- *Binary Choice Models*
  - Linear Probability Model
  - Index Models for Binary Response
  - Latent variable threshold (LVT) model
  - Random utility models
  - Marginal Effects and Average Partial Effects
  - Simple specification tests
  - Goodness of Fit

In many applications, the variate of interest is binary, i.e., takes only the values 0 and 1.

Examples:

- Labour force participations.

$$Y = \begin{cases} 1 & \text{if employed} \\ 0 & \text{otherwise} \end{cases}.$$

  We would like to study how labour force participation depends on the characteristics of the individuals.

- House ownership

$$Y = \begin{cases} 1 & \text{if a person owns her house} \\ 0 & \text{otherwise} \end{cases}.$$

  We would like to study how house ownership depends on the characteristics of the individuals.

- Denote $X = (X_1, ..., X_k)'$.
- The objective of a regression model is to estimate $E(Y|X)$.

- $E(Y|X) = \mathcal{P}(Y = 1|X)$, when Y is a binary variable.
- In the *linear probability model* we assume that

$$\mathcal{P}(Y = 1|X) = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

- So, the interpretation of β_j is the change in the probability of success when x_j changes:

$$\frac{\partial \mathcal{P}(Y = 1|X)}{\partial X_j} = \beta_j, j = 2, ..., k$$

- The predicted Y is the predicted probability of success.
- The linear probability model is estimated using OLS, that is regressing Y on $X_2, ..., X_k$ (with an intercept).

- Potential problem that the fitted values can be outside $[0, 1]$.
- Even without predictions outside of $[0, 1]$, we may estimate effects that imply a change in x changes the probability by more than $+1$ or -1.
- This model will violate assumption of homoskedasticity, so will affect inference. Notice that

$$\begin{aligned} Var(Y|X) &= \mathcal{P}(Y = 1|X)(1 - \mathcal{P}(Y = 1|X)) \\ &= (\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k) \times \\ &\quad (1 - \beta_1 - \beta_2 X_2 - \ldots - \beta_k X_k). \end{aligned}$$

- Therefore we should use the Eicker-Huber-White robust standard errors to make inference.

- An alternative is to assume that $E[Y|X] = \mathcal{P}(Y = 1|X) = G(X'\beta_0)$, where the function $G(.)$ is <u>known</u> $0 < G(.) < 1$, $\beta_0 = (\beta_1, ..., \beta_k)'$ thus

$$Y = \begin{cases} 1 & \text{with probability } G(X'\beta_0) \\ 0 & \text{with probability } 1 - G(X'\beta_0) \end{cases}$$

- In most applications, $G(.)$ is a cumulative distribution function.
- The framework is similar to the case of the Bernoulli random variable (conditional on the regressors). The Log-Likelihood function is given by

$$\log\{L(\beta)\} = \textstyle\sum_{i=1}^{n} Y_i \log(G(X_i'\beta)) + \sum_{i=1}^{n} (1 - Y_i) \log(1 - G(X_i'\beta)).$$

Differentiating with respect to β we have that the MLE estimator $\hat{\beta}_{ML}$ solves

$$\begin{aligned} \frac{\partial \log\{\mathcal{L}(\hat{\beta}_{ML})\}}{\partial \beta} &= 0 \\ \sum_{i=1}^{n} \left\{ \frac{Y_i - G(X_i'\hat{\beta}_{ML})}{G(X_i'\hat{\beta}_{ML})\left(1 - G(X_i'\hat{\beta}_{ML})\right)} g(X_i'\hat{\beta}_{ML}) X_i \right\} &= 0 \end{aligned}$$

where $g(z) = \partial G(z)/\partial z$.

- Define the **generalized residuals** as

$$\hat{\varepsilon}_i^G = \frac{Y_i - G(X_i'\hat{\beta}_{ML})}{G(X_i'\hat{\beta}_{ML})\left[1 - G(X_i'\hat{\beta}_{ML})\right]} g(X_i'\hat{\beta}_{ML})$$

- Likelihood equations are then given by:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^G X_i = 0.$$

This condition requires $\hat{\varepsilon}_i^G$ and X_i are uncorrelated.

Remarks:

- This is a system of non-linear equations hence we have to resort to numerical methods to solve it. There is no closed form solution for this estimator.
- Consistency and asymptotic normality follow from the general results described for the Maximum Likelihood estimator under some regularity conditions.
- Essentially the main requirement for consistency is that $E[Y|X] = \mathcal{P}(Y = 1|X) = G(X'\beta_0)$.

- In correctly specified models $\hat{\beta}$ is consistent and asymptotically normally distributed with variance-covariance matrix $\left[\mathcal{I}(\beta_0)\right]^{-1}$, that is

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} \mathcal{N}(0, \left[\mathcal{I}(\beta_0)\right]^{-1})$$

where

$$\mathcal{I}(\beta_0) = E\{\frac{g(X'\beta_0)^2 XX'}{G(X'\beta_0)[1 - G(X'\beta_0)]}\}$$

- An estimator for $\mathcal{I}(\beta_0)$ is

$$\mathcal{I}_n(\hat{\beta}_{ML}) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{g(X_i'\hat{\beta}_{ML})^2 X_i X_i'}{G(X_i'\hat{\beta}_{ML})[1 - G(X_i'\hat{\beta}_{ML})]}\}$$

- Inference is done using the Wald, likelihood ratio and Lagrange multiplier statistics.

- The most popular forms of $G(X'\beta_0)$ that are considered in the literature
  - The Logit Model:
  $$G(X'\beta_0) = \frac{\exp(X'\beta_0)}{1 + \exp(X'\beta_0)}.$$
  - The Probit Model:
  $$G(X'\beta_0) = \Phi(X'\beta_0),$$
  where
  $$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du.$$
  is the Standard Normal Distribution Function.
- Both the probit and logit models are nonlinear and require maximum likelihood estimation.
- No real reason to prefer one over the other

Other possible models:

- $G(X'\beta_0) = \exp(-\exp(X'\beta_0))$ [the log-Weibull distribution]
- $G(X'\beta_0) = 1 - \exp(-\exp(X'\beta_0))$ [the Gompertz distribution, known as the Complementary log-log model]
- $G(X'\beta_0) = \Phi(X'\beta_0)^{\tau}, \tau > 0$
- $G(X'\beta_0) = 1 - \left(1 + \omega \exp(X'\beta_0)\right)^{-\frac{1}{\omega}}, \omega > 0$. Note that for $\omega = 1$ we have the logit model and $\lim_{\omega \to 0} G(X'\beta_0) = 1 - \exp(-\exp(X'\beta_0))$.

Remark on the Logit Model

- In statistics a common interpretation of the coefficients is in terms of marginal effects on the odds ratio rather than on the probability.

$$\begin{aligned}\mathcal{P}(Y &= 1|X) = \frac{\exp(X'\beta_0)}{1+\exp(X'\beta_0)} \\ &\Rightarrow \frac{\mathcal{P}(Y=1|X)}{1-\mathcal{P}(Y=1|X)} = \exp(X'\beta_0) \\ &\Rightarrow \log(\frac{\mathcal{P}(Y=1|X)}{1-\mathcal{P}(Y=1|X)}) = X'\beta_0\end{aligned}$$

- $\mathcal{P}(Y=1|X)/(1-\mathcal{P}(Y=1|X))$ measures the probability that $Y=1$ relative to the probability that $Y=0$ and is called the odds ratio or relative risk.
- Example, consider a pharmaceutical drug study where $Y=1$ denotes survival and $Y=0$ denotes death. An odds ratio of 2 means that the probability of survival is twice the probability of death.
- For the logit model the log-odds ratio is linear in the regressors.

Latent variable threshold (LVT) model

- A possible motivation for the specification $E[Y|X] = \mathcal{P}(Y = 1|X) = G(X'\beta_0)$ can be given by considering the *latent variable threshold model*
- Define a latent random variable:

$$Y^* = X'\beta_0 + \varepsilon,$$

  where Y^* is unobserved $\Rightarrow$ **latent variable.**
- Assume: ε independent of X, $E[\varepsilon] = 0$ and $var(\varepsilon) = \sigma^2$ and distribution function $F(.)$
- **Observation rule:**

$$Y = \begin{cases} 1 & \text{if } Y^* > \lambda \\ 0 & \text{if } Y^* \leq \lambda \end{cases} .$$

  That is, the option is chosen if $Y^* > \lambda$, where λ is a threshold
- **Interpretation:** Y^* propensity of an individual towards option, or net benefit from choosing option.

Latent variable threshold (LVT) model

- Probability of choosing the option:

$$
\begin{aligned}
\mathcal{P}[Y=1|X] &= \mathcal{P}[Y^* > \lambda|X] \\
&= \mathcal{P}\left[X'\beta_0 + \varepsilon > \lambda|X\right] \\
&= \mathcal{P}\left[\varepsilon > -X'\beta_0 + \lambda|X\right] \\
&= 1 - \mathcal{P}\left[\varepsilon \leq -X'\beta_0 + \lambda|X\right] \\
&= 1 - F\left(-X'\beta_0 + \lambda\right). \\
&= G(X'\beta_0)
\end{aligned}
$$

with $G(z) = 1 - F(-z+\lambda)$.

- Note that:
$$\mathcal{P}[Y=1|X] = 1 - F\left(-X'\beta_0 + \lambda\right)$$
- If $X_1 = 1$, i.e. there is an intercept in the model, it is not possible to identify separately the intercept and $\lambda \Rightarrow$ solution: set $\lambda = 0$.
- **Remark:** If $\lambda = 0$ and ε has a symmetric distribution around zero (as in the Probit or Logit) $G(z) = F(z)$ as in this case
$1 - F(-z) = F(z)$

Latent variable threshold (LVT) model

Second identification problem:

- Divide Y^* by $a > 0$
$$\frac{Y^*}{a} = X'\beta_0^* + \frac{\varepsilon}{a}$$
where $\beta_0^* = \beta_0 \, / a$
- Note that the definition of the observable variable Y doesn't change. That is
$$\begin{aligned} Y &= \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases} \\ &= \begin{cases} 1 & \text{if } \frac{Y^*}{a} > 0 \\ 0 & \text{if } \frac{Y^*}{a} \leq 0 \end{cases} \end{aligned}$$
where $\beta_0^* = \beta_0 \, / a$.
- This implies that we cannot identify the variance of ε.
- For given β_0^*, value of β_0 depends on a.
- β_0 identified up to a scale factor.
- **Solution:** normalise distribution of ε- Fix σ^2 at a given number $\Rightarrow$ Assume σ^2 is **known**.

- **Example:**
  - Suppose $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$.
  - $P[Y=1|X] = \Phi(X'\frac{\beta_0}{\sigma}) = \Phi(X'\beta_0^*)$.
  - In the case of Probit model we fix $\sigma^2 = 1$ thus $\varepsilon \sim \mathcal{N}(0,1)$ and:

$$\mathcal{P}[Y=1|X] = \Phi(X'\beta_0).$$

Random utility models

- Suppose that an individual has to choose between alternatives a and b, with utilities U^a and U^b.
- The researcher does not observe the utilities, but observes some characteristics of the observation, and writes

$$\begin{aligned} U^a &= X'\beta_a + u_a, \\ U^b &= X'\beta_b + u_b. \end{aligned}$$

  - The researcher observes the chosen alternative, say a, which is indicated by $Y = 1$.
  - Then, we know that

$$\begin{aligned} \mathcal{P}(Y=1|X) &= \mathcal{P}\left(U^a > U^b|X\right) = \Pr\left(X'\beta_a + u_a > X'\beta_b + u_b|X\right) \\ &= \mathcal{P}\left(u_a - u_b > X'\left(\beta_b - \beta_a\right)|X\right) \\ &= \mathcal{P}\left(\varepsilon > -X'\beta_0|X\right) = 1 - F\left(-X'\beta_0\right). \end{aligned}$$

  where $\varepsilon = u_a - u_b$ and $\beta_0 = \beta_a - \beta_b$

- **Whatever the interpretation**, we have to make inference about $\mathcal{P}(Y = 1|x)$.

Marginal Effects and Average Partial Effects

Whatever motivation is used, it is important to note that the parameters of the model, like those of any nonlinear regression model, are not necessarily the marginal effects. In general, the partial effect associated with variable k, (for example) X_{ik}, is given by

$$\Delta \mathrm{E}\left(Y_i \mid X_i\right) \equiv \Delta \mathcal{P}\left(Y_i = 1 \mid X_i\right) \simeq \frac{\partial \mathrm{E}\left(Y_i \mid X_i\right)}{\partial X_{ik}} \Delta X_{ik} = \frac{\partial G\left(X_i'\beta_0\right)}{\partial X_{ik}} \Delta X_{ik}.$$

Let $w = X_i'\beta_0 = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{iK}$. Then

$$\frac{\partial \mathrm{E}\left(Y_i \mid X_i\right)}{\partial X_{ik}} = \frac{dG\left(w\right)}{d\left(w\right)} \frac{\partial\left(w\right)}{\partial X_{ik}} = g\left(w\right)\beta_k = g\left(X_i'\beta_0\right)\beta_k$$

where $g(w) = \partial G(w)/\partial w$, specifically

$$\begin{aligned} \text{Probit} &: \quad g\left(X_i'\beta_0\right) = \phi\left(X_i'\beta_0\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(X_i'\beta_0\right)^2} \\ \text{Logit} &: \quad g\left(X_i'\beta_0\right) = \lambda\left(X_i'\beta_0\right) = \Lambda\left(X_i'\beta_0\right)\left(1 - \Lambda\left(X_i'\beta_0\right)\right) \end{aligned}$$

Marginal Effects and Average Partial Effects

Thus, we need to evaluate

$$\Delta \mathrm{E}\left(Y_i \middle| X_i'\beta_{0i} \right) \equiv \Delta \mathcal{P}\left(Y_i = 1 \middle| X_i \right) \simeq g\left(X_i'\beta_0 \right) \beta_k \Delta X_{ik}.$$

How to deal with $X_i'\beta_0$ for computing marginal effects of variable X_k? Assuming that X_{ik} is continuous we have three options:

- Marginal effects at a particular value $X_i = X_0$

  $$g\left(X_0'\hat{\beta}_{ML} \right) \hat{\beta}_{ML,k} \Delta X_{ik}.$$

  Most common choices of X_0:

  - the vector sample means (doesn't make sense is some of the regressors are dummy variables)
  - the vector of sample medians.

- an alterative is to compute Average Partial Effects, i.e the sample average of the individual marginal effects,

  $$\frac{1}{n} \sum_{i=1}^{n} g\left(X_i'\hat{\beta}_{ML} \right) \hat{\beta}_{ML,k} \Delta X_{ik}.$$

Marginal Effects and Average Partial Effects

Options to calculate the marginal (or partial) effect of a dummy variable d:

- Marginal effects at a particular value $X_i = X_0$. Compute

$$\widehat{\mathcal{P}\left(Y_i = 1 \middle| X_0, d_i = 1\right)} - \widehat{\mathcal{P}\left(Y_i = 1 \middle| X_0, d_i = 0\right)}.$$

  Here X_0 denotes all variables in the model at the particular value X_0 except d.(as before the most common choices of X_0 are the vector of sample means or the vector of sample medians)

- Average Partial Effects:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{P}\left(Y_i = 1 \middle| X_i, d_i = 1\right)} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{P}\left(Y_i = 1 \middle| X_i, d_i = 0\right)}$$

  Here X_i denotes all variables in the model except d.

Simple specification tests

As pointed out above if $G(.)$ is misspecified, then $\hat{\beta}_{ML}$ is inconsistent. Some simple specification tests are available:

- A RESET-type test can be performed by testing $H_0 : \delta_1 = \delta_2 = 0$ in the model

$$E[Y_i|X_i] = G(X_i'\beta_0 + \delta_1(X_i'\hat{\beta}_{ML})^2 + \delta_2(X_i'\hat{\beta}_{ML})^3), i = 1, ..., n$$

  * This is actually a normality test in the probit.

- The model can be tested against more general parametric specifications, which include additional shape parameters. Examples:
  - Consider $G(X'\beta_0) = \Phi(X'\beta_0)^\tau$, and use the score statistic to test $H_0 : \tau = 1$ (Probit)
  - Consider $G(X'\beta_0) = 1 - \left(1 + \omega \exp(X'\beta_0)\right)^{-\frac{1}{\omega}}$, $\omega > 0$. and use the score statistic to test $H_0 : \omega = 1$ (Logit).

Note that heteroskedasticity in the LVT model leads to misspecification of the conditional mean of Y: Define a latent random variable:

$$Y^* = X'\beta_0 + k \times h(Z'\gamma_0)\varepsilon,$$

where Y^* is unobserved. Assume ε independent of X, $E[\varepsilon] = 0$ and $var(\varepsilon) = 1$ and distribution function $F(.)$, Z are a vector function of X of size d and h any function with $h > 0$, $h(0) = 1$, $h'(0) \neq 0$

- $k = 1$ for probit; $k = \sqrt{\pi^2/3}$ for logit.
- Observation rule:

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases} .$$

- In this case

$$
\begin{aligned}
\mathcal{P}[Y=1|X] &= \mathcal{P}[Y^* > 0|X] \\
&= \mathcal{P}\left[X'\beta_0 + kh(Z'\gamma_0)\varepsilon > 0|X\right] \\
&= \mathcal{P}\left[\varepsilon > -\frac{X'\beta_0}{kh(Z'\gamma_0)}|X\right] \\
&= 1-\mathcal{P}\left[\varepsilon \leq -\frac{X'\beta_0}{kh(Z'\gamma_0)}|X\right] \\
&= 1-F\left(-\frac{X'\beta_0}{kh(Z'\gamma_0)}\right). \\
&= G(\frac{X'\beta_0}{kh(Z'\gamma_0)}) \neq G(X'\beta_0)
\end{aligned}
$$

- To test the hypothesis $H_0 : \gamma_0 = 0$ (homoskedasticity), we can construct a LM test based on the so called generalized residuals

- LM test statistic can be calculated as

$$\xi_{LM} = \iota' S(S'S)^{-1} S' \iota \sim \chi^2(d)$$

where ith row of S equal to

$$S_i = (\hat{\varepsilon}_i^G X_i', \hat{\varepsilon}_i^G (X_i' \hat{\beta}_{ML}) Z_i')$$

where $\hat{\varepsilon}_i^G$ are the Generalised residuals and ι is a $(k+d)$ $-$vector of ones.

- This is asymptotically equivalent to testing $H_0 : \gamma_0 = 0$ in the model

$$E[Y_i | X_i] = G(X_i' \beta_0 + (X_i' \hat{\beta}_{ML}) Z_i' \gamma_0)), i = 1, ..., n.$$

- Unlike the Linear Probability Model, where we can compute an R^2 to judge goodness of fit, we need new measures of goodness of fit
- One possibility is a pseudo R^2 based on the log likelihood and defined as $1 - \log(\mathcal{L}_{ur})/\log(\mathcal{L}_r)$. Where $\log(\mathcal{L}_r)$ corresponds to the log-likelihood computed only with the intercept.
- Can also look at the percent correctly predicted – if predict a probability $> .5$ then that matches $Y = 1$ and vice versa.