Lecture 4: Binary Outcomes

Professor: Mauricio Sarrias

Universidad de Talca

2020

1. The Linear Probability Model
   - Introduction to LPM
   - WLS
   - Example
2. Non-Linear Probability Model
   - Latent Approach
   - Probit and Logit
   - Estimation
   - Marginal Effects
3. Covariance Matrix Estimation
   - Asymptotic Distribution
   - Marginal Effects and Average Partial Effects
   - Goodness-of-Fit
4. Example
   - Stata Example
   - Programming in R

Goals and readings

Goals

1. To understand the pros and cons of estimating a Linear probability model.
2. To derive the Probit and Logit model.
3. To understand how to obtain the ME under a binary model.
4. To be able to interpret different measures of binary models.

Binary Dependent Variable

We will study method for estimating model with binary dependent variable:

$$y_i = \begin{cases} 1 & \text{if some event occurs} \\ 0 & \text{if the even does not occurs} \end{cases}$$

Some examples:

- Is an adult a member of the labor force?
- Did a citizen vote in the last election?
- Does a high school student decide to go to college?
- Is a consumer more likely to buy the same brand or try a new brand?
- Does the individual migrate?

Binary Dependent Variable

So, the question is: **How to estimate a model then the dependent variable is binary?**

The first approach is to apply OLS as if the dependent **variable is continuous**.

What if OLS ... ?

The linear probability model is the regression model applied to a binary dependent variable. The structural model is:

$$y_i = \mathbf{x}_i^{\intercal}\boldsymbol{\beta} + \epsilon_i,$$

where

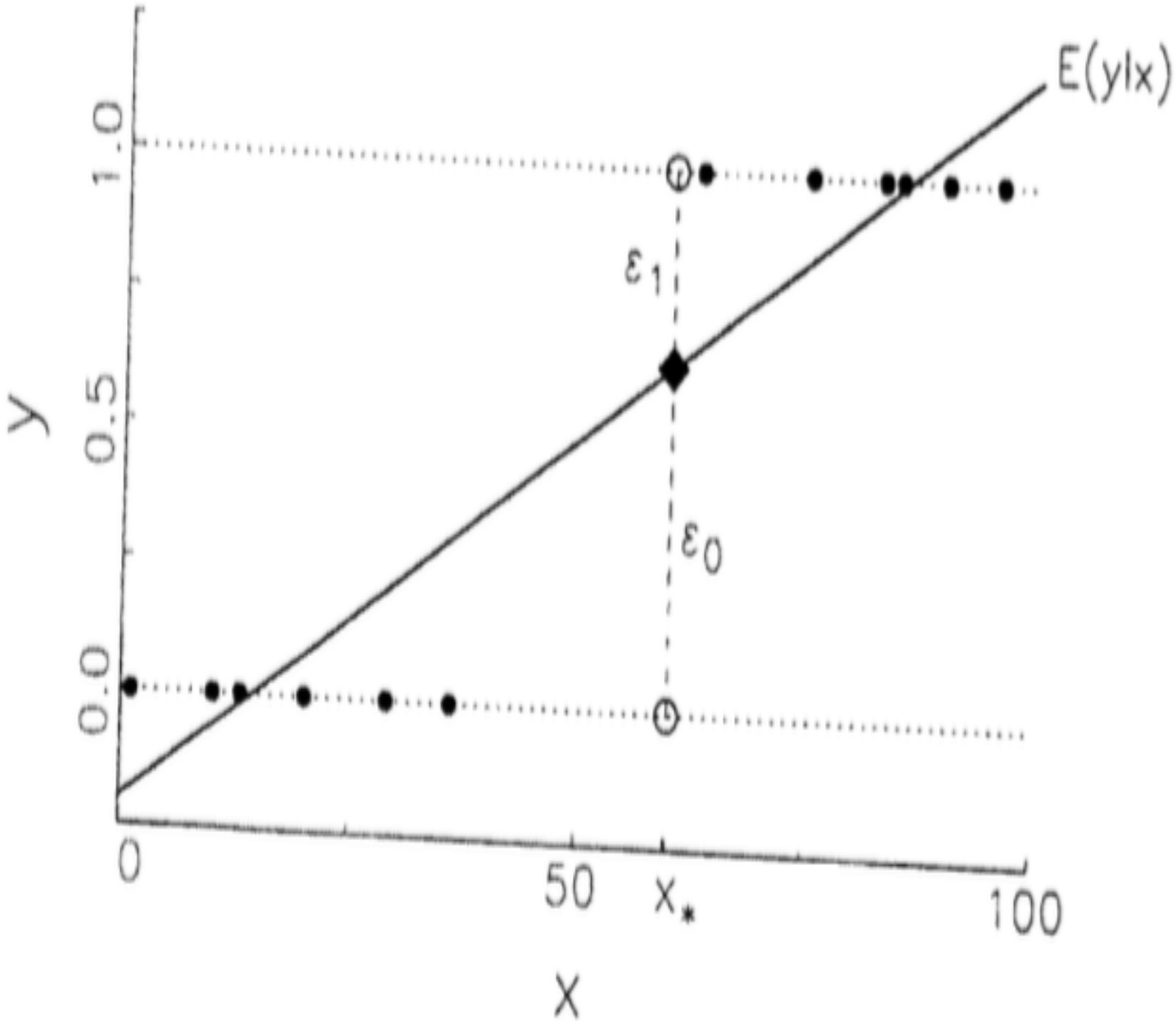
$$y_i = \begin{cases} 1 & \text{if some event occurs} \\ 0 & \text{if the even does not occurs} \end{cases}$$

When y is a **binary random** variable, then:

$$\mathbb{E}\left(y_i|\,\mathbf{x}_i\right) = \left[1 \times \Pr\left(y_i = 1|\,\mathbf{x}_i\right)\right] + \left[0 \times \Pr\left(y_i = 0|\,\mathbf{x}_i\right)\right] = \Pr\left(y_i = 1|\,\mathbf{x}_i\right) = \mathbf{x}_i^{\intercal}\boldsymbol{\beta}$$

Linear Probability Model

For a Single Independent Variable



Problems with the LPM

- **Heterokedasticity**: If a binary random variable has mean μ, then its variance is $\mu(1-\mu)$ **(Prove this!)**. Then:

$$\text{Var}\left(y_i|\,\mathbf{x}_i\right) = \Pr\left(y_i = 1|\,\mathbf{x}_i\right)\left[1 - \Pr\left(y_i = 1|\,\mathbf{x}_i\right)\right] = \mathbf{x}_i^\top \boldsymbol{\beta}\left(1 - \mathbf{x}_i^\top \boldsymbol{\beta}\right),$$

  which implies that the variance of the errors depend on the $\mathbf{x}$'s and is not constant.
- **Nonsensical Predictions**: The LPM predicts values of y that are negative or greater than 1.
- **Functional Form**: Since the model is linear, a unit increase in x_k results in change of β_k in the probability of an event. The increase is the same regardless of the current value of $\mathbf{x}$.

Problems with the LPM

More on Functional Form

- Consider the following two specifications:

$$y_i = \alpha + \beta x_i + \delta d_i + \epsilon_i$$
$$y_i = F\left(\alpha + \beta x_i + \delta d_i\right)$$

where d_i is a dummy variable.

- The discrete change in y as d changes from 0 to 1, holding x constant as:

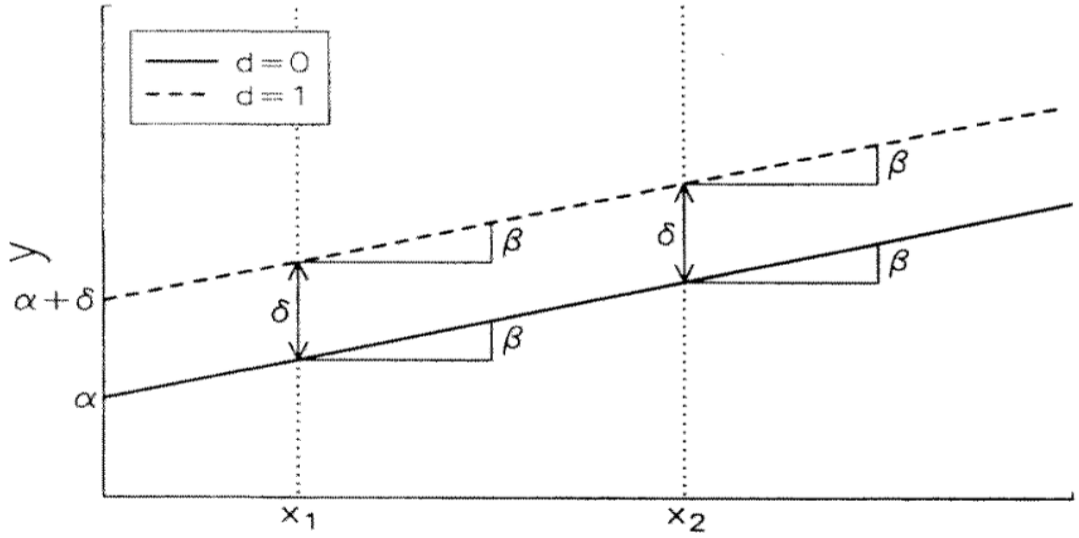
$$\frac{\Delta y}{\Delta d} = (\alpha + \beta x_i + \delta \cdot 1) - (\alpha + \beta x_i + \delta \cdot 0) = \delta$$

- For our second function, the discrete change is?

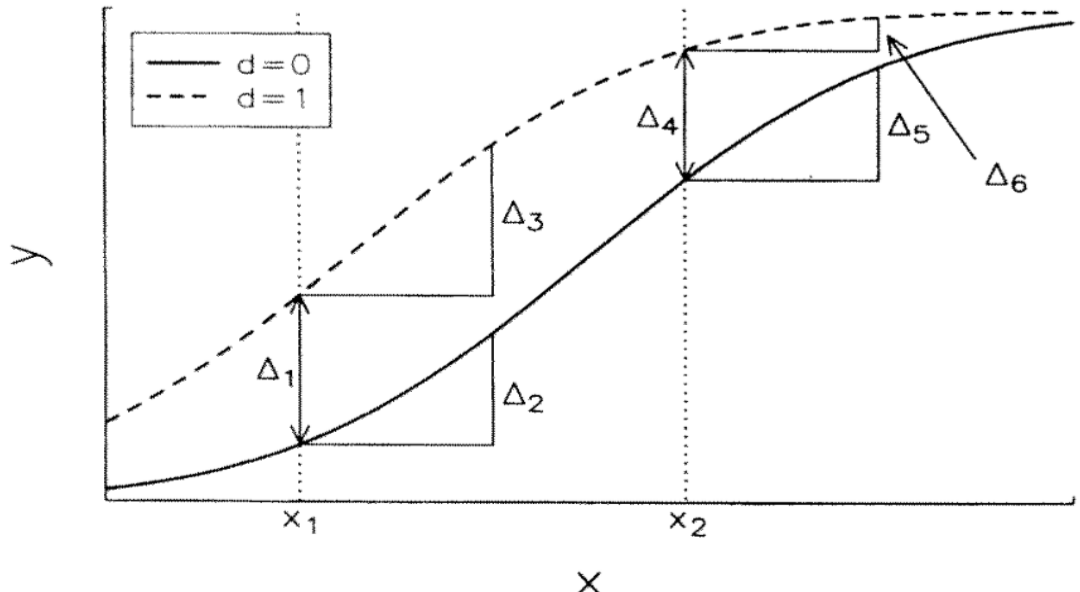
Problems with the LPM

More on Functional Form

Panel A: Linear Model



Panel B: Nonlinear Model



WLS

Since we know the exact form of the heterokedasticity function we can use Weighted-Least-Square (WLS)

In this case:

$$\mathbb{E}\left(\epsilon_i^2|\mathbf{X}\right) = \text{Var}\left(\epsilon_i|\mathbf{X}\right) = \sigma_0^2 h_i(\mathbf{X}) \tag{1}$$

where

$$h_i(\mathbf{X}) = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} \left(1 - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}\right) \tag{2}$$

Then we can estimate the WLS estimator:

$$\widehat{\boldsymbol{\beta}}_{WLS} = \left[\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^\top\right]^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{x}_i y_i\right], \quad w_i = 1/h_i \tag{3}$$

Review your Econometric Notes!

Determinants of Personal Computer Ownership

Consider the following model:

$$\texttt{PC}_i = \beta_0 + \beta_1 \texttt{hsGPA}_i + \beta_2 \texttt{ACT}_i + \beta_3 \texttt{parcoll}_i + \epsilon_i \tag{4}$$

where:

- `PC`: binary indicator equal to unity if the student owns a computer, zero otherwise.
- `hsGPA`: High school GPA.
- `ACT`: is achievement test score.
- `parcoll`: binary indicator equal to unity if at least one parent attended college.[1]

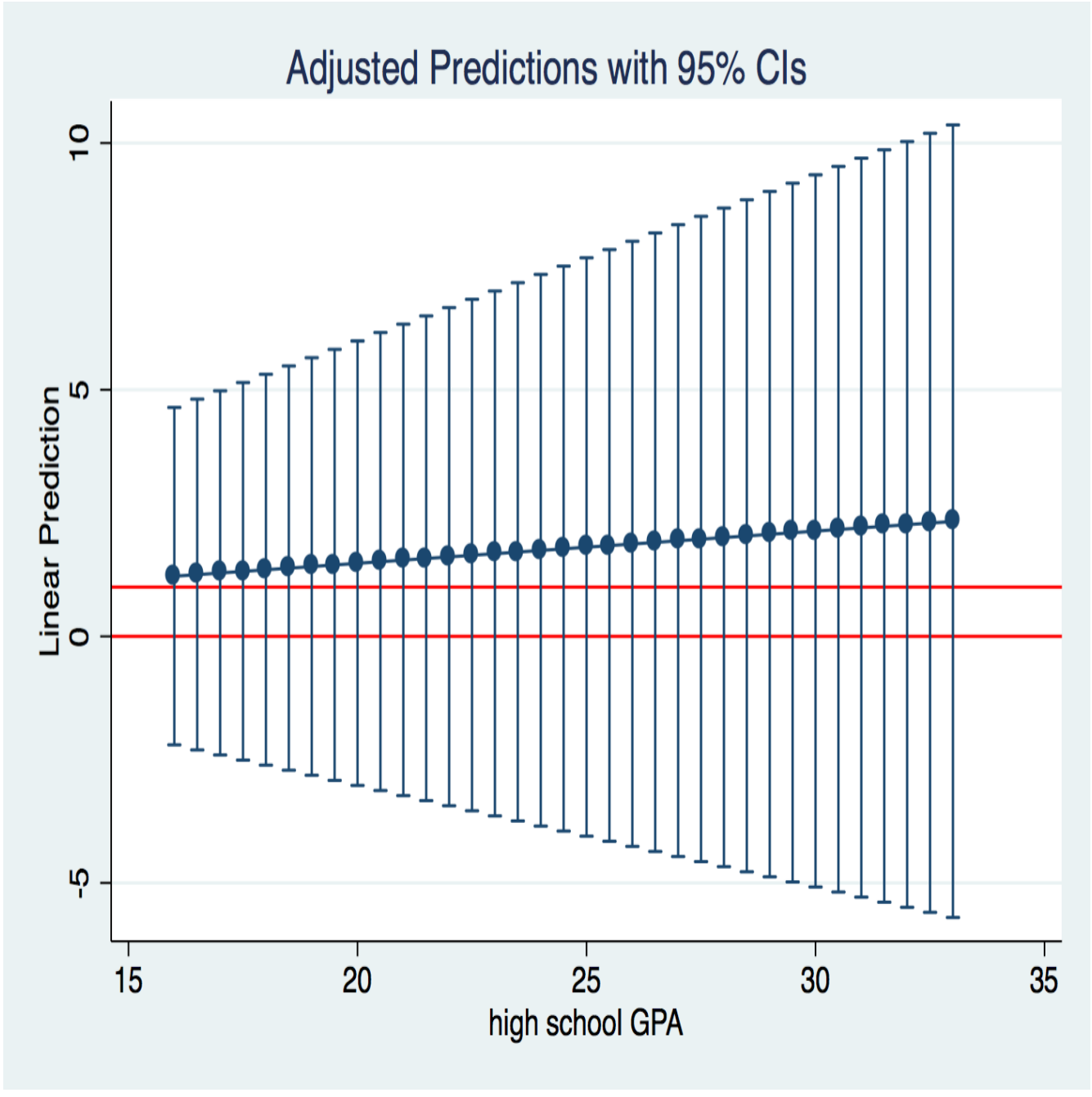
We use data in GPA1.dta to estimate the probability of owning a computer.

[1] Separate college indicators for the mother and father do not yield individually significant results, as these are pretty highly correlated.

```
* Open Data
cd "/Users/mauriciosarrias/Documents/Clases/Discrete Choice Models/P
use "GPA1.DTA", clear

* Gen parcoll
quietly{
gen parcoll = 0
replace parcoll =1 if fathcoll == 1 | mothcoll == 1
reg PC hsGPA ACT parcoll
}

* Plot Predicted Values
quietly margins, at(hsGPA = (16(0.5)33)) atmeans
marginsplot, yline(0, lcolor(red)) yline(1, lcolor(red))
```


Adjusted Predictions with 95% CIs
Linear Prediction
10
5
0
-5
15
20
25
30
35
high school GPA

```
* Predicted value
 qui reg PC hsGPA ACT parcoll
 predict yhat, xb
 sum yhat

Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    yhat |       141    .3971631    .1000667   .1700624   .4974409

 * Weights
 gen h_hat = yhat * (1 - yhat)
```

```
* Models
quietly eststo ols : reg PC hsGPA ACT parcoll
quietly eststo olsr: reg PC hsGPA ACT parcoll, robust
quietly eststo wls : reg PC hsGPA ACT parcoll [aweight = 1/ h_hat]
esttab ols olsr wls, b se ///
mtitle("OLS" "OLS Rob" "WLS")

------------------------------------------------------------
                    (1)             (2)             (3)
                    OLS         OLS Rob             WLS
------------------------------------------------------------
hsGPA             0.0654          0.0654          0.0327
                 (0.137)         (0.141)         (0.130)

ACT              0.000565         0.000565         0.00427
                 (0.0155)        (0.0161)        (0.0155)

parcoll           0.221*          0.221*          0.215*
                 (0.0930)        (0.0880)        (0.0863)

_cons            -0.000432       -0.000432          0.0262
                 (0.491)         (0.496)         (0.477)
------------------------------------------------------------
N                   141             141             141
------------------------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

WLS

Final Remarks

- WLS and Robust Standard Errors help us to lead with Heterokedasticity.
- ... but it does not solve the problems related to interpretation.

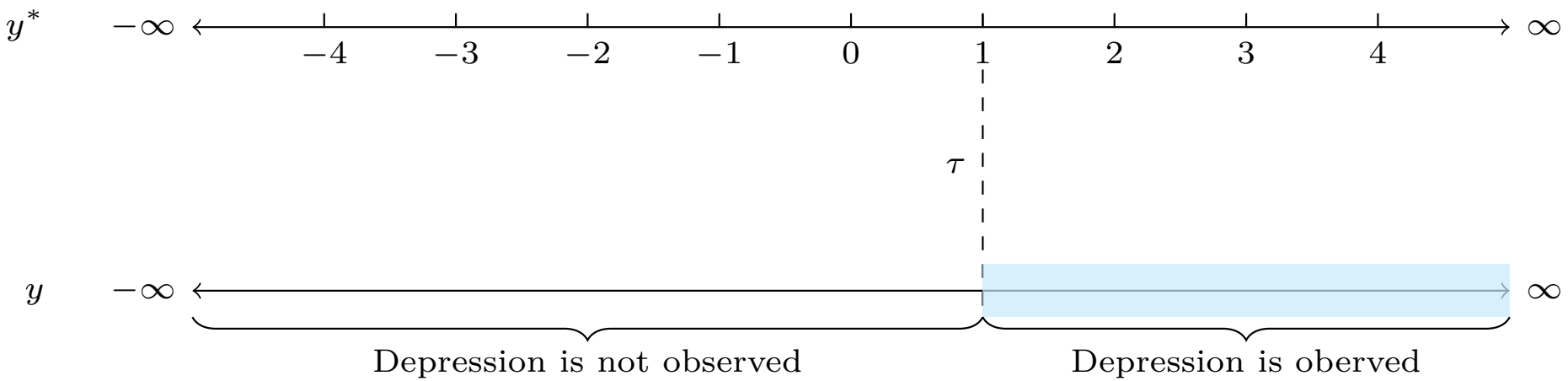
Latent Variable Approach

- Suppose there is an unobserved or latent variable y^* ranging from $-\infty$ to ∞ that generates the observed y, such that those who have larger values of y^* are observed as $y = 1$.
- For example, consider individuals' mental health as the observed y which equals 1 if the individual has depression and 0 otherwise.
- We might assume that there exists an underlying and continuous variable y^* that indicates the propensity of having depression.

Latent Variable Approach

- Let the latent variable y^* be the propensity of having depression, so that higher values implies that people have lower mental health and more likely to actually have depression.
- We do not observe y^*, but we can also assume that if $y^* \geq \tau$, where τ is an arbitrary threshold or cutpoint, then we observed an individual with $y = 1$ and $y = 0$ otherwise.

Figure: Latent process



Latent Variable Approach

The latent $\mathbf{y}^*$ is assumed to be linearly related to the observed $\mathbf{x}$'s through the structural model:

$$y_i^* = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i \tag{5}$$

Definition (Latent process)

Consider the latent process given in Equation (5) and assume that $\mathbb{E}(\epsilon_i|\mathbf{x}_i) = \mathbf{0}$ for all $i = 1, ..., n$. The expected value of y_i^* is:

$$\mathbb{E}(y_i^*|\mathbf{x}_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta},$$

and its variance is:

$$\text{Var}(y_i^*|\mathbf{x}_i) = \text{Var}(\epsilon_i|\mathbf{x}_i) = \sigma_\epsilon^2$$

where σ_ϵ^2 is the variance of the error term.

Latent Variable Approach

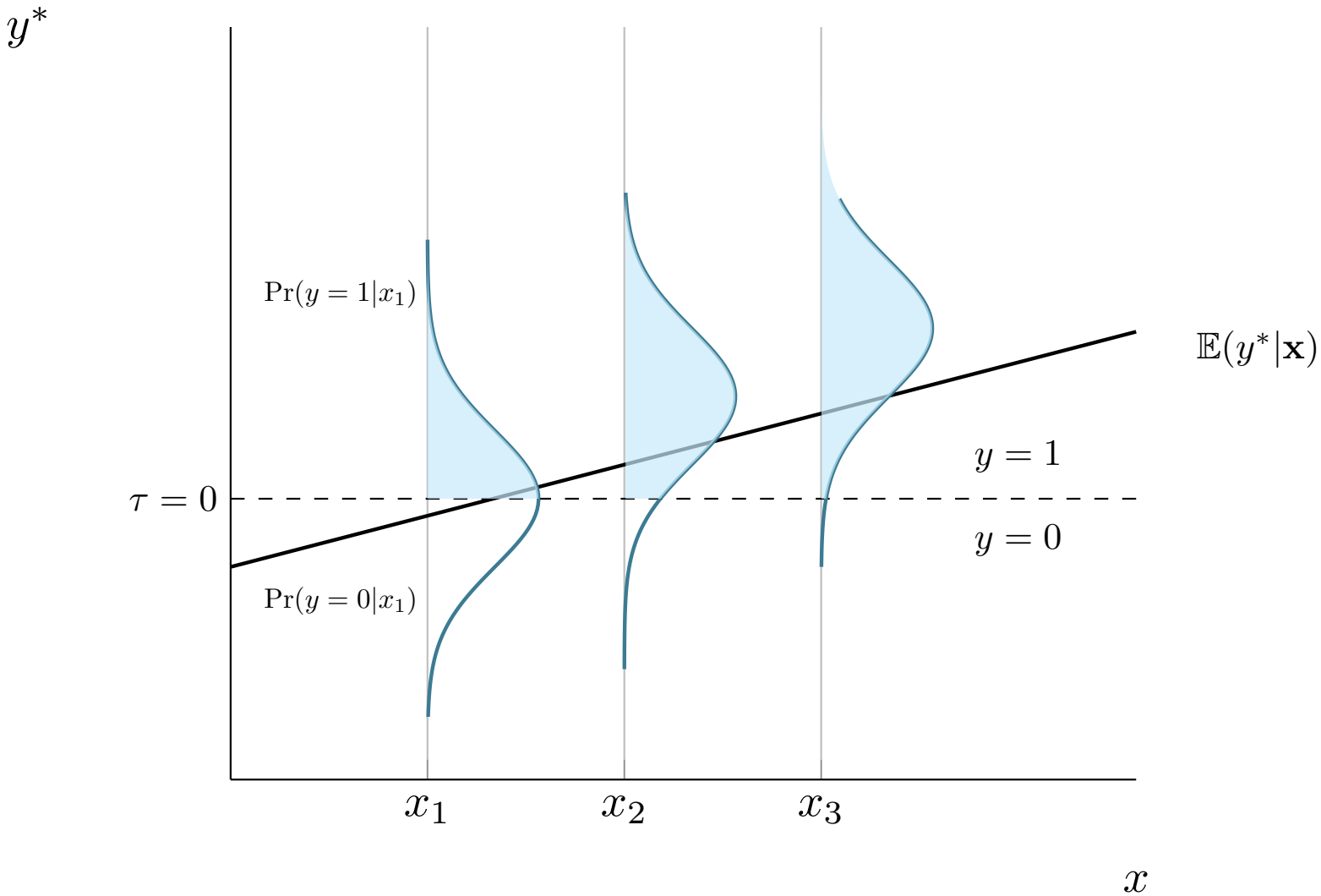
The latent variable y^* is linked to the observed binary variable y by the measurement equation:

$$y_i = \begin{cases} 1, & \text{if } y_i^* > \tau \\ 0, & \text{if } y_i^* \leq \tau \end{cases} \tag{6}$$

A figure might help to understand the role of τ.

Latent Variable Approach

Figure: The distribution of y^* given x in the binary response model



Latent Variable Approach

Caution:

Unfortunately, $\boldsymbol{\beta}$, σ^2 and τ are not separately identified.

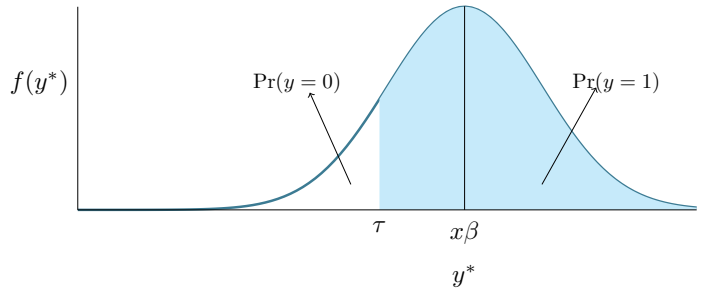
Latent Variable Approach

Let $\mathbf{x}_i^\intercal \boldsymbol{\beta} = \alpha + \mathbf{x}_i^\intercal \boldsymbol{\delta}$:

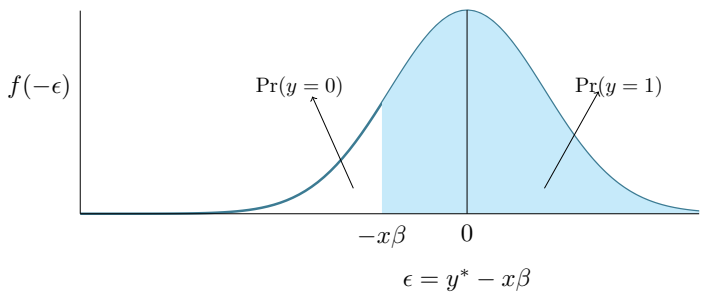
$$
\begin{aligned}
\Pr\left(y_i = 1 | \mathbf{x}_i\right) &= \Pr\left(y_i^* > \tau \middle| \mathbf{x}_i\right) \\
&= \Pr\left(\alpha + \mathbf{x}_i^\intercal \boldsymbol{\delta} + \epsilon_i > \tau \middle| \mathbf{x}_i\right) \\
&= \Pr\left(\epsilon_i > -(\alpha - \tau) - \mathbf{x}_i^\intercal \boldsymbol{\delta} \middle| \mathbf{x}_i\right) \\
&= 1 - \Pr\left(\epsilon_i \leq -(\alpha - \tau) - \mathbf{x}_i^\intercal \boldsymbol{\delta} \middle| \mathbf{x}_i\right) \\
&= \Pr\left(\epsilon_i \leq (\alpha - \tau) + \mathbf{x}_i^\intercal \boldsymbol{\delta} \middle| \mathbf{x}_i\right) \\
&= \Pr\left(\frac{\epsilon_i}{\sigma_\epsilon} \leq \frac{(\alpha - \tau)}{\sigma_\epsilon} + \frac{\mathbf{x}_i^\intercal \boldsymbol{\delta}}{\sigma_\epsilon} \middle| \mathbf{x}_i\right) \\
&= \int_{-\infty}^{\mathbf{x}_i^\intercal \boldsymbol{\beta}} f(\epsilon_i) d\epsilon_i \\
&= F\left(\frac{(\alpha - \tau)}{\sigma_\epsilon} + \frac{\mathbf{x}_i^\intercal \boldsymbol{\delta}}{\sigma_\epsilon}\right)
\end{aligned}
$$

where $F(\cdot)$ is the 'standardized' cumulative density function (CDF) of the error term.

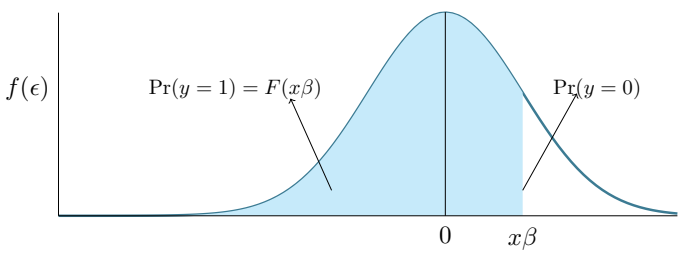
(a) Original axis



(b) Shift the axis



(c) Flip the axis



Latent Variable Approach

Thus...

Thus, in general we set σ^2 to be fixed and $\tau = 0$, such that

$$\begin{aligned}\Pr\left(y_i = 1 \middle| \mathbf{x}_i\right) &= F(\mathbf{x}_i^\intercal \boldsymbol{\beta}) \\ &= \int_{-\infty}^{\mathbf{x}_i^\intercal \boldsymbol{\beta}} f(\epsilon_i) d\epsilon_i\end{aligned}$$

Probit

Error term distributed as Normal

When $\boldsymbol{\varepsilon}$ is normal with $\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\mathbf{X}\right)=0$ and $\mathrm{Var}\left(\boldsymbol{\varepsilon}|\,\mathbf{X}\right)=\mathbf{I}$, the pdf is

$$\phi(\boldsymbol{\varepsilon})=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\boldsymbol{\varepsilon}^2}{2}\right)$$

and the cumulative distribution function (cdf) is:

$$\Phi(\boldsymbol{\varepsilon})=\int_{-\infty}^{\epsilon}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

Logit

Error term distributed as Logistic

In the Logistic model, the errors are assumed to have a standard logistic distribution with mean 0 and variance $\pi^2/3$. This unusual variance is chosen because it results in a particularly simple equation for the pdf:

$$\lambda(\epsilon) = \frac{\exp(\epsilon)}{[1 + \exp(\epsilon)]^2}$$

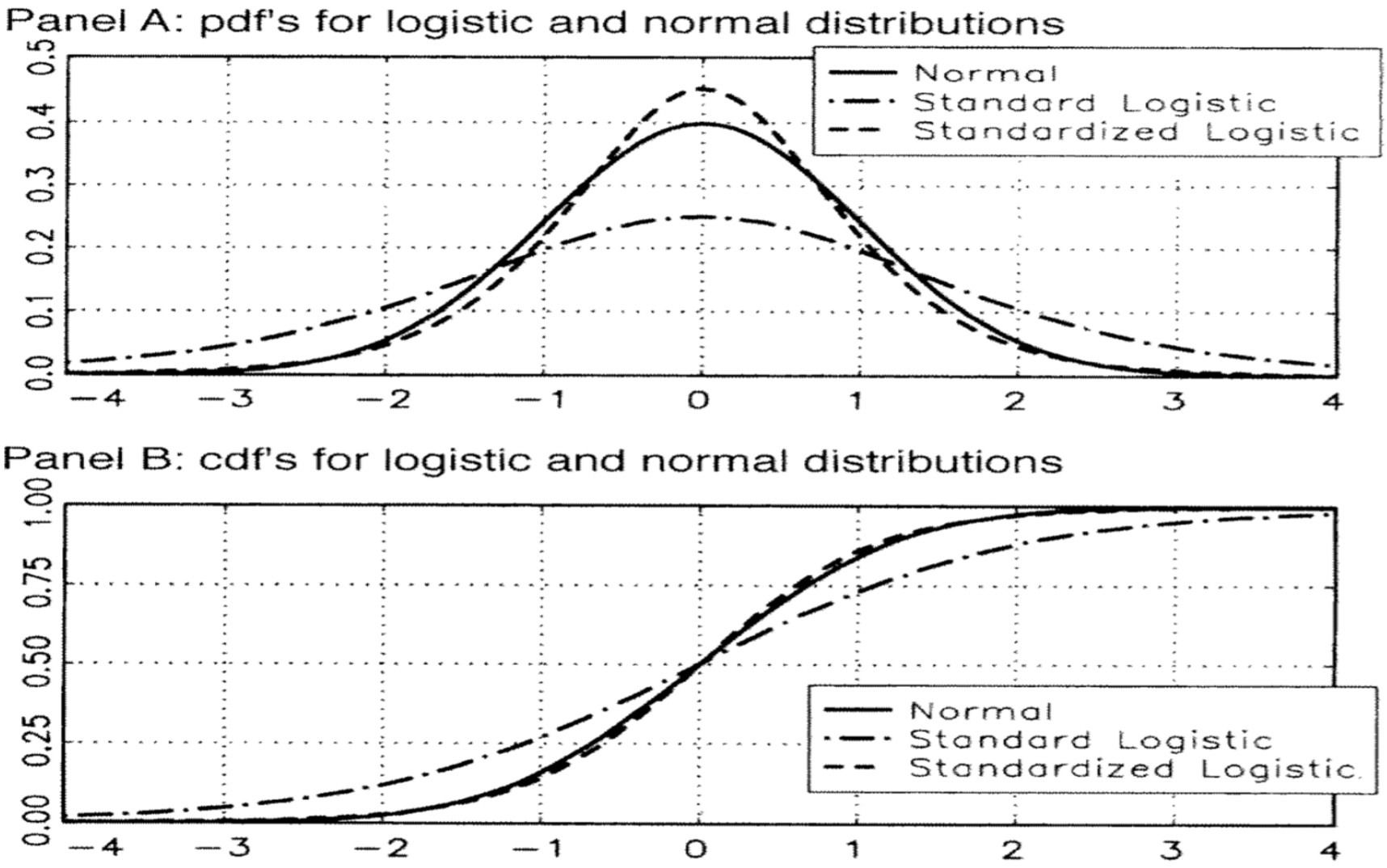
and an even simpler equation for the cdf:

$$\Lambda(\epsilon) = \frac{\exp(\epsilon)}{1 + \exp(\epsilon)}$$

Normal and Logistic Distribution

- The probit model, like many other statistical models using the normal distribution, may be justified by appealing to a central limit theorem.
- A justification for the logit model is that the logistic distribution is similar to the nomral distribution function but has a much simpler form.

Normal and Logistic Distribution



Probit and Logit

Probit

If the ϵ_i's are independently and normally distributed, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, then

$$
\begin{aligned}
\Pr\left(y_i = 1 \middle| \mathbf{x}_i\right) &= \Pr\left(\frac{\epsilon_i}{\sigma} > -\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}_i\right) \\
&= 1 - \Phi\left(-\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \\
&= \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \quad \text{by symmetry of standard normal distribution} \\
&= \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \quad \text{setting } \sigma^2 = 1 \text{ for identification}
\end{aligned}
$$

Probit and Logit

Logit

If the ϵ_i's are independently and logistically distributed, $\epsilon_i \sim \Lambda(0, \pi^2/3)$, then

$$\begin{aligned}\Pr\left(y_i = 1 | \mathbf{x}_i\right) &= \Lambda\left(\frac{\mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \quad \text{since } \sigma = \pi/\sqrt{3}\end{aligned}$$

ML Estimation

The outcome is Bernoulli distributed, the binomial distribution with just one trial. A very convenient compact notation for the density of y_i, or more formally its **probability mass function**, is:

$$f\left(y_i|\,\mathbf{x}_i\right) = P_i^{y_i}(1-P_i)^{1-y_i}, \qquad y_i = 1, 0$$

where $P_i = F(\mathbf{x}_i'\boldsymbol{\beta})$. This yields probabilities P_i and $(1-P_i)$ since $f(1) = P^1(1-P)^0 = P$ and $f(0) = P^0(1-P)^1 = 1-P$. Assuming that each probability is independent of that other, the joint probability function (or Likelihood function) is:

$$\Pr\left(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n|\,\mathbf{X}\right) = \prod_{i=1}^{n} f\left(y_i|\,\mathbf{x}_i\right)$$

The likelihood function for a sample of n obvervations can be written as

$$L\left(\boldsymbol{\beta}|\underbrace{\mathbf{y}, \mathbf{X}}_{data}\right) = \prod_{i=1}^{n}\left[F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{y_i}\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{1-y_i}$$

Log-Likelihood Function

Taking logs, we obtain the Log-Likelihood function, which must be maximized

$$\begin{aligned}\log L\left(\boldsymbol{\beta}|\,\text{data}\right) &= \log\left(\prod_{i=1}^{n}\left[F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{y_i}\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{1-y_i}\right)\\ &= \sum_{i=1}^{n}\left\{\log\left(\left[F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{y_i}\right)+\log\left(\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^{1-y_i}\right)\right\}\\ &= \sum_{i=1}^{n}\left\{y_i\log\left[F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]+(1-y_i)\log\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]\right\}\end{aligned}$$

Useful Trick

If the distribution is symmetric, as the normal and logistic, then $1-F(\mathbf{x}_i^\top\boldsymbol{\beta})=F(-\mathbf{x}_i^\top\boldsymbol{\beta})$. Let $q_i=2y_i-1$. Then:

$$\log L\left(\boldsymbol{\beta}|\,\text{data}\right)=\sum_{i=1}^{n}\log F\left(q_i\mathbf{x}_i^\top\boldsymbol{\beta}\right)$$

Gradient

FOC:

$$\underbrace{\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}}_{(K\times 1)} = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ y_i \log\left[F(\mathbf{x}_i^\top \boldsymbol{\beta})\right] + (1-y_i)\log\left[1-F(\mathbf{x}_i^\top \boldsymbol{\beta})\right] \right\}$$

$$= \sum_{i=1} \left\{ y_i \frac{f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i}{F(\mathbf{x}_i^\top \boldsymbol{\beta})} + (1-y_i)\frac{-f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i}{1-F(\mathbf{x}_i^\top \boldsymbol{\beta})} \right\} \quad \because \frac{\partial F(\mathbf{x}_i^\top \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i$$

$$= \sum_{i=1} \left\{ \frac{y_i}{F(\mathbf{x}_i^\top \boldsymbol{\beta})} - \frac{(1-y_i)}{1-F(\mathbf{x}_i^\top \boldsymbol{\beta})} \right\} f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i$$

$$= \sum_{i=1} \left\{ \frac{y_i(1-F(\mathbf{x}_i^\top \boldsymbol{\beta})) - (1-y_i)F(\mathbf{x}_i^\top \boldsymbol{\beta})}{F(\mathbf{x}_i^\top \boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top \boldsymbol{\beta})\right]} \right\} f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i$$

$$= \sum_{i=1} \left\{ \frac{y_i - F(\mathbf{x}_i^\top \boldsymbol{\beta})}{F(\mathbf{x}_i^\top \boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top \boldsymbol{\beta})\right]} \right\} f(\mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i$$

Gradient

Therefore, the ML estimator $\widehat{\boldsymbol{\beta}}_{ML}$ is given by the solution of:

$$\underbrace{\sum_{i=1}\left\{\frac{y_i - F(\mathbf{x}_i^\top \boldsymbol{\beta})}{F(\mathbf{x}_i^\top \boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\right]}\right\}}_{(1\times 1)} \underbrace{f(\mathbf{x}_i^\top \boldsymbol{\beta})}_{(1\times 1)} \underbrace{\mathbf{x}_i}_{(K\times 1)} = \underbrace{\mathbf{0}}_{K\times 1}$$

This equation does not have analytic solution of $\widehat{\boldsymbol{\beta}}_{ML}$ and we have to solve it by **numerical computation**

Using our previous trick, we have:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left[\frac{q_i f(q_i \mathbf{x}_i^\top \boldsymbol{\beta})}{F(\mathbf{x}_i^\top \boldsymbol{\beta})}\right]\mathbf{x}_i = \sum_{i=1}^{n}\lambda_i \mathbf{x}_i = \mathbf{0}$$

Hessian

The Hessian is:

$$\mathbf{H} = \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{N} \left[\frac{q_i^2 f'(q_i \mathbf{x}_i^\top \boldsymbol{\beta})}{F(q_i \mathbf{x}_i^\top \boldsymbol{\beta})} - \lambda_i^2 \right] \mathbf{x}_i \mathbf{x}_i^\top$$

Again, note that $F(\cdot)$ depend upon the distribution of the error term. If we assume a probit model, then:

$$\begin{aligned} F(\epsilon) &= \Phi(\epsilon) = \int_{-\infty}^{\epsilon} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t}{2}\right) dt \\ f(\epsilon) &= \phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2}{2}\right) \\ f'(\epsilon) &= -\epsilon\phi(\epsilon) \end{aligned}$$

and for the logit model, we have:

$$\begin{aligned} F(\epsilon) &= \Lambda(\epsilon) \\ f(\epsilon) &= \Lambda(\epsilon)\left[1 - \Lambda(\epsilon)\right] \\ f'(\epsilon) &= \Lambda(\epsilon)\left[1 - \Lambda(\epsilon)\right]\left[1 - 2\Lambda(\epsilon)\right] \end{aligned}$$

Some Comments

- The Hessian is always negative definite, so the log-likelihood is **globally concave**.
- Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations unless the data are especially badly conditioned.

Marginal Effects

Recall that:

$$\mathbb{E}\left(y_i|\,\mathbf{x}_i\right) = [1 \times \Pr\left(y_i = 1|\,\mathbf{x}_i\right)] + [0 \times \Pr\left(y_i = 0|\,\mathbf{x}_i\right)] = \Pr\left(y_i = 1|\,\mathbf{x}_i\right) = F(\mathbf{x}_i^\top \boldsymbol{\beta})$$

The marginal effect is given by:

Marginal Effects

$$\underbrace{\frac{\partial \mathbb{E}\left(y_i|\,\mathbf{x}_i\right)}{\partial \mathbf{x}_i}}_{(K\times 1)} = \left[\frac{dF(\mathbf{x}_i^\top \boldsymbol{\beta})}{d(\mathbf{x}_i^\top \boldsymbol{\beta})}\right]\boldsymbol{\beta} = \underbrace{f(\mathbf{x}_i^\top \boldsymbol{\beta})}_{(1\times 1)}\underbrace{\boldsymbol{\beta}}_{(K\times 1)}$$

where $f(\cdot)$ is the probability density function.

So:

$$\text{Probit} \implies \frac{\partial \mathbb{E}\left(y_i|\,\mathbf{x}_i\right)}{\partial \mathbf{x}_i} = \phi(\mathbf{x}_i^\top \boldsymbol{\beta}_i)\boldsymbol{\beta}$$

$$\text{Logit} \implies \frac{\partial \mathbb{E}\left(y_i|\,\mathbf{x}_i\right)}{\partial \mathbf{x}_i} = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta})\left[1 - \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta})\right]\boldsymbol{\beta}$$

Marginal Effects

Comments

Some aspects are important to note:

- MEs will vary with the values of $\mathbf{x}$.
- Current practices:
  - Calculate MEs at means of the variables.
  - Calculate MEs at specific values.
  - Evaluate MEs at every observation and use the sample average of the individual MEs.

Marginal Effects

Dummy variable

The appropriate ME for a binary independent variable, say, d, would be:

$$\text{ME} = \left[\Pr\left(y_i = 1|\bar{\mathbf{x}}_{(d)}, d_i = 1\right)\right] - \left[\Pr\left(y_i = 1|\bar{\mathbf{x}}_{(d)}, d_i = 0\right)\right]$$

where $\bar{\mathbf{x}}_{(d)}$ denotes the means of all the other variables in the model.

Marginal Effects

Elasticities

It is common to report elasticities of probabilities, rather than derivatives. These are computed as:

$$\begin{aligned}\epsilon_{i,k} &= \frac{\partial \ln \Pr(y_i = 1|\mathbf{x})}{\partial \ln x_{i,k}} \\ &= \frac{\partial \Pr(y_i = 1|\mathbf{x})}{\partial x_{i,k}} \frac{x_{i,k}}{\Pr(y_i = 1|\mathbf{x})}\end{aligned}$$

Since it is a ratio of percentage changes, the elasticity is not likely to be useful for dummy variables.

Asymptotic Distribution of Probit and Logit Model

Recall that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}^{-1})$$

where $\mathbf{I}$ is the sample Fisher Information defined as:

$$\mathbf{I} = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})\right]$$

Using **Information equality**

$$\mathbf{I} = -\mathbb{E}\underbrace{\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})\right]}_{\text{Hessian}} = \mathbb{E}\underbrace{\left[\left(\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right)\right]}_{\text{Outer product}}$$

We have derivatives:

$$\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\} f(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i$$

$$\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top} = \left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\} f(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i^\top$$

Asymptotic Distribution of Probit and Logit Model

$$\mathbf{I} = \mathbb{E}\left[\underbrace{\left(\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)}_{(K\times 1)}\underbrace{\left(\frac{\partial^2 \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\right)}_{(1\times K)}\right]$$

$$= \mathbb{E}\left[\left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\} f(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\} f(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i^\top\right]$$

$$= \mathbb{E}\left[\left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\}^2 f(\mathbf{x}_i^\top\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i^\top\right]$$

$$= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}_i|\mathbf{x}_i}\left(\left\{\frac{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\right\}^2 f(\mathbf{x}_i^\top\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i^\top \middle| \mathbf{x}_i\right)\right] \quad \text{By LIE}$$

$$= \mathbb{E}_{\mathbf{x}}\left[\frac{\mathbb{E}_{y_i|\mathbf{x}_i}\left(\left\{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right\}^2 \middle| \mathbf{x}_i\right)}{F(\mathbf{x}_i^\top\boldsymbol{\beta})^2\left[1-F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^2} f(\mathbf{x}_i^\top\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i^\top\right]$$

Asymptotic Distribution of Probit and Logit Model

Using definition of conditional variance:

$$\mathbb{E}_{y_i|\mathbf{x}_i}\left(\left\{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right\}^2\Big|\,\mathbf{x}_i\right) = \mathrm{Var}_{y_i|\mathbf{x}_i}\left(\left\{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right\}\Big|\,\mathbf{x}_i\right) + \mathbb{E}_{y_i|\mathbf{x}_i}\left(y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\Big|\,\mathbf{x}_i\right)^2$$

where

$$\begin{aligned}\mathbb{E}_{y_i|\mathbf{x}_i}\left(y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\Big|\,\mathbf{x}_i\right) &= \mathbb{E}_{y_i|\mathbf{x}_i}\left(y_i|\,\mathbf{x}_i\right) - F(\mathbf{x}_i^\top\boldsymbol{\beta}) \\ &= F(\mathbf{x}_i^\top\boldsymbol{\beta}) - F(\mathbf{x}_i^\top\boldsymbol{\beta}) = 0\end{aligned}$$

and

$$\mathrm{Var}_{y_i|\mathbf{x}_i}\left(\left\{y_i - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right\}\Big|\,\mathbf{x}_i\right) = F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]$$

Then:

$$\mathbf{I} = \mathbb{E}_{\mathbf{x}}\left[\frac{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}{F(\mathbf{x}_i^\top\boldsymbol{\beta})^2\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]^2} f(\mathbf{x}_i^\top\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i^\top\right] = \mathbb{E}_{\mathbf{x}}\left[\frac{f^2(\mathbf{x}_i^\top\boldsymbol{\beta})}{F(\mathbf{x}_i^\top\boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top\boldsymbol{\beta})\right]}\mathbf{x}_i\mathbf{x}_i^\top\right]$$

Asymptotic Distribution of Probit and Logit Model

Then, an estimator of the **expected value of the Hessian** (or information matrix), is given by:

$$\begin{aligned}\mathbf{I}(\boldsymbol{\beta}) &= -\mathbb{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{f^2(\mathbf{x}_i^\top \boldsymbol{\beta})}{F(\mathbf{x}_i^\top \boldsymbol{\beta})\left[1 - F(\mathbf{x}_i^\top \boldsymbol{\beta})\right]} \mathbf{x}_i \mathbf{x}_i^\top\end{aligned} \tag{7}$$

We can also use the following estimators:

- $\left(-\sum_{i=1}^{n} \mathbf{H}(\mathbf{w}_i, \widehat{\boldsymbol{\beta}})\right)^{-1}$
- or the outer product of the gradients.

What about small samples?

Griffiths et al. (1987)

- We have three estimators.
  - NR: inverse of the negative of the Hessian matrix from the log-likelihood function.
  - Scoring: inverse of the information matrix is used.
  - BHHH: the inverse of the outer product of the first derivatives of the log-likelihood function.
- They are asymptotically equivalent, but their performance can vary in finite samples.

They find that, on average, the Hessian matrix and the information matrix give almost identical results and lead to more accurate estimates of the asymptotic covariance matrix than does the estimator based on first derivatives.

ME and APE

Theorem (Delta Method)

Suppose $\{\boldsymbol{\theta}_n\}$ *is a sequence of* K*-dimensional random vectors such that* $\boldsymbol{\theta}_n \xrightarrow{p} \boldsymbol{\theta}$ *(Consistency) and*

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{z}$$

and suppose $\mathbf{f}(\cdot) : \mathbb{R}^K \to \mathbb{R}^r$ *has a continuous first derivatives with* $\mathbf{F}(\boldsymbol{\theta})$ *denoting the* $r \times K$ *matrix of first derivatives evaluated at* $\boldsymbol{\beta}$*:*

$$\mathbf{F}(\boldsymbol{\theta}) \equiv \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top}$$

Then:

$$\sqrt{n}\left[\mathbf{f}(\boldsymbol{\theta}_n) - \mathbf{f}(\boldsymbol{\theta})\right] \xrightarrow{d} \mathbf{F}(\boldsymbol{\theta})\mathbf{z}$$

In particular, if:

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

then:

$$\sqrt{n}\left[\mathbf{f}(\boldsymbol{\theta}_n) - \mathbf{f}(\boldsymbol{\theta})\right] \xrightarrow{d} N(\mathbf{0}, \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{F}(\boldsymbol{\theta})^\top)$$

ME and APE

Consider the following nonlinear estimators:

- predicted probabilities: $F(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) = \widehat{F}$
- estimated partial effects: $f(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) \times \widehat{\boldsymbol{\beta}} = \widehat{f}\widehat{\boldsymbol{\beta}}$

Using Delta Method, we have:

$$\text{Var}(\widehat{F}) = \left[\frac{\partial \widehat{F}}{\partial \widehat{\boldsymbol{\beta}}}\right]^\top \mathbf{V} \left[\frac{\partial \widehat{F}}{\partial \widehat{\boldsymbol{\beta}}}\right]$$

where $\mathbf{V}$ is the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$. Let $z = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}$. Then:

$$\frac{\partial \widehat{F}}{\partial \widehat{\boldsymbol{\beta}}} = \frac{d\widehat{F}}{dz}\frac{\partial z}{\partial \widehat{\boldsymbol{\beta}}} = \widehat{f}\mathbf{x}$$

Then:

$$\text{Var}(\widehat{F}) = \widehat{f}^2 \mathbf{x}^\top \mathbf{V} \mathbf{x}$$

ME and APE

Note the ME for a dummy variable we have:

$$\Delta\widehat{F} = \left[\widehat{F}|d=1\right] - \left[\widehat{F}|d=0\right]$$

The asymptotic variance would be

$$\text{Var}(\Delta\widehat{F}) = \left[\frac{\partial\Delta\widehat{F}}{\partial\widehat{\boldsymbol{\beta}}}\right]^{\top} \mathbf{V} \left[\frac{\partial\Delta\widehat{F}}{\partial\widehat{\boldsymbol{\beta}}}\right]$$

where

$$\left[\frac{\partial\Delta\widehat{F}}{\partial\widehat{\boldsymbol{\beta}}}\right] = \widehat{f}_1(1, \bar{\mathbf{x}}_{(d)}) - \widehat{f}_0(0, \bar{\mathbf{x}}_{(d)})$$

ME

For the other marginal effects, let $\widehat{\boldsymbol{\gamma}} = \widehat{f}\widehat{\boldsymbol{\beta}}$. Then:

$$\text{Var}(\widehat{\boldsymbol{\gamma}}) = \left[\frac{\partial \widehat{\boldsymbol{\gamma}}}{\partial \widehat{\boldsymbol{\beta}}}\right]' \mathbf{V} \left[\frac{\partial \widehat{\boldsymbol{\gamma}}}{\partial \widehat{\boldsymbol{\beta}}}\right]$$

Average Partial Effects

Current practice: "average partial effects". The quantity of interest is:

$$APE = \mathbb{E}_x \left[\frac{\partial \mathbb{E}(y|\mathbf{x})}{\partial \mathbf{x}} \right]$$

In practical terms, this suggests the computation:

$$\widehat{APE} = \bar{\widehat{\boldsymbol{\gamma}}} = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\beta}}.$$

Goodness-of-Fit

- A number of suggestions have been made for how to evaluate the overall quality of a binary response model.
  - First approach: to mimic the R^2 measure.
  - Assess the predictive performance of the model

R^2 are not directly applicable in non-linear models such as binary response models, since **we do not have a proper variance decomposition result**

The so-called **pseudo** R^2 measures have been suggested.

Goodness-of-Fit

Let:

- $\log L(\widehat{\boldsymbol{\beta}}_r)$: the value of the maximized log-likelihood function in the constant-only model.
- $\log L(\widehat{\boldsymbol{\beta}}_u)$: is the maximized log-likelihood value in the full model.

Note that the value of the log-likelihood function is always negative, so:

$$\log L(\widehat{\boldsymbol{\beta}}_u) \geq \log L(\widehat{\boldsymbol{\beta}}_r) \implies \left|\log L(\widehat{\boldsymbol{\beta}}_u)\right| \leq \left|\log L(\widehat{\boldsymbol{\beta}}_r)\right|$$

So that:

$$0 \leq 1 - \frac{\log L(\widehat{\boldsymbol{\beta}}_u)}{\log L(\widehat{\boldsymbol{\beta}}_r)} = R^2_{\text{McFadden}} \leq 1$$

The McFadden R^2 will be zero if the full model has no explanatory power.

Goodness-of-Fit

Another measure if the McKelvey and Zavoina (1975). It is based on the latent linear model $y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$. In particular, if we let $\widehat{y}_i^* = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$, then we can write:

$$R^2_{MZ} = \frac{SSE^*}{SSR^* + SSE^*} = \frac{\sum_{i=1}(\widehat{y}_i^* - \bar{\widehat{y}^*})^2}{n\sigma^2 + \sum_{i=1}(\widehat{y}_i^* - \bar{\widehat{y}^*})^2}$$

where SSE^* denotes the explained sum of squared, and SSR^* denotes the "residual" sum of the squares of the latent model.

- For probit $\sigma^2 = 1$
- For logit $\sigma^2 = \pi^2/3$

Note

See Scott section 4.3.2 for more goodness-of-fit measures.

Information Measures

Definition (Akaike's Information Criterion (AIC))

Akaike's (1973) information criterion is defined as

$$AIC = \frac{-2 \log \widehat{L} + 2K}{n} \tag{8}$$

where $\log \widehat{L}$ is the likelihood of the model and K is the number of parameters in the model.

- Larger values of $\log \widehat{L}$ indicates a better fit.
- $-2 \log \widehat{L}$ ranges from to 0 $+\infty$ with smaller values indicating a better fit.
- As K increases, $-2 \log \widehat{L}$ becomes smaller since **more parameters make what is observed more likely**.
- $2K$ is added as a penalty.
- All else being equal, smaller values suggest a better fitting model.
- Use to compare models across different samples or to compare nonested models.

Information Measures

Definition (Bayes Information Criterion (BIC))

BIC information criterion is defined as

$$BIC = \frac{-2 \log \widehat{L} + K \log n}{n} \tag{9}$$

where $\log \widehat{L}$ is the likelihood of the model and K is the number of parameters in the model and n is the number of individuals.

It is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

Stata Example

Open `BinaryStata.do`

Code in R

Open `CodeProbit.R` and `Phat.R`