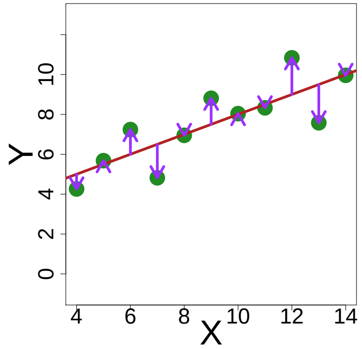
  or

$$y_i = \hat{y}_i + \varepsilon_i$$

  where ε_i represents the part of y_i that isn't predictable by knowing x.
- If we had the model already, our prediction would be off by ε.

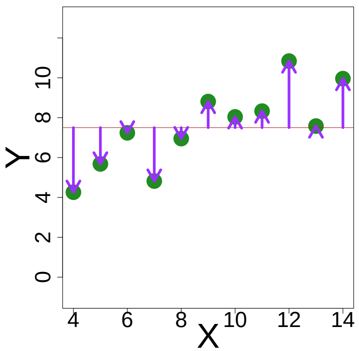
An Experiment

What's a Good Prediction?



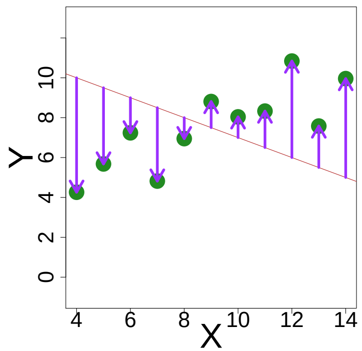
- Every line will have a different set of errors associated with it.

What's a Good Prediction?



- Every line will have a different set of errors associated with it.

What's a Good Prediction?



- Every line will have a different set of errors associated with it.
- Which is best?
- Intuitively, we want to minimize the overall "distance" between the line and the points.

Least Squares

- Remember our geometric analogy for standard deviation?
- Can think about total distance between predicted and observed values with a generalization of the Pythagorean theorem:

$$\begin{aligned} \text{Distance} &= \sqrt{(y_1 - \hat{y}1))^2 + \cdots + (y_n - \hat{y}_n)^2} \\ &= \sqrt{(\varepsilon_1^2 + \cdots + \varepsilon_n^2)} \end{aligned}$$

- Or, if you prefer: tennis...
- Or springs...

Least Squares

$$\begin{aligned} \text{Distance} &= \sqrt{(y_1 - \hat{y}1))^2 + \cdots + (y_n - \hat{y}_n)^2} \\ &= \sqrt{(\varepsilon_1^2 + \cdots + \varepsilon_n^2)} \end{aligned}$$

- To minimize this distance, we can minimize the sum of squared errors.
  - Why can we ignore the square root?
  - Why can we work with the sum rather than the average?

Least Squares

- Our full equation is:

$$\begin{aligned} y_i &= \hat{y}_i + \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i \end{aligned}$$

- Putting ε on one side and adding "hats" to denote that we're really *estimating* the "true" line and errors:

$$\begin{aligned} \hat{\varepsilon}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ \hat{\varepsilon}_i^2 &= (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ \sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

The Least Squares Regression Equations

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make this as small as possible.
- Using calculus to find minima, we get:

(Least Squares Regression Equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The Prediction Equation

- Putting these into our line equation, we get the prediction function:

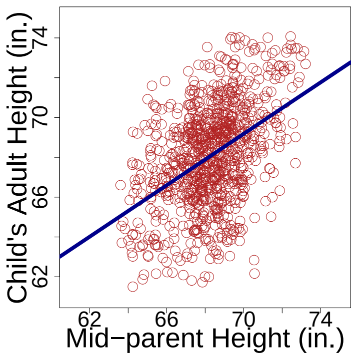
(Prediction Function)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Challenge Exercise

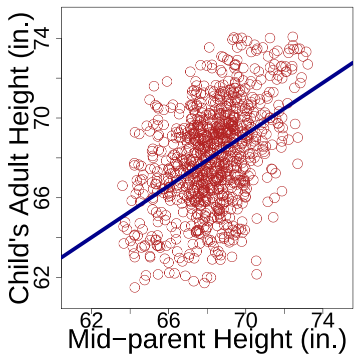
1. Show that if we have only two data points, these equations always give us the line that passes through them (provided the x values are distinct)
   - Hint: With two data points, the mean is half way between them.

Regression Example



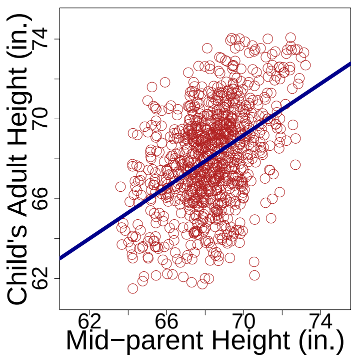
- The "father of regression", Francis Galton, looked at parents' and children's heights.
- Here's his data, with the associated regression line.

Regression Example



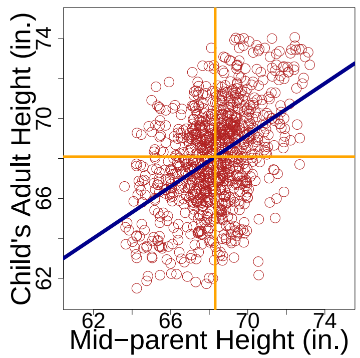
- What's $\hat{\beta}_1$ (approximately)?
- What's $\hat{\beta}_0$?
- We have $\hat{y} = 23.94 + 0.646x$.
- What does the 23.94 mean?
- What does the 0.646 mean?

Regression Example



- What would you predict for a child whose parents' average height is 64 in.?
- How about if the parents' average height is 72 in.?
- What about for the average parents' height?

Regression Example



- Notice that the prediction for the average x is the average y.

Challenge Exercises

1. Show that if we have only two data points, these equations always give us the line that passes through them (provided the x values are distinct)
   - Hint: With two data points, the mean is half way between them.
2. Show that if we plug $\bar{x}$ into the regression equation, we always get $\hat{y} = \bar{y}$.

Regression with Deviations

- Notice that we can reframe our regression equation directly in terms of deviations from the mean:

$$
\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
\hat{y}_i - \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} \\
&= \hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) \\
&= \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x} \\
&= \hat{\beta}_1 (x_i - \bar{x})
\end{aligned}
$$

Regression of Deviations

(Regression of Deviations)

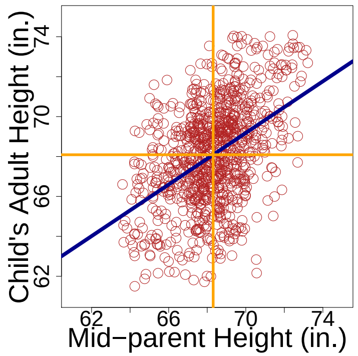
$$(\hat{y}_i - \bar{y}) = \hat{\beta}_1 (x_i - \bar{x})$$

- If we let d_{x_i} and d_{y_i} stand for x and y deviations, this is just:

$$d_{y_i} = \hat{\beta}_1 d_{x_i}$$

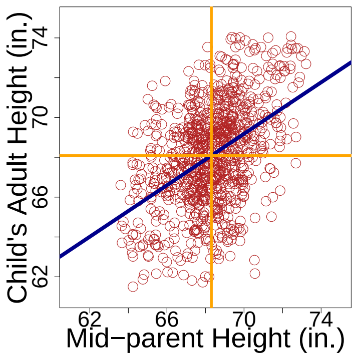
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} d_{x_i} d_{y_i}}{\sum_{i=1}^{n} d_{x_i}^2}$$

Regression to the Mean



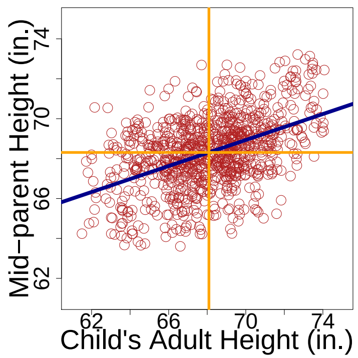
- Since the x and y have the same units, the slope less than 1 means that, on average, children are closer to average than their parents. Why?

Regression to the Mean



- Galton called this phenomenon “reversion to the mean”. Later changed to “regression to the mean”.
- This is the origin of the term.

Regression to the Mean



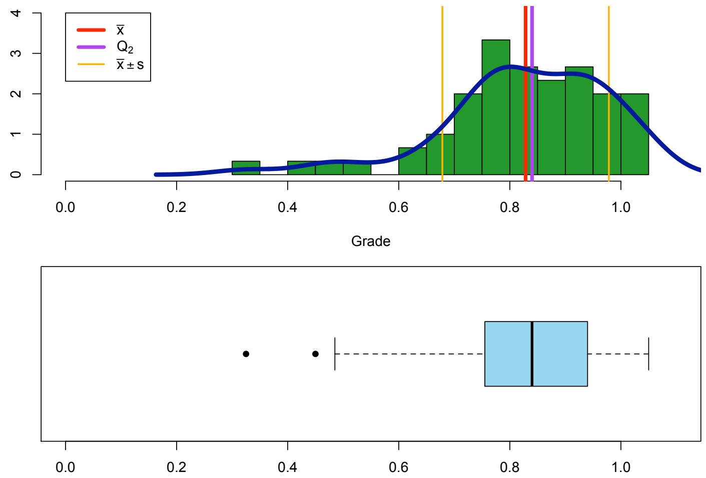
- If we try to predict parent from child, we get the following regression equation:

$$\hat{x}_i = 46.14 + 0.33 y_i$$

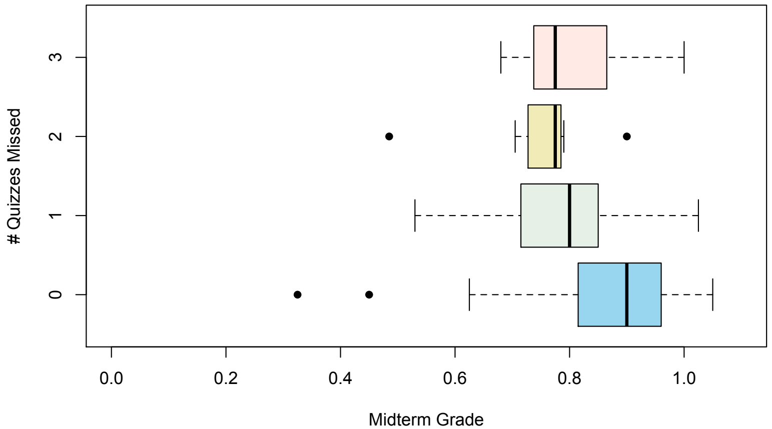
- In other words, on average, parents are closer to average than their children. Um...what?

An Experiment, Part 2

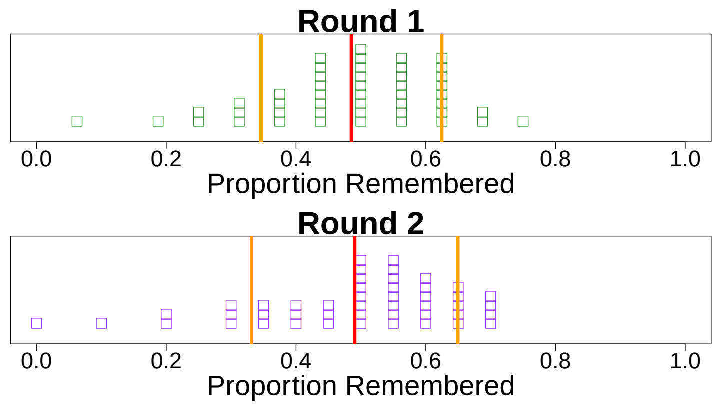
Midterm Results



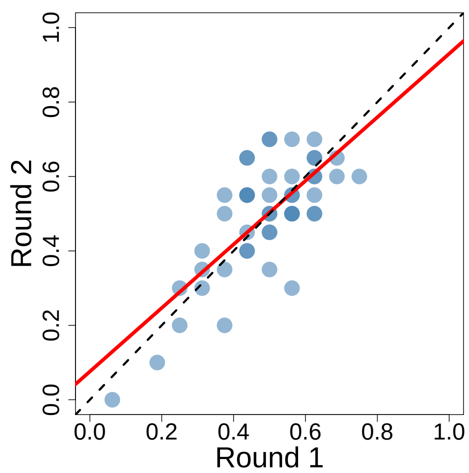
Midterm Results by Attendance Group



Memory Test Results

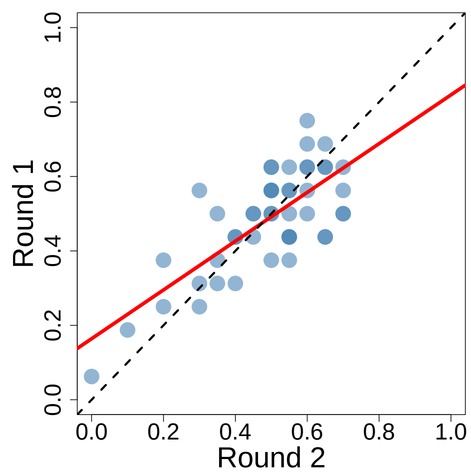


Memory Test Results



ISTA 116: Statistical Foundations for the Information Age
└The Method of Least Squares

Memory Test Results



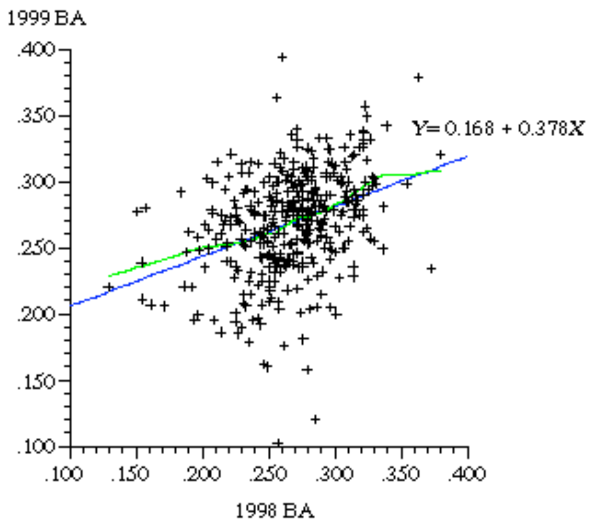
ISTA 116: Statistical Foundations for the Information Age
└The Method of Least Squares

Regression to the Mean

- Many variables have a systematic and random part
- E.g., “Skill” and “Luck”
- If you had a really high score the first time, there’s a good chance you had both
- If you try again, you would expect your skill to carry over, but not your luck; so your score would go down
- Conversely, low scores are likely partly the result of bad luck, so they should go up.

ISTA 116: Statistical Foundations for the Information Age
└The Method of Least Squares

Example: Batting Average in Successive Seasons



Why our children's future no longer looks so bright

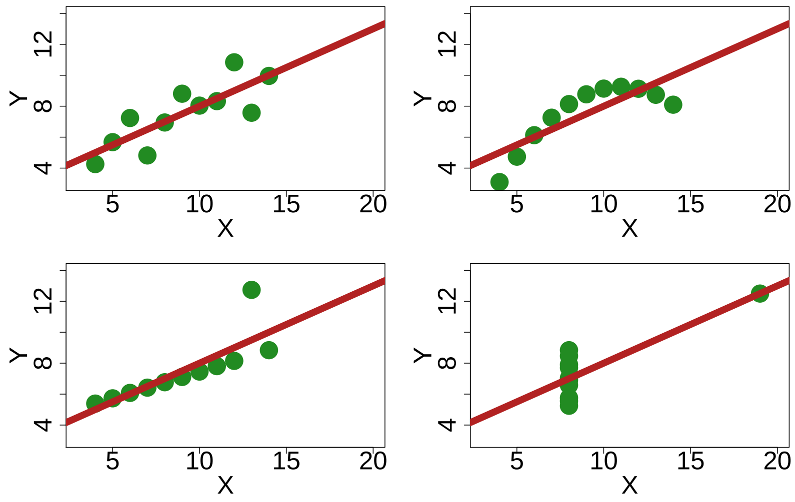
Text Size Print E-mail Reprints

By Robert J. Samuelson, Published: October 16

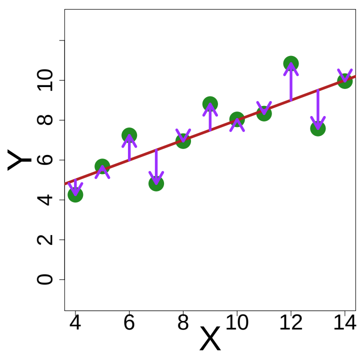
Aspecter haunts America: downward mobility. Every generation, we believe, should live better than its predecessor. By and large, Americans still embrace that promise. A Pew survey earlier this year found that 48 percent of respondents felt that their children's living standards would exceed their own. Although that's down from 61 percent in 2002, it's on a par with the mid-1990s. But these expectations could be dashed. For young Americans, the future could be dimmer.

Along with jobs, the 2012 presidential election could be fought over this issue. "Can the Middle Class Be Saved?" worried a recent cover story in the Atlantic. Pessimism rises with schooling. In the Pew poll, 54 percent of respondents with a high-school diploma or less felt their children would do better; only 35 percent of graduate school alums agreed. "A kind of depression has set in," writes Washington Post columnist Richard Cohen. "We've lost our mojo, our groove."

Not Every Best Line is a Good Line



Residuals



- Every line will have a different set of errors associated with it.
- These errors are called **residuals**.

Residuals

- Remember we said that we could think of y as having a "systematic" and a "random" part?
- The residuals are the "random" part.
- Remember that the sum of squared residuals is what we minimized to get our regression line: try to put as much as we can into the systematic part.

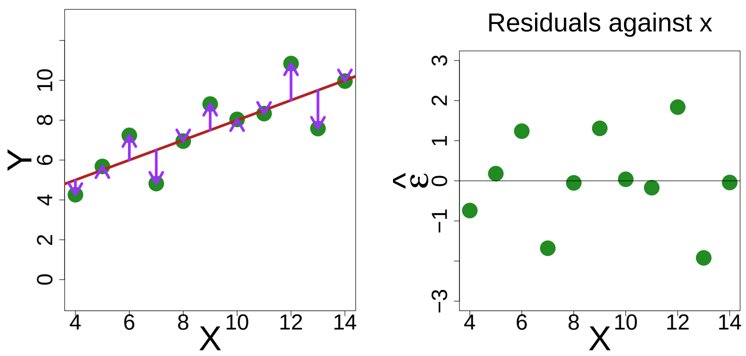
Randomness of Residuals

- If our regression model is a good one, we shouldn't be able to predict the residuals from anything else: they should be truly random.
- This suggests a way to diagnose whether we have a good model. What could we do?
- If the residuals are random, they should be unrelated to both x and $\hat{y}$.

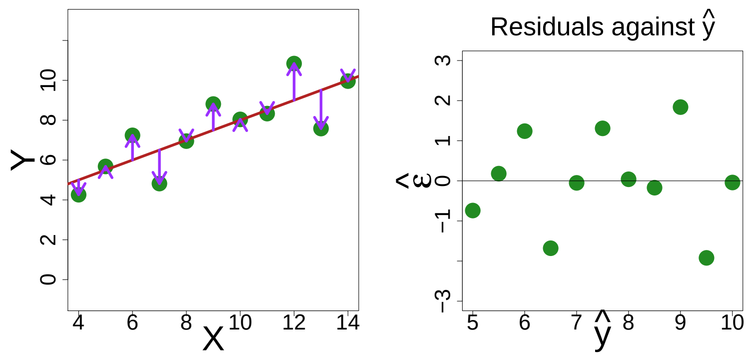
Residual Plots

- Two useful plots:
  1. Residuals against x
  2. Residuals against $\hat{y}$
- If residuals are random, these should both look like an unstructured "cloud".

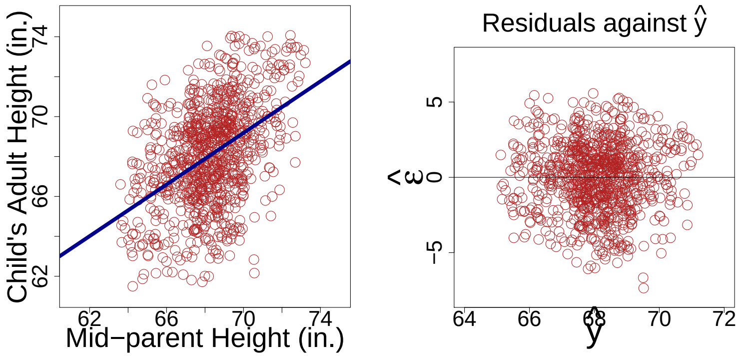
Residual Plots



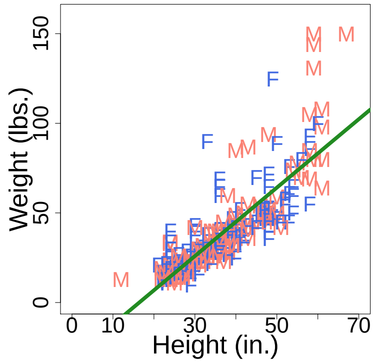
Residual Plots



Residual Plots

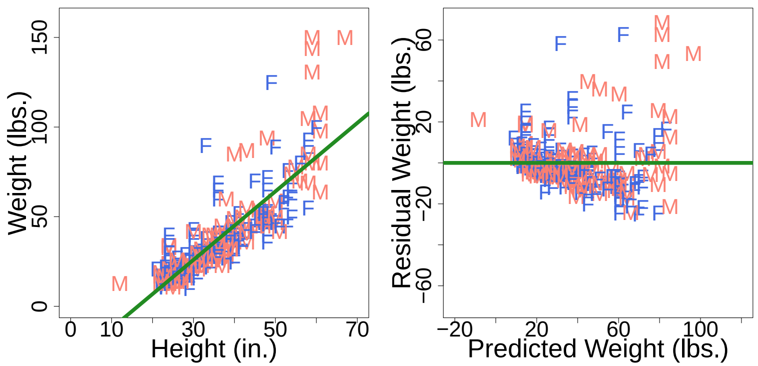


Nonlinear Residual Plots



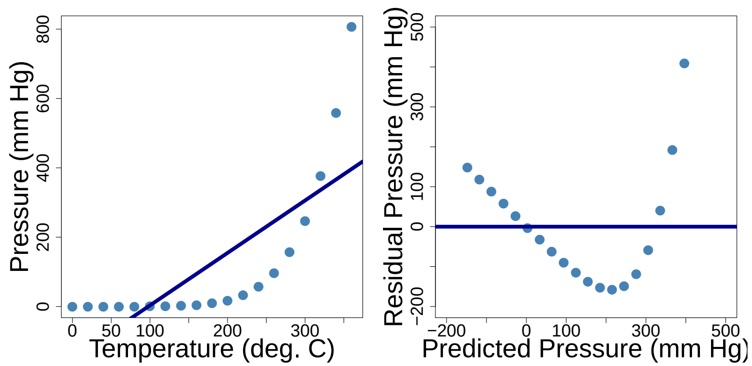
- What will the residual plot look like if the actual relationship is curved?

Nonlinear Residual Plots



Nonlinear Residual Plots

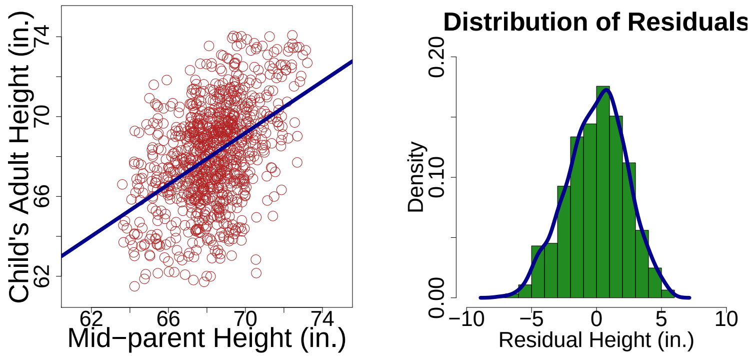
- Even more strikingly...



The Residual Distribution

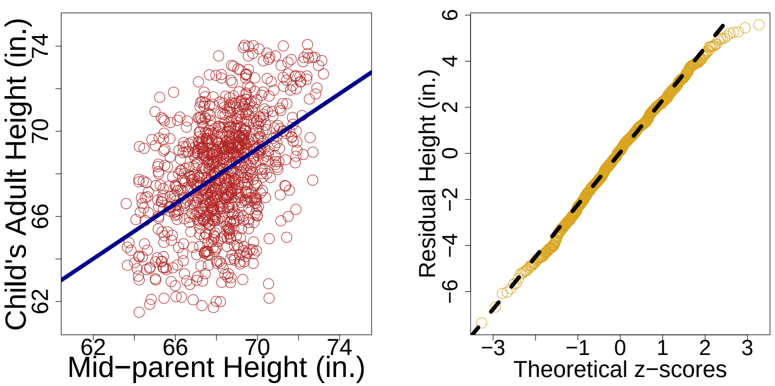
- It's also useful to look at the **distribution** of residuals.
- In many cases, if the residuals really just capture randomness, they will have a bell-shaped distribution.
- We can apply our univariate techniques to the residual distribution to assess this.

The Residual Distribution



The Residual Distribution

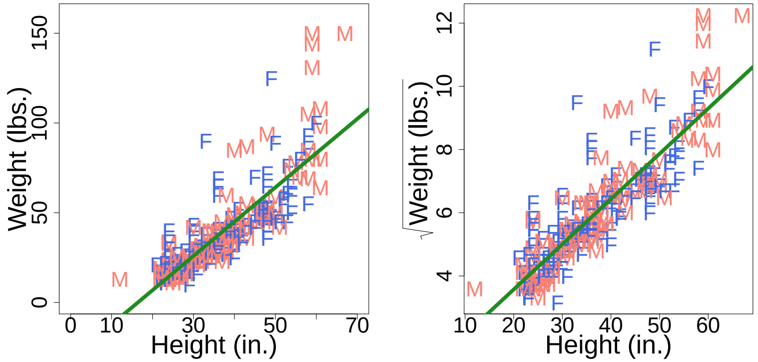
- We can also use a QQ Plot against a hypothetical bell curve:



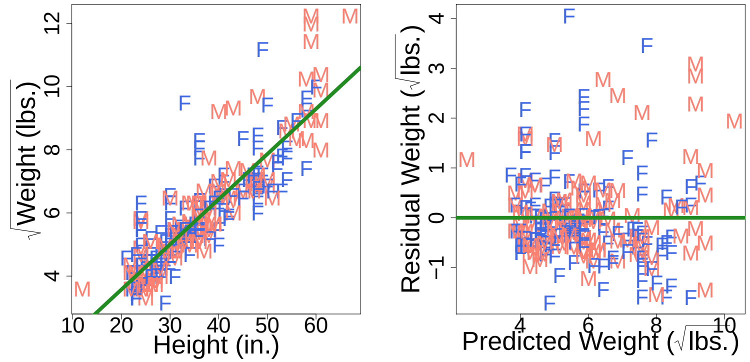
- We can use residual plots as a diagnostic tool, to see when a linear model is inadequate.
- What might we do in these cases?
- Sometimes we need a more complex model (e.g., a higher order polynomial; one with other predictors)
- Sometimes we can create a linear relationship via a **transformation** of one or both variables.

Example of a Nonlinear Relationship

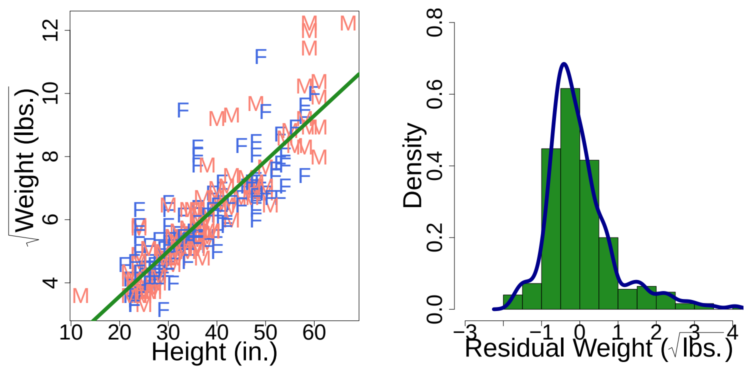
- Suggestions?



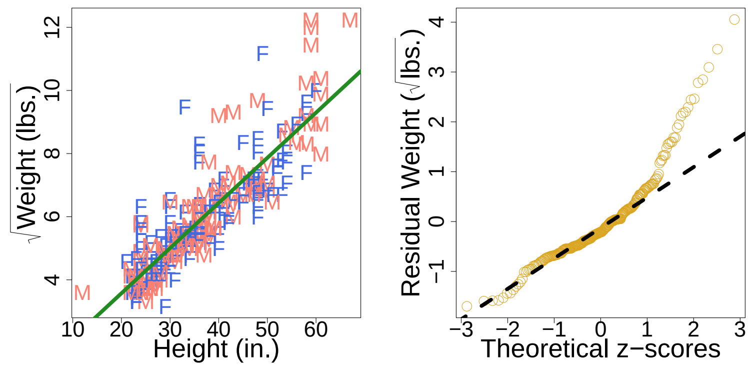
After Square Root Transformation



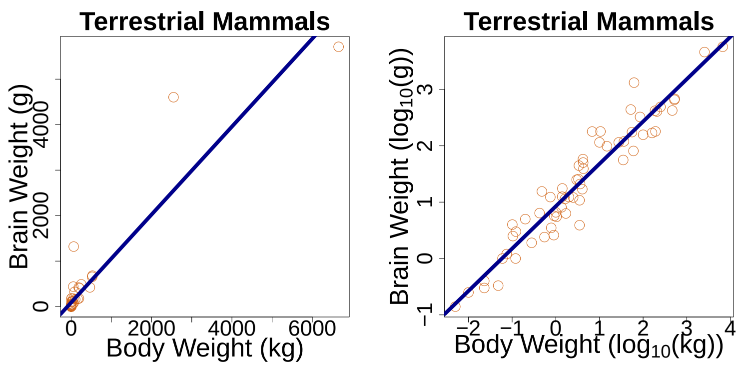
After Square Root Transformation



After Square Root Transformation



What About This?



Preview of Part 2

A Video!