Exercises 3

Due Monday, February 15, 2016

(1) Polynomial regression and prediction intervals

For this question, you will return to the "utilities.csv" data set from class and the course packet. Recall that each row has information about a monthly utility bill for a house in Minnesota. The variable "gasbill" is the gas bill for that month, measured in dollars. The "temp" variable depicts the average temperature, in degrees F, for the billing period.

Compute the daily average gas bill, so that we can compare months with different numbers of days. Then fit first-order through third-order polynomial regression models for the daily average gas bill (Y) versus temperature (X). That is, fit the models

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i \\
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \\
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i
\end{aligned}
$$

For each model, report the fitted equation, together with the model's R^2.

Now pick your favorite model of these three, and briefly explain why you chose it. Use this favorite model to generate a naïve prediction interval for the gas bill during a month in which the average temperature is 50 degrees Fahrenheit. Make sure you state the empirical coverage level of your interval (i.e. quantify its likely forecasting accuracy).

(2) 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 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

[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.

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.

(3) Life expectancy and economic development

After completing the "House prices" walkthrough and reading through page 95 of the course packet, 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. Address two questions:

(A) Does the group variable seem to modulate the relationship between GDP and life expectancy?

(B) What would you predict the life expectancy to be for an OECD country with a GDP of $20,000 per person? 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.