Problem Set 2

Directions: Answer all questions. Clearly label all answers. Show all of your code. Turn in m-file(s), a log file (using `diary` or from off the cluster) and a writeup to me via your dropbox in Sakai (in a folder labeled 'MatlabPS1.2') by 11:59 p.m. on Thursday, July 12, 2012. Put the names of all group members at the top of your writeup (each student must turn in his/her own materials).

1. Practice with Matlab's functional optimizers using `nlsw88.mat` (from PS1)

   (a) Estimate the following model:

   $$\ln(wage_i) = \beta_1 + \beta_2 age_i + \beta_3 black_i + \beta_4 other_i + \beta_5 collgrad_i + \varepsilon_i \quad (1)$$

   under the assumption that ε_i is mean-zero and well-behaved (i.e., use OLS). Note that $black_i$ corresponds to $race_i = 2$ and $other_i$ corresponds to $race_i = 3$. *Be sure to drop observations for all variables where any of the variables are missing. Also, report the sum of squared residuals at convergence, and the estimation sample size.*

   i. Estimate $\hat{\beta}$ and s^2 (variance of ε_i) using `fminsearch` with default convergence tolerances, and starting values of $U[0,1]$. Set the `rand` seed at 1234 (just like last problem set). For all optimizations in this problem set, set the maximum iterations at 10^6 and the maximum function evaluations at 10^{12}

   ii. Estimate $\hat{\beta}$ and s^2 using `fminunc` with default convergence tolerances and $U[0,1]$ starting values

   iii. Estimate $\hat{\beta}$ and s^2 using `fmincon` with default convergence tolerances and with $\beta_3 < 0$ as the only restriction. Use $U[0,1]$ starting values.

   iv. How do your answers differ when using each of the optimizers? How different are your answers from the closed-form solution for OLS? ($\hat{\beta} = (X'X)^{-1}X'y$)

   (b) Now estimate the same model from (a), but assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. In this case, the log likelihood function looks like

   $$\ell(X_i;\beta,\sigma) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(\ln(wage_i) - X_i\beta\right)^2 \right\} \quad (2)$$

   where $X_i\beta$ is the right-hand side of equation (1) (except ε, of course). *Report the log likelihood value at convergence, and sample size. Remember to take the negative of the likelihood, since Matlab's optimizers are minimizers and you want the maximum likelihood estimator.*

i. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ (variance of ε_i) using `fminsearch` with default convergence tolerances and $U[0,1]$ starting values

ii. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminunc` with default convergence tolerances and $U[0,1]$ starting values

iii. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fmincon` with default convergence tolerances and with $\beta_3 < 0$ and $\sigma > 0$ as the only restrictions. Use $U[-.25,.25]$ as the starting values.

iv. How do your answers differ when using each of the optimizers? How sensitive is $\hat{\beta}$ to the normal distribution assumption? How close are s^2 and $\hat{\sigma}^2$?

(c) Now estimate the following model:

$$\begin{aligned}\ln(wage_i) &= \beta_1 + \beta_2 age_i + \beta_3 black_i + \beta_4 other_i + \beta_5 collgrad_i + \\ &\quad \beta_6 grade_i + \beta_7 married_i + \beta_8 south_i + \beta_9 c\_city_i + \\ &\quad \beta_{10} union_i + \beta_{11} ttl\_exp_i + \beta_{12} tenure_i + \beta_{13} age_i^2 + \\ &\quad \beta_{14} hours_i + \beta_{15} never\_married_i + \varepsilon_i \end{aligned} \tag{3}$$

Again, be sure to drop observations for all variables where any of the variables are missing. Also, report the sum of squared residuals and/or log likelihood at convergence, and the estimation sample size.

i. Estimate $\hat{\beta}$ and s^2 using `fminsearch` with default convergence tolerances, assuming ε_i is mean-zero. Use "noisy but good" starting values, i.e. set the starting values equal to $\hat{\beta}_{OLS,closed\_form} + U\left[-\alpha\hat{\beta}_{OLS,closed\_form}, \alpha\hat{\beta}_{OLS,closed\_form}\right]$ with $\alpha = .75$.

ii. Estimate $\hat{\beta}$ and s^2 using `fminunc` with default convergence tolerances, assuming ε_i is mean-zero. Use the same starting values as in part (i).

iii. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminsearch` with default convergence tolerances, assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i).

iv. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminunc` with default convergence tolerances, assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i).

v. How does `fminsearch` compare to `fminunc` when the dimension of the parameter vector increases?

vi. How do the estimators in (i) through (iv) perform when the starting values are 0.01 for all parameters?

2. Practice with Matlab's functional optimizers using `nhanes2d.mat` (data from the National Health and Nutritional Examination Survey—NHANES). Details on the variables:

`ID`	:	unique individual identifier
`age`	:	age (in years)
`hct`	:	hematocrit percentage (% of red blood cells in the blood)
`heartatk`	:	binary variable indicating heart attack history
`height`	:	height (in cm)
`highbp`	:	binary variable indicating history of high blood pressure
`houssiz`	:	size of household
`race`	:	1 = white, 2 = black, 3 = other
`region`	:	1=northeast, 2=midwest, 3=south, 4=west
`sex`	:	1=male, 2=female
`smsa`	:	1=central city, 2=non central city, 4=rural
`weight`	:	weight (in kg)

(a) Estimate the following model:

$$\begin{aligned} hct_i &= \beta_1 + \beta_2 age_i + \beta_3 black_i + \beta_4 other_i + \beta_5 heartatk_i + \\ &\quad \beta_6 female_i + \beta_7 highbp_i + \beta_8 northeast_i + \beta_9 midwest_i + \\ &\quad \beta_{10} south_i + \beta_{11} non\_central\_city_i + \beta_{12} rural_i + \beta_{13} height_i + \\ &\quad \beta_{14} weight_i + \beta_{15} houssiz_i + \varepsilon_i \end{aligned} \tag{4}$$

Be sure to drop observations for all variables where any of the variables are missing. Also, report the sum of squared residuals and/or log likelihood at convergence, number of iterations to convergence, and the estimation sample size.

i. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminsearch` with default convergence tolerances, assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i) of 1(c), but now with $\alpha = 1.5$.

ii. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminsearch` with `TolX` and `TolFun` each set to 10^{-8} (instead of the default), assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i) of 2(a).

iii. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminunc` with default convergence tolerances, assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i) of 2(a).

iv. Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ using `fminunc` with `TolX` and `TolFun` each set to 10^{-8} (instead of the default), assuming $\varepsilon_i \overset{iid}{\sim} N(0,\sigma)$. Use the same starting values as in part (i) of 2(a).

v. How do your answers change when the convergence tolerance changes? How many more iterations did the optimization require under the stricter tolerances? How different are your answers depending on the optimizer?

3. Summarize your findings regarding Matlab's different optimizers. When is `fminsearch` the best optimizer to use? When is `fminunc` the best to use? How important are starting values in finding a solution? How important are convergence tolerances?