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.'

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

Fisher's Exact Test

	Cat. 1	Cat. 2
Successes	A	B
Failures	C	D

> Hypothesis: (Assume we know the total # of successes)

- Null: The proportion of successes in Category 1 is not less than that in Category 2.
- Alternative: The proportion of successes in Category 1 is less than in Category 2.

We observe this outcome

0	5	1	4	2	3	3	2	4	1	5	0
5	2	4	3	3	4	2	5	1	6	0	7

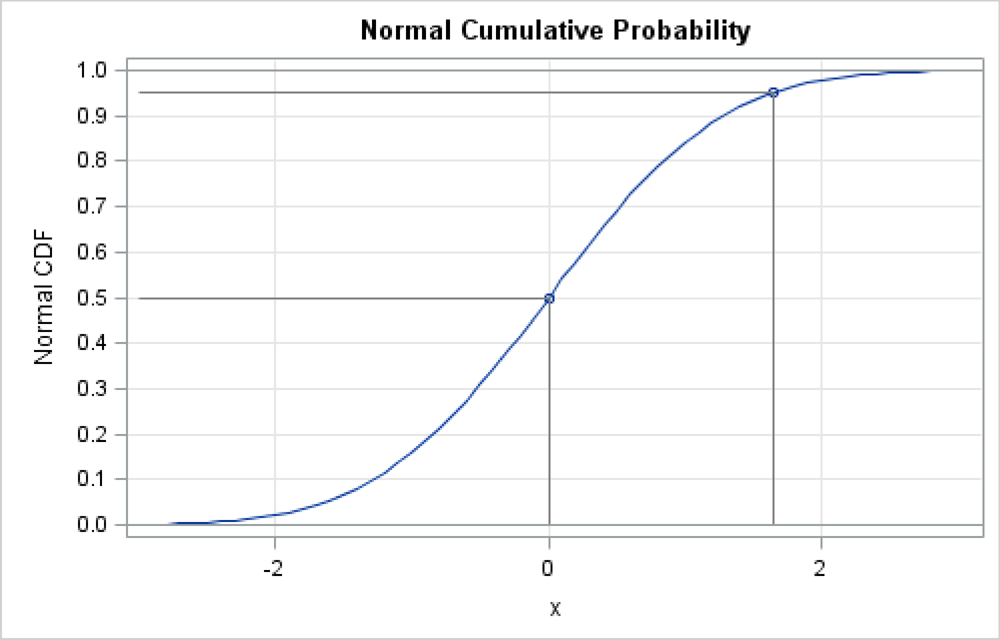
Same outcome or worse, with the same marginals (row sums).

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?

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.

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^{st} moment = mean
    - 2^{nd} moment = variance
    - 3^{rd} moment = skewness
    - 4^{th} 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 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:

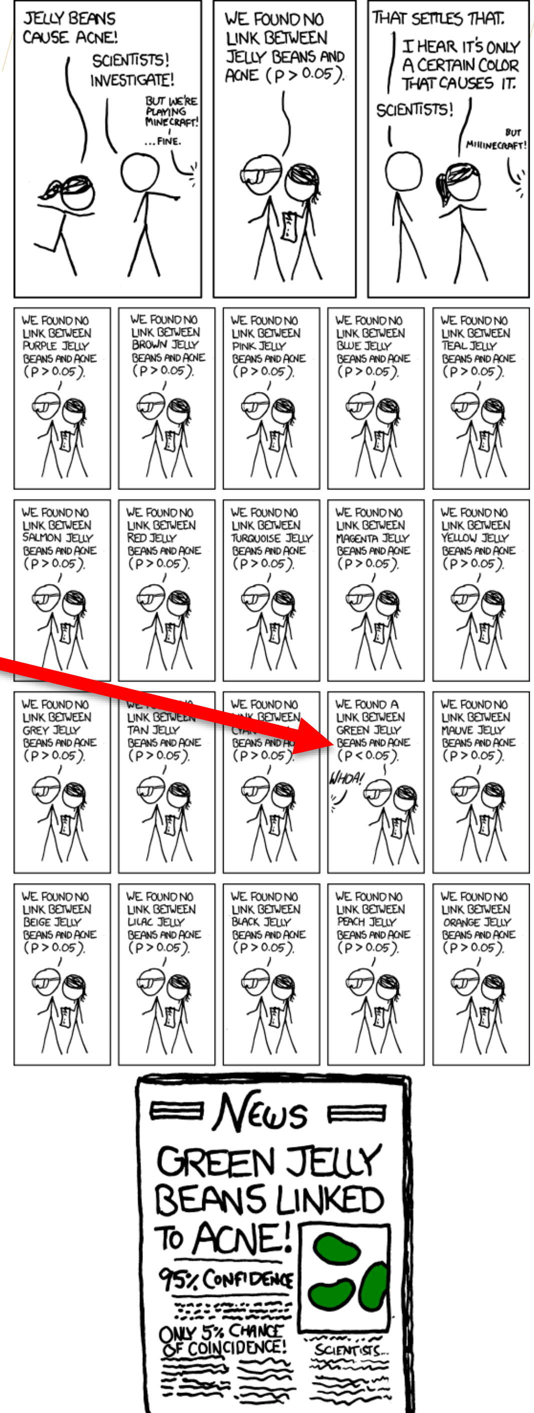
$$statistic \sim \frac{between\ group\ variability}{within\ group\ variability}$$

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.

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

- 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



JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO
BEANS AND A
(P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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.

$\bar{X}$=sample mean.

$$\bar{X} \sim N(mean, \frac{st.dev}{\sqrt{n}})$$

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

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

Central Limit Theorem

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

- 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

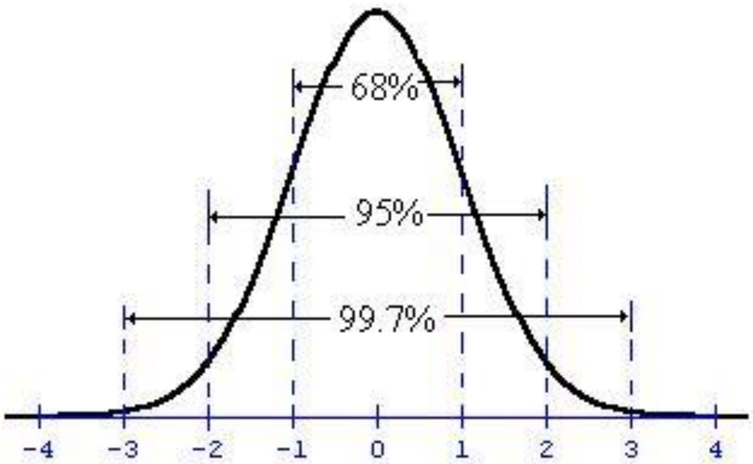
- Confidence intervals are a way to express uncertainty in *population* parameters, as estimated by the sample.
- E.g. If we create a 95% confidence interval for the population mean, say $\hat{\mu} = \bar{X} = 10 \pm 5$
  - Then we say that the true population mean, μ, has a 95% chance of being between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

It is **<u>not</u>** correct to say:

- ~~"95% of the sample values are in this range."~~
- ~~"There is a 95% chance that the mean of another sample will be in this range."~~

Confidence Intervals

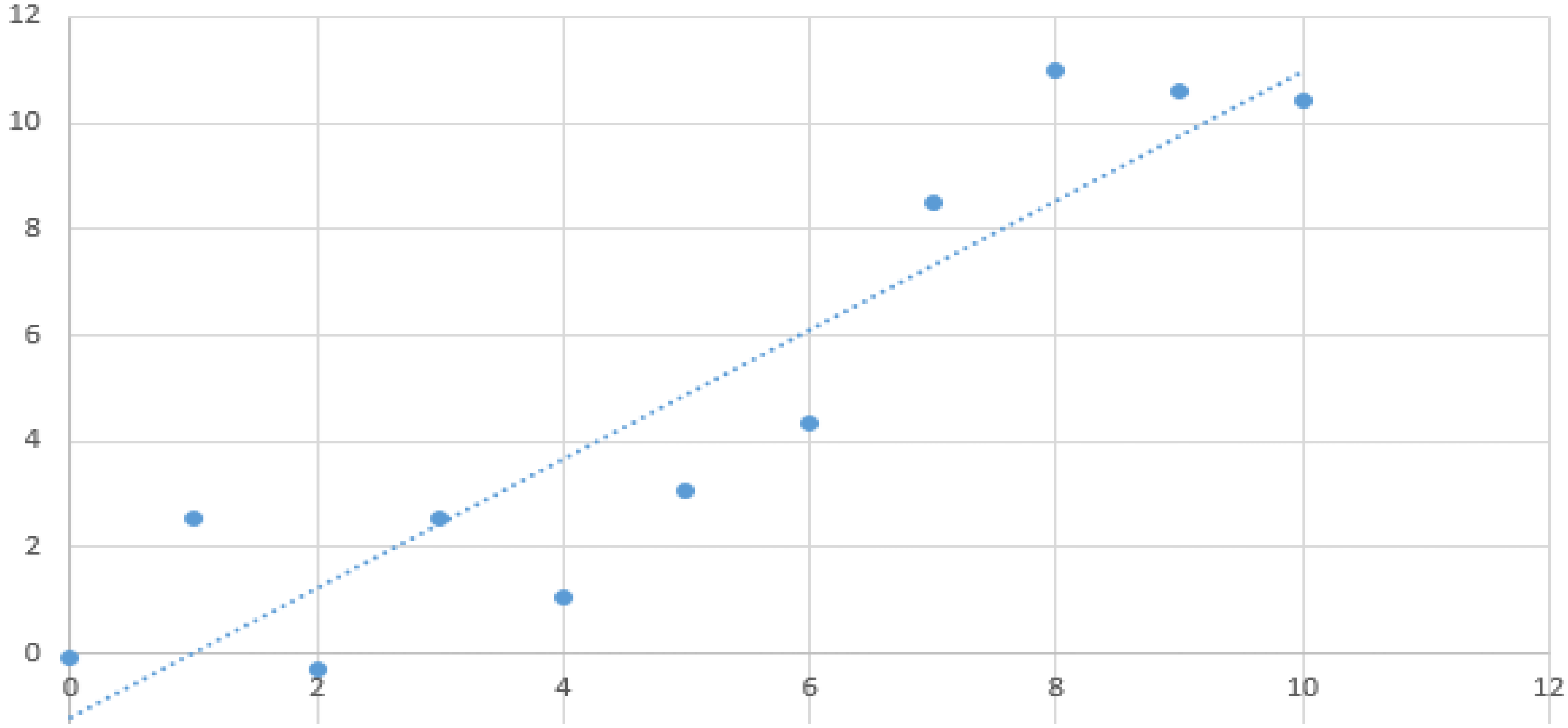
- To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.
  - Repeatedly sample from the population.
  - Calculate the mean for each sample.
  - Use the average of the sample means as the population estimate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



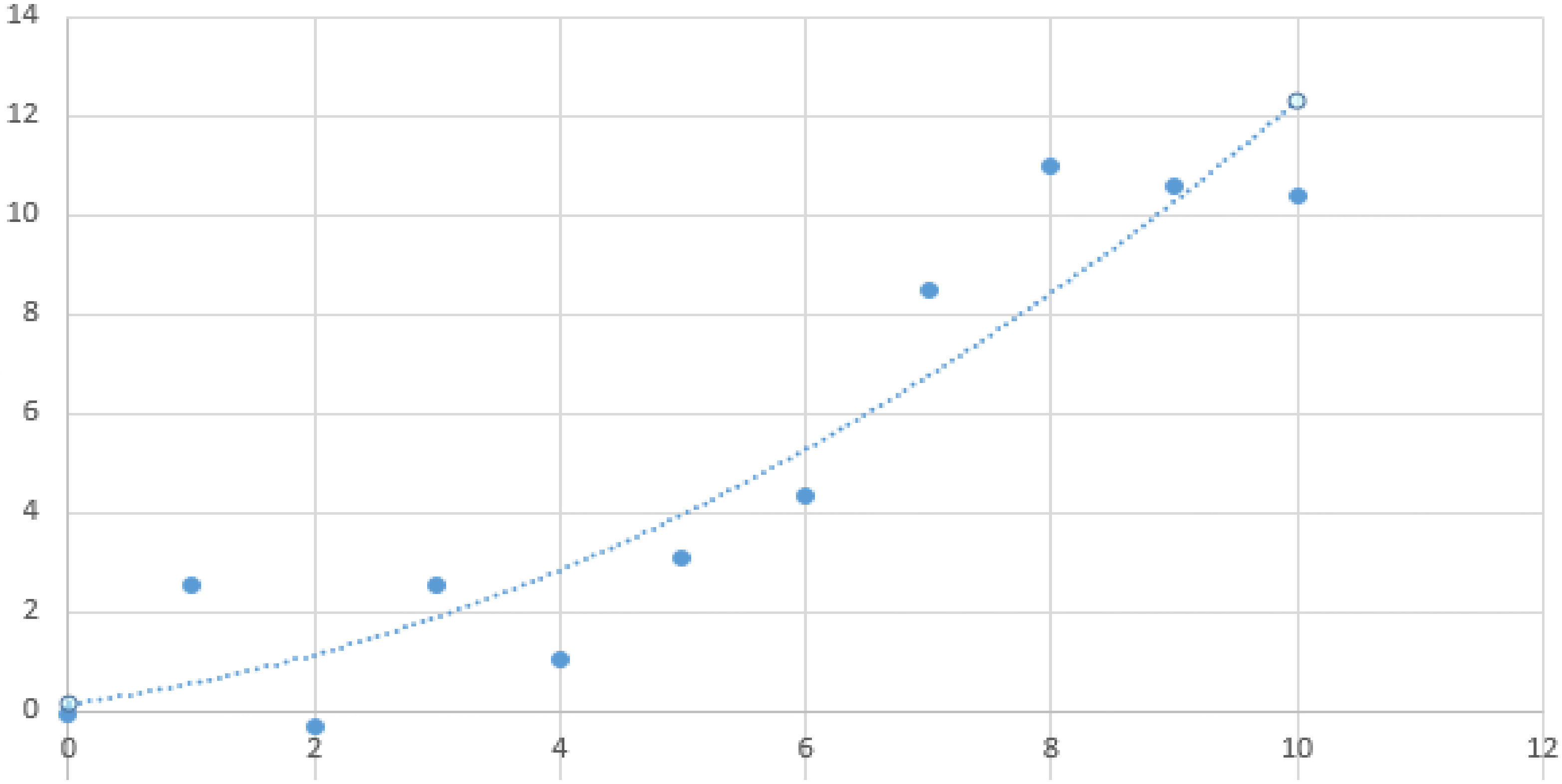
Regression Models

- 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.

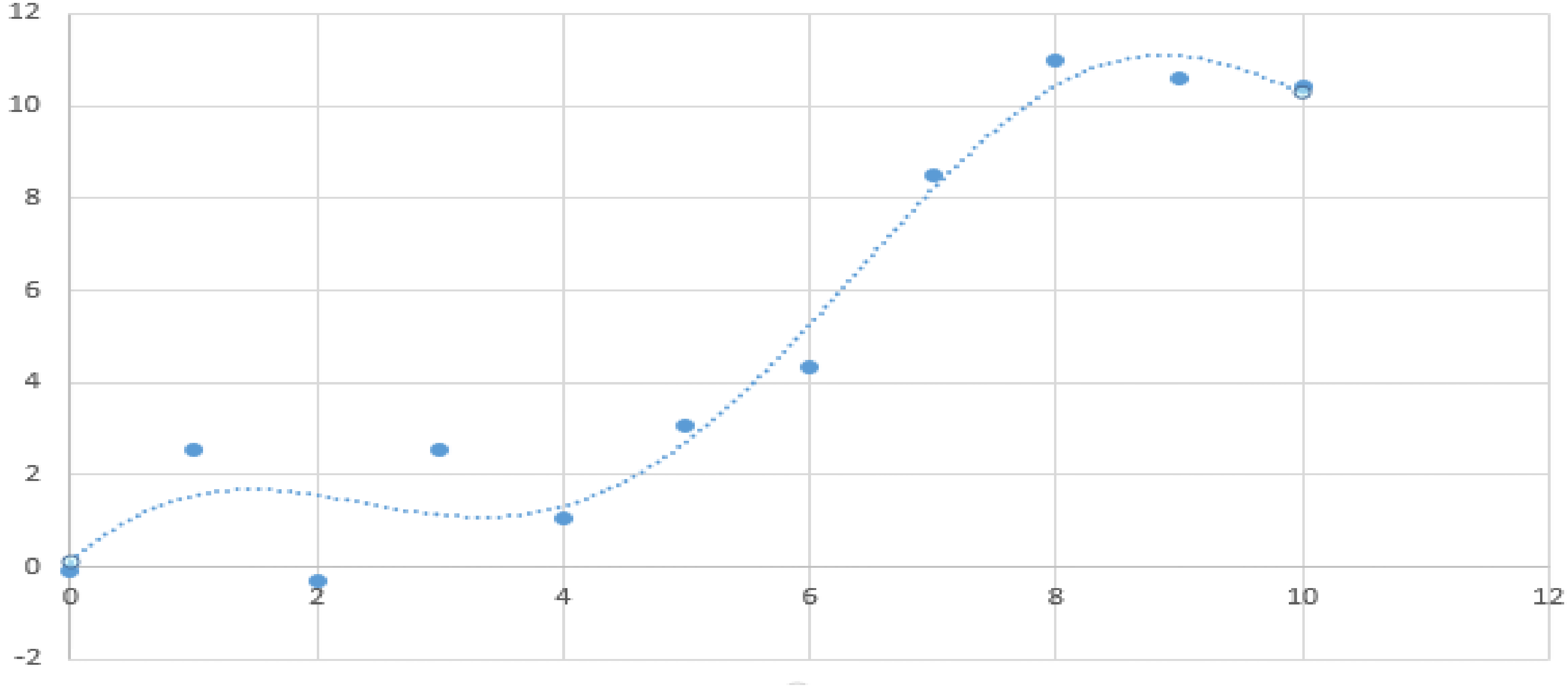
Regression: Linear



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.

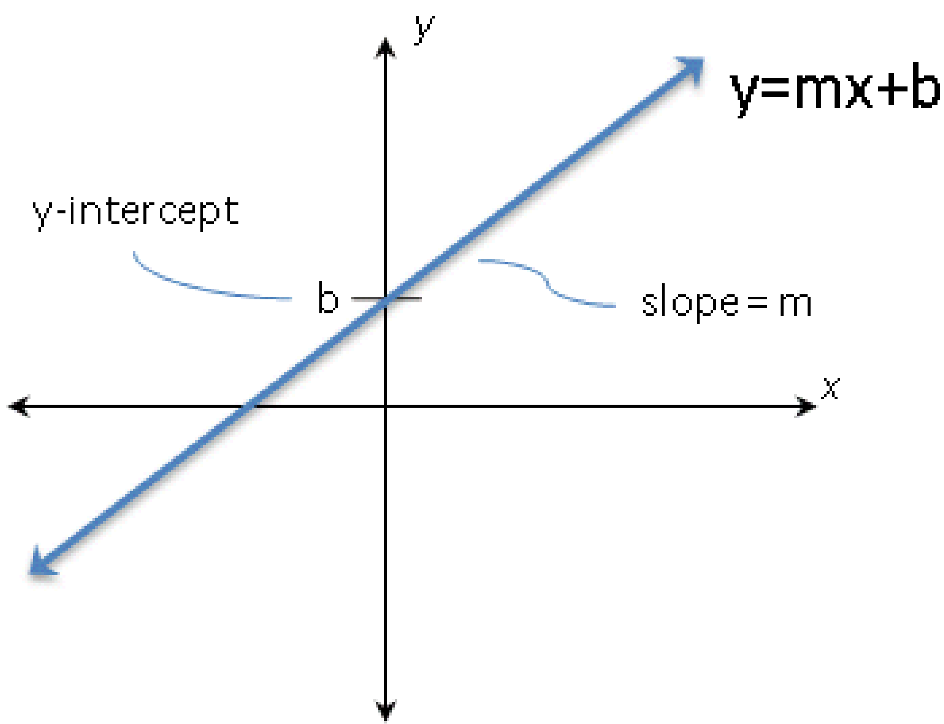
- We don’t know if these models will generalize

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

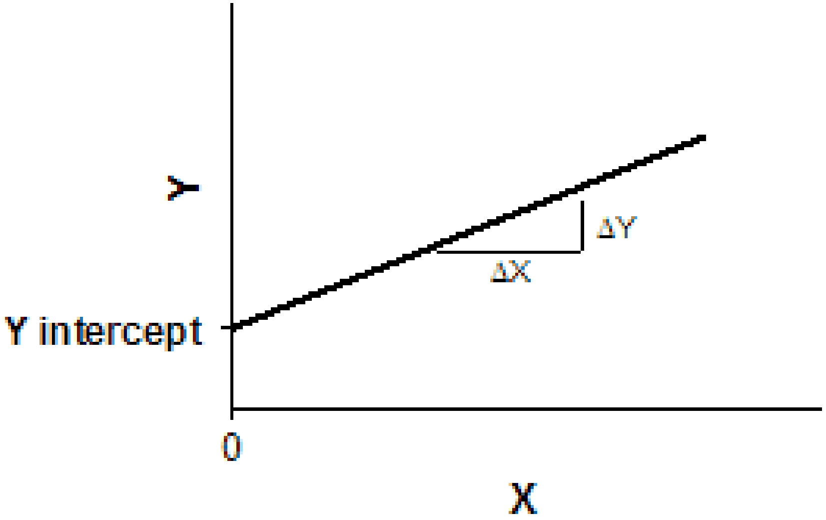
Linear Regression

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



Linear Regression

- Interpret slope: $m = \frac{rise}{run} = \frac{\Delta y}{\Delta x}$
  - If x changes by Δx, then y must change by Δy in order for the slope to stay the same (and it must).



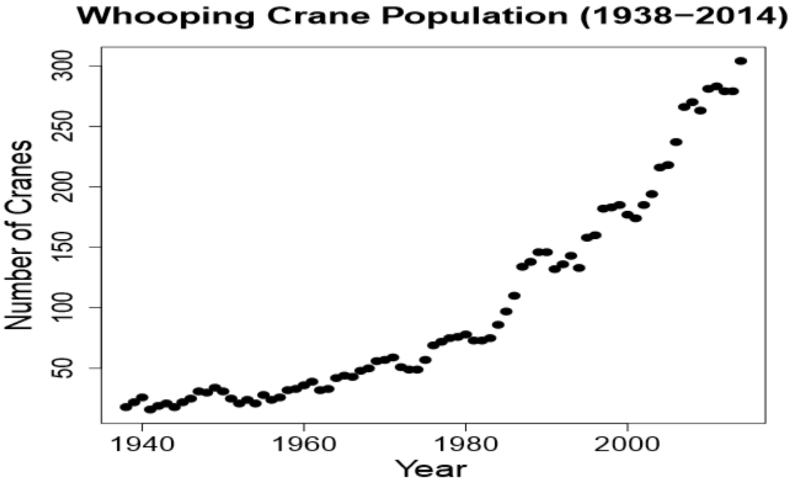
Given two points, (x1,y1), (x2,y2)

$$m = \frac{(y2 - y1)}{(x2 - x1)} = \frac{(y1 - y2)}{(x1 - x2)}$$

R example!

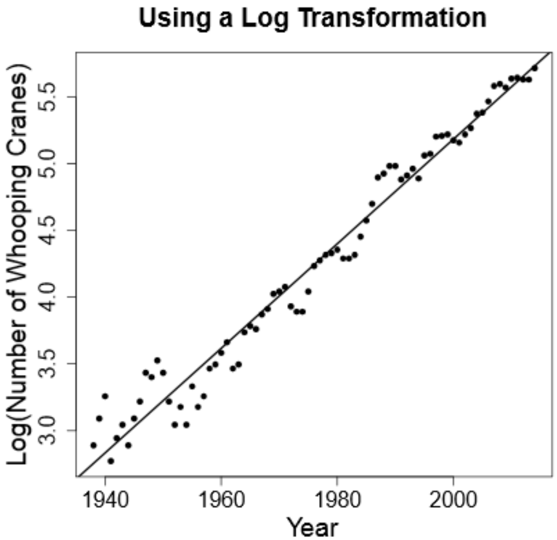
Linear Regression

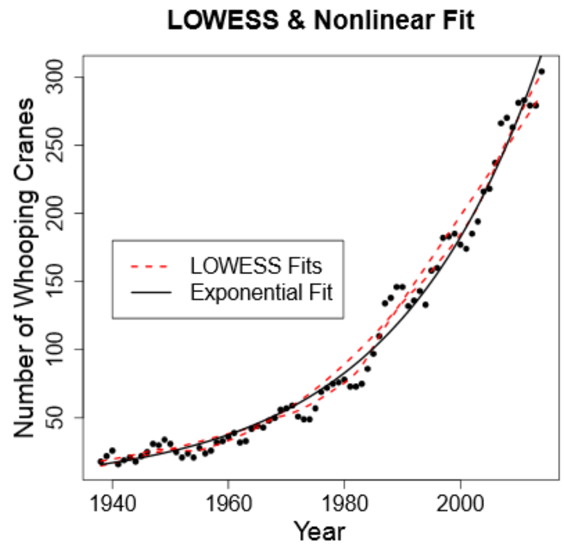
> Consider the relationship between Whooping Crane population and year below.



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

Linear Regression





> How would we decide on a ‘best’ model?

Linear Regression

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

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

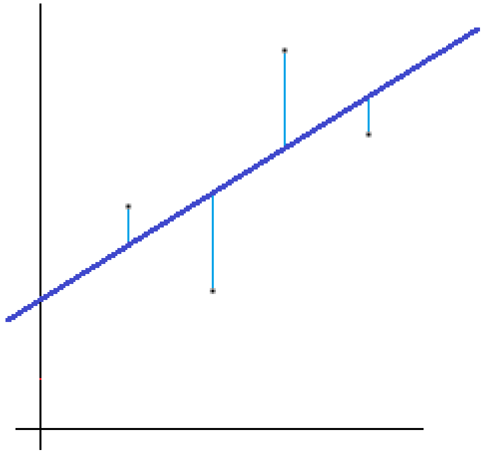
- Explicit solutions exist (using calculus).
- Computers are really good at finding minimums of equations. We let them do this for us.

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.

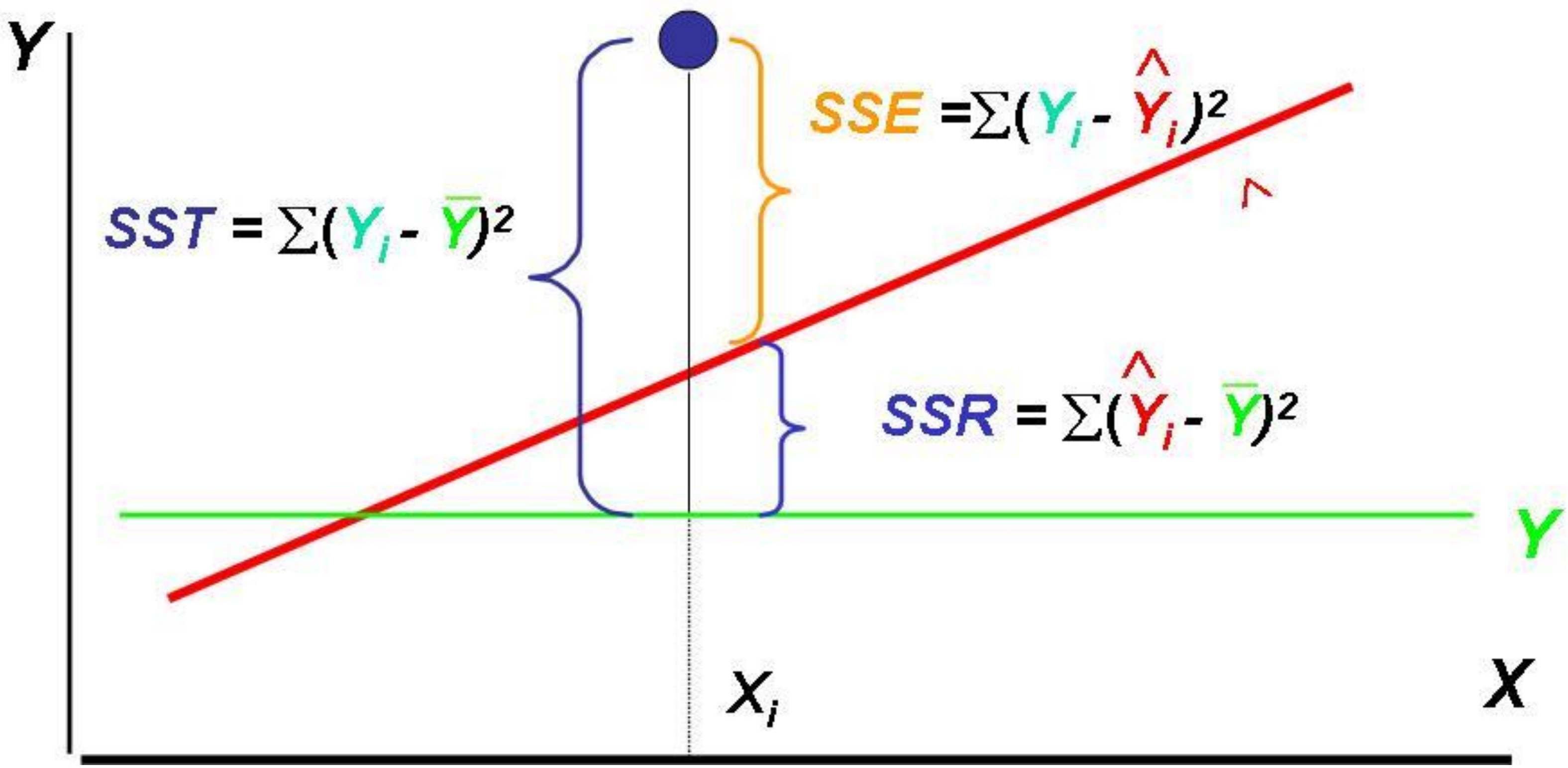


Least Squares Linear Regression

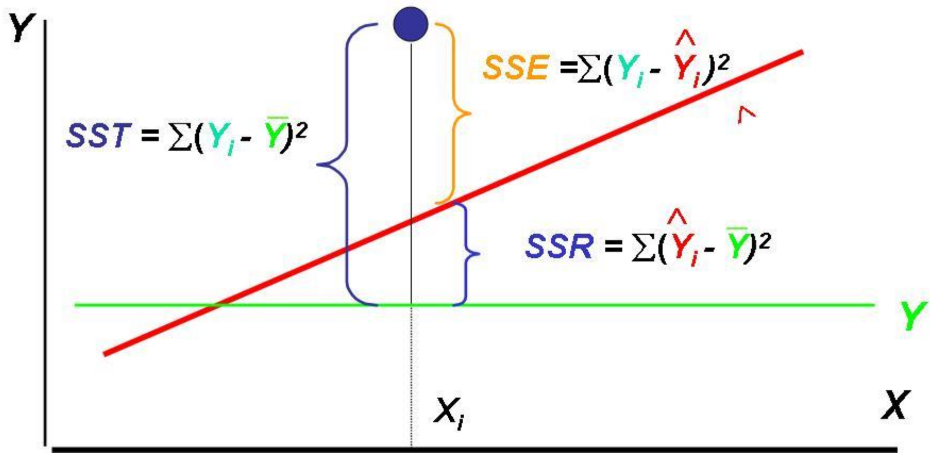
> We use the sum of the squared error between the predicted and actual.
> R demo.

Linear Regression

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



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:

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

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
      - Δ Num. Hospitalizations vs. Δ 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.