Review mapping and linear regression

Announcement: midterm exam

Thursday during class time (9-10:15am)

- 60 minutes for the exam, 15 minutes to upload it to Gradescope

Open notes, slides, etc.

Can use the internet to look up R syntax and LaTeX symbols **only**

TAs will have office hours early next week to answer your questions

Practice questions will also be posted soon

- Real exam will be a little different

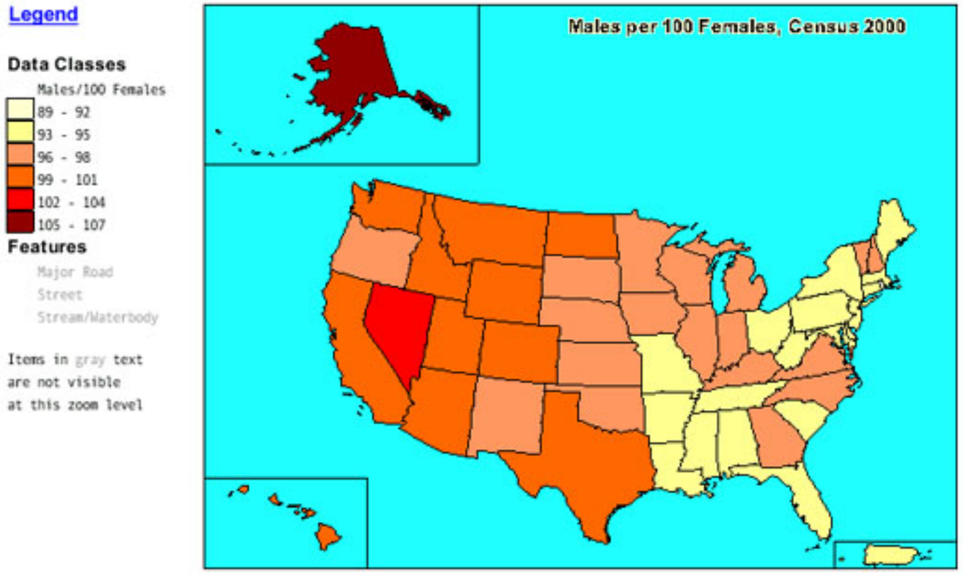
Contact me if you have accommodations or are in a different timezone

Maps

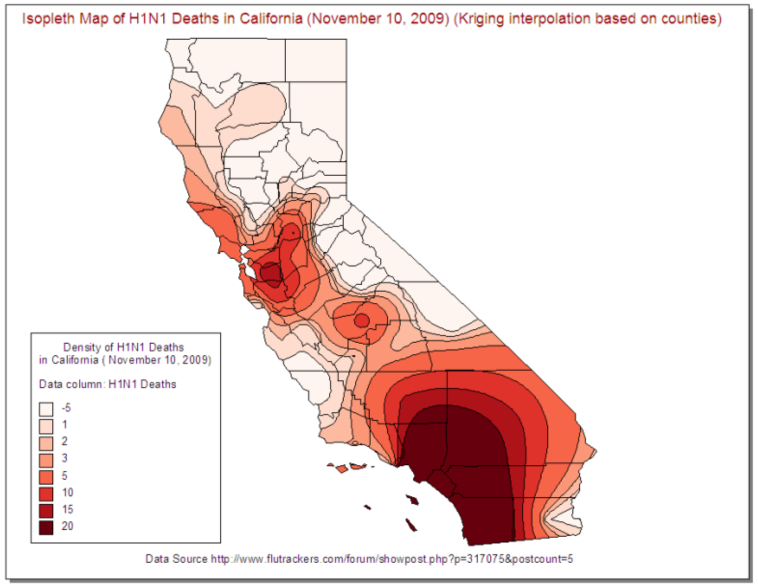
Choropleth maps: shades/colors in predefined areas based on properties of a variable

Isopleth maps: creates regions based on constant values

Choropleth map

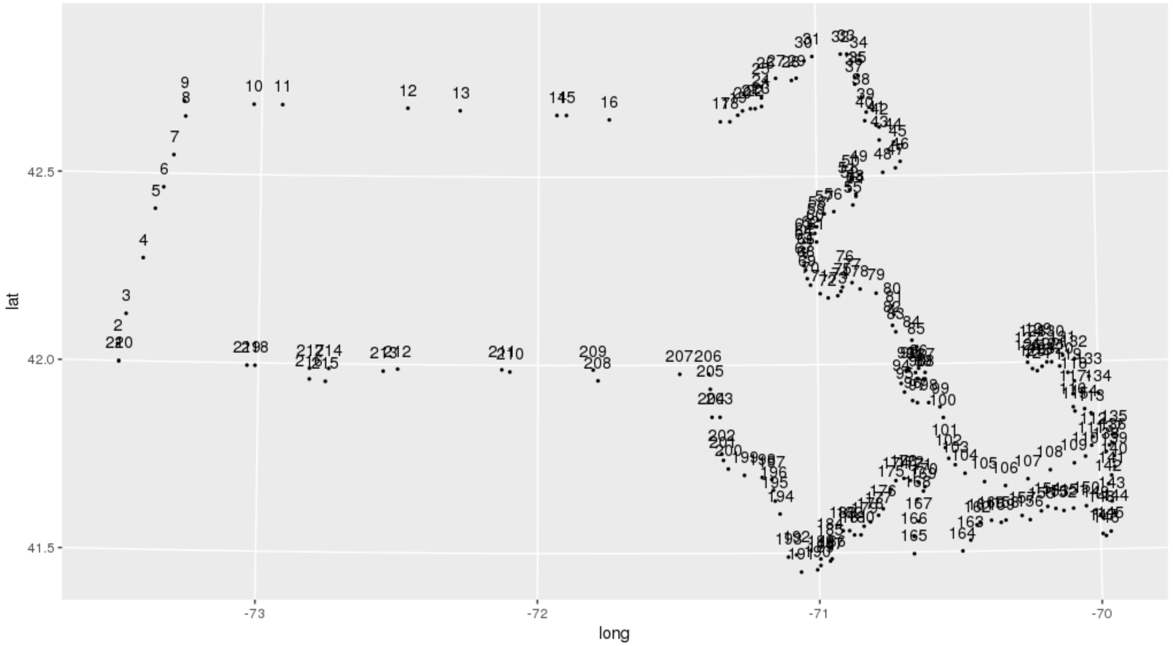


Isopleth map



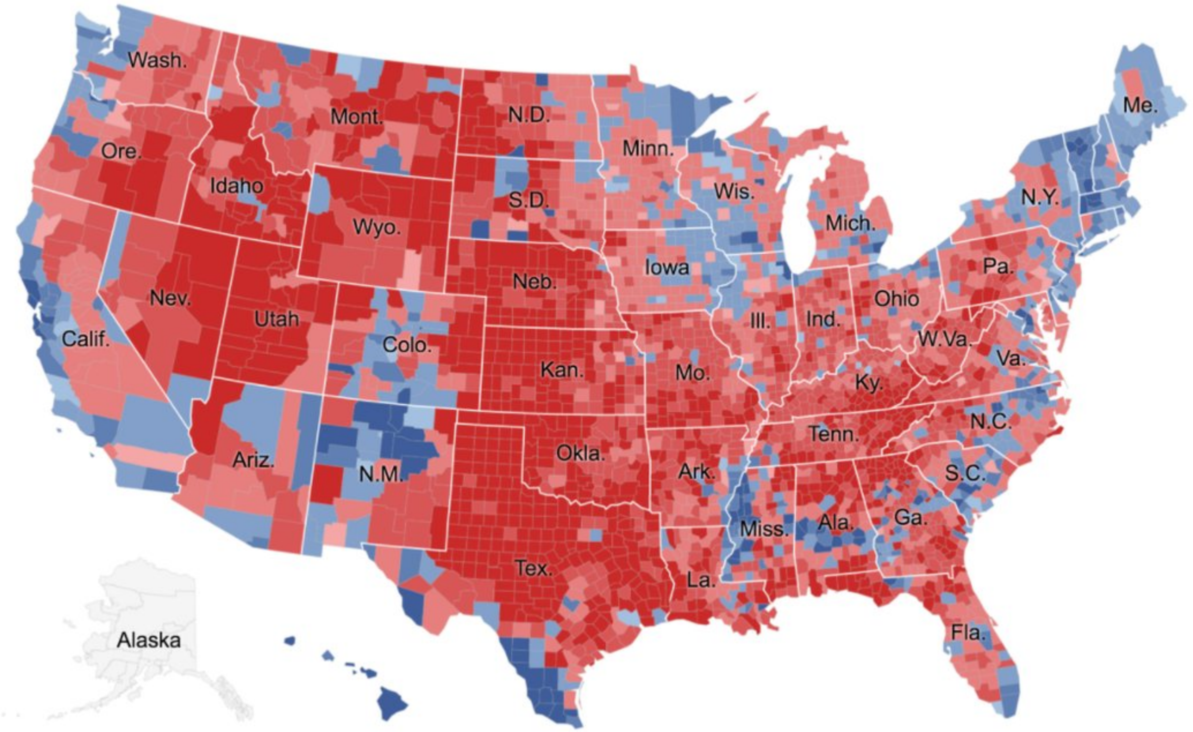
Choropleth maps

geom_polygon() works by connecting the dots:



Often need to arrange points first: arrange(states_map, group, order)

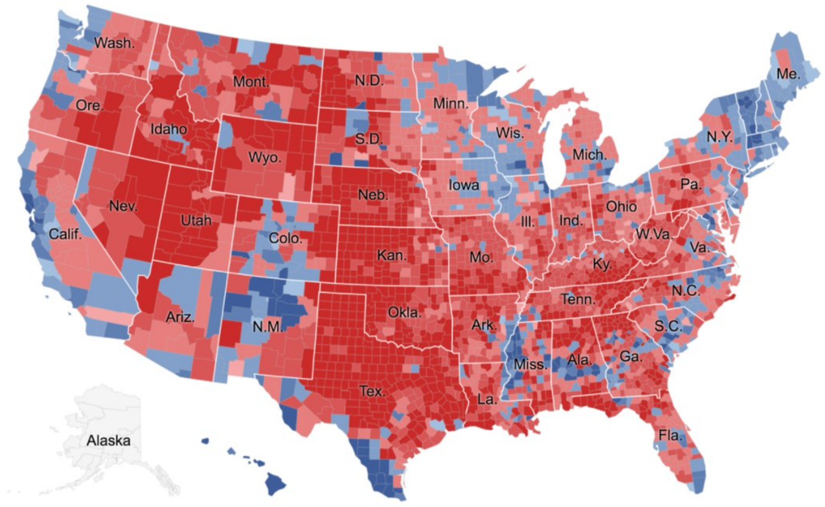
Survey question 1: in what way could this map be misleading?



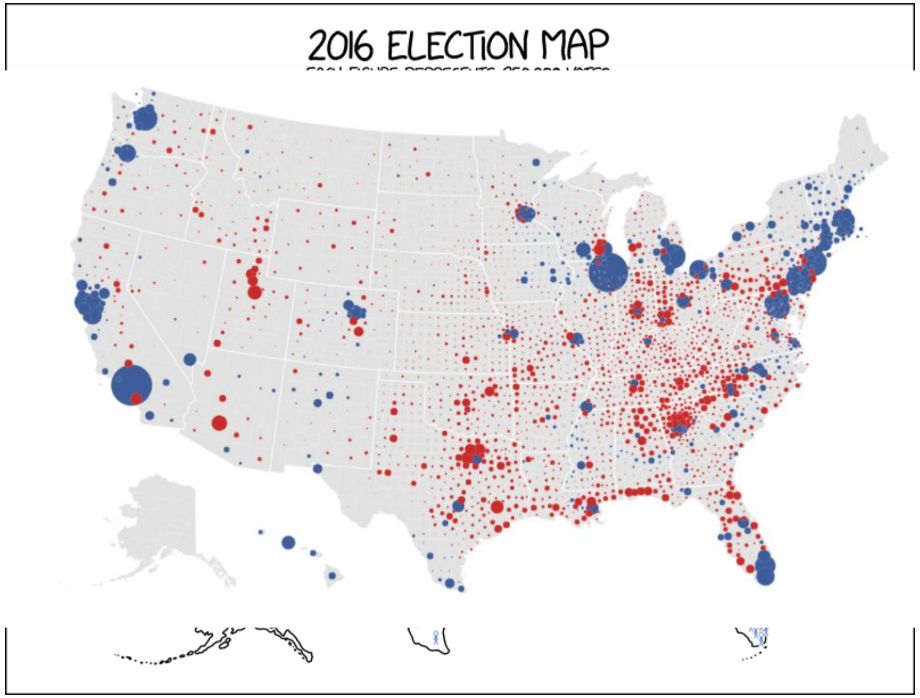
Darker red: county had higher % Trump vote

Darker blue: county had higher % Clinton vote

Cloropleth maps could be misleading



Looks like most of the country voted republican



Land doesn't vote

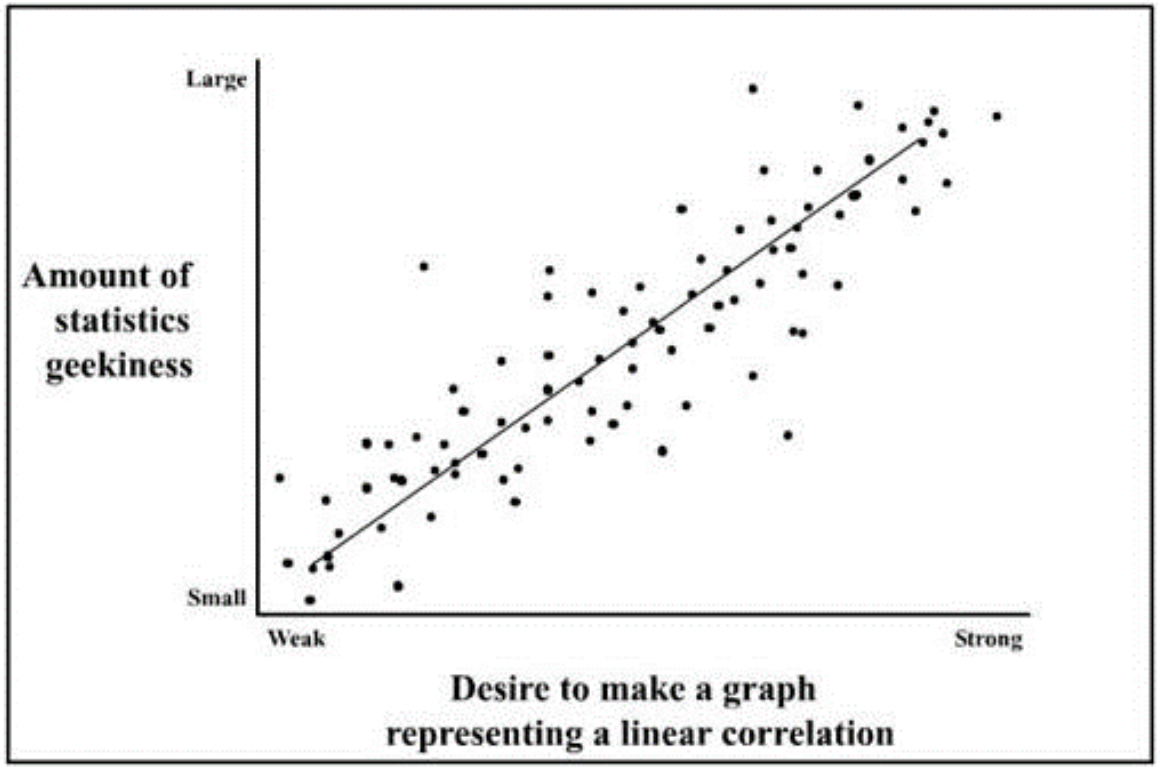
Linear regression

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

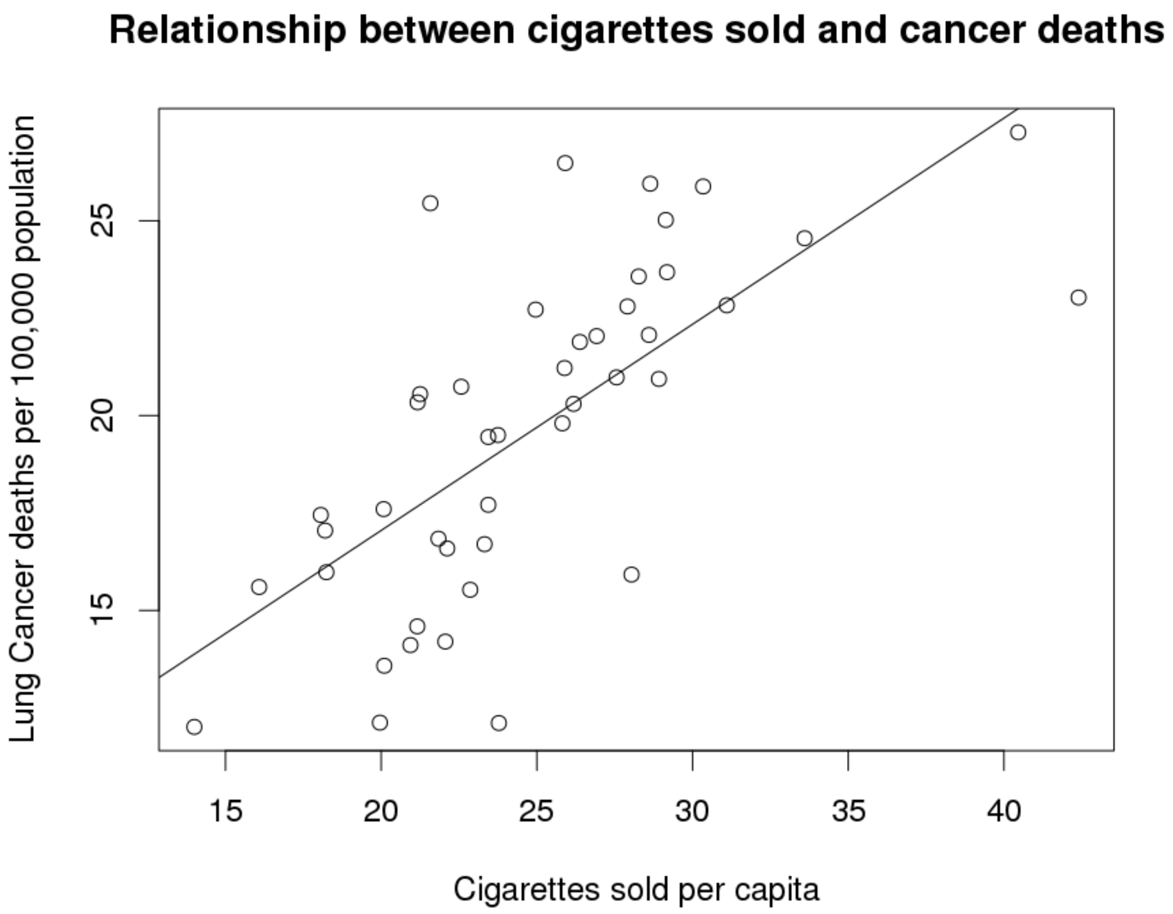
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

- In simple linear regression, we use a single variable x, to predict y



Cancer smoking regression line



$\hat{y} = b_0 + b_1 \cdot x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

R: `lm(y ~ x)`

$b_0 = 6.47$

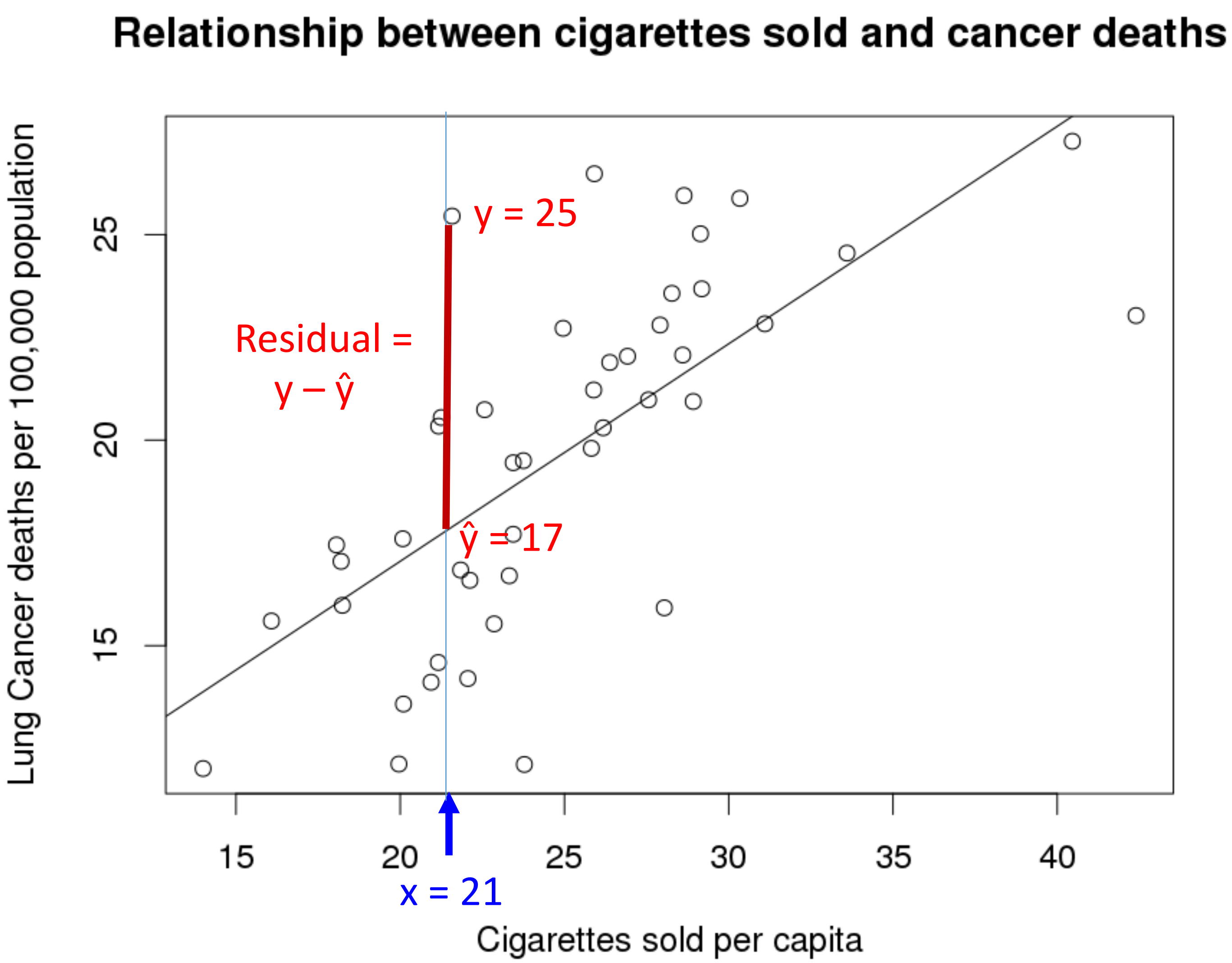
$b_1 = 0.53$

$\hat{y} = 6.47 + .53 \cdot x$

Residuals

The **residual** at a data value is the difference between the observed (y) and predicted value of the response variable

$$\text{Residual} = \text{Observed} - \text{Predicted} = y - \hat{y}$$

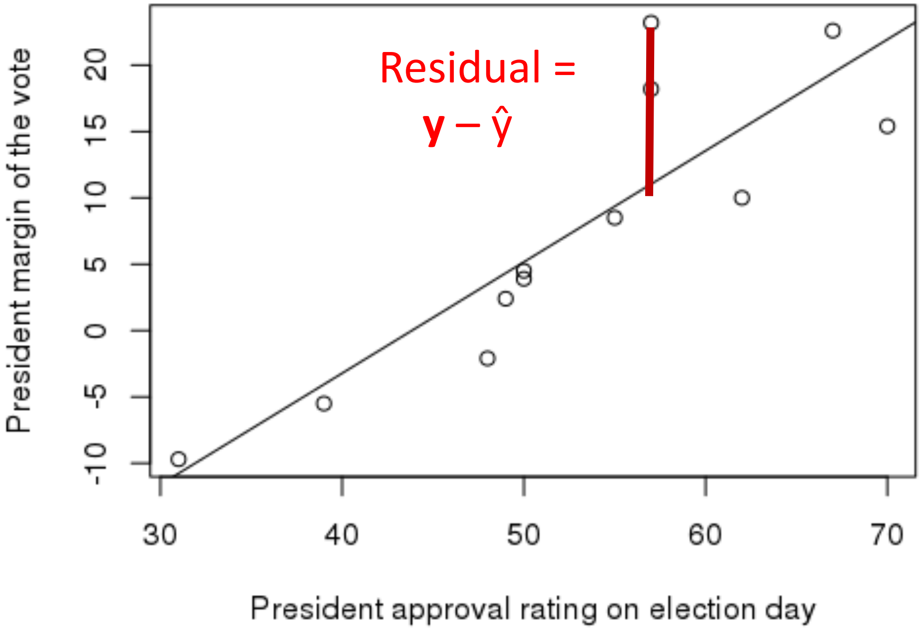


The **least squares line**, is the line which <u>minimizes the sum of squared residuals</u>

Minimizing the sum of the squared residuals to find the regression coefficients

To find the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_0$ we minimize the **residual sum of squares (RSS)**

- The residual sum of squares is also called the **error sum of squares (SSE)**



$$residual = e_i$$

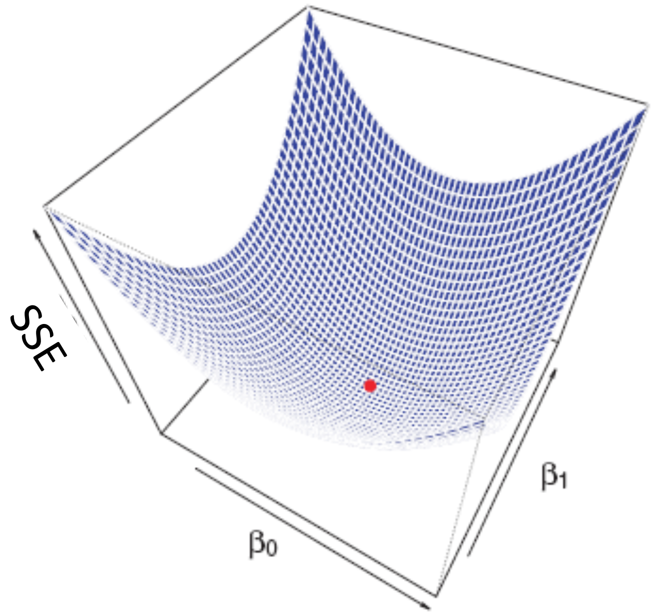
$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{f}(x))^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$

R: `lm(y ~ x)`

How do we minimize the SSE?

$$SSE \quad = \quad \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x)^2$$



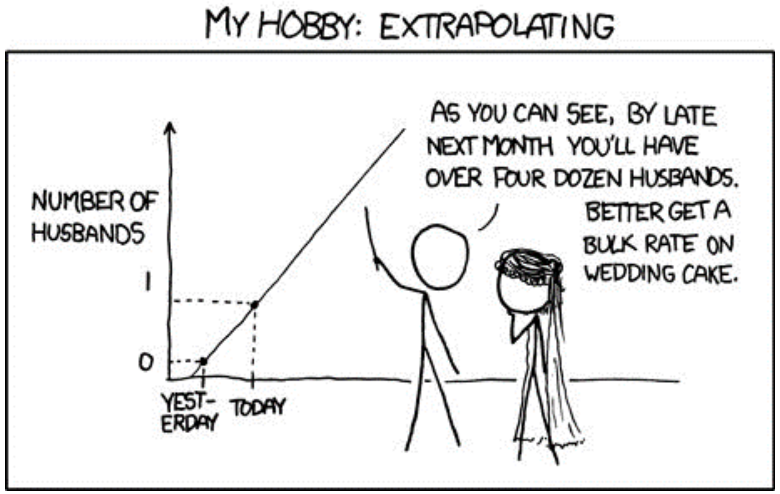
How do we find $\hat{\beta}_0, \hat{\beta}_1$?

Calculus and linear algebra:

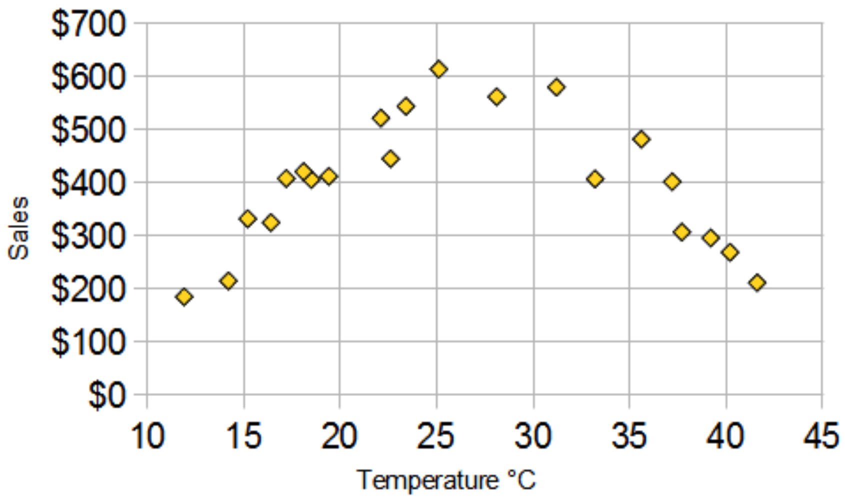
- Take the derivative, set to 0 and solve
- This mathematical convenience is why the squared loss is so commonly used

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Regression caution #1: Avoid trying to apply the regression line to predict values far from those that were used to create the line.



Regression caution #2: Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3: Be aware of outliers and high leverage points. They can have an huge effect on the regression line.

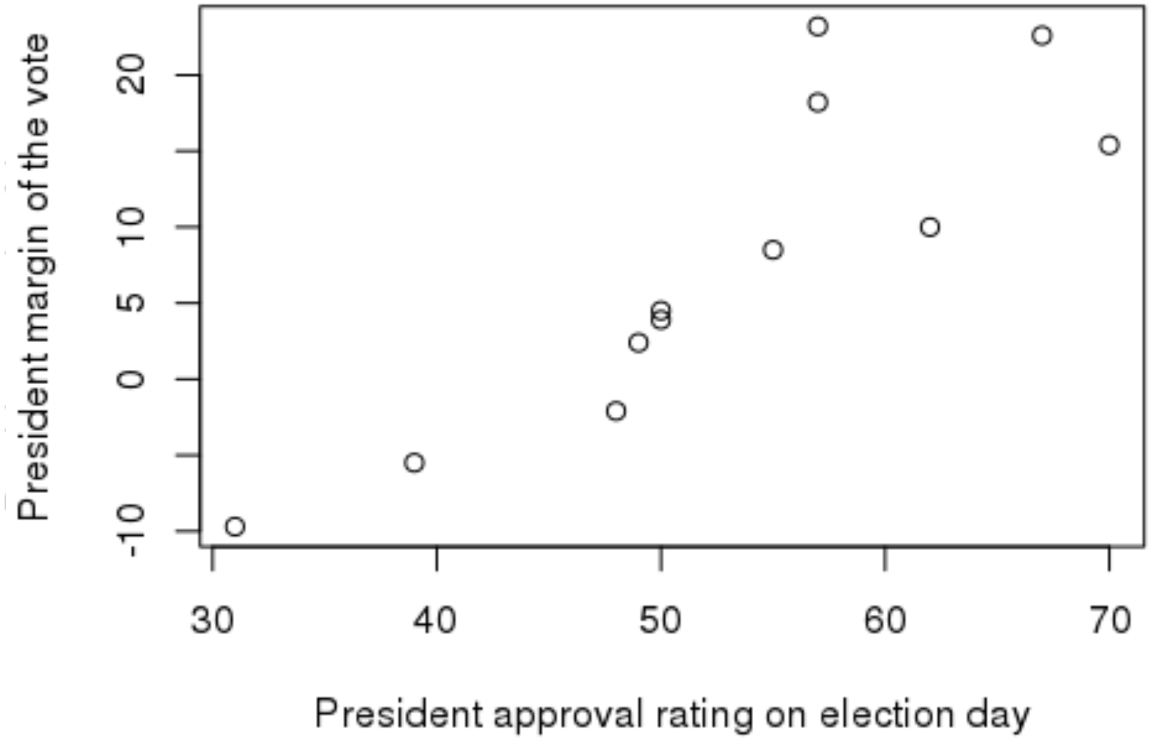
Outlier: big $| y - \bar{y} |$

Leverage: big $| x - \bar{x} |$

Influential point: big outlier and leverage

Approval rating vote margin regression line

From last 12 US president's running for reelection



$\hat{y} = b_0 + b_1 \cdot x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

R: `lm(y ~ x)`

$\hat{\beta}_0 = b_0 = -36.76$

$\hat{\beta}_1 = b_1 = 0.84$

$\hat{y} = -36.76 + .84 \cdot x$

Approval rating vote margin survey questions

1. If a president had a 0% approval rating, what percent of the vote margin does this model predict the president would get?

   A: would have a margin of -36.76% of the vote

2. If a president's approval rating increased by 1%, how much of would the president's margin of the vote increase by?

   A: .84 increase in the margin of the vote

3. At what presidential approval level would there be an exactly even split of the vote?

   A: 36.76/.84 = 43.76% approval rating

$$\hat{y} = b_0 + b_1 \cdot x$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

R: `lm(y ~ x)`

$$\hat{\beta}_0 = b_0 = -36.76$$

$$\hat{\beta}_1 = b_1 = 0.84$$

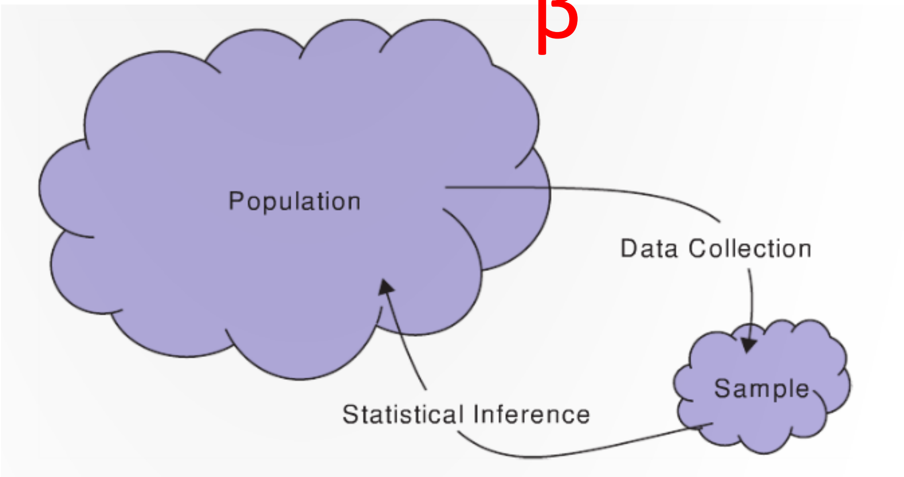
$$\hat{y} = -36.76 + .84 \cdot x$$

After the exam: Inference for simple linear regression

The letter **b** or $\hat{\beta}$ is typically used to denote the slope ***of the sample***

The Greek letter **β** is used to denote the slope ***of the population***

β



b $\hat{\beta}$

Any questions about simple linear regression?