Problem Set 3 Solutions

Directions: Answer all questions. Clearly label all answers. Show all of your code. Turn in the following to me via your dropbox (in a folder labeled 'MatlabPS1.3') in Sakai by 11:59 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) See Figure 1. The data look quite normal.

   (b) See Figure 2. Once we look at the histogram, the data don't look nearly as normally distributed as with the smoothed graph.

   (c) See Figure 3. Males clearly stochastically dominate females, meaning that, at any point in their respective distributions, a male has more red blood cells (as a percentage of blood volume) than a female.

   (d) See Figure 4. There is no such pattern by region.

   (e) See Figure 5. Whites/other stochastically dominate blacks in terms of hematocrit percentage.

2. Viewing model fit graphically

   (a) Graphing a predicted OLS plane

      i. See Figure 6.

   (b) Graphing actual data vs predicted OLS plane

      i. See Figure 7. The model fit is not good at all. Even after conditioning on 14 covariates and an intercept, there is still a large amount of variation in hematocrit percentage that is unexplained.

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

   (a) Logit estimates are found in Table 1.

   (b) Probit estimates are found in Table 2.

   (c) Table 3 compares the two sets of estimates. Some coefficients are much closer than others, but they are still quite different overall. Intepreting the model, high blood pressure is associated with being older, black, having higher hematocrit, having a larger household and being taller and heavier. Obviously, there isn't a model establishing causality for any of these, so all we can say in terms of interpretation is that the observed effects are associated with high blood pressure. I was also surprised that having a heart attack was negatively related to high blood pressure. This is probably due to heart attack victims paying closer attention to their biometric data after suffering a heart attack.

   (d) See Table 4 for a summary of the model fit. The logit fits better in terms of $\overline{P}$, but the probit has a higher log likelihood value.

Figure 1: Kernel Density Plot of Hematocrit Percentage

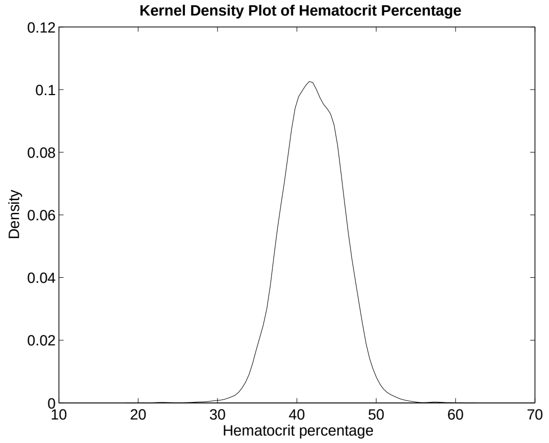

Figure 2: Distribution of Hematocrit Percentage

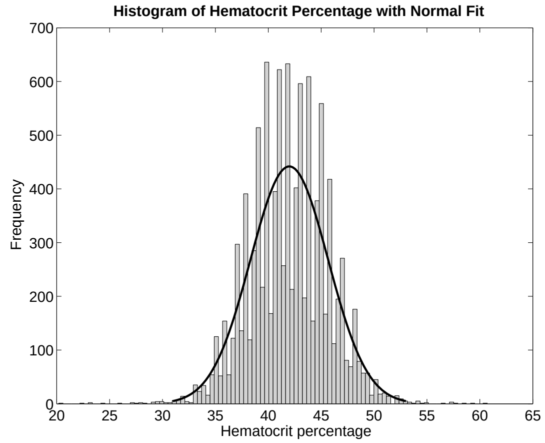

Figure 3: Empirical CDF of Hematocrit Percentage by Gender

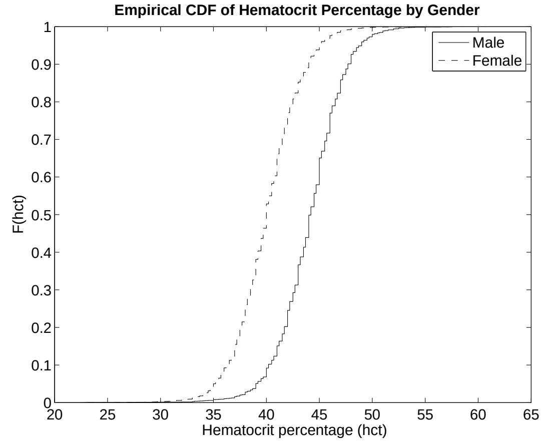

Figure 4: Empirical CDF of Hematocrit Percentage by Census Region

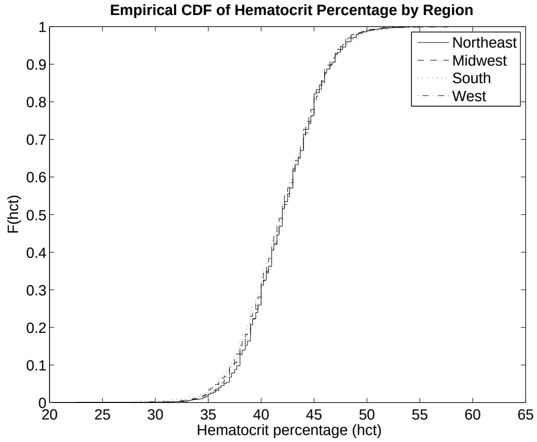

Figure 5: Empirical CDF of Hematocrit Percentage by Race

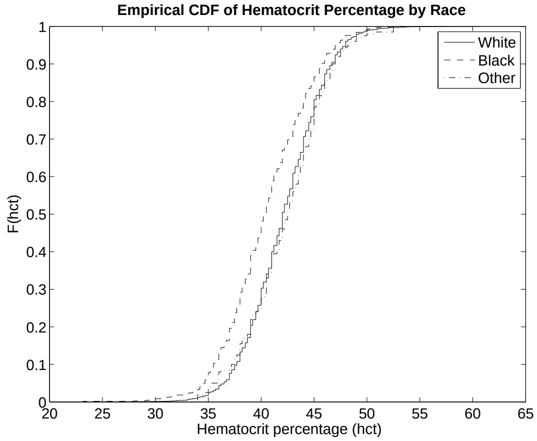

Figure 6: OLS plane predicting hematocrit percentage by height and weight

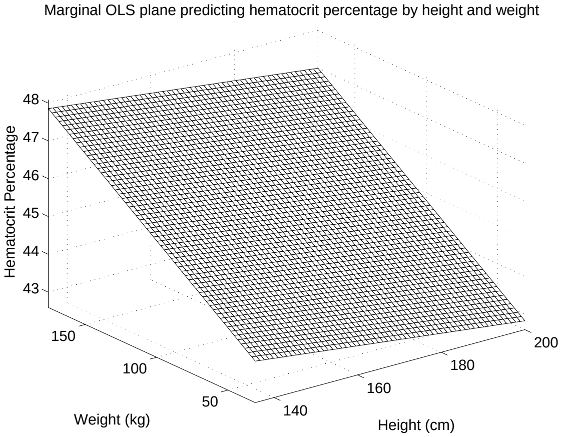

Figure 7: OLS plane predicting hematocrit percentage by height and weight, with actual data values

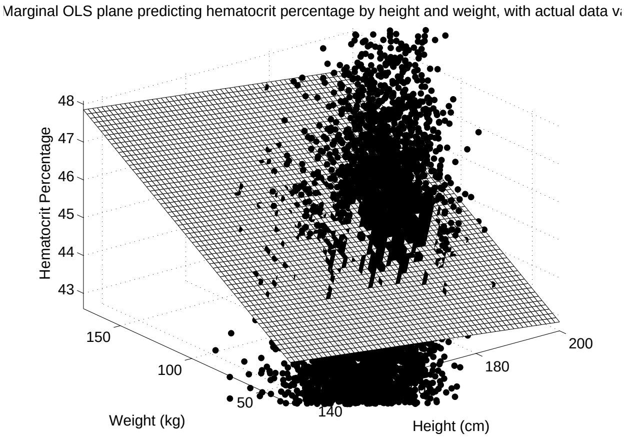

Table 1: Logit MLE Estimates

Variable	$\hat{\beta}$	Std Err.
Constant	-4.825	1.017
age	0.047	0.003
black	0.536	0.104
other	0.338	0.249
heartatk	-0.266	0.132
female	-0.091	0.099
hematocrit %	0.031	0.010
NE	0.145	0.095
MW	0.039	0.090
S	0.106	0.090
non central city	0.124	0.091
rural	0.127	0.084
height	-0.029	0.005
weight	0.048	0.002
household size	0.049	0.021
log likelihood	-3,414.90	
iterations	169	
N	10,349	

Table 2: Probit MLE Estimates

Variable	$\hat{\beta}$	Std Err.
Constant	-2.539	0.552
age	0.025	0.001
black	0.307	0.057
other	0.189	0.133
heartatk	-0.150	0.074
female	-0.056	0.054
hematocrit %	0.017	0.006
NE	0.074	0.052
MW	0.014	0.049
S	0.047	0.049
non central city	0.068	0.050
rural	0.080	0.046
height	-0.016	0.003
weight	0.027	0.001
household size	0.023	0.011
log likelihood	-3,401.55	
iterations	7,757	
N	10,349	

Table 3: Logit vs. Probit

Variable	$\hat{\beta}_{logit}/1.6$	$\hat{\beta}_{probit}$
Constant	-3.016	-2.539
age	0.029	0.025
black	0.335	0.307
other	0.211	0.189
heartatk	-0.166	-0.150
female	-0.057	-0.056
hematocrit %	0.020	0.017
NE	0.090	0.074
MW	0.024	0.014
S	0.066	0.047
non central city	0.077	0.068
rural	0.079	0.080
height	-0.018	-0.016
weight	0.030	0.027
household size	0.031	0.023

Table 4: Model Fit

	$\overline{P}$	log likelihood
Data	.1290	—
Logit Model	.1290	-3,414.90
Probit Model	.1287	-3,401.55