5/5/2022

Regression and Correlation (Ch 12)
Hypothesis Testing
Confidence intervals
Prediction Intervals
Outliers

https://phet.colorado.edu/en/simulations/least-squares-regression
https://phet.colorado.edu/en/simulations/curve-fitting

Hypothesis Testing:
For a regression line (simple linear regression): The null hypothesis is either that the slope of the regression line is 0, or that the correlation is 0.

The regression line from the data: $y = b_0 + b_1 x + \varepsilon$
The population regression line: $y = \beta_0 + \beta_1 x$

$$H_0: \beta_1 = 0$$

Or

$$H_0: \rho = 0$$

The alternative is that it's not zero. (You can do inequalities, but uncommon in regression lines).

The data Analysis tool pack output tests this in two ways: does an ANOVA test for the entire model (which depends in our case on just the slope and correlation). And, there is test on each coefficient below that with their own p-values and test statistics.

The confidence intervals are related to the values of the coefficients, specifically to the slope coefficient. We will talk about confidence intervals on the slope coefficient.

The prediction intervals are intervals on our predicted y-values. If we were to measure an x-value and then want to predict not just the mean value, but a range of values with a certain confidence level, how do we calculate that?

Outliers (partly through calculation, and partly through residual graphs).
For regression, the outliers are considered based on their distance from the regression line.
Influential points and how they might differ from outliers.

Go to excel for examples.

residual plot is a plot of the x-values vs. the residuals (the difference between the predictions of the regression line and the actual observations). We want residuals to appear to be random. They should be evenly spread out and have no patterns.

outliers are far from the line in the scatterplot, and show up as very large residuals (far from 0) compared to the other residuals

errors are supposed to be normally distributed with fixed standard deviation.

The prediction intervals use the model standard error and a t-value for the confidence level with the mean as the predicted value from the regression equation.

Typical outlier standard is 2 standard deviations. Anything outside 2 standard deviations is considered "unusual". Extreme outliers are outside 3 standard deviations.

Another way is to make a boxplot. Excel will mark outliers in the boxplot automatically, and so we can flag them that way.

This is the end of the new material.

Review for Final
The exam is comprehensive: but material from Chapters 1-8 will be very similar to questions on the first two exams.
The rest of the exam will cover hypothesis testing (1 and 2-sample problems, chapter 9 and 10), ANOVA and Tests of Independence (from ch 11.3, ch 13.1), and regression (ch 12).
The quizzes from Quiz 9-11 are the best models for questions from these last chapters.