

UNIVERSITY *of* WASHINGTON

# Data Science UW

# Methods for Data

# Analysis

---

More on Hypothesis Testing, The Central Limit Theorem,  
And an introduction to Regression

Lecture 4

Nick McClure





Excellent health statistics - smokers are less likely to die of age related illnesses.'

W

# Topics



- > Review
- > More on hypothesis testing
- > Central Limit Theorem
- > Introduction to Regression



# Review

- > Sampling Methods
- > Law of Large Numbers
- > Hypothesis Testing
  - Normal testing
  - One tailed vs Two tailed
  - P-values
  - T-test (Student's, Welch's)
  - Chi-Squared
  - Fisher's Exact
- > Outliers



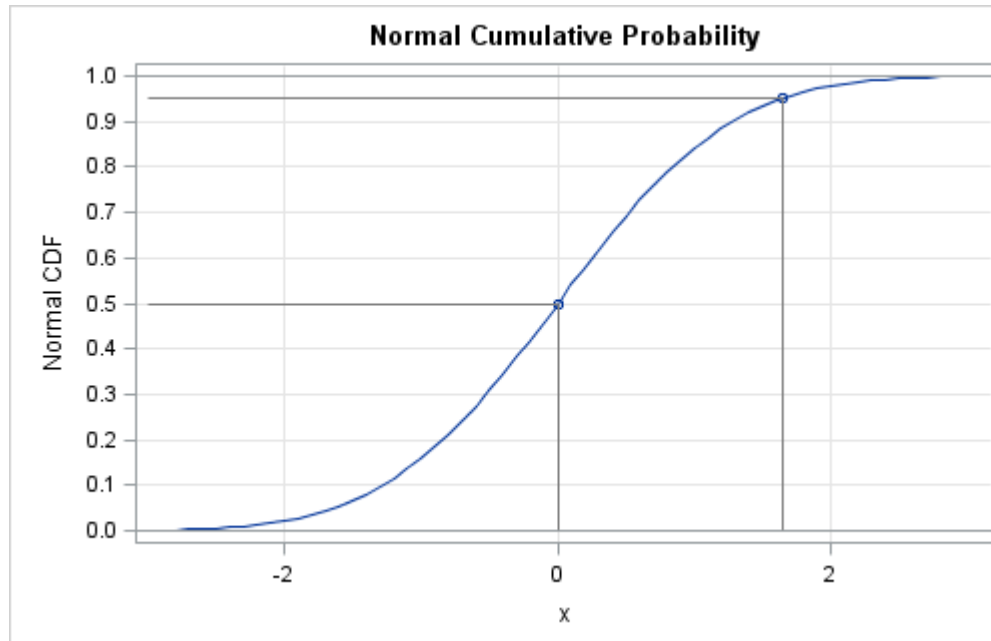
# Hypothesis Testing Summary (so far)

- > If data is normal,
  - If you know population mean and variance,
    - > Use standard normal 'z-test'.
  - If you just know population mean,
    - > Use t-test (unpaired data).
    - > Use Welch's t-test (paired data).
- > For categorical comparison tests,
  - If the sample/subgroup size is large enough,
    - > Use Chi-squared test
  - If the sample/subgroup size is small,
    - > Use Fisher's Exact test.
- > How do we know the data is normal?

**W**

# Testing for Normality

- > Kolmogorov-Smirnov test (K-S test).
  - Tests if two distributions are similar.
- > Consider the Normal Cumulative Distribution Function (CDF).



- > Any similar distribution should have a similar CDF.

W

# Testing for Normality

- > The K-S statistic is just the maximum vertical distance between two CDFs.
- > Note: the K-S test can test departure from any hypothetical distribution, not just normal.
- > R-demo



# Testing for Normality

- > Also, the Shapiro-Wilk test can tell us a test statistic for normality.
  - Tests the difference in expected and sample ‘moments’.
  - Moments:
    - > 1<sup>st</sup> moment = mean
    - > 2<sup>nd</sup> moment = variance
    - > 3<sup>rd</sup> moment = skewness
    - > 4<sup>th</sup> moment = kurtosis
    - > ...
  - Slightly more powerful than the K-S test.





# Testing for Normality

- > ALWAYS do a qq-plot to look at normality. (qqnorm())
- > R-demo.



# Testing Between Multiple Groups

- > What if we had multiple groups and we wanted to compare their means?
- > Why can't we just do multiple two-sample t-tests for all pairs?
  - Results in increased probability of accepting a false hypothesis.
  - E.g., if we had 7 groups, there would be  $(7 \text{ Choose } 2) = 21$  pairs to test. If our alpha cutoff is 5%, then we are likely to accept about 1 false hypothesis ( $21 * 0.05$ ).



# Testing Between Multiple Groups

## > Null Hypothesis:

- All groups are just samples from the same population.

## > Alternative Hypothesis:

- At least one group has a statistically different mean.

## > This type of analysis is called “Analysis of Variants”, or ANOVA.

- We make data independence and normality assumptions first.
- Our test statistic is based on:

$$\text{statistic} \sim \frac{\text{between group variability}}{\text{within group variability}}$$



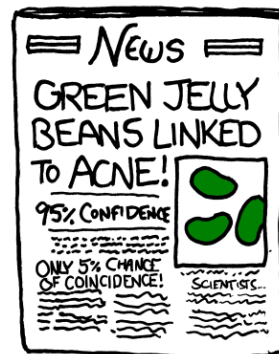
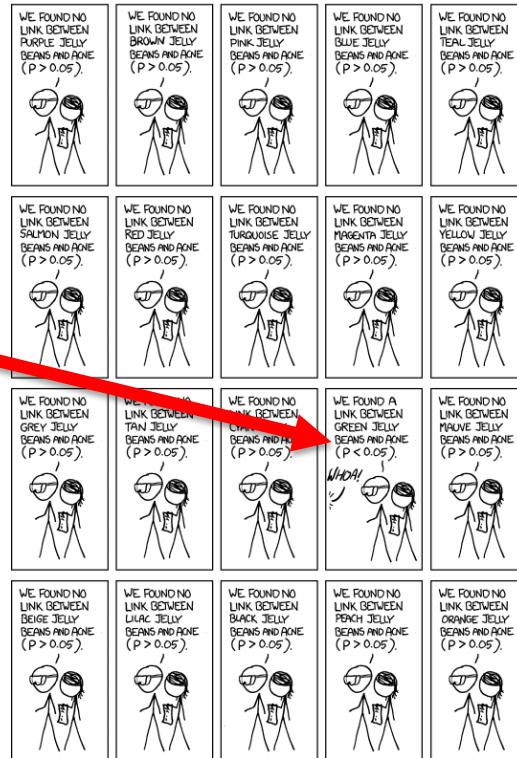
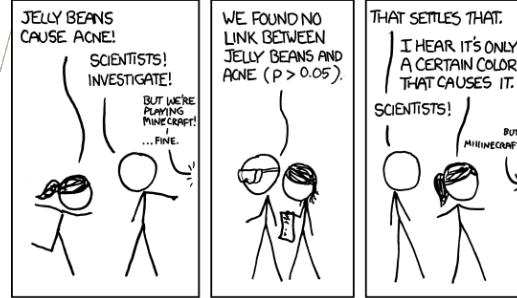
# Performing Multiple Hypothesis Tests

- > For non-ANOVA methods, remember that performing many hypothesis tests increases our risk of incorrectly rejecting a null-hypothesis.
- > To compensate for this we decrease the p-value cutoff.
- > The most common way of doing this is with the Bonferroni Correction.

$$p' = \frac{p}{(\# \text{ of Hypotheses})}$$

- > This correction is argued to be too strong and other approximations for a new-p can be used instead.
  - Tukey's Range Test
- > This is VERY important in genetics/bioinformatics.
- > R Example

**W**



W

# Additional Hypothesis Tests

- > Parametric test types:
  - Mean comparison
  - Variance comparison
  - More distribution comparisons



# Central Limit Theorem

- > If we sample a population over and over, the set of means of all samples are normally distributed, regardless of the population distribution.

$$\bar{X} \sim N(\text{mean}, \frac{\text{variance}}{\sqrt{n}})$$

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- > Compare to Law of Large numbers ('proof' by R), shown in previous class.



# Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- > We can use this central limit theorem to generate confidence intervals on expressing the population mean.
- > We know the sample mean, sample variance, and number of samples.
- > Then we know how our estimate of the population mean is distributed (from above formula).
- > We can then generate 90%, 95%, ... confidence intervals around our sample mean.





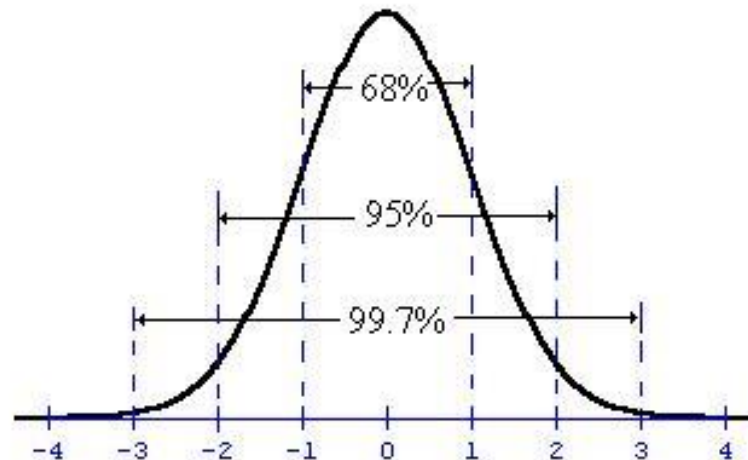
# Confidence Intervals

- > Confidence intervals are a way to express uncertainty in *population* parameters, as estimated by the sample.
- > E.g. If we create a 95% confidence interval for the population mean, say  $\hat{\mu} = \bar{X} = 10 \pm 5$ 
  - Then we say that the true population mean,  $\mu$ , has a 95% chance of being between 5 and 15.



# Confidence Intervals

- > To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.
  - Repeatedly sample from the population.
  - Calculate the mean for each sample.
  - Use the average of the sample means as the population estimate and create a C.I. based on the s.d. of the sample means.
  - R demo



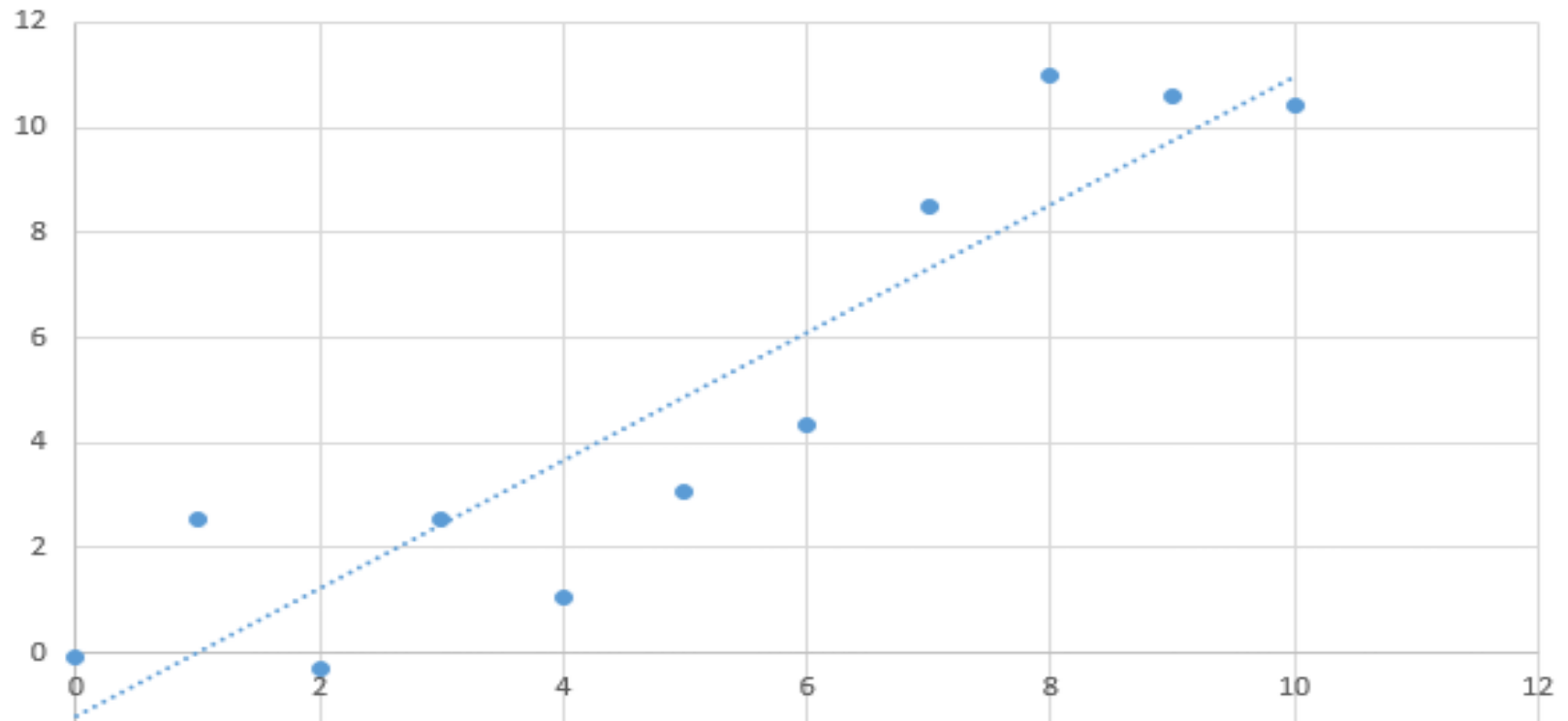
**W**

# Regression Models

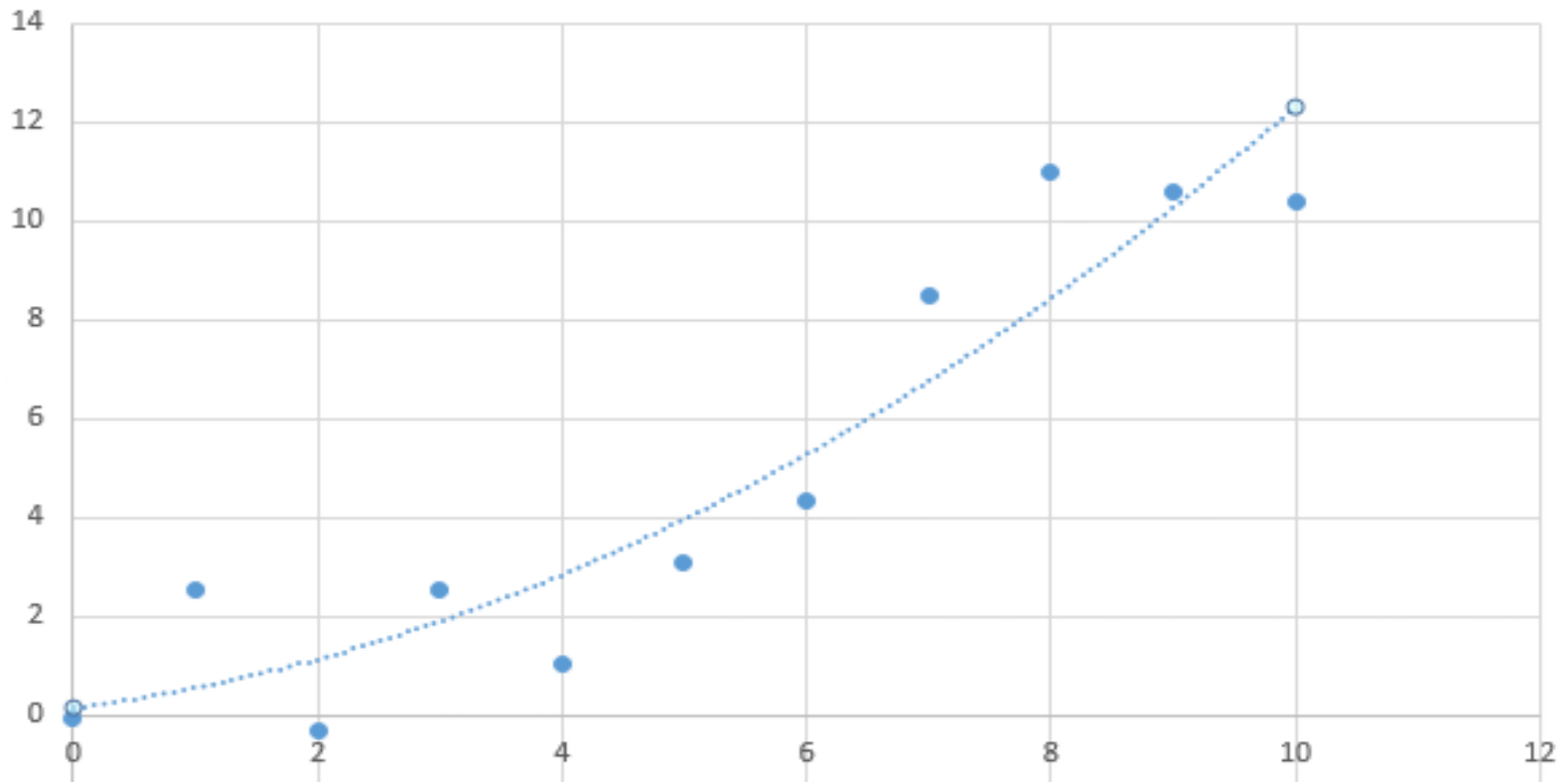
- > The goal of regression is to produce a model that represents the best fit to some observed data.
- > Typically the model is a function describing some type of curve (lines, parabolas, etc.) that is determined by a set of parameters (e.g., slope and intercept).
- > “Best fit” means that there is an optimal set of parameters according to an evaluation criteria we choose.



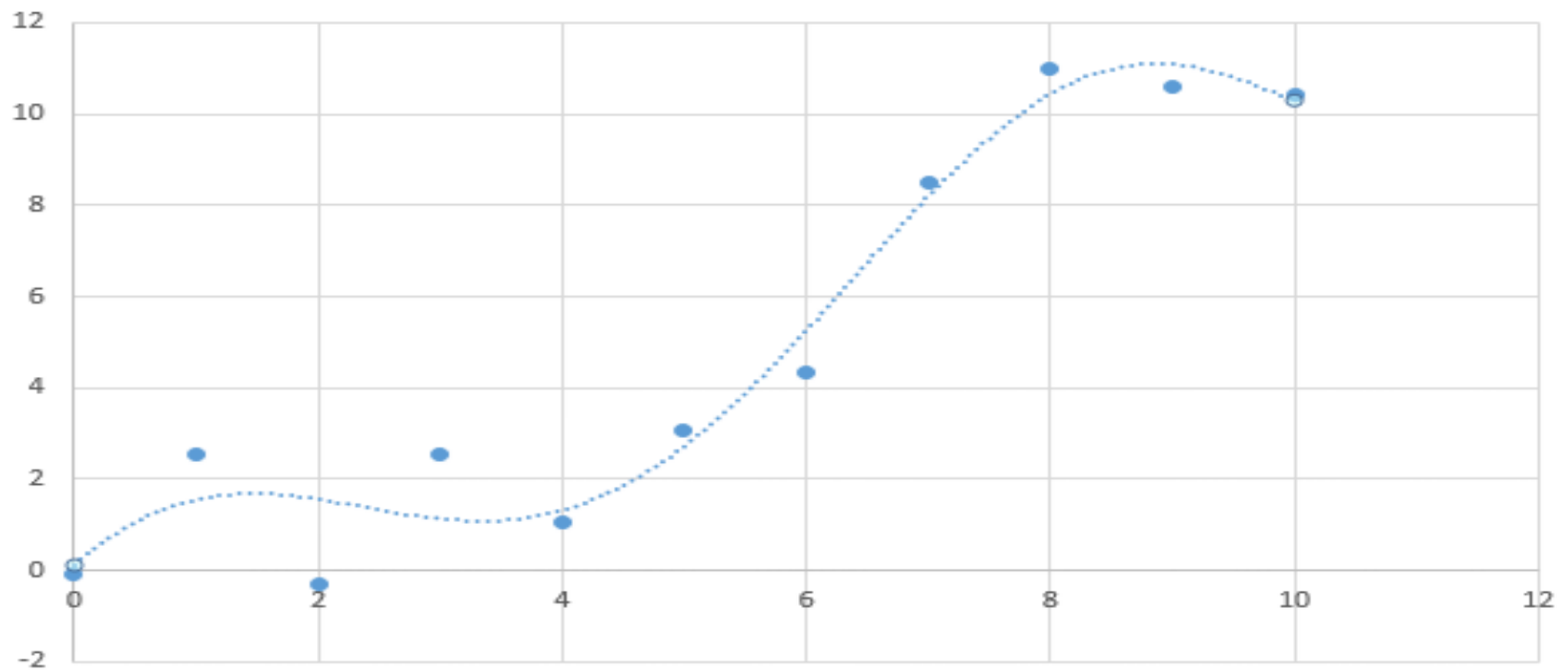
# Regression: Linear



# Regression: quadratic



# Regression: High Order



# Regression Models

- > Which one of the preceding examples is correct?
- > In a sense, all of them are. They all give decent approximations to the data.
- > It's hard to tell, just from looking at these plots whether any of them in fact will continue to perform well as more data is received.
- > We don't know if these models will generalize



# Linear Regression

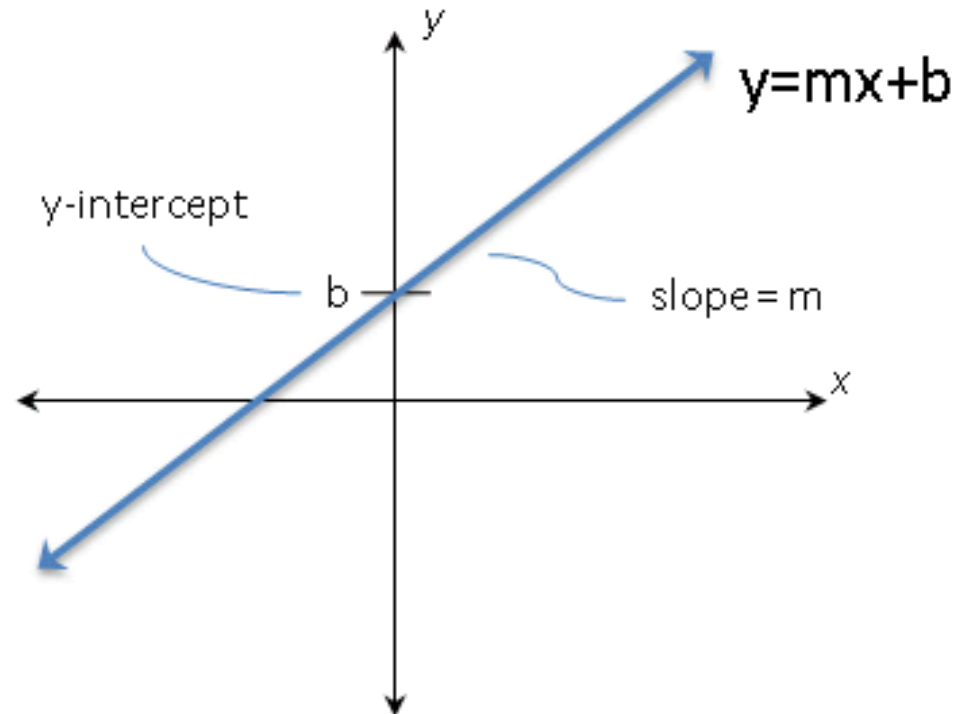
- > Response (Dependent) variable: the variable of primary interest in a study- the one you are trying to predict or explain.
- > Explanatory (Independent) variable: the variable that attempts to explain the observed outcomes of the response variable.
- > There are two types of parameters in linear models:
  - The intercept (y-intercept).
  - The slope, rise over run, or change in Y divided by change in X





# Linear Regression

- > When  $x = 0$ , then  $y = b$ .
- > When  $x = -(b/m)$  then  $y = 0$ .

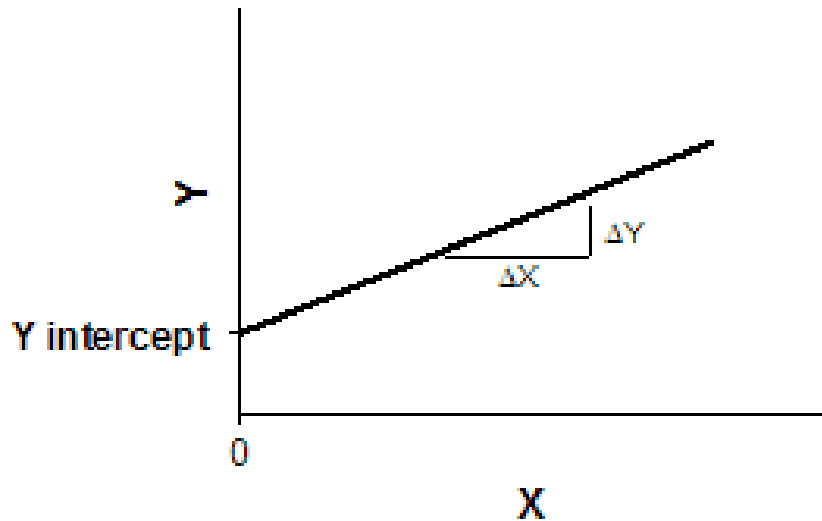


**W**

# Linear Regression

> Interpret slope:  $m = \frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x}$

- If  $x$  changes by  $\Delta x$ , then  $y$  must change by  $\Delta y$  in order for the slope to stay the same (and it must).



Given two points,  $(x_1, y_1)$ ,  $(x_2, y_2)$

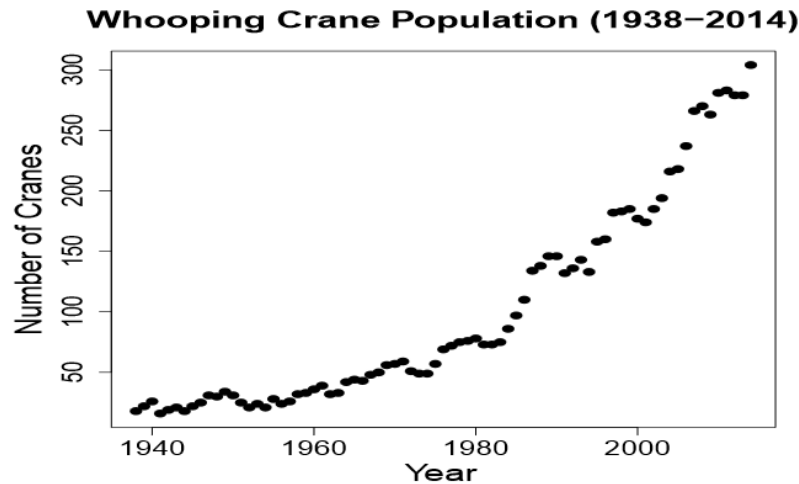
$$m = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(y_1 - y_2)}{(x_1 - x_2)}$$

R example!

W

# Linear Regression

- > Consider the relationship between Whooping Crane population and year below.

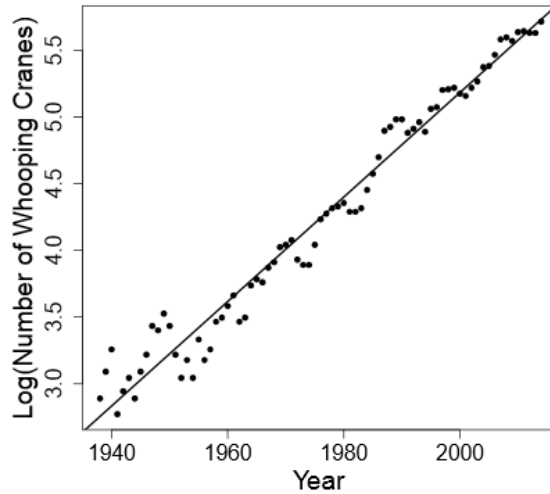


- > Possible regression solutions:
  - Transform the response variable, to linearize the relationship.
  - Fit a nonlinear regression model to the data.

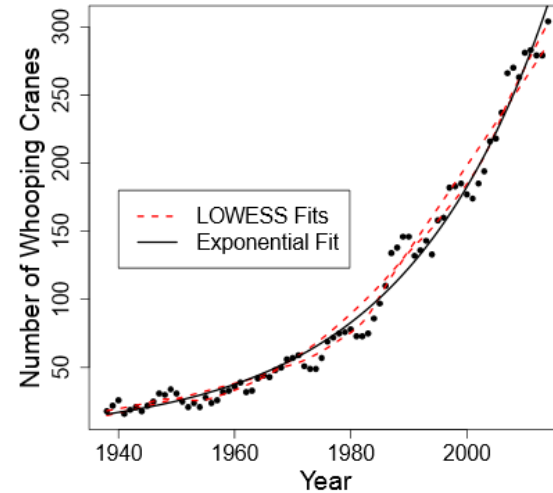


# Linear Regression

Using a Log Transformation



LOWESS & Nonlinear Fit



> How would we decide on a 'best' model?

W

# Linear Regression

- > We use the method of least squares to find the best fit line:  $y_i = mx_i + b + \varepsilon_i$

$$\min_{m, b} \sum_{i=1}^n (\varepsilon_i)^2 = \min_{a, b} \sum_{i=1}^n (y_i - (mx_i + b))^2$$

- > Explicit solutions exist (using calculus).
- > Computers are really good at finding minimums of equations. We let them do this for us.



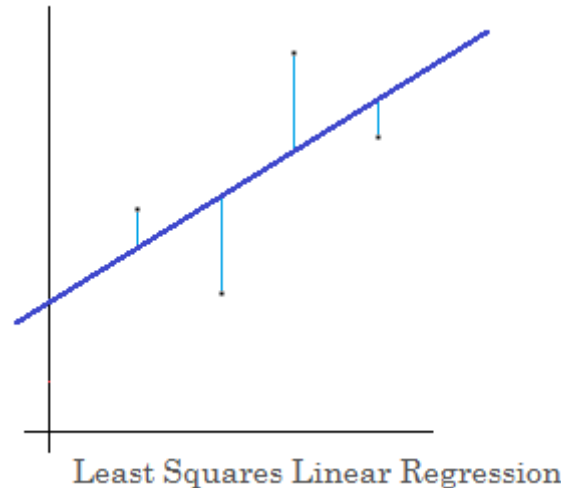
# Linear Regression

- > The method of least squares finds the best fit line.
  - The mean of the errors from the best fit line is zero.
  - This means there is no 'bias' in our prediction.



# Linear Regression

- > Linear regression is the most common.
  - Fit a line (2D), plane (3D), or a hyperplane to the observed data.
- > We need to define an error metric for a line through points.

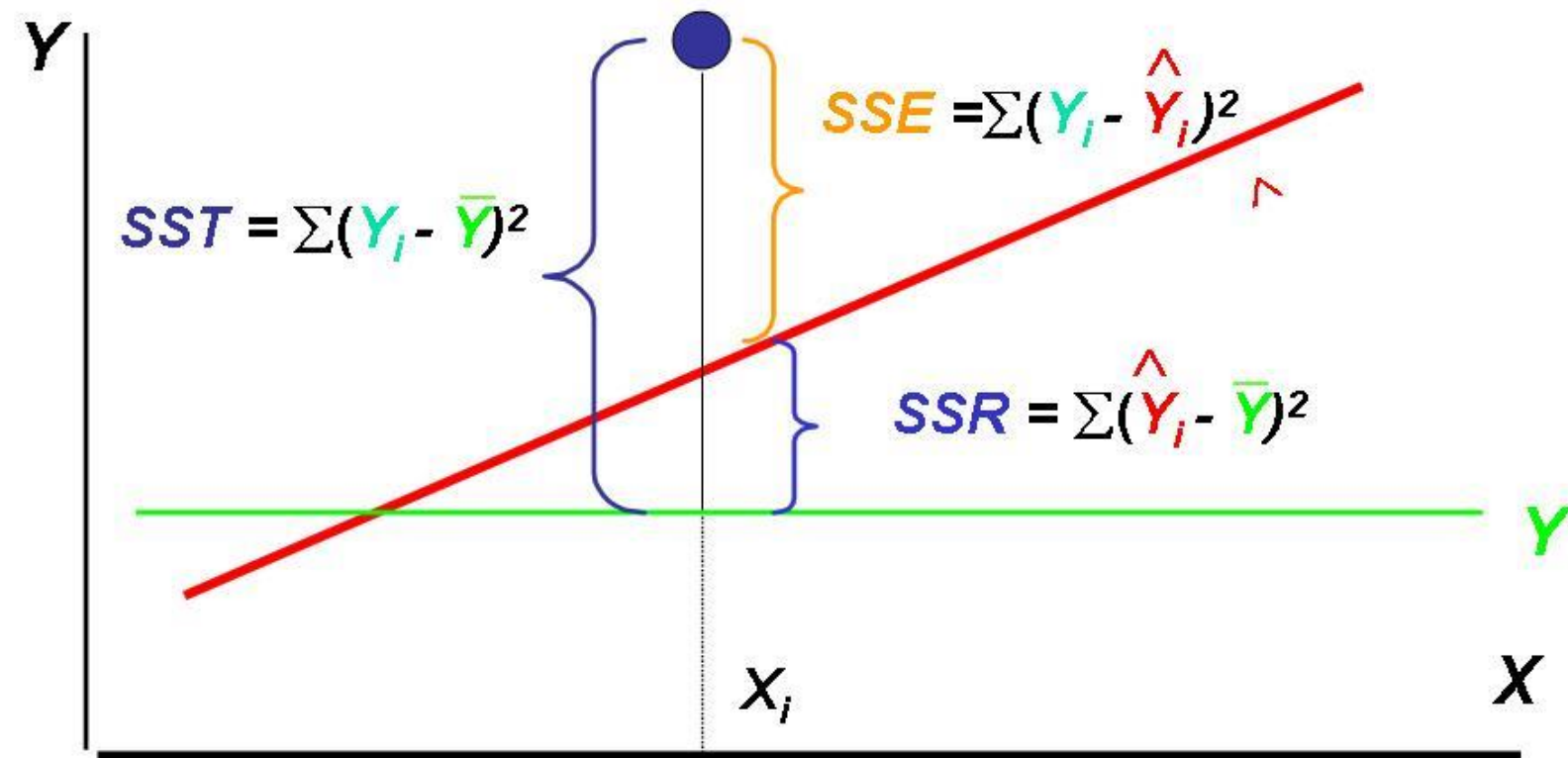


- > We use the sum of the squared error between the predicted and actual.
- > R demo.

W

# Linear Regression

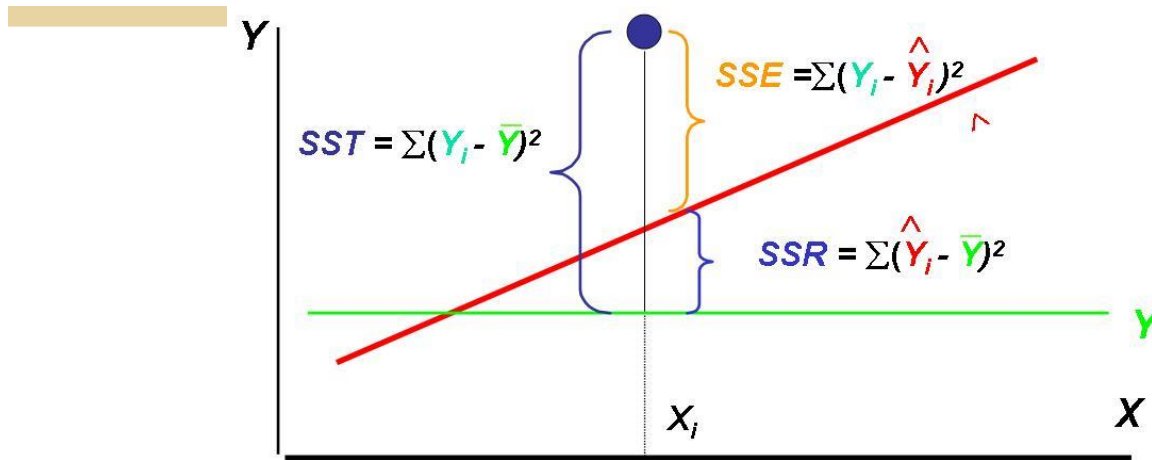
> With modeling, we are interested in the SSE, SSR, SST.



W



# Linear Regression



- > R-squared is called the coefficient of determination.
- > It indicates how well the data fits a specified model.
- > For linear models, we define this as:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

W

# Linear Regression

- > We can also measure accuracy of the line using Root Mean Squared Error (RMSE).
  - Using this as an estimate of the error means we are losing one more degree of freedom than the standard deviation, so we write the RMSE as

$$RMSE = \frac{SSE}{n - 2}$$



# Assignment

## > Complete Homework 4:

- Fit a linear model to the Chicago Diabetes Hospital data as follows:
  - > Transform the data to sum across all zip codes.
  - > Then plot and fit a line to the following:
    - Num. Hospitalizations vs. Crude Admittance Rate
    - $\Delta$  Num. Hospitalizations vs.  $\Delta$  Crude Admittance Rate
  - > Summarize your finding of each model.
    - Be sure to interpret the slope for each model (think of the units).
- You should submit:
  - > Just one R-script.
  - > Read Statistical Thinking for Programmers Chapters 6 and 7.

