Data 8, Final Review

Review schedule:

- Day 1: Confidence Intervals, Center and Spread (CLT, Variability of Sample Mean)
- Day 2: **Regression**, Regression Inference**,** Classification

Your friendly reviewers today:
Hari Subbaraj, Rohan Narain, Ryan Roggenkemper,
Howard Ki, Claire Zhang

Regression

Why estimate the average of a population at all?

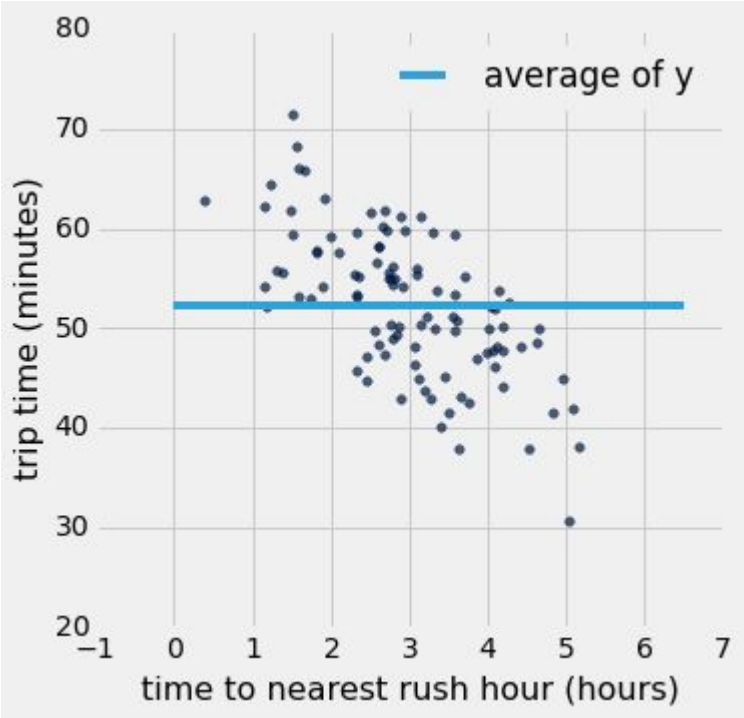
One reason: Sometimes most things are near the average.

Example: Time to drive from Berkeley to Mountain View.

Reasonable guess: Overall average driving time.

But what if we know it’s close to rush hour? Overall average looks like this:

Regression

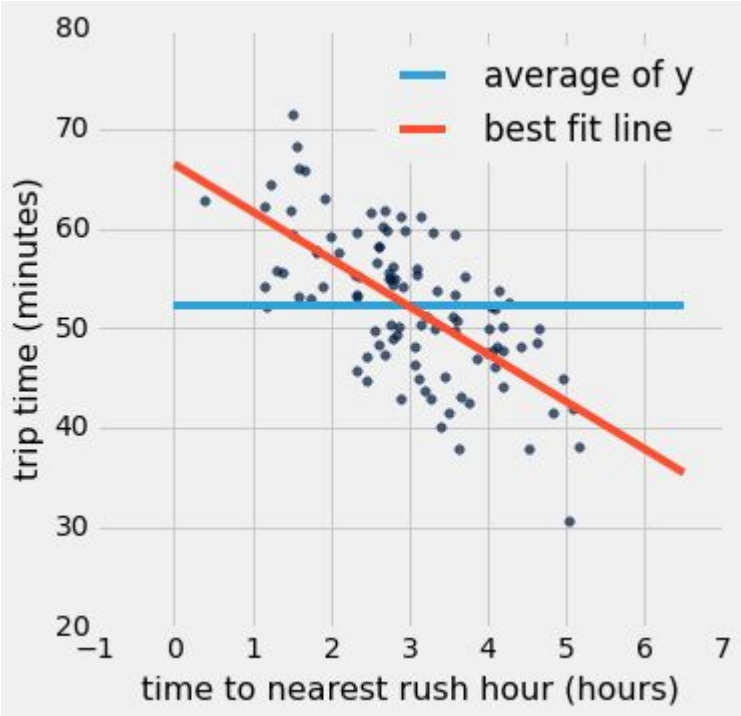

average of y
trip time (minutes)
time to nearest rush hour (hours)
80
70
60
50
40
30
20
−1
0
1
2
3
4
5
6
7

Regression

Better idea: Predict that it will take the average driving time *for trips around this close to rush hour.*

Linear regression assumes the averages are on a line:

average trip time = a*(proximity to rush hour) + b



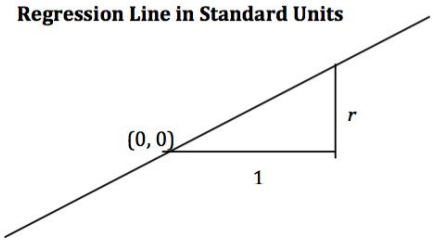
Steps to the regression line

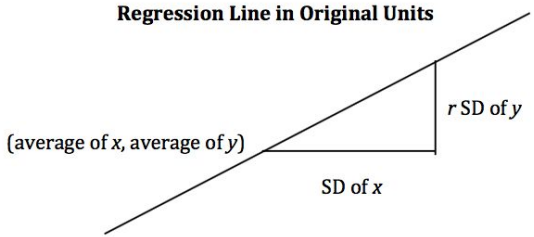
In this class, we first taught correlation and standard units

We used correlation and realized that the line of best fit for two variables in standard units is the line with slope=correlation and intercept=0

Computing Slopes and Intercepts with Correlation

- (estimate of y in standard units) = r * (x in standard units)
- (estimate of y − average of y) / SD of y = r * (x − average of x) / SD of x





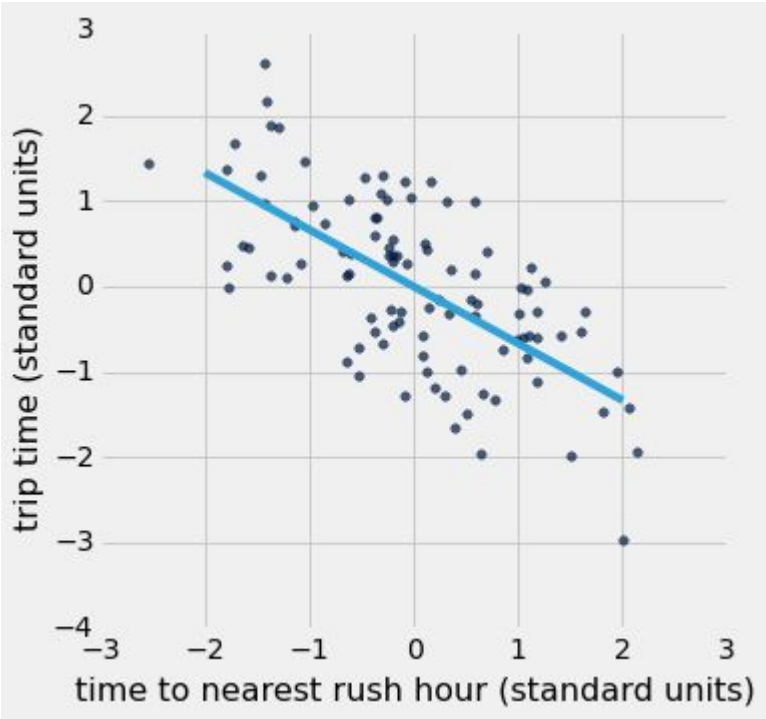
If the regression line is, indeed, a line,

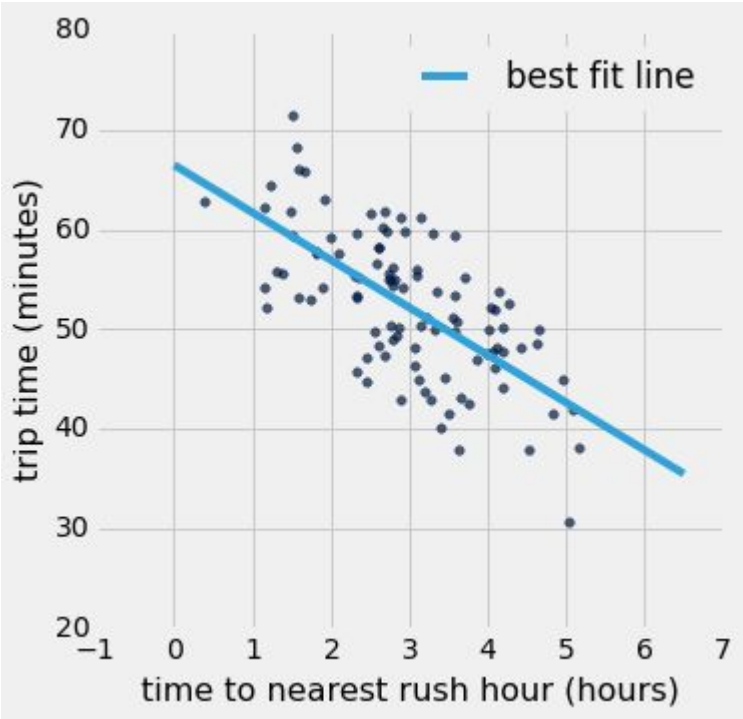
Then its form: y = slope*x + intercept.

Slope: r*SD(y)/SD(x)

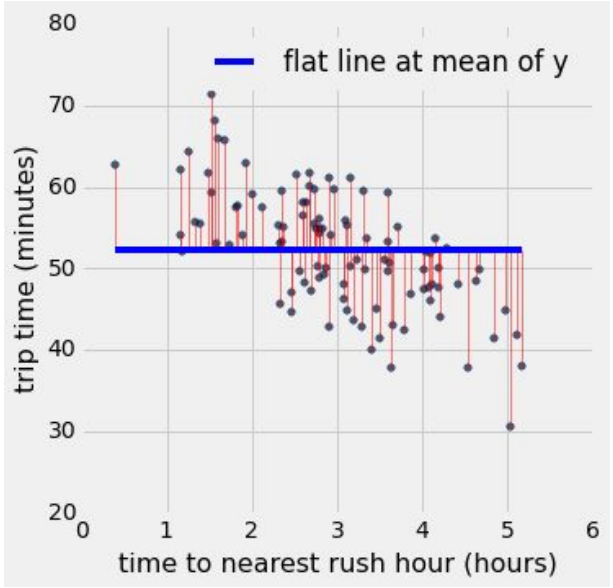
Intercept: mean(y) - slope*mean(x)

Computing Slopes and Intercepts with Correlation

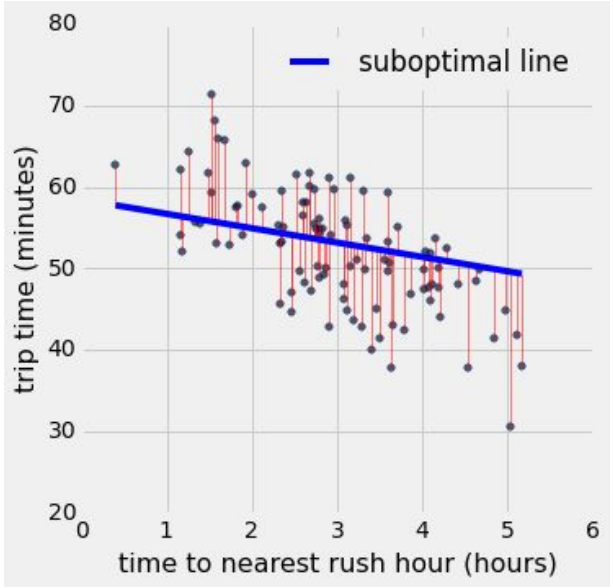




Computing Slopes and Intercepts via Optimization



Average of squared errors: 53.37



Average of squared errors: 39.99

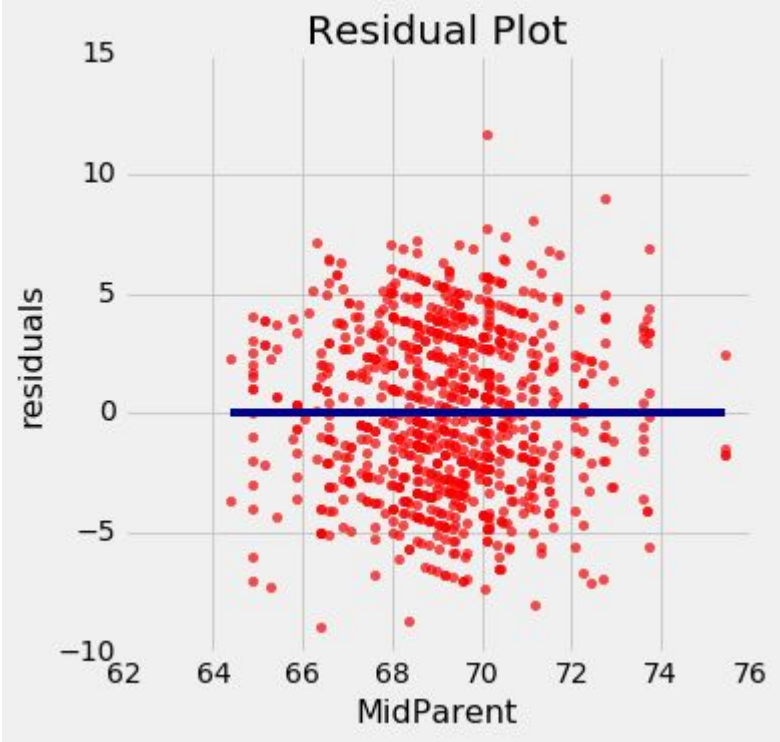
Average of squared errors: 29.68

Residual Plots

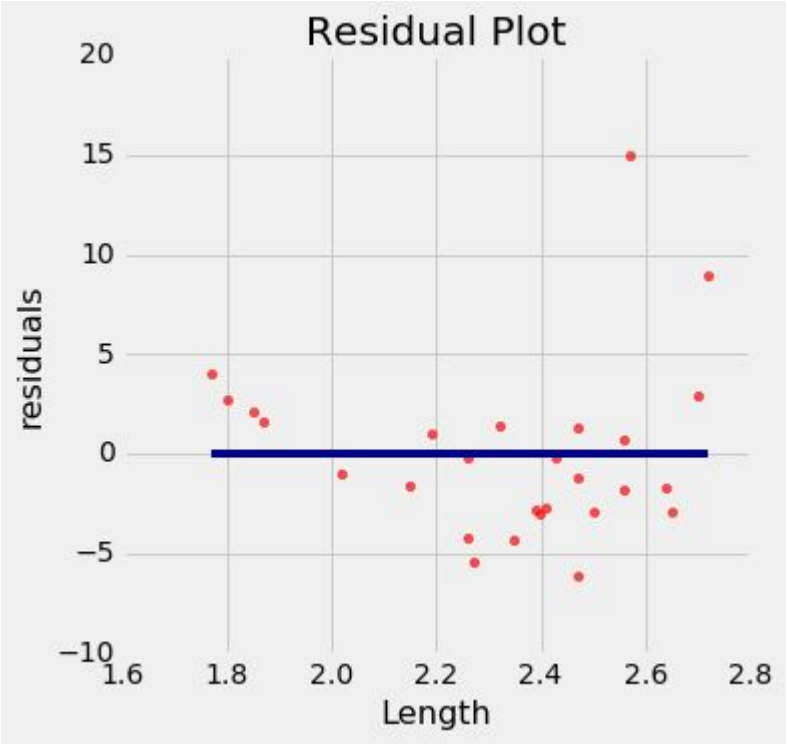
Residuals are the observed value - the predicted estimate by regression

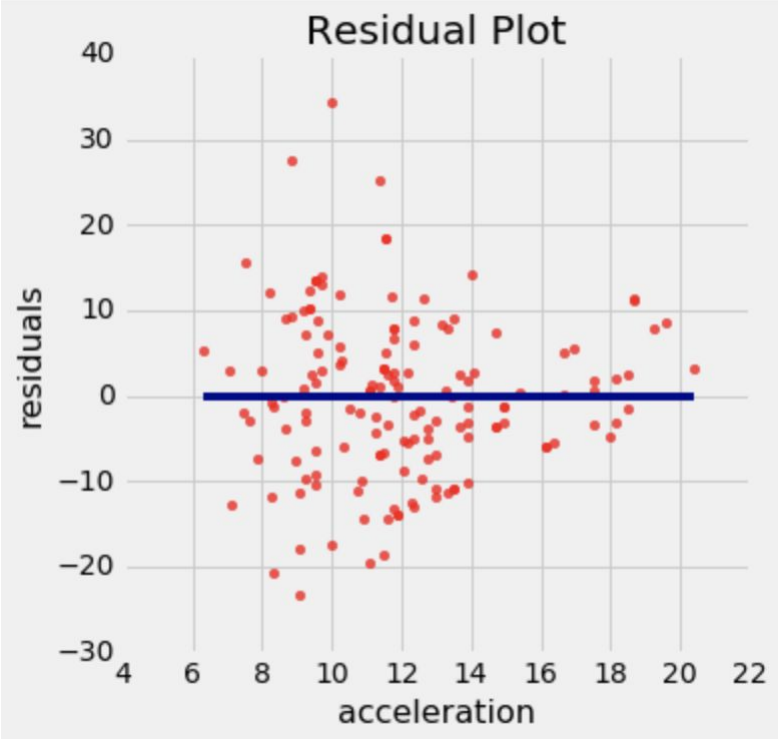
Looking at residual plots; the x value versus the residual, help us decide whether regression was a good fit for our graph

The residual plot of a good regression shows **no pattern**, so the residuals look about the same above and below a horizontal line at 0



Is Regression a good idea?





Variance of Residuals

$$\frac{\text{Variance of residuals}}{\text{Variance of } y} = 1 - r^2$$

$$\text{SD of residuals} = \sqrt{(1 - r^2)}\ \text{SD of } y$$

Standard Deviation of Residuals

- Two important formulas

$$\frac{\text{SD of predicted values}}{\text{SD of y}} = |r|$$

$$\text{SD of residuals} = \sqrt{1 - r^2} * \text{SD of y}$$

Data 8, Final Review

Review schedule:

- Day 1: Confidence Intervals, Center and Spread (CLT, Variability of Sample Mean)
- Day 2: Regression, **Regression Inference,** Classification

Your friendly reviewers today:
Hari Subbaraj, Rohan Narain, Ryan Roggenkemper,
Howard Ki, Claire Zhang

Regression Inference

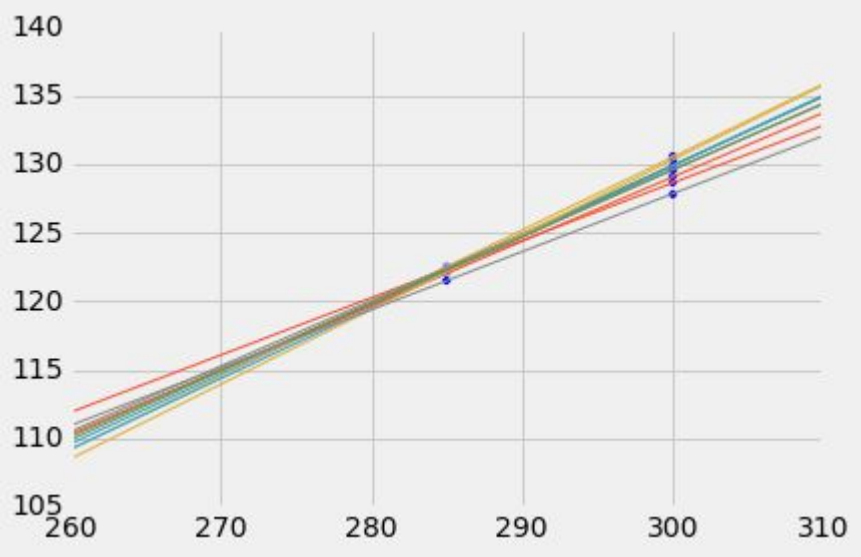
Compute a confidence interval over your regression line.

Recall: the regression line

y (sd units) = r * x (sd units)

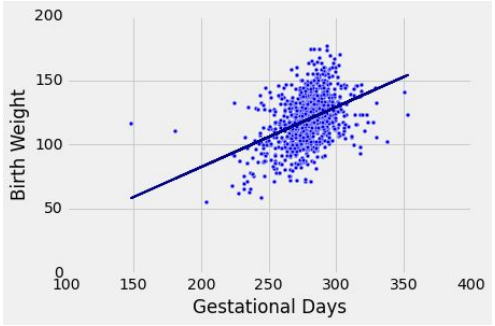
Idea: use bootstrap samples to randomly generate regression lines.

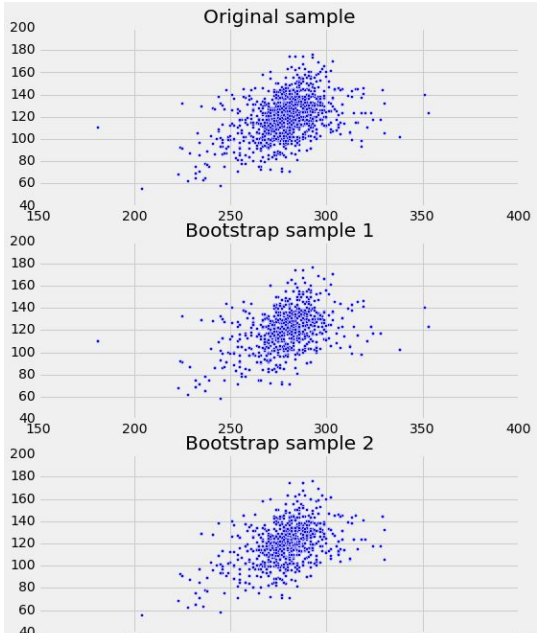
Why: Compute a 95% confidence interval over your predictions.

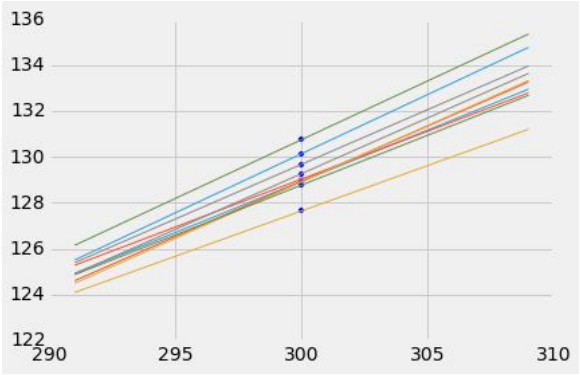


Regression Inference

Compute a confidence interval over your regression line.







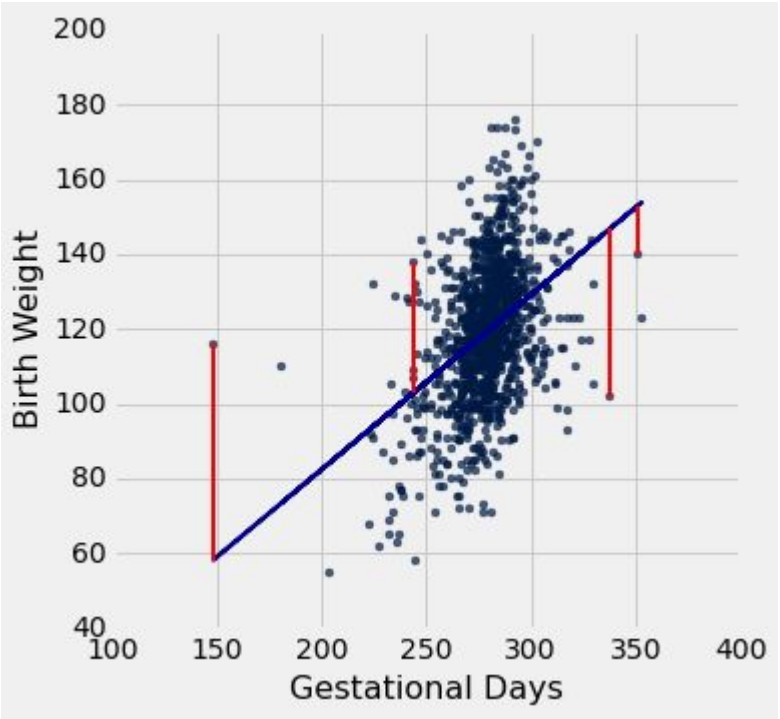
Regression Inference

The generative model:

y = (mx + b) + error (ie residual)

Data is generated according to **some true line,** with errors drawn at random with replacement from a normal distribution with mean zero and fixed variance.

Use **residual plot** for diagnostics.



Regression Inference

Questions

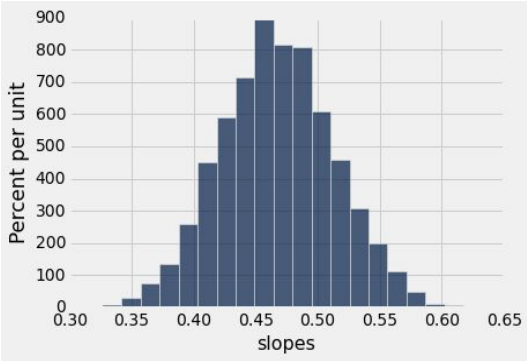
1. When can you determine whether the slope of the true line is zero?
2. How can you determine whether variables are truly linearly related?

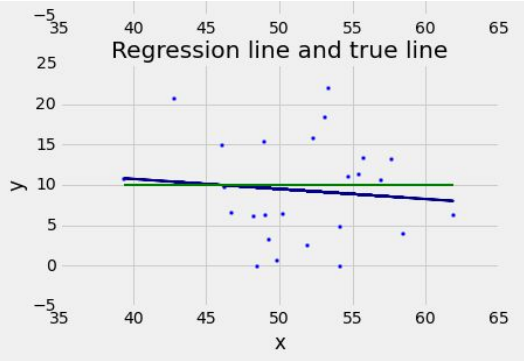
Regression Inference

Questions

1. When can you determine whether the slope of the true line is zero?

Use the bootstrap to compute a 95% confidence interval over your slope; then *check whether it contains zero*. Null is that the slope is 0, while alternative is that the slope is not 0



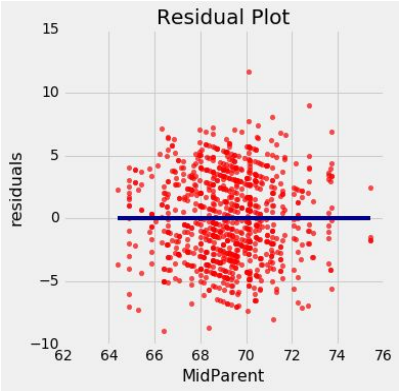


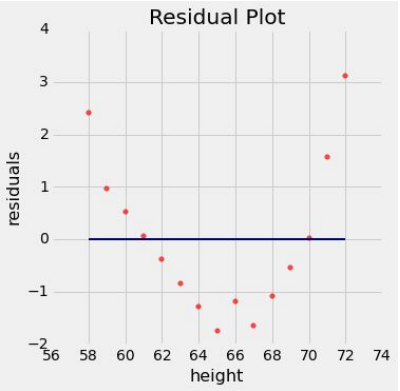
Regression Inference

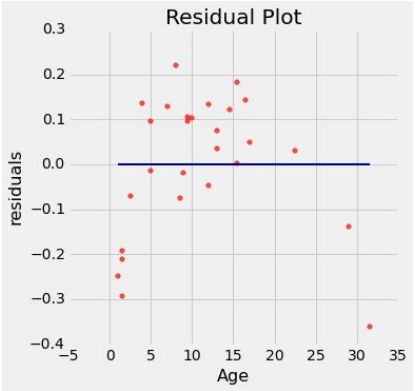
Questions

2. How can you determine whether variables are truly linearly related?

Draw a residual plot and check that the errors are appear to be drawn randomly from a normally distributed population.







Data 8, Final Review

Review schedule:

- Day 1: Confidence Intervals, Center and Spread (CLT, Variability of Sample Mean)
- Day 2: Regression, Regression Inference, **Classification**

Your friendly reviewers today:
Hari Subbaraj, Rohan Narain, Ryan Roggenkemper,
Howard Ki, Claire Zhang

Classification

Observations: Single instances or situations that we would like to be able to classify.

- Observations have **attributes**, or different features about them
- Attribute/feature selection is crucial to classification

Classification (contd.)

Examples:

- Genre of movie (Action/Romance)
- Presence of Breast Cancer
- Student standing (lower division/upper division)

Classification (contd.)

Why do we use a training set and a test set? How should they be picked?

In order to test our classifier, and testing on the training set might give a false impression. They need to be picked randomly!

Why is it bad not to partition your data into a test set and a training set?

You'll overfit! With a one-nearest neighbor classifier, you'll have 100% accuracy, which is not a good model for the real world.

Should the test set be used to tune your model?
No! The test set should be used to see how well your model does in real world applications and should not be used during training

K-nearest neighbors classifier

1. Get some features/attributes which you think would be useful to classify something (often times the hardest part)
2. Gather some data which you know the values of the features as well as the true classification/label of those data points
3. When you encounter a new data point which you don't know, calculate the 'closest' k neighbors and from those k data points, take the majority.

How to calculate the 'closest': Distance formula is often useful

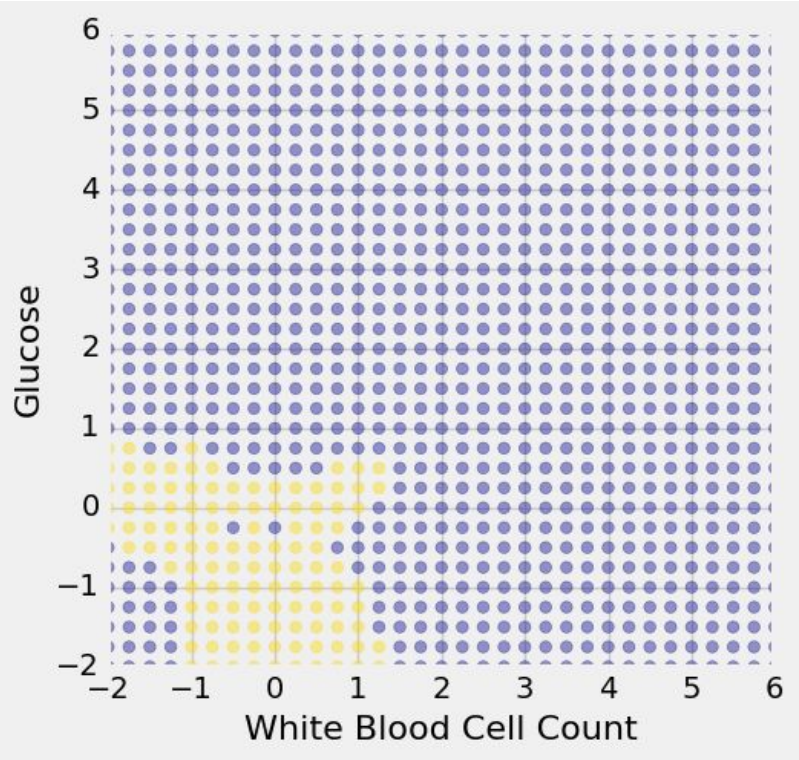
Distance Formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Mrs. E Teaches Math

K-nearest neighbors classifier example

1. Choose some features that you think would distinguish between someone who has chronic kidney disease and not
2. Find a relevant data set
3. When you encounter a new person's attributes which you don't know, calculate the 'closest' k neighbors and from those k data points, take the majority and classify as either with CKD or without



Machine Learning

- Supervised
  - Use labeled data to make a prediction about an unlabeled example
- Unsupervised
  - Look at unlabeled data to recognize underlying patterns