Problem Sets 5 and 6
ECON 60303

Bill Evans
Spring 2014

Problem basics

In this problem set, I want you to use Matlab to write a program that uses a quasi-Newton search to maximize a log-likelihood function. You are to read in the data, write subroutines that calculate the log-likelihood function, the gradient and the hessian, a hill-climbing program that obtains the maximum likelihood estimate (MLE), plus print out the results to an external file.

The problem set will be done in stages. By Friday, March 21, I want you to turn in a printout of the three subroutines that generate the log-likelihood function, the gradient and the hessian, plus a printout of the comparison between the analytic and numeric values of the gradient and the elements along the main diagonal of the hessian for starting values of the parameters. Once you have verified that the subroutines are internally consistent (i.e., the analytic and numeric derivatives are similar), Friday, March 28, I want you to write the program that obtains the MLE. Use my sample programs on the web to read in the data, compare the analytic and numeric derivatives, and print out the results.

Problem context

You are to estimate a model that explains a dichotomous choice problem (a logit model) where you are trying to describe the correlates of whether someone smokes or not. In this example, let y_i =1 if the person smokes and 0 otherwise. Let x_i be a (1 x k) vector of observed characteristics. We can model the probability that someone smokes by the use of a logistic model where

$$\Pr(y_i = 1) = \frac{\mathrm{e}^{x_i\beta}}{1+\mathrm{e}^{x_i\beta}} \text{ and } \Pr(y_i = 0) = \frac{1}{1+\mathrm{e}^{x_i\beta}}$$

Given a sample of n observations, the log-likelihood is therefore

$$\ell = \sum_{i=1}^{n}\left[y_i \ln[\Pr(y_i = 1)] + (1-y_i)\ln[\Pr(y_i = 0)]\right] \text{ or}$$

$$\ell = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right) + (1-y_i)\ln\left(\frac{1}{1+e^{x_i\beta}}\right)\right] = \sum_{i=1}^{n}\left[y_i(x_i\beta) - \ln\left(1+e^{x_i\beta}\right)\right]$$

This is a globally concave function so there is a unique maximum and the search routine should quickly move to the top of the hill.

For starting values and parameters values to compare the analytic and numeric derivatives, use OLS estimates: $\hat{\beta} = (x'x)^{-1}x'y$.

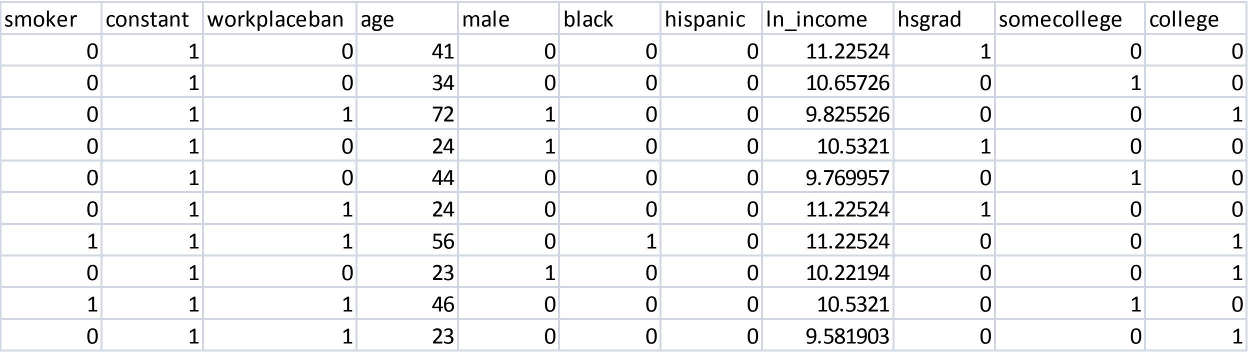
Data

The data for this project contains information on 1626 workers. The data is contained in the excel data set workplacesmoking.xlsx. There are 11 variables. The variable in column 1 is the dependent variable (smoker), the variable in column 2 is the intercept and the next 9 variables are the covariates for the model. The first ten lines of the data set are listed below and the table below defines the variables. The data set has been used in the past to examine whether workplace smoking bans decrease smoking among workers.

Variable definitions: workplacesmoking.xlsx

Column	Variable name	Description
A	Smoker	Dummy variable, =1 if the worker smokes, =0 otherwise
B	Constant	Constant term, =1 for all observations
C	Workplaceban	Dummy variable, =1 is the employee works in an establishment that bans smoking, =0 otherwise.
D	Age	Age in years
E	Male	Dummy variable, =1 if the respondent is male, =0 otherwise.
F	Black	Dummy variable, =1 if the respondent is black, non-Hispanic, =0 otherwise. The reference group is white, non-Hispanics.
G	Hispanic	Dummy variable, =1 if the respondent is Hispanic, =0 otherwise. The reference group is white, non-Hispanics.
H	ln_income	Natural log of family income
I	Hsgrad	Dummy variable, =1 if the respondent is a high school graduate, =0 otherwise. The reference group is high school dropouts.
J	somecollege	Dummy variable, =1 if the respondent has some college education but not a degree, =0 otherwise. The reference group is high school dropouts.
K	college	Dummy variable, =1 if the respondent is a college graduate, =0 otherwise. The reference group is high school dropouts.

First 10 observations from workplacesmoking.xlsx



smoker	constant	workplaceban	age	male	black	hispanic	ln_income	hsgrad	somecollege	college
0	1	0	41	0	0	0	11.22524	1	0	0
0	1	0	34	0	0	0	10.65726	0	1	0
0	1	1	72	1	0	0	9.825526	0	0	1
0	1	0	24	1	0	0	10.5321	1	0	0
0	1	0	44	0	0	0	9.769957	0	1	0
0	1	1	24	0	0	0	11.22524	1	0	0
1	1	1	56	0	1	0	11.22524	0	0	1
0	1	0	23	1	0	0	10.22194	0	0	1
1	1	1	46	0	0	0	10.5321	0	1	0
0	1	1	23	0	0	0	9.581903	0	0	1

What variables to print out:
You are to print out the following results to an external file: the number of iterations it took to get to the top of the hill, the final convergence criteria, the value of the log-likelihood function at the top of the hill, plus the parameter estimates, standard errors, z-score, and p-value on the test of the hypothesis that the parameter equals 0 for all 10 covariates in the model (a Wald test).