Exercises 5 · Uncertainty

Due Monday, February 29, 2016

(1) Case study: cheese sales and promotional displays

This question considers data on sales volume, price, and advertisting display activity for packages of Borden sliced cheese, available as "cheese.csv" on the course website. For each of 88 stores (store) in different US cities, we have repeated observations of the weekly sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display during that week (disp = 1 for display). Altogether there are 5,555 observations in the data set.

Address the following questions thoroughly but concisely. Make sure to include the appropriate plots, statistical summaries, and measures of uncertainty to illustrate and support your conclusions.

(A) Ignoring price, do the in-store displays appear to have an effect on sales volume? Use an appropriate transformation for modeling consumer demand. In light of your analysis, complete the following two sentences. "I estimate that in-store displays increase/decrease sales by —%. I am 95% confident that this quantity is between —% and —%."

(B) Is there reason to suspect that your result in (A) is confounded by pricing strategies? Show evidence either way. If the answer is yes, propose a model that allows you to adjust for price in assessing the marginal effect of in-store displays on sales volume. Remember back to our milk sales-versus-price data that a typical model for price elasticity of demand is of the form $\hat{y}_i = Kx_i^{\beta}$, where $\hat{y}$ is expected sales, x is price, K is a constant, and β is the elasticity—that is, the marginal effect of price on sales volume. You should recall how to use linear least squares to fit such a model; now modify it to account for the effect of in-store displays.

As above, in light of your analysis, complete the following two sentences. "I estimate that in-store displays increase/decrease sales by —%, once the effect of price is accounted for. I am 95% confident that this quantity is between —% and —%." Again, make sure you properly account for differences in overall sales volume among stores.

(C) Does price elasticity for Borden cheese appear to be changed by the presence of in-store advertisement? (Hint: remember about inter-

action terms in models with numerical and categorical predictors.) As above, quote an appropriate confidence interval that addresses this question. Can you think of a possible economic explanation for your result here?

(D) What should Kroger's in Dallas/Ft. Worth charge for cheese in no-display weeks? What should they charge in display weeks? Assume that the wholesale cost of cheese is $1.50 per unit.

(2) Bootstrapped prediction intervals

For this problem, use the data set "shocks.csv." This data was taken by Monroe Shocks and Struts, a company that manufactures high-performance shock absorbers for top-end cars. Monroe offers a range of shock absorbers for cars of various sizes. These different shocks are distinguished from one another by their "rebound," a number which describes how aggressively the vibrations from the road are absorbed by the shock. Having an accurate understanding of a shock's rebound is important for safety; you don't want to put shocks designed for an SUV on a small car, or vice versa.

As part of its manufacturing process, Monroe tests each shock absorber to make sure it performs to the required rebound specification. They have one very accurate test of the shock's rebound, but this test is expensive. They also have a cheaper test, but this is less accurate.

In "shocks.csv," you have rebound readings on 35 different shock absorbers for both the expensive test and the cheap test. If the cheap test can accurately predict the result of the expensive test with minimal uncertainty, then it's OK to use the cheap test. But if it can't, then the expensive test must be used instead.

(A) Suppose the company is willing to use the cheap test as long it can predict at least 90% of the total variation in the readings given by the expensive test. In light of this data, should they use the cheap test? Why or why not?

(B) Now suppose the company adopts a more specific standard, and decides it is willing to use the cheap test if both of the following criteria are met. First, the slope of the regression line for the expensive test, given the cheap test, is close to 1, as measured by a 95% confidence interval. Second, the 95% prediction interval for the value of the expensive test, given the cheap test, is no wider than 18 units of rebound, as measured from center to endpoint. (Or, measured from endpoint to endpoint, the interval can be no wider

than 33 units of rebound.) This criterion must be met for readings of the cheap test (x) in the low (510), middle (550), and high (590) end of the rebound scale. That is, if the prediction interval for y is too wide at any of these three different x values, then the cheap test is not precise enough and cannot be used. Moreover, the prediction intervals must take into account parameter uncertainty. A naïve prediction interval may therefore be misleading here.

In light of the data and these criteria, should the company use the cheap test? If not, what criterion was missed and how? Use bootstrapping to account for your parameter and prediction uncertainty in forming your prediction intervals, and describe your methodology and results in a careful write-up. Make sure you use enough bootstrapped samples so that you can address the question without Monte Carlo error substantially affecting your results.

(3) The PREDIMED trial: a first look

For this problem, we'll revisit the PREDIMED trial, described in the course packet. For details, see this paper. The data is in predimed.csv from the course website.

The main goal of the trial was to understand the relationship between a Mediterranean diet and the likelihood of experiencing a major cardiovascular event (stroke, heart attack, or death from heart-related causes). Trial participants were assigned to one of three treatment arms, described in the paper as: "a Mediterranean diet supplemented with extra-virgin olive oil, a Mediterranean diet supplemented with mixed nuts, or a control diet (advice to reduce dietary fat)."

The predimed.csv file has data on many variables on each trial participant; we'll focus only on two:

- group: which treatment arm the person was assigned to
- event: yes or no, did the person experience a cardiac event during the study period

If you look at a contingency table for these two categorical variables, you get the following.

```
> xtabs(~event + group, data=predimed)
     group
event Control MedDiet + Nuts MedDiet + VOO
  No     1945           2030          2097
  Yes      97             70            85
```

Thus there is a hint that cardiact events happened at a slightly higher rate among participants in the control group.

Your task is to use a permutation test to assess whether this difference in event rates across the dietary categories could be explained due to chance. Note: you've seen a walkthrough of this kind of thing for a 2x2 table, but this is a 3x2 table with three levels of the predictor. You will have to define your own test statistic that collapses the association across categories to a single number. You have considerable freedom to choose a test statistic here; just make sure you are clear about what you are doing and why.