Third Example Set

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

1 Amazon.com

The `amazon.csv` data file has, for a sample of households in the US, total purchases in 2008 from *amazon.com* (`bought`) and a series of demographic variables:

`region`: US census region; 1=North East, 2=Mid West, 3=South, 4=West.

`size`: Number of people in household

`oldest`: Age of oldest resident; 1=18-20, 2=21-24, 3=25-29,...11=65+.

`income`: Household income; 1= 15k-, 2=15-25k, 3=25-35k, 4=35-50k, 5=50-75K, 6=75-100k, 7=100k+.

`children`: 1 if there are children, 0 if not.

`race`: 1= White, 2=Black, 3=Asian, 5=Other/Mixed.

`connection`: 1 if broadband, 0 if dial-up.

`hispanic`: 1 if yes, 0 if not.

(i) How does faster (broadband) internet access affect purchasing behavior on amazon.com?

(ii) This data only consider people who made purchases. How might that bias our conclusions?

2 Washington State Police Traffic Stops

We will investigate whether or not there is a systematic racial bias in determining who is stopped by Washington State Police (WSP) officers. Many researchers suggest that a difference between the racial distribution of persons stopped by police and the racial distribution of the population at risk of being stopped would constitute evidence of racial profiling. This implicit definition reveals the key empirical problem in testing for racial profiling: measuring the risk set, or the *benchmark* racial distribution, against which to compare the racial distribution of traffic stops.

We focus on radar initiated traffic stops as a benchmark for the population that is *at-risk* to be stopped by the WSP. These drivers are selected from passing motorists based upon driving characteristics, and there is very little chance of racial bias. If members of a particular race are actively stopped at a different rate than predicted by this benchmark, we have evidence of racial bias. The data (in the file `WSPTrafficStops.csv` on the course site) consist of both self-initiated traffic stops (e.g., without a radar trap or a crash) and radar initiated traffic stops for each recorded racial group between November, 1, 2005 and September 30, 2006 for 34 autonomous patrol areas (APAs).

(i) We will first evaluate the appropriateness of $Radar$ as a benchmark for stops.

- Plot the data and determine the appropriate SLR model for the relationship between $Stops$ and $Radar$. Use residual plots to evaluate model fit, and interpret the parameters.
- Do you think that $Radar$ is a good predictor for the baseline level of active $Stops$?

(ii) Test for a racial bias in traffic stops: consider an expanded model that allows for additional bias effects on the conditional expectation for active stops of each race given our benchmark.

- Plot the data to illustrate potential race effects, and fit a model that allows for racial bias.
- Evaluate model fit and test for whether or not the race effects are significant.

(iii) Can the data be better explained by higher than expected active stops for any single racial group?

- Consider you data plot in (ii) and devise a potential model.
- Test whether this model is more appropriate than the full racial bias model in (ii).

(iv) Using the best model you've found, plot and compare conditional expected $Stops$ given $Radar$ and $Race$. What is your final conclusion?