Exercises 3

(1) Solder skips

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/cases/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], and explain what you have learned fro your data analysis. Write a short report summarizing your analysis and conclusions.

Preparing your report. Writing up the results of a data analysis is not a skill that anyone is born with. It requires practice and—at least here in the beginning—a bit of help.

When writing your report, organization will set you free. A good outline is: 1) overview of the problem, 2) your data and modeling approach, 3) the results of your data analysis (plots, numbers, etc), and 4) your substantive conclusions.

Overview: Describe the problem. What substantive question are you trying to address?

Data and model: What data did you use to address the question, and how did you do it? When describing your approach, be specific.

- Don't say, "I ran a regression" when you instead can say, "I fit a linear regression model to predict price that included a house's size and neighborhood as predictors."
- Justify your modeling approach. This need not take a lot of space. For example: "Neighborhood was included as a categorical predictor in the model because Figure 2 indicated clear differences in price across the neighborhoods."

Sometimes your Data and Model section will contain plots or tables, and sometimes it won't. If you feel that a plot helps the reader understand the problem or data set itself—as opposed to your results—then go ahead and include it.[1]

Results: In your results section, include any figures and tables necessary to make your case. Label them (Figure 1, 2, etc), give them informative captions, and refer to them in the text by their numbered

[1] A great example here is Tables 1 and 2 in the paper on the PREDIMED study, at `http://www.nejm.org/doi/pdf/10.1056/NEJMoa1200303`. These tables help the reader understand some important properties of the data and approach, but not the results of the study itself.

labels where you discuss them. Typical things to include here may include: pictures of the data; pictures and tables that show the fitted model; tables of model coefficients and summaries

Conclusions: what did you learn from the analysis? What is the answer, if any, to the question you set out to address?

Make the sections as short or long as they need to be. For example, a conclusions section is often pretty short, while a results section is usually a bit longer.

Here are some further general guidelines:

- It's OK to use the first person to avoid awkward or bizarre sentence constructions, but try to do so sparingly.
- Do not include computer code unless explicitly called for.
- When in doubt, use shorter words and sentences.
- A very common way for reports to go wrong is when the writer simply narrates the thought process he or she followed: "First I did this, but it didn't work. Then I did something else, and I found A, B, and C. I wasn't really sure what to make of B, but C was interesting, so I followed up with D and E. Then having done this. . . ." Do not do this. The desire for specificity is admirable, but the overall effect is one of amateurism. Follow the recommended outline above.

(2) Life expectancy and economic development

After completing the "House prices" walkthrough, download the data on life expectancy (LifeExpectancy.csv) from the class website. 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.[2]

Build a regression model that relates life expectancy (the response) to GDP. Use a transformation if necessary, and think carefully about whether the "Group" variable seems to modulate the relationship between GDP and life expectancy. Address two questions. 1) What would you predict the life expectancy to be for an OECD country with a GDP of $20,000 per person? 2) What about for an African country with a GDP of $1000 per person? Note: make sure to provide an interval prediction (not just a point prediction) and to provide some measure of the accuracy/coverage of your interval.

(3) Sampling variability and regression modeling

Time to run some simulations 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 "simdata03samp.csv" and "simdata03pop.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. Briefly summarize your understanding of the relationship here, quoting the coefficients, residual standard deviation, and R^2.

(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." 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.

Of course, on real problems, we can't look at the whole population.

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

But in this problem, you can. That means you can actually investigate what kinds of answers other samples might have given you. You'll do this by simulation. To simulate one random sample of size 50, try this:

```
npop = nrow(simdata03pop)
nsamp = 50
mysample = sample(1:npop, size=nsamp)
lmsub = lm(y~x, data=simdata03pop, subset = mysample)
coef(lmsub)
```

The random sample of size 50 generated by this code is no better or worse than the one in "simdata03samp.csv." It's just a different sample!

Of course, that's just one sample. You might be able to generate 5 or 10 like this before you get sick of it. Even this would give you some idea of sampling variability for the estimates of β_0, β_1, and R^2. If you want to stop there, that's OK. But can you set up a "for loop" to quickly generate 10,000 different samples from the population, each of size 50? (See the montecarlo.R file from the website for an idea of where to start. If you're inexperienced with programming, there's no need to go past Method 1 in this script.)

Whatever you decide to do here, make sure you provide some answer to the fundamental question at issue: how much can you trust the estimates of the model parameters arising from the sample in Part A? You get to define a specific measure of precision or trustworthiness here—make sure you say explicitly what measure you are using.

(C) Try exploring the effect of different sample sizes on the resulting uncertainty of an estimate. Make a plot of "size of random sample" on the x axis and "my measure of precision about the slope of the true regression line" on the y axis.