Universität Zürich[UZH]

Soziologisches Institut

Data Analysis – Advanced Statistics with Python

Dr. Julia Jerke

jerke@soziologie.uzh.ch

Thursday, 12.15pm – 13.45pm, AND 2.46

Session 5 – Advanced regression methods

Agenda

1. Multinomial regression
2. Ordinal regression
3. Count data regression

Overview

- In general, regression models aim at investigating the relationship between a dependent variable Y and several explanatory variables $X_1, X_2, X_3, \ldots, X_p$
- However, there are countless different regression methods
- Main decision criteria for the choice of the regression model is the **characterization of the dependent variable**, e.g.:

Y is a continuous variable	⟶	Linear regression
Y is a binary variable	⟶	Logistic regression
Y is a categorical variable	⟶	Multinomial regression
Y is a rank variable	⟶	Ordinal regression
Y is a count variable	⟶	Poisson regression or NegBin
Y is an event history variable	⟶	Cox regression

Overview

- In general, regression models aim at investigating the relationship between a dependent variable Y and several explanatory variables $X_1, X_2, X_3, \ldots, X_p$
- However, there are countless different regression methods
- Another criteria for the choice of the regression model is the **structure of the data**, e.g.:

The data is nested in clusters	⟶	Multilevel regression
Longitudinal data	⟶	Panel regression
Time series	⟶	Time series regression

- Depending on the scale of the dependent variable, these models can be combined, e.g.
  - A *multilevel multinomial regression:* for clustered data with a categorical dependent variable
  - A *logistic panel regression:* for longitudinal data with a binary dependent variable

Disclaimer

- So far, we covered linear and logistic regression
- These are the most often applied regression methods
- However, they are often incorrectly applied, e.g. in situations in which other regression designs would have been more appropriate
- This session ***IS***:
    - an overview over various specific regression methods
    - a demonstration of the range of regression methods/designs that can be used as soon as you can't apply standard regressions any more
- This session ***IS NOT***:
    - an in-depth statistical presentation
    - A how-to turorial that will teach you how to use these methods in detail

1. Multinomial regression

Multinomial vs logistic regression

- Specifying a (non-)linear relationship between a dependent **categorical** variable Y and one or more explanatory variables $X_1, X_2, X_3, \ldots$
  - *Simple multinomial regression:* one explanatory variable X_1
  - *Multiple multinomial regression:* two or more explanatory variables $X_1, X_2, \ldots, X_p$
- Basis: the dependent variable has outcomes $\mathrm{k} = \{1, 2, \ldots, K\}$, whereas $\#k \geq 3$, and there is no order
- Transferring the principle of the logistic regression to the categorical case:
  - Prediction of probabilities for each category
  - Choosing one reference category, e.g. $k = 1$, and predicting the probability of falling in categories $\mathrm{k} = \{2, \ldots, K\}$ compared to the probability of $k = 1$

Multinomial regression model

Objective: Modelling the probability of falling in categories $\mathrm{k} = \{2, \dots, K\}$ compared to the category $k = 1$

- Just like for the logistic regression, we predict for each $k \in \{2, \dots, K\}$ and the reference category $k = 1$:

$$\ln\left(\frac{P(y=k)}{P(y=1)}\right) = \beta_0^{\{k\}} + \beta_1^{\{k\}} * \mathrm{x}_1 + \beta_2^{\{k\}} * \mathrm{x}_2 + \cdots + \beta_\mathrm{p}^{\{k\}} * \mathrm{x}_\mathrm{p},$$

  whereas again:

  - β_0 is the constant (i.e., the predicted value if $x_1, \dots, x_p = 0$)
  - $\beta_1, \dots, \beta_p$ are the regression coefficients

- The result are $K - 1$ different regression equations with individual coefficients $\beta_0^{\{k\}}, \dots, \beta_p^{\{k\}}$ each
- There is one constraint – the predicted probabilities for $\mathrm{k} = \{1, 2, \dots, K\}$ must sum to one:

$$\sum_{k=0}^{K} P(y=k) = 1$$

- The model parameters are estimated with Maximum Likelihood
- Note: since $(K-1) * (p-1)$ coefficients have to be estimated small data sets might be problematic

Interpretation of the coefficients from a multinomial regression

- Basically, the interpretation is pretty much the same as for logistic regression
- Different coefficients:
  - *Log Odds*: $\boldsymbol{\beta_p}$
    - These are the raw coefficients from the regression model
    - They have a linear interpretation
    - The sign of the coefficient indicates the direction of the effect
  - *Odds (Ratio)*: $\boldsymbol{e^{\beta_p}}$
    - The odds are changed by the factor of e^{β_p}
    - Respectively, the odds ratio is $\boldsymbol{e^{\beta_p}}$
  - *Marginal Effects:* ***a*verage marginal effect, marginal effect at the mean**, etc.
    - AME: the average of the effects for each indivdual observation
    - MEM: the effect of the «average» observation (the mean of all $X_1, X_2, \ldots, X_p$)

Evaluation of effects in a multinomial regression

- It is not as straightforward as in simpler regression models to assess, whether a variable has an effect
  - Since there are $K-1$ coefficients for each variable (from $K-1$ different regression equations), we cannot only look at the single coefficients
  - In principle, for an effect to be zero, all $K-1$ coefficients might have to be zero
- Instead: comparing models with and without the specific variable as a whole
  - AIC and BIC
  - Log-Likelihood Ratio

2. Ordinal regression

Ordinal vs linear and logistic regression

- Specifying a (non-)linear relationship between a dependent **rank** variable Y and one or more explanatory variables $X_1, X_2, X_3, \ldots$
  - *Simple logistic regression:* one explanatory variable X_1
  - *Multiple logistic regression:* two or more explanatory variables $X_1, X_2, \ldots, X_p$
- Basis: the dependent variable has outcomes $\mathrm{k} = \{0, 1, 2, \ldots, K\}$, whereas $\#k \geq 3$, and they are ordered
- Assumption: the dependent variable is based on a latent continuous variable that has been discretized
- Linear regression cannot be used since it requires continuous data and makes continuous predictions
- Transferring the principle of the logistic regression to the ordinal case:
  - Prediction of cumulative probabilities for each category
  - What is the probability of being in a specific category or below it given $X_1, X_2, \ldots, X_p$?

Ordinal regression

- Assumption: basis of Y is a latent continuous variable Z

$$y = f(x) = \begin{cases} 0, & z \leq \theta_1 \\ 1, & \theta_1 < z \leq \theta_2 \\ 2, & \theta_2 < z \leq \theta_3 \\ & \vdots \\ K, & \theta_K < z \end{cases}$$

- Note that we have: $\theta_1 < \theta_2 < \cdots < \theta_K$
- We want to predict the **cumulative probabilities** for each category: given the covariates, what is the probability of being in a given category or below?

$$\ln\left(\frac{P_k}{1 - P_k}\right) = \ln\left(\frac{P(y \leq k)}{P(y > k)}\right)$$

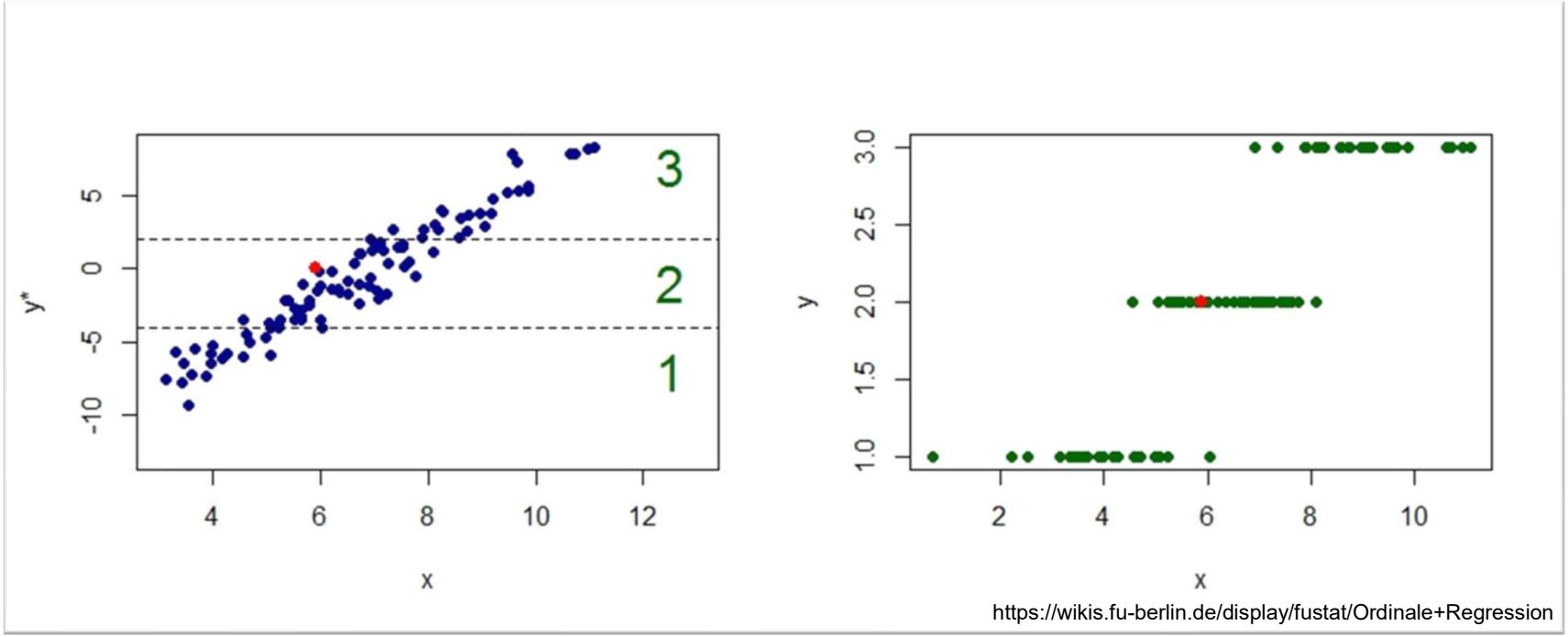
  - P_0 is the probability of the outcome $k = \{0\}$
  - P_1 is the probability of the outcome $k = \{0, 1\}$, hence $k \leq 1$
  - P_{K-1} is the probability of the outcome $k = \{0, 1, 2, \ldots, K-1\}$, hence $k \leq K - 1$
  - P_K can be calculated with $P_1, P_1, \ldots, P_{K-1}$

Ordinal regression

- **Underlying assumption**: the latent variable Z can be modelled with a standard linear regression model

$$z = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_p * x_p$$

- We then have to estimate the coefficients $\beta_1, \ldots, \beta_p$ and the thresholds $\theta_0, \ldots, \theta_{k-1}$



https://wikis.fu-berlin.de/display/fustat/Ordinale+Regression

Ordinal regression

- Predicting the **cumulative probabilities:**

$$\ln\left(\frac{P_k}{1-P_k}\right) = \ln\left(\frac{P(y \leq k)}{P(y > k)}\right)$$

- In theory, we have K different regression equations
- In practice, these regressions share the same coefficients $\beta_1, \ldots, \beta_p$ and only differ by their intercept β_0
- **Proportional Odds Assumption**: the regression coefficients are the same for all categories of y, the increase in probability between two categories is always the same

$$\ln\left(\frac{P(y \leq k)}{P(y > k)}\right) = \theta_k - (\beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_p * x_p)$$

  Whereas again:

  - $\beta_1, \ldots, \beta_p$ are the regression coefficients, they are the same for each category k
  - θ_k is the threshold (intercept) for the k^{th} regression equation

- Note: usually we are not really interested in the threshold θ_k (they are necessary for the prediction), rather we are interested in the regression coefficients $\beta_1, \ldots, \beta_p$

Ordinal regression example

Example taken from:

Elgar, F. J., Craig, W., Boyce, W., Morgan, A., & Vella-Zarb, R. (2009). Income inequality and school bullying: multilevel study of adolescents in 37 countries. *Journal of Adolescent Health*, *45*(4), 351-359.

Dependent variable: school bullying

> ‘‘How often have you bullied others at school in the past couple of months?’’
>
> Responses were on a five-point ordinal scale:
>
> 0 - not at all,
>
> 1 - once or twice
>
> 2 - two or three times per month
>
> 3 - once a week
>
> 4 - several times a week

Independent variable: income inequality

Sample: 66’910 11-year-olds in 37 countries

Ordinal regression example

Table 4
Hierarchical ordinal regression analysis for variables predicting bullying

	Model 1		Model 2		Model 3		Model 4	
Variable	*B* (SE)	OR (95% CI)	*B* (SE)	OR (95% CI)	*B* (SE)	OR (95% CI)	*B* (SE)	OR (95% CI)
1. Males (N = 32,942)								
Country wealth	−.20 (.01)	.82 (.79–.85)	−.20 (.02)	.82 (.79–.85)	−.20 (.02)	.82 (.79–.86)	−.19 (.02)	.83 (.80–.86)
Individual wealth	.04 (.01)	1.04 (1.01–1.07)	.04 (.02)	1.04 (1.02–1.07)	.04 (.02)	1.04 (1.01–1.07)	.05 (.01)	1.05 (1.02–1.08)
Income inequality	.16 (.02)	1.17 (1.12–1.21)	.15 (.02)	1.16 (1.12–1.20)	.14 (.02)	1.16 (1.11–1.20)	.15 (.02)	1.16 (1.12–1.20)
Family support			−.05 (.01)	.96 (.93–.98)				
Peer support					.01 (.01)	1.01 (.98–1.04)		
School support							−.23 (.02)	.80 (.77–.82)
Thresholds:								
1 (once or twice)	.60 (.02)		.59 (.02)		.60 (.02)		.60 (.02)	
2 (two or three times)	2.04 (.02)		2.05 (.02)		2.06 (.02)		2.05 (.02)	
3 (once a week)	2.73 (.03)		2.74 (.03)		2.75 (.03)		2.75 (.03)	
4 (several times a week)	3.39 (.04)		3.40 (.04)		3.42 (.04)		3.41 (.04)	
2. Females (N = 33,875)								
Country wealth	−.19 (.03)	.83 (.79–.87)	−.19 (.03)	.83 (.79–.87)	−.18 (.03)	.84 (.80–.88)	−.17 (.03)	.85 (.80–.89)
Individual wealth	−.03 (.02)	.97 (.94–1.00)	−.03 (.02)	.97 (.94–1.01)	−.03 (.02)	.97 (.94–1.01)	−.01 (.02)	.99 (.96–1.02)
Income inequality	.22 (.02)	1.24 (1.19–1.29)	.22 (.02)	1.24 (1.19–1.29)	.22 (.02)	1.24 (1.19–1.29)	.21 (.02)	1.23 (1.19–1.28)
Family support			−.11 (.02)	.89 (.86–.92)				
Peer support					.02 (.02)	1.02 (.99–1.06)		
School support							−.31 (.02)	.74 (.71–.76)
Thresholds:								
1 (once or twice)	1.22 (.02)		1.24 (.02)		1.22 (.02)		1.23 (.02)	
2 (two or three times)	2.83 (.03)		2.86 (.03)		2.85 (.03)		2.85 (.03)	
3 (once a week)	3.41 (.04)		3.44 (.04)		3.44 (.04)		3.43 (.04)	
4 (several times a week)	4.10 (.05)		4.13 (.05)		4.16 (.05)		4.13 (.05)	

Note: R^2 (Cox and Snell) = .03 to .05.
OR = odds ratio; CI = confidence interval; SE = standard error.

$OR = e^B$

Thresholds
$\theta_1, \theta_2, \theta_3, \theta_4$

3. Count data regression

What are count data?

Technically speaking:

- Discrete data that only takes non-negative integer values $(0,1,2,3,4, \dots)$
- These values are the result of *counting* instead of *ranking*
- Usually, they represent the number of occurrences of an event within a fixed period
- In general, we have rare events:
    - Usually strongly right-skewed data, with a large proportion of values in the lower range
    - Most of the data are concentrated on a few small discrete values
    - Often, we have a lot of zeros in the data

Examples

- Length of hospital stay
- Number of doctor visits
- Number of murders
- Number of awards received by students

Overview

The count regression toolbox

- Poisson regression model
- Negative binomial regression model
- Zero-truncated model (poisson and negative binomial)
- Zero-inflated model (poisson and negative binomial)
- Hurdle models

(Note: this list covers the most common methods, but is not exhaustive)

Poisson distribution

Basis

- We observe how often a certain event occurs within specific time period
- We then want to investigate the relationship between the expected number of occurrences of the event and some covariates

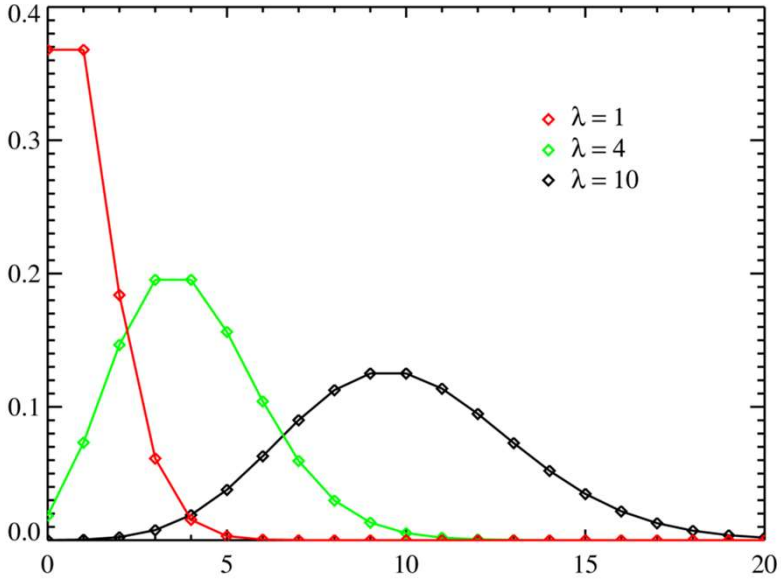
Poisson distribution

- The Poisson distribution is often used to represent count data
- Specification:

$$P(Y = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \qquad k = 0, 1, 2, 3, \ldots$$

whereas:

  - $\lambda = E(Y) = Var(Y)$
  - λ is called the *intensity parameter* and represents the expected number of occurrences in a fixed period of time



Poisson regression model

The Poisson regression is the standard model for count data

Objective: Modelling the expected number $E(Y|X)$ of occurrences of the event of interest given the independent variables $X_1, X_2, \ldots, X_p$

- Specification of the regression model:

$$E(Y_i|X_i) = e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\cdots+\beta_p*x_p}$$

or in its log-linear form:

$$\ln\left(E(Y_i|X_i)\right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_p * x_p$$

whereas again:

  - β_0 is the constant (i.e., the predicted value if $x_{i1}, \ldots, x_{ip} = 0$)
  - $\beta_1, \ldots, \beta_p$ are the regression coefficients (i.e., the effect of a one unit change in $X_1, X_2, \ldots, X_p$, ceteris paribus)

- The model parameters are estimated with Maximum Likelihood

Poisson regression: interpretation of coefficients

- As for the logistic regression, coefficients from the linear form are difficult to interpret (what does a change in $\ln(E(Y|X))$ even mean?)
- But we can draw on the non-linear form:

$$
\begin{aligned}
E\left(Y\middle|X_{[x_p+1]}\right) &= e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\cdots+\beta_p*(x_p+1)} \\
&= e^{\beta_0} * e^{\beta_1 \cdot x_1} * e^{\beta_2 \cdot x_2} * \ldots * e^{\beta_p \cdot (x_p+1)} \\
&= e^{\beta_0} * e^{\beta_1 \cdot x_1} * e^{\beta_2 \cdot x_2} * \ldots * e^{\beta_p \cdot x_p} * e^{\beta_p} \\
&= E\left(Y\middle|X_{[x_p]}\right) * e^{\beta_p}
\end{aligned}
$$

$$
E\left(Y\middle|X_{[x_p+1]}\right) - E\left(Y\middle|X_{[x_p]}\right) = E\left(Y\middle|X_{[x_p]}\right) * (e^{\beta_p} - 1) \quad \text{or} \quad \frac{E\left(Y\middle|X_{[x_p+1]}\right)}{E\left(Y\middle|X_{[x_p]}\right)} = e^{\beta_p}
$$

- **a one-unit change in X_j will proportionately change the *expected number of events* by $e^{\beta_p} - 1$ (however, for small values of β_p is often a good approximation)**

Poisson regression

Assumptions

- The dependent variable is follows a Poisson distribution
- Observations are independent of each other, i.e. the occurrence of one event does not affect the occurrence of another event
- Constant intensity parameter λ, i.e. the expected number of occurrences does not vary over time
- Linear relationship between $\ln(E(Y|X))$ and the independent variables
- No homoscedasticity required!

Equidispersion

- An implication of the model and these assumptions is e*quidispersion*: the mean of the independent variable equals its variance, $E(Y) = Var(Y)$
- However, *equidispersion* is often violated because the variance exceeds the mean, $Var(Y) \gg E(Y)$
- In case of such *overdispersion*, standard errors and t scores are likely to be biased
- There are alternative ways to model the count regression in these situations

Example

Example taken from:

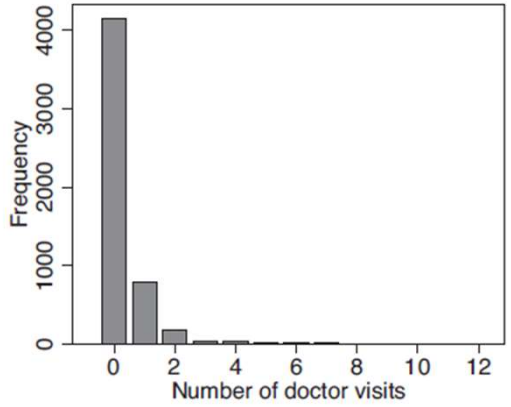
Cameron, Adrian Colin, and P. K. Trivedi. Regression Analysis of Count Data . Second edition. Cambridge: Cambridge University Press, 2013.

Dependent variable: number of doctor visits in the past two weeks

Sample: single-adults from the Australian Health Survey 1977–78, sample of size is 5'190

Table 3.1. *Doctor visits: Actual frequency distribution*

Count	0	1	2	3	4	5	6	7	8	9
Frequency	4,141	782	174	30	24	9	12	12	5	1
Relative frequency	0.798	0.151	0.033	0.006	0.005	0.002	0.002	0.002	0.001	0.000



Example

Independent Variables

Table 3.2. *Doctor visits: Variable definitions and summary statistics*

Variable	Definition	Mean	Standard deviation
DVISITS	Number of doctor visits in past two weeks	0.302	0.798
SEX	Equals 1 if female	0.521	0.500
AGE	Age in years divided by 100	0.406	0.205
AGESQ	*AGE* squared	0.207	0.186
INCOME	Annual income in tens of thousands of dollars	0.583	0.369
LEVYPLUS	Equals 1 if private health insurance	0.443	0.497
FREEPOOR	Equals 1 if free government health insurance due to low income	0.043	0.202
FREEREPA	Equals 1 if free government health insurance due to old age, disability, or veteran status	0.210	0.408
ILLNESS	Number of illnesses in past two weeks	1.432	1.384
ACTDAYS	Number of days of reduced activity in past two weeks due to illness or injury	0.862	2.888
HSCORE	General health questionnaire score using Goldberg's method	1.218	2.124
CHCOND1	Equals 1 if chronic condition not limiting activity	0.403	0.491
CHCOND2	Equals1 if chronic condition limiting activity	0.117	0.321

Example

Results

- Estimation results are in column 1 (estimated with quasi maximum likelihoood)
- Recent health meassures have a strong positive effect on the number of doctor visits
- Long-term health status has a positive effect as well
- Women have a higher number of doctor visits

Table 3.3. *Doctor visits: Poisson* QMLE *with different standard error estimates*

	Coefficient	Standard errors						*t* statistic
Variable	Poisson PMLE	RS	MLH	MLOP	NB1	NB2	Boot	RS
ONE	−2.224	0.254	0.190	0.144	0.219	0.207	0.271	−8.74
SEX	0.157	0.079	0.056	0.041	0.065	0.062	0.076	1.98
AGE	1.056	1.364	1.001	0.750	1.153	1.112	1.391	0.77
AGESQ	−0.849	1.460	1.078	0.809	1.242	1.210	1.477	−0.58
INCOME	−0.205	0.129	0.088	0.062	0.102	0.096	0.129	−1.59
LEVYPLUS	0.123	0.095	0.072	0.056	0.083	0.077	0.100	1.29
FREEPOOR	−0.440	0.290	0.180	0.116	0.207	0.188	0.293	−1.52
FREEREPA	0.080	0.126	0.092	0.070	0.106	0.102	0.132	0.63
ILLNESS	0.187	0.024	0.018	0.014	0.021	0.021	0.024	7.81
ACTDAYS	0.127	0.008	0.005	0.004	0.006	0.006	0.008	16.33
HSCORE	0.030	0.014	0.010	0.007	0.012	0.012	0.014	2.11
CHCOND1	0.114	0.091	0.067	0.051	0.077	0.071	0.087	1.26
CHCOND2	0.141	0.123	0.083	0.059	0.096	0.092	0.120	1.15
−ln L	3355.5							

Note: Different standard error estimates are due to different specifications of ω, the conditional variance of y. RS, unspecified ω robust sandwich estimate; MLH, $\omega = \mu$ Hessian estimate; MLOP, $\omega = \mu$ summed outer product of first derivatives estimate; NB1, $\omega = \phi\mu = (1+\alpha)\mu$ where $\alpha = 0.382$; NB2, $\omega = \mu + \alpha\mu^2$where $\alpha = 0.286$; and Boot, unspecified ω bootstrap estimate.

Example

Interpretation of the raw coefficients I

- Being ill leads to a 0.206 ($e^{0.187} - 1$) proportionate increase in the number of expected doctor visits, or 20.06%
- For example:
  - for someone who has not been ill and who went to the doctor 2 times, we would expect about 2.4 doctor visits if that person had been ill ($2.4 * e^{0.187}$)
  - for someone who has not been ill and who went to the doctor 5 times, we would expect about 6 doctor visits if that person had been ill ($5 * e^{0.187}$)
- The example demonstrates that the change in doctor visits depends on the values of X

Table 3.6. *Doctor visits: Poisson QMLE mean effects and scaled coefficients*

Variable	Coefficient QMLE	Mean effect AME	Mean effect MEM	Mean effect OLS	Scaled Coeffs Elast	Scaled Coeffs SSC	Summary Statistics Mean	Summary Statistics Standard deviation
ONE	−2.224							
SEX	1.157	0.047	0.035	0.034	0.082	0.078	0.521	0.500
AGE	1.056	0.319	0.241	0.203	0.430	0.216	0.406	0.205
AGESQ	−0.849	−0.256	−0.193	−0.062	−0.176	−0.157	0.207	0.186
INCOME	−0.205	−0.062	−0.047	−0.057	−0.120	−0.076	0.583	0.369
LEVYPLUS	0.123	0.037	0.028	0.035	0.055	0.061	0.443	0.497
FREEPOOR	−0.440	−0.133	−0.100	−0.103	−0.019	−0.089	0.043	0.202
FREEREPA	0.080	0.024	0.018	0.033	0.017	0.033	0.210	0.408
ILLNESS	0.187	0.056	0.043	0.060	0.268	0.259	1.432	1.384
ACTDAYS	0.127	0.038	0.029	0.103	0.109	0.366	0.862	2.888
HSCORE	0.030	0.009	0.007	0.017	0.037	0.064	1.218	2.124
CHCOND1	0.114	0.034	0.026	0.004	0.046	0.056	0.403	0.491
CHCOND2	0.141	0.043	0.032	0.042	0.016	0.045	0.117	0.321

Note: AME, average over sample of effect of y of a one-unit change in x; MEM, effect on y of a one-unit change in x evaluated at average regressors; OLS, OLS coefficients; Elast, coefficients scaled by sample mean of x; SSC, coefficients scaled by standard deviation of x.

Example

Interpretation of the raw coefficients II

- Since we are dealing with proportionate changes that are scale-free, we can directly compare the coefficients
- For example: The effect of a one-day increase in inactive days is more than four times as high as the effect of a one-point increase of the health score

Table 3.6. *Doctor visits: Poisson QMLE mean effects and scaled coefficients*

Variable	Coefficient	Mean effect			Scaled Coeffs		Summary Statistics	
	QMLE	AME	MEM	OLS	Elast	SSC	Mean	Standard deviation
ONE	−2.224							
SEX	1.157	0.047	0.035	0.034	0.082	0.078	0.521	0.500
AGE	1.056	0.319	0.241	0.203	0.430	0.216	0.406	0.205
AGESQ	−0.849	−0.256	−0.193	−0.062	−0.176	−0.157	0.207	0.186
INCOME	−0.205	−0.062	−0.047	−0.057	−0.120	−0.076	0.583	0.369
LEVYPLUS	0.123	0.037	0.028	0.035	0.055	0.061	0.443	0.497
FREEPOOR	−0.440	−0.133	−0.100	−0.103	−0.019	−0.089	0.043	0.202
FREEREPA	0.080	0.024	0.018	0.033	0.017	0.033	0.210	0.408
ILLNESS	0.187	0.056	0.043	0.060	0.268	0.259	1.432	1.384
ACTDAYS	0.127	0.038	0.029	0.103	0.109	0.366	0.862	2.888
HSCORE	0.030	0.009	0.007	0.017	0.037	0.064	1.218	2.124
CHCOND1	0.114	0.034	0.026	0.004	0.046	0.056	0.403	0.491
CHCOND2	0.141	0.043	0.032	0.042	0.016	0.045	0.117	0.321

Note: AME, average over sample of effect of y of a one-unit change in x; MEM, effect on y of a one-unit change in x evaluated at average regressors; OLS, OLS coefficients; Elast, coefficients scaled by sample mean of x; SSC, coefficients scaled by standard deviation of x.

Example

Interpretation of the marginal effects

- As for the logistic regression, we need marginal effects to make more general statements about the effect
- Average Marginal Effect (AME):
  - The average of the effects for each indivdual observation
- Marginal Effect at the Mean (MEM):
  - The effect of the «average» observation (the mean of all $X_1, X_2, \ldots, X_p$)

Table 3.6. *Doctor visits: Poisson QMLE mean effects and scaled coefficients*

Variable	Coefficient QMLE	Mean effect AME	Mean effect MEM	Mean effect OLS	Scaled Coeffs Elast	Scaled Coeffs SSC	Summary Statistics Mean	Summary Statistics Standard deviation
ONE	−2.224							
SEX	1.157	0.047	0.035	0.034	0.082	0.078	0.521	0.500
AGE	1.056	0.319	0.241	0.203	0.430	0.216	0.406	0.205
AGESQ	−0.849	−0.256	−0.193	−0.062	−0.176	−0.157	0.207	0.186
INCOME	−0.205	−0.062	−0.047	−0.057	−0.120	−0.076	0.583	0.369
LEVYPLUS	0.123	0.037	0.028	0.035	0.055	0.061	0.443	0.497
FREEPOOR	−0.440	−0.133	−0.100	−0.103	−0.019	−0.089	0.043	0.202
FREEREPA	0.080	0.024	0.018	0.033	0.017	0.033	0.210	0.408
ILLNESS	0.187	0.056	0.043	0.060	0.268	0.259	1.432	1.384
ACTDAYS	0.127	0.038	0.029	0.103	0.109	0.366	0.862	2.888
HSCORE	0.030	0.009	0.007	0.017	0.037	0.064	1.218	2.124
CHCOND1	0.114	0.034	0.026	0.004	0.046	0.056	0.403	0.491
CHCOND2	0.141	0.043	0.032	0.042	0.016	0.045	0.117	0.321

Note: AME, average over sample of effect of y of a one-unit change in x; MEM, effect on y of a one-unit change in x evaluated at average regressors; OLS, OLS coefficients; Elast, coefficients scaled by sample mean of x; SSC, coefficients scaled by standard deviation of x.

Common issues of count regression

Overdispersion

- The variance exceeds the mean
- That imbalance cannot be reduced by including the covariates

Solution

- The standard model that accommodates overdispersion is the *negative binomial regression*
- The negative binomial regression is a generalization of the Poisson regression
- It gets rid of the very restrictive requirement that the mean equals the variance

Common issues of count regression

Zero truncation

- In some cases we observe no zeros at all
- Often this is the results of how the data is collected or how the dependent variable is defined
- Example: length of a hospital stay
  - Analyzing factors that influence how many days a person has to stay in hospital
  - By definition, we have no observations in the sample that stayed zero days in hospital

Solution

- Zero-truncated Poisson or negative binomial regression

Common issues of count regression

Zero inflation

- Excess zeros, i.e. the presence of more zeros than the Poisson model would predict
- Often this is the result of two different processes that are at work:
  - one determining whether there are zero events or any events
  - and a Poisson process determining how many events there are
  - Example: number of cigarettes smoked in an hour within a group where some are non-smokers
- This results in two different types of zeros:
  - True zeros: observations for which the event did not occur (e.g. smokers that did not smoke in that hour)
  - Excess zeros: observations for which the event cannot occur (e.g. non-smokers)

Solution

- Zero-inflated Poisson or negative binomial regression
- Hurdle models (modelling the two processes separately: 1) a binomial model for zero or not, 2) a zero-truncated model for the positive values)