Matematické Metody v Ekonometrii 11.

GLM

Blanka Šedivá

KMA

zimní semestr 2016/2017

General model for discrete choice

- $$P\left(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}\right) = 1 - F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right)$$
$$P\left(Y_i = 0|\boldsymbol{x}_i, \boldsymbol{\beta}\right) = F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right),$$

- $$\mathrm{E}\left(Y_i|\boldsymbol{x}, \boldsymbol{\beta}\right) = 1 \cdot P\left(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}\right) + 0 \cdot P\left(Y_i = 0|\boldsymbol{x}_i, \boldsymbol{\beta}\right) = 1 - F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right)$$

- $$Y_i = 1 - F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right)$$

- $$L(\boldsymbol{\beta}) = \prod_i P(y_i) = \prod_{i=1}^{n} \left(1 - F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)^{y_i} \left(F\left(-\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)^{1-y_i}$$

Several type of models

- Probit model

$$P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_i\boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i\boldsymbol{\beta})$$

- Logit model

$$P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_i\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}$$

- Gumbel model
  $P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = 1 - F(-\boldsymbol{x}_i\boldsymbol{\beta}) = 1 - (1 - \exp(-\exp(-\boldsymbol{x}_i\boldsymbol{\beta}))) = \exp(-\exp(-\boldsymbol{x}_i\boldsymbol{\beta}))$

Poisson regression I

- The Poisson distribution for a random variable Y has the following probability mass function for a given value $Y = y$:
$$\mathrm{P}(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \text{ for } y = 0, 1, 2, \ldots$$
- Poisson distribution is characterized by the single parameter λ, which is the mean rate of occurrence for the event being measured. For the Poisson distribution, it is assumed that large counts (with respect to the value of λ) are rare.
- In Poisson regression the dependent variable Y is an observed count that follows the Poisson distribution. The rate λ is determined by a set of $p-1$ predictors $\boldsymbol{X} = (X_1, \ldots, X_{p-1})$. The expression relating these quantities is
$$\lambda = \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right).$$

Poisson regression II

- Thus, the fundamental Poisson regression model for observation i is

$$\mathrm{P}(Y_i = y_i | \mathbf{X}_i, \boldsymbol{\beta}) = \frac{e^{-\exp(\boldsymbol{X}\boldsymbol{\beta})} \exp(\boldsymbol{X}\boldsymbol{\beta})^{y_i}}{y_i!}$$

- for a sample of size n, the likelihood for a Poisson regression is :

$$L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \frac{e^{-\exp(\boldsymbol{X}_i\boldsymbol{\beta})} \exp(\boldsymbol{X}_i\boldsymbol{\beta})^{y_i}}{y_i!}.$$

- the log likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{X}_i \boldsymbol{\beta} - \sum_{i=1}^{n} \exp(\boldsymbol{X}_i\boldsymbol{\beta}) - \sum_{i=1}^{n} \log(y_i!)$$

- maximizing the likelihood (or log likelihood) has no closed-form solution, so a technique like iteratively reweighted least squares is used to find an estimate of the regression coefficients

Deviance Test for Poisson regression

- Changes in the deviance can be used to test the null hypothesis that any subset of the β's is equal to 0. Suppose we test that $r < p$ of the β's are equal to 0. Then the deviance test statistic is

$$G^2 = -2l(\hat{\beta}^{(0)}) - (-2l(\hat{\beta}))$$

- where $-2l(\hat{\beta})$ is the deviance of the fitted (full) model and $-2l(\hat{\beta}^{(0)})$ is the deviance of the model specified by the null hypothesis evaluated at the maximum likelihood estimate of that reduced model.
- This test statistic has a χ^2 distribution with $p - r$ degrees of freedom.
- This test procedure is analagous to the general linear $F-$test procedure for multiple linear regression.

Goodness-of-Fit

- Overall performance of the fitted model can be measured by two different chi-square tests.
- There is the Pearson statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right))^2}{\exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right)}$$

- and the deviance statistic

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right)}\right) - (y_i - \exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right))\right]$$

- Both of these statistics are approximately χ^2 distributed with $n - p$ degrees of freedom.

Pseudo R^2

- The value of R^2 used in linear regression also does not extend to Poisson regression.
- $R^2 = \frac{\ell(\hat{\boldsymbol{\beta}}_0) - \ell(\hat{\boldsymbol{\beta}})}{\ell(\hat{\boldsymbol{\beta}}_0)} = 1 - \frac{-2\ell(\hat{\boldsymbol{\beta}})}{-2\ell(\hat{\boldsymbol{\beta}}_0)}$,
  where $\ell(\hat{\boldsymbol{\beta}}_0)$ is the log likelihood of the model when only the intercept is included.
- The pseudo R^2 goes from 0 to 1 with 1 being a perfect fit.

Residuals for Poisson regression I

- The **raw residual** is the difference between the actual response and the estimated value from the model.
  Remember that the variance is equal to the mean for a Poisson random variable. Therefore, we expect that the variances of the residuals are unequal. This can lead to difficulties in the interpretation of the raw residuals, yet it is still used.

$$r_i = y_i - \exp\left(\boldsymbol{X}_i \boldsymbol{\beta}\right)$$

- The **Pearson residual** corrects for the unequal variance in the raw residuals by dividing by the standard deviation.

$$p_i = \frac{r_i}{\sqrt{\hat{\phi} \exp\left(\boldsymbol{X}_i \boldsymbol{\beta}\right)}},$$

where $\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right))^2}{\exp\left(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}\right)}$

Residuals for Poisson regression II

- **Deviance residuals** are also popular because the sum of squares of these residuals is the deviance statistic.

$$d_i = \mathrm{sgn}(y_i - \exp\left(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right))\sqrt{2\left(y_i \ln\left(\frac{y_i}{\exp\left(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right)}\right) - \left(y_i - \exp\left(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right)\right)\right)}$$

- The hat matrix serves the same purpose as in the case of linear regression - to measure the influence of each observation on the overall fit of the model $H = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{XWX})^{-1}\mathbf{XW}^{1/2}$, where W is an $n \times n$ diagonal matrix with the values of $\exp\left(\mathbf{X}_i\hat{\beta}\right)$ on the diagonal
- The Studentized Pearson residuals $sp_i = \frac{p_i}{\sqrt{1-h_{i,i}}}$
- Studentized deviance residuals $sp_i = \frac{d_i}{\sqrt{1-h_{i,i}}}$

Generalized Linear Models (GLM)

- Generalized linear models provides a generalization of ordinary least squares regression that relates the random term (the response Y) to the systematic term (the linear predictor) via a link function (denoted by $g(\cdot)$).
$$\mathrm{E}(Y) = \mu = g^{-1}(\mathbf{X}\boldsymbol{\beta}),$$
- Some common link functions are:
  - identity link: $g(\mu) = \mu = \mathbf{X}\boldsymbol{\beta}$, which is used in traditional linear regression.
  - logit link: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\boldsymbol{\beta} \Rightarrow \mu = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1+e^{\mathbf{X}\boldsymbol{\beta}}}$, which is used in logistic regression.
  - log link: $g(\mu) = \log(\mu) = \mathbf{X}\boldsymbol{\beta} \Rightarrow \mu = e^{\mathbf{X}\boldsymbol{\beta}}$, which is used in Poisson regression.
  - ...

Distribution of variable Y I

- The random variable Y is assumed to belong to an exponential family distribution where the density can be expressed in the form

$$f(y;\theta;\varphi) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right),$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are specified functions, θ is a parameter related to the mean of the distribution, and ϕ is called the dispersion parameter.

  - the normal distribution is used for traditional linear regression
    - the binomial distribution is used for logistic regression:
    $f(y;\pi) = \begin{pmatrix} m \\ y \end{pmatrix} \pi^y (1-\pi)^{m-y}$
  - the Poisson distribution is used for Poisson regression:
    $f(y;\lambda) = \exp(y_i \ln \lambda - \lambda \ln y_i!)$

Distribution of variable Y II

- the exponential distribution :
  $f(y;\pi;\mu) = \exp\left(y \ln\left(\frac{\mu}{1-\mu}\right) + m \ln(1-\pi) + \ln\left(\begin{array}{c} m \\ y \end{array}\right)\right)$
- gamma regression, inverse Gaussian (normal) regression, and negative binomial regression,

- the variance is typically a function of the mean and is often written as

$$\mathrm{Var}(Y) = V(\mu) = V(g^{-1}(\mathbf{X}\beta))$$

Estimation of parameters

- The unknown parameters, β, are typically estimated with maximum likelihood techniques (in particular, using iteratively reweighted least squares), Bayesian methods, or quasi-likelihood methods.
- maximum likelihood techniques is based on log likehood function

$$\ell(\theta) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right)$$

$\mu_i = EZ_i = b'(\theta_i)$ $\qquad\qquad$ $\eta_i = g(\mu_i)$

- The quasi-likelihood is a function which possesses similar properties to the log-likelihood function and is most often used with count or binary data.
- Specifically, for a realization y of the random variable Y, it is defined as $Q(\mu; y) = \int_y^\mu \frac{y-t}{\sigma^2 V(t)} dt$, where σ^2 is a scale parameter. There are also tests using likelihood ratio statistics for model development to determine if any predictors may be dropped from the model.

MATLAB

- **b,dev,stats = glmfit(X,y,distr,param1,val1,param2,val2,...)**
- **distr** can be any of the following strings: 'binomial', 'gamma', 'inverse gaussian', 'normal' (the default), and 'poisson'
- parameter **'link'**:

'identity', default for the distribution 'normal'	$\mu = \boldsymbol{X}\boldsymbol{\beta}$
'log', default for the distribution 'poisson'	$\ln \mu = \boldsymbol{X}\boldsymbol{\beta}$
'logit', default for the distribution 'binomial'	$\ln(\mu(1-\mu)) = \boldsymbol{X}\boldsymbol{\beta}$
'probit'	$norminv(\mu) = \boldsymbol{X}\boldsymbol{\beta}$
'comploglog'	$\ln(-\ln(1-\mu)) = \boldsymbol{X}\boldsymbol{\beta}$
'reciprocal', default for the distribution 'gamma'	$1/\mu = \boldsymbol{X}\boldsymbol{\beta}$
'loglog'	$\ln(-\ln(\mu)) = \boldsymbol{X}\boldsymbol{\beta}$
p (a number), default for the distribution 'inverse gaussian' (with $p = -2$)	$\mu^p = \boldsymbol{X}\boldsymbol{\beta}$
cell array of the form FL FD FI, containing three function handles, created using @, that define the link (FL), the derivative of the link (FD), and the inverse link (FI).	$FL(\mu)$, $FD = \frac{\mathrm{d}FL(\mu)}{\mathrm{d}\mu}$, $FI = FL^{-1}$