Block 6
Limited dependent variables
(repetition from BSc. courses)

Advanced econometrics 1 4EK608
Pokročilá ekonometrie 1 4EK416

Vysoká škola ekonomická v Praze

Content

Binary dependent variables: LPM, Logit, Probit

Binary dependent variables: $y_i \in \{0; 1\}$

Linear probabilistic model (LPM) – quick repetition

- ✓ Simple estimation through OLS
  easy β_j-parameter interpretation

- ! $\hat{y}_i$ may get beyond the interpretable $\langle 0, 1 \rangle$ probabilistic interval

- ! Marginal (partial) effects of the regressors are constant.

- ! heteroscedastic error term: $\hat{u}_i \sim Bi(0, [\hat{y}_i(1 - \hat{y}_i)])$

Binary dependent variables: LPM, Logit, Probit

Binary dependent variables: Logit and Probit

$$P(y=1|\boldsymbol{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = G(\boldsymbol{x}'\boldsymbol{\beta}) = G(z); \ 0 < G(z) < 1$$

"Success" probability

CDF: function of regressor and the β-parameters

Simplified notation

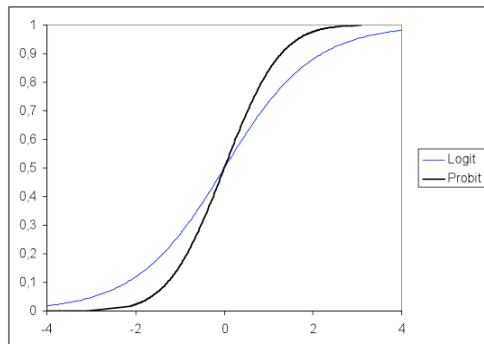Cumulative distribution functions (CDFs) for Logit and Probit

Logit: $G(z) = \Lambda(z) = \exp(z)/[1 + \exp(z)]$

Probit: $G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv; \quad \phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$

Logit vs Probit

Logit: $G(z) = \Lambda(z) = \exp(z)/[1 + \exp(z)]$

Probit: $G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$

$$G(z_i \to -\infty) \qquad \to 0$$
$$G(z_i = 0) \qquad = 0.5$$
$$G(z_i \to +\infty) \qquad \to 1$$

- There are no convincing arguments to prefer one model type over the other.

- Logit is used more often, partly due to simpler estimate calculation.

- In economics, we often assume normal distribution of variables, which would imply the use of Probit.

- With Probit, β_j parameters have no interpretation except for their signs and significance levels.

  (Logit parameters discussed next.)

$$P_i = P(y_i = 1 | \boldsymbol{x}_i) = \hat{y}_i = G(\boldsymbol{x}_i \boldsymbol{\beta}) = G(z) = \frac{e^z}{1+e^z},$$

also:

$\frac{P_i}{1-P_i}$ is the relative chance of success: **Odds ratio**

Example:

$P_i = 0.8 \leftrightarrow (1 - P_i) = 0.2$

Odds ratio: $\frac{0.8}{0.2} = \frac{4}{1}$

Interpetation: the relative chance of success is 4 to 1.

Logit – interpretation of model coefficents β

- $P_i = \frac{e^z}{1+e^z}$

- $1 - P_i = 1 - \frac{e^z}{1+e^z} = \frac{(1+e^z)-e^z}{1+e^z} = \frac{1}{1+e^z}$

- $\frac{P_i}{1-P_i} = \frac{\frac{e^z}{1+e^z}}{\frac{1}{1+e^z}} = e^z$

- $\frac{P_i}{1-P_i} = e^z = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...)}$

- $\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$

- β coefficients have a linear interpretation towards the logarithm of relative chance of success (log-odds) $\ldots$
  Also, $\hat{\beta}_j \approx (\%\Delta)\frac{P_i}{1-P_i}$.

Density function of y_i (given x_i, assuming random sampling):

$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) = [G(\boldsymbol{x}_i\boldsymbol{\beta})]^{y_i}[1 - G(\boldsymbol{x}_i\boldsymbol{\beta})]^{1-y_i}$

Maximum Likelihood Estimation (MLE)
is used to estimate coefficients $\hat{\boldsymbol{\beta}}$.

Computers and iterative methods are used to maximize:

$$LL(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}}\{\textstyle\sum_{i=1}^{N}(y_i \log[G(\boldsymbol{x}_i\boldsymbol{\beta})] + (1 - y_i) \log[1 - G(\boldsymbol{x}_i\boldsymbol{\beta})])\},$$

where $LL(\hat{\boldsymbol{\beta}})$ is the maximized log-likelihood function,

$\qquad G(\cdot)$ is the CDF for Logit or Probit,

$\qquad y_i = \{0; 1\}$

Logit and Probit – estimation

Example: Married women's labor force participation

Dependent Variable: *inlf*			
Independent Variables	**LPM (OLS)**	**Logit (MLE)**	**Probit (MLE)**
nwifeinc	$-.0034$ (.0015)	$-.021$ (.008)	$-.012$ (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
exper^2	$-.00060$ (.00018)	$-.0032$ (.0010)	$-.0019$ (.0006)
age	$-.016$ (.002)	$-.088$ (.015)	$-.053$ (.008)
kidslt6	$-.262$ (.032)	-1.443 (.204)	$-.868$ (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constant	.586 (.151)	.425 (.860)	.270 (.509)
Percentage correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	-401.77	-401.30
Pseudo R-squared	.264	.220	.221

Coefficients are not comparable across models

Often, Logit estimated coefficients $\approx$ 1.6 times Probit estimated because $g_{Logit}(0)/g_{Probit}(0) \approx 1/1.6$.

The biggest difference between the LPM and Logit/Probit is that partial effects are nonconstant in Logit/Probit:

$$\hat{P}(working|\overline{\boldsymbol{x}}, kidslt6 = 0) = .707$$

$$\hat{P}(working|\overline{\boldsymbol{x}}, kidslt6 = 1) = .373$$

$$\hat{P}(working|\overline{\boldsymbol{x}}, kidslt6 = 2) = .117$$

(Larger decrease in probability for the first child)

Logit and Probit – estimation

Given random sampling, MLE is consistent, asymptotically efficient and asymptotically normal estimation function.

Asymptotic variances / standard errors can be used to test hypotheses such as $H_0 : \beta_j = 0$.

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) \equiv \left(\sum_{i=1}^{n} \frac{[g(\boldsymbol{x}_i \hat{\boldsymbol{\beta}})]^2 \boldsymbol{x}_i' \boldsymbol{x}_i}{G(\boldsymbol{x}_i \hat{\boldsymbol{\beta}})[1 - G(\boldsymbol{x}_i \hat{\boldsymbol{\beta}})]} \right)^{-1}$$

$\leftarrow$ This is a $k \times k$ matrix

Multiple (compound) hypotheses may be tested using the **Likelihood ratio** statistics:

$$LR = 2(\log L_{ur} - \log L_r) \sim \chi_q^2$$

where L_{ur} is the $LL(\hat{\boldsymbol{\beta}})$ of the unrestricted (ur) model

$\qquad L_r$ is the $LL(\hat{\boldsymbol{\beta}})$ of the restricted (r) model

$\qquad q$ is the number of restrictions imposed.

The effect of a change in x_j on the outcome (probability of success) may be described as follows:

- Continuous x_j regressors:

$$\frac{\partial P(y=1|\boldsymbol{x}_i)}{\partial x_j} = g(\boldsymbol{x}_i\boldsymbol{\beta})\beta_j \quad \text{where} \quad g(z) \equiv \frac{dG}{dz}(z)$$

- Binary x_k regressors:

$$\Delta P_i = \Delta G(\cdot) = G(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{k-1,i} + \beta_k)-$$
$$- G(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{k-1,i} x_{k-1,i})$$

The effect (of a change in x_{ik}) differs for each individual (CS unit) i and it is a non-linear function of all parameters from the vector $\boldsymbol{\beta}$ and all regressor-values in the row vector $\boldsymbol{x}_i$.

Average partial effects (alternatively labelled as marginal effects) are used to summarize the expected effect of a change in a given regressor on the conditional success probability.

Average Partial Effect (APE):

For a binary regressor x_k , we "simply" average the individual effects across the whole sample:

$$\widehat{\text{APE}}_k = n^{-1} \sum_{i=1}^{N} [G(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_{k-1,i} x_{k-1,i} + \hat{\beta}_k) - \\ - G(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_{k-1,i} x_{k-1,i})]$$

For a continuous regressor x_j:

$$\widehat{\text{APE}}_j = n^{-1} \sum_{i=1}^{n} g(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}) \hat{\beta}_j$$

Estimated model and partial/marginal effects

Example: Estimated model (Logit) – APE

Wooldridge, MROZ data file, dependent variable: labor force participation of married women.

$$logitmfx(\text{formula} = inlf \sim nwifeinc + educ + \text{exper} + age+$$
$$+ kidslt6 + kidsge6, \ \text{data} = \text{mroz}, \ \text{atmean} = F\,)$$

Marginal Effects:

	dF/dx	Std. Err.	z	$P > |z|$	
$nwifeinc$	-0.0036634	0.0015295	-2.3951	0.01662	*
$educ$	0.0411306	0.0085924	4.7868	1.694e-06	***
$exper$	0.0216992	0.0030857	7.0322	2.033e-12	***
age	-0.0165062	0.0029573	-5.5816	2.383e-08	***
$kidslt6$	-0.2608333	0.0428538	-6.0866	1.153e-09	***
$kidsge6$	0.0105417	0.0133312	0.7907	0.42909	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

As an alternative to APE, we may use average x_j regressor value (sample average) to produce another type of "summary effect":

Partial Effect at the Average (PEA)

For any given continuous explanatory variable: x_j:

$$\widehat{\text{PEA}}_j = g(\overline{\boldsymbol{x}}_i \hat{\boldsymbol{\beta}}) \hat{\beta}_j$$

For discrete/binary regressors x_k, the calculation is analogous.

> Warning:
> Averaging discrete variables $\Rightarrow$ non-representative values.

> Example: x_k is *female* and we have 47% of women in the sample. Thus, $\overline{x}_k = 0.47$

R^2: not a suitable statistics for the explained variability of y_i.

McFadden pseudo R^2: $\tilde{R}^2 = 1 - \log L_0 / \log L_{ur}$

where L_0 is $LL(\hat{\boldsymbol{\beta}})$ of the trivial model, often given as:
$$y_i = \beta_0 + u_i$$
(values such as 0.2 or 0.4 – and higher – are regarded as "good")

Correct prediction ratio may be calculated based on:
$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}) > .5 \\ 0 & \text{otherwise} \end{cases}$$

Correlation between y_i and prediction:
$$Corr(y_i, \tilde{y}_i), \ Corr(y_i, G(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}))$$

Estimated LDV model: Confusion matrix

	Predicted 1	Predicted 0
True 1	TP	FN
True 0	FP	TN

		True diagnosis		
		Positive	Negative	Total
Screening test	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Total	$a+c$	$b+d$	N

$Accuracy = (\text{TP} + \text{TN})/N = (a+d)/N$

$Sensitivity = True\ positive\ rate = \text{TP}/Actual\ condition\ positive =$
$= \text{TP}/(\text{TP+FN}) = a/(a+c)$

$Specificity = True\ negative\ rate = \text{TN}/Actual\ condition\ negative =$
$= \text{TN}/(\text{FP+TN}) = d/(b+d)$

$False\ discovery\ rate = \text{FP}/(\text{TP+FP}) = b/(a+b)$

$Prevalence = Condition\ positive/N$

Estimated LDV model: Confusion matrix

		Condition (as determined by "Gold standard")		
Total population		Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population
Test outcome positive		True positive	False positive (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
Test outcome negative		False negative (Type II error)	True negative	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
Positive likelihood ratio (LR+) = TPR/FPR		True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population
Negative likelihood ratio (LR−) = FNR/TNR		False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative	
Diagnostic odds ratio (DOR) = LR+/LR−				
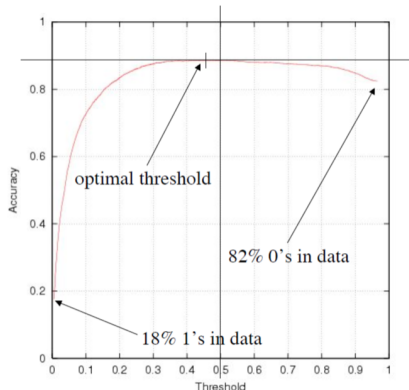
$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}) > .5 \\ 0 & \text{otherwise} \end{cases}$$

By changing the success prediction threshold, we can influence the overal prediction accuracy (classification effectiveness of the model).

To maximize *Accuracy*, we can find an optimal "threshold" within the $\langle 0; 1 \rangle$ interval.

Problem: FP and FN may be associated with different "costs".

Solution: use weighted statistics or use ROC.

Estimated LDV model: ROC curve

- Receiver Operator Characteristic (ROC) Curve
- Developed in WWII to statistically model false positive and false negative detections of radar operators.
- Suitable for evaluating prediction/classification performance of different models and estimation methods used (Logit, Probit, ...)

ROC:
- True Positive Rate vs. False Positive Rate
- Sensitivity vs. (1 – Specificity)

Estimated LDV model: ROC curve
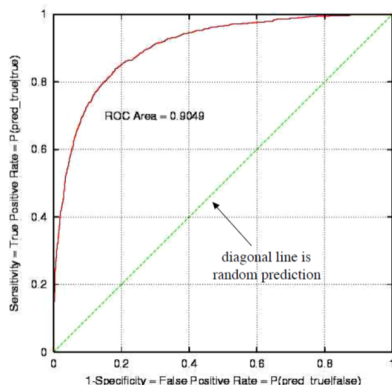
ROC Area – Area under curve (AUC)
1.0 – perfect prediction
0.9 – very good prediction
0.6 – poor prediction
0.5 – random prediction
< 0.5 – something is wrong!

Estimated LDV model: ROC curve

- ROC: slope is non-increasing
- Each point on ROC represents different trade-off (cost ratio) between false positives and false negatives
- AUC (ROC Area) represents model performance averaged over all possible cost ratios
- ROC may not be generalized for multinomial models (possible for *Accuracy*)

Comparison and evaluation of ROC curves:

- If two ROC curves do not intersect, one model (method) dominates the other.
- If two ROC curves intersect, one model (method) is better for some cost ratios, and other model is better for other cost ratios.

Count dependent variables

Introduction

The dependent variable is a count variable, takes on nonnegative integer values: $0, 1, 2, 3, \ldots$

Specifically, Poisson regression applies best when y_i takes on relatively few values (including zero).

The count variable describes the number of specific events that occur during a preset period of time.
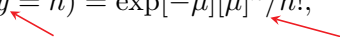
Not to be confused with:

- Binary data: $y \in \{0; 1\}$
- Ordered multinomial data: ranking of the observed events is relevant/important.

Poisson regression model

- Models the relationship between a Poisson-distributed response variable and one or more explanatory variables.

- The explanatory variables can be either continuous, discrete or categorical (factors/dummies).

- The Poisson model predicts the number of occurences of an event with a mean that depends on a set of exogenous regressors $\boldsymbol{x}_i$.

Poisson regression model

$$P(y = h) = \exp[-\mu][\mu]^h/h!, \quad h = 0, 1, 2, \dots$$

Probability that y takes on some integer value of h

Poisson distribution fuction, where: $\mu = E(y) > 0$

Model the mean of the dependent variable as a function of explanatory variables:

$$\mu(\boldsymbol{x}) = \exp(\boldsymbol{x\beta}) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) > 0$$

The Poisson regression model models a count variable as a function of explanatory variables:

$$P(y = h|\boldsymbol{x}) = \exp[-\exp(\boldsymbol{x\beta})][\exp(\boldsymbol{x\beta})]^h/h!, \quad h = 0, 1, 2, \dots$$

Poisson regression model

$$P(y = h|\boldsymbol{x}) = \exp[-\exp(\boldsymbol{x\beta})][\exp(\boldsymbol{x\beta})]^h/h!, \quad h = 0, 1, 2, \ldots$$

May also be re-written as:

$$P(y = h|\boldsymbol{x}) = \frac{e^{-\mu}\mu^h}{h!}$$

where: $\mu = \exp(\boldsymbol{x'\beta})$

and $\mu = E(y|\boldsymbol{x}) = var(y|\boldsymbol{x})$

Poisson regression model

Interpretation of the coefficients of Poisson regression:

$$\frac{\partial \mu(\boldsymbol{x})}{\partial x_j} = \exp(\boldsymbol{x\beta})\beta_j = \mu(\boldsymbol{x})\beta_j \Rightarrow \beta_j = \frac{\frac{\partial \mu(\boldsymbol{x})}{\mu(\boldsymbol{x})}}{\partial x_j}$$

$(\%\Delta)E(y)$ ... by what percentage does the mean (expected) outcome change if x_j is increased by one?

MLE: ML estimation of the Poisson regression model

$$\max \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log P(y = y_i | \boldsymbol{x}_i) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i \boldsymbol{\beta} - \exp(\boldsymbol{x}_i \boldsymbol{\beta})$$

A limitation of the model is that it assumes: $E(y|\boldsymbol{x}) = var(y|\boldsymbol{x})$

ML estimators in the Poisson regression model are consistent and asymptotically normal even if the Poisson distribution assumption (equidispersion) does not hold.

Poisson regression model

	Dependent Variable: narr86	
Independent Variables	**Linear (OLS)**	**Exponential (Poisson QMLE)**
pcnv	−.132	−.402
	(.040)	(.085)
avgsen	.011	−.024
	(.012)	(.020)
tottime	.012	.024
	(.009)	(.015)
*ptime*86	−.041	−.099
	(.009)	(.021)
*qemp*86	−.051	−.038
	(.014)	(.029)
*inc*86	−.0015	−.0081
	(.0003)	(.0010)
black	.327	.661
	(.045)	(.074)
hispan	.194	.500
	(.040)	(.074)
*born*60	−.022	−.051
	(.033)	(.064)
constant	.577	−.600
	(.038)	(.067)
Log-likelihood value	—	-2,248.76
R-squared	.073	.077
$\hat{\sigma}$	.829	1.232

Wooldridge (crime1) example:

The expected number of 1986 arrests falls by 2.4 % if "*avgsen*" increases by one unit (one month).

Note the different values and different meanings of the OLS-generated coefficients.

Poisson regression model

Violation of the **equidispersion** assumption: $E(y|\boldsymbol{x}) = var(y|\boldsymbol{x})$

- Often violated in real (economic) data: $E(y|\boldsymbol{x}) < var(y|\boldsymbol{x})$
- Most often caused by highly skewed dependent variables.

Impacts of **overdispersion**:

- Underestimated standard errors lead to overestimated significance of the estimated parameters.
- Predicted outcomes (counts / realizations of the dependent variable) are not necessarily skewed.

Negative Binomial and Poisson distributions

Poisson vs. NB Distribution with μ = 10
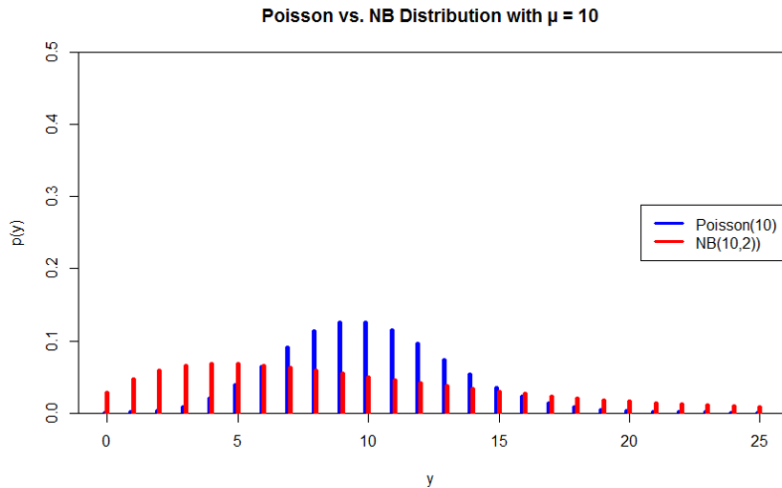
Overdispersion

Observed variance > theoretical Poisson variance

$var(y) = \mu + \alpha \times f(\mu)$, ($\alpha$: dispersion parameter)

- Observed variance & implications

$\alpha = 0$ Poisson distribution, Poisson model

$\alpha > 0$ Overdispersion (common count variable behavior),
NB distribution, NB model

$\alpha < 0$ Underdispersion (not common)

- $y \sim \mathrm{NB}(\mu, \alpha)$

- $\mu = E(y|\boldsymbol{x}) = \exp(\boldsymbol{x}'\boldsymbol{\beta})$

- $\alpha \ldots$ denotes the dispersion parameter (not variance)

- $var(y|\boldsymbol{x}) = \mu + \alpha\mu^2$, alternatively $var(y|\boldsymbol{x}) = \mu + \alpha\mu$

Negative binomial distribution can be used to model count data
with overdispersion if Poisson model is not suitable ($\alpha > 0$):

$$P(Y = y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y}, \; y = 0, 1, 2, \ldots$$

where $\Gamma(\cdot)$ is the CDF for Gamma distr., defined for $r > 0$
by the following expression:

$$\Gamma(r) = \int_0^\infty z^{r-1}\exp(-z)dz.$$

where $\mu_i = \exp(\boldsymbol{x}'\boldsymbol{\beta}) = E(y_i|\boldsymbol{x}_i)$

$$\mu_i = \exp(\boldsymbol{x}'\boldsymbol{\beta}) = E(y|\boldsymbol{x}_i)$$

Maximum Likelihood for Poisson regression

$$LL = \sum_{i=1}^{n} [-\mu_i + y_i \log(\mu_i) \underbrace{- \log(y_i!)}]$$

This drops out from the *LL*-maximization as its not a function of β

Maximum Likelihood for Negative-Binomial regression

$$LL = \sum_{i=1}^{n} \left[\log \left(\frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})} \right) - (y_i + \alpha^{-1}) \log(1 + \alpha\mu_i) + y_i \log(\alpha\mu_i) \right]$$

Estimated parameters: Poisson/NB model

Poisson regression

$\beta_0, \beta_1, \ldots, \beta_k$

NB regression

$\beta_0, \beta_1, \ldots, \beta_k$ **and** α

Zero-inflated and Hurdle models

Used to handle the "excess zeroes" issue in the dependent variable.

Zero-Inflated models:
Instead of assuming that count outcome comes from a single data generating process, we can assume that the count outcome is generated by two systematically different statistical processes:

1. Logit model
   (possibility of non-zero outcome is modelled)

2. Poisson or Negative-Binomial model
   (model expected outcome for the group where non-zero outcome is assumed possible)

Hurdle models:
Often used for modelling DGP associated with two-stage decision making: In the first stage, individuals decide on participation (purchasing activity) and in the second stage, they decide upon intensity (how many "items" are purchased).

Zero-inflated and Hurdle models

ZI models

Ex.1: Internet usage at work
We model hours spent on the internet per day
Some people don't have internet at work.
For those who have, Poiss/NB is used.

Ex.2: Marathons per year
People are/are not active athletes.
For athletes, # marathons is modelled by Poiss/NB.

Hurdle models

Ex.1: Predict purchased amounts
Decision upon purchase.
Decision upon (non-zero) amount.

Ex.2: Doctor visits per year
Does the person visit a doctor at all? If yes, how may times?
(You decide on GP "first" visit, they may send you to specialists.)

Zero-inflated models

We assume two types of individuals (C-S units) in the sample

Type A: Outcome is always zero.
$\text{Prob}(y_i = 0) = 1, \ \text{Prob}(y_i > 0) = 0.$

Type non-A: Non-zero chance of positive count value.
Probability is variable (given regressor values) and follows
Poison/NB distributions.

Binary regime indicator:
$z = 0$ for type A, $z = 1$ for type non-A.

$\text{Prob}(y_i = 0 | \boldsymbol{x}_i) = \text{Prob}(z_i = 0) + \text{Prob}(y_i = 0 | \boldsymbol{x}_i, z_i = 1) \cdot \text{Prob}(z_i = 1)$
$\text{Prob}(y_i = j | \boldsymbol{x}_i) = \text{Prob}(y_i = j | \boldsymbol{x}_i, z_i = 1) \cdot \text{Prob}(z_i = 1), \quad j = 1, 2, \ldots$

Zero-inflated models

$\text{Prob}(y_i = 0 | \boldsymbol{x}_i) = \text{Prob}(z_i = 0) + \text{Prob}(y_i = 0 | \boldsymbol{x}_i, z_i = 1) \cdot \text{Prob}(z_i = 1)$

$\text{Prob}(y_i = j | \boldsymbol{x}_i) = \text{Prob}(y_i = j | \boldsymbol{x}_i, z_i = 1) \cdot \text{Prob}(z_i = 1), \quad j = 1, 2, \ldots$

May be extended by modelling z-regime:

1. Use logit/probit to model group membership (A vs non-A)

2. Use Poisson or NB to model counts for the non-A group

ZI model (Logit/Poisson) example:

$\text{Prob}(z_i = 0 | \boldsymbol{w}_i) = \frac{\exp(\boldsymbol{w}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{w}_i'\boldsymbol{\beta})}, \qquad$ for $z_i = 0$, y_i is always zero.

$\text{Prob}(y_i = j | \boldsymbol{x}_i, z_i = 1) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}, \quad$ for $z_i = 1$, y_i a count variable.

Testing for ZI redundancy (H_0 of Poisson vs. H_1 ZI Poisson) is non-trivial. The two distributions are different, e.g. H_0 and H_1 are non-nested. See Greene (ch. 18) and R for available tests.

Like ZI models, Hurdle count models are two-component models

- Hurdle component models the zero counts.
- Truncated count component for positive counts.

Unlike ZI models, there is only one source of zeros.

Count model is only employed if the hurdle for modeling the occurrence of zeros is exceeded.

Count model is typically a truncated Poisson or NB regression

Hurdle models

Hurdle model (Logit/Poisson) example:

$$\text{Prob}(y_i = 0 | \boldsymbol{w}_i) = \frac{\exp(\boldsymbol{w}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{w}_i'\boldsymbol{\beta})} = \Lambda(\boldsymbol{w}_i'\boldsymbol{\beta})$$

$$\text{Prob}(y_i = j | \boldsymbol{x}_i, \boldsymbol{w}_i, y_i > 0) = \frac{[1-\Lambda(\boldsymbol{w}_i'\boldsymbol{\beta})]\exp(-\lambda_i)\lambda_i^j}{j![1-\exp(-\lambda_i)]}.$$

Again, it is complicated to test for Hurdle model redundancy H_0 of Poisson vs. H_1 Hurdle (say, logit+Poisson)

The two distributions are non-nested.
See Greene (ch. 18) and R for available tests.

Count data models in `R`

Poisson model   R: {stats} package

```
glm(y ~ x, family = poisson, ...)
```

NB model   R: {MASS} package

```
glm.nb(y ~ x, ...)
```

Zero-Inflated model   R: {pscl} package

```
zeroinfl(y ~ x|z, link = "logit", dist = "poisson", ...)
zeroinfl(y ~ x|z, link = "logit", dist = "negbin", ...)
```
    $x1 + x2$ are the count regressors and $z1 + z2$ are regressors of
    the ZI component. [overlap among x and z is possible]

Hurdle model   R: {pscl} package

```
hurdle(y ~ x|z, dist="poisson", zero.dist="binomial" ...)
```
    $x1 + x2$ are the count regressors and $z1 + z2$ are regressors of
    the zero hurdle model. [overlap among x and z is possible]

Introduction

Multinomial dependent variables $\leftrightarrow$ discrete response models for > 2 outcomes

Two basic types of outcomes / "choices" / dependent variables:

Unordered alternatives (nominal outcome)
  Examples: occupational choice, commuting & transportation choice . . .

Ordered alternatives (ordered outcome)
  Examples: Bond ratings, credit ratings . . .

Logit and Probit (MLE for binary dependent variables)

- May be extended to cases with more than 2 possible outcomes.

- Ordered and unordered outcomes
  – differ in both model estimation and result interpretation.

- Alternatives (choices, outcomes) for the i-th individual: $y_i \in \{0, 1, \ldots, J\}$

  ... i.e. y takes on $J + 1$ "values"
       (both ordered & unordered types).

$y_i \in \{0, 1, \ldots, J\}$

Note: outcome ordering is arbitrary, as well as the choice of the base outcome, i.e. the choice of i where $y_i = 0$.

Two basic model-types for unordered responses exist:

Multinomial logit (MNL)

y is an unordered response and we have a set of conditioning variables $\boldsymbol{x}$, that change by unit but not by alternative. For example, we assume that health-plan choice depends on the age and wage of individuals but not on features (e.g. costs) of alternative health-plans.

Probabilistic choice model (Conditional logit model: CL)

Conditioning variables $\boldsymbol{x}$ can change by units and alternatives. Transportation choice: individual's characteristics & time and cost of travel.

In this model, we are interested in the response (outcome) probabilities:

$$p_j(\boldsymbol{x}) = P(y = j | \boldsymbol{x}), \qquad j = 0, 1, \ldots, J.$$

Since exactly one choice is possible
(each individual / CS unit chooses one alternative),

$$p_0(\boldsymbol{x}) + p_1(\boldsymbol{x}) + \cdots + p_J(\boldsymbol{x}) = 1$$

allways holds, for all $\boldsymbol{x}$.

MNL response probabilities are given as

$$P(y = j|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta}_j)}{[1+\sum_{h=1}^{J} \exp(\boldsymbol{x}\boldsymbol{\beta}_h)]}, \quad j = 1, \ldots, J$$

$$P(y = 0|\boldsymbol{x}) = \frac{①}{[1+\sum_{h=1}^{J} \exp(\boldsymbol{x}\boldsymbol{\beta}_h)]}$$

vector $\boldsymbol{x}$ usually contains an intercept.

The vector of coefficients for the base outcome is normalized to $\boldsymbol{\beta}_0 = \mathbf{0}$ (zero vector)

$\ldots \exp(\boldsymbol{x}\mathbf{0}) \ldots e^0 = 1$

$\ldots$ despite the fact that $\boldsymbol{\beta}_0 = \mathbf{0}$, $P(y = 0|\boldsymbol{x}) \neq 0$

Once $P(y = j|\boldsymbol{x})$ for $j = 1, \ldots, J$ are known (estimated), we may simply calculate $P(y = 0|\boldsymbol{x})$: it adds up to 1.

Maximum likelihood estimation of the $\boldsymbol{\beta}_j$ is straightforward.
The log likelihood for random draw of $(\boldsymbol{x}_i, y_i)$ is:

$$LL_i(\boldsymbol{\beta}) = \sum_{j=0}^{J} 1[y_i = j] \log[p_j(\boldsymbol{x}_i \boldsymbol{\beta}_j)]$$

Standard inference applies.

Partial effect interpretation is complicated

(continuous regressor x_k example):

$$\frac{\partial p_j(\boldsymbol{x})}{\partial x_k} = p_j(\boldsymbol{x}) \left\{ \beta_{jk} - \frac{\left[\sum_{h=1}^{J} \beta_{hk} \exp(\boldsymbol{x}\boldsymbol{\beta}_h) \right]}{\left[1 + \sum_{h=1}^{J} \exp(\boldsymbol{x}\boldsymbol{\beta}_h) \right]} \right\}$$

In this equation, RHS may differ in sign from β_{jk}

Easier to interpret:

$$\frac{p_j(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_j)$$

MDVs: Multinomial logit (MNL)

$\frac{p_j(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_j)$ motivation

j-th outcome probability given by:

$P(y=j|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta}_j)}{\left[1+\sum_{h=1}^{J}\exp(\boldsymbol{x}\boldsymbol{\beta}_h)\right]}, \quad j = 1, \ldots, J$

For outcomes j and h, we can write:

$P(y=j|\boldsymbol{x} \cup y=h|\boldsymbol{x}) = p_j(\boldsymbol{x}, \boldsymbol{\beta}) + p_h(\boldsymbol{x}, \boldsymbol{\beta}),$

Hence:

$P(y=j \mid y=j \cup y=h, \boldsymbol{x}) = \frac{p_j(\boldsymbol{x},\boldsymbol{\beta})}{\left[p_j(\boldsymbol{x},\boldsymbol{\beta})+p_h(\boldsymbol{x},\boldsymbol{\beta})\right]} = \boldsymbol{G}(\boldsymbol{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h))$

which may be re-written (if we arbitrarily set $h = 0$):

$P(y=j \mid y=j \cup y=0, \boldsymbol{x}) = \frac{p_j(\boldsymbol{x},\boldsymbol{\beta})}{\left[p_j(\boldsymbol{x},\boldsymbol{\beta})+p_0(\boldsymbol{x},\boldsymbol{\beta})\right]} = \boldsymbol{G}(\boldsymbol{x}\boldsymbol{\beta}_j)$

where $\boldsymbol{G}(\cdot)$ is the logistic function.

Interpretation based on the odds-ratio
(relative chance of success):

Odds-ratio for the j-th and h-th outcome

$$\frac{p_j(\boldsymbol{x})}{p_h(\boldsymbol{x})} = \frac{P(y=j|\boldsymbol{x})}{P(y=h|\boldsymbol{x})} = \exp[\boldsymbol{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)] = e^{\boldsymbol{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)},$$

Odds-ratio for the j-th and base outcome
($\boldsymbol{\beta}_0 = \boldsymbol{0}$ by definition; choice of base outcome is arbitrary)

$$\frac{p_j(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \frac{P(y=j|\boldsymbol{x})}{P(y=0|\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_j) = e^{\boldsymbol{x}\boldsymbol{\beta}_j}$$

For $j = 0, 1, \ldots, J$ (i.e. for $J + 1$ alternatives), we estimate J vectors $\boldsymbol{\beta}_j$ of the MNL model by MLE. Also, we have the $\boldsymbol{\beta}_0 = \mathbf{0}$ vector.

Now, we can calculate and interpret the relative chances:

$$\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_1)$$

$$\frac{p_2(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_2)$$

$\ldots$

$$\frac{p_J(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_J)$$

From $\frac{p_j(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_j)$, we can see that $\log\left[\frac{p_j(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right] = \boldsymbol{x}\boldsymbol{\beta}_j$.

Coefficient β_{jk} measures the linear partial effect of x_{ik} on the log odds of j-outcome relative to base outcome.

Hence, the β_{jk} coefficient also measures the expected relative change in odds ratio (chance of choosing outcome j instead of outcome 0):

For a continuous regressor x_k:

$\hat{\beta}_{jk} \approx (\%\Delta)\frac{p_j}{p_0}$,

or

$(\exp(\hat{\beta}_{jk}) - 1) \times 100 = (\%\Delta)\frac{p_j}{p_0}$

MDVs: Multinomial logit (MNL)

Base choice (outcome)

Additional grade-point in "science" test results in lowering the odds ratio P(Major2)/P(Major1) by approx. 2.4 %

Additional grade-point in "science" test increases the odds ratio P(Major3)/P(Major1) by approx. 2,3 %

Major1	Coef.	Std. Err.	z	P > |z|	[95% Conf. Interval]
Major2						
science	-.0235647	.0209747	-1.12	0.261	-.0646744	.017545
socst	-.0389243	.0195165	-1.99	0.046	-.0771759	-.0006726
female	.8166202	.3909813	2.09	0.037	.050311	1.582929
__cons	1.912256	1.127256	1.70	0.090	-.2971258	4.121638
Major3						
science	.022922	.0208718	1.10	0.272	-.0179861	.0638301
socst	.0430036	.0198894	2.16	0.031	.0040211	.081986
female	-.032862	.3500153	-0.09	0.925	-.7188793	.6531553
__cons	-4.057323	1.222939	-3.32	0.001	-6.45424	-1.660407

$[\exp(0.82) - 1] \times 100$
...odds ratio increased by 126 %.

Odds ratio decreases by 3.3 %.

Unordered nominal outcomes described by y_i

$\boldsymbol{x}_{ij}$ can change over i (individuals, CS units)
and across j (otcomes, choices).

For example, we assume that transportation choice depends on
wealth, income, gender and age of individuals – as well as on
time and cost of different transportation choices.

In fact, MNL is a special case of the CL model.

Quantitative analysis can be based on a model maximizing the latent utility function:

$$y_{ij}^* = \boldsymbol{x}_{ij}\boldsymbol{\beta} + a_{ij}, \ \ j = 0, \ldots, J,$$

where $\boldsymbol{x}_{ij}$ change over i and j,

vector $\boldsymbol{\beta}$ does not depend on the outcome j,

a_{ij}: unobservable individual effects (random elements).

If the $\{a_{ij} : j = 0, 1, \ldots, J\}$ are independent, identically distributed with the *type I* extreme value distribution, that is, with cdf $F(a) = \exp[-\exp(-a)]$, then it can be shown that

$$P(y_i = j | \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})}{\left[\sum_{h=0}^{J} \exp(\boldsymbol{x}_{ih}\boldsymbol{\beta})\right]}, \quad j = 0, 1, \ldots, J.$$

This is the probabilistic choice model, often called the conditional logit model (name given by McFadden)

Note: The *type I* extreme value distribution is not especially natural because it is not symmetric - it has a thicker right tail. But it does roughly have a "bell shape".

For interpretation in the CL model, we may use the odds ratio:

$$\frac{p_j(\boldsymbol{x}_{ij})}{p_h(\boldsymbol{x}_{ih})} = \frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{ih}\boldsymbol{\beta})} = \exp[(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ih})\boldsymbol{\beta}]$$

A key restriction of the CL model is
independence from irrelevant alternatives (IIA).

IIA: probability of selecting between alternatives j and h (odds ratio) does not depend on characteristics of other choices – that is, regressors for the m-th alternative $\boldsymbol{x}_{im}$ for $m \notin \{j, h\}$ do not appear in the odds ratio.

Independence from irrelevant alternatives (IIA assumption)

- Problematic for similar alternatives (some outcomes are close substitutes, a_{ij} may not be independent among different j values).

Blue bus / red bus (extreme) example

Say, the transportation choice is between a car and a blue bus. The probability of each is 0.5. According to the IIA assumption, the introduction of a red bus (equivalent to blue bus in all aspects but color) will not change the odds ratio between blue bus and car, hence probabilities for the three choices must change to $1/3$ (assuming commuters do not have systematic preferences concerning bus color).

Means to relax IIA exist (e.g. using multinomial Probit instead of Logit).

Hausman test for IIA assumption validity

H_0: IIA holds (odds ratio for outcomes j and h is independent of irrelevant alternatives)

Test:

1) Drop one "irrelevant" outcome (dependent variable category, other than j or h)

2) Re-estimate the model and check for significant changes in coefficients.

3) Repeat for all "irrelevant" alternatives.

If H_0 is rejected:

$\rightarrow$ We cannot use the CL model

$\rightarrow$ Use multinomial Probit or hierarchical model (Logit).
(see Wooldridge, Econometric analysis of cross section and panel data)

CL: Model prediction effectiveness evaluation
   example: transport choice

```
Confusion Matrix and Statistics


          Reference
Prediction air train bus car
     air   173    4  37   0
     train  23  132  70   0
     bus    14   74 103   0
     car     0    0   0 210
```

	Class: air	Class: train	Class: bus	Class: car
Sensitivity	0.8238	0.6286	0.4905	1.00
Specificity	0.9349	0.8524	0.8603	1.00
Pos Pred Value	0.8084	0.5867	0.5393	1.00
Neg Pred Value	0.9409	0.8732	0.8351	1.00
Prevalence	0.2500	0.2500	0.2500	0.25
Detection Rate	0.2060	0.1571	0.1226	0.25
Detection Prevalence	0.2548	0.2679	0.2274	0.25
Balanced Accuracy	0.8794	0.7405	0.6754	1.00

Ordered Logit/Probit models are used to estimate relationships between an ordinal dependent variable and a set of independent variables.

An ordinal variable is a variable that is categorical and ordered, for instance, "poor", "good", and "excellent".

Ordinal variables – interpretation:

Say, y is a credit rating on a scale from '0' to '6' (6 is the best rating). Then, credit ratings have ordinal meanings only! (6 better than 2)

We cannot say that the difference between '4' and '2' is twice as important/prominent as the difference between '6' and '5'.

Ordered Probit model may be derived from a latent variable model.

$$y^* = \boldsymbol{x}\boldsymbol{\beta} + e, \ e|\boldsymbol{x} \sim Normal(0,1)$$

Interpretation example:

y_i^* latent variable describing reported health,

$\boldsymbol{x}_i$ individual health factors.

The latent index y_i^* measures respondent-specific scale of health. This latent scale may differ across individuals i !

Once y_i^* crosses certain - individual-specific - value, respondent report corresponding "observed" outcomes, such as:
$y_i \in \{$ bad, poor, good, very good, excellent $\}$

MDVs: Ordered responses

Ordered Probit Model

$$y^* = \boldsymbol{x\beta} + e, \ e|\boldsymbol{x} \sim N(0,1)$$

Observation rules (the relationship between latent and observed y):

$$y_i = 0 \text{ for } \qquad y_i^* \leq \alpha_1$$
$$y_i = 1 \text{ for } \quad \alpha_1 < y_i^* \leq \alpha_2$$
$$y_i = 2 \text{ for } \quad \alpha_2 < y_i^* \leq \alpha_3$$
$$\dots$$
$$y_i = J \text{ for } \qquad y_i^* > \alpha_J$$

where $\boldsymbol{x}$ regressors, excludes constant (!)

$\quad y \quad$ $y_i \in \{0, 1, 2, \dots, J\}$ $\dots J + 1$ ordered alternatives

$\quad \alpha_j \quad$ unknown (individual) cutpoints/thresholds (ordered); $j = 1, \dots, J$; $\alpha_1 < \alpha_2 < \cdots < \alpha_J$

       (distances between successive cutpoints may differ)

Threshold values $(\alpha_1, \alpha_2, \ldots, \alpha_J)$ are unknown. We do not know the value of the index necessary to "push" someone from reporting *good* to *very good*.

Threshold values $(\alpha_1, \alpha_2, \ldots, \alpha_J)$ are different for each individual (at least in theory).

Ordered Probit/Logit methods not only estimate the vector $\boldsymbol{\beta}$, but also the thresholds $\boldsymbol{\alpha}$ – i.e. their averages across individuals.

$$P(y = 0|\boldsymbol{x}) + P(y = 1|\boldsymbol{x}) + \cdots + P(y = J|\boldsymbol{x}) = 1$$

For Ordered Probit: different outcome probabilities are given as:

$$P(y = 0|\boldsymbol{x}) = P(\boldsymbol{x}\boldsymbol{\beta} + e \leq \alpha_1|\boldsymbol{x}) = \Phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta})$$

$$P(y = 1|\boldsymbol{x}) = P(\alpha_1 < \boldsymbol{x}\boldsymbol{\beta} + e \leq \alpha_2|\boldsymbol{x}) = \Phi(\alpha_2 - \boldsymbol{x}\boldsymbol{\beta}) - \Phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta})$$

$$\vdots$$

$$P(y = J - 1|\boldsymbol{x}) = \Phi(\alpha_J - \boldsymbol{x}\boldsymbol{\beta}) - \Phi(\alpha_{J-1} - \boldsymbol{x}\boldsymbol{\beta})$$

$$P(y = J|\boldsymbol{x}) = 1 - \Phi(\alpha_J - \boldsymbol{x}\boldsymbol{\beta})$$

For Ordered Logit, we simply substitute $\Phi(\cdot)$ for the Logistic CDF: $\Lambda(\cdot)$.

MLE is used for estimation. The log-likelihood function for the i-th individual (CS unit) is as follows:

$$LL_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1[y_i = 0] \log[\Phi(\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta})]+$$
$$+ 1[y_i = 1] \log[\Phi(\alpha_2 - \boldsymbol{x}_i\boldsymbol{\beta}) - \Phi(\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta})]+$$
$$+ \cdots + 1[y_i = J] \log[1 - \Phi(\alpha_J - \boldsymbol{x}_i\boldsymbol{\beta})]$$

When $J = 1$, $P(y = 0|\boldsymbol{x}) = \Phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta}) = 1 - \Phi(\boldsymbol{x}\boldsymbol{\beta} - \alpha_1)$,
$P(y = 1|\boldsymbol{x}) = \Phi(\boldsymbol{x}\boldsymbol{\beta} - \alpha_1)$, hence - α is the intercept.

Interpreting coefficients requires some care:

$p_0 = P(y = 0|\boldsymbol{x})$; i.e. expression describes the change in P of the "worst" outcome

change in P of the "best" outcome

$$\frac{\partial p_0(\boldsymbol{x})}{\partial x_k} = -\beta_k \phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta}), \quad \frac{p_J(\boldsymbol{x})}{\partial x_k} = \beta_k \phi(\alpha_J - \boldsymbol{x}\boldsymbol{\beta})$$

$$\frac{\partial p_j(\boldsymbol{x})}{\partial x_k} = \beta_k[\phi(\alpha_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - \phi(\alpha_j - \boldsymbol{x}\boldsymbol{\beta})]$$

change in P of an intermediate outcome j

For a small change in x_k, the sign of the resulting effect on $P(y = 0|\boldsymbol{x})$ and $P(y = J|\boldsymbol{x})$ is unambiguously determined by the sign of β_k.

Interpreting coefficients requires some care:

$$\frac{\partial p_0(\boldsymbol{x})}{\partial x_k} = -\beta_k \phi(\alpha_1 - \boldsymbol{x\beta}), \quad \frac{p_J(\boldsymbol{x})}{\partial x_k} = \beta_k \phi(\alpha_J - \boldsymbol{x\beta})$$

$$\frac{\partial p_j(\boldsymbol{x})}{\partial x_k} = \beta_k [\phi(\alpha_{j-1} - \boldsymbol{x\beta}) - \phi(\alpha_j - \boldsymbol{x\beta})]$$

change in P of a given intermediate outcome j

For a small change in x_k, the sign of the resulting effect on

$P(y = j|\boldsymbol{x})$ is ambiguous, it depends on the ratio of

$|\alpha_{j-1} - \boldsymbol{x\beta}|$ vs. $|\alpha_j - \boldsymbol{x\beta}|$

... (remember, $\phi(\cdot)$ is symmetric around zero).

MDVs: Ordered responses

Interpreting coefficients:

$$\frac{\partial p_0(\boldsymbol{x})}{\partial x_k} = -\beta_k \phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta}), \quad \frac{p_J(\boldsymbol{x})}{\partial x_k} = \beta_k \phi(\alpha_J - \boldsymbol{x}\boldsymbol{\beta})$$

$$\frac{\partial p_j(\boldsymbol{x})}{\partial x_k} = \beta_k[\phi(\alpha_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - \phi(\alpha_j - \boldsymbol{x}\boldsymbol{\beta})]$$

Example:

For 3 ordered outcomes, $y_i \in \{0, 1, 2\}$, we have $\beta_k > 0$. Then:

$\partial p_0(\boldsymbol{x})/\partial x_k < 0$,

$\partial p_2(\boldsymbol{x})/\partial x_k > 0$,

$\partial p_1(\boldsymbol{x})/\partial x_k$ can have either positive or negative sign.

In ordered Probit/Logit estimation, $\hat{\boldsymbol{\beta}}$ coefficients are usually of limited interest (have limited interpretation).

Instead, we are interested in response probabilities $P(y = j|\boldsymbol{x})$

As in other nonlinear models, we can compute and interpret PEAs or APEs. (Bootstrap standard errors)

Confusion matrix may be used for prediction (classification) accuracy evaluation.

Many generalizations and extension to ordered models are possible. For example, it makes sense to work with "cumulative" probabilities of "at least" or "at best" defined outcomes:

$$P(y \leq j|\boldsymbol{x}) = P(y^* \leq a_j|\boldsymbol{x}) = G(a_j - \boldsymbol{x}\boldsymbol{\beta}), \;\; j = 0, 1, \ldots, J - 1$$

Corner solution response data: Tobit model

The Tobit model for corner solution responses

- Corner solution response (many zero outcomes + roughly continuous positive outcomes).

- In many economic contexts, decision problems are such that either a positive amount or a zero amount is chosen (e.g. demand for alcohol)

- A linear regression model may be inadequate in such cases as predictions may be negative and effects of explanatory variables are linear

- The Tobit model makes use of a latent variable formulation
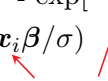
Definition of the Tobit model

Conditional on the values of the explanatory variables, the error term is homoscedastic normally distributed

$$y^* = \boldsymbol{x}\boldsymbol{\beta} + u, \ \ u|\boldsymbol{x} \sim N(0, \sigma^2)$$
$$y = \max\left(0, y^*\right)$$

The observed outcome of the dependent variable is positive or zero

Maximum likelihood estimation of the Tobit model

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma) = \begin{cases} (2\pi\sigma^2)^{-\frac{1}{2}} \exp[-(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2/(2\sigma^2)] & \text{if } y_i > 0 \\ 1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma) & \text{if } y_i = 0 \end{cases}$$

For positive outcomes, the normal density (applied to $y_i - \boldsymbol{x}_i\boldsymbol{\beta}$) is used, for zero outcomes the probability is one minus the probability that the latent variable is greater than zero (see Probit).

Maximization of the log-likelihood:

$$\max LL(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma) \rightarrow \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}$$

As in the Logit/Probit case, the maximization problem is highly nonlinear. It cannot be solved analytically and has to be solved with the help of computer software using e.g. Newton-Raphson methods.

Interpretation of the coefficients in the Tobit model

Conditional mean for all outcomes:

$$E(y|\boldsymbol{x}) = P(y > 0|\boldsymbol{x}) \cdot E(y|y > 0, \boldsymbol{x}) + P(y = 0|\boldsymbol{x}) \cdot 0 =$$
$$= \Phi(\boldsymbol{x\beta}/\sigma) \cdot E(y|y > 0, \boldsymbol{x})$$

The mean for all outcomes is a scaled version of the mean for only the positve outcomes (this is the reason why a regression using only the positive outcomes would yield wrong results)

Conditional mean for positive outcomes:

$$E(y|y > 0, \boldsymbol{x}) = \boldsymbol{x\beta} + \sigma\lambda(\boldsymbol{x\beta}/\sigma)$$

$$\lambda(c) = \phi(c)/\Phi(c) > 0$$

This is the so called
inverse Mills ratio

The mean for only the positive outcomes is the usual linear regression but plus an extra term (this is again a reason why an ordinary linear regression would yield wrong results)

Partial effects of interest in the Tobit model

On the probability of a nonzero outcome:

$\frac{\partial P(y>0|\boldsymbol{x})}{\partial x_j} = (\beta_j/\sigma)\phi(\boldsymbol{x}\boldsymbol{\beta}/\sigma)$ ← Note that all partial effects depend on the explanatory variables and the error variance

On the mean of positive outcomes:

This adjustment factor can be shown to lie between zero and one

$\frac{\partial E(y|y>0,\boldsymbol{x})}{\partial x_j} = \beta_j \overbrace{\{1 - \lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma)[\boldsymbol{x}\boldsymbol{\beta}/\sigma + \lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma)]\}}$

On the mean of all possible outcomes including zero:

$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j \Phi(\boldsymbol{x}\boldsymbol{\beta}/\sigma)$ ← Note that this adjustment factor also lies between zero and one

Estimation of average partial effects in the Tobit model

APE on the probability of a nonzero outcome:

$$\hat{\text{APE}}_{1,j} = n^{-1} \sum_{i=1}^{n} (\hat{\beta}_j/\hat{\sigma})\phi(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$$

Analogous formulas are available for partial effects at the average (PEA) but they have the aforementioned disadvantages

APE on the mean of positive outcomes:

$$\hat{\text{APE}}_{2,j} = n^{-1} \sum_{i=1}^{n} \hat{\beta}_j\{1 - \lambda(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})[\boldsymbol{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma} + \lambda(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})]\}$$

APE on the mean of all possible outcomes including zero:

$$\hat{\text{APE}}_{3,j} = n^{-1} \sum_{i=1}^{n} \hat{\beta}_j\Phi(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$$

Specification and other advanced topics in Tobit/Logit/Probit models

- In the Tobit model, regressors influence the values of observed positive outcomes and the probability of positive outcome observations at the same time.

- This may be unrealistic in many cases, for example, when modeling the relationship between the amount of life insurance and person's age.

- For such cases, more advanced so-called hurdle models can be used.

- As in Logit/Probit models, heteroscedasticity may arise in Tobit.

- ML estimates may be wrong if distributional assumptions do not hold.

- There are methods to deal with endogeneity in Logit/Probit/Tobit.

- Logit/Probit/Tobit models are available for panel/time series data.

Corner solution response data: Tobit model

Tobit example: Annual hours worked by married women

	Dependent Variable: *hours*	
Independent Variables	Linear (OLS)	Tobit (MLE)
nwifeinc	−3.45 (2.54)	−8.81 (4.46)
educ	28.76 (12.95)	80.65 (21.58)
exper	65.67 (9.96)	131.56 (17.28)
$exper^2$	−.700 (.325)	−1.86 (0.54)
age	−30.51 (4.36)	−54.41 (7.42)
kidslt6	−442.09 (58.85)	−894.02 (111.88)
kidsge6	−32.78 (23.18)	−16.22 (38.64)
constant	1,330.48 (270.78)	965.31 (446.44)
Log-likelihood value	-	-3,819.09
R-squared	.266	.274
$\hat{\sigma}$	750.18	1,122.02

Because of the different scaling factors involved, Tobit coefficients are not comparable to OLS coefficients.

To compare Tobit and OLS, one has to compare average partial effects (or partial effects at the average). It turns out that partial effects of Tobit and OLS are different in a number of cases.

Another difference between Tobit and OLS is that, due to the linearity of the model, OLS assumes constant partial effects, whereas partial effects are nonconstant in Tobit.

In the given example, OLS yields negative predictions of hours worked for 39 out of 753 women. This is not much but it may be a reason to view the linear model as misspecified.

Censored data/models (response variable is censored above and/or below some threshold; random sampling holds).

Often, dependent variable is censored in the sense that values are only reported up to a certain level (e.g. top coded wealth).

Censored normal regression model:

True outcome (unobserved)

Observed outcome $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i, \ u_i|\boldsymbol{x}_i, c_i \sim N(0, \sigma^2)$

$w_i = \min(y_i, c_i)$

If the true outcome exceeds the censoring threshold, only the threshold is reported

Regressing y_i on $x_i \Rightarrow$ correct results; but y_i is unobserved.

Regressing w_i on x_i will yield incorrect results (even if only the uncensored observations are used in this regression).

ML estimation of the censored regression model

$$P(w_i = c_i | \boldsymbol{x}_i) = P(y_i \geq c_i | \boldsymbol{x}_i) =$$
$$= P(u_i \geq c_i - \boldsymbol{x}_i\boldsymbol{\beta}) = 1 - \Phi[(c_i - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma]$$

Probability/density function of observed outcome conditional
on explanatory variables:

> If the censoring threshold does not bind,
> the density of the outcome is normal

$$f(w_i | \boldsymbol{x}_i, c_i; \boldsymbol{\beta}, \sigma) = \begin{cases} (1/\sigma)\phi[(w_i - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma] & \text{if } w_i < c_i \\ 1 - \Phi[(c_i - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma] & \text{if } w_i = c_i \end{cases}$$

Maximization of log-likelihood:

$$\max LL(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \log f(w_i | \boldsymbol{x}_i, c_i) \;\rightarrow\; \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}$$

Censored regression example: estimation of criminal recidivism

Dependent Variable: log(*durat*)	
Independent Variables	Coefficient (Standard Error)
workprg	−.063
	(.120)
priors	−.137
	(.021)
tserved	−.019
	(.003)
felon	.444
	(.145)
alcohol	−.635
	(.144)
drugs	−.298
	(.133)
black	−.543
	(.117)
married	.341
	(.140)
educ	.023
	(.025)
age	.0039
	(.0006)
constant	4.099
	(.348)
Log-likelihood	-1,597.06
value $\hat{\sigma}$	1.810

The variable *durat* measures the time in months until a prison inmate is arrested after being released from prison. Of 1,445 inmates, 893 had not been arrested during the time they were followed. Their time out of prison is censored (because a follow-up project ended, if there were some later arrests, they were not observed).

For example, if the time in prison was one month longer, this reduced the expected duration until the next arrest by about 1.9 %.

In the censored regression model, the coefficients can be directly interpreted. This is contrary to the Tobit model. The censored regression model and the Tobit model have similar structures, but in the Tobit model, the outcome is of a nonlinear nature whereas in the censored regression model, the outcome is linear with respect to β, but incompletely observed.

Truncated data

- Truncated data/regression models (for some observations, response variable is missing due to non-random selection)
- In a truncated regression model, the outcome and regressors are only observed if the outcome is less or equal some value c_i
- In this case, the sample is not a random sample from the population (systematically, some units will never be a part of the sample)
- Truncated normal regression model:

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i, \ u_i | \boldsymbol{x}_i \sim N(0, \sigma^2)$$
$$(y_i, \boldsymbol{x}_i) \ \underline{\text{only observed if}} \ y_i \leq c_i$$

- OLS would not yield correct results:
  MLR.2 (random sampling) is violated

ML estimation of the truncated regression model

Density of an observed outcome conditional on explanatory variables and the threshold c_i

$$g(y_i|\boldsymbol{x}_i, c_i) = \frac{f(y_i|\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)}{P(y_i \leq c_i|\boldsymbol{x}_i)} = \frac{f(y_i|\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)}{F(c_i|\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)}$$

Density and distribution function of a normal distribution with mean $\boldsymbol{x}_i\boldsymbol{\beta}$ and variance σ^2

Likelihood maximization:

$$\max LL(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \log g(y_i|\boldsymbol{x}_i, c_i) \rightarrow \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}$$

As in the censored regression model, nonnormality or heteroscedasticity in the truncated regression model lead to inconsistency.

Sample selection corrections

The question is under which assumptions a sample with nonrandom sample selection can be used to infer relationships in the population

When is OLS on the selected sample consistent?

$y = \boldsymbol{x}\boldsymbol{\beta} + u, \ E(u|\boldsymbol{x}) = 0$ ← Population model

s_i ← Sample selection indicator, $s_i = 1$ if observation selected/included in the sample, $s_i = 0$ otherwise

$s_i y_i = s_i \boldsymbol{x}_i \boldsymbol{\beta} + s_i u_i$ ← Regression based on the selected sample

Condition for consistency of OLS:

$E(su) = 0$

$E[(sx_j)(su)] = E(sx_j u) = 0$ (because $s^2 = s$ as s is binary)

Sample selection corrections

Unbiasedness: $E(u|x_1, x_2, \ldots, x_k) = 0$ vs. $E(su|sx_1, sx_2, \ldots, sx_k) = 0$

Conditions for OLS consistency on the selected sample:

1. s is independent of explanatory variables and the error term. $E(sx_j u) = E(s)E(x_j u) = 0$ because $E(x_j u) = 0$.

2. s is completely determined by explanatory variables. $sx_j = f(\boldsymbol{x})$. Hence, $\text{corr}(sx_j, u) = 0$ and $E(su|sx_1, sx_2, \ldots, sx_k) = 0$

3. s depends on the explanatory variables and other factors that are uncorrelated with the error term.

Similar conditions apply to IV/2SLS estimation

Besides explanatory variables $\boldsymbol{x}$, conditions extend to the full set of IVs: $\boldsymbol{z}$ as well.

Sample selection and nonlinear models estimated by ML

MLE (probit, logit) produces consisten etstimates if s is fully determined by regressors (cond. 2.).

Sample selection corrections (non-random selection)

Incidental truncation (Heckman model)

Special case of nonrandom selection: regressors are always observed, but the observation of the dependent variable is non-random (the same variables that determine the outcome of y_i also determine whether it is observed).

Example: Wage offer function using a sample of working women

- We are interested in the wage a woman with certain characteristics would be offered on the labor market if she decided to work.

- Unfortunately, we only observe the wages of women who actually work, i.e. who have accepted the wage offered to them. The sample is truncated because women who do not work (but who would be offered a wage if they asked for it) are never in the sample.

- Truncation of this kind is called incidental truncation because y_i observation depends on another variable (here: labor force participation).

Definition of Heckman model

$y = \boldsymbol{x\beta} + u$ ← Main equation (e.g. wage equation)

$s = 1[\boldsymbol{z\gamma} + v \geq 0]$ ← Selection equation (e.g. whether working)

$(u, v)|\boldsymbol{x}, \boldsymbol{z} \sim N(0, \sigma^2), \ \rho$ ← The error terms of both equations are jointly normally distributed (independent of the explanatory variables) with correlation coefficient ρ

Selection bias in OLS

$$E(y|\boldsymbol{x}, \boldsymbol{z}, s = 1) = \boldsymbol{x\beta} + \rho\lambda(\boldsymbol{z\gamma})$$

← For example, if the correlation of unobserved wage determinants and unobserved determinants of the work decision is positive, the women observed working will have higher wages than expected from their characteristics (= positive selection)

Running the regression on the truncated sample suffers from omitted variable bias

Estimation of Heckman model

1) Estimation of correction term:

$$P(s = 1 | \boldsymbol{z}) = \Phi(\boldsymbol{z\gamma})$$

Estimate Probit for work decision using all observations (working and nonworking women)

$$\hat{\lambda} = \phi(\boldsymbol{z}\hat{\boldsymbol{\gamma}}) / \Phi(\boldsymbol{z}\hat{\boldsymbol{\gamma}})$$

Calculate inverse Mills ratio using Probit coefficients (only $\hat{\lambda}_i$ for $s_i = 1$ are necessary)

2) Include estimated correction term in regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \rho \hat{\lambda}_i + error$$

If this coefficient is different from zero, there is nonignorable sample selection bias

There have to be explanatory variables in the selection equation that are not in the main equation (exclusion restrictions), otherwise there is multicollinearity because the inverse Mills ratio is almost linear in z

Sample selection corrections: Heckit model

Heckit example: Wage offer for married women

Dependent Variable: $\log(\textbf{wage})$		
Independent Variables	**OLS**	**Heckit**
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
exper^2	−.00081 (.00039)	−.00086 (.00044)
constant	−.522 (.199	−.578 (.307)
$\hat{\lambda}$	-	.032 (.134)
Sample size	428	428
R-squared	.157	.157

The standard errors of the two-step Heckman method are actually wrong and have to be corrected (not done here). One can also use a maximum likelihood procedure.

There is no significant sample selection bias. This is the reason why OLS and Heckman estimates are so similar.