Understanding Statistical Concepts in Clinical Research

Other types of regression

Nick Counsell, Medical Statistician
Cancer Research UK & UCL Cancer Trials Centre

Introduction

- Have seen outcome measures based on:
  - Taking measurements on people (continuous data); linear regression
  - Counting people (yes/no); logistic regression
  - Time-to-event data; Cox regression

- There are other types of regression:
  - Counting the number of events that have occurred (i.e. per person or object)

    eg. number of days spent in hospital (if very skewed)
  - The outcome has ≥3 categories that have a natural order to them (called ordered categories)

    eg. disease severity (mild, moderate, severe); agreement (strongly agree, agree, neutral, disagree, strongly disagree)

Ordered categorical data

- For ordered categories the researcher often chooses to divide the variable into two groups and apply binary logistic regression
- Although not incorrect, this method does not utilise all of the available information within the outcome data
- Ordinal logistic regression is an extension of binary logistic regression which is appropriate for ordered categorical outcome variables
- The model is based on the notion that there is some underlying quantitative scale

Logistic regression (reminder)

- In the binary case, we look at the probability p of a given response. This can only lie between 0 and 1 so we use the logit of p instead, i.e. log(p/1-p), when deriving the regression (because this then looks similar to linear regression):

$$\text{logit}(p) = \log(p/1-p) = a + bX_1 + cX_2 + dX_3 + \ldots.$$

- b can be interpreted as increasing/decreasing the log-odds of an event, and exp(b) is used as the odds ratio for a unit increase/decrease in factor b
- For ordered categories we can look at the probability p_j of having a response less than or equal to a given group, or a higher response (i.e.1-p_j)

Ordinal logistic regression

- The model is similar to binary logistic regression, but the probabilities represent categories of a response:

$$\log\left[\frac{p_j}{1-p_j}\right] = a_j + bX_1 + cX_2 + dX_3 + \ldots.$$

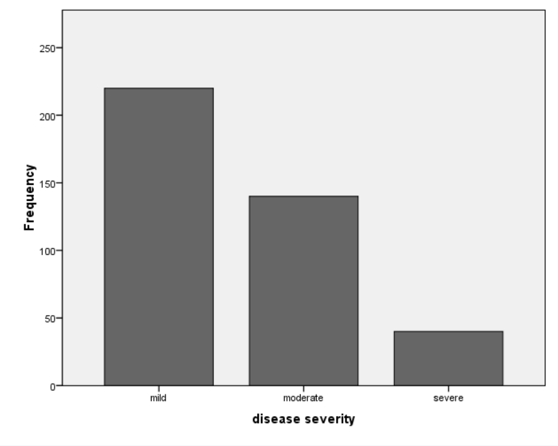
- The thresholds (a_j's) correspond to the intercept in simpler models, these depend only on which category's probability is being predicted
- The prediction part of the model depends only on the factors and is independent of the outcome category

Choosing regression covariates

- The process of choosing variables for the model is similar to the process of selecting them in other types of regression models
- Both theoretical and empirical considerations should be taken into account when selecting variables to be included
- Individual continuous or binary variables can be assessed through the use of Wald tests
- Likelihoods can be used to compare nested models for categorical variables with ≥3 levels

Example of ordinal data

- Collect data on the severity of 400 patients with depression



	Frequency	Percent
mild (=0)	220	55.0
moderate (=1)	140	35.0
severe (=2)	40	10.0
Total	400	100.0

We can describe the data using contingency tables and bar charts

---

- Information is also collected on whether each patient has a history of depression (binary – yes/no) and their score from a baseline questionnaire (continuous scale)

- We aim to be able to see the relationship between the clinical severity of depression, medical history and the questionnaire score

- i.e. disease severity is the outcome variable (coded '0'/'1'/'2' for the three levels), medical history and questionnaire score are factors

- logit(p_j) = a_j + (b x Score)
- logit(p_j) = a_j + (b x History)

Interpreting a continuous covariate (eg. score)

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Threshold	[severity = 0]	4.404	1.089	16.345	1	.000	2.269	6.539
	[severity = 1]	8.105	1.301	38.819	1	.000	5.555	10.655
Location	score	.928	.350	7.040	1	.008	.242	1.613

- 'Threshold' is analogous to the intercept terms for linear, logistic and Cox regression models
- The estimate is the odds ratio for the outcome on the natural log-scale, the odds ratio for 'score' is exp(0.928) = 2.53
- For a one unit increase in questionnaire score, the odds of being in a higher severity group increases by 2.53 times. That is:
  - The odds of being either moderate or severe compared to mild are 2.53 times greater
  - The odds of being severe compared to either mild or moderate are 2.53 times greater

Interpreting a binary covariate (eg. history)

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Threshold	[severity = 0]	2.176	.773	7.935	1	.005	.662	3.690
	[severity = 1]	4.272	.798	28.683	1	.000	2.708	5.835
Location	[history = 1]	1.046	.268	15.200	1	.000	.520	1.571
	[history = 0]	0[a]	.	.	0	.	.	.

- The odds ratio for 'history' is exp(1.046) = 2.85
- The odds of being in a higher severity group increases by 2.85 times among those with a history of depression compared to those without. That is:
  - The odds of being either moderate or severe compared to mild are 2.85 times greater for those with a history compared to those without history
  - The odds of being severe compared to either mild or moderate are 2.85 times greater for those with a history compared to those without history

Assumptions of ordinal regression

- The only assumption to be fulfilled when applying ordinal logistic regression is that the parameters are the same across all categories
- A test can help you assess whether this assumption is reasonable, called ‘the test of parallel lines’
- It compares the estimated model with coefficients for all categories, to a model with a separate set of coefficients for each category
- A small *p*-value indicates that the general model (with separate parameters for each category) gives a significant improvement, i.e. the above assumption is not reasonable

Are the odds the same across categories?

- In the above examples, we assume that the odds ratio is the same across categories of the ordered response
- The analysis can test whether this assumption is valid

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	494.903			
General	494.067	.836	2	.658

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

- Test of parallel lines indicates that the assumption is reasonable ($p > .05$)

Key points

- Ordinal regression is an extension of the binary case, appropriate for ordered categorical outcome variables
- The most common method compares the proportional log-odds of outcome categories
- If the different categories have no natural ordering other methods exist (multinomial or polychotomous logistic regression)
- A difficult decision needs to be made on ordinal variables with a large number of categories - can the data be considered continuous? (number of categories, spread of data, normality)

Introduction to Poisson regression

- Counts are another form of numeric outcome variable. Sometimes, we can treat these as a continuous measure and use other methods such as linear regression (but this is not always appropriate)
- Counts can be rare events, such as the number of:
  - new disease cases occurring in a population over a period of time
  - hospital admissions per day
- Poisson regression can be used when:
  - the subjects may have the same duration of exposure (then we're just interested in the observed counts)
  - the subjects have a different amount of exposure (eg. length of follow-up, so we're interested in the counts after allowing for the time in the study)

- We face a constraint: counts are all positive integers. We therefore work with the log of the counts (in a similar way to working with the log of the odds for logistic regression)

- The natural logarithm of the response variable is linked to the covariates:

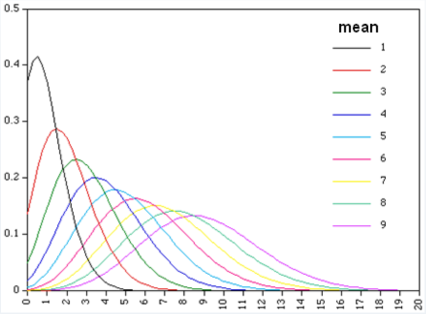
  $\log(Y) = a + bX_1 + cX_2 + dX_3 + ...$

- The Poisson distribution works with the mean of the outcome measure (a single parameter represents both the mean and the variance)

- This distribution is a natural fit for count data

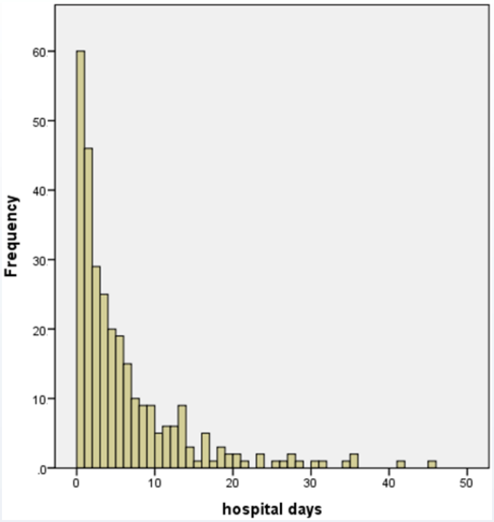
When can we use linear regression instead?

- First look at the mean of the counts. It is a skewed distribution if the mean is small, but becomes more symmetrical as the mean increases
- For large means, the Poisson distribution approaches Normality. We might then be able to use linear regression; but can also look at a Normal probability plot, in case the skewed distribution has just been shifted to the right



http://paulbourke.net/miscellaneous/functions/

Looking at the distribution



Summary:

Mean = 5.7 days
Median = 3.0 days
Range = 0 to 45 days

What the output gives you

- For each covariate we get the following statistical output:
  - estimated Poisson regression coefficient (e.g. *b*)
  - associated standard error
  - confidence limits
  - estimated relative rate, e.g. exp(*b*)
  - Wald test statistic, testing *b* = 0 or relative rate = 1
  - associated *p*-value
- For the overall model:
  - goodness-of-fit information
  - can be used when comparing models

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	Upper	Hypothesis Test Wald Chi-Square	df	Sig.
(Intercept)	2.749	.0920	2.569	2.930	893.265	1	.000
[gender=0]	-.417	.0502	-.516	-.319	69.132	1	.000
[gender=1]	0[a]	.	.	.	.	.	.
age	.017	.0017	.013	.020	93.366	1	.000

- The Poisson regression can be used to predict the number of days spent in hospital, given gender and age
- Model: log(Days) = 2.749 – (0.417 x Gender) + (0.017 x Age)

  [where Gender=0 for females and Gender=1 for males]
- If people have been in the study for very different lengths of time, we could include a covariate in the model to allow for this (i.e. for each subject you have length of time, and this is included as a factor in the regression model)

Parameter Estimates

Parameter	B	Std. Error	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	Upper
(Intercept)	2.749	.0920	893.265	1	.000	15.632	13.053	18.721
age	.017	.0017	93.366	1	.000	1.017	1.013	1.020
[gender=0]	-.417	.0502	69.132	1	.000	.659	.597	.727
[gender=1]	0	.	.	.	.	1	.	.

- The relative rate for gender is 0.66, with 95% CI 0.60 to 0.73
- The number of days spent in hospital is lower for women than men (after allowing for age); i.e. the rate is 34% lower for women (95% CI: 27 to 40%). [Gender=0 for females and 1 for males]
- The number of hospital days increases as age increases (after allowing for gender). As age increases by one year the rate increases by 2%

Assumptions of Poisson regression

- The assumptions include:
  - Logarithm of the response is approximately a straight line (analogous to log-odds for logistic regression)
  - At each level of the covariates the number of cases has variance equal to the mean
  - Observations are independent

- The same diagnostics can be used to identify violations of these assumptions in the case of Poisson Regression
  - Use plots of residuals against fitted values (i.e. how well does the model fit the observed data values?)

Is the Poisson model reasonable?

- If the variance is greater than the mean of the data, the data is said to be overdispersed. This can occur when:
  - There are outliers
  - Missing important covariates
  - There is a tendency for observations to cluster

- Overdispersed data have standard errors and *p*-values that are too small, and narrow confidence limits

- The Pearson adjustment (which can be specified in a stats package) can be used to correct the standard errors and give more accurate *p*-values (otherwise use more complex regression, called Negative Binomial)

Key points

- We can analyse count data by fitting Poisson regression models to the individual frequency of events
- The natural logarithm of the response variable is linked to the covariates
- Different lengths of exposure time can be accounted for in the model
- Also, variables that change over time can be incorporated by dividing up the follow-up time of each individual (eg. 5 years smoking status for each individual gives 5 rows of data)