Discrete response models

- Dependent variable qualitative, number of possible outcomes small

- Leading case binary response models

Example 1. *Model explaining the factors influencing the unemployment. Sample: people active on the labor market. Dependent variable* $y = 0$ *person unemployed,* $y = 1$ *person employed. Explanatory variables: socioeconomic characteristics of the persons. Traditionally:* 1 *success,* 0 *failure*

- In binary response models the interest really is in response probability:

$$p(\boldsymbol{x}) = \Pr(y = 1 | \boldsymbol{x}) = \Pr(y = 1 | x_1, \ldots, x_K)$$

- For binary (Bernoulli) random variables $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = p\left(\boldsymbol{x}\right)$ and $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = p\left(\boldsymbol{x}\right)\left[1 - p\left(\boldsymbol{x}\right)\right]$

- For continuous variable x_j the partial effect of x_j on the probability of success is
$$\frac{\partial\,\mathrm{E}\left(y|\,\boldsymbol{x}\right)}{\partial x_j} = \frac{\partial\,\mathrm{Pr}\left(y = 1|\,\boldsymbol{x}\right)}{\partial x_j} = \frac{\partial p\left(\boldsymbol{x}\right)}{\partial x_j}$$

- The partial effect describes the effect of the small change of x_j on probability of success

- For binary explanatory variable x_K the partial effect is calculated as

$$p\left(x_1, x_2, \ldots, x_{K-1}, 1\right) - p\left(x_1, x_2, \ldots, x_{K-1}, 0\right)$$

Linear probability model (LPM)

- We assume that

$$p\left(\boldsymbol{x}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K = \boldsymbol{x}\boldsymbol{\beta}$$

- Partial effects $\frac{\partial p(\boldsymbol{x})}{\partial x_j} = \beta_j$

- Easy to estimate by OLS

- This model cannot be a good description of probability of success, unless the range of $\boldsymbol{x}$ is severely restricted, $p\left(\boldsymbol{x}\right)$ cab take values outside unit interval!

- It can only be treated as an approximate model - however it is often a good approximate

- It should be noticed that in this model we must have heteroscedasticity

$$
\begin{aligned}
\mathrm{E}\left(y|\,\boldsymbol{x}\right) &= \boldsymbol{x\beta} \\
\mathrm{Var}\left(y|\,\boldsymbol{x}\right) &= \boldsymbol{x\beta}\left(1-\boldsymbol{x\beta}\right)
\end{aligned}
$$

- Then the heteroscedasticity robust variance matrix should be used for inference

- As the form of variance is known the weighted OLS can be used to obtain more efficient estimates.

  1. We estimate LPM with OLS
  2. Find the estimates of variances $\widehat{\sigma}_i^2 = \boldsymbol{x}_i\boldsymbol{b}\left(1-\boldsymbol{x}_i\boldsymbol{b}\right) = \widehat{y}_i\left(1-\widehat{y}_i\right)$

3. Make regression of $y_i/\widehat{\sigma}_i$ on $x_{1i}/\widehat{\sigma}_i, \ldots, x_{Ki}/\widehat{\sigma}_i$

- Some ad hoc adjustments are needed for $\widehat{\sigma}_i^2$ if $\widehat{\sigma}_i^2 \notin (0,1)$

- For LPM model partial effects are constant. This cannot be literally true as the big increase in x_j would drive the probability outside the unit interval

- Usually LPM model gives good estimates of the partial effects near the center of the distribution of $\boldsymbol{x}$. For extreme values of $\boldsymbol{x}$ the estimates are usually poor

Example 2. *(Wooldridge) Married women and labor force participation. Dependent variable: labor force participation (inlf) of married women. Explanatory variables: age, education, experience, nonwife income in thousands, number of children less than six years old (kidslt6), number of* *kids between 6 and 18 (kidsge6).*

$$\begin{array}{lcccccccc} inlf = & \underset{\substack{[.154]\\[.151]}}{.586} & - & \underset{\substack{[.0014]\\[.0015]}}{.0034} & \textit{nwifeinc} + & \underset{\substack{[.007]\\[.007]}}{.038} & \textit{educ} + & \underset{\substack{[.006]\\[.006]}}{.039} & \textit{exper} \\ + & \underset{\substack{[.00018]\\[.00019]}}{.00060} & \textit{exper}^2 - & \underset{\substack{[.002]\\[.002]}}{0.016} & \textit{age} - & \underset{\substack{[.034]\\[.032]}}{.262} & \textit{kidslt6} + & \underset{\substack{[.013]\\[.013]}}{.013} & \textit{kidsge6} \end{array}$$

The second standard errors are robust standard errors. Not really different from OLS standard errors in this case.

Example 3. *Dependence of the labor activity on age and education (educa). Data: polish Labour Force Survey (BAEL). Question: had you done any work in last week?*

```
      Source |       SS       df       MS              Number of obs =   47860
-------------+------------------------------           F(  9, 47850) =  993.61
       Model |  1827.14253     9  203.015837           Prob > F      =  0.0000
    Residual |  9776.79083 47850  .204321648           R-squared     =  0.1575
-------------+------------------------------           Adj R-squared =  0.1573
       Total |  11603.9334 47859  .242460841           Root MSE      =  .45202

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0036288   .0001188   -30.55   0.000    -.0038616   -.0033961
   _Ieduca_2 |  -.1155233   .0141975    -8.14   0.000    -.1433507    -.087696
   _Ieduca_3 |  -.1575525   .0082029   -19.21   0.000    -.1736303   -.1414747
   _Ieduca_4 |  -.3616109   .0098378   -36.76   0.000     -.380893   -.3423287
   _Ieduca_5 |   -.196475   .0078076   -25.16   0.000    -.2117779    -.181172
  _Ieduca_60 |  -.7496461   .0163531   -45.84   0.000    -.7816984   -.7175937
  _Ieduca_61 |  -.4814629   .0077523   -62.11   0.000    -.4966576   -.4662683
  _Ieduca_70 |  -.5166902   .0162939   -31.71   0.000    -.5486266   -.4847539
  _Ieduca_71 |   -.594044   .0302346   -19.65   0.000    -.6533043   -.5347837
       _cons |   .8569807   .0084361   101.59   0.000     .8404459    .8735155
------------------------------------------------------------------------------
```

- Results of heteroscedasticity test

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of _Iworks_1

         chi2(1)      =   961.62
         Prob > chi2  =   0.0000
```

- Some of the predicted values:

```
       +-----------+
       |      yhat |
       |-----------|
    1. | -.1974881 |
    2. | -.1793439 |
    3. | -.1720862 |
    4. | -.1466843 |
    5. | -.1398648 |
```

- But from 47860 observation only $144 \notin (0,1)$

Index models: Probit and logit

- We model the probability of the choices
- Assume that
$$\Pr\left(y=1|\,\boldsymbol{x}\right)=G\left(\boldsymbol{x\beta}\right)=p\left(\boldsymbol{x}\right)$$
- This model is called index model because it restricts the probability $p\left(\boldsymbol{x}\right)$ to depend only on index $\boldsymbol{x\beta}$
- In most applications G is a cumulative distribution fuction (cdf)
- Index model where G is a cdf can be derived form an underling latent variable model
$$y^{*}=x\beta+e,\qquad y=1\left[y^{*}>0\right]$$

where $1\left[y^{*}>0\right]=\left\{\begin{array}{ccc}1 & \text{if} & y^{*}>0 \\ 0 & \text{if} & y \leq 0\end{array}\right|$ is an indicator function.

- If we assume that e has a symmetric distribution than

$$\Pr\left(y=1|\boldsymbol{x}\right)=\Pr\left(y^{*}>0|\boldsymbol{x}\right)=\Pr\left(e>-\boldsymbol{x\beta}|\boldsymbol{x}\right)=1-G\left(-\boldsymbol{x\beta}\right)=G\left(\boldsymbol{x\beta}\right)$$

- For probit model we assume that e has standard normal distribution. Than $G\left(\boldsymbol{x\beta}\right)=\Phi\left(\boldsymbol{x\beta}\right)$ where $\Phi\left(\cdot\right)$ is cdf of standard normal distribution

- For logit model we assume that e has logistic distribution. Than $G\left(\boldsymbol{x\beta}\right)=\Lambda\left(\boldsymbol{x\beta}\right)$ where $\Lambda\left(\boldsymbol{x\beta}\right)=\frac{\exp(\boldsymbol{x\beta})}{1+\exp(\boldsymbol{x\beta})}$

- The partial effects of x_j on probability of success is given by

$$\frac{\partial p\left(\boldsymbol{x}\right)}{\partial x_j}=g\left(\boldsymbol{x\beta}\right)\beta_j$$

where $g\left(\boldsymbol{x\beta}\right)=\frac{\partial G(z)}{\partial z}$.

- Then the value of β_j has no direct interpretation but the sign has. The positive sign means that the influence of x_j on probability of success is positive as $g\left(\boldsymbol{x\beta}\right)>0$ (density function is always positive)

- Similarly, values of β_j and β_h can be interpreted as they are equal to relative partial effects
$$\frac{\partial p\left(\boldsymbol{x}\right)}{\partial x_j}\Bigg/\frac{\partial p\left(\boldsymbol{x}\right)}{\partial x_h}=\frac{\beta_j}{\beta_h}$$

- The partial effects of binary variable x_K are calculated as
$$G\left(\beta_1+\beta_2x_2+\ldots+\beta x_{K-1}+\beta_k\right)-G\left(\beta_1+\beta_2x_2+\ldots+\beta x_{K-1}\right)$$
it is interpreted as a change of response probability resulting from the change of x_k from 0 to one

- The density an observation i for the binary response index model is given by

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

- The log likelihood for an observation i is then given by

$$\ell_i\left(\boldsymbol{\beta}\right) = y_i \ln\left[G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)^{y_i}\right] + \left(1-y_i\right)\ln\left[1-G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]$$

- The loglikelihood function $\ell\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \ell_i\left(\boldsymbol{\beta}\right)$ is maximized numerically. In this way we obtain Maximum Likelihood Estimators (MLE).

- Testing the hypotestis in probit and logit models can be easily done with LR or $Wald$ statistics. Sometimes the LM statistics is also useful.

Reporting results for probit and logit

- The various measures of the goodness of fit were proposed
- One of them is percent of correctly predicted:

		Observed		
		0	1	Total
Predicted	0	n_{00}	n_{01}	$n_{00} + n_{01}$
	1	n_{10}	n_{11}	$n_{10} + n_{11}$
	Total	$n_{00} + n_{10}$	$n_{01} + n_{11}$	

- We assume that the model predicts $y_i = 1$ if $p\left(\boldsymbol{x}\widehat{\boldsymbol{\beta}}\right) > 0.5$

- Another measure of fit which is often used is $pseudo$-R defined as

$$pseudo\text{-}R^2 = 1 - \frac{\ell_{ur}}{\ell_o}$$

where ℓ_{ur} is the loglikelihood at maximum of an unrestricted model and ℓ_o is loglikelihood in the model with only intercept.

- $pseudo$-R^2 satisfy $0 \leq pseudo\text{-}R^2 \leq 1$

- The problem with interpreting the partial effects in probit and logit is related to dependance of the partial effects on $\boldsymbol{x}$

- The estimated partial effect of the change Δx_j on probability is

$$\Delta \Pr\left(y = 1 | \boldsymbol{x}\right) \approx \left[g\left(\boldsymbol{x}\boldsymbol{\beta}\right)\beta_j\right]\Delta x_j$$

- Therefore must calculate the partial effects for some $\boldsymbol{x}$. Usually we choose mean of the $\boldsymbol{x}$ in the sample $\overline{\boldsymbol{x}}$.

- If $\overline{\boldsymbol{x}}\boldsymbol{\beta}$ is close to 0 than $g\left(0\right) \approx .4$ for probit, $g\left(0\right) \approx .25$ for logit, and $g\left(0\right) = 1$ for LPM. Therefore it is approximately often the case that logit estimates are $\frac{.4}{.25} = 1.6$ larger than probit estimates and 4 times larger than LPM estimates.

- For binary variable x_j the $\overline{x}_j$ is equal to fraction of 1 in the sample. Than the result does not correspond to any individual but it can be interpreted as the result for ”mean” individual.

Example 4. *Dependence of the labor activity on age and education*

- Coefficients:

```
Probit estimates                                  Number of obs   =      47860
                                                  LR chi2(9)      =    8274.24
                                                  Prob > chi2     =     0.0000
Log likelihood = -28311.095                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0113366   .0003666   -30.93   0.000     -.012055   -.0106181
   _Ieduca_2 |  -.3254287   .0404874    -8.04   0.000    -.4047825   -.2460749
   _Ieduca_3 |   -.433603    .023857   -18.18   0.000     -.480362   -.3868441
   _Ieduca_4 |  -.9675114   .0286116   -33.82   0.000    -1.023589   -.9114338
   _Ieduca_5 |  -.5330219   .0227775   -23.40   0.000    -.5776649   -.4883789
  _Ieduca_60 |  -2.513776   .0730264   -34.42   0.000    -2.656905   -2.370647
  _Ieduca_61 |  -1.343109   .0233078   -57.62   0.000    -1.388792   -1.297427
  _Ieduca_70 |  -1.678283   .0650819   -25.79   0.000    -1.805842   -1.550725
  _Ieduca_71 |  -2.321871   .1804895   -12.86   0.000    -2.675624   -1.968118
       _cons |   1.027194   .0256271    40.08   0.000     .9769656    1.077422
------------------------------------------------------------------------------
```

- Partial effects:

```
Probit estimates                                        Number of obs =  47860
                                                        LR chi2(9)    =8274.24
                                                        Prob > chi2   = 0.0000
Log likelihood = -28311.095                             Pseudo R2     = 0.1275

------------------------------------------------------------------------------
_Iwork~1 |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
    wiek |  -.0043578   .0001406   -30.93   0.000   43.9481  -.004633 -.004082
_Ieduc~2*|  -.1181235   .0136421    -8.04   0.000   .027267  -.144862 -.091385
_Ieduc~3*|  -.1582424   .0081461   -18.18   0.000    .19135  -.174209 -.142276
_Ieduc~4*|  -.3006814   .0063921   -33.82   0.000   .082804   -.31321 -.288153
_Ieduc~5*|  -.1946251   .0077792   -23.40   0.000    .26728  -.209872 -.179378
_Iedu~60*|  -.4090495   .0025428   -34.42   0.000   .020079  -.414033 -.404066
_Iedu~61*|  -.4366777   .0058916   -57.62   0.000   .290639  -.448225  -.42513
_Iedu~70*|  -.3783463   .0048507   -25.79   0.000   .020664  -.387854 -.368839
_Iedu~71*|  -.3921597   .0035269   -12.86   0.000   .004952  -.399072 -.385247
---------+--------------------------------------------------------------------
  obs. P |   .4131425
 pred. P |   .3926327  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| are the test of the underlying coefficient being 0
```

Example 5. *Dependence of labor activity market on age and education (cont.)*

- Logit model - coeffcients

```
Logit estimates                                   Number of obs   =      47860
                                                  LR chi2(9)      =    8275.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -28310.607                       Pseudo R2       =     0.1275

------------------------------------------------------------------------------
   _Iworks_1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |  -.0186039   .0006038   -30.81   0.000    -.0197873   -.0174204
   _Ieduca_2 |  -.5221753   .0655267    -7.97   0.000    -.6506053   -.3937453
   _Ieduca_3 |  -.6942371   .0389682   -17.82   0.000    -.7706134   -.6178608
   _Ieduca_4 |  -1.561855   .0471004   -33.16   0.000     -1.65417    -1.46954
   _Ieduca_5 |  -.8542947   .0372593   -22.93   0.000    -.9273216   -.7812678
  _Ieduca_60 |  -4.372065   .1570768   -27.83   0.000     -4.67993     -4.0642
  _Ieduca_61 |  -2.207483   .0391494   -56.39   0.000    -2.284215   -2.130752
  _Ieduca_70 |   -2.88925   .1303508   -22.17   0.000    -3.144732   -2.633767
  _Ieduca_71 |  -4.194055    .416003   -10.08   0.000    -5.009406   -3.378704
       _cons |   1.666699   .0423531    39.35   0.000     1.583688    1.749709
------------------------------------------------------------------------------
```

- Partial effects

```
Marginal effects after logistic
      y  = Pr(_Iworks_1) (predict)
         =  .38568338
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    wiek |  -.0044078      .00014  -30.94   0.000  -.004687 -.004129   43.9481
_Ieduc~2*|  -.1148651      .06553   -1.75   0.080  -.243295  .013565   .027267
_Ieduc~3*|  -.1539078      .03897   -3.95   0.000  -.230284 -.077532    .19135
_Ieduc~4*|  -.2864085       .0471   -6.08   0.000  -.378724 -.194093   .082804
_Ieduc~5*|   -.189642      .03726   -5.09   0.000  -.262669 -.116615    .26728
_Iedu~60*|   -.398101      .15708   -2.53   0.011  -.705966 -.090236   .020079
_Iedu~61*|  -.4279628      .03915  -10.93   0.000  -.504694 -.351231   .290639
_Iedu~70*|  -.3641805      .13035   -2.79   0.005  -.619663 -.108698   .020664
_Iedu~71*|  -.3810388        .416   -0.92   0.360  -1.19639  .434312   .004952
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

- Percent correctly predicted:

```
  Logistic model for _Iworks_1

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |     12568          7211  |      19779
     -     |      7205         20876  |      28081
-----------+--------------------------+-----------
   Total   |     19773         28087  |      47860

Classified + if predicted Pr(D) >= .5
True D defined as _Iworks_1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   63.56%
Specificity                     Pr( -|~D)   74.33%
Positive predictive value       Pr( D| +)   63.54%
Negative predictive value       Pr(~D| -)   74.34%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   25.67%
False - rate for true D         Pr( -| D)   36.44%
False + rate for classified +   Pr(~D| +)   36.46%
False - rate for classified -   Pr( D| -)   25.66%
--------------------------------------------------
Correctly classified                        69.88%
--------------------------------------------------
```

Specification issues in binary response models

- Neglected heterogeneity. Assume that in probit

  $$\Pr\left(y=1|\,\boldsymbol{x},c\right)=\Phi\left(\boldsymbol{x\beta}+\gamma c\right)$$

  where c is not observable and $e|\,\boldsymbol{x},c\sim N\left(0,1\right)$, $c\sim N\left(0,\tau^2\right)$ and is independent on $\boldsymbol{x}$.

- The variance of the composite error term $e+\gamma c$ is then equal to $\sigma^2=1+\gamma^2\tau^2$ and

  $$\Pr\left(y=1|\,\boldsymbol{x}\right)=\Pr\left(e+\gamma c>-\boldsymbol{x\beta}\right)=\Phi\left(\boldsymbol{x\beta}/\,\sigma\right)$$

- It implies that the omitted variable (neglected heterogeneity) will influence the estimates of $\boldsymbol{\beta}$ even if omitted variable is not correlated with variables in the regression (attenuation bias)

- However, if we are not interested in $\boldsymbol{\beta}$ calculated for $c = 0$, but in average partial effect calculated for an $\boldsymbol{x}^o$ than as

  $$\mathrm{E}\left(\frac{\partial p\left(\boldsymbol{x}^o,c\right)}{\partial x_j}\right) = \mathrm{E}\left[\phi\left(\boldsymbol{x}^o\,\boldsymbol{\beta}+\gamma c\right)\beta_j\right] = \left(\beta_j/\,\sigma\right)\phi\left(\boldsymbol{x}^o\boldsymbol{\beta}/\,\sigma\right)$$

  the estimates from probit will be consistent.

- The omitted heterogeneity in probit is not a problem as long as it is not correlated with x and we are only interested in average partial effects

- The heteroskedasticity and nonnormality problem is much more serious in probit that in linear model.

- If e is heteroscedastic or nonnormal than $\Pr\left(y=1|\,\boldsymbol{x}\right) \neq \Phi\left(\boldsymbol{x\beta}\right)$ and functional for of the model is misspecified

- In practice the distribution functions used are often so similar that the estimates does not differ much

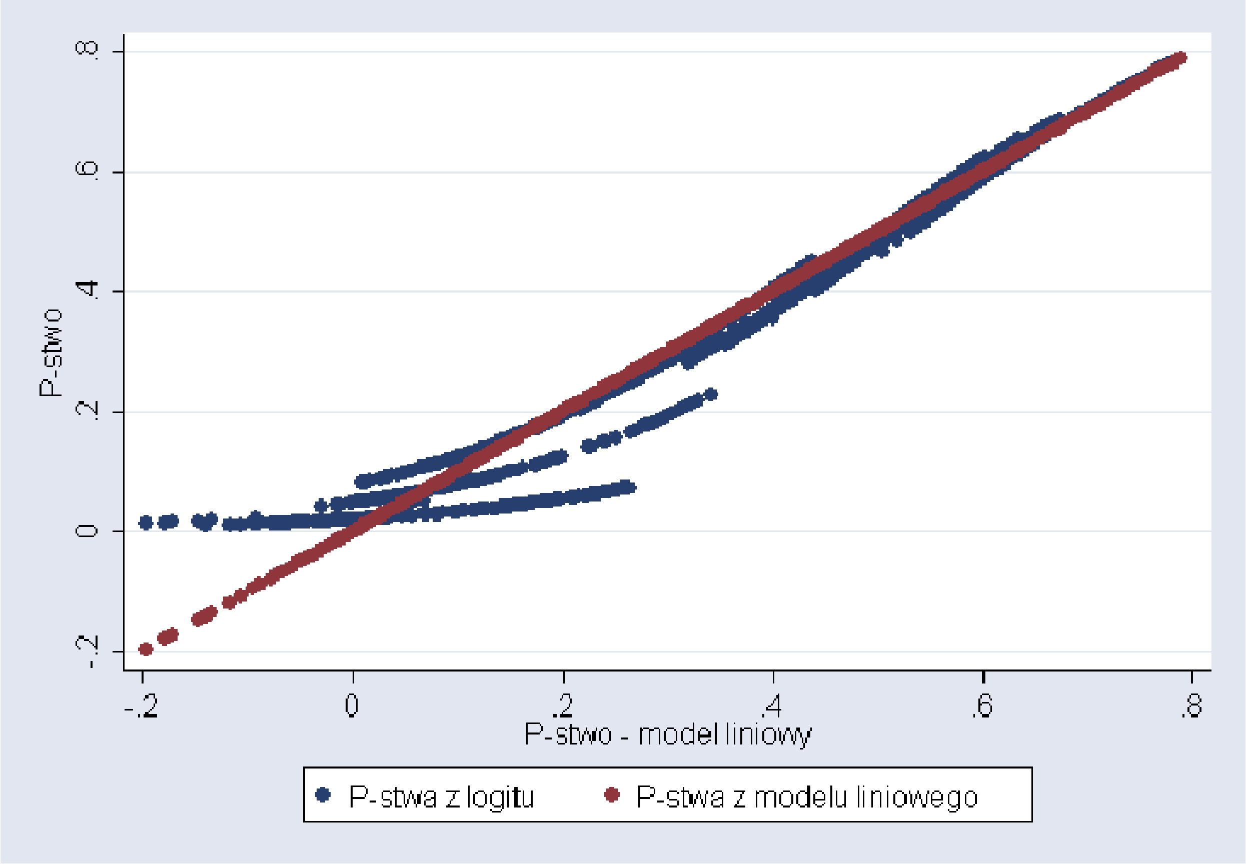

Figure 1: Predicted probabilities of outcomes from linear and logit models

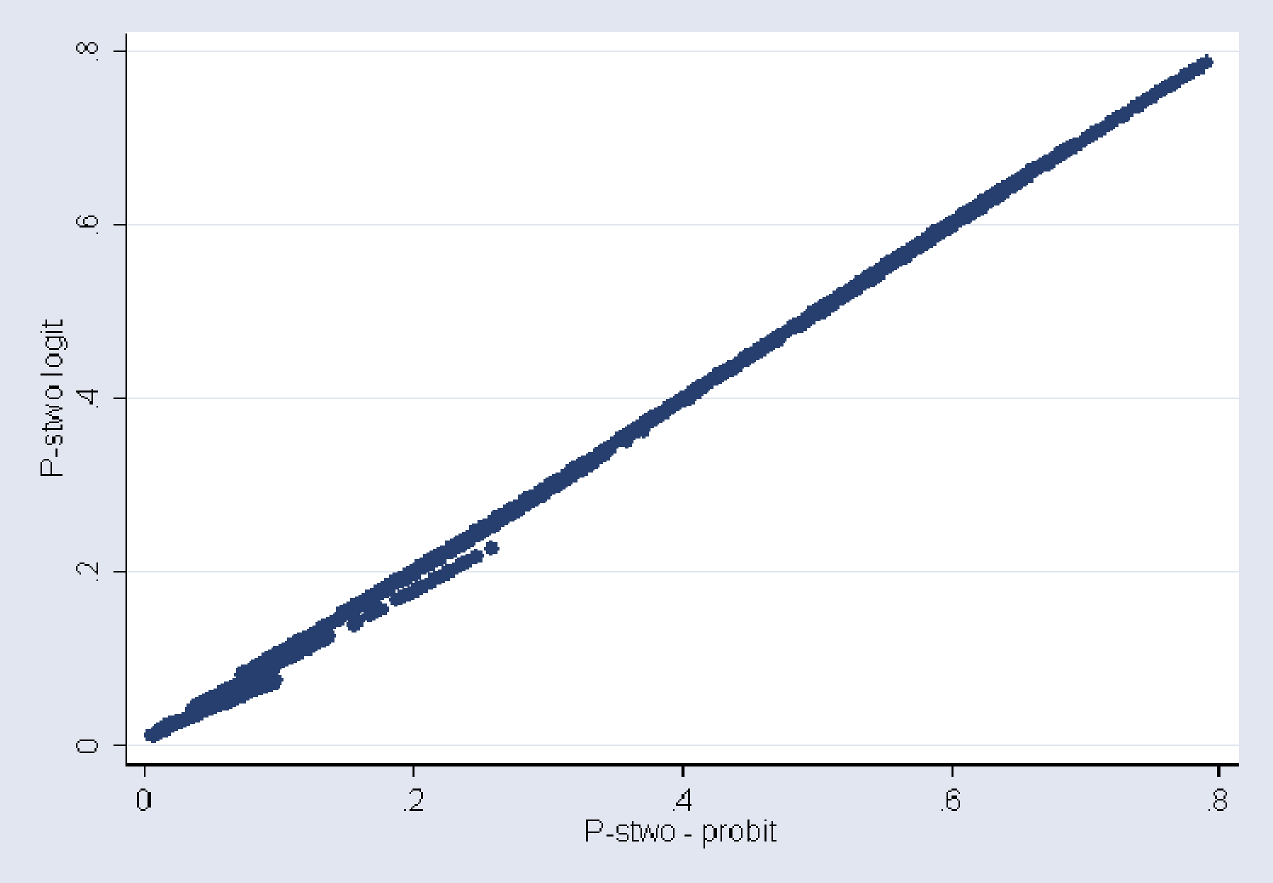

Figure 2: Predicted probabilities of outcomes from probit and logit models

- Panel binary response model

$$y_{it}^{*} = x_{it}\beta + c_i + e_{it}, \quad i = 1, \ldots, n \quad t = 1, \ldots, T$$
$$y_{it} = 1\left[y_{it}^{*} > 0\right]$$

- Consistent estimates for the panel probit and logit models can be found by running pooled probit or logit.
  - In order to obtain the correct estimates of variance matrix we should define individuals as clusters

- It is also possible to assume some distribution of c_i, derive $f\left(y_i | \boldsymbol{x}_i\right) = \mathrm{E}\left[f\left(y_i | \boldsymbol{x}_i, c_i\right)\right]$ and find the maximum likelihood estimator (random effect probit)

- As in linear model pooled probit or logit, and random effect model is only consistent if c_i are independent on $\boldsymbol{x}_i$

- The fixed effect panel estimation done by estimating the individual effects as parameters is not correct for probit and logit models.
  - Incidental parameters problem - as the estimates of the $\boldsymbol{\beta}$ are related to estimates of c_i and estimates of c_i are not consistent if T finite, the estimates of $\boldsymbol{\beta}$ are also not consistent

- However it is possible to find the model analogous to fixed effect for logit model (fixed effect logit) by conditioning on $n_i = \sum_{t=1}^{T} y_{it}$. Only observations used in this case are the observations for individuals which were observed to change state y_i (only this values can tell as something about parameters $\boldsymbol{\beta}$)

Multinomial response models

- In some problems we model the qualitative variable with outcomes $\{0, 1, \ldots, J\}$

- These model is usually used if responses have no logical order (e.g. decision to work in state sector, private sector or be self-employed)

- The most popular model used in this context is the multinomial logit model

$$p_j\left(\boldsymbol{x}\right) = \Pr\left(y = j \middle| \boldsymbol{x}\right) = \frac{\exp\left(\boldsymbol{x}\boldsymbol{\beta}_j\right)}{1 + \sum_{h=1}^{J} \exp\left(\boldsymbol{x}\boldsymbol{\beta}_h\right)} \text{ for } j = 1, \ldots, J$$

$$p_0\left(\boldsymbol{x}\right) = \Pr\left(y = 0 \middle| \boldsymbol{x}\right) = \frac{1}{1 + \sum_{h=1}^{J} \exp\left(\boldsymbol{x}\boldsymbol{\beta}_h\right)} \text{ for } j = 0$$

- The simplest interpretation of the coefficients is related to odds (probability of outcome j relative to outcome 0)

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

  and odds ratios (ratio of odds if x changes)

$$OR_j = \frac{Odds_j\left(\boldsymbol{x} + \boldsymbol{\Delta x}\right)}{Odds_j\left(\boldsymbol{x}\right)} = \exp\left(\boldsymbol{\Delta x}\boldsymbol{\beta}_j\right)$$

- So the positive $\boldsymbol{\beta}_{j,k}$ means that if $x_k$the odds ratio for alternative j also rises.

- We can also calculate partial effects of for alternative j instead of odds ratios

Example 6. *(Wooldridge). We model following choices of men: being enrolled in school, not in school and not working, working. The base category is to be enrolled in school.*

Explanatory variable	home	work
	(status=1)	(status=2)
educ	-.674	-.315
exper	-.106	.849
	(.070)	(.065)
exper^2	-.013	-0.077
	(.173)	(.157)
black	.813	.311
	(.303)	(.282)
constant	10.28	5.54
	(1.13)	(1.09)
number of observations	1.717	
Pecent correctly predicted	79.6	
Log-likelihood value	-907.86	
Pseudo R-squared	.243	

Interpretation: one more year in school reduces the log-odds between at home and enrolled in school by $-.674$ and log-odds of being at work versus being at school by $-.315$. Log odds between at home and enrolled in school is $.813$ higher for black man. Signs of these coefficients can be interpreted but not the magnitudes. We can either calculate the partial effects or compare the differences in probabilities: black man with 16 of education has an employment probability that higher by $.042$ than a man with 12 years of education and at home probability lower by $.072$.

Probabilistic choice models

- Some times it is possible to construct the model for which the choice is based on underling utility

- Assume that the utility from choosing alternative j for individual i is given by

$$y_{ij}^{*} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + a_{ij}$$

  where a_{it} are unobservables effecting tests.

- Important: the $\boldsymbol{x}_{ij}$ are different between individuals but also between alternatives

- Individual is maximizing his utility than it is choosing the alternative y_i such that

$$y_i = \arg\max\left(y_{i0}^*, y_{i1}^*, \ldots, y_{iJ}^*\right)$$

- It is possible to prove that if a_{ij} for $j = 0, \ldots, J$ are independently distributed with cdf given by $F\left(a_{ij}\right) = \exp\left[-\exp\left(-a_{ij}\right)\right]$ (type I extreme value distribution) than

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

- This model is called conditional logit

- So for all the individuals we observe choices, characteristics of alternatives and characteristics of individuals

Example 7. *We are modelling the choices of the modes of transportation to work (car, bus). The possible characteristics of the modes of transportations is cost (ticket, patrol+parking) and time of travel. The characteristics of individuals is are income. The data set will have the following form*

individual	*choice*	*decision*	*cost*	*time*	*income*
1	*bus*	*1*	*2*	*25*	*2000*
1	*car*	*0*	*3*	*15*	*2000*
2	*bus*	*0*	*2*	*25*	*3000*
2	*car*	*1*	*3*	*15*	*3000*
3	*bus*	*1*	*2*	*25*	*1000*
3	*car*	*0*	*3*	*15*	*1000*

- For individual i the ratio of probabilities for alternative j and h is equal to

$$\frac{p_{ij}\left(\boldsymbol{x}_{ij}\right)}{p_{ih}\left(\boldsymbol{x}_{ih}\right)} = \frac{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{x}_{ih}\boldsymbol{\beta}\right)} = \exp\left[\left(\boldsymbol{x}_{ij}-\boldsymbol{x}_{ih}\right)\boldsymbol{\beta}\right]$$

The relative probability for any two alternative depends only on attributes of these alternatives

- Independence from irrelevant alternatives (IIA) assumption: adding an third alternative or changing its attributes does not affect the odd ratio. This assumption is often not realistic for similar alternatives.

Example 8. *(McFadden 1974) Assume that commuters are choosing between car and red bus. Suppose that the probability of choice these two alternatives are equal so that* $p_{car} = p_{red\,bus} = \frac{1}{2}$ *and odds ratio is equal* $\frac{p_{car}}{p_{red\,bus}} = 1$*. Now third alternative is considered: blue bus. If passengers are indifferent between the blue an red buses than* $p_{red\,bus} = p_{blue\,bus}$*. As the odds ratio* $\frac{p_{car}}{p_{red\,bus}} = 1$ *then probability of choosing car has to fall to* $\frac{1}{3}$ *as a result of introducing red buses. This effect seems not very realistic!*

- Possible solutions for this problem:

1. Mulinomial probit. The IIA is the result of the special form of the distribution used for multinomial logit. If we use the multivariate normal distribution for a_{ij} the IIA will not hold. Problem: to estimate multivariate probit model by ML it is necessary to use the multivariate normal cdf which is difficult approximate.
2. Hierarchical model - nested logit. The similar alternatives are grouped. The consumer choice is assumed to be done in stages: first we choose the group (say s) and than given the alternative within the group (say j). We specify the probability of the choice of the group $\Pr\left(y \in G_s | \boldsymbol{x}\right)$ and the choice of the alternative $j \in G_s$ given that G_s was chosen $\Pr\left(y = j | \boldsymbol{x}, G_s\right)$. Unconditional probability of choice j is equal to $\Pr\left(y = j | \boldsymbol{x}\right) = \Pr\left(y \in G_s | \boldsymbol{x}\right) \Pr\left(y = j | \boldsymbol{x}, G_s\right)$. With this probability we define the likelihood function. For nested logit the probabilities has the

form

$$\Pr\left(y\in G_s\middle|\,\boldsymbol{x}\right)=\frac{\alpha_s\left[\sum_{j\in G_s}\exp\left(\rho_s^{-1}\boldsymbol{x}_j\boldsymbol{\beta}\right)\right]^{\rho_s}}{\sum_{r=1}^{S}\left\{\alpha_r\left[\sum_{j\in G_s}\exp\left(\rho_r^{-1}\boldsymbol{x}_j\boldsymbol{\beta}\right)\right]^{\rho_r}\right\}}$$

$$\Pr\left(y=j\middle|\,\boldsymbol{x},G_s\right)=\frac{\exp\left(\rho_s^{-1}\boldsymbol{x}_j\boldsymbol{\beta}\right)}{\sum_{h\in G_s}\exp\left(\rho_s^{-1}\boldsymbol{x}_h\boldsymbol{\beta}\right)}$$

Ordered choice models

- The responses $\{1, 2, \ldots, J\}$ we analyze are ordered. This means that the values of the discrete dependent variable have same logical order (e.g. credit ratings, exam marks, education degrees).

- Ordered probit

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

We have J *unknown* cut points $\alpha_1 < \alpha_2 < \ldots < \alpha_j$

$$\begin{array}{lll} y = 0 & \text{if} & y^* \leq \alpha_1 \\ y = 1 & \text{if} & \alpha_1 < y^* \leq \alpha_2 \\ \quad\vdots & & \quad\vdots \\ y = J & \text{if} & y^* > \alpha_J \end{array}$$

The probabilities for choices are given by

$$
\begin{aligned}
\Pr\left(y=0\middle|\,\boldsymbol{x}\right) &= \Pr\left(y^{*}\leq\alpha_{1}\middle|\,\boldsymbol{x}\right)=\Phi\left(\alpha_{1}-\boldsymbol{x\beta}\right)\\
\Pr\left(y=1\middle|\,\boldsymbol{x}\right) &= \Pr\left(\alpha_{1}<y^{*}\leq\alpha_{2}\middle|\,\boldsymbol{x}\right)=\Phi\left(\alpha_{2}-\boldsymbol{x\beta}\right)-\Phi\left(\alpha_{1}-\boldsymbol{x\beta}\right)\\
&\vdots\\
\Pr\left(y=J\middle|\,\boldsymbol{x}\right) &= \Pr\left(y^{*}>\alpha_{J}\middle|\,\boldsymbol{x}\right)=1-\Phi\left(\alpha_{J}-\boldsymbol{x\beta}\right)
\end{aligned}
$$

- These probabilities can be used to define the likelihood function.

- Ordered logit: we obtain this model if instead of taking normal distribution $cdf\ \Phi\left(\cdot\right)$ we use logistic distribution $\Lambda\left(\cdot\right)$

- The partial effects for ordered probit are give by

$$\frac{\partial p_0(\boldsymbol{x})}{\partial \boldsymbol{x}} = -\boldsymbol{\beta}_k \phi(\alpha_1 - \boldsymbol{x}\boldsymbol{\beta}), \ \frac{\partial p_J(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{\beta}_k \phi(\alpha_J - \boldsymbol{x}\boldsymbol{\beta})$$
$$\frac{\partial p_j(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{\beta}_k [\phi(\alpha_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - \phi(\alpha_j - \boldsymbol{x}\boldsymbol{\beta})], \quad 1 < j < J$$

  The signs of these effects is only determined by signs of β for $j = 1$ and $j = J$. For intermediate choices it depends also on the sign of $\phi(\alpha_{j-1} - \boldsymbol{x}\boldsymbol{\beta}) - \phi(\alpha_j - \boldsymbol{x}\boldsymbol{\beta})$

- For these model we can calculated percent correctly predicted (predicted choice is the most probable choice according to model)

- Interval coded data: we do not know the exact values of the dependent variable but we know the interval in which it is located.

- This model has the exactly the same structure as the ordered choice model but the cut points $\alpha_1 < \alpha_2 < \ldots < \alpha_j$ are *known*. In this case $\boldsymbol{\beta}$ has the same interpretation as in classical regression model.

- Similar model for which $e|\,\boldsymbol{x}$ has logistic distribution is known as ordered logit model or Proportional Odds Model (POM)

- Probability of the cases in ordered logit model are

$$
\begin{aligned}
\Pr\left(y = 0|\,\boldsymbol{x}\right) &= \Lambda\left(\alpha_1 - \boldsymbol{x\beta}\right) \\
\Pr\left(y = 1|\,\boldsymbol{x}\right) &= \Lambda\left(\alpha_2 - \boldsymbol{x\beta}\right) - \Lambda\left(\alpha_1 - \boldsymbol{x\beta}\right) \\
&\vdots \\
\Pr\left(y = J|\,\boldsymbol{x}\right) &= 1 - \Lambda\left(\alpha_J - \boldsymbol{x\beta}\right)
\end{aligned}
$$

- But this implies that the odds of observing y larger or equal to k is (POM assumption):

$$Odds\left(\boldsymbol{x}\right)=\frac{\Pr\left(y\geq k|\,\boldsymbol{x}\right)}{\Pr\left(y<k|\,\boldsymbol{x}\right)}=\frac{\sum_{i=k}^{J}\Pr\left(y=i|\,\boldsymbol{x}\right)}{\sum_{i=0}^{k-1}\Pr\left(y=i|\,\boldsymbol{x}\right)}=\frac{1-\Lambda\left(\alpha_k-\boldsymbol{x\beta}\right)}{\Lambda\left(\alpha_k-\boldsymbol{x\beta}\right)}=\exp\left(\boldsymbol{x\beta}\right)$$

and is the same for every k

- Therefore also in the case of the ordered logit model we can use the intepretation of parameters based on odds ratios

$$\frac{Odds\left(\boldsymbol{x}\right)}{Odds\left(\boldsymbol{x}+\Delta\boldsymbol{x}\right)}=\exp\left(\Delta\boldsymbol{x\beta}\right)$$

- If $\Delta\boldsymbol{x}$ has all elements equal to 0 except $\Delta x_j=1$ (x_j changes by 1, all ather variables stay constant) this odds ratio is equal to $\exp\left(\beta_j\right)$ and can be intepreted as the increase of the relative probability of observing higher categories.

- POM assumption can be tested with suitable diagnostic test.

Count data

- The *count variable* is the variable which takes the nonnegative integer values (e.g. number of children, number of cigarettes smoked a day, number of strikes for a given year for a firm)

- In some cases the count variable has no logical upper bound but sometimes it has (e.g. number of children in a family being high school graduates is smaller or equal to the total number of children)

- The simplest method of estimation is to assume that $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \boldsymbol{x\beta}$ and estimate the model using OLS.

- But: this method has a shortcoming that the some of the predicted values can $\widehat{y} = \boldsymbol{xb}$ can prove to be negative, which make no sense as count variable is necessarily positive

- Another choice it to use the log transformation and estimate assume that $\log \mathrm{E}\left[y|\,\boldsymbol{x}\right] = \boldsymbol{x\beta}$. Then predicted $\widehat{y} = \exp\left(\boldsymbol{xb}\right) > 0$. This approach is however not applicable for cases when for nontrivial fraction of data $y = 0$ (e.g. number of children)

- Then the best choice is often to directly model the probabilities of discrete choices

Poisson model

- The Poisson model is the most popular model for count data as it simple and has same desirable properties

- We assume that the conditional mean y is given by $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \mu\left(\boldsymbol{x}\right)$ (the most popular choice is $\mu\left(\boldsymbol{x}\right) = \exp\left(\boldsymbol{x\beta}\right)$)

- The conditional distribution of y is given by Poisson distribution for $\lambda = \mu\left(\boldsymbol{x}\right)$
$$f\left(y|\,\boldsymbol{x}\right) = \frac{e^{-\lambda}\lambda^{y}}{y!} = \frac{\exp\left[-\mu\left(\boldsymbol{x}\right)\right]\left[\mu\left(\boldsymbol{x}\right)\right]^{y}}{y!},\ \lambda = \mu\left(\boldsymbol{x}\right) > 0$$

- Indeed for Poisson distribution $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \lambda = \mu\left(\boldsymbol{x}\right)$

- Partial effects for $\mu(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta})$ are the same as in loglinear model (semielasticities)

$$\frac{\partial \log[\mu(\boldsymbol{x})]}{\partial x_j} = \beta_j \approx \left. \frac{\Delta\mu(\boldsymbol{x})}{\mu(\boldsymbol{x})} \right/ \Delta x_j$$

- The parameters of this model can be consistently and efficiently estimated with ML.

- The most important limitation of Poisson model. For the Poisson distribution

$$\mathrm{E}(y|\,\boldsymbol{x}) = \mathrm{Var}(y|\,\boldsymbol{x}) = \lambda = \mu(\boldsymbol{x})$$

- This so called Poisson variance assumption is often violated in practice (usually there is no theoretical base to claim that the variance and expected value of the dependent variable are equal)

- Sometimes we make the weaker assumption that $\operatorname{Var}(y|\,\boldsymbol{x}) = \sigma^2 \operatorname{E}(y|\,\boldsymbol{x})$. We say that we have the overdispersion in the model if $\sigma^2 > 1$ and underdispersion if $\sigma^2 < 1$

- It is possible to prove that if $\operatorname{E}(y|\,\boldsymbol{x}) = \mu(\boldsymbol{x})$ the (Quasi) Maximum Likelihood ($QMLE$) estimator of $\boldsymbol{\beta}$ obtained from Poisson model is consistent for large class of the count models if conditional mean $\operatorname{E}(y|\,\boldsymbol{x})$ is correctly specified even if $\operatorname{Var}(y|\,\boldsymbol{x}) \neq \mu(\boldsymbol{x})$. In this sense Poisson model gives the robust estimates.

- However: in this case the standard ML estimates of variance covariance matrix is not valid. We have to use the robust formulas for $QMLE$ estimates of variance matrix. With such estimate of variance matrix, test statistics have standard asymptotic distributions. Inference is somewhat simplified if $\operatorname{Var}(y|\,\boldsymbol{x}) = \sigma^2 \operatorname{E}(y|\,\boldsymbol{x})$.

Example 9. *(Wooldridge). Effects of Education on fertility. Data is for* *women in Botswana. Dependent variable: number of children*

Explanatory variable	Linear (OLS)	Exponential (Poisson $QLME$)
educ	-.0644 (.0063)	-.0217 (.0025)
age	.272 (.017)	.337 (.009)
age^2	-.0019 (.0003)	-.0041 (.0001)
evermarr	.682 (.052)	.315 (.021)
urban	-.228 (.046)	-.086 (.019)
electric	-.262 (.076)	-.121 (.034)
tv	-.250 (.090)	-.145 (.041)
constant	-3.394 (1.13)	-5.375 (1.09)
Log-likelihood value		-6497,060
R-squared	.590	.598
$\widehat{\sigma}$	1.424	.867

$\mathrm{E}\left(y|\,\boldsymbol{x}\right)$The coefficients differ as the first model is linear ($\frac{\partial\mu(\boldsymbol{x})}{\partial x_j} = \beta_j$) and the second exponential ($\frac{\partial\log[\mu(\boldsymbol{x})]}{\partial x_j} = \beta_j$). The value of $\sigma^2 < 1$ implies that for this model we have underdispersion.

- Specification tests: in the context of Poisson model the most important tests are the test validity of conditional mean specification $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \mu\left(\boldsymbol{x}\right)$ and tests for validity of the specification of variance $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \mathrm{E}\left(y|\,\boldsymbol{x}\right)$ or $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \sigma^2\,\mathrm{E}\left(y|\,\boldsymbol{x}\right)$

Negative binomial model

- Two types of negative binomial model NegBin type I and NegBin type II

- NegBin type I is the a special kind of parametrization of negative binomial model for which $\mathrm{E}\left(y|\,\boldsymbol{x}\right) = \mu\left(\boldsymbol{x}\right)$ and $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = \sigma^2\,\mathrm{E}\left(y|\,\boldsymbol{x}\right)$ and $\sigma^2 = 1+\eta^2$. This implies that negative NegBin type I can only be used if we have overdispersion.

- ML estimates of the NegBin type I of $\boldsymbol{\beta}$ are generally not consistent if $\mathrm{E}\left(y|\,\boldsymbol{x}\right)$ is correctly specified but distribution is not of NegBin type I. In this sense this model is less robust that Poisson model.

- NegBin type II model

$$y_i|\,\boldsymbol{x}_i, c_i \sim Poisson\left[c_i m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]$$

- If we assume that c_i and $\boldsymbol{x}_i$ are independent and c_i has gamma distribution than distribution of y_i can be shown to be negative binomial with

$$\begin{aligned} \mathrm{E}\left(y|\,\boldsymbol{x}\right) &= m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) \\ \mathrm{Var}\left(y|\,\boldsymbol{x}\right) &= m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) + \eta^2\left[m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]^2 = m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\left[1+\eta^2 m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right] \end{aligned}$$

- The NegBin type II model can also be only used for the case of overdispersion.

- It is possible to prove that for *fixed* η^2 the NegBin type II estimates of $\boldsymbol{\beta}$ is as robust as Poisson model. Therefore the two step procedure

  – estimate β for some η^2 (e.g. $\eta^2 = 1$)
  – estimate η^2 using estimates of β from first stage

  is also as robust as Poisson estimates - only requires correct specification of $\mathrm{E}\left(y|\,\boldsymbol{x}\right)$. It is however more efficient that Poisson if $\mathrm{Var}\left(y|\,\boldsymbol{x}\right) = m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\left[1+\eta^2 m\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]$

Binomial regression model

- Sometimes the discrete responses has an upper bound (e.g. children mortality in a family has upper bound equal to number of children ever born)

- In this case it is natural to assume that y_i has the binomial distribution

$$y_i \sim \text{Binominal}\left[n_i, p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right]$$

where $0 \leq p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) \leq 1$ is the probability of success in each trial and n_i is the number of trials.

- Conditional mean and variance for this model are equal to

$$\mathrm{E}\left(y_i | \boldsymbol{x}_i, n_i\right) = n_i p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) = m\left(\boldsymbol{x}_i \boldsymbol{\beta}\right)$$
$$\mathrm{Var}\left(y_i | \boldsymbol{x}_i, n_i\right) = n_i p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\left[1 - n_i p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right]$$

- Typically $p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) = G\left(\boldsymbol{x}_i \boldsymbol{\beta}\right)$ where $G\left(\cdot\right)$ is a standard distribution function such as normal or logistic distribution.

- The probability function for an observation

$$f\left(y | \boldsymbol{x}, n\right) = \binom{n}{y} p\left(\boldsymbol{x}, \boldsymbol{\beta}\right)^y \left[1 - p\left(\boldsymbol{x}, \boldsymbol{\beta}\right)\right]^{n-y}$$

- The loglikelihood function for one observation in this model:

$$\ell_i\left(\boldsymbol{\beta}\right) = y_i \log\left[p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right] + \left(n_i - y_i\right) \log\left[1 - p\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)\right]$$

first term is dropped as it does not depend on $\boldsymbol{\beta}$.

- This model can be estimated with use of ML or Quasi ML method.