Spatial mapping and simple linear regression

Overview

Creating maps

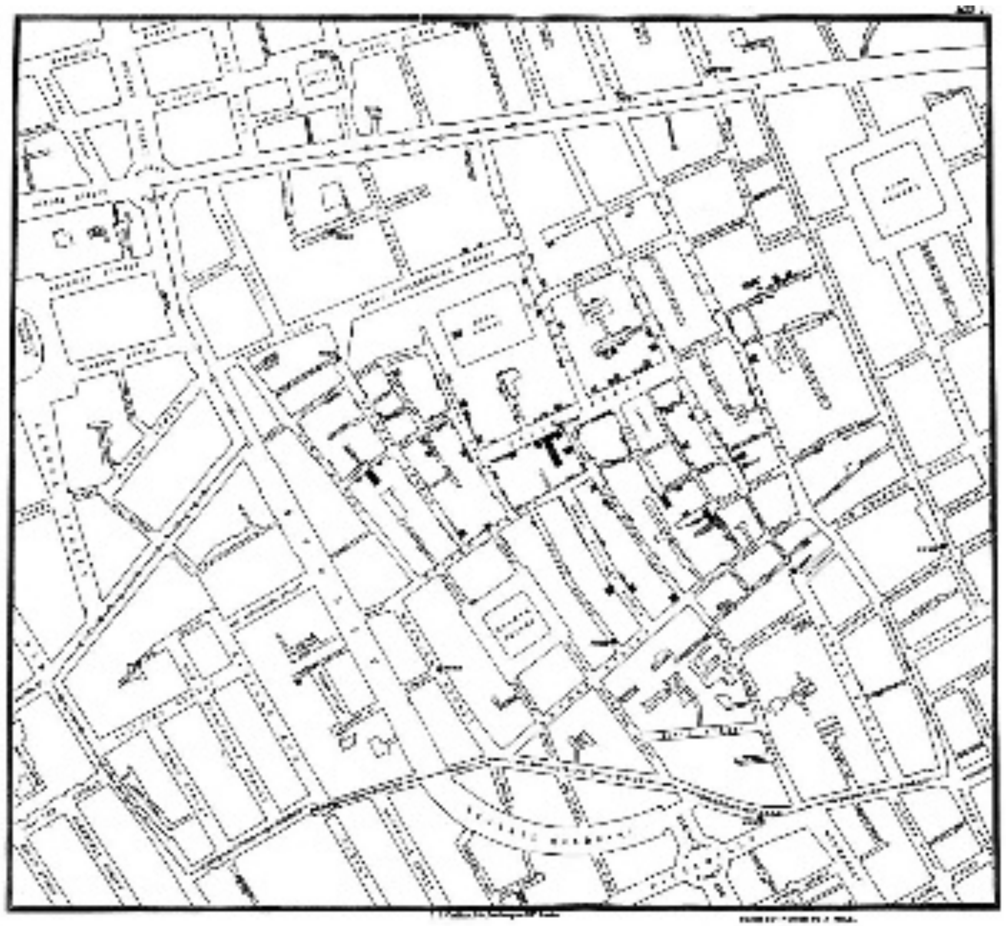
Creating maps in R

Simple linear regression

Simple linear regression in R

ggplot bonus features

Spatial mapping

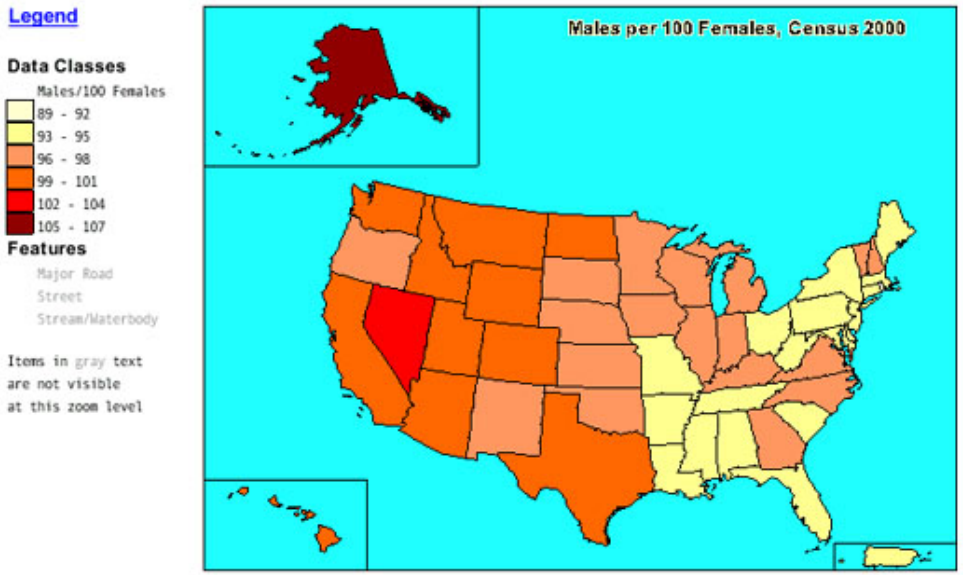


Maps

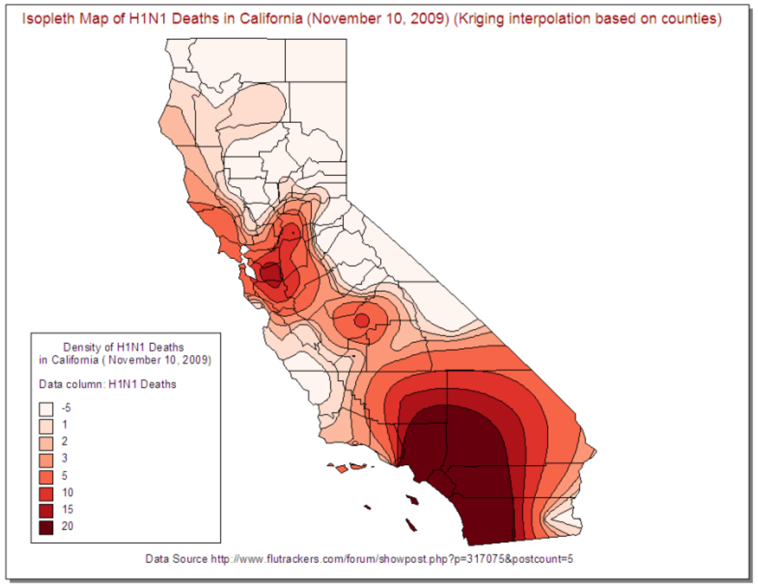
Choropleth maps: shades/colors in predefined areas based on properties of a variable

Isopleth maps: creates regions based on constant values

Choropleth map



Isopleth map



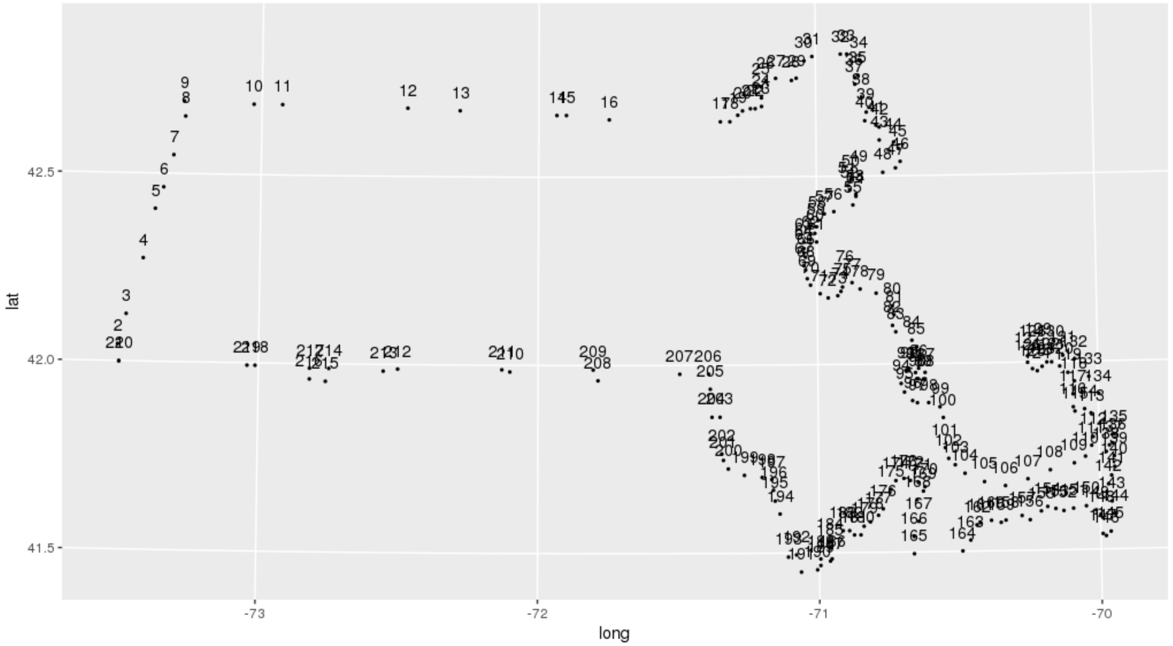
Choropleth maps

```
# has the coordinates for several maps
> library('maps')

# get a data frame with coordinates of states
> states_map <- map_data("state")
```

	long	lat	group	order	region	subregion
1	-87.46201	30.38968	1	1	alabama	NA
2	-87.48493	30.37249	1	2	alabama	NA
3	-87.52503	30.37249	1	3	alabama	NA
4	-87.53076	30.33239	1	4	alabama	NA
5	-87.57087	30.32665	1	5	alabama	NA

Choropleth maps

geom_polygon() works by connecting the dots:



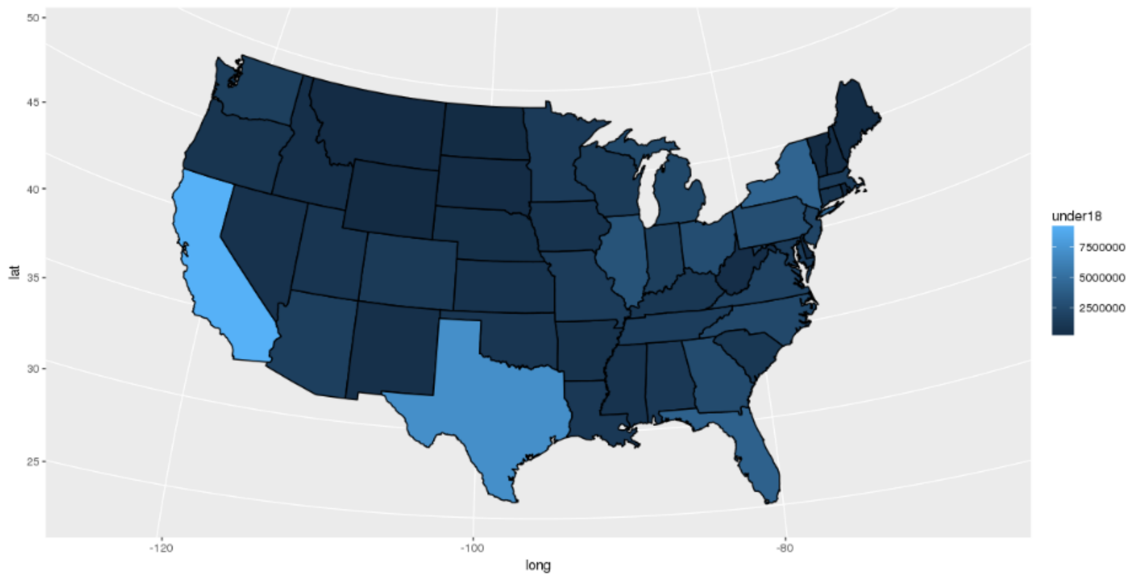
Often need to arrange points first: arrange(states_map, group, order)

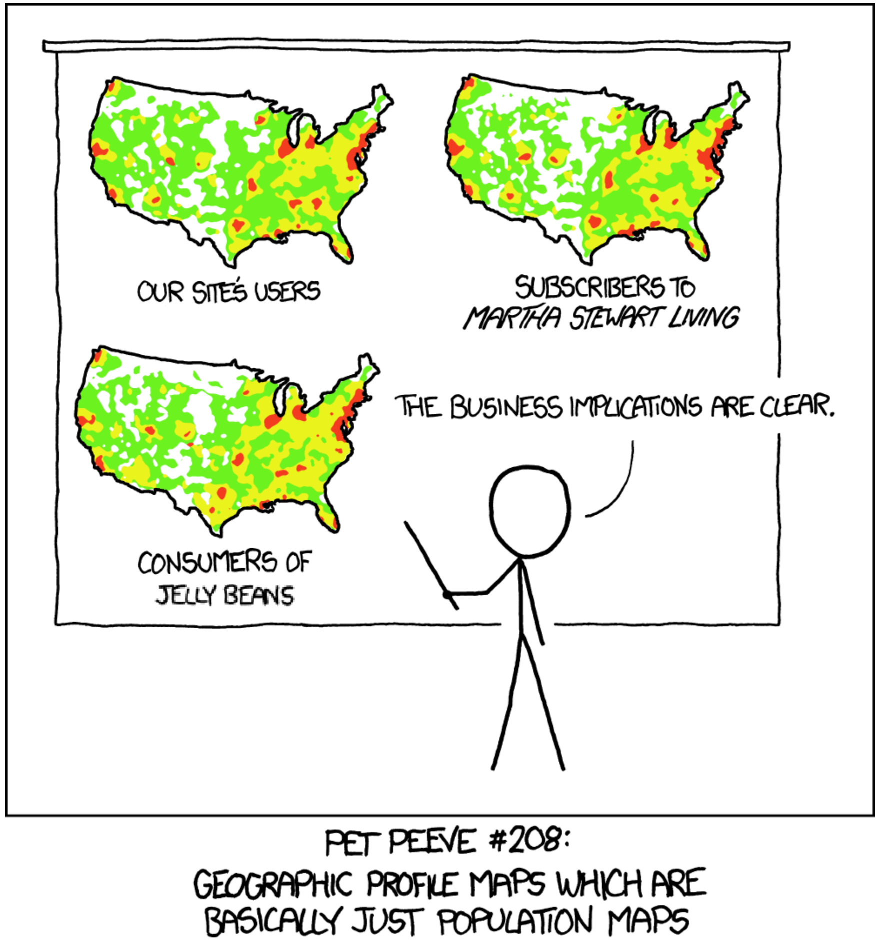
Choropleth maps

```
# has the coordinates for several maps
> library('maps')

# get a data frame with coordinates of states
> states_map <- map_data("state")

# filled white states with black borders
> ggplot(states_map,
         aes(x = long, y = lat, group = group)) +
         geom_polygon(fill = "white", color = "black")
```

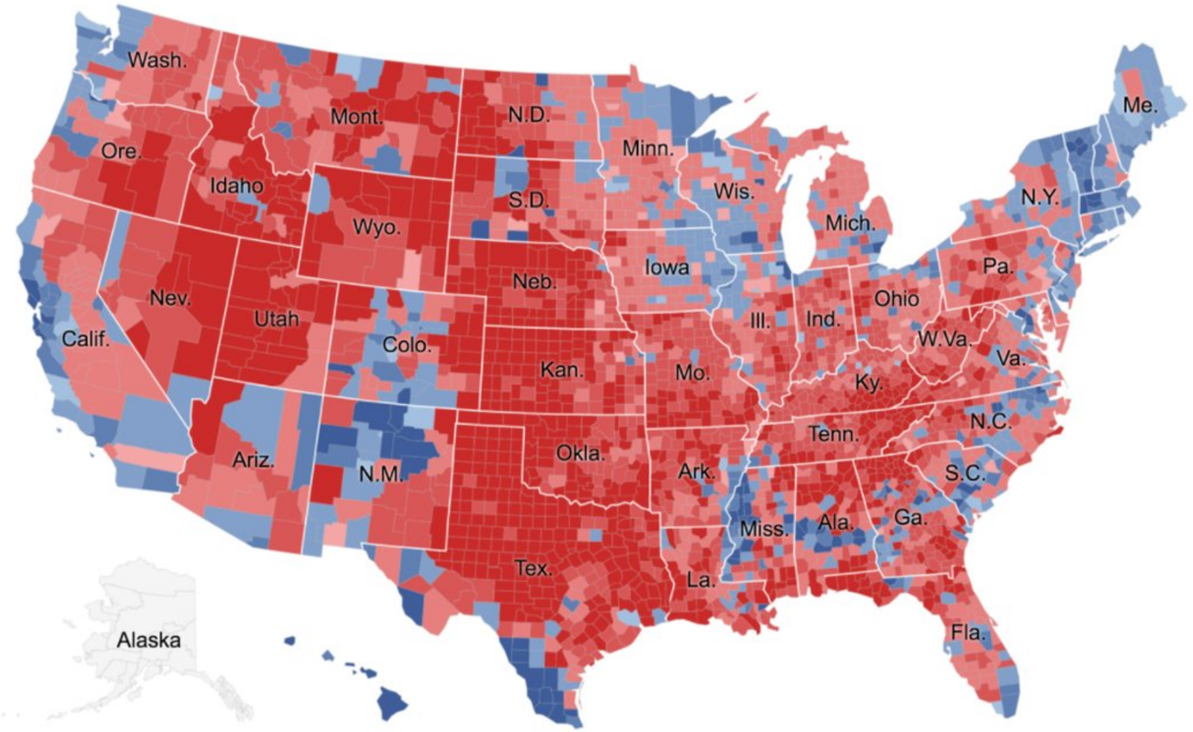
Let’s try it in R!

Pet Peeve #208





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

More maps

Animated map of the 2018 US elections

https://www.nytimes.com/interactive/2018/11/07/us/politics/how-democrats-took-the-house.html

A site with lots of fun data to map:

https://howmuch.net/

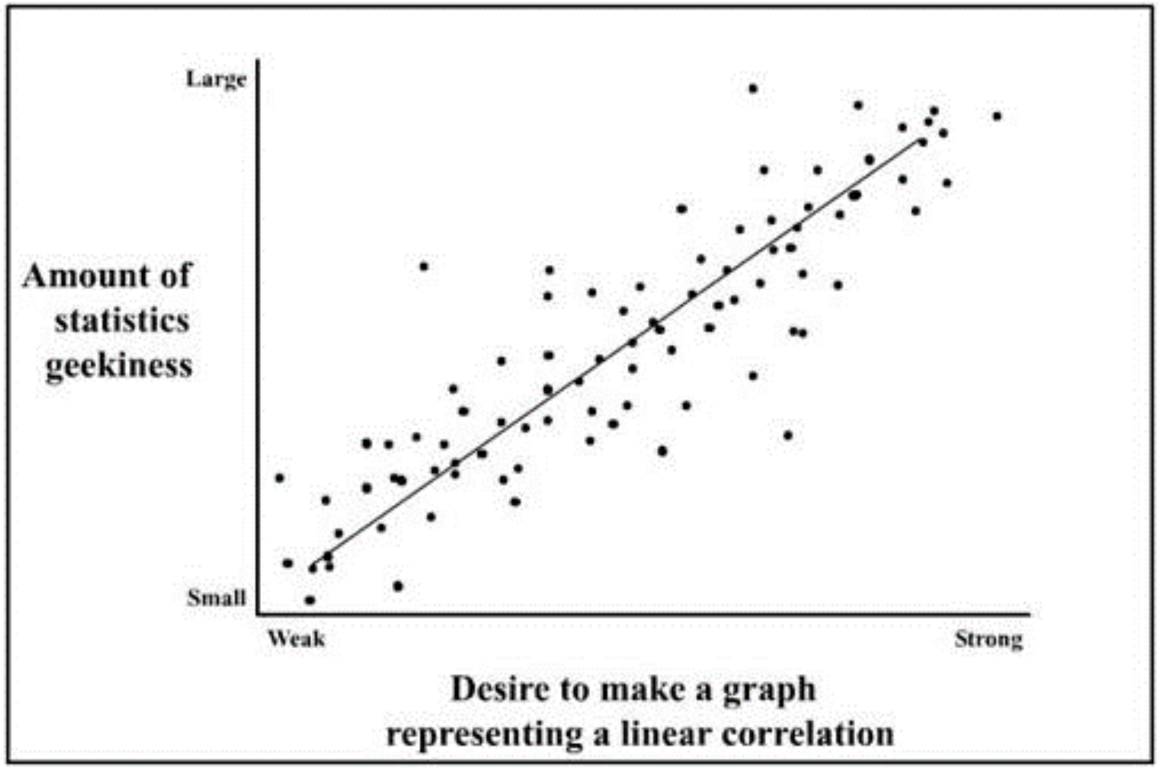
Linear regression

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

$$\hat{y} = f(x)$$

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



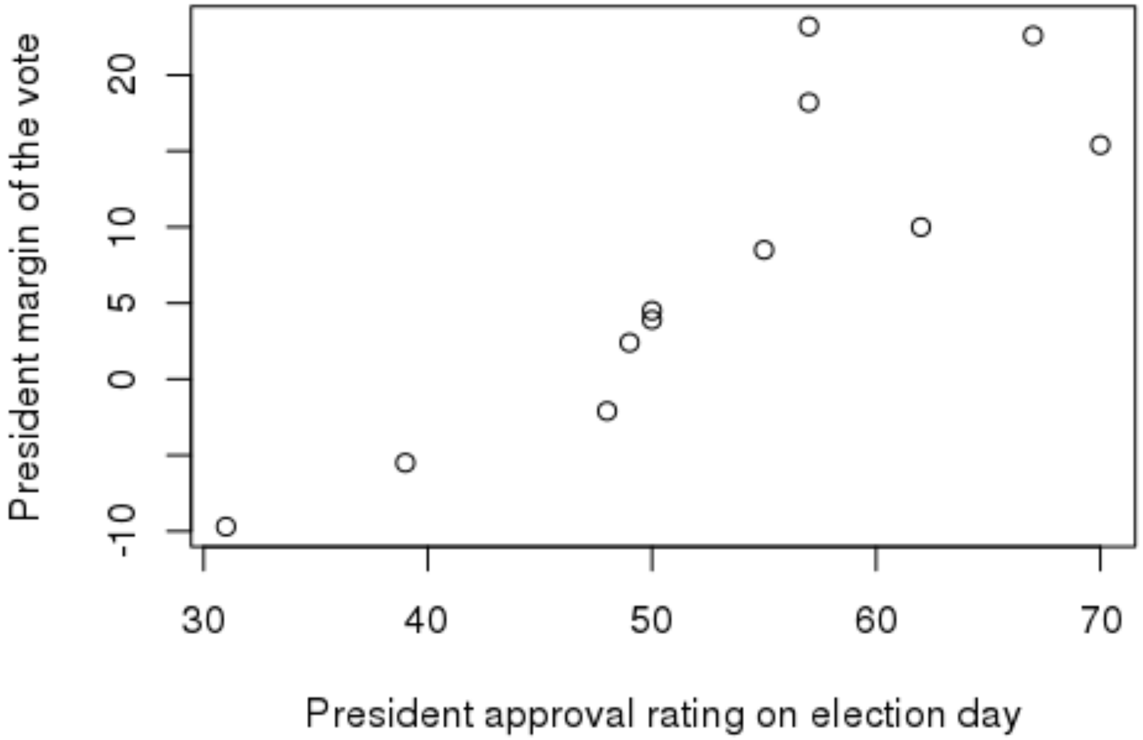
Motivation: Predicting the 2020 election

Predict the margin of the popular vote based on the president's approval rating

Data from an article on the 2012 electon on the Five Thirty Eight website

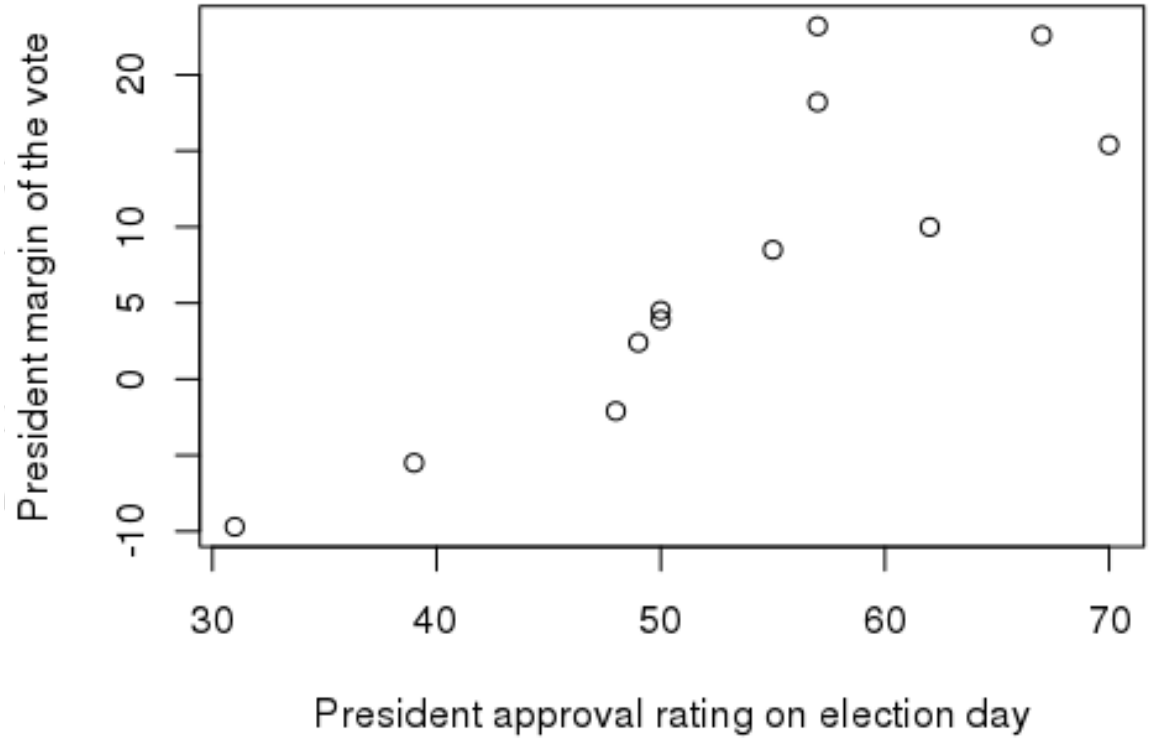
Approval rating vote margin regression line

From last 12 US president's running for reelection



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)`

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

$\hat{\beta}_1$ = 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?

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

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

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

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

R: `lm(y ~ x)`

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

$\hat{\beta}_1 = 0.84$

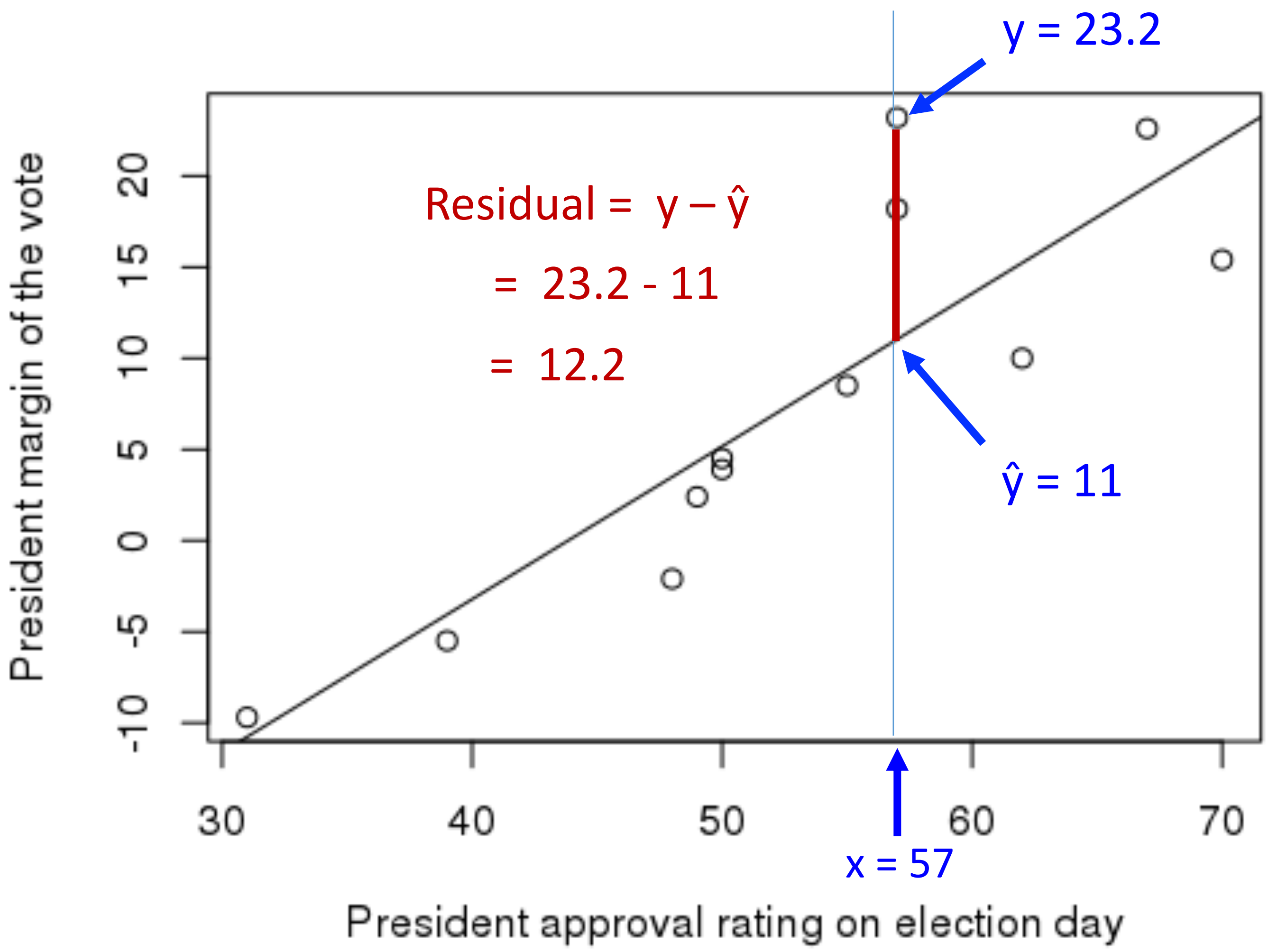
$\hat{y} = -36.76 + .84 \cdot x$

Residuals

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

$$\textit{Residual} \quad = \quad \textit{Observed} - \textit{Predicted} \quad = \quad y - \hat{y}$$

Approval rating vote margin regression line



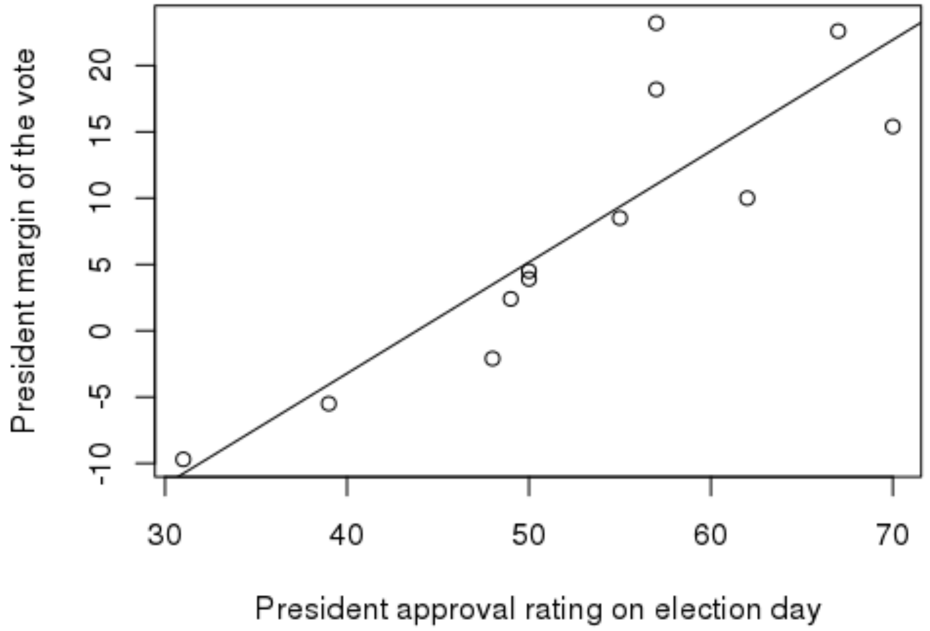
Approval rating vote margin regression line

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

Approval (x)	**Margin obs (y)**	**Margin pred (ŷ)**	**Residuals (y - ŷ)**
62	10	15.23	-5.23
50	4.5	5.17	-0.67
70	15.4	21.94	-6.54
67	22.6	19.43	3.17
57	23.2	11.04	12.16
48	-2.1	3.49	-5.59
31	-9.7	-10.76	1.06
57	18.2	11.04	7.16

Line of ‘best fit’

The **least squares line**, also called **‘the line of best fit’**, is the line which minimizes the sum of squared residuals



Try to find the line of best fit

Approval rating vote margin regression line

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

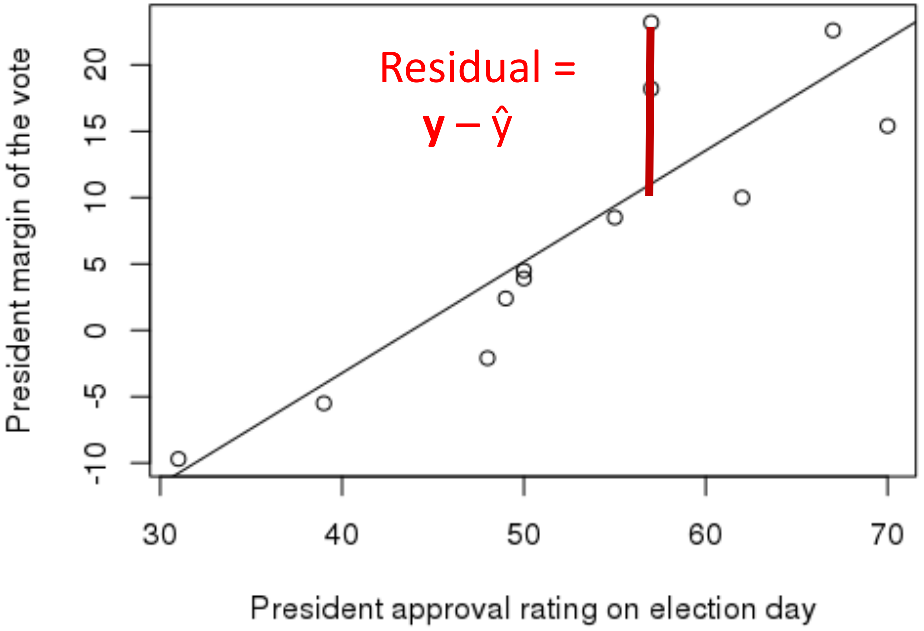
Approval (x)	**Margin obs (y)**	**Margin pred (ŷ)**	**Residuals (y - ŷ)**	**Residuals² (y - ŷ)²**
62	10	15.23	-5.23	27.40
50	4.5	5.17	-0.67	0.45
70	15.4	21.94	-6.54	42.81
67	22.6	19.43	3.17	10.07
57	23.2	11.04	12.16	147.84
48	-2.1	3.49	-5.59	31.29
31	-9.7	-10.76	1.06	1.13
57	18.2	11.04	7.16	51.25

Q: Why do we minimize the sum of ***squared*** residuals rather than just the sum of residuals?

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

To find the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ we minimize the **residual sum of squares**

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



$$residual = e_i = y_i - \hat{y_i}$$

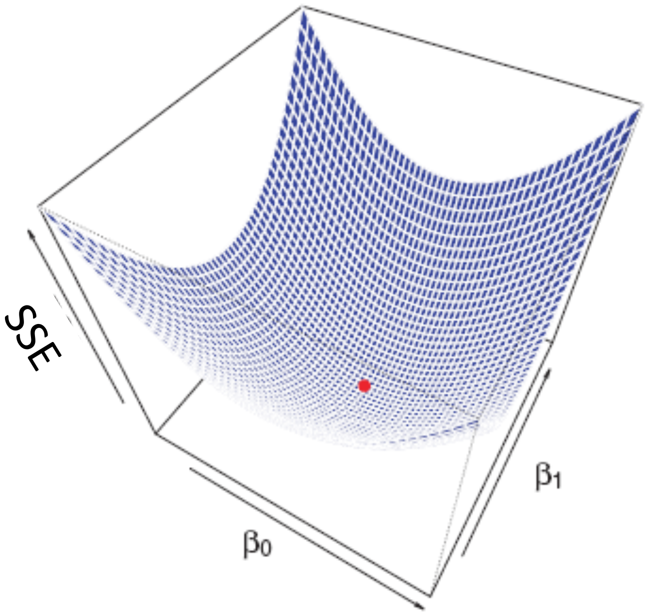
$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

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

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:

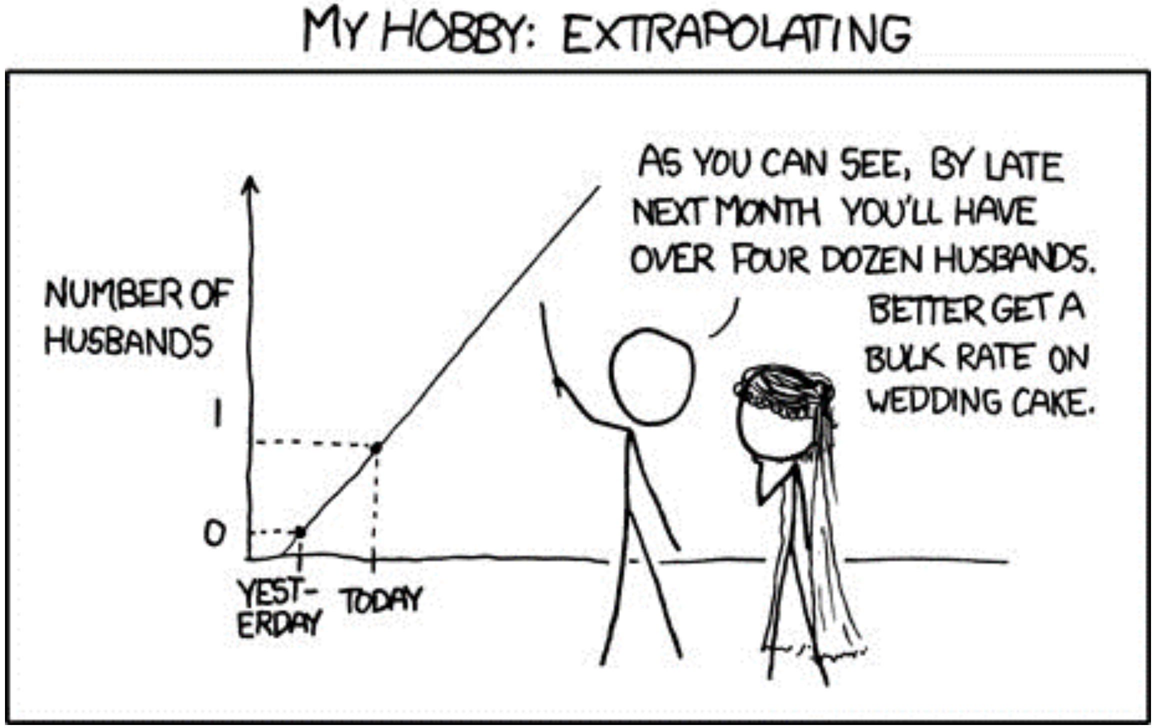
- Take the derivative, set to 0 and solve
- This mathematical convenience is why the squared loss is so commonly used

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

Regression caution # 1

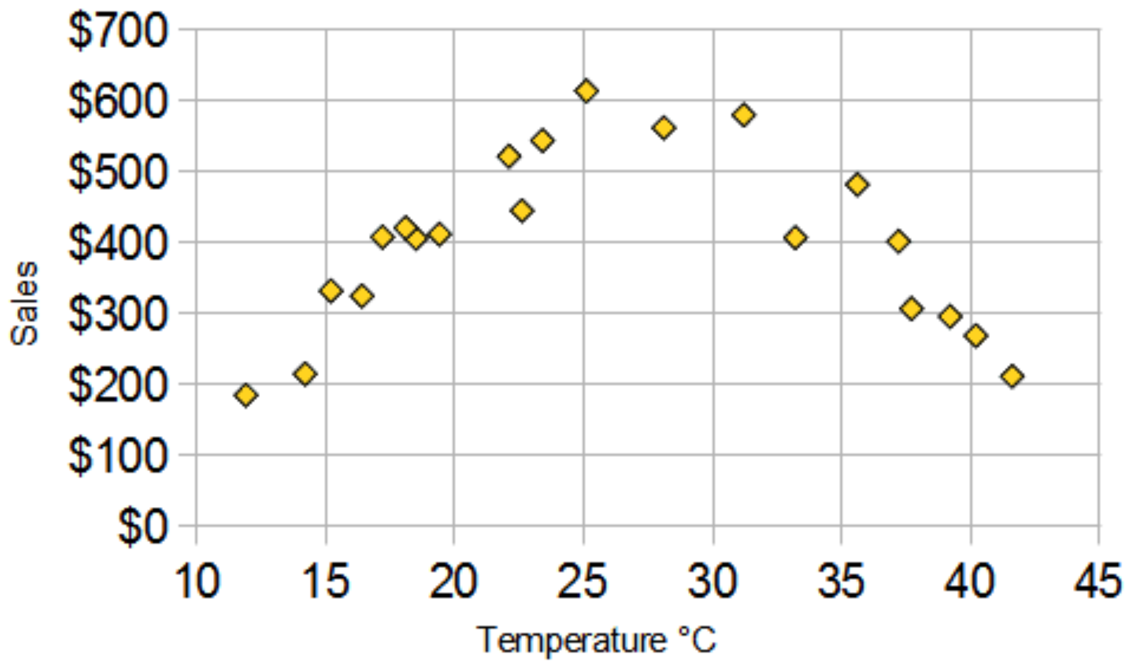
Avoid trying to apply the regression line to predict values far from those that were used to create the line.

- i.e., do not extrapolate too far



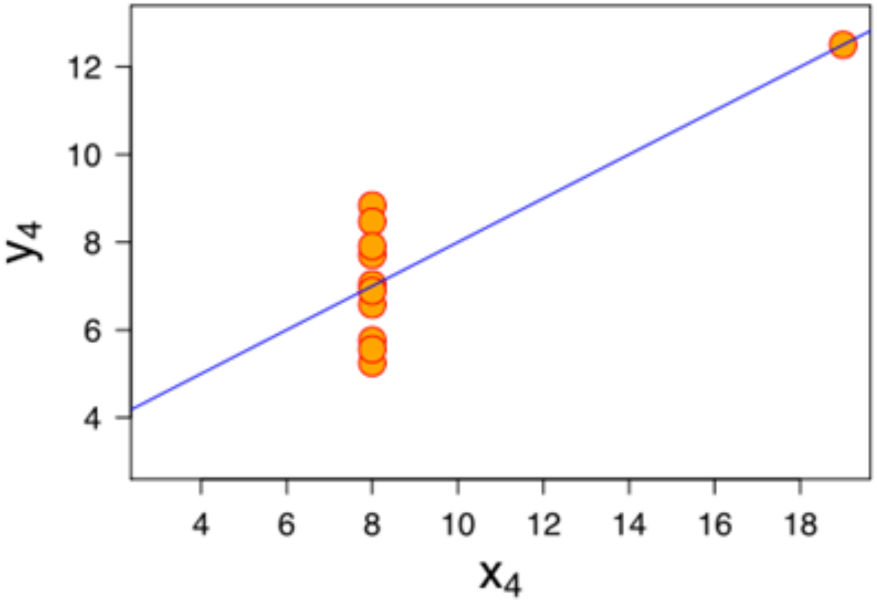
Regression caution # 2

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 a large effect on the regression line.



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

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

Influential point: big outlier and leverage

There are statistics that quantify/describe these concepts

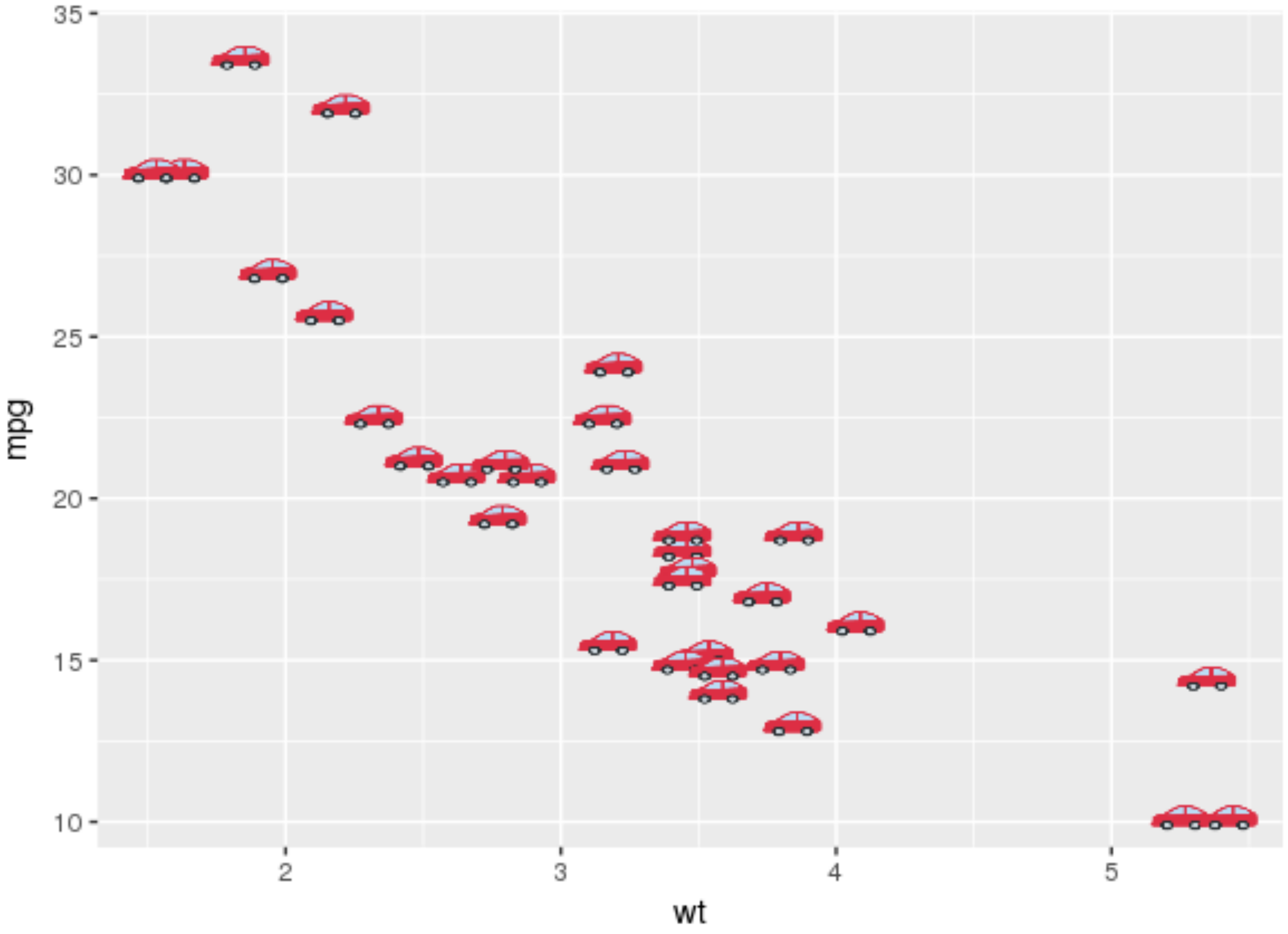
Let’s try simple linear regression in R…

ggplot bonus features

Plotly – interactive plots

```
> library(plotly)

p <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp,
        size = pop, col = continent, frame = year)) +
        geom_point() +
        scale_x_log10()

ggplotly(p)
```

Additional geometries: emoji text



Animation

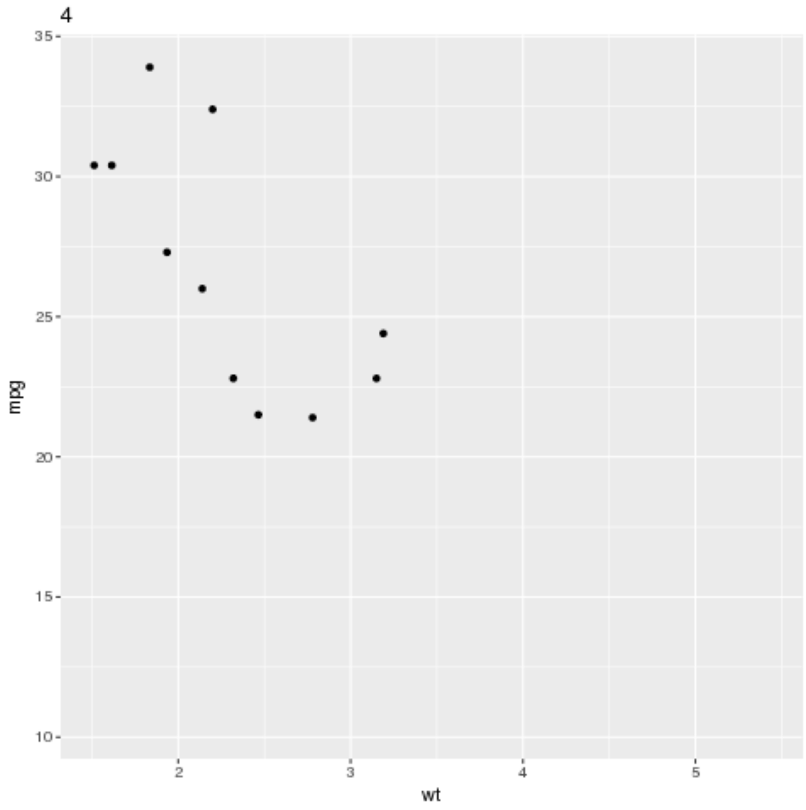
We can create animated images (gifs) using the gganimate package

> library(gganimate)

> library(gapminder)

In the gapminder video, Hans had the following mapping:

- x = gpd per capita
- y = life expectancy
- size = population
- color = continent
- frame = year



Recreating gapminder plot

```
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_x_log10() +
  facet_wrap(~continent) +
   # Here comes the gganimate specific bits
   labs(title = 'Year: {frame_time}’,
          x = 'GDP per capita', y = 'life expectancy') +
   transition_time(year) +
   ease_aes('linear')
```