Exercises 4 · Sampling distributions

Due Monday, February 22, 2016

(1) Sampling variability and regression modeling

In this problem, you will use Monte Carlo simulation to build your intuition about the effect of sampling variability on estimates of parameters in statistical models. To do this, you'll need the files `simdata_samp.csv` and `simdata_pop.csv`.

(A) First look at the data in "simdatasamp.csv." This is a sample of size 50 from a much larger population (a situation that arises often in statistics). Fit a regression model for y versus x to this data set.

(B) You may have guessed already that the sample from Part A is, in fact, a random sample from the 10,000 observations in "simdatapop.csv" (which we imagine to be the whole population). The question at issue here is: *how much can you trust the estimates of the model parameters arising from the sample in Part A?*

In statistics, we often equate the trustworthiness of an estimate with the degree to which that estimate might change under different hypothetical random samples. If we'd taken a different sample of 50 individuals from the population, and gotten drastically different estimates of the model parameters, then our original estimate isn't very trustworthy. If, on the other hand, pretty much any sample of 50 individuals would have led to the same estimates, then our answers for *this particular* subset of 50 are likely to be accurate.

On real problems, we can't look at the whole population. But because we're using simulated data on this problem, you can. That means you can actually investigate what kinds of answers other samples might have given you.

Complete the "Gone fishing" walk-through on the course website: `http://jgscott.github.io/teaching/r/gonefishing/gonefishing.html`. Apply the techniques you learn in this walkthrough to set up a Monte Carlo simulation that approximates the sampling distribution of the least-squares estimator you calculated in Part A (i.e. using a sample of size 50 from the wider population). You should use more Monte Carlo samples for this simulation: at least 1000.

For this problem, turn in the following items:

(i.) A brief summary of your understanding of the relationship you fit to the sample in Part A, including the coefficients, residual standard deviation, and R^2.

(ii.) Histograms of the sampling distributions for the intercept and slope that you simulated in Part B.

(iii.) The R code you used to produce these histograms. You can print this out directly from RStudio, or copy and paste into a Word document. If you copy/paste into Word, make sure you show the code in a fixed-width font like Courier or Monaco.

(iv.) A paragraph that describes, in your own words, what the sampling distributions in Part B represent, and why they are useful for quantifying the uncertainty in the answer to Part A.

(2) Bootstrapping

Complete the "Creatinine, revisited" walkthrough on the class website.[1] This will introduce you to the idea of bootstrapping as a way to approximate a sampling distribution when you cannot simulate samples from the population (as you did in the previous question). Make sure you also read up through page 116 in Chapter 5 of the course packet.

Once you've done this:

(A) Return to the data set "ut2000.csv" on SAT scores from UT students across all 10 undergraduate colleges. Calculate an approximate 95% confidence interval for the difference in mean SAT math (SAT.Q) scores between students in the colleges of architecture and liberal arts. I can think of at least two ways you could accomplish this, so make sure you describe precisely what you did and why, and report the interval.

(B) Fit a regression model for graduating GPA in terms of SAT combined score (SAT.C) and College (with no interaction term), and provide a 95% confidence interval for the slope of the SAT score.

(C) In your own words, briefly describe the idea of bootstrapping (both what we do and why we do it).

[1] http://jgscott.github.io/teaching/r/creatinine/creatinine_bootstrap.html