Introduction to Statistics (MAT/SST 115.03 2008S)

Class 40: Topic 29: Inference for Correlation and Regression

Held: Wednesday, 7 May 2008

Summary: We conclude our exploration of statistics by examining what regression lines between two variables from a sample of a population tell us about the relationships between the variables within the populations from which the sample was drawn.

Notes:

- I still hope to distribute tentative grades on Friday, but, given the number of administrative tasks that have been thrown at me this week, I'm less hopeful than I was.
- The final remains Thursday, May 15, 2-5 p.m.
- Katherine will hold a review session on Wednesday evening, May 14.
- I should be around on the morning of May 15 to answer questions. I can also hold an earlier review session.
- For the final, you may bring your own cheat sheet: One hand-written, double-sided, 8.5x11 inch (or A4) sheet of notes.
- We will stop at 29-3. (You are not responsible for 29-4 or 29-5.)
- And yes, you can ask questions before the exam-lette.
- I will be leaving early again today to pick up my youngest after school.
- *There is no homework for Friday! However, I do expect you to show up.*

- Handouts: Applet: Sampling Regression Lines; R Notes for Topic 29.
- Due: 28-5, 28-10, 28-24, 28-25.

Overview:

- Making inferences about populations (once again).
- Important R.

The Progression Continues: What Samples Tell us About Populations

- We've been studying the relationship between two quantitative variables taken from one sample.
- Of course, once we start to work with samples, we should ask ourselves whether or not they represent the population.
- What are our normal strategies?
- A test of significance
- A confidence interval

- Test of significance:
  - Choose a null hypothesis.
  - Choose an alternate hypothesis.
  - Using the two hypotheses and the sample, generate a test statistic.
  - Look up the test statistic to get a *p*-value.
- For regression models, we do all of this in terms of the slope of the regression line.
- The null hypothesis: There is no relationship between the variables (so the slope is 0).
- The alternatives: It's negative, it's positive, or it's just not zero.
- The test statistic: $b/\text{SE}(b)$
- The table: Use the *t*-table with $n-2$ degrees of freedom. (Why $n-2$? It seems to work well.)
- What should we use as the standard error? We let the computer figure it out for us.
  - (Another thing in which almost no one wants to see the formula)

- Confidence interval
  - $b +/- t^* \text{SE}(b)$
- How do we compute the critical value? We look it up in the *t*-table.
- How do we compute the standard error? We rely on our statistical software.

Getting Information from R

- So, how do we get R to tell us the standard error (and other things)?
- We continue to use `lm`.
- We filter the results of `lm` through `summary`.
- For example,

```
> TBP = read.csv("/home/rebelsky/Stats115/Data/TextbookPrices.csv")
> summary(lm(TBP$Price ~ TBP$Pages))

Call:
lm(formula = TBP$Price ~ TBP$Pages)

Residuals:
     Min       1Q   Median       3Q      Max
-65.4748 -12.3241  -0.5838  15.3037  72.9909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.42231   10.46374  -0.327    0.746
TBP$Pages    0.14733    0.01925   7.653 2.45e-08 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 29.76 on 28 degrees of freedom
Multiple R-Squared: 0.6766,     Adjusted R-squared: 0.665
F-statistic: 58.57 on 1 and 28 DF,  p-value: 2.452e-08
```
