Logistic, multinomial, and ordinal regression

I. Žežula[1]

[1]Faculty of Science
P. J. Šafárik University, Košice

Herľany 2010
14.4. – 16.4.2010, Herľany

Obsah

1) Logistic regression

- **Introduction**

Testing a risk factor: we want to check whether a certain factor adds to the probability of outbreak of a disease.

This corresponds to the following contingency table:

risk factor	disease status		sum
	ill	not ill	
exposed	n_{11}	n_{12}	n_{10}
unexposed	n_{21}	n_{22}	n_{20}
sum	n_{01}	n_{02}	n

1) Logistic regression

- **Introduction**

  Testing a risk factor: we want to check whether a certain factor adds to the probability of outbreak of a disease.

  This corresponds to the following contingency table:

risk factor	disease status		sum
	ill	not ill	
exposed	n_{11}	n_{12}	n_{10}
unexposed	n_{21}	n_{22}	n_{20}
sum	n_{01}	n_{02}	n

- Odds of the outbreak for both groups are:

$$o_1 = \frac{n_{11}}{n_{10} - n_{11}} = \frac{\frac{n_{11}}{n_{10}}}{1 - \frac{n_{11}}{n_{10}}} = \frac{\hat{p}_1}{1 - \hat{p}_1}, \quad o_2 = \frac{\hat{p}_2}{1 - \hat{p}_2}$$

1) Logistic regression

- As a measure of the risk, we can form **odds ratio**:

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$
$$o_1 = \mathbf{OR} \cdot o_2$$

1) Logistic regression

- As a measure of the risk, we can form **odds ratio**:

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$

$$o_1 = \mathbf{OR} \cdot o_2$$

- Then, it holds

$$\log(o_1) = \log(o_2) + \log(\mathbf{OR}),$$

1) Logistic regression

- As a measure of the risk, we can form **odds ratio**:

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$
$$o_1 = \mathbf{OR} \cdot o_2$$

- Then, it holds

$$\log(o_1) = \log(o_2) + \log(\mathbf{OR}),$$

- or

$$y = \log(o_2) + \log(\mathbf{OR}) \cdot x,$$

where $x \in \{0, 1\}$ and $y \in \{\log(o_1), \log(o_2)\}$.

1) Logistic regression

- This is the basic logistic model. Formally it is a regression model

$$y = \beta_0 + \beta_1 x$$

with baseline $\beta_0 = \log(o_2)$ and slope $\beta_1 = \log(\mathbf{OR})$ – effect of the exposure.

1) Logistic regression

- This is the basic logistic model. Formally it is a regression model

$$y = \beta_0 + \beta_1 x$$

with baseline $\beta_0 = \log(o_2)$ and slope $\beta_1 = \log(\mathbf{OR})$ – effect of the exposure.

- If we denote the probability of the event (outbreak of the disease) by p, then

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

in either group.

1) Logistic regression

- This is the basic logistic model. Formally it is a regression model

$$y = \beta_0 + \beta_1 x$$

with baseline $\beta_0 = \log(o_2)$ and slope $\beta_1 = \log(\mathbf{OR})$ – effect of the exposure.

- If we denote the probability of the event (outbreak of the disease) by p, then

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

in either group.

- It follows

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

1) Logistic regression

- Estimation is done via ML-method. Likelihood of observed frequencies for given β's is

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1 - p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1 - p_2)^{n_{20}-n_{21}}$$

1) Logistic regression

- Estimation is done via ML-method. Likelihood of observed frequencies for given β's is

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1-p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1-p_2)^{n_{20}-n_{21}}$$

- Then

$$\ell(\beta) = \log L(\beta) = n_{11} \log(p_1) + n_{12} \log(1-p_1) + \\ + n_{21} \log(p_2) + n_{22} \log(1-p_2)$$

1) Logistic regression

- Estimation is done via ML-method. Likelihood of observed frequencies for given β's is

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1-p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1-p_2)^{n_{20}-n_{21}}$$

- Then

$$\ell(\beta) = \log L(\beta) = n_{11} \log(p_1) + n_{12} \log(1-p_1) + \\ + n_{21} \log(p_2) + n_{22} \log(1-p_2)$$

- It is easy to derive that ML-equations are

$$\frac{\partial \ell}{\partial \beta_1} = n_{11} - n_{10} p_1 = 0\,, \qquad \frac{\partial \ell}{\partial \beta_0} = n_{11} - n_{10} p_1 + n_{21} - n_{20} p_2 = 0$$

1) Logistic regression

- In this case we already know the solution, $\hat{p}_1 = \dfrac{n_{11}}{n_{10}}$ and $\hat{p}_2 = \dfrac{n_{21}}{n_{20}}$ (and corresponding β's).

1) Logistic regression

- In this case we already know the solution, $\hat{p}_1 = \dfrac{n_{11}}{n_{10}}$ and $\hat{p}_2 = \dfrac{n_{21}}{n_{20}}$ (and corresponding β's).
- Asymptotic variance matrix of $\hat{\beta}$'s can also be established. In this case it is

$$\operatorname{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = J^{-1} = \frac{1}{ab}\begin{pmatrix} a & -a \\ -a & a+b \end{pmatrix},$$

where $a = n_{10}\hat{p}_1\left(1-\hat{p}_1\right)$, $b = n_{20}\hat{p}_2\left(1-\hat{p}_2\right)$.

1) Logistic regression

- In this case we already know the solution, $\hat{p}_1 = \dfrac{n_{11}}{n_{10}}$ and $\hat{p}_2 = \dfrac{n_{21}}{n_{20}}$ (and corresponding β's).
- Asymptotic variance matrix of $\hat{\beta}$'s can also be established. In this case it is

$$\operatorname{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = J^{-1} = \frac{1}{ab}\begin{pmatrix} a & -a \\ -a & a+b \end{pmatrix},$$

  where $a = n_{10}\hat{p}_1\,(1-\hat{p}_1)$, $b = n_{20}\hat{p}_2\,(1-\hat{p}_2)$.
- $J = \begin{pmatrix} a+b & a \\ a & a \end{pmatrix}$ is the information matrix.

1) Logistic regression

- **Example:** Baystate Medical Center in Springfield, MA, studied factors influencing low birth weights of babies. Let us take as a risk factor smoking during pregnancy. We get the following contingency table:

smoking	low birth weight		sum
	yes	no	
exposed	30	44	74
unexposed	29	86	115
sum	59	130	159

1) Logistic regression

- **Example:** Baystate Medical Center in Springfield, MA, studied factors influencing low birth weights of babies. Let us take as a risk factor smoking during pregnancy. We get the following contingency table:

smoking	low birth weight		sum
	yes	no	
exposed	30	44	74
unexposed	29	86	115
sum	59	130	159

- Odds for the unexposed is $o_1 = 30/44 = 0.681818$, for the exposed $o_2 = 29/86 = 0.337209$, **OR** $= 2.021644$. Sakoda coefficient $S = 0.225344$ indicates moderate association, but statistically significant ($\chi^2 = 4.92 > 3.84 = \chi^2_1(0.05)$).

1) Logistic regression

- As $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, we get the model

$$y = -1.087 + 0.704\,x\,.$$

1) Logistic regression

- As $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, we get the model

$$y = -1.087 + 0.704\,x.$$

- Software gets

	coefficient	*std. error*	*z*	$P > \|z\|$	*95% conf. interval*	
smoking	*0.7040592*	*0.319639*	*2.20*	*0.028*	*0.077579*	*1.330539*
constant	*-1.087051*	*0.21473*	*-5.06*	*0.000*	*-1.50791*	*-0.66619*

1) Logistic regression

- As $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, we get the model

$$y = -1.087 + 0.704\,x\,.$$

- Software gets

	coefficient	*std. error*	*z*	$P > \lvert z \rvert$	*95% conf. interval*	
smoking	*0.7040592*	*0.319639*	*2.20*	*0.028*	*0.077579*	*1.330539*
constant	*-1.087051*	*0.21473*	*-5.06*	*0.000*	*-1.50791*	*-0.66619*

- We can also get 95% CI for the **OR**:

$$\left[e^{0.077579}; e^{1.330539}\right] = [1.08;\ 3.78]$$

1) Logistic regression

General categorical predictor:

There are more probabilities to estimate. Let the predictor have m categories:

- $o_i = n_{i1}/n_{i2}\ ,\ i = 1, \ldots, m$

1) Logistic regression

General categorical predictor:

There are more probabilities to estimate. Let the predictor have m categories:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \dots, m$
- $\log\left(\mathbf{OR}_i\right) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \dots, m-1$

1) Logistic regression

General categorical predictor:

There are more probabilities to estimate. Let the predictor have m categories:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \ldots, m$
- $\log(\mathbf{OR}_i) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \ldots, m-1$

Then,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1}\,,$$

where all $x_i \in \{0, 1\}$, but only one of them can be 1 at a time.

1) Logistic regression

General categorical predictor:

There are more probabilities to estimate. Let the predictor have m categories:

- $o_i = n_{i1}/n_{i2}\ ,\ i = 1, \ldots, m$
- $\log(\mathbf{OR}_i) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \ldots, m-1$

Then,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1}\ ,$$

where all $x_i \in \{0, 1\}$, but only one of them can be 1 at a time.
One category out of m has to be a reference category. One explanatory variable is replaced by m˘1 indicator variables of other categories, which are mutually exclusive. Significance of a regression coefficient indicates significant difference between corresponding category and the reference category (differential effect of the category).

1) Logistic regression

Example: Let us take mother's weight (grouped in 3 categories) as a risk factor of child's low birth weight. We get

weight group	*low birth weight*		*sum*	*row percentages*	
	yes	*no*			
≤ 110	*25*	*28*	*53*	*47.2%*	*52.8%*
$(110; 150]$	*27*	*73*	*100*	*27.0%*	*73.0%*
> 150	*7*	*29*	*36*	*19.4%*	*80.6%*
sum	*59*	*130*	*189*	*31.2%*	*68.8%*

Changing row percentages show that mother's weight can be a risk factor.

1) Logistic regression

Software output for the 3rd category as the reference:

variable	coefficient	std. error	Wald	df	p-value
wt_groups			9.073891	2	0.010706
wt_groups(1)	1.308056995	0.503044916	6.761449	1	0.009315
wt_groups(2)	0.426763105	0.477572579	0.798537	1	0.371531
constant	-1.42138568	0.421117444	11.39246	1	0.000737

Software output for the 1st category as the reference:

variable	coefficient	std. error	Wald	df	p-value
wt_groups			9.073891	2	0.010706
wt_groups(2)	-0.88129389	0.355598021	6.142184	1	0.013199
wt_groups(3)	-1.308057	0.503044916	6.761449	1	0.009315
constant	-0.11332869	0.27516229	0.16963	1	0.680441

Symbolically, for the impact of weight groups holds $1 \neq \{2, 3\}$ at 5% level. Weight as a factor also is significant.

1) Logistic regression

Quantitative predictor:

If we have quantitative explanatory variable influencing the probability of the outcome, we simply assume that p is a (continuous) function of x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

1) Logistic regression

Quantitative predictor:

If we have quantitative explanatory variable influencing the probability of the outcome, we simply assume that p is a (continuous) function of x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

The regression coefficient of x is interpreted as effect of 1 unit change in x on the outcome.

1) Logistic regression

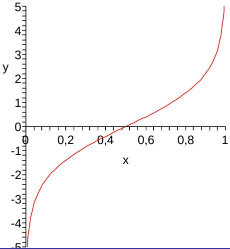
Quantitative predictor:

If we have quantitative explanatory variable influencing the probability of the outcome, we simply assume that p is a (continuous) function of x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

The regression coefficient of x is interpreted as effect of 1 unit change in x on the outcome.

Logistic transformation $p \to \log\left(\frac{p}{1-p}\right)$ is from $(0;1)$ to $(-\infty;+\infty)$, so that it removes restrictions harming regression.



1) Logistic regression

Example: Let us take mother's weight as a continuous risk factor of child's low birth weight.
Software gets:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
lwt	*-0.01405826*	*0.006169588*	*5.192193*	*1*	*0.022689*
constant	*0.998314313*	*0.78529092*	*1.616119*	*1*	*0.203634*

Mother's weight is again significant.

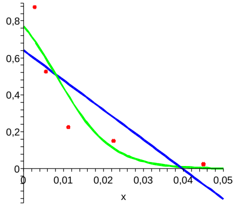
1) Logistic regression

Example: Let us take mother's weight as a continuous risk factor of child's low birth weight.
Software gets:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
lwt	*-0.01405826*	*0.006169588*	*5.192193*	*1*	*0.022689*
constant	*0.998314313*	*0.78529092*	*1.616119*	*1*	*0.203634*

Mother's weight is again significant.

Example: Effect of anti-pneumococcus serum on survival of ill mice was studied. Five different doses were administered to five groups of 40 mice. Plot shows the death rates, simple linear regression line, and logistic regression curve.



1) Logistic regression

General logistic regression model:

- We have dichotomic dependent variable Y and explanatory variables $X_1, X_2, \ldots, X_k$ of any type. We want to explain and/or predict the behaviour of Y using the explanatory variables.

1) Logistic regression

General logistic regression model:

- We have dichotomic dependent variable Y and explanatory variables $X_1, X_2, \dots, X_k$ of any type. We want to explain and/or predict the behaviour of Y using the explanatory variables.
- Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

or (taking $X_0 = \mathbf{1}$)

$$p_i = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} = \frac{1}{1 + \exp\left(-\sum_{j=0}^{k} \beta_j x_{ij}\right)}$$

1) Logistic regression

General logistic regression model:

- We have dichotomic dependent variable Y and explanatory variables $X_1, X_2, \dots, X_k$ of any type. We want to explain and/or predict the behaviour of Y using the explanatory variables.
- Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

or (taking $X_0 = \mathbf{1}$)

$$p_i = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} = \frac{1}{1 + \exp\left(-\sum_{j=0}^{k} \beta_j x_{ij}\right)}$$

- Estimation of parameters is done via ML-method

1) Logistic regression

- Likelihood of observed frequencies for given β's is

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

1) Logistic regression

- Likelihood of observed frequencies for given β's is

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

- Then

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

1) Logistic regression

- Likelihood of observed frequencies for given β's is

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

- Then

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

- The ML-equations are

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p_i(\beta)) x_{ij} = 0 \quad \forall j$$

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).
- Let us denote $X = \{x_{ij}\}$, $V = \text{diag}\,\{\hat{p}_1, \ldots, \hat{p}_n\}$. Information matrix is

$$J = X'VX$$

and J^{-1} is the asymptotic variance matrix of $\hat{\beta}$.

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).
- Let us denote $X = \{x_{ij}\}$, $V = \text{diag}\,\{\hat{p}_1, \ldots, \hat{p}_n\}$. Information matrix is
$$J = X'VX$$
and J^{-1} is the asymptotic variance matrix of $\hat{\beta}$.
- If all predictors are categorical ones, it is possible to reformulate the model in terms of binomial variables.

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).
- Let us denote $X = \{x_{ij}\}$, $V = \operatorname{diag}\{\hat{p}_1, \dots, \hat{p}_n\}$. Information matrix is

$$J = X'VX$$

and J^{-1} is the asymptotic variance matrix of $\hat{\beta}$.
- If all predictors are categorical ones, it is possible to reformulate the model in terms of binomial variables.
- Pros & cons:

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).
- Let us denote $X = \{x_{ij}\}$, $V = \text{diag}\,\{\hat{p}_1, \ldots, \hat{p}_n\}$. Information matrix is

$$J = X'VX$$

  and J^{-1} is the asymptotic variance matrix of $\hat{\beta}$.
- If all predictors are categorical ones, it is possible to reformulate the model in terms of binomial variables.
- Pros & cons:
  - no closed form formulas, iterative estimation

1) Logistic regression

- In general, these equations are solved numerically (Newton-Raphson type algorithm).
- Let us denote $X = \{x_{ij}\}$, $V = \text{diag}\,\{\hat{p}_1, \ldots, \hat{p}_n\}$. Information matrix is

$$J = X'VX$$

  and J^{-1} is the asymptotic variance matrix of $\hat{\beta}$.
- If all predictors are categorical ones, it is possible to reformulate the model in terms of binomial variables.
- Pros & cons:
  - – no closed form formulas, iterative estimation
  - + approximate variances, p-values and CI available

1) Logistic regression

- Wald test of a coefficient:
  If $H_0: \ \beta_i = 0$ holds, then

$$Z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

  has asymptotically $N(0;1)$.
  There is alternative chi-square form (Z^2).

1) Logistic regression

- Wald test of a coefficient:
  If $H_0: \ \beta_i = 0$ holds, then
  $$Z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$
  has asymptotically $N(0;1)$.
  There is alternative chi-square form (Z^2).
- Likelihood ratio test:
  Notions:
  - estimated model – model with predictors
  - empty model – model without predictors, just with intercept
  - full model – model predicting $n_{i0}\hat{p}_i = n_{i1}\ \forall i$, t.j. $\hat{y}_i = y_i\ \forall i$

1) Logistic regression

Deviance:

$$\hat{\ell}_m = \log \hat{L}_m = \log \hat{L}(\text{estimated model}),$$
$$\hat{\ell}_f = \log \hat{L}_f = \log \hat{L}(\text{full model})$$
$$D_m = 2\left(\hat{\ell}_f - \hat{\ell}_m\right) = 2\log\left(\hat{L}_f/\hat{L}_m\right)$$

For binomial data, D_m is a measure of goodness-of-fit of the model. Asymptotically,

$$D_m \approx \chi^2_{n-k-1}$$

1) Logistic regression

Deviance:

$$\hat{\ell}_m = \log \hat{L}_m = \log \hat{L}(\text{estimated model}),$$
$$\hat{\ell}_f = \log \hat{L}_f = \log \hat{L}(\text{full model})$$
$$D_m = 2\left(\hat{\ell}_f - \hat{\ell}_m\right) = 2\log\left(\hat{L}_f/\hat{L}_m\right)$$

For binomial data, D_m is a measure of goodness-of-fit of the model. Asymptotically,

$$D_m \approx \chi^2_{n-k-1}$$

For all kinds of models, deviance difference is used for comparison of nested models (LRT of significance of added predictors):

$$D_{m_1} - D_{m_2} = 2\left(\hat{\ell}_{m_2} - \hat{\ell}_{m_1}\right) \approx \chi^2_{k_2-k_1}$$

1) Logistic regression

Example: Let us consider low birth weight data with risk factor smoking. Software tell us that

$$\hat{\ell}_1 = \log \hat{L}_1 = \log \hat{L}(\text{estimated model}) = -114.9023,$$
$$\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{empty model}) = -117.336.$$

If $H_0 : \ \beta_1 = 0$ holds, then

$$2\left(\hat{\ell}_1 - \hat{\ell}_0\right) = 2\log\left(\frac{\hat{L}_1}{\hat{L}_0}\right) \approx \chi^2_1$$

We have

$$\Delta D = 2(-114.9023 + 117.336) \ = \ 4.8674 > 3.84 = \chi^2_1(0.05),$$

so that the association between smoking and low birth weight is significant.

1) Logistic regression

Interactions

Logistic regression model allows to include (and test) interactions of categorical variables. If variable X has c categories and variable Z d categories, then the interaction $X * Z$ has $(c-1)(d-1)$ categories. They are all possible combinations of non-reference categories of X and Z.

1) Logistic regression

Interactions

Logistic regression model allows to include (and test) interactions of categorical variables. If variable X has c categories and variable Z d categories, then the interaction $X * Z$ has $(c-1)(d-1)$ categories. They are all possible combinations of non-reference categories of X and Z.

Interactions are tested via Wald test or LRT as any other variable.

1) Logistic regression

Interactions

Logistic regression model allows to include (and test) interactions of categorical variables. If variable X has c categories and variable Z d categories, then the interaction $X * Z$ has $(c-1)(d-1)$ categories. They are all possible combinations of non-reference categories of X and Z.

Interactions are tested via Wald test or LRT as any other variable.

If an interaction is significant, the significance of the original constituent variables has no interpretation, nor meaning. The effects are crossed, and we have only two ways to handle the situation:

1. Perform stratification and do separate analyses in different strata.
2. Introduce a new variable, which operates on the cross-product of the crossed variables (the interaction variable), and omit the interacting variables.

1) Logistic regression

Measures of goodness-of-fit

Let $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{empty model})$.

- McFadden $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$

1) Logistic regression

Measures of goodness-of-fit

Let $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{empty model})$.

- McFadden $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$
- Cox & Snell $R^2_{CS} = 1 - \left(\dfrac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$

1) Logistic regression

Measures of goodness-of-fit

Let $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{empty model})$.

- McFadden $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$
- Cox & Snell $R^2_{CS} = 1 - \left(\dfrac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$
- Nagelkerke $R^2_N = \dfrac{R^2_{CS}}{1 - \hat{L}_0^{\frac{2}{n}}}$

1) Logistic regression

Measures of goodness-of-fit

Let $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{empty model})$.

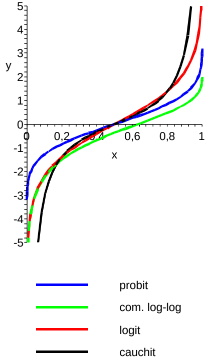
- McFadden $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$
- Cox & Snell $R^2_{CS} = 1 - \left(\dfrac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$
- Nagelkerke $R^2_N = \dfrac{R^2_{CS}}{1 - \hat{L}_0^{\frac{2}{n}}}$
- Hosmer-Lemeshow test (chi-square test of goodness-of-fit in contingency table between the outcome variable and groups of predicted values)

1) Logistic regression

Alternatives

Logistic function is not the only one used for transformation of probabilities of binary outcomes. Most used are:

- logistic function $\log\left(\frac{p}{1-p}\right)$
- probit function $\Phi^{-1}(p)$
- complementary log-log function $\log(-\log(1-p))$
- negative log-log function $-\log(-\log(p))$
- cauchit function $\tan\left(\left(p-\frac{1}{2}\right)\pi\right)$



They are called link functions.

2) Multinomial regression

- **Example:** From 1991 U.S. General Social Survey data, we want to check whether sex of a respondent influences the probabilities of life satisfaction feelings.

  We get the following contingency table:

sex	life satisfaction			sum
	exiting	routine	dull	
male	213	200	12	425
female	221	305	29	555
sum	434	505	41	980

2) Multinomial regression

- **Example:** From 1991 U.S. General Social Survey data, we want to check whether sex of a respondent influences the probabilities of life satisfaction feelings.

  We get the following contingency table:

sex	life satisfaction			sum
	exiting	routine	dull	
male	213	200	12	425
female	221	305	29	555
sum	434	505	41	980

- Sometimes it may have sense to solve such a problem by a series of binary models:

2) Multinomial regression

- **Example:** From 1991 U.S. General Social Survey data, we want to check whether sex of a respondent influences the probabilities of life satisfaction feelings.

  We get the following contingency table:

sex	life satisfaction			sum
	exiting	routine	dull	
male	213	200	12	425
female	221	305	29	555
sum	434	505	41	980

- Sometimes it may have sense to solve such a problem by a series of binary models:
  1. life is exiting – not exiting
  2. not exiting life: routine – dull

2) Multinomial regression

- Or, we have to consider more probabilities and more odds. In our example, we have to consider two multinomial distributions (p_{11}, p_{12}, p_{13}) and (p_{21}, p_{22}, p_{23}), describing the probabilities of life satisfaction feelings for males and females, respectively. The simplest way is to choose one response category as a reference – say exciting life – because one of the probabilities in each row is redundant.

2) Multinomial regression

- Or, we have to consider more probabilities and more odds. In our example, we have to consider two multinomial distributions (p_{11}, p_{12}, p_{13}) and (p_{21}, p_{22}, p_{23}), describing the probabilities of life satisfaction feelings for males and females, respectively. The simplest way is to choose one response category as a reference – say exciting life – because one of the probabilities in each row is redundant.
- Then, the model is

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) = \beta_{0j} + \beta_{1j}x_i\,, \quad j = 2, 3,$$

where $x_i \in \{0, 1\}$ is the indicator of sex.

2) Multinomial regression

- Or, we have to consider more probabilities and more odds. In our example, we have to consider two multinomial distributions (p_{11}, p_{12}, p_{13}) and (p_{21}, p_{22}, p_{23}), describing the probabilities of life satisfaction feelings for males and females, respectively. The simplest way is to choose one response category as a reference – say exciting life – because one of the probabilities in each row is redundant.
- Then, the model is

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) = \beta_{0j} + \beta_{1j}x_i\,, \quad j = 2, 3,$$

where $x_i \in \{0, 1\}$ is the indicator of sex.
- The previous formula coincides with the simple linear logistic model in the case of dichotomic outcome.

2) Multinomial regression

Based on our frequencies, we get the following odds and log-odds:

odds		*log-odds*	
$200/213 = 0,938967$	$12/213 = 0,056338$	*-0,06297*	*-2,87639*
$305/221 = 1,38009$	$29/221 = 0,131222$	*0,322149*	*-2,03087*

2) Multinomial regression

Based on our frequencies, we get the following odds and log-odds:

odds		*log-odds*	
$200/213 = 0,938967$	$12/213 = 0,056338$	*-0,06297*	*-2,87639*
$305/221 = 1,38009$	$29/221 = 0,131222$	*0,322149*	*-2,03087*

The differences of the log-odds in the last two columns are $-0,385123874$ and $-0,845518644$. Thus, we can write two models

$$y = 0.322 - 0.385x\,, \quad y = -2.031 - 0.846x$$

for routine and dull feeling, respectively.

2) Multinomial regression

Based on our frequencies, we get the following odds and log-odds:

odds		*log-odds*	
$200/213 = 0,938967$	$12/213 = 0,056338$	*-0,06297*	*-2,87639*
$305/221 = 1,38009$	$29/221 = 0,131222$	*0,322149*	*-2,03087*

The differences of the log-odds in the last two columns are $-0,385123874$ and $-0,845518644$. Thus, we can write two models

$$y = 0.322 - 0.385x, \quad y = -2.031 - 0.846x$$

for routine and dull feeling, respectively. Software gets:

Life		B	Std. Error	Wald	df	Sig.	Exp(B)	95% CI	
Routine	Intercept	0,322149	0,088338	13,29904	1	0,000266			
	[sex=1]	-0,38512	0,132282	8,476221	1	0,003598	0,680366	0,524982	0,881741
	[sex=2]	0			0				
Dull	Intercept	-2,03087	0,197504	105,7336	1	0			
	[sex=1]	-0,84552	0,356421	5,627561	1	0,01768	0,429335	0,213506	0,86334
	[sex=2]	0			0				

2) Multinomial regression

General model

Let

- the response variable Y have r categories, and $X_1, \ldots, X_k$ be explanatory variables;

2) Multinomial regression

General model

Let

- the response variable Y have r categories, and $X_1, \ldots, X_k$ be explanatory variables;
- $y_i = (y_{i1}, \ldots, y_{ir})$ be the response values in i-th subgroup having multinomial distribution $Mn\,(n_i, p_{i1}, \ldots, p_{ir})$;

2) Multinomial regression

General model

Let

- the response variable Y have r categories, and $X_1, \ldots, X_k$ be explanatory variables;
- $y_i = (y_{i1}, \ldots, y_{ir})$ be the response values in i-th subgroup having multinomial distribution $Mn\,(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ be regression coefficients for the j-th response category with respect to the j^*-th (reference) one;

2) Multinomial regression

General model

Let

- the response variable Y have r categories, and $X_1, \ldots, X_k$ be explanatory variables;
- $y_i = (y_{i1}, \ldots, y_{ir})$ be the response values in i-th subgroup having multinomial distribution $Mn\,(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ be regression coefficients for the j-th response category with respect to the j^*-th (reference) one;
- $x_i = (1, x_{i1}, \ldots, x_{ik})'$ be actual values of explanatory variables for the i-th subgroup.

2) Multinomial regression

General model

Let

- the response variable Y have r categories, and $X_1, \ldots, X_k$ be explanatory variables;
- $y_i = (y_{i1}, \ldots, y_{ir})$ be the response values in i-th subgroup having multinomial distribution $Mn\,(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ be regression coefficients for the j-th response category with respect to the j^*-th (reference) one;
- $x_i = (1, x_{i1}, \ldots, x_{ik})'$ be actual values of explanatory variables for the i-th subgroup.

Then, general multinomial regression model is

$$\log\left(\frac{p_{ij}}{p_{ij^*}}\right) = x_i'\beta_j\,, \quad j \neq j^*$$

2) Multinomial regression

- Inverse formulas are

$$p_{ij} = \frac{\exp\left(x_i'\beta_j\right)}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}, \; j \neq j^*$$

and

$$p_{ij^*} = \frac{1}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}.$$

2) Multinomial regression

- Inverse formulas are

$$p_{ij} = \frac{\exp\left(x_i'\beta_j\right)}{1 + \sum\limits_{\substack{k=1\\k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}, \; j \neq j^*$$

and

$$p_{ij^*} = \frac{1}{1 + \sum\limits_{\substack{k=1\\k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}.$$

- Estimation is done via ML-method.

2) Multinomial regression

- The log-likelihood function is

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij}!}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log\left(p_{ij}\right)$$

2) Multinomial regression

- The log-likelihood function is

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij}!}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log(p_{ij})$$

- The ML-equations are

$$\frac{\partial \ell}{\partial \beta_{mj}} = \sum_{i=1}^{n} x_{im}(y_{ij} - n_i p_{ij}) = 0, \quad j \neq j^*,\ m = 0, \ldots, k$$

2) Multinomial regression

- The log-likelihood function is

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij}!}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log\left(p_{ij}\right)$$

- The ML-equations are

$$\frac{\partial \ell}{\partial \beta_{mj}} = \sum_{i=1}^{n} x_{im}\left(y_{ij} - n_i p_{ij}\right) = 0, \quad j \neq j^*,\ m = 0, \ldots, k$$

- Hessian matrix of the estimates of $\beta = \left(\beta_j{}', j = 1, \ldots, r, j \neq j^*\right)'$ is

$$H = -\sum_{i=1}^{n}\left(I_{r-1} \otimes x_i\right)\hat{V}_i^*\left(I_{r-1} \otimes x_i\right)',$$

where $\hat{V}_i^* = n_i\left(\mathrm{diag}\left(\hat{p}_i^*\right) - \hat{p}_i^*\hat{p}_i^{*\prime}\right)$ and $\hat{p}_i^*$ is the vector of all estimates of probabilities p_{ij} except p_{ij^*}

2) Multinomial regression

- Chi=square of the estimated model is

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$$

Since total number of non-redundant parameters in the model is (r-1)(k+1), it holds $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.

2) Multinomial regression

- Chi=square of the estimated model is

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{\left(y_{ij} - n_i \hat{p}_{ij}\right)^2}{n_i \hat{p}_{ij}}$$

  Since total number of non-redundant parameters in the model is (r-1)(k+1), it holds $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.

- Deviance of the estimated model is

$$D_m = 2\left(\ell_f - \ell_m\right) = \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \log\left(\frac{y_{ij}}{n_i \hat{p}_{ij}}\right)$$

  Here also $D_m \approx \chi^2_{(n-k-1)(r-1)}$.

2) Multinomial regression

- Chi=square of the estimated model is

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$$

  Since total number of non-redundant parameters in the model is (r-1)(k+1), it holds $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.
- Deviance of the estimated model is

$$D_m = 2(\ell_f - \ell_m) = \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \log\left(\frac{y_{ij}}{n_i \hat{p}_{ij}}\right)$$

  Here also $D_m \approx \chi^2_{(n-k-1)(r-1)}$.
- The same pseudo-R^2 statistics as in logistic regression model can be used.

2) Multinomial regression

If actual variance matrix of y_i are substantially larger than $V_i = n_i\,(\mathrm{diag}\,(p_i) - p_i p_i')$ (given by the multinomial model), we speak about **overdispersion**. Then, we introduce scale parameter σ^2, such that $\mathrm{var}\, y_i = \sigma^2 V_i$.

2) Multinomial regression

If actual variance matrix of y_i are substantially larger than $V_i = n_i\,(\mathrm{diag}\,(p_i) - p_i p_i')$ (given by the multinomial model), we speak about **overdispersion**. Then, we introduce scale parameter σ^2, such that $\mathrm{var}\, y_i = \sigma^2 V_i$.

- Usual (asymptotically unbiased) estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

(or D_m instead of χ^2).

2) Multinomial regression

If actual variance matrix of y_i are substantially larger than $V_i = n_i\left(\operatorname{diag}(p_i) - p_i p_i'\right)$ (given by the multinomial model), we speak about **overdispersion**. Then, we introduce scale parameter σ^2, such that $\operatorname{var} y_i = \sigma^2 V_i$.

- Usual (asymptotically unbiased) estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

  (or D_m instead of χ^2).
- Use of σ^2 does not change estimators of β.

2) Multinomial regression

If actual variance matrix of y_i are substantially larger than $V_i = n_i \left(\operatorname{diag}(p_i) - p_i p_i'\right)$ (given by the multinomial model), we speak about **overdispersion**. Then, we introduce scale parameter σ^2, such that $\operatorname{var} y_i = \sigma^2 V_i$.

- Usual (asymptotically unbiased) estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

  (or D_m instead of χ^2).
- Use of σ^2 does not change estimators of β.
- Variance of the estimators is

$$\operatorname{var} \hat{\beta} = \hat{\sigma}^2 \left[\sum_{i=1}^{n} \left(I_{r-1} \otimes x_i\right) \hat{V}_i^* \left(I_{r-1} \otimes x_i\right)' \right]^{-1}$$

2) Multinomial regression

Tests

- For any $L_{q \times k+1}$ of full rank, it holds

$$\hat{\beta}_j' L' \left[L \operatorname{var} \hat{\beta}_j L' \right]^{-1} L \hat{\beta}_j \approx \chi_q^2$$

under $H_0 : L\beta_j = 0$. This allows testing of separate regression coefficients and/or their linear combinations.

2) Multinomial regression

Tests

- For any $L_{q\times k+1}$ of full rank, it holds

$$\hat{\beta}_j' L' \left[L \operatorname{var} \hat{\beta}_j L' \right]^{-1} L \hat{\beta}_j \approx \chi^2_q$$

  under $H_0 : L\beta_j = 0$. This allows testing of separate regression coefficients and/or their linear combinations.

- LR test of nested models with k_1 and k_2 ($k_1 < k_2$) regression parameters is based on

$$\frac{1}{\hat{\sigma}^2}\left(D_{m_1} - D_{m_2}\right) = \frac{2}{\hat{\sigma}^2}\left(\hat{\ell}_{m_2} - \hat{\ell}_{m_1}\right) \approx \chi^2_{k_2-k_1}$$

2) Multinomial regression

Example: We have used sex as a predictor in our life satisfaction study model. This first model has 4 non-redundant parameters (intercept and one non-reference sex category for each of 2 non-reference life feelings).

2) Multinomial regression

Example: We have used sex as a predictor in our life satisfaction study model. This first model has 4 non-redundant parameters (intercept and one non-reference sex category for each of 2 non-reference life feelings).
If we add race (white, black, other) as another possible predictor, we have 4 more non-redundant parameters – 2 non-reference races within 2 non-reference life feelings.

2) Multinomial regression

Example: We have used sex as a predictor in our life satisfaction study model. This first model has 4 non-redundant parameters (intercept and one non-reference sex category for each of 2 non-reference life feelings).
If we add race (white, black, other) as another possible predictor, we have 4 more non-redundant parameters – 2 non-reference races within 2 non-reference life feelings.
Overdispersion is not observed. Software gets:

$$\hat{\ell}_{m_1} = -25,8165, \quad \hat{\ell}_{m_2} = -24.332$$

Therefore,

$$\Delta D_m = 2(-24.332 + 25,8165) = 2.969 < 9.488 = \chi^2_{8-4}(0.05)$$

Corresponding p-value is 0.563. The race factor proves to be non-significant.

3) Ordinal regression

Example: Random sample of Vermont citizens was asked to rate the work of criminal judges in the state. The scale was Poor (1), Only fair (2), Good (3), and Excellent (4). At the same time, they had to report whether somebody of their household had been a crime victim within the last 3 years.

3) Ordinal regression

Example: Random sample of Vermont citizens was asked to rate the work of criminal judges in the state. The scale was Poor (1), Only fair (2), Good (3), and Excellent (4). At the same time, they had to report whether somebody of their household had been a crime victim within the last 3 years.

The question was, whether people with personal experience with crime and people without it share the same view of criminal justice performance.

3) Ordinal regression

Example: Random sample of Vermont citizens was asked to rate the work of criminal judges in the state. The scale was Poor (1), Only fair (2), Good (3), and Excellent (4). At the same time, they had to report whether somebody of their household had been a crime victim within the last 3 years.

The question was, whether people with personal experience with crime and people without it share the same view of criminal justice performance.

The data:

Household victim	Judges' performance				sum
	Poor	Only fair	Good	Excellent	
Yes	14	28	31	3	76
No	38	170	248	34	490
sum	52	198	279	37	566

3) Ordinal regression

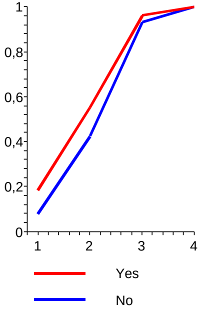
With ordinal data, it is natural to consider probabilities of cumulative events, like specific score or worse. Table of cumulative frequencies is as follows:

Household victim	Judges' performance			
	Poor	Only fair or worse	Good or worse	Excellent or worse
Yes row percentage	14 18,42%	42 55,26%	73 96,05%	76 100,00%
No row percentage	38 7,76%	208 42,45%	456 93,06%	490 100,00%

3) Ordinal regression

With ordinal data, it is natural to consider probabilities of cumulative events, like specific score or worse. Table of cumulative frequencies is as follows:

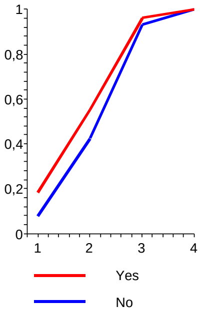
Household victim	Judges' performance			
	Poor	Only fair or worse	Good or worse	Excellent or worse
Yes row percentage	14 18,42%	42 55,26%	73 96,05%	76 100,00%
No row percentage	38 7,76%	208 42,45%	456 93,06%	490 100,00%



3) Ordinal regression

With ordinal data, it is natural to consider probabilities of cumulative events, like specific score or worse. Table of cumulative frequencies is as follows:

Household victim	Judges' performance			
	Poor	Only fair or worse	Good or worse	Excellent or worse
Yes row percentage	14 18,42%	42 55,26%	73 96,05%	76 100,00%
No row percentage	38 7,76%	208 42,45%	456 93,06%	490 100,00%



The graph suggests that having a crime victim in the household implies more negative opinion on judges' performance.
The lines must meet at 100%. Otherwise they look almost parallel. That suggest model with common slope for both categories.

3) Ordinal regression

Let us denote $p_{ij}^c = P(\text{score} \leq j)$, $i = 1(\text{No}), 2(\text{Yes})$, $j = 1, 2, 3$ the non-trivial cumulative probabilities. Then, our model is

$$\log\left(\frac{p_{1j}^c}{1 - p_{1j}^c}\right) = \alpha_j \quad \text{and} \quad \log\left(\frac{p_{2j}^c}{1 - p_{2j}^c}\right) = \alpha_j + \beta\,,$$

3) Ordinal regression

Let us denote $p^c_{ij} = P(\text{score} \le j)$, $i = 1(\text{No}), 2(\text{Yes})$, $j = 1, 2, 3$ the non-trivial cumulative probabilities. Then, our model is

$$\log\left(\frac{p^c_{1j}}{1 - p^c_{1j}}\right) = \alpha_j \quad \text{and} \quad \log\left(\frac{p^c_{2j}}{1 - p^c_{2j}}\right) = \alpha_j + \beta\,,$$

or

$$\log\left(\frac{p^c_j(x)}{1 - p^c_j(x)}\right) = \alpha_j + \beta x \quad \forall j\,, x \in \{0, 1\}.$$

3) Ordinal regression

Let us denote $p_{ij}^c = P(\text{score} \leq j)$, $i = 1(\text{No}), 2(\text{Yes})$, $j = 1, 2, 3$ the non-trivial cumulative probabilities. Then, our model is

$$\log\left(\frac{p_{1j}^c}{1 - p_{1j}^c}\right) = \alpha_j \quad \text{and} \quad \log\left(\frac{p_{2j}^c}{1 - p_{2j}^c}\right) = \alpha_j + \beta\,,$$

or

$$\log\left(\frac{p_j^c(x)}{1 - p_j^c(x)}\right) = \alpha_j + \beta x \quad \forall j\,, x \in \{0, 1\}.$$

Software gets $\alpha_1 = -2.39$, $\alpha_2 = -0.32$, $\alpha_2 = 2.59$, $\beta = 0.63$. Using standard inverse formula for logits, we obtain the following estimates:

	≤ 1	≤ 2	≤ 3
Yes	14,69%	57,85%	96,18%
No	8,38%	42,15%	93,04%

3) Ordinal regression

Proportional odds model

Let

- Y be ordinal response variable with possible values $1, \ldots, r$

3) Ordinal regression

Proportional odds model

Let

- Y be ordinal response variable with possible values $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ be independent predictor variables

3) Ordinal regression

Proportional odds model

Let

- Y be ordinal response variable with possible values $1, \dots, r$
- $X = (X_1, \dots, X_k)$ be independent predictor variables
- $\alpha_1, \dots, \alpha_{r-1}$ and $\beta = (\beta_1, \dots, \beta_k)$ be unknown regression coefficients

3) Ordinal regression

Proportional odds model

Let

- Y be ordinal response variable with possible values $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ be independent predictor variables
- $\alpha_1, \ldots, \alpha_{r-1}$ and $\beta = (\beta_1, \ldots, \beta_k)$ be unknown regression coefficients

Model: logits of cumulative probabilities $p_j^c(x) = P\left(Y \leq j | X = x\right)$ satisfy

$$\log\left(\frac{p_j^c(x)}{1 - p_j^c(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \ldots, r-1$$

3) Ordinal regression

Proportional odds model

Let

- Y be ordinal response variable with possible values $1, \dots, r$
- $X = (X_1, \dots, X_k)$ be independent predictor variables
- $\alpha_1, \dots, \alpha_{r-1}$ and $\beta = (\beta_1, \dots, \beta_k)$ be unknown regression coefficients

Model: logits of cumulative probabilities $p_j^c(x) = P(Y \leq j | X = x)$ satisfy

$$\log\left(\frac{p_j^c(x)}{1 - p_j^c(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \dots, r-1$$

Because log of cumulative odds ratio of making the same responses at different x-points is proportional to the distance of the points, the model is called proportional odds model:

$$\log\left(\frac{p_j^c(x_1)}{1 - p_j^c(x_1)} \cdot \frac{1 - p_j^c(x_2)}{p_j^c(x_2)}\right) = \beta' (x_1 - x_2)$$

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:
  1. at first estimating $p_j^c(x)$, and then calculating $p_j(x) = p_j^c(x) - p_{j-1}^c(x) \quad$ (taking $p_0(x) \equiv 0$)

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:
  1. at first estimating $p_j^c(x)$, and then calculating $p_j(x) = p_j^c(x) - p_{j-1}^c(x) \quad$ (taking $p_0(x) \equiv 0$)
  2. estimating directly $p_j(x)$

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:
  1. at first estimating $p_j^c(x)$, and then calculating $p_j(x) = p_j^c(x) - p_{j-1}^c(x) \quad$ (taking $p_0(x) \equiv 0$)
  2. estimating directly $p_j(x)$
- Likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[p_j(x_i) - p_{j-1}(x_i)\right]^{y_{ij}}$$

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:
  1. at first estimating $p_j^c(x)$, and then calculating $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (taking $p_0(x) \equiv 0$)
  2. estimating directly $p_j(x)$
- Likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} [p_j(x_i) - p_{j-1}(x_i)]^{y_{ij}}$$

- All standard goodness-of-fit measures apply.

3) Ordinal regression

- Estimation is done again via ML-method. There are two options:
  1. at first estimating $p_j^c(x)$, and then calculating $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (taking $p_0(x) \equiv 0$)
  2. estimating directly $p_j(x)$
- Likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[p_j(x_i) - p_{j-1}(x_i)\right]^{y_{ij}}$$

- All standard goodness-of-fit measures apply.
- The fit is different than separate logit models for all j's.

3) Ordinal regression

Example: Software output for Vermont crime data:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,39221	0,15177	248,44332	1	0,00000	-2,68968	-2,09475
	[rating = 2]	-0,31651	0,09082	12,14637	1	0,00049	-0,49451	-0,13852
	[rating = 3]	2,59316	0,17163	228,28667	1	0,00000	2,25678	2,92955
Location	[hhcrime=1]	-0,63298	0,23198	7,44539	1	0,00636	-1,08765	-0,17831
	[hhcrime=2]	0	.	.	0	.	.	.

Notice opposite sign of the coefficient β (hhcrime=1). Many work with the model $\alpha_j - \beta x$ because of interpretation reasons: in such a case, higher coefficients indicate association with higher scores.

3) Ordinal regression

Example: Software output for Vermont crime data:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,39221	0,15177	248,44332	1	0,00000	-2,68968	-2,09475
	[rating = 2]	-0,31651	0,09082	12,14637	1	0,00049	-0,49451	-0,13852
	[rating = 3]	2,59316	0,17163	228,28667	1	0,00000	2,25678	2,92955
Location	[hhcrime=1]	-0,63298	0,23198	7,44539	1	0,00636	-1,08765	-0,17831
	[hhcrime=2]	0	.	.	0	.	.	.

Notice opposite sign of the coefficient β (hhcrime=1). Many work with the model $\alpha_j - \beta x$ because of interpretation reasons: in such a case, higher coefficients indicate association with higher scores.

Let us now add another predictor variable, sex:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,57419	0,17641	212,93519	1	0,00000	-2,91995	-2,22844
	[rating = 2]	-0,48730	0,12326	15,62868	1	0,00008	-0,72890	-0,24571
	[rating = 3]	2,43740	0,18672	170,40298	1	0,00000	2,07143	2,80336
Location	[hhcrime=1]	-0,62074	0,23228	7,14177	1	0,00753	-1,07599	-0,16548
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,34145	0,16030	4,53709	1	0,03317	-0,65563	-0,02726
	[sex=2]	0	.	.	0	.	.	.

3) Ordinal regression

We suspect that sex may influence sensitivity to crime victims, so that we add the interaction:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[sex=2]	0	.	.	0	.	.	.
	[hhcrime=1] * [sex=1]	0,95889	0,46413	4,26832	1	0,03883	0,04921	1,86857
	[hhcrime=1] * [sex=2]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=1]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

3) Ordinal regression

We suspect that sex may influence sensitivity to crime victims, so that we add the interaction:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[sex=2]	0	.	.	0	.	.	.
	[hhcrime=1] * [sex=1]	0,95889	0,46413	4,26832	1	0,03883	0,04921	1,86857
	[hhcrime=1] * [sex=2]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=1]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

But, since the interaction is significant, the two individual variables don't have good meaning any more:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1] * [sex=1]	-0,64690	0,32950	3,85460	1	0,04961	-1,29270	-0,00110
	[hhcrime=1] * [sex=2]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2] * [sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

Redundant parameters are not estimated, so that interaction itself is enough. This model has the same χ^2, deviance, and pseudo-R^2 as the previous one.

3) Ordinal regression

Other ordinal regression models

- General cumulative logit model is

$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta_j' x \quad \forall j = 1,\ldots,r-1$$

  Thus, every group has its own slope. Proportional odds model is a special case, and can be tested by LR test.

3) Ordinal regression

Other ordinal regression models

- General cumulative logit model is
$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta_j' x \quad \forall j = 1, \dots, r-1$$
Thus, every group has its own slope. Proportional odds model is a special case, and can be tested by LR test.
- Adjacent categories model is
$$\log\left(\frac{p_j(x)}{p_{j+1}(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \dots, r-1$$
This model recognizes the ordering, since
$$\log\left(\frac{p_j(x)}{p_r(x)}\right) = \sum_{m=j}^{r} \alpha_m + \beta'(r-j)x \quad \forall j$$

The suffering is over...

Thank you for your attention!

Literature

- D. Collett: Modeling binary data, 2nd ed., Chapman & Hall CRC, 2003
- A. Agresti: Categorical data analysis, 2nd ed., Wiley, 2002
- D.W. Hosmer, S. Lemeshow: Applied logistic regression, 2nd ed., Wiley, 2000
- M. Norusis: SPSS Advanced Statistical Procedures Companion, Prentice Hall, 2005
- J.L. Schafer: Lecture notes for Categorical data analysis course, Pensylvania State University, 2006
- D.G. Garson: Statnotes - Topics in Multivariate Analysis, North Carolina State University, 2009
- SPSS documentation