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 **marginal contrast** (or **partial effect** or **marginal effect**):

$$\mathrm{MC}_1 := \frac{\partial \mathrm{P}(x)}{\partial x_1}$$

  which is the marginal contrast with respect to the first element of x (i.e., how much higher the probability of $Y = 1$ is for a 1 unit increase in X_1, holding all other covariates fixed). As we did earlier in the semester, one could also consider marginal contrasts for discrete and/or binary regressors. Notice that, in general, the marginal contrast depends on the values of all the regressors. You could report this for particular values of the regressors of interest; alternatively, you might want to aggregate this into something lower-dimensional (and, therefore, easier to report).

- The **average marginal contrast**

$$\mathrm{AMC}_1 := \mathbb{E}\left[\frac{\partial \mathrm{P}(X)}{\partial x_1}\right]$$

  The average marginal contrast, well, averages the marginal contrasts across the distribution of the covariates. This is a single number. The R package `margins` is useful for computing marginal contrasts.

As earlier in the semester, one motivation for thinking about marginal contrasts 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,

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

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 $\mathrm{P}(x)$ is by imposing/assuming a **linear probability model**. That is, $\mathrm{P}(x) = \mathbb{E}[Y|X = x] = x'\beta$. 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 marginal contrast for 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 – which does not make any sense. A related drawback is that the marginal contrasts are constant (e.g., the marginal contrast for 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 that the marginal contrast for age on marital status is likely to be much different at 25 than at 65. And, just to be clear, this is a general "issue" here; in most applications, you would think that marginal contrasts are likely to be smaller at values of the covariates where $\mathrm{P}(x)$ is close to 0 or 1 than when $\mathrm{P}(x)$ is, say, close to 0.5.

A next class of models are **single-index models** where one would impose/assume that $\mathrm{P}(x) = G(x'\beta)$ 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 $\mathrm{P}(x)$ cannot be outside of $[0, 1]$.

In this case, the marginal contrast and average marginal contrast of X_1 are given by

$$\mathrm{MC}_1 = g(x'\beta)\beta_1 \quad \text{and} \quad \mathrm{AMC}_1 = \mathbb{E}[g(X'\beta)\beta_1]$$

respectively, where the expression for the marginal contrast follows from the chain rule and where g is the derivative of G. Notice that the marginal contrasts 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, and $\frac{\partial \mathrm{P}(x)}{\partial x_1} = \phi(x'\beta)\beta_1$ where ϕ is the pdf 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, and $\frac{\partial \mathrm{P}(x)}{\partial x_1} = \lambda(x'\beta)\beta_1$ where $\lambda(u) = \frac{\exp(u)}{(1+\exp(u))^2}$.

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.

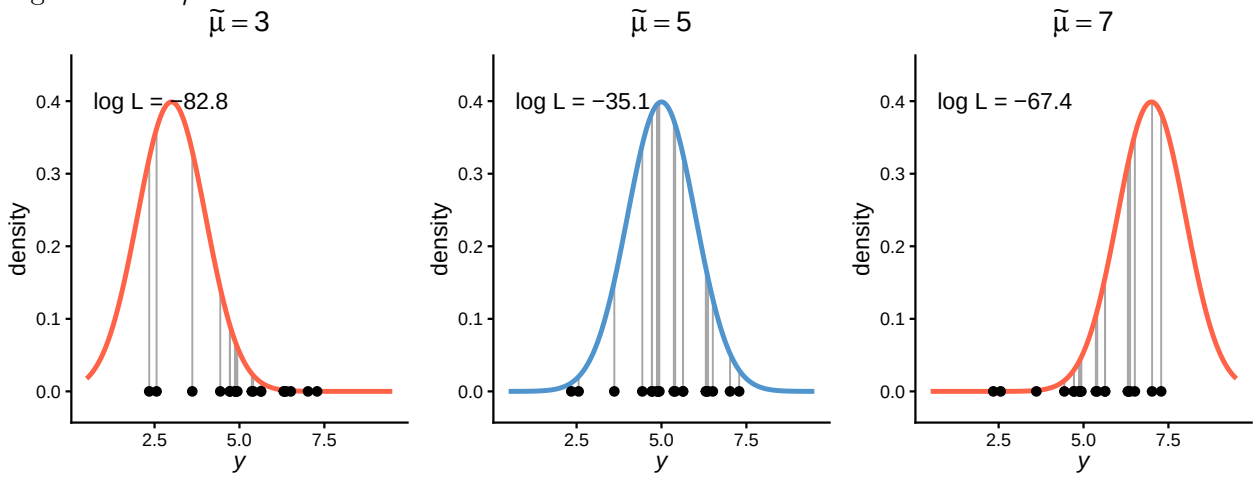
The **likelihood** is the joint density of the observed data viewed as a function of the parameters. The maximum likelihood estimator is the value of the parameter which maximizes the likelihood function. The **likelihood function** is given by

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

where the second equality holds under iid data.

Example: Maximum Likelihood for Normal

Before proceeding, let me give a concrete example. Suppose that we (somehow) know that $Y \sim \mathcal{N}(\mu, 1)$. We don't know what μ is, but we (somehow) know that it is either equal to 3, 5, or 7; we'll denote the candidates for μ by $\tilde{\mu} \in \{3, 5, 7\}$. Finally, suppose we have 20 observations of Y. The idea of maximum likelihood is to find which possible value of $\tilde{\mu}$ makes the observed data most likely. The figure below plots 20 data points drawn from $\mathcal{N}(5, 1)$ along with the $\mathcal{N}(\tilde{\mu}, 1)$ density for our three candidate values of $\tilde{\mu}$. The vertical segments show the density $f(y_i \mid \tilde{\mu})$ at each observation—these are the individual likelihood contributions. You can see the log-likelihood is highest when $\tilde{\mu} = 5$.



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

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

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

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

Using this expression, we have that

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

In other words, we are aiming to estimate β in the probit model by choosing the value of b that maximizes the likelihood of observing the data that we have conditional on that value of the parameter.

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

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

Once you have figured out the likelihood function and the score, estimation is just a matter of getting the computer to maximize the likelihood function (or equivalently, solve for the root of the

score function). Next, we will talk about how to conduct inference for $\hat{\beta}$.

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. M-estimators are estimators that come from minimizing some function. So this class of functions includes the least squares regression estimators that we have talked about extensively. Noting that you can maximize a function by minimizing the negative of the function, the maximum likelihood estimators we have been talking about above also fit into this class.

For this part, I am just going to sketch the argument showing why M-estimators are asymptotically normal. [As a side-comment, the textbook also seems to slightly change notation at this point. I'm going to follow the discussion in the textbook below, but try to use notation similar to what we have used throughout the semester (as well as bit less overall notation)].

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

For maximum likelihood estimation, $\rho(Y_i, X_i, \tilde{\theta}) = -\log f(Y_i | X_i; \tilde{\theta})$. Likewise, the population parameter θ solves

$$\theta = \underset{\tilde{\theta} \in \Theta}{\operatorname{argmin}} \mathbb{E}[\rho(Y, X, \tilde{\theta})]$$

Following the textbook, define the following notation

$$\psi(Y, X, \tilde{\theta}) := \frac{\partial \rho(Y, X, \tilde{\theta})}{\partial \tilde{\theta}}$$

which is a $k \times 1$ vector (where k is the dimension of θ). Thus, the first order condition for $\hat{\theta}$ that minimizes the sample objective function is

$$0 = \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta})$$

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,

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \theta) + \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi(Y_i, X_i, \bar{\theta})}{\partial \tilde{\theta}'} \right) \sqrt{n}(\hat{\theta} - \theta)$$

where $\frac{\partial \psi(Y_i, X_i, \bar{\theta})}{\partial \tilde{\theta}'}$ is a $k \times k$ matrix. Re-arranging terms implies that

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &= - \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi(Y_i, X_i, \bar{\theta})}{\partial \tilde{\theta}'} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \theta) \\ &= -\mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \theta) + o_p(1) \end{aligned} \tag{2}$$

where $\mathbf{Q} := \mathbb{E}\left[\frac{\partial \psi(Y, X, \theta)}{\partial \tilde{\theta}'} \right]$; to see the second equality, you can show that $\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi(Y_i, X_i, \bar{\theta})}{\partial \tilde{\theta}'} \xrightarrow{p} \mathbf{Q}$ (this is not surprising since $\bar{\theta}$ is between $\hat{\theta}$ and θ, but there are some technical details omitted here on uniform convergence of this term; these complications arise because it is an average of a function (Y_i, X_i) and $\bar{\theta}$). Next, since θ minimizes $\mathbb{E}[\rho(\tilde{\theta})]$, it follows that the first order condition for the population objective function is $0 = \mathbb{E}[\psi(Y, X, \theta)]$. Thus, the $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \theta)$ term in Equation 2 is a sum of iid terms that have mean 0, and we can therefore apply the CLT. It converges to $\mathcal{N}(0, \mathbf{\Omega})$ where $\mathbf{\Omega} := \mathbb{E}[\psi(Y, X, \theta)\psi(Y, X, \theta)']$. Thus,

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

where $\mathbf{V} = \mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}$ (note that the negative in front of $\mathbf{Q}$ in Equation 2 cancels in the expression for the asymptotic variance).

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

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

where the negative sign arises because we are maximizing the likelihood (to fit into the theory of

M-estimation above, we minimize the negative of the likelihood function). Thus,

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

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 then recalling that $\mathbb{E}[Y|X] = \Phi(X'\beta)$) and canceling terms, the fourth equality by factoring and canceling in the numerator and denominator.

Next, for $\mathbf{Q}$, you can show that

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

Although it seems like it would be quite tedious to calculate $\mathbf{Q}$ (because $\psi(Y, X, b)$ is highly nonlinear and b shows up in four places), it is very helpful to notice that the expected value of most of the terms that arise from the derivative are equal to 0 which holds by the law of iterated expectations and because $\mathbb{E}[Y|X] = \Phi(X'\beta)$ (this means that the only derivative term that ends up being non-zero is the first one in the numerator and that is where the expression for $\mathbf{Q}$ comes from). Next, notice that $\mathbf{Q} = \mathbf{\Omega}$. Thus, $\mathbf{V} = (\mathbf{\Omega})^{-1}\mathbf{\Omega}(\mathbf{\Omega})^{-1} = \mathbf{\Omega}^{-1} = \mathbb{E}\left[\frac{\phi(X'\beta)^2}{\Phi(X'\beta)(1-\Phi(X'\beta))}XX'\right]^{-1}$. That $\mathbf{Q} = \mathbf{\Omega}$ is actually a property of correctly specified maximum likelihood estimators; it is called information matrix equality.

To conduct inference, you can estimate $\mathbf{\Omega}$ by

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

and use this to construct $\hat{\mathbf{V}} = \hat{\mathbf{\Omega}}^{-1}$, and use this to get standard errors, t-statistics, confidence intervals, Wald statistics, etc. in the usual way.

Estimating Poisson Regression Model with MLE

In Poisson regression, we model count outcomes $Y_i \in \{0, 1, 2, ...\}$ by assuming

$$Y_i \mid X_i \sim \text{Poisson}(\lambda_i), \qquad \lambda_i = \exp(X_i'\beta)$$

The exponential link guarantees $\lambda_i > 0$, and the conditional mean is $\mathbb{E}[Y_i|X_i] = \lambda_i = \exp(X_i'\beta)$. The conditional pmf is

$$f(y|x;b) = \frac{e^{-\exp(x'b)}\,\exp(x'b)^y}{y!}$$

so the log-likelihood is

$$\ell_n(b) = \sum_{i=1}^{n} [Y_i X_i' b - \exp(X_i'b) - \log(Y_i!)]$$

The last term is constant in b and can be dropped for optimization. Setting the derivative equal to zero gives the score:

$$S_n(b) = \frac{\partial \ell_n(b)}{\partial b} = \sum_{i=1}^{n} (Y_i - \exp(X_i'b)) X_i$$

To derive the asymptotic distribution, we apply the M-estimator results from above with $\rho(Y_i, X_i, b) = \exp(X_i'b) - Y_i X_i'b + \log(Y_i!)$. Since $\psi(Y_i, X_i, b) = (\exp(X_i'b) - Y_i)X_i$, we have

$$\frac{\partial \psi(Y_i, X_i, b)}{\partial b'} = \exp(X_i'b)\, X_i X_i'$$

so $\mathbf{Q} = \mathbb{E}[\lambda_i X_i X_i']$. Evaluating ψ at the true β,

$$\mathbf{\Omega} = \mathbb{E}\big[(\lambda_i - Y_i)^2 X_i X_i'\big] = \mathbb{E}\big[\mathbb{E}\big[(\lambda_i - Y_i)^2 \mid X_i\big]\, X_i X_i'\big] = \mathbb{E}[\mathrm{Var}(Y_i \mid X_i)\, X_i X_i'] = \mathbb{E}[\lambda_i X_i X_i']$$

where the third equality uses $\mathbb{E}[\lambda_i - Y_i \mid X_i] = 0$ and the fourth uses the Poisson equidispersion property $\mathrm{Var}(Y_i \mid X_i) = \lambda_i$. Hence $\mathbf{Q} = \mathbf{\Omega}$ (information matrix equality), and

$$\mathbf{V} = \mathbf{Q}^{-1} = \left(\mathbb{E}[\lambda_i X_i X_i']\right)^{-1}$$

Replacing λ_i with $\hat{\lambda}_i = \exp(X_i'\hat{\beta})$, the estimated variance of $\hat{\beta}$ is

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \left(\sum_{i=1}^{n} \hat{\lambda}_i\, X_i X_i'\right)^{-1}$$

and standard errors are the square roots of the diagonal elements.

Code

```
library(AER)
data("NMES1988")
```

```r
Y <- NMES1988$visits
X <- model.matrix(~ chronic + hospital + health + insurance, data = NMES1988)

neg_ll <- function(b) {
    lam <- exp(X %*% b)
    -sum(Y * log(lam) - lam)
}

neg_score <- function(b) {
    lam <- exp(X %*% b)
    as.numeric(-t(X) %*% (Y - lam))
}

fit_manual <- optim(
    par     = rep(0, ncol(X)),
    fn      = neg_ll,
    gr      = neg_score,
    method  = "BFGS",
    control = list(maxit = 500)
)

# Estimated variance: (sum_i lambda_hat_i * Xi Xi')^{-1}
lam_hat <- exp(X %*% fit_manual$par)
vcov_manual <- solve(t(X) %*% (c(lam_hat) * X))
se_manual <- sqrt(diag(vcov_manual))

# Compare with glm
fit_glm <- glm(visits ~ chronic + hospital + health + insurance,
    data = NMES1988, family = poisson
)
se_glm <- sqrt(diag(vcov(fit_glm)))

round(cbind(
    coef_manual = fit_manual$par,
    coef_glm = coef(fit_glm),
    se_manual = se_manual,
    se_glm = se_glm
), 4)
```

	coef_manual	coef_glm	se_manual	se_glm
(Intercept)	1.2134	1.2134	0.0171	0.0171
chronic	0.1459	0.1459	0.0046	0.0046
hospital	0.1630	0.1630	0.0060	0.0060
healthpoor	0.2195	0.2195	0.0177	0.0177
healthexcellent	-0.3430	-0.3430	0.0303	0.0303
insuranceyes	0.2646	0.2645	0.0161	0.0161