# *Exercises 4 · Categorical predictors in regression · Sampling distributions*

## Due Monday, February 16, 2015

### *(1) Practice using categorical predictors in regression models*

In each of the following two parts (A and B below), I ask you to conduct a data analysis and write a short report. Here are some guidelines to follow, both here and in the future.

1. Imagine that you are writing for a statistically savvy audience who is not familiar with the specific problem at hand.
2. Start by summarizing the overarching problem, the specific question(s) to be answered, and your modeling strategy.[1] This whole summary need not take more than a single substantial paragraph.
3. Focus on the problem, your conclusions, and your evidence. Do not simply narrate your process used to arrive at your conclusions.
4. Include any figures and tables necessary to make your case. Label them and refer to them in the text where appropriate.
5. Write well. Use fewer words. Use shorter words. Use fewer sentences. Use shorter sentences.

(A) After completing the "Reaction time in video games" walkthrough on the class website, head over to the case study on quality control in the supply chain for circuit-board manufacturing: `http://jgscott.github.io/teaching/r/solder/solder.html`

Read about the problem and work your way through the introductory commands I've posted. Then address the question I pose at the bottom of the page:

> Build a model to predict solder skips using these three predictor variables [Opening, Solder, and Mask]. Include whatever combination of main effects and interaction terms you deem appropriate. Justify your final choice of model. By this, I mean that if you include a term in the model, explain why. Similarly, if you decide to leave a possible term out of the model, explain why.

Write a short report summarizing your conclusions. Do not exceed two typed pages including tables and figures.[2]

(B) After completing the "House prices" walkthrough, download the data on life expectancy (LifeExpectancy.csv) from the class website.

[1] For example, if you're building a regression model for the housing data set with price as the response and neighborhood/size as predictors, say this.

[2] This is an upper limit, not necessarily a target.

Life expectancy is often used as an indicator for the well-being of a country. Experts on economic development are interested in the relationship between a country's life expectancy and its economic well-being.

This data set has the following variables:

*Country:* the name of the country
*PPGDP:* per-person gross domestic product in US dollars
*LifeExp:* life expectancy at birth in that country
*Group:* whether the country is in the OECD, Africa, or neither (labeled as "other")

To clarify the "group" variable, the OECD is the Organization for Economic Cooperation and Development:

> The Organisation for Economic Co-operation and Development (OECD) . . . is an international economic organisation of 34 countries founded in 1961 to stimulate economic progress and world trade. It is a forum of countries describing themselves as committed to democracy and the market economy, providing a platform to compare policy experiences, seeking answers to common problems, identify good practices and coordinate domestic and international policies of its members.[3]

Build a regression model that relates life expectancy (the response) to GDP. Use a transformation if necessary. Address the question of whether it is necessary to include dummy variables and/or interaction terms for the "Group" variable in order to model the data adequately. As above, keep your answer to no more than two typed pages including tables and figures.

*(2) 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 "simdata03samp.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 "simdata03pop.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?*

[3] Wikipedia, `http://en.wikipedia.org/wiki/Organisation_for_Economic_Co-operation_and_Development`, accessed 7 Feb 2015.

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).[4] In the walkthrough, we used 365 Monte Carlo samples to simulate a year of fishing. 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.

[4] You will also find pages 99–104 of Chapter 5 of the course packet helpful for conceptual background.

### *(3) Bootstrapping*

Complete the "Creatinine, revisited" walkthrough on the class website.[5] 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).

Once you've done this, use what you've learned to address the following questions. None should require more than a single paragraph.

(A) In an earlier set of exercises, an interesting issue came up on the problem about the CAPM (capital-asset pricing model): was your estimated "beta" for Wal-Mart the same as that reported by Yahoo Finance? You may have seen a slight discrepency here. Can this discrepancy be explained by sampling variability? After all, you only looked at a sample of data, rather than entire history of Wal-Mart's returns. Could it be that you got a different beta simply because you used a different sample than Yahoo did? Use what you know about sampling distributions and bootstrapping to address whether this seems like a plausible explanation in light of the data. Show your evidence in a picture.

(B) 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. (This should be easy if you remember about dummy variables!) Concisely describe what you did, and report the interval.

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

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