Exercises 5 · Multiple regression; hypothesis testing

(1) 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? 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 —%."

Note: make sure you properly account for differences in overall sales volume among stores in evaluating the relationship between display and sales volume. Also notice that I'm asking for a percentage (multiplicative) change due to the display, rather than an absolute (additive) change. Think carefully about why we should expect the change to be multiplicative, and about what kind of transformation would be appropriate for answering this question with a single number.[1]

(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 both price and store differences in assessing the marginal effect of in-store displays on sales volume. Remember back to our milk sales-versus-price data: 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 this model to account for the effect of in-store displays and store-level differences on sales.

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

[1] Hint: if $\log y_1 = a$ and $\log y_2 = a + b$, then what is the *ratio* y_2/y_1, expressed in terms of a and b?

that this quantity is between —% and —%." As in (A), 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 interaction 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 display weeks? Should their price change when they're not running a display ad? Assume that the wholesale cost of cheese is $1.50 per unit.

(2) Hypothesis testing

The National Football League modified its rules for overtime games in 2012, to try to reduce the unfair advantage associated with winning the coin toss in overtime. In case you care, you can read more about the rules here: `http://www.nfl.com/news/story/09000d5d827ee2c0/article/nfl-overtime-rules`

In the three years following that rule change, there were 70 overtime games, and the team that won that coin toss ended up winning 38 of them (54.2%). That looks to be a slight advantage in favor of the team winning the coin toss.

On the basis of this evidence, can you reject the null hypothesis that each team (both the coin toss winner and the coin toss loser) has an equal chance of winning a game that goes to overtime? Provide evidence either way, and briefly describe your process for answering the question.

(3) The PREDIMED study

For this problem, we'll look at data from 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 simple: 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, with two levels for the predictor, using relative risk as a test statistic. But this is a 3x2 table, with three levels of the predictor. You can proceed in two ways here:

1. You could define your own test statistic that describes the association between diet and event outcome in terms of 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.
2. Or you could create a new category (say MedDietAny) that collapses the two Mediterrenean diet categories into one, and proceed as for a 2-by-2 table.

Whatever you do, just explain it clearly.