Solutions to Problem Set 1
Economics 551, Yale University
Professor John Rust

Question 1 The choice probability for the general case when

$$\begin{pmatrix} \epsilon_0 \\ \epsilon_1 \end{pmatrix} \sim N(\mu, \Sigma),$$

where

$$\begin{aligned} \mu &= \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} \\ \Sigma &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}. \end{aligned}$$

The choice probability is given by:

$$\begin{aligned} \Pr(1|x, \theta, \mu, \Sigma) &= \Pr\{u(1,x,\theta) + \epsilon_1 \geq u(0,x,\theta) + \epsilon_0\} \\ &= \Pr\{\epsilon_1 - \epsilon_0 \geq u(0,x,\theta) - u(1,x,\theta)\} \\ &= \Phi\left(\frac{u(1,x,\theta) - u(0,x,\theta) + \mu_1 - \mu_0}{\sqrt{\sigma_{11} - 2\sigma_{12} + \sigma_{22}}}\right), \end{aligned} \tag{1}$$

where we use the fact that $\epsilon_1 - \epsilon_0 \sim N(\mu_1 - \mu_0, \sigma_{11} - 2\sigma_{12} + \sigma_{22})$, so we standardized $\epsilon_1 - \epsilon_0$ by subtracting $\mu_1 - \mu_0$ from both sides of the inequality in the probability in the second line of the above equation and divided both sides by its standard deviation, allowing us to use the standard normal CDF $\Phi(x)$ in the last line. Note that when $\mu_1 = \mu_0 = 0$ and $\sigma_{11} = 1/2 = \sigma_{22}$ and $\sigma_{12} = 0$, this equation reduces to

$$\begin{aligned} \Pr(1|x, \theta, \mu, \Sigma) &= \Pr\{u(1,x,\theta) + \epsilon_1 \geq u(0,x,\theta) + \epsilon_0\} \\ &= \Phi\left(u(1,x,\theta) - u(0,x,\theta)\right). \end{aligned} \tag{2}$$

We note that we must make identifying normalizations of the μ and Σ parameters of this model, since there are infinitely many different combinations of the 5 free parameters in μ and Σ that yield the same conditional probability in equation (1), and are thus *observationally equivalent*. For example, let $\mu_0 = \mu_1 = \mu \neq 0$ and let σ_{11}, σ_{12} and σ_{22} be any parameters satisfying 1) Σ is positive semidefinite, and 2) $\sqrt{\sigma_{11} - 2\sigma_{12} + \sigma_{22}} = 1$ (one example is $\sigma_{11} = 1.5 = \sigma_{22}$ and $\sigma_{12} = -1$). This model has the same choice probability as the model in equation (2) where $\mu_0 = \mu_1 = 0$, and $\sigma_{11} = \sigma_{22} = .5$ and $\sigma_{12} = 0$. Therefore we need to impose arbitrary *identifying normalizations* in order to estimate the model. One common normalization is that $\mu_0 = \mu_1 = 0$ and $\sigma_{11} = 1/2 = \sigma_{22}$ and $\sigma_{12} = 0$. An alternative identifying normalization is $\mu_0 = \mu_1 = 0$ and $\sigma_{11} = 1 = \sigma_{22}$ and σ_{12} a free parameter to be estimated. However whether it is possible to identify the covariance term σ_{12} depends on the specification of the utilities $u(i,x,\theta)$, $i=0,1$. We will generally need to impose additional identifying normalizations to estimate the parameters of the utilities, $u(i,x,\theta)$, $i=0,1$. For example if the utility function is linear in parameters, i.e. $u(i,x,\theta)=x\theta_i$, $i=0,1$ with $\theta=(\theta_0,\theta_1)$, then it is easy to see that without further restrictions it is not possible to identify θ and σ_{12} simultaneously, even with the normalization that $\sigma_{11}=\sigma_{22}=1$. To see this, note that for the linear in parameters specification we have:

$$\begin{aligned}
\Pr(1|x,\theta,\mu,\Sigma) &= \Pr\{u(1,x,\theta)+\epsilon_1 \geq u(0,x,\theta)+\epsilon_0\} \\
&= \Phi\left(\frac{u(1,x,\theta)-u(0,x,\theta)}{\sqrt{2+2\sigma_{12}}}\right) \\
&= \Phi\left(\frac{x(\theta_1-\theta_0)}{\sqrt{2-2\sigma_{12}}}\right)
\end{aligned}$$

It should be clear that any combination of $(\theta_0,\theta_1,\sigma_{12})$ such that

$$\frac{(\theta_1-\theta_0)}{\sqrt{2-2\sigma_{12}}}=\beta,$$

for a fixed vector β are observationally equivalent, and that there are infinitely many such combinations. Therefore we must make a further normalization of the θ coefficients. A typical normalization is that $\theta_0=0$ and that one of the components of θ_1 is normalized to 1. Since we are free to choose different normalizations, when interpreting the estimation results from the probit model we need to keep the underlying normalization in mind. For the rest of this problem set we will use the normalization $\sigma_{11}=1/2=\sigma_{22}$ and $\sigma_{12}=0$, and $\theta_0=0$. Under this normalization the choice probability is given by $\Psi(x,\theta)=\Pr\{1|x,\theta\}=\Phi(x\theta)$, so the estimated value of θ is interpreted as the impact of an additional unit of x on the incremental utility of choosing alternative 1, i.e.

$$\theta=\frac{\partial}{\partial x}[u(1,x,\theta)-u(0,x,\theta)].$$

Question 2 See answer to questions 7 and 8 of 1997 Econ 551 problem set 3.

Question 3 The "true model" used to generate the data in model 3 was a probit model. Table 1 below presents the true θ coefficients and the logit and probit estimates of these values which were estimated using the shell program `estimate.gpr` using two procedures, `log_mle.g` and `prb_mle.g` that compute the log-likelihood, gradients and hessians for the logit and probit specifications, respectively. Both log-likelihoods have the following general form:

$$L_N(\theta)=\frac{1}{N}\sum_{i=1}^{N} I\{y_i=1\}\ln\left[\Psi(x_i|\theta)\right]+I\{y_i=0\}\ln\left[1-\Psi(x_i|\theta)\right]. \qquad (3)$$

Table 1: Maximum Likelihood Estimates of θ using `data3.asc`

Parameter	θ_0	θ_1	θ_2	θ_3
True Value	0.500	0.200	0.100	-0.050
Probit MLE	0.543	0.173	0.083	-0.038
Std. Dev.	0.031	0.040	0.020	0.011
Logit MLE	0.877	0.295	0.144	-0.065
Std. Dev.	0.051	0.068	0.034	0.019

Notice that the logit parameter estimates are "further" from the true parameters than the probit estimates. One "metric" for measuring this distance is the Wald test statistic $\hat{W}$ of the hypothesis that the estimated logit parameters equals the true parameters

$$\hat{W} = N(\hat{\theta}_N - \theta^*)'\hat{\Sigma}_N^{-1}(\hat{\theta}_N - \theta^*),$$

where $\hat{\Sigma}_N$ is the estimated misspecification-consistent covariance matrix for $\sqrt{N}(\hat{\theta}_N - \theta^*)$:

$$\hat{\Sigma}_N = H_N^{-1}(\hat{\theta}_N)\mathcal{I}_N(\hat{\theta}_N)H_N^{-1}(\hat{\theta}_N),$$

where $H_N(\theta)$ and $\mathcal{I}_N(\theta)$ are the sample analogs of the hessian and information matrix of the log-likelihood, respectively. Computing the Wald statistic for the misspecified logit model we have $\hat{W} = 102.15$, which corresponds to a marginal significance level of 3.4×10^{-21} given that under the null hypothesis $\hat{W} \Longrightarrow \chi^2(4)$, a Chi-squared random variable with 4 degrees of freedom. The Wald test statistic that the estimated probit parameters equals the true values is $\hat{W} = 2.788$, which corresponds to a marginal significance level of 0.594. Thus we can clearly reject the hypothesis that the logit model is correctly specified, but we do not reject the hypothesis that the probit model is correctly specified. However our ability to compute this statistic requires *prior knowledge the true parameters* θ^*. Of course in nearly all "real" applications we do not know θ^* so this type of Wald test is infeasible.

Later in Econ 551 we will consider general specification tests, such as White's (1982) *Econometrica* Information matrix test statistic (which is not necessarily a consistent test), or Bieren's (1990) *Econometrica* specification test statistic of functional form (which is a consistent test). These allow us to test whether the parametric model $\Psi(x, \theta)$ is correctly specified (i.e. whether there exists a θ^* such that $\Psi(x, \theta^*) = \Pr\{y = 1|x\}$, where $\Pr\{y = 1|x\}$ is the true conditional choice probability) without any prior knowledge of θ^* or, indeed, without any prior information about what the true model really is.

However the estimation results suggest that the power of these "omnibus" specification test statistics may be low, even with samples as large as $N = 3000$. To see how hard it might be to test this hypothesis, consider figure 1 below. For example comparing $-H_N(\hat{\theta}_N)$ and $\mathcal{I}_N(\hat{\theta}_N)$ we find that they are very close in both the probit and logit specifications. Tables 2 and 3 present the estimated values of $-H_N(\hat{\theta}_N)$ and $\mathcal{I}_N(\hat{\theta}_N)$ for the probit and logit specifications, respectively.

Table 2: Estimates of $-H_N(\hat{\theta}_N)$ and $\mathcal{I}_N(\hat{\theta}_N)$ for Probit Model

$$\mathcal{I}_N(\hat{\theta}_N) = \begin{matrix} 1643.540 & -14.797 & 1529.713 & 141.053 \\ -14.797 & 1529.713 & 141.053 & 4344.458 \\ 1529.713 & 141.053 & 4344.458 & 1556.723 \\ 141.053 & 4344.458 & 1556.723 & 20823.807 \end{matrix}$$

$$-H_N(\hat{\theta}_N) = \begin{matrix} 1643.599 & -17.351 & 1516.827 & 134.246 \\ -17.351 & 1516.827 & 134.246 & 4258.982 \\ 1516.827 & 134.246 & 4258.982 & 1998.199 \\ 134.246 & 4258.982 & 1998.199 & 20659.827 \end{matrix}$$

Table 3: Estimates of $-H_N(\hat{\theta}_N)$ and $\mathcal{I}_N(\hat{\theta}_N)$ for Logit Model

$$\mathcal{I}_N(\hat{\theta}_N) = \begin{matrix} 581.582 & -18.152 & 514.050 & 41.983 \\ -18.152 & 514.050 & 41.983 & 1401.355 \\ 514.050 & 41.983 & 1401.355 & 627.199 \\ 41.983 & 1401.355 & 627.199 & 6622.895 \end{matrix}$$

$$-H_N(\hat{\theta}_N) = \begin{matrix} 581.619 & -18.848 & 513.331 & 37.097 \\ -18.848 & 513.331 & 37.097 & 1404.316 \\ 513.331 & 37.097 & 1404.316 & 664.932 \\ 37.097 & 1404.316 & 664.932 & 6756.945 \end{matrix}$$

Figure 1 below plots the true conditional choice probability $\Pr\{y = 1|x\} = \Phi(x\theta^*)$, i.e. the probit model evaluated at the true parameters and at the (sorted) x values in the data file `data3.asc`, the estimated probit and logit models, and the logit model evaluated at the true parameter values θ^*. We see that even thought the estimated parameter values $\hat{\theta}_N$ for the logit and probit models are significantly different from each other, the *estimated choice probabilities are nearly identical for each x in the sample.* Indeed the estimated logit and probit choice probabilities are visually virtually indistinguishable. Maximum likelihood is doing its best (in the presence of noise) to try to fit the true choice probability $\Pr\{y = 1|x\} = \Phi(x\theta^*)$, and we see that both the logit and probit models are sufficiently flexible functional forms that we can approximate the data about equally well with either specification. As a result the maximized value of the log-likelihood is almost identical for both models, i.e. $L_N(\hat{\theta}_N) = -.5751$ for both the probit and logit specifications. Recalling the discussion of neural networks in our presentation of non-parametric estimation methods, both the logit and probit models can be regarded as simplified neural networks with a single hidden unit and the logistic and normal cdf's as "squashing functions." Given that neural networks can approximate a wide variety of functions, it isn't so surprising that the logit and probit choice probabilities can approximate each other very well, with each yielding virtually the same overall fit. Thus, one can imagine it would be very hard for an omnibus specification test statistic to discern which of these models is the true model generating the data.

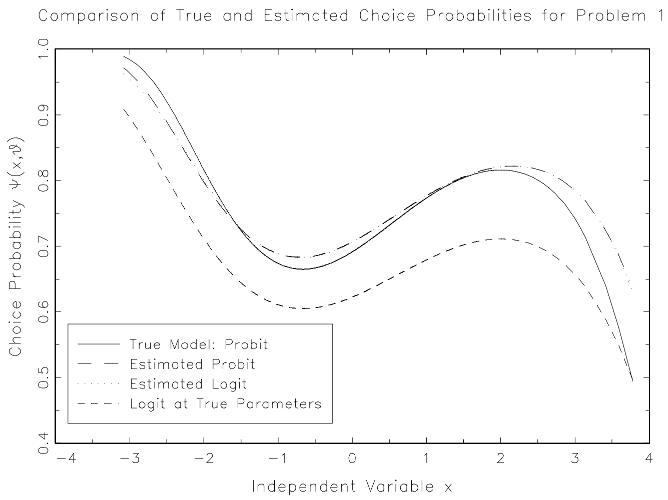

Figure 1: Comparison of True and Estimated Choice Probabilities

Figure 1 also plots the predicted logit choice probabilities that result from evaluating the logit model at the true parameter values θ^*. We can see that in this case the choice probabilities of the logit model are quite different from the choice probabilities of the true probit model. However the logit maximum likelihood estimates are not converging to θ^* when the model is misspecified. Instead, the misspecified maximum likelihood estimator is converging to the parameter vector θ^0 which minimizes the Kullback-Liebler distance between the chosen parametric specification and the true choice probability:

$$\begin{aligned}\theta^0 &= \underset{\theta \in R^4}{argmin} \int_x \int_y \ln\left[\frac{g(y,x)}{f(y,x|\theta)}\right] g(y,x)dydx \\ &= \underset{\theta \in R^4}{argmin} \int_x \left[\ln\left[\frac{\Phi(x\theta^*)}{\Psi(x|\theta)}\right]\Phi(x\theta^*) + \ln\left[\frac{1-\Phi(x\theta^*)}{1-\Psi(x|\theta)}\right](1-\Phi(x\theta^*))\right]\phi(x)dx,\end{aligned}$$

where $\phi(x)$ is the standard normal density, the marginal density of the x variables used to generate the data in this problem. Given the flexibility of the logit specification, we find that $\Psi(x|\theta^0) = \exp\{x\theta^0\}/(1+exp(x\theta^0))$ is almost identical to the true probit specification $\Phi(x\theta^*)$ even though θ^0 and θ^* are fairly different parameter vectors.

Question 4 We can also use nonlinear least squares to consistently estimate θ^*, assuming the specification of the choice probability is correct, since by definition of the conditional choice probability we have:

$$E\{y|x\} = 1 \times \Pr\{y=1|x\} + 0 \times \Pr\{y=0|x\} = \Pr\{y=1|x\} = \Phi(x\theta^*).$$

Thus, $\Phi(x\theta^*)$ is the true conditional expectation function, so even though the dependent variable y only takes on the values $\{0,1\}$ we still have a valid regression equation:

$$y = \Phi(x\theta^*) + \eta,$$

where the error term also takes on two possible values $\{-\Phi(x\theta^*), 1-\Phi(x\theta^*)\}$, but satisfies $E\{\eta|x\} = 0$ by construction. By the general uniform consistency arguments presented in class, it is easy to show that the nonlinear least squares estimator $\hat{\theta}_N$ defined by:

$$\hat{\theta}_N = \underset{\theta \in R^4}{argmin}\, \mathrm{SSR}_N(\theta) \equiv \frac{1}{2N}\sum_{i=1}^{N}\left[y_i - \Psi(x_i|\theta)\right]^2, \tag{4}$$

will be a consistent estimator of θ^* if the model is correctly specified, i.e. if $\Psi = \Phi$, where Φ is the standard normal CDF, but if the choice probability is misspecified, then with probability 1 we have $\hat{\theta}_N \to \theta^0$ where θ^0 is given by:

$$\theta^0 = \underset{\theta \in R^4}{argmin}\, E\left\{[\Psi(x|\theta) - \Pr\{y=1|x\}]^2\right\}.$$

Question 5 Table 4 presents the NLS estimates of θ for the logit and probit specifications of $\Pr\{y=1|x\}$ using the estimation program `estimate.gpr` and the procedures `log_nls.g` and `prb_nls.g` respectively.

Table 4: Nonlinear Least Squares Estimates of θ using `data3.asc`

Parameter	θ_0	θ_1	θ_2	θ_3
True Value	0.500	0.200	0.100	-0.050
Probit NLS	0.542	0.174	0.085	-0.038
Std. Dev.	0.030	0.040	0.020	0.011
Logit NLS	0.876	0.295	0.145	-0.064
Std. Dev.	0.051	0.068	0.034	0.019

Comparing Tables 1 and 4 we see that the MLE and NLS estimates of θ^* are virtually identical for each specification. The standard errors are virtually the same in this case as well. In general the NLS estimator is less efficient than the MLE since the latter attains the Cramer-Rao lower bound when the model is correctly specified. In this case the MLE and NLS estimates happen to be amazingly close to each other, and the standard errors of the NLS estimates are actually minutely smaller than the standard errors of the MLE estimates (for example for the MLE estimator of θ_1 we have $\mathrm{std}(\hat{\theta}_1) = .030680$ whereas for the NLS estimator we have $\mathrm{std}(\hat{\theta}_1) = .030548$). This anomaly is probably not due to a programming error on my part (since running the gradient and hessian check options in `estimate.gpr` reveals that the analytic formulas I programmed match the numerical values quite closely), but probably due to to a combination of roundoff error and estimation noise. Although the Cramer-Rao lower bound holds asymptotically, it need not hold in finite samples for the sample analog estimates of the covariance matrix, which can be potentially quite noisy estimates of the asymptotic covariance matrix. It is straightforward to show that for the NLS has asymptotic covariance matrix Σ_{nls} given by:

$$\Sigma_{\text{nls}} = [-H(\theta^*)]^{-1}\mathcal{I}(\theta^*)[-H(\theta^*)]^{-1} \tag{5}$$

where

$$-H(\theta^*) = E\left\{\frac{\partial}{\partial\theta}\Phi(x\theta^*)\frac{\partial}{\partial\theta'}\Phi(x\theta^*)\right\},$$

and

$$\mathcal{I}(\theta^*) = E\left\{\Phi(x\theta*)[1-\Phi(x\theta^*)]\frac{\partial}{\partial\theta}\Phi(x\theta^*)\frac{\partial}{\partial\theta'}\Phi(x\theta^*)\right\},$$

whereas for the correctly specified probit model the Cramer-Rao lower bound, Σ_{mle}, is given by

$$\Sigma_{\text{mle}} = E\left\{\frac{\frac{\partial}{\partial\theta}\Phi(x\theta^*)\frac{\partial}{\partial\theta'}\Phi(x\theta^*)}{\Phi(x\theta*)[1-\Phi(x\theta^*)]}\right\}^{-1} \tag{6}$$

Thus we have $\Sigma_{\text{nls}} \geq \Sigma_{\text{mle}}$ unless the model is homoscedastic, i.e. when $\sigma^2(x) = \Phi(x\theta^*)[1-\Phi(x\theta^*)] = \sigma^2$ for all x, which implies that $\Phi(x\theta^*)$ is a constant for all x, which is almost never the case in any "interesting" application. We conclude that the MLE estimator of θ^* is necessarily more efficient than the NLS estimator, and the only reason why the NLS has slightly smaller estimated standard errors in this example is due to round-off and estimation error. For other sample sizes, say $N = 500$, we do find that the estimated standard deviations of the MLE estimator are smaller than the NLS estimator. For example when $N = 500$ the NLS estimator of $\theta^*2 = 0.1$ is $\hat{\theta}_1 = 0.095360$ and its standard error is $\text{std}(\hat{\theta}_2) = 0.102891$, whereas the MLE estimator is $\hat{\theta}_2 = 0.096739$ and its standard error is $\text{std}(\hat{\theta}_2) = 0.094864$. Thus, while we do see an efficiency gain to doing maximum likelihood, it is far from overwhelming in this particular case.

Question 6 It is easy to see that the errors $\{\eta_i\}$ in the regression formulation of the binary choice model, $y_i = \Psi(x_i|\theta^*) + \eta_i$ are heteroscedastic with conditional variance $\sigma^2(x_i)$ given by:

$$\sigma^2(x_i) = \Psi(x_i|\theta^*)[1-\Psi(x_i|\theta^*)],$$

(Too see this, note that the conditional variance of η_i and y_i given x_i are the same, and the latter is a Bernoulli random variable that takes on the value $y_i = 1$ with probability $p = \Phi(x_i|\theta^*)$. As is well known, a Bernoulli random variable has variance $p(1-p)$). Thus, we have a case where heteroscedasticity has a known functional form and we can make use of it to compute feasible generalized least squares (FGLS) estimates of θ^*. In the first stage we compute the NLS estimates of θ^* and using these estimates, call them $\hat{\theta}_N^1$, we compute estimated conditional variance $\hat{\sigma}^2(x_i)$ given by the formula above but with the first stage NLS estimates θ_N^1 in place of θ^*. Then in the second stage we compute the FGLS estimates $\hat{\theta}_N^2$ as the solution to the following weighted least squares problem:

$$\hat{\theta}_N = \underset{\theta \in R^4}{argmin}\, \mathrm{WSSR}_N(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{[y_i - \Psi(x_i|\theta)]^2}{2\Psi(x_i|\theta_N^1)[1-\Psi(x_i|\theta_N^1)]} \tag{7}$$

The FGLS estimates of θ^*, computed by `log_fgls` and `prb_fgls.g` in the logit and probit cases, respectively, are virtually identical to the NLS estimates of θ^*, which are in turn virtually identical to the maximum likelihood estimates in the logit and probit specifications presented in Table 1 so I didn't bother to present them here.

Should we conclude from this that there isn't much heteroscedasticity in this problem? Figure 2 below plots $\sigma^2(x)$ for this problem and we see that there is indeed substantial heteroscedasticity, with fairly large variation in the effective weighting of the observations. However by plotting the relative contribution of the terms in the weighted and unweighted sum of squared residuals, you will find that except for a small number of observations with the lowest values of x_i which the FGLS estimator assigns very high weights to, the relative sizes of the vast majority of the true squared residuals in both the FGLS and NLS estimators are very similar. This explains why the FGLS and NLS estimators are not very different even though there appears to be substantial heteroscedasticity in this problem.

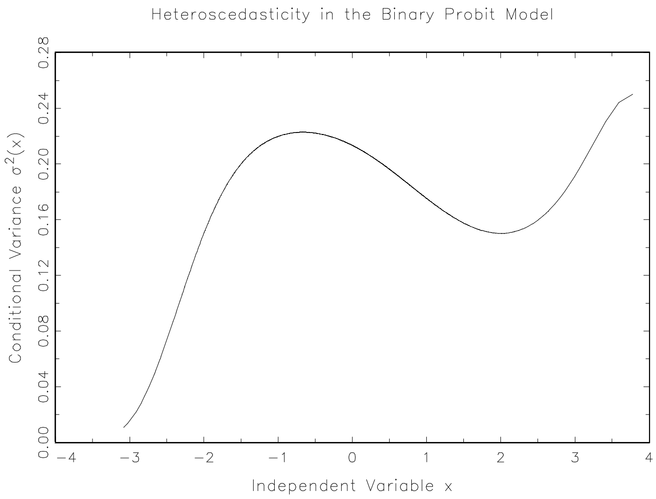

Figure 2: Conditional Heteroscedasticity in the Probit Model

Question 7 The FGLS estimator is asymptotically equivalent to the maximum likelihood estimator, a result suggested by the fact that the likelihood function and the weighted sum of squared residuals happen to have the same first order conditions:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} L_N(\theta) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{I\{y_i=1\}}{\Psi(x_i|\theta)} - \frac{I\{y_i=0\}}{1-\Psi(x_i|\theta)}\right]\frac{\partial}{\partial\theta}\Psi(x_i|\theta) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i-\Psi(x_i|\theta)}{\Psi(x_i|\theta)[1-\Psi(x_i|\theta)]}\frac{\partial}{\partial\theta}\Psi(x_i|\theta) \\ &= \frac{\partial}{\partial\theta}\mathrm{WSSR}_N(\theta). \end{aligned}$$

Actually, the first order conditions are only identical for the *continuously updated* version of the FGLS estimator, where instead of using a first stage NLS estimate $\hat{\theta}_N^1$ to make an estimated correction for heteroscedasticity, we continually update our estimate of the heteroscedasticity as θ changes, so the same θ appears in the numerator and denominator terms in the third equation above whereas in the FGLS estimator $\hat{\theta}_N^1$ appears in the denominator terms. However recalling the logic of the "Amemiya correction" we need to consider whether it is necessary to account for the estimation noise in the first stage estimates $\hat{\theta}_N^1$ in deriving the asymptotic distribution of the FGLS estimator, $\hat{\theta}_N^2$. It will turn out that there is a form of "block diagonality" here which enables the FGLS estimator to be "adaptive" in the sense that the asymptotic distribution of the FGLS estimator $\hat{\theta}_N^2$ does not depend on whether we use the noisy first stage NLS estimator to compute a noisy estimate of the conditional variance $\hat{\sigma}^2(x) = \Phi(x\hat{\theta}_N^1)[1-\Phi(x\hat{\theta}_N^1)]$ to use as weights, or if we use the true conditional variance $\sigma^2(x) = \Phi(x\theta^*)[1-\Phi(x\theta^*)]$.

Before we show this, we first show that if we did use the true conditional variance as the weights in the FGLS estimator, it would be as efficient as maximum likelihood: i.e. the FGLS estimator attains the Cramer-Rao lower bound. To see this do a Taylor-series expansion of the first order condition for the FGLS estimator about θ^*:

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i-\Phi(x_i\hat{\theta}_N^2)}{\Phi(x_i\theta^*)[1-\Phi(x_i\theta^*)]}\frac{\partial}{\partial\theta}\Phi(x_i\hat{\theta}_N^2) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i-\Phi(x_i\theta^*)}{\Phi(x_i\theta^*)[1-\Phi(x_i\theta^*)]}\frac{\partial}{\partial\theta}\Phi(x_i\theta^*) + H_N(\tilde{\theta}_N^2)[\hat{\theta}-\theta^*] \qquad (8) \end{aligned}$$

where $\tilde{\theta}_N^2$ is on the line segment between $\hat{\theta}_N^2$ and θ^* and

$$H_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{-\partial\Phi(x_i\theta)/\partial\theta \; \partial\Phi(x_i\theta)/\partial\theta' + [y_i - \Phi(x_i\theta)]\partial^2\Phi(x_i\theta)/\partial\theta\partial\theta'}{\Phi(x_i\theta^*)[1-\Phi(x_i\theta^*)]}. \tag{9}$$

By the uniform Strong Law of Large Numbers, we have that $\|H_N(\theta) - H(\theta)\| \to 0$ with probability 1 where

$$H(\theta) = E\left\{\frac{-\partial\Phi(x\theta)/\partial\theta \; \partial\Phi(x\theta)/\partial\theta' + [y - \Phi(x\theta)]\partial^2\Phi(x\theta)/\partial\theta\partial\theta'}{\Phi(x\theta^*)[1-\Phi(x\theta^*)]}\right\}. \tag{10}$$

Since $\tilde{\theta}_N^2 \to \theta^*$ with probability 1, it follows that $H_N(\tilde{\theta}_N^2) \to H(\theta^*)$ with probability 1. Using the law of iterated expectations we can show that the second term in the above expectation is zero when $\theta = \theta^*$ so that

$$H(\theta^*) = E\left\{\frac{-\partial\Phi(x\theta^*)/\partial\theta \; \partial\Phi(x\theta^*)/\partial\theta'}{\Phi(x\theta^*)[1-\Phi(x\theta^*)]}\right\}. \tag{11}$$

The Central Limit Theorem implies that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{y_i - \Psi(x_i|\theta^*)}{\Psi(x_i|\theta^*)[1-\Psi(x_i|\theta^*)]}\frac{\partial}{\partial\theta}\Psi(x_i|\theta^*) \Longrightarrow N(0, \mathcal{I}(\theta^*)), \tag{12}$$

where it is easy to see that $\mathcal{I}(\theta^*) = -H(\theta^*)$. Combining all results in equations $(8), \ldots, (12)$ we see that the asymptotic distribution of the FGLS estimator is given by:

$$\sqrt{N}(\hat{\theta}_N^2 - \theta^*) \Longrightarrow N\left(0, [-H(\theta^*)]^{-1}\mathcal{I}(\theta^*)[-H(\theta^*)]^{-1}\right) = N\left(0, \mathcal{I}^{-1}(\theta^*)\right). \tag{13}$$

Thus the asymptotic covariance matrix of the FGLS estimator is the inverse of the information matrix (see equation 6), so it is asymptotically efficient.

Now we need to show that if we computed the FGLS estimator using the (inverse of) the estimated conditional variance $\hat{\sigma}^2(x) = \Phi(x\hat{\theta}_N^1)[1 - \Phi(x\hat{\theta}_N^1)]$ instead of the true conditional variance as weights, the asymptotic distribution is still the same as that given in (13) above. We do this using the same logic as for the general derivation of the "'Amemiya correction", Taylor expanding the FGLS FOC in both variables $\hat{\theta}_N^1$ and $\hat{\theta}_N^2$ about their limiting value θ^*. That is, if we define the function $F_N(\alpha, \beta)$ by

$$F_N(\alpha,\beta) = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i - \Phi(x_i\beta)}{\Phi(x_i\alpha)[1-\Phi(x_i\alpha)]}\frac{\partial}{\partial\beta}\Phi(x_i\beta) \tag{14}$$

then we have the following joint Taylor series expansion for $F_N(\hat{\theta}_N^1, \hat{\theta}_N^2)$ about $F_N(\theta^*, \theta^*)$

$$F_N(\hat{\theta}_N^1, \hat{\theta}_N^2) = F_N(\theta^*, \theta^*) + \frac{\partial}{\partial\alpha}F_N(\tilde{\theta}_N^1, \tilde{\theta}_N^2)(\hat{\theta}_N^1 - \theta^*) + \frac{\partial}{\partial\beta}F_N(\tilde{\theta}_N^1, \tilde{\theta}_N^2)(\hat{\theta}_N^2 - \theta^*) \tag{15}$$

We know that the NLS estimator is asymptotically normal, so $\sqrt{N}[\hat{\theta}_N^1 - \theta^*] = O_p(1)$, i.e. it is bounded in probability. Thus, the FGLS estimator that uses estimated conditional variance as weights will have the same asymptotic distribution as the (infeasible) FGLS estimator that of the true conditional variance as weights if we can show that with probability 1 we have:

$$\begin{aligned} \frac{\partial}{\partial \alpha} F_N(\tilde{\theta}_N^1, \tilde{\theta}_N^2) &\rightarrow 0 \\ \frac{\partial}{\partial \beta} F_N(\tilde{\theta}_N^1, \tilde{\theta}_N^2) &\rightarrow H(\theta^*) \end{aligned}$$

But this follows from the USLLN and the consistency of $\hat{\theta}_N^1$ and $\hat{\theta}_N^2$.

Question 8 Figure 3 presents a comparison of the true choice probability and nonparametric estimates of this probability using both kernel and series estimators from the program `kernel.gpr`.

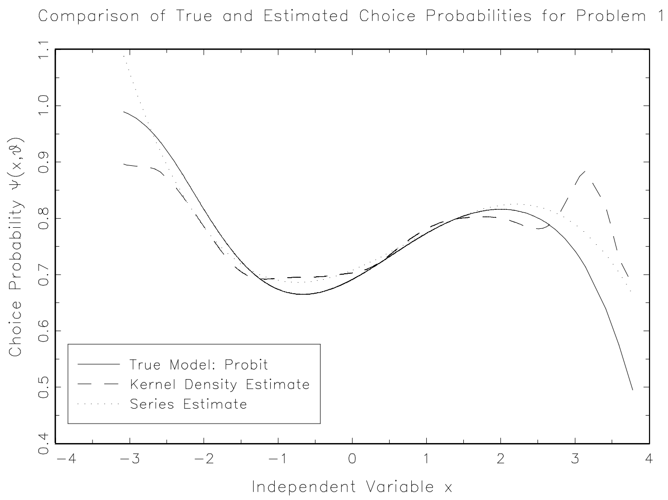

Figure 3: True vs. Nonparametric Estimates of Choice Probabilities

The series estimator seems to provide a better estimator of the true choice probability than the kernel density estimator in this case. The series estimator is just the predicted $\hat{y}_i$ value from a simple OLS regression of the $\{y_i\}$ on a constant and the first 3 powers of x_i:

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \theta_3 x_i^3 + \eta_i$$

and the kernel estimator is the standard Nadaraya-Watson estimator

$$\hat{\Psi}(x) = \frac{\sum_{i=1}^N K_h(x_i - x)\, y_i}{\sum_{i=1}^N K_h(x_i - x)},$$

where $K_h(x_i - x) = \frac{1}{h} K\left(\frac{x_i - x}{h}\right)$ and $K(\cdot)$ is defined to be a Gaussian density function. For the choice of a bandwidth parameter, h, a rule of thumb is used: $\widehat{h} = c * std(x) * n^{-\frac{1}{5}}$, with $c \in [0.5, 2.5]$. In this case the automatically chosen bandwith turned out to be $\hat{h} = .3037$. The series estimator is much faster to compute than the kernel estimator, since the above summations must be carried out for each of the $N = 3000$ observations in the sample in order to plot the estimated choice probability for each observation. Comparing the fit of the parametric and nonparametric models in figures 1 and 2, we see that even though the logit and probit models are "parametric", they have sufficient flexibility to enable them to provide a better fit than either the kernel density or series estimators. This conclusion is obviously specific to this example where the true conditional choice probability was generated by a probit model, and as we saw from figure 1, one can adjust the parameters to make the predicted probabilities of the logit and probit models quite close to each other.

Figure 4 plots the estimated choice probabilities produced by both the probit and logit maximum likelihood estimates and the kernel and series nonparametric estimates. We see that except for the "hump" in the kernel density estimate, all the estimates are very close to each other. It would appear to be quite difficult to say which estimate was the "correct" one: instead we conclude that 4 different ways of estimating the conditional choice probability give approximately the same results.

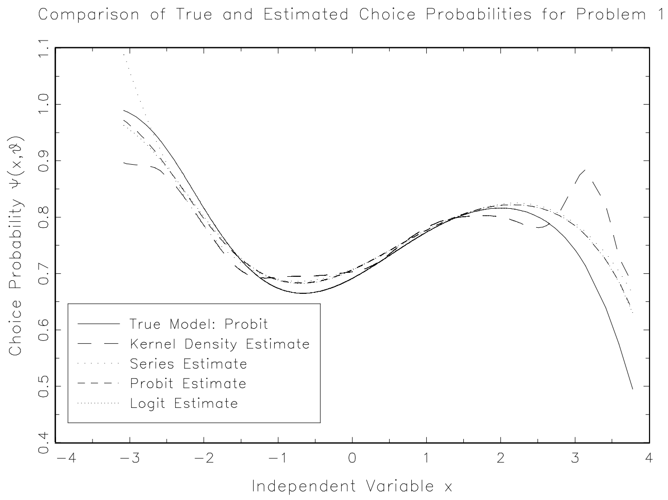

Figure 4: True vs. Nonparametric Estimates of Choice Probabilities