Problem Set 3

Directions: Answer all questions. Clearly label all answers. Show all of your code. Turn in the following to me via the assignment submission in Sakai ('Matlab 1 PS 3' assignment) by 11:55 p.m. on Thursday, July 19, 2012:

- m-file(s)
- a log file (from off the cluster)
- matsub.oXXXXX file
- pdf version of your writeup with its LaTeXsource code

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 graphics using `nhanes2d.mat` (from PS2) — visualizing descriptive evidence

   (a) Generate a kernel density plot of the variable `hct`. How "normal" does the variable look? Be sure to add axis labels and a title to your graph.

   (b) Now generate a histogram of `hct` with a normal pdf overlaying it. Again, label axes and assign a title appropriately. Do you think `hct` still looks "normal"?

   (c) On the same figure, plot the empirical cdf of `hct` by gender. Plot males and females with different line styles and include a legend. Does one gender stochastically dominate another?

   (d) Now repeat (c) but break out the distribution by region instead of gender.

   (e) Now repeat (c) but break out the distribution by race instead of gender.

2. Viewing model fit graphically

   (a) Graphing a predicted OLS plane

      i. Estimate the following model (from question 2(a) of Problem Set 2):

$$\begin{aligned} hct_i &= \beta_1 + \beta_2 age_i + \beta_3 black_i + \beta_4 other_i + \beta_5 heartatk_i + \\ &\quad \beta_6 female_i + \beta_7 highbp_i + \beta_8 northeast_i + \beta_9 midwest_i + \\ &\quad \beta_{10} south_i + \beta_{11} non\_central\_city_i + \beta_{12} rural_i + \beta_{13} height_i + \\ &\quad \beta_{14} weight_i + \beta_{15} houssiz_i + \varepsilon_i \end{aligned} \tag{1}$$

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.

Use `ezsurf` to graph the "marginal" OLS plane for the intercept and the variables `height` and `weight`. Display the graph over the bounds $[\min(height), \max(height)] \times [\min(weight), \max(weight)]$. Correctly label all axes, and title your graph appropriately.

(b) Graphing actual data vs predicted OLS plane

i. Add actual data points to your graph in part (i) of question (a). Make sure your dimensions are correctly lined up. How well do height and weight jointly predict hematocrit percentage (conditional on all other covariates in the OLS data matrix)?

3. Maximum likelihood estimation for a discrete dependent variable (high blood pressure)

(a) Using `fminunc` (with convergence tolerances at 10^{-8}), estimate the following model:

$$\begin{aligned} highbp_i &= \beta_1 + \beta_2 age_i + \beta_3 black_i + \beta_4 other_i + \beta_5 heartatk_i + \\ &\quad \beta_6 female_i + \beta_7 hct_i + \beta_8 northeast_i + \beta_9 midwest_i + \\ &\quad \beta_{10} south_i + \beta_{11} non\_central\_city_i + \beta_{12} rural_i + \beta_{13} height_i + \\ &\quad \beta_{14} weight_i + \beta_{15} houssiz_i + \varepsilon_i \end{aligned} \tag{2}$$

assuming $\varepsilon_i \sim$logistic. For this problem, the log likelihood looks like

$$\ell(X_i;\beta) = \sum_{i=1}^{n} \{1[highbp_i = 1] \ln(P_i) + 1[highbp_i = 0] \ln(1 - P_i)\}$$

where $1[highbp_i = 1]$ is a dummy for whether or not $highbp_i$ is 1, and

$$P_i = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}.$$

Be sure to drop observations for all variables where any of the variables are missing. Also, report the standard errors of $\hat{\beta}$, the log likelihood at convergence, number of iterations to convergence, and the estimation sample size.

(b) Repeat (a), but now assume that $\varepsilon_i \sim N(0,1)$ (i.e., estimate the probit model). In this case, the log likelihood is the same as in (a), but with

$$P_i = \Phi(X_i\beta),$$

where Φ is the CDF of the standard normal distribution.

Note: If you're like me and have trouble getting your probit likelihood to converge, try estimating it first with `fminsearch` (starting from your logit answers), then with `fminunc` (starting from where `fminsearch` ended). If this is still problematic, consider modifying your likelihood function so that Matlab doesn't attempt to evaluate the log of 0. This can be done, e.g., by setting $P_i = 10^{-200}$ if $P_i = 0$.

(c) Divide your logit $\hat{\beta}$ estimates by 1.6 and compare with your probit $\hat{\beta}$ estimates. How different are the two vectors? Briefly discuss some of the main results from this estimation.

(d) Compare the fit of the models estimated in (a) and (b) in two ways: *(i)* compare the average of *highbp* with the average $\hat{P}_{logit}$ and $\hat{P}_{probit}$[1]; *(ii)* compare the maximum likelihood values. Which model does better at fitting the average probability? Which model has a higher likelihood value?

[1] $\hat{P}_i = f\left(X_i\hat{\beta}\right)$