These notes come from Chapters 22 and 25 of the textbook and provide an introduction to nonlinear models, particularly binary outcome models.

Binary Choice Models

H: 25.1, H 25.2

So far this semester, the examples that we have considered have all been for the case where the outcome is continuous. Now, let's consider the case where the outcome is binary; that is, $Y \in \{0, 1\}$. Typically, in this case, one would be interested in either

- The **response probability** of Y conditional on $X = x$, that is, $\mathrm{P}(x) := \mathrm{P}(Y = 1|X = x)$
- The **partial effect** or **marginal effect**:

$$\frac{\partial \mathrm{P}(x)}{\partial x_1}$$

which is the the partial effect with respect to the first element of x. As we did earlier in the semester, one could also consider partial effects of discrete and/or binary regressors. Notice that, in general, the partial effect depends on the values of all the regressors.

- The **average partial effect**

$$\mathrm{E}\left[\frac{\partial \mathrm{P}(X)}{\partial x_1}\right]$$

The average partial effect, well, averages the partial effects across the distribution of the covariates. This is a single number.

As earlier in the semester, one motivation for thinking about partial effects is that, under unconfoundedness conditions, they correspond to causal effects — these arguments continue to go through in this case.

Another thing that is worth pointing out: when Y is binary,

$$\begin{aligned}
\mathrm{E}[Y|X = x] &= \sum_{y \in \{0,1\}} y\mathrm{P}(Y = y|X = x) \\
&= 0 \times \mathrm{P}(Y = 0|X = x) + 1 \times \mathrm{P}(Y = 1|X = x) \\
&= \mathrm{P}(Y = 1|X = x)
\end{aligned}$$

where the first equality holds from the definition of expectation when Y is discrete. This means that the response probability is equal to the conditional expectation.

Models for the Response Probability

H: 25.3

One common way to estimate $P(x)$ is by imposing/assuming a **linear probability model**. That is, $P(x) = \mathrm{E}[Y|X = x] = x'\beta$. One can estimate β by just running a regression of Y on X. This is exactly the same as we have done many times before (and just amounts to essentially ignoring that Y is binary). This is quite common in applications.

One advantage of this approach is that it is very simple. In this case, unless there are interactions and/or higher order terms, the partial effect of X_1 is β_1 which does not vary across different values of the regressors.

A main drawback of this approach is that the linear probability model does not respect that probabilities must be between 0 and 1. The textbook gives the example of the probability of being married conditional on age. When this is estimated with CPS data, the estimated probability of being married is increasing in age and greater than 1 for ages over 67. A related drawback is that the partial effects are constant (e.g., the partial effect of X_1 is equal to β_1 regardless of the values of the other covariates). In many applications, this is unrealistic. In the same application as above, it seems unlikely that the partial effect of age on marital status is likely to be much different at 25 than at 65.

A next class of models are **single-index models** where one would impose/assume that $P(x) = G(x'\beta)$ where G is called a **link function** and $x'\beta$ is a **linear index**. The link function is chosen to be a cdf so that $0 \leq G(u) \leq 1$ for all possible u. This means that $P(x)$ cannot be outside of $[0, 1]$.

In this case, partial effects are given by

$$\frac{\partial \mathrm{P}(x)}{\partial x_1} = g(x'\beta)\beta_1$$

which follows from the chain rule and where g is the derivative of G. Notice that the partial effects depend on the covariates in this case.

The two most common single-index models are

- **Probit** — in this case, $\mathrm{P}(x) = \Phi(x'\beta)$ where Φ is the cdf of a standard normal random variable
- **Logit** — in this case, $\mathrm{P}(x) = \Lambda(x'\beta)$ where $\Lambda(u) = \frac{\exp(u)}{1+\exp(u)}$ which is the logistic cdf.

Maximum Likelihood Estimation

Maximum likelihood estimation is a major class of estimation strategies. In general, these involve imposing/assuming parametric models that completely specify the (conditional) pdf of the data. Often, maximum likelihood estimation involves relatively strong assumptions, but typically results in efficient estimators. An example is the normal regression model that we skipped in chapter 5 of the textbook.

The **likelihood** is the joint density of the observed data viewed as a function of the parameters. The maximum likelihood estimator is the value which maximizes the likelihood function. The **likelihood function** is given by

$$L_n(\theta) = f(Y_1, \ldots, Y_n | X_1, \ldots, X_n; \theta) = \Pi_{i=1}^n f(Y_i | X_i; \theta)$$

It is common to work with the **log-likelihood function** which is just the log of the previous likelihood function and is given by

$$l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^n \log(f(Y_i|X_i;\theta))$$

For the notes below, I'll focus on probit, but the arguments are very similar for logit. Because Y is binary note that its conditional pmf can be written as

$$f(y|x) = \mathrm{P}(Y=1|X=x)^y (1 - \mathrm{P}(Y=1|X=x))^{(1-y)} \qquad \text{for } y \in \{0,1\}$$

This is just the pmf for a Bernoulli random variable — and, notice that, when $y=1$, $f(1|x) = \mathrm{P}(Y=1|X=x)$, and when $y=0$, $f(0|x) = 1 - \mathrm{P}(Y=1|X=x)$. Now, plugging in the probit model, we have that

$$f(y|x;\beta) = \Phi(x'\beta)^y (1-\Phi(x'\beta))^{(1-y)} \qquad \text{for } y \in \{0,1\}$$

Using this expression, we have that

$$\begin{aligned}
\hat{\beta} &= \underset{b}{\text{argmax}}\ l_n(b) \\
&= \underset{b}{\text{argmax}}\ \sum_{i=1}^n \log\left(\Phi(X_i'b)^{Y_i}(1-\Phi(X_i'b))^{(1-Y_i)}\right) \\
&= \underset{b}{\text{argmax}}\ \sum_{i=1}^n Y_i \log(\Phi(X_i'b)) + (1-Y_i)\log(1-\Phi(X_i'b))
\end{aligned}$$

Unlike the regression estimators that we have talked about this semester, this is not a problem that has an explicit solution. That said, this turns out to be an easy problem for the computer to solve.

To estimate a probit model in R, you can run the following sort of code `glm(Y~X, family=binomial(link="probit"))`. If your professor asks you to write the code manually for this, then you need to actually get the computer to maximize the above function. In this case, a helpful function is the `optim` function.

For optimizing this kind of function (and conducting inference), it is often helpful to know the vector of first derivatives of the log likelihood function (the "score"). For probit, these are given by

$$
\begin{aligned}
S_n(b) &:= \frac{\partial l_n(b)}{\partial b} \\
&= \sum_{i=1}^{n} Y_i \frac{\phi(X_i'b)}{\Phi(X_i'b)} X_i - (1 - Y_i) \frac{\phi(X_i'b)}{1 - \Phi(X_i'b)} X_i \\
&= \sum_{i=1}^{n} \frac{(Y_i - \Phi(X_i'b))\phi(X_i'b)}{\Phi(X_i'b)(1 - \Phi(X_i'b))} X_i \qquad (1)
\end{aligned}
$$

M-estimators

H 22.1 - H 22.6

Next, we'd like to establish the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$ for probit. To do this, I am going to follow the traditional approach (well, at least this is what it does in the book and what my professor taught me in graduate school) of considering the more general class of M-estimators.

For this part, I am just going to sketch the argument showing why M-estimators are asymptotically normal. The textbook also seems to slightly change notation at this point. For this section, θ is "generic" and usually an argument to a function while θ_0 is the population parameter that we are interested in estimating.

$$
\begin{aligned}
\hat{\theta} &= \underset{\theta \in \Theta}{\operatorname{argmin}}\ M_n(\theta) \\
M_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta)
\end{aligned}
$$

where $M_n(\theta)$ is called the objective function. For maximum likelihood estimation, $\rho(Y_i, X_i, \theta) = -\log f(Y_i | X_i, \theta)$. Likewise, the population parameter θ_0 solves

$$
\begin{aligned}
\theta_0 &= \underset{\theta \in \Theta}{\operatorname{argmin}}\ M(\theta) \\
M(\theta) &= \mathrm{E}[\rho(Y, X, \theta)]
\end{aligned}
$$

Following the textbook, define the following notation

$$
\begin{aligned}
\psi(Y, X, \theta) &:= \frac{\partial \rho(Y, X, \theta)}{\partial \theta} \\
\bar{\psi}_n(\theta) &:= \frac{\partial M_n(\theta)}{\partial \theta} \\
\psi(\theta) &:= \frac{\partial M(\theta)}{\partial \theta}
\end{aligned}
$$

The first order condition for $\hat{\theta}$ that minimizes $M_n(\theta)$ is

$$
0 = \bar{\psi}_n(\hat{\theta})
$$

This is just like solving for $\hat{\beta}$ in the context of regression; however, here it would typically be the case that we cannot come up with an explicit solution for $\hat{\theta}$. From a mean value theorem type of argument (similar to what we have used before in the context of the delta method) where $\bar{\theta}$ is between $\hat{\theta}$ and θ_0,

$$\begin{aligned}\bar{\psi}_n(\hat{\theta}) &= \bar{\psi}_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'}M_n(\bar{\theta})(\hat{\theta}-\theta_0) \\ &= \bar{\psi}_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'}M_n(\theta_0)(\hat{\theta}-\theta_0) + o_p(n^{-1/2})\end{aligned}$$

Multiplying by $\sqrt{n}$ and re-arranging implies that

$$\sqrt{n}(\hat{\theta}-\theta_0) = \left(\frac{\partial^2}{\partial\theta\partial\theta'}M_n(\theta_0)\right)^{-1}\sqrt{n}\bar{\psi}_n(\theta_0)$$

Next, notice that

$$\frac{\partial^2}{\partial\theta\partial\theta'}M_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\rho(Y_i,X_i,\theta_0) \xrightarrow{p} \mathrm{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\rho(Y,X,\theta_0)\right] =: Q$$

and that

$$\sqrt{n}\bar{\psi}_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Y_i,X_i,\theta_0) \tag{2}$$

and further recalling that, since θ_0 minimize $M(\theta) = \mathrm{E}[\rho(\theta)]$, the first order condition for this minimization problem is $0 = \psi(\theta_0) = \mathrm{E}[\psi(Y,X,\theta_0)]$. Thus, we apply the CLT to the term in Equation 2, and it converges to $N(0,\Omega)$ where $\Omega = \mathrm{E}[\psi(Y,X,\theta_0)\psi(Y,X,\theta_0)']$. Thus,

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,V)$$

where $V = Q^{-1}\Omega Q^{-1}$.

Now, returning to probit, the previous results hold; we just need to figure out expressions for Ω and Q. Using essentially the same arguments as around Equation 1, it follows that

$$\psi(Y,X,b) = -\frac{(Y-\Phi(X'b))\phi(X'b)}{\Phi(X'b)(1-\Phi(X'b))}X_i$$

so that

$$\begin{aligned}
\Omega &= \mathrm{E}\left[\left(\frac{(Y-\Phi(X'\beta))\phi(X'\beta)}{\Phi(X'\beta)(1-\Phi(X'\beta))}\right)^2 XX'\right] \\
&= \mathrm{E}\left[\frac{(Y-2Y\Phi(X'\beta)+\Phi(X'\beta)^2)\phi(X'\beta)^2}{\Phi(X'\beta)^2(1-\Phi(X'\beta))^2} XX'\right] \\
&= \mathrm{E}\left[\frac{(\Phi(X'\beta)-\Phi(X'\beta)^2)\phi(X'\beta)^2}{\Phi(X'\beta)^2(1-\Phi(X'\beta))^2} XX'\right] \\
&= \mathrm{E}\left[\frac{\phi(X'\beta)^2}{\Phi(X'\beta)(1-\Phi(X'\beta))} XX'\right]
\end{aligned}$$

where the second equality holds mainly because $Y = Y^2$ due to Y being binary, the third equality by the law of iterated expectations and canceling terms, the fourth equality by factoring and canceling in the numerator and denominator.

Next, for Q, you can show (this is tedious but not otherwise complicated) that

$$Q = -\mathrm{E}\left[\frac{\phi(X'\beta)^2}{\Phi(X'\beta)(1-\Phi(X'\beta))} XX'\right]$$

and notice that $Q = -\Omega$. Thus, $V = (-\Omega)^{-1}\Omega(-\Omega)^{-1} = \Omega = \mathrm{E}\left[\frac{\phi(X'\beta)^2}{\Phi(X'\beta)(1-\Phi(X'\beta))} XX'\right]$, which you can estimate by

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi(X_i'\hat{\beta})^2}{\Phi(X_i'\hat{\beta})(1-\Phi(X_i'\hat{\beta}))} X_i X_i'$$