Lecture 23 – Residuals & Regression Inference
Announcements

- No lab this week

- **Homework 9 - Regression Inference**
  - Due Monday 04/25

- Course Evaluations:
  - 

- **Project 3:**
  - Due Monday 05/02
Rubrics

- **Rubric 1:**
  - Projects (not final project): 45%
  - Homeworks: 25%
  - Participation: 5%
  - Project 3 required

- **Rubric 2:**
  - Projects (not final project): 30%
  - Homeworks: 35%
  - Participation 10%
  - Project 3 optional

*We will compute scores for both Rubrics and then use whichever is best for each student*
Linear Regression
Finding the best-fit line

- Compute correlation coefficient \( (r) \)
  - Prediction in standard units

- Find slope and intercept of the data
  - Prediction in original units
  - slope = \( r \times \frac{sd(y)}{sd(x)} \)
  - intercept = \( \text{mean}(y) - \text{slope} \times \text{mean}(x) \)

- Numerical Optimization:
  - Use a compute to find slope and intercept to minimize \( y \)
  - \( y = \text{slope} \times x + \text{intercept} \)
Residuals

- Error in regression estimate

- One residual corresponding to each point \((x, y)\)

- residual
  \[= \text{observed } y - \text{regression estimate of } y\]
  \[= \text{observed } y - \text{height of regression line at } x\]
  \[= \text{vertical distance between the point and line}\]
Regression Diagnostics
Residual Plot

A scatter diagram of residuals

- For linear relations, plotted residuals should look like an unassociated blob
- For non-linear relations, the plot will show patterns
- Used to check whether linear regression is appropriate
- Look for curves, trends, changes in spread, outliers, or any other patterns
Properties of residuals

- The mean of residuals is always 0

- Variance is standard deviation squared

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

- \( \frac{\text{(Variance of fitted values)}}{\text{(Variance of y)}} = r^2 \)

- Variance of y = \( (\text{Variance of fitted values}) + (\text{Variance of residuals}) \)
We just said

- \( \frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2 \)
- variance is standard deviations squared,

So:

- \( \frac{\text{SD of fitted values}}{\text{SD of } y} = |r| \)
- \( \text{SD of fitted values} = |r| \times (\text{SD of } y) \)

- \( \frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2 \)
Variance of Fitted (Predicted) Values

- Variance = Square of the SD
  = Mean Square of the Deviations

- Variance has weird units, but good math properties

\[
\frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2
\]
By definition,

\[ y = \text{fitted values} + \text{residuals} \]

\[ \text{Var}(y) = \text{Var}(\text{fitted values}) + \text{Var}(\text{residuals}) \]
A Variance Decomposition

Var(y) = Var(fitted values) + Var(residuals)

\[
\frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2
\]

\[
\frac{\text{Variance of residuals}}{\text{Variance of } y} = 1 - r^2
\]
A Variance Decomposition

\[ \text{Var}(y) = \text{Var}(\text{fitted values}) + \text{Var}(\text{residuals}) \]

- \[ \frac{\text{SD of fitted values}}{\text{Variance of } y} = |r| \]
- \[ \frac{\text{SD of residuals}}{\text{Variance of } y} = \sqrt{1 - r^2} \]
Residual Average and SD

- The average of residuals is always 0

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

- SD of residuals = SD of y, not \( \sqrt{1 - r^2} \)
Question 1

**Midterm:** Average 70, SD 10

**Final:** Average 60, SD 15

\[ r = 0.6 \]

The SD of the residuals is _____.
Midterm: Average 70, SD 10
Final: Average 60, SD 15
\[ r = 0.6 \]

For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within ____________ points.
Regression Model
A “Model”: Signal + Noise

Distance drawn at random from the normal distribution with mean 0

Another distance drawn independently from the same distribution
What we get to see
Prediction Variability
If the data come from the regression model,
And if the sample is large, then:

- The regression line is close to the true line
- Given a new value of $x$, predict $y$ by finding the point on the regression line at that $x$
Confidence Interval for Prediction

- Bootstrap the scatter plot
- Get a prediction for $y$ using the regression line that goes through the resampled plot
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at $y$. 
Since $y$ is correlated with $x$, the predicted values of $y$ depend on the value of $x$.

The width of the prediction’s CI also depends on $x$.
  - Typically, intervals are wider for values of $x$ that are further away from the mean of $x$. 

Inference about the True Slope
Bootstrap the scatter plot.

Find the slope of the regression line through the bootstrapped plot.

Repeat.

Draw the empirical histogram of all the generated slopes.

Get the “middle 95%” interval.

That’s an approximate 95% confidence interval for the slope of the true line.
Test Whether There Really is a Slope

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it’s not.
- **Method:**
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn’t contain 0, the data are more consistent with the alternative.
  - If the interval does contain 0, the data are more consistent with the null.
minimize() works no matter what*!

Define a function that computes the prediction you want, then the error you want, for example:

- Nonlinear functions of $x$
- Multiple columns of the table for $x$
- Other kinds of error instead of RMSE

Nonlinear functions can get complicated, fast!
Attributes (features) of an example

Classifier

Model association between attributes and labels

Population

Sample

Labels

Training Set

Test Set

Estimate classifier’s accuracy
Attributes (features) of an example

NN Classifier: Use the label of the most similar training example

Predicted label of the example

Population

Sample

Labels

Training Set

Test Set

Copyright © 2016 Barnard College
Pythagoras’ Formula

\[ D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}. \]
Distance Between Two Points

- Two attributes $x$ and $y$:
  \[ D = \sqrt{ ((x_0 - x_1)^2 + (y_0 - y_1)^2 ) } \]

- Three attributes $x$, $y$, and $z$:
  \[ D = \sqrt{ ((x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2 ) } \]