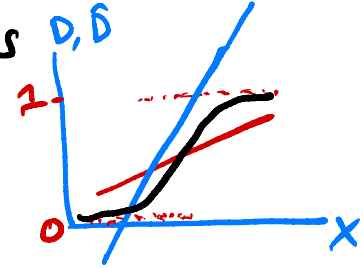
Lecture Note 7: Binary dependant variables



$$D_i = \begin{cases} 1 \\ 0 \end{cases} \qquad D_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots + \varepsilon_i$$

$$= X_i' \beta + \varepsilon_i$$

Interpretation: $\hat{D}_i = X_i' \beta$ as expected $Pr[D_i = 1 | X_i]$

Linear probability model

Coefficients: $\beta_k = \dfrac{\partial Pr[D_i = 1]}{\partial X_{ki}}$ if X_k is continuous

β_k = difference in conditional expectations if X_k binary

Problems:

1. $Pr[D_i = 1]$ may lie outside $[0, 1]$
2. Heteroskedasticity

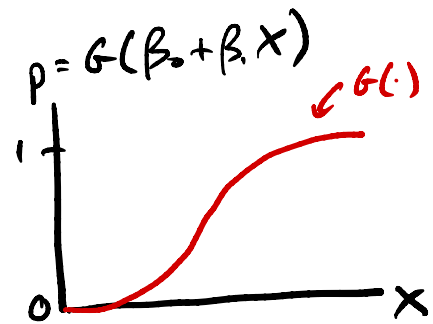
Probit and logit models

For Bernoulli:

$$L = p^S (1-p)^{N-S} = \prod_{i=1}^{N} p^{D_i} (1-p)^{1-D_i}$$

Now we want to estimate some function $p = G(X_i'\beta)$

$$L = \prod_{i=1}^{N} G(X_i'\beta)^{D_i} \left[1 - G(X_i'\beta)\right]^{1-D_i}$$



We will use CDFs as $G(\cdot)$:

1. Std. normal dist → Probit
2. Logistic dist → Logit

Probit regression:

- $\Pr[D_i = 1 \mid X_i] = \Phi[X_i'\beta] = \int_{-\infty}^{X_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x'\beta)^2} dx$
- In R: feglm( ) with family = 'probit'

Logit regression:

- $\Pr[D_i = 1 \mid X_i] = \Lambda[X_i'\beta] = \frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}} = \frac{1}{1 + e^{-X_i'\beta}}$
- In R: feglm( ) with family = 'logit'

Latent variables

Think of D_i as crude discretization of latent variable:

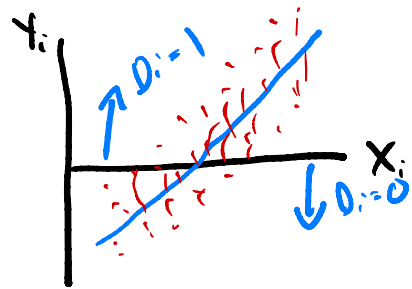
$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots = X_i'\beta + \varepsilon_i$$

$\varepsilon_i \sim N(0,1)$

$\varepsilon_i \sim$ logistic

Only observe D_i:

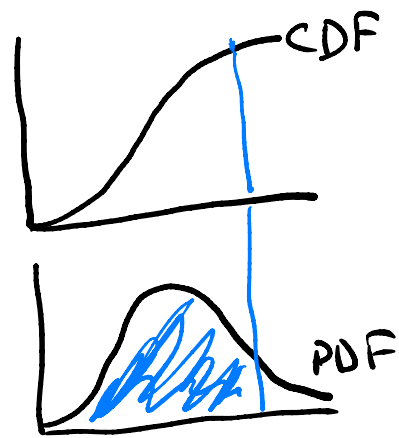
$$D_i = \begin{cases} 1 & \text{if} \quad Y_i > 0 \\ 0 & \text{if} \quad Y_i \leq 0 \end{cases}$$



What is $\Pr[D_i = 1 | X_i]$?

$$\begin{aligned}
\Pr[D_i = 1 | X_i] &= \Pr[Y_i > 0 | X_i] \\
&= \Pr[X_i'\beta + \varepsilon_i > 0 | X_i] \\
&= \Pr[\varepsilon_i > X_i'\beta | X_i] \\
&= 1 - F[-X_i'\beta] \\
&= F[X_i'\beta]
\end{aligned}$$

$G \rightarrow$



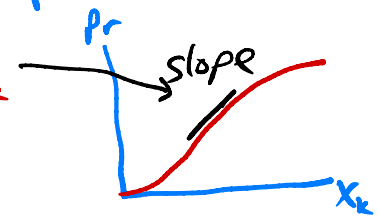
Marginal effect

$$\frac{\partial \Pr[D_i = 1 \mid X_i]}{\partial X_k}$$

← LPM automatically gives us β_k = derivative

↖ Logit/probit models more complex:

$$\frac{\partial \Pr[D_i = 1 \mid X_i]}{\partial X_k} = \frac{\partial F[X_i'\beta]}{\partial X_k} = f(X_i'\beta)\,\beta_k$$



Two options for marg effects in logit/probit:

1. marginal effect at $\bar{X}$
2. average marginal effect across i

Very similar answers

In R: mfx package, logitmfx() and probitmfx()

Percentage points <u>not</u> percent!

Odds ratio

- Odds of event: $\frac{p}{1-p}$
- logit model: $p = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$
- log odds:

$$\ln(\text{odds}) = \ln\left(\frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{1-\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}\right) = \ln\left(e^{x_i'\beta}\right) = X_i'\beta = \beta_0 + \beta_1 X_{1i} + \cdots$$

- so β_k measures derivative of log-odds w.r.t. X_k
- rewrite in terms of odds:

$$\text{odds} = e^{X_i'\beta} = e^{\beta_0 + \beta_1 X_{1i} + \cdots}$$

- odds ratio: $\dfrac{\dfrac{Pr[D_i=1 \mid X_{1i}=1, X_2 \cdots]}{1-Pr[D_i=1 \mid X_{1i}=1, X_2 \cdots]}}{\dfrac{Pr[D_i=1 \mid X_{1i}=0, X_2 \cdots]}{1-Pr[D_i=1 \mid X_{1i}=0, X_2 \cdots]}} = e^{\beta_1} = \dfrac{\text{odds for } X_{1i}=1}{\text{odds for } X_{1i}=0}$

not rel risk