Today: How well does your data fit a line?
- Talk about linear in detail, look at some more complicated ones in R
- Eyeballing is just not rigorous enough

Basic model:  
\[ y_i = b_0 + b_1 x_i + e_i \]

- \( y_i \) is the prediction
- \( b_0 \) is the y-intercept
- \( b_1 \) is the slope
- \( x_i \) is the predictor
- \( e_i \) is the error

Which of these are random variables?
* A: All but \( x_i \), the \( bs \) are estimated from random variables, \( e \) is difference between random variables
* So, we can compute statistics on them

Two criteria for getting \( bs \)
- Zero total error
- Minimize SSE (sum of squared errors)
- Example of why one is not enough: two points, infinite lines with zero total error
- Squared errors always positive, so this criterion alone could overshoot or undershoot

Deriving \( b_0 \) is easy
- Solve for \( e_i \): \( y_i - (b_0 + b_1 x_i) \)
- Take the mean over all \( i \): \( \bar{x} = \bar{y} - b_0 - b_1 \bar{x} \)
- Set mean error to 0 to get \( b_0 = \bar{y} - b_1 \bar{x} \)
- Now we just need \( b_1 \)

Deriving \( b_1 \) is harder
- \( SSE = \) sum of errors squared over all \( i \)
- We want a minimum value for this
- It's a function with one local maximum
- So we can differentiate and look for zero
- \( s_y^2 = 2b_1s_{xy} + b_1^2s_x^2 \), then take derivative
- $s_{xy}$ is correlation coefficient of $x$ and $y$ (see p. 181)
- In the end, gives us $b_1 = \frac{s_{xy}}{s_x^2}$
  * Correlation of $x$ and $y$ divided by variance of $x$
  * $\frac{\sum xy - n\bar{xy}}{\sum x^2 - n(x)^2}$

- SS*
  - SSE = Sum of squared errors
  - SST = total sum of squares (TSS): difference from mean
  - SSo = square $\bar{y}$ $n$ times
  - SSY = square of all $y$, so SST = SSY - SSo
  - SSR = Error explained by regression: SST - SSE

- Point of above: we can talk about two sources that explain variance: sum of squared difference from mean, and sum of errors
  - $R^2 = \frac{SSR}{SST}$
  - The ratio is the amount that was explained by the regression - close to 1 is good (1 is max possible)
  - If the regression sucks, SSR will be close to 0

- Remember, our error terms and $b$s are random variables
  - We can calculate stddev, etc. on them
  - Variance is $s_e^2 = \frac{SSE}{n-2}$ - MSE, mean squared error
  - Confidence intervals, too
  - What do confidence intervals tell us in this case?
    * A: Our confidence in how close to the true slope our estimate is
    * For example: How sure are we that two slopes are actually different
    - When would we want to show that the confidence interval for $b_1$ includes zero?

- Residuals
  - AKA error values
  - We can expect several things from them if our assumptions about regressions are correct
  - They will not show trends: why would this be a problem
    * Tells us that an assumption has been violated
    * If not randomly distributed for different $x$, tells us there is a systematic error at high or low values - error and predictor not independent
  - Q-Q plot of error distribution vs. normal distribution
  - Want the spread of stddev to be constant across range

- Switch to R
  - Show example of linear fitting (good fit)
  - Show example of linear fitting (bad fit)
  - Show example of polynomial fit (intercept and 3 coefficients)
• For next time
  – I won’t be here week after spring break
  – papers3 due Tuesday of spring break week
  – On Thursday, we will have some guest students talk about paper writing process
  – lab2 now due Friday after spring break
    * I want some more from you now, so be sure to update your fork
    * Mainly, I want to know how you will improve the graph you are reproducing, and to actually look a bit at the code you find