Econometrics **August 18, 2026**

Qualitative Response Models

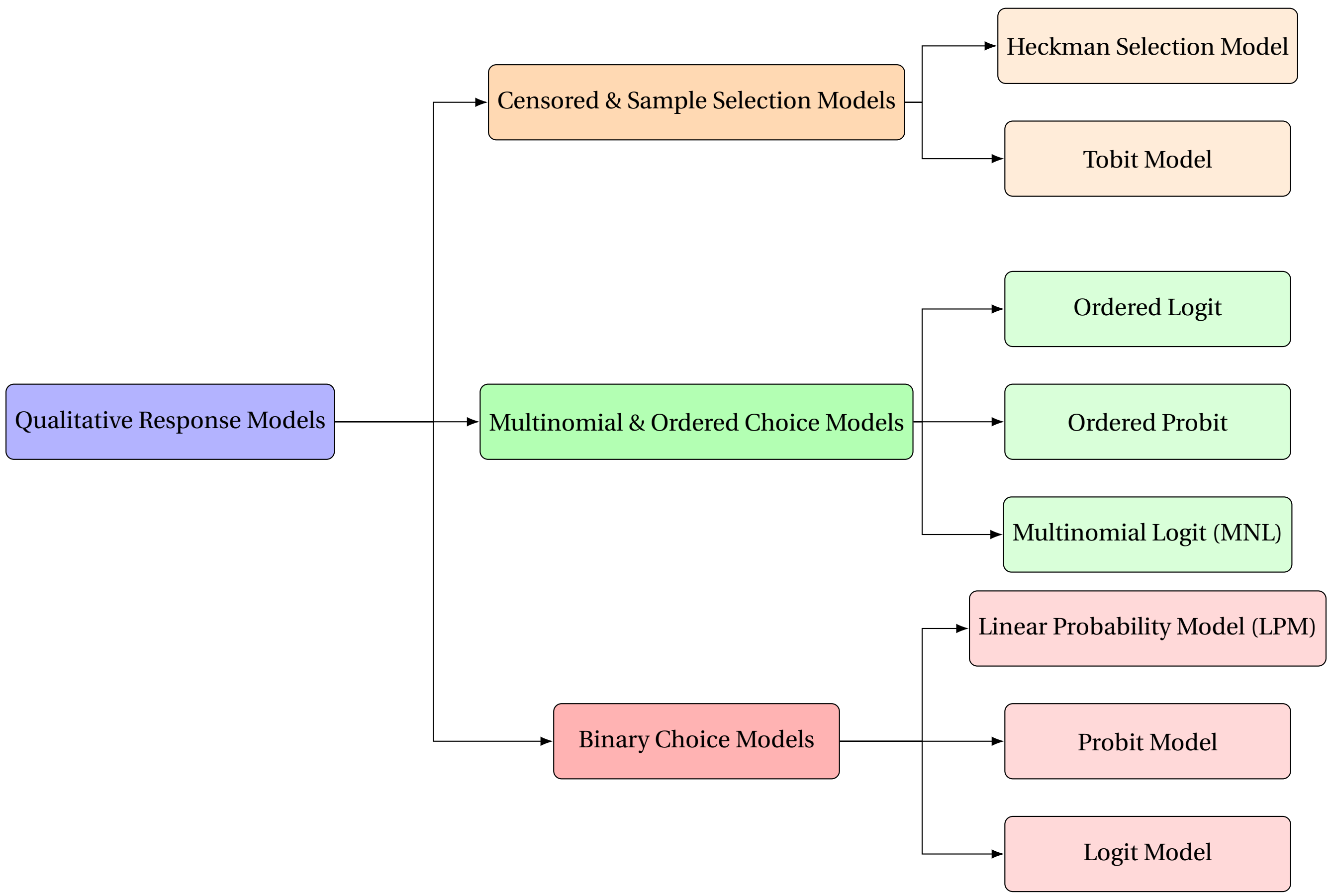
Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

Contents

1 Qualitative Response Models

Qualitative response econometric models are used when the dependent variable Y is categorical. So far, we have seen models where the outcome is continuous.

Figure 1: Taxonomy of Qualitative Response Model



1.1 Linear Probability Model (LPM)

$$Y_i = \begin{cases} 0 & \text{probability } (1 - p_i) \\ 1 & \text{probability } (p_i) \end{cases}$$

$$\begin{aligned} &\boxed{\text{LPM:} \quad Y_i = \beta_i + \beta_1 X_i + \varepsilon_i} \\ &E\left[Y_i|X_i\right] = \beta_i + \beta_1 X_i \\ &\Pr\left(Y_i = 1 \mid X_i\right) = E\left[Y_i \mid X_i\right] = p_i = \beta_0 + \beta_1 (X_i) \end{aligned} \tag{1}$$

1.2 Issues with LPM

1. Non-normality of ε_i (can be sorted by using large 'N')
2. Heteroskedatic ε_i (can be sorted by WLS / GLS)
3. Possibility of $\hat{Y}_i$ lying outside [0,1]
4. Generally lower R^2 values.

Fundamental Problem with LPM is probability p_i rises linearly with X_i. Ideally, though we need $p_i = \hat{Y}_i$ to be constrained between [0,1].

2 Logistic Model

Logistic Regression Model

Logistic regression applies when the dependent variable Y takes on only two values, often coded as $Y \in \{0,1\}$, and the goal is to model the conditional probability $\Pr(Y = 1 \mid X)$ as a function of a vector of covariates X. Typical examples include labor-force participation (working or not), loan default (default or not), or program participation (participate or not). In these settings, applying ordinary least squares to a linear probability model may yield fitted probabilities outside $[0,1]$ and implies a constant marginal effect of each regressor on the probability, which can be unrealistic. (Wooldridge, 2016, ch. 17)

- The logistic regression (logit) model specifies the conditional probability as

$$\Pr(Y = 1 \mid X) = \Lambda(X'\beta) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)},$$

  where $\Lambda(\cdot)$ is the logistic cumulative distribution function and β is a vector of unknown parameters.

- Logistic distribution function: $\boxed{p_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1(X_i)]}} = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}}}$

- Odds Ratio: $\boxed{\frac{p_i}{(1 - p_i)} = e^{Z_i}}$

- Log(Odds ratio) = L_i is the logit/sigmoid function $L_i = log\left(\frac{p_i}{(1 - p_i)}\right) = Z_i = \beta_0 + \beta_1(X_i)$

The model is estimated by maximum likelihood, using the Bernoulli log-likelihood based on independent observations $\{(Y_i, X_i)\}_{i=1}^{n}$. Under standard regularity conditions, the maximum likelihood estimator is consistent and asymptotically normal. (Wooldridge, 2016, ch. 17)

Since Y_i is a *bernoulli random variable*. The joint probability (in this case, a likelihood) of observing $(Y_1, Y_2, \cdots Y_n)$ becomes that of a binomial distribution, where ($Y = 1$ is a success and $Y = 0$ is a failure). Therefore,

$$f(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} [p_i^{Y_i}(1-p_i)^{1-Y_i}]$$

Taking a log and solving yields the following Likelihood function $\mathcal{L}$:

$$\mathcal{L} = \sum_{i}^{n} [Y_i (\beta_0 + \beta_1 X_i)] - \sum_{i}^{n} \left[\log\left(1 + e^{\beta_0 + \beta_1 X_i}\right)\right]$$

Logistic regression is appropriate in empirical settings where:

- The outcome is binary (yes/no, success/failure, participate/not participate).
- The probability that $y = 1$ is assumed to vary smoothly with the covariates and must lie in the unit interval.
- We are interested in odds ratios or marginal effects of regressors on the probability of the event.

Compared with linear probability models, logit naturally restricts fitted probabilities to $[0, 1]$ and allows marginal effects to depend on the values of the regressors. (Wooldridge, 2016, ch. 17)

2.1 Estimating a Logit Model - Max. Log Likelihood

Log-likelihood function (LLF) $\mathcal{L}$:

$$\mathcal{L} = \sum_{i}^{n} [Y_i (\beta_0 + \beta_1 X_i)] - \sum_{i}^{n} \left[\log\left(1 + e^{\beta_0 + \beta_1 X_i}\right)\right]$$

F.O.C:

1. $\dfrac{\partial \mathcal{L}}{\partial \beta_0} = 0 \Rightarrow \sum Y_i - \sum \left(\dfrac{1}{1 + e^{\beta_0 + \beta_1 X_i}} e^{\beta_0 + \beta_1 X_i}\right) = 0.$

2. $\dfrac{\partial \mathcal{L}}{\partial \beta_1} = 0 \Rightarrow \sum Y_i X_i - \sum \left(\dfrac{1}{1 + e^{\beta_0 + \beta_1 X_i}} \times e^{\beta_0 + \beta_1 X_i} \times X_i\right) = 0$

Solving F.O.C is difficult since the resulting expressions in β_0, β_1 are non-linear. So, a numerical trial and error method is employed.

2.2 Interpreting Marginal effects in a Logit Model

$L_i = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X$

$$\frac{1}{(p/1-p)} \times \frac{dP}{dX} \times \frac{(1-p)-(-p)}{(1-p)^2} = \beta_1 \Rightarrow \frac{dP}{dX} = \beta_1 p\,(1-p)$$

Marginal effects in a Logit Model

The change in probability w.r.t X depends on the level of probability 'p' at which it is measured. So it follows that the values of vector of covariates $\vec{X}$ (which builds the value of 'p') matters to get the marginal effect

2.3 Logit Example: Married Women's Labor-Force Participation

Labor force participation (cross-sectional) data

Cross-sectional labor force participation data from T.A. Mroz (1987) A concrete example discussed in Wooldridge's *Introductory Econometrics: A Modern Approach* is the labor-force participation decision of married women, using the `mroz` data set. The binary dependent variable `inlf` equals one if a married woman is in the labor force and zero otherwise.
The logit model can be written as

$$\Pr(\texttt{inlf}_i = 1 \mid x_i) = \Lambda\big(\beta_0 + \beta_1\texttt{nwifeinc}_i + \beta_2\texttt{educ}_i$$
$$= +\beta_3\texttt{exper}_i + \beta_4\texttt{expersq}_i + \beta_5\texttt{age}_i$$
$$= +\beta_6\texttt{kidslt6}_i + \beta_7\texttt{kidsge6}_i\big).$$

In this setting, logistic regression is used to quantify how factors such as education, potential earnings, and presence of young children affect the probability that a married woman participates in the labor force. (Wooldridge, 2016, Example 17.1)

- **inlf**: =1 if in labor force, 1975 (dependent variable)
- **hours**: hours worked, 1975
- **kidslt6**: # kids < 6 years
- **kidsge6**: # kids 6-18
- **age**: woman's age in years
- **educ**: years of schooling
- **wage**: estimated wage from earnings, hours
- **repwage**: reported wage at interview in 1976
- **hushrs**: hours worked by

husband, 1975

- **husage**: husband's age
- **huseduc**: husband's years of schooling
- **huswage**: husband's hourly wage, 1975
- **faminc**: family income, 1975
- **mtr**: federal marginal tax rate facing woman
- **motheduc**: mother's years of schooling
- **fatheduc**: father's years of schooling
- **unem**: unemployment rate in county of residence
- **city**: =1 if living in SMSA
- **exper**: actual labor market experience
- **nwifeinc**: (faminc - wage*hours)/1000
- **lwage**: log(wage)
- **expersq**: exper^2

```r
# R code (to be run separately, e.g. in R or knitr)
library(wooldridge)   # for the mroz data
library(stargazer)    # for LaTeX regression tables

data("mroz")

# Fit logistic regression: inlf on standard covariates
logit_mod <- glm(inlf ~ nwifeinc + educ + exper + I(exper^2) +
                   age + kidslt6 + kidsge6,
                 data = mroz,
                 family = binomial(link = "logit"))

# Show usual summary in the console
summary(logit_mod)

##
## Call:
## glm(formula = inlf ~ nwifeinc + educ + exper + I(exper^2) + age +
##     kidslt6 + kidsge6, family = binomial(link = "logit"), data = mroz)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.425452   0.860365   0.495  0.62095
## nwifeinc    -0.021345   0.008421  -2.535  0.01126 *
## educ         0.221170   0.043439   5.091 3.55e-07 ***
## exper        0.205870   0.032057   6.422 1.34e-10 ***
## I(exper^2)  -0.003154   0.001016  -3.104  0.00191 **
## age         -0.088024   0.014573  -6.040 1.54e-09 ***
## kidslt6     -1.443354   0.203583  -7.090 1.34e-12 ***
```

```
## kidsge6      0.060112   0.074789   0.804  0.42154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  803.53  on 745  degrees of freedom
## AIC: 819.53
##
## Number of Fisher Scoring iterations: 4
```

Table 1: Logit Model for Married Women's Labor-Force Participation

	Dependent variable:
	inlf
nonwife income	-0.021^{**}
	(0.008)
education	0.221^{***}
	(0.043)
experience	0.206^{***}
	(0.032)
experience squared	-0.003^{***}
	(0.001)
age	-0.088^{***}
	(0.015)
kids younger than 6	-1.443^{***}
	(0.204)
kids 6 or older	0.060
	(0.075)
Constant	0.425
	(0.860)
Observations	753
Log Likelihood	-401.765
Akaike Inf. Crit.	819.530
Note:	*p<0.1; **p<0.05; ***p<0.01

3 Probit Model

If a variable X follows normal distribution with mean (μ) & variance $\left(\sigma^2\right)$.

Probability density function

$$\text{pdf} = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

Cumulative Distribution Function

$$\text{CDF} = \int_{-\infty}^{X_0} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

Say, home ownership decision of the i^{th} family depends on an unobservable utility index (I_i) (*latent variable*) that is determined by one or more Explanatory variables (X), where larger the I_i, greater the probability of owning house.

$$I_i = \beta_0 + \beta_1 X_i$$

It is reasonable to assume that I_i can have a threshold level of the index; call it (I_i^*)

- If $I_i < I_i^*$; Home ownership $Y = 0$
- If $I_i > I_i^*$; Home ownership $Y = 1$

$$\begin{aligned}
P_i &= P(Y = 1 \mid X) = P(I_i^* \leq I_i) \\
\Rightarrow P(I_i^* \leq I_i) &= P(Z_i \leq \beta_0 + \beta_1 X_i) \\
&= F(\beta_0 + \beta_1 X_i) \\
Z_i &\sim N(0, \sigma^2) \\
CDF = F(I_i) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-Z^2/2} dZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0+\beta_1(X_i)} e^{-Z^2/2} \mathrm{d}Z \\
P_i &= F_i(I_i) [\text{ probability of owning a home depends on CDF of latent variable } \quad I_i] \\
I_i &= F^{-1}(P_i)
\end{aligned}$$

3.1 Interpreting Probit Estimates

Marginal probability is: $\dfrac{dP_i}{dX_i} = f\left(\beta_0 + \beta_1 X_i\right) \beta_1$.

$f(\beta_0 + \beta_1 X_i)$ is the standard normal probability density function.

Probit model Marginal effects

The value of $\frac{dP}{dX_i}$ depends on the particular value of X_i variables

Logit vs Probit functions for probability

Comparing estimates in binary response models

A simple rule of thumb for scaling up the probit estimates to make them comparable to the logit estimates: multiply the probit estimates by 1.6 Wooldridge, 2010, 594.

3.2 Probit Models in Practice (R)

```
# Load necessary libraries
library(tidyverse)
library(modelsummary)

## Error in library(modelsummary):  there is no package called 'modelsummary'

library(wooldridge)

# Load Married Woman Labour Force Participation dataset by MROZ
data("mroz")

# Run models
lpm_model <- lm(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6
logit_model <- glm(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kids
                   data = mroz, family = binomial(link = "logit"))
probit_model <- glm(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kid
                    data = mroz, family = binomial(link = "probit"))

# Store models in a list
models <- list("LPM (OLS)" = lpm_model,
               "Logit" = logit_model,
               "Probit" = probit_model)
```

```
# Function to calculate Pseudo R² for Logit & Probit
pseudo_r2 <- function(model) {
  if ("glm" %in% class(model)) {
    return(1 - model$deviance / model$null.deviance)  # McFadden's Pseudo R²
  } else {
    return(NA)  # NA for OLS
  }
}

# Extract Pseudo R² values
pseudo_r2_values <- c(NA, pseudo_r2(logit_model), pseudo_r2(probit_model))

# Define custom goodness-of-fit (GOF) statistics
gof_custom <- tibble(
  raw = "Pseudo R²",
  clean = "McFadden's Pseudo R²",
  fmt = 3,
  `LPM (OLS)` = pseudo_r2_values[1],
  Logit = pseudo_r2_values[2],
  Probit = pseudo_r2_values[3]
)
print(gof_custom)

## # A tibble: 1 x 6
##   raw       clean                  fmt `LPM (OLS)` Logit Probit
##   <chr>     <chr>                <dbl>       <dbl> <dbl>  <dbl>
## 1 Pseudo R² McFadden's Pseudo R²     3          NA 0.220  0.221
```

3.2.1 LPM, Logit and Probit, Side by Side

```
# Display results using Stargazer
stargazer(lpm_model, logit_model, probit_model,
          type = "latex",
          title = "LPM, Logit, and Probit Models",
          column.labels = c("LPM (OLS)", "Logit", "Probit"),
          dep.var.labels = "Pr(inlf = 1)",
          covariate.labels = c("Nonwife income", "Education", "Experience",
                               "Experience$^2$", "Age",
                               "Kids <6", "Kids $\\geq$6"),
```

```
# Pseudo R² as custom statistic (stargazer way)
add.lines = list(c("Pseudo R$^2$",
                   sprintf("%.3f", pseudo_r2_values[1]),
                   sprintf("%.3f", pseudo_r2_values[2]),
                   sprintf("%.3f", pseudo_r2_values[3]))),
omit.stat = c("ser", "f", "ll"),
digits = 3,
notes = "Pseudo R$^2$ = 1 - Deviance/Null Deviance")
```

Table 2: LPM, Logit, and Probit Models

	Dependent variable:		
	Pr(inlf = 1)		
	OLS	*logistic*	*probit*
	LPM (OLS)	Logit	Probit
	(1)	(2)	(3)
Nonwife income	-0.003^{**}	-0.021^{**}	-0.012^{**}
	(0.001)	(0.008)	(0.005)
Education	0.038^{***}	0.221^{***}	0.131^{***}
	(0.007)	(0.043)	(0.025)
Experience	0.039^{***}	0.206^{***}	0.123^{***}
	(0.006)	(0.032)	(0.019)
Experience^2	-0.001^{***}	-0.003^{***}	-0.002^{***}
	(0.0002)	(0.001)	(0.001)
Age	-0.016^{***}	-0.088^{***}	-0.053^{***}
	(0.002)	(0.015)	(0.008)
Kids <6	-0.262^{***}	-1.443^{***}	-0.868^{***}
	(0.034)	(0.204)	(0.118)
Kids $\geq$6	0.013	0.060	0.036
	(0.013)	(0.075)	(0.044)
Constant	0.586^{***}	0.425	0.270
	(0.154)	(0.860)	(0.508)
Pseudo R^2	NA	0.220	0.221
Observations	753	753	753
R^2	0.264		
Adjusted R^2	0.257		
Akaike Inf. Crit.		819.530	818.604
Note:	*p<0.1; **p<0.05; ***p<0.01		
	Pseudo R^2 = 1 - Deviance/Null Deviance		

4 Summary

Table 3: Interpretation of Marginal effects (β) in Qualitative response models

Model	**Interpretation of β in the model**
LPM	Change in average value of Y for unit change in X (*ceteris paribus*)
Logit	Change in log(odds) for unit change in X Rate of change of probability of event happening $\frac{dp}{dx_{ji}} = \beta_j P_i(1-P_i)$ β_j is partial correlation coeff. from $L_i = \beta_i + \beta_i x_i + \cdots \beta_j x_{ji} + \cdots + u_i$ calculating P_i requires the value of all $x_{1\ldots n}$
Probit	Rate of change of probability of event happening $\frac{dp}{dx_{ji}} = \beta_j f(Z_i)$ β_j is partial correlation coeff. from $Z_i = \beta_i + \beta_i x_i + \cdots \beta_j x_{ji} + u_i$ $f(Z_i)$ is density function of std. normal dist.

References

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press.

The End