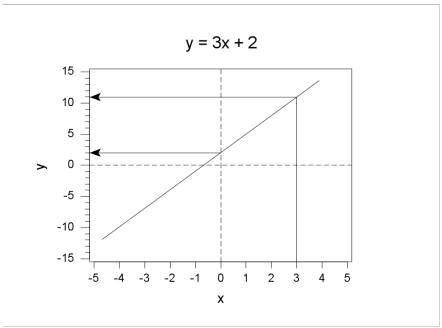
Math 107

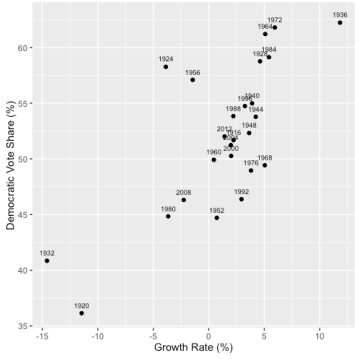
Intro to Simple Linear Regression
(Section 2.6)

Review: Equation of a line



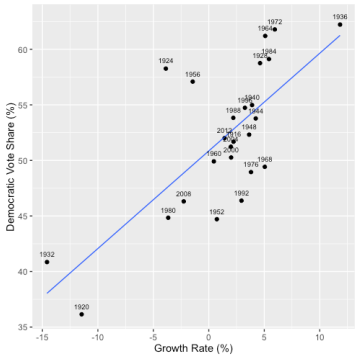
$$\mathbf{y = mx + b}$$

Linear Regression



- **Goal:** predict the Democratic share of the two-party vote based on the growth rate of output per person (real per capita GDP)
- **Target:** determine the “line of best fit”

Notation

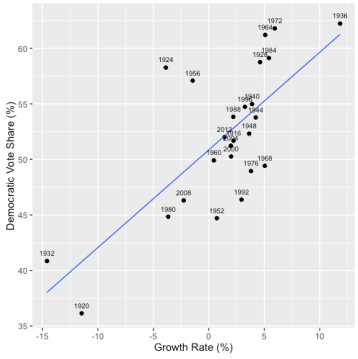


$\mathbf{y}$

$\widehat{\mathbf{y}}$

$\mathbf{e}$

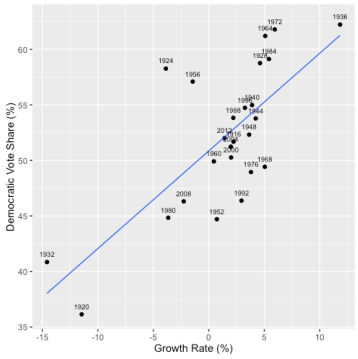
Regression Equation



The estimated regression line is

$$\widehat{\mathbf{y}} = \mathbf{a} + \mathbf{bx}$$

Least Squares



To find the line of best fit we need to **minimize** the sum of the squared errors; i.e., we must minimize

$$\sum \mathbf{e_i^2} = \sum (\mathbf{y_i} - \widehat{\mathbf{y}}_\mathbf{i})^\mathbf{2}$$

R

```
elections <- read.csv("elections.csv")
lm(IVS ~ G, data = elections)

Call:
  lm(formula = IVS ~ G, data = elections)

Coefficients:
(Intercept)            G
    50.8552       0.8787
```

Interpreting the Slope

- The slope tells us how y is predicted to change based on a one unit change in x
- Interpretation:

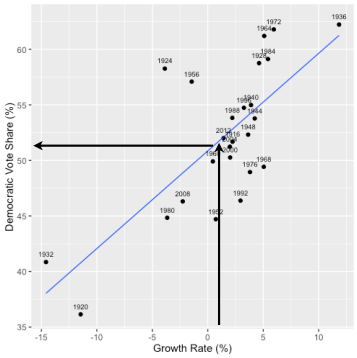
  A one unit increase in the x-variable is associated with an expected/predicted |b| unit increase/decrease in the y-variable.

Interpreting the Intercept

- The intercept is the value of the y-variable with the x-variable is 0
- Interpretation:

  When the x-variable is 0, the y-variable is expected/predicted to be a units.

Prediction



The regression equation can be used to predict y for a given value of x

$$\widehat{y} = 50.855 + 0.879x$$

Suppose we think the growth rate will be 0.87 and want to make a prediction for November.

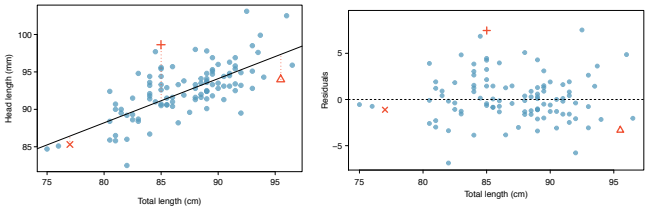
Residuals

$$\mathbf{e = observed - predicted = y - \widehat{y}}$$

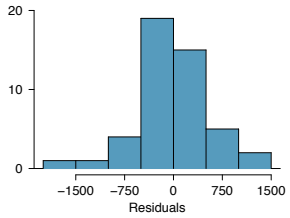
- error we make by using the regression line for prediction
- unless a y-value falls on the regression line, a residual will be either
  - e < 0
  - e > 0

Residual Plots

- If the fitted model really explains the relationship between the two variables, then the residuals should not contain any extra structure.

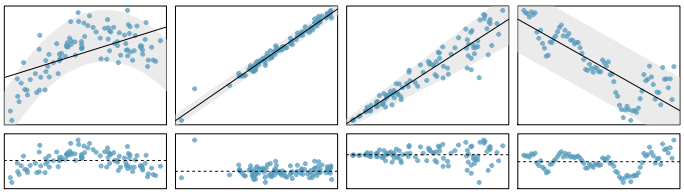


Residual Distributions



A histogram of the residuals should be roughly bell-shaped

What can go wrong?



Caution: Don't Extrapolate

Don't make predictions (far) outside the range of your observed x-values!

Example: Suppose we can describe the financial aid a student receives at a small liberal arts college using the following regression equation:

$$\widehat{\textbf{aid}} = \mathbf{24.3} - \mathbf{0.0431} \times (\textbf{family income})$$

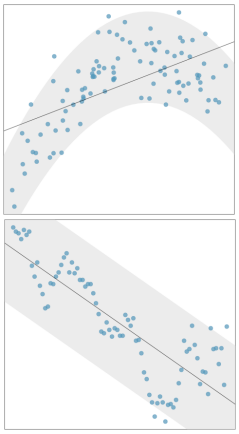
Predict the amount of aid a student will receive if their family income is $1 million.

Caution: Don't Extrapolate

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

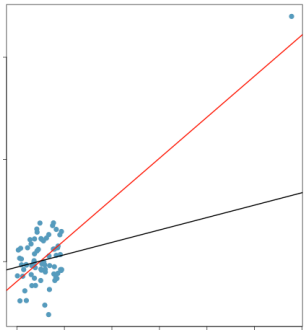
Stephen Colbert
April 6th, 2010

Caution: Plot the Data



Make sure that the relationship between x and y is actually linear!

Caution: Beware of Outliers



Outliers can "hijack" your analysis!