***A tour of large dataset logistic regression in R***

Slide 1: “Goals”
R has been criticized for the speed with which it handles large datasets, which has led to the development of specialized packages, such as “biglm” that mimic the way SAS handles memory in order to better process large datasets. Another technique is to install the 64 bit version of R rather than the recommended 32 bit version, since the 32 bit version stores all memory in RAM and is generally faster, but for large datasets the additional memory handling capacity of the 64 bit version will come in handy. It is also possible to speed up R by using a LINUX operating system rather than Windows. With this in mind, I proceed to run a logistic regression model on a large dataset ( n > 1 M = population of Idaho).

Slide 2: “Money is Skewed”
Our only continuous variables are two forms of money. Income from self-employment (SEMP) and wages (RWAGP). This plot nicely depicts the skewed nature of these two variables. However, they are even more skewed than the plot suggests since the plot is cut off at $ = 100,000 for each.

Slide 3: “Dependent Variable”
The first attempt at defining a person as an entrepreneur was to define them as such if the percentage of their income from self employment was larger than 38%. However, this definition suffers from the deficiency that an individual could possibly earn very little of either, but a relatively larger amount from self employment and be defined as an entrepreneur when, in fact, they may not be. For example, many CEO’s may only officially make $1 in wages as most of their income is derived from stock options or other non-monetary sources.

Slide 4: Dependent Variable”
These brilliant graphs show how if you adjust the percentage up or down (25%, 65% and 38% shown), you include more or less individuals in your definition of entrepreneur. Again the axes are circumscribed to only include $0-100,000 to make the graphs more understandable.

Slide 5: “Dependent Variable”
Unfortunately, the 38% entrep definition yielded about 1% of the population as entrepreneurs. Upon further investigation, a variable, “cow”, or class of worker, was found where individuals self-defined themselves as entrepreneurs. This provided a more reasonable result where nearly 10% of the population were defined as entrepreneurs.

Slide 6: no title
These bar graphs break age into 5-year intervals, with 0-15 cut out of the dataset and 76+ lumped together. This categorization of age allows for the bar chart and also allows the logistic regression to be cast in terms of all categorical variables, which will allow for a goodness of fit test. The top bar graph shows the entire population broken into age categories with self-defined entrepreneurs representing the red portions of each strip. These entrepreneurs are their own population in the bottom chart, with red and blue designating the two genders now. It seems the entrepreneurial population becomes more male dominant as it ages.

Slide 7: no title
Break down entrepreneurs by the various educational categories. 9 represents high school.

Slide 8: no title
This is interesting because it appears that the population of entrepreneurs is slightly different than non-entrepreneurs in terms of their marital relationships. Perhaps entrepreneurs require a stable base in order to devote themselves to their businesses.

Slide 9: no title
Perhaps citizenship has an effect on the probability of being an entrepreneur.

Slide 10: no title
It was also conjectured that the wealthy may be more likely to become entrepreneurial since they had the means to pursue building their own businesses. Indeed it does appear that more entrepreneurs are more concentrated in the wealthier property value brackets.

Slide 11: “Model Selection”
Unfortunately, yes, R had trouble with this dataset, which was only 250,000 observations after incorporating all the variables and their respective missing values. However, I already had the “bestglm” package loaded so I went ahead and ran a simple example just to compare it to the leaps package which is used for simpler OLS models.

Slide 12: “Interpreting Coefficients”
There are 3 ways to look at the coefficients in a logistic regression by various manipulations of the dv.

Slide 13: no title
These are the three manipulations of the dv. You can see that we might compare a coefficient, $β\_{i}$, to zero because it would have no effect on the logged odds. Similarly we might compare $e^{β\_{i}}$ to 1 since it would have no effect on the odds. You have to input Xi values to compute a given p…these are usually called “marginal effects”

Slide 14: “Probabilities: ref: marginals”
There are at least two ways to compute marginal effects, this code illustrates them.

Slide 15: “Logits”
Logits are the standard output of many statistical programs. We can also look at them like an anova table, using type II or III effects.

Slide 16: “Assessing the Model”
Similar to lecture, we have a measure of R2 that we can use that I call D2. Here is a handy function in R to use on multiple models.

Slide 17: “Assessing the Model”
If you type plot(logistic glm object) you get the following plots. They are not as easy to interpret as their OLS counterparts.