# YData: An Introduction to Data Science

## Lecture 36: Multiple Regression

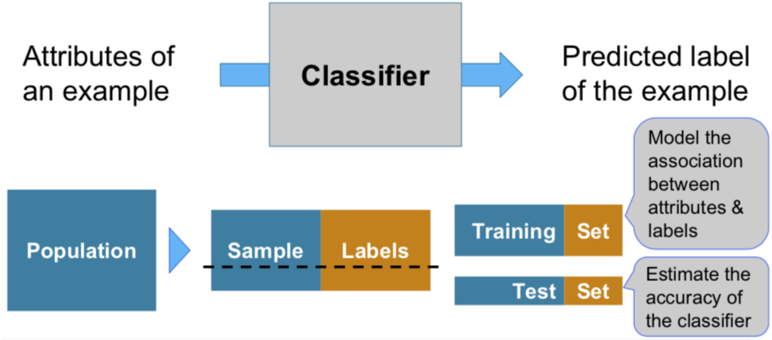
Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

# Keep in Mind

- Project 3 due Friday 4/30 (tommorow)
- Assignment 11 out; due next Thursday 5/6
- We'll have info on prep for the final exam next week
- We'll compile "provisional grades" next week

# Previously: Classifiers



# Finding the $k$ Nearest Neighbors

To find the $k$ nearest neighbors of an example:

- Find the distance between the example and each example in the training set
- Augment the training data table with a column containing all the distances
- Sort the augmented table in increasing order of the distances
- Take the top $k$ rows of the sorted table

# The Classifier

To classify a point:

- Find its $k$ nearest neighbors
- Take a majority vote of the $k$ nearest neighbors to see which of the two classes appears more often
- Assign the point the class that wins the majority vote

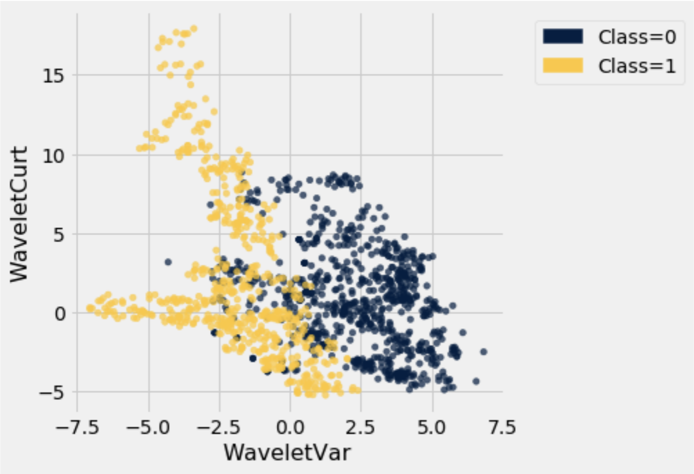
# Evaluation

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly

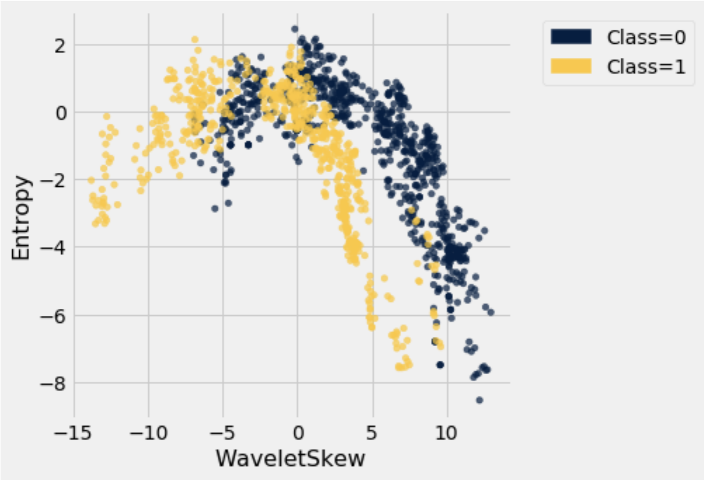
Need to compare classifier predictions to true labels

If the labeled data set is sampled at random from a population, then we can infer accuracy on that population

# k-NN Intuition

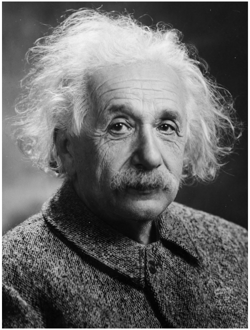


# k-NN Intuition



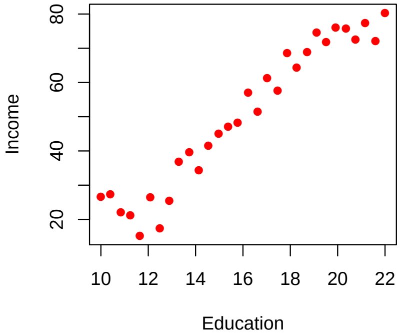
# For today: Multiple linear regression

- Multiple linear regression = multiple predictors
- Foundation for more advanced topics, such as neural networks
- Usually a good place to start — Bay Area traffic story



*Everything should be made as simple as possible, but no simpler.*

# Simulated income dataset

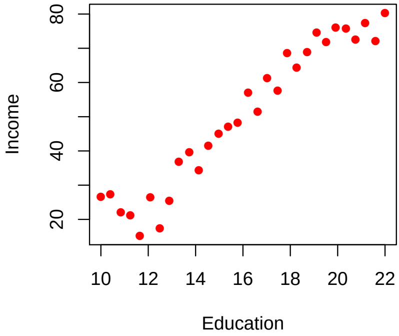



Goal: Predict income($Y$) using education ($X$).

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

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30})$$

# Simulated income dataset

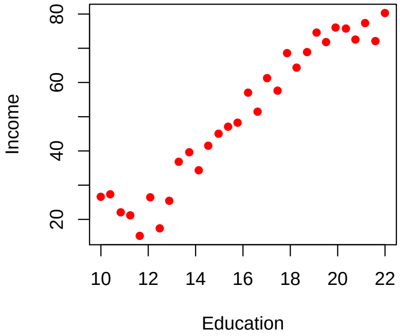



Goal: Predict income($Y$) using education ($X$).

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

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30})$$

# Simulated income dataset




$$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30})$$

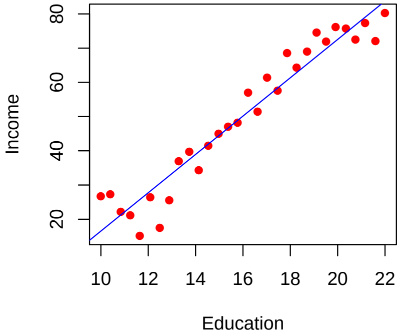
Goal: Predict income($Y$) using education ($X$).

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

Linear model:

$$f(X) = \beta_0 + \beta_1 X$$

# Simulated income dataset




Goal: Predict income($Y$) using education ($X$).

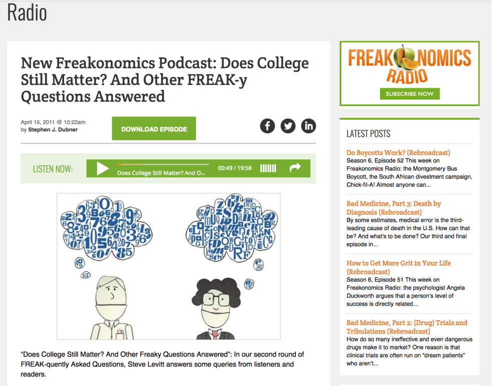
$$Y = f(X) + \epsilon$$

Linear model:

$$f(X) = \beta_0 + \beta_1 X$$

Find coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ s.t. $\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ is reasonably close to $Y$.

# How does education impact earnings?




*Every extra year of education translates to 8% increase in earnings over lifetime.*

http://freakonomics.com/podcast/new-freakonomics-podcast-does-college-still-matter-and-other-freak-y-questions-answered/

# Estimating the coefficients

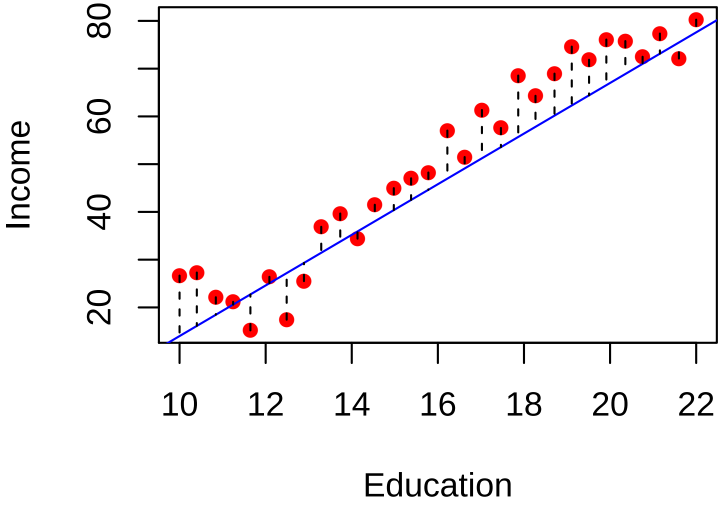
For any $\hat{\beta}_0, \hat{\beta}_1$, we predict $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We call these **fitted values**.

# Estimating the coefficients

For any $\hat{\beta}_0, \hat{\beta}_1$, we predict $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We call these **fitted values**.

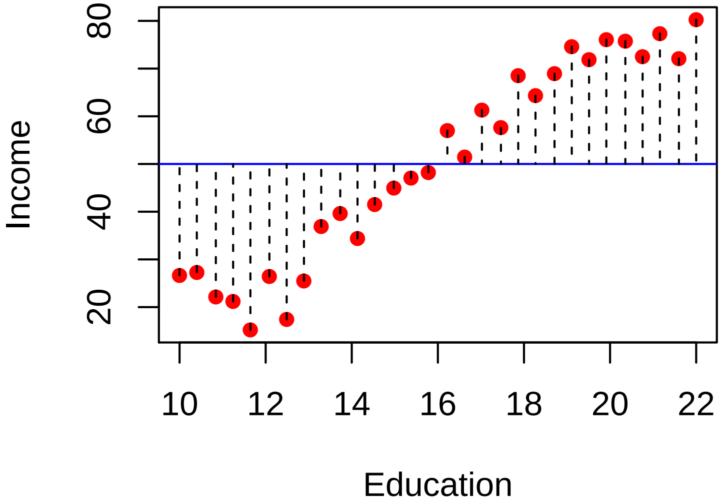
The **residual** $e_i = y_i - \hat{y}_i$ is difference between the $i$-th observed value and its fitted value.

# Some candidate lines (and residuals)



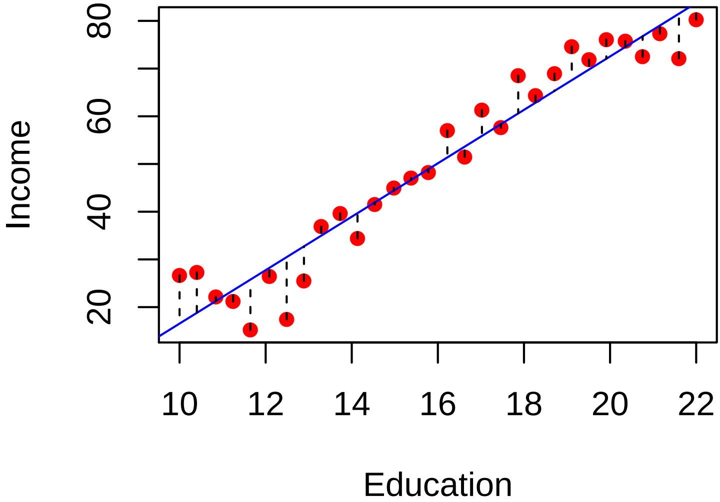

$\hat{\beta}_0 = -39, \hat{\beta}_1 = 5.3$

# Some candidate lines (and residuals)




$\hat{\beta}_0 = 50, \hat{\beta}_1 = 0$

# Some candidate lines (and residuals)




$\hat{\beta}_0 = -39.4, \hat{\beta}_1 = 5.6$

# Estimating the coefficients

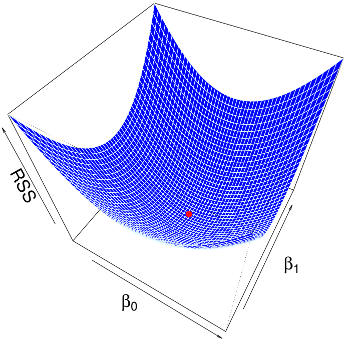
The **least squares** approach selects coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

# Estimating the coefficients

The **least squares** approach selects coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = (y_1 - \beta_0 - \beta_1 x_1)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2.$$



# Estimating the coefficients

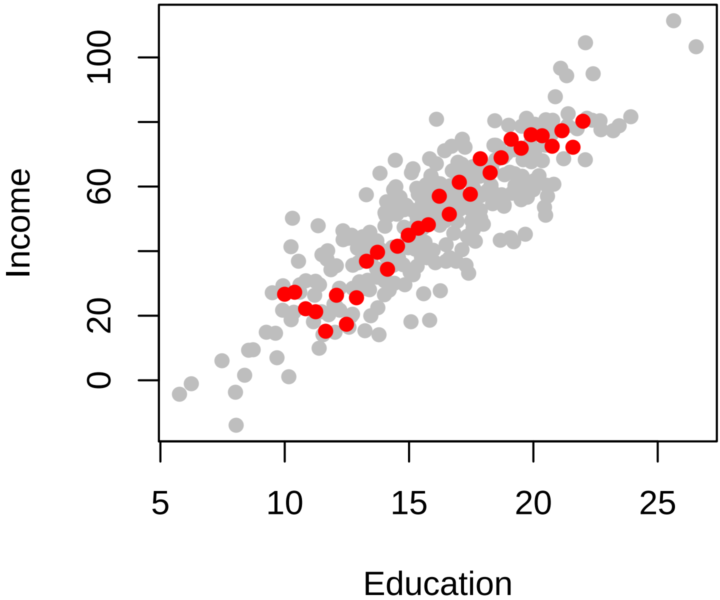
The **least squares** approach selects coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = (y_1 - \beta_0 - \beta_1 x_1)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2.$$

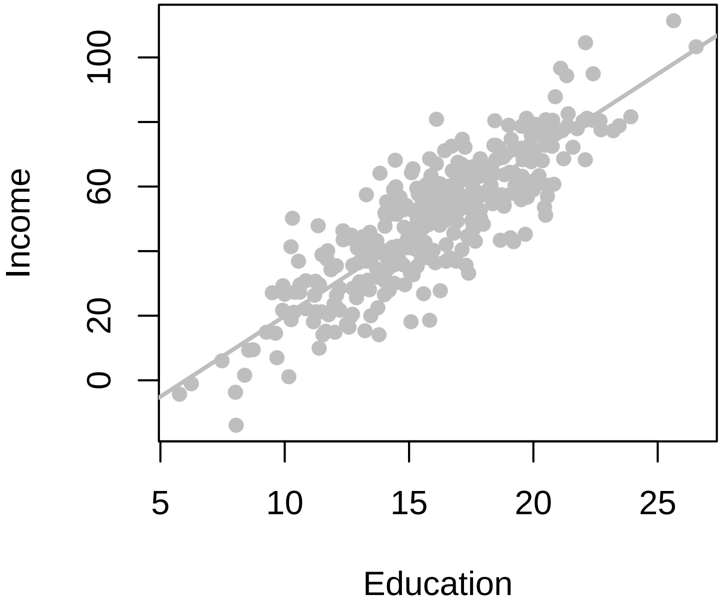
How do we find the minimum?

- Apply a formula...
- Use numerical optimization!

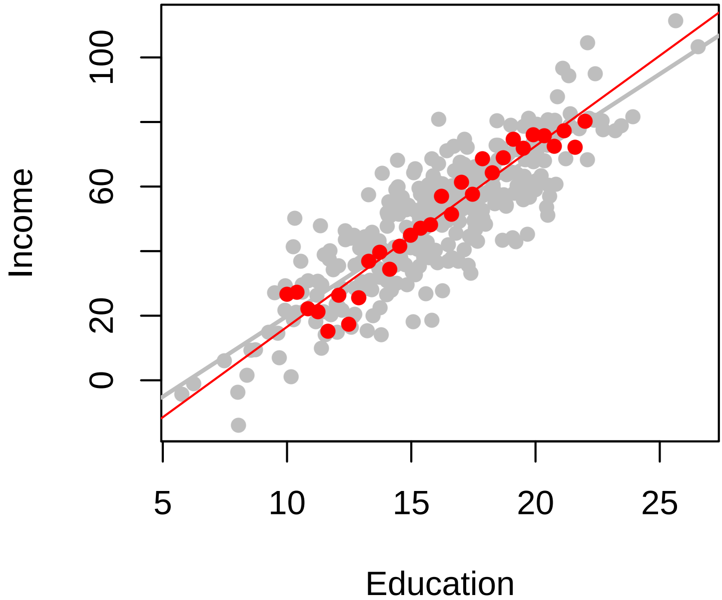
# Reminder: Population vs. sample



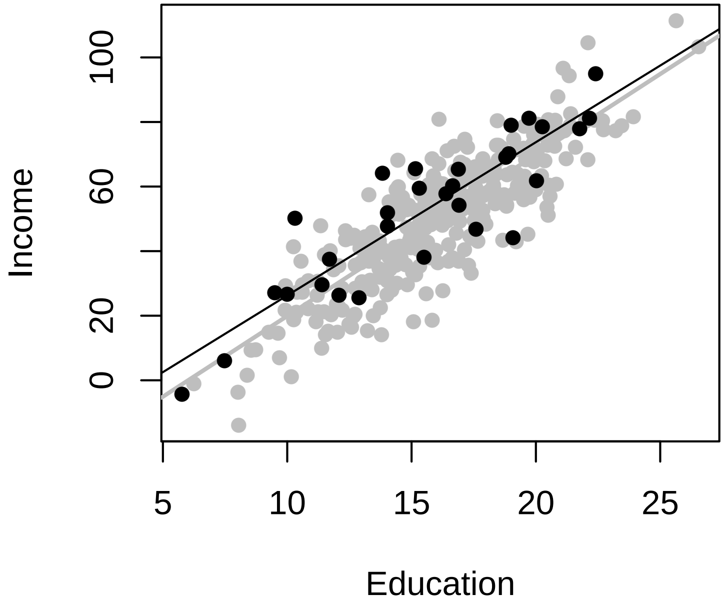
# Reminder: Population vs. sample



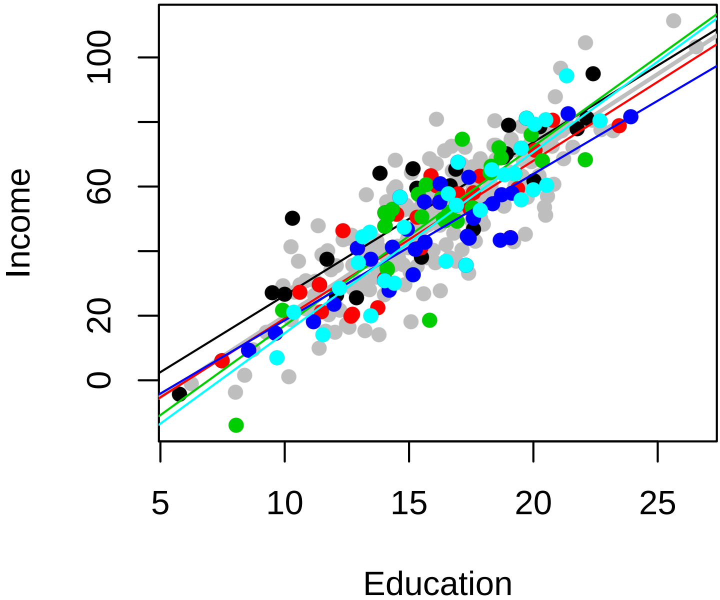
# Reminder: Population vs. sample



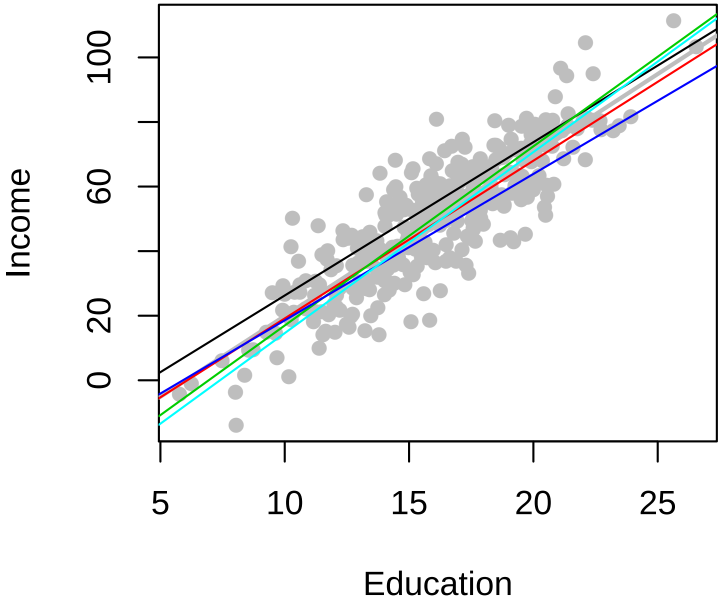
# Different samples



# Different samples



# Different samples



# Inference for linear regression

**Standard errors** of the coefficients describe how the coefficients vary under repeated sampling.

# Inference for linear regression

**Standard errors** of the coefficients describe how the coefficients vary under repeated sampling. A 95% confidence interval for $\beta_i$ is approximately:

$$\hat{\beta}_i \pm 2 \cdot SE(\hat{\beta}_i)$$

# Inference for linear regression

**Standard errors** of the coefficients describe how the coefficients vary under repeated sampling. A 95% confidence interval for $\beta_i$ is approximately:

$$\hat{\beta}_i \pm 2 \cdot SE(\hat{\beta}_i)$$

Can be estimated using the bootstrap!

# Sums of squares and $R^2$

Partitioning the sums of squares:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{total sum of squares}(TSS)} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares}(ESS)} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{residual sum of squares}(RSS)}$$

for least squares linear regression

# Sums of squares and $R^2$

Partitioning the sums of squares:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{total sum of squares}(TSS)} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares}(ESS)} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{residual sum of squares}(RSS)}$$

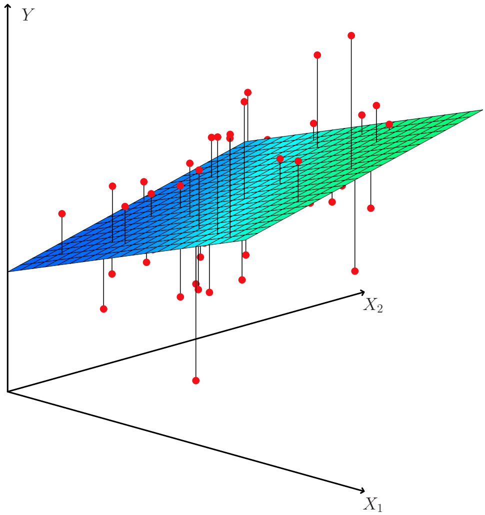
for least squares linear regression

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

We can interpret $R^2$ (**multiple R-squared**) as the proportion of variability in $y$ explained by the model.

- Between 0 and 1
- Doesn't depend on the scale of $Y$.

# Multiple linear regression



# General form for linear regression

With $p$ predictors $x_1, \dots, x_p$,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $\epsilon$ indicates an error term. In matrix notation,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \ddots & & x_{2,p} \\ \vdots & & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

# Estimating $\beta$

Two options:

- Apply a formula
- Use numerical optimization

Numerical optimization is the more powerful, flexible, and "modern" approach!

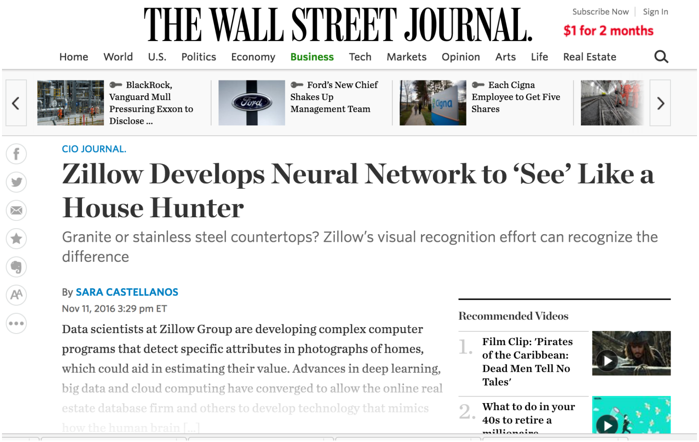
# Interpretation of formula

The coefficients are just the correlations between the variables $X_j$ and the data $Y$—*after* the variables are "whitened" to become uncorrelated.

# Discussion

Questions?

# Predicting home values



THE WALL STREET JOURNAL.

Subscribe Now | Sign In

$1 for 2 months

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

BlackRock, Vanguard Mull Pressuring Exxon to Disclose ...

Ford's New Chief Shakes Up Management Team

Each Cigna Employee to Get Five Shares

CIO JOURNAL.

## Zillow Develops Neural Network to 'See' Like a House Hunter

Granite or stainless steel countertops? Zillow's visual recognition effort can recognize the difference

By SARA CASTELLANOS

Nov 11, 2016 3:29 pm ET

Data scientists at Zillow Group are developing complex computer programs that detect specific attributes in photographs of homes, which could aid in estimating their value. Advances in deep learning, big data and cloud computing have converged to allow the online real estate database firm and others to develop technology that mimics how the human brain [...]

Recommended Videos

1. Film Clip: 'Pirates of the Caribbean: Dead Men Tell No Tales'

2. What to do in your 40s to retire a millionaire

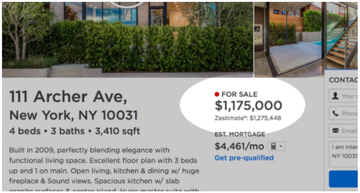
“The Seattle-based firm has amassed a database of 115 million homes across the country. Zestimates are used to estimate each property’s valuation, based on statistical and machine learning models that examine hundreds of data points on each home, including square footage, lot size, number of transactions in a geographical area, and soon, hundreds of thousands of photos. Since 2005, the company has reduced its valuation error rate from 14% to 4.5% through iterations of its algorithm, and it’s betting that estimates could be even more accurate with sophisticated neural networks.”

# $\$1M$ question

https://www.kaggle.com/c/zillow-prize-1 https://www.zillow.com/promo/zillow-prize-first-round/

I'm excited to share the launch of [Zillow Prize: Home Value Prediction (Zestimate) Competition](). In this million-dollar competition, participants will develop an algorithm that makes predictions about the future sale prices of homes.

Zillow's Zestimate home valuation shook up the U.S. real estate industry when it was first released 11 years ago. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of information at no cost.




This million dollar contest is structured into two rounds. In the qualifying round, opening today, you'll be building a model to improve the Zestimate residual error. The top 100 ranking teams in this round will advance to the final round. In the final round, competitors will be challenged with building a home valuation algorithm from the ground up, using external data sources to help engineer new features that give your model an edge over the competition. The first place prize in the final round is $1,000,000 USD.

**Join the competition**

# Demo

Let's do a simple version of this using (multiple) linear regression!

DEMO

# Summary

- Least squares coefficients correspond to minimum of a "bowl shapped" surface
- Confidence intervals computed using the bootstrap
- $R^2$ is a scale-invariant accuracy measure — proportion of variance in $Y$ explained by the model
- Multiple linear regression (many predictors) estimated by numerical optimization.
- Most of what we learned in the 1-dimensional case carries over—except the formula for the slope and intercept