Second Example Set

Matt Taddy – Chicago Booth
A Three Day Course in Applied Regression Analysis

1 Orange Juice

Consider 9649 observations of weekly orange juice sales at various locations of the Dominick's supermarket chain. In detail, the file `OJprice.csv` contains sales volume and price for Minute Maid orange juice (`minutevol` in fl.oz and `minuteprice` in dollars per oz), as well as the price for two competing brands: premium juice from Tropicana (`tropicprice`) and budget juice from Dominick's in-house brand (`dmnckprice`). We also have indicators for whenever each brand was featured in the store's marketing that week (`minutead`, `tropicad`, and `dmnkad`).

We are interested in understanding the influence of price on sales for our Minute Maid OJ.

(i) Plot the data and explain how what you see relates to the problem at hand.

(ii) What is the relationship between Minute Maid sales and price, ignoring the other brands? Does it appear that Minute Maid pricing is a function of other brands' pricing? How so, and why?

(iii) Does pricing for the other brands also influence sales? How does considering other brand prices change the effect of Minute Maid pricing on sales? Can you explain this?

(iv) What is the effect of advertisement on sales? How does it change price elasticity? Why?

2 Income and Vote

The data for this example, contained in `Election2008byState.csv` on the course webpage, include results by state for the 2008 presidential election as well as demographic, geographic, and income information for each state/district in the lower continental US.

In detail, the variables are $OBAMA$: number of votes for the Democratic candidate Obama, $MCCAIN$: number of votes for Republican candidate McCain, INC: Median income in dollars, AGE: Median age, $POPDENS$: Population density, $PBLACK$: Census percentage Black or African American, $PHISP$: Census percentage Hispanic, and $CHURCH$: Gallup poll percentage who report attending church/synagogue at least once a week.

We will consider the marginal effect of median income on relative vote-share *for* Barack Obama. That is, the response variable of interest is $V_{BO} = OBAMA/(MCCAIN + OBAMA)$.

(i) Consider a regression of V_{BO} onto income, and investigate the linear model fit. Are there any observations which you feel justified in removing as outliers?

(ii) Now consider adding state characteristic covariates into your model, with the goal of improving both fit and predictive ability. Are there any interaction terms?

(iii) Finally, present your conclusions about the marginal effect on income on voting patterns.

A Twist: Plot the CNN exit poll data in the file `Income-ObamaVoteShare.csv`, which includes Obama's percentage share of the popular vote – in the entire USA, and for California, Ohio, and Texas – divided by 8 income groupings (valued in multiples of $1000 per year). How does this data compare to your results for the income effect across states? Can you explain what you see?