Biostatistics & Epidemiological Data Analysis using R

11

Regression models for binary and count data

Juliana Schneider

Health Intervention Analytics Group, HPI

January 11, 2023

HPI Hasso Plattner Institut
Digital Engineering • Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

Overview

1. Regression for binary & count data
   - Logistic regression
   - Poisson regression

2. Generalized linear model
   - Overview
   - Further regression models

Learning objectives of today

- Introduction to logistic regression and Poisson regression for modeling binary and count outcomes.
- Introduction to theory of generalized linear models (GLMs).

Review: multiple linear regression

- Aim: Predict a variable Y by multiple variables $X_1, \ldots, X_k$ under the assumption of a linear relationship.
- Model equation:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- The regression coefficient β_j describes the effect of X_j on Y (increase of Y by β_j when X_j is increased by 1 unit), while holding all other X variables in the model constant.
- I.e. the estimated effect of each X_j on Y is adjusted for the effect of the other X variables in the model.

Logistic regression - regression for binary outcomes

Motivation for logistic regression

- What are arguments against computing a linear regression model for a binary outcome variable?

Overview

Aim

Predict a binary outcome variable Y by k variables X_j.

Example research questions

- Which variables are associated with the risk of Alzheimer's disease?
- Does the daily drinking of milk as a child have a positive effect on the probability of having osteoporosis later in life?
- Which factors are predictive of whether a person takes part in study xyz?

Derivation of model equation

Derivation

1. Construct latent variable Z as linear combination of the X_j:

$$Z = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

2. Link Z to Y (or rather to $p = P(Y = 1)$) using the logit function:

$$logit(p) = log\left(\frac{p}{1-p}\right) = Z \Leftrightarrow p = \frac{1}{1+e^{-z}}$$

3. Then, you can predict Y from Z (i.e. X) using a cut-off c, e.g. $c = 0$:

$$\widehat{Y} = \begin{cases} 1 & \text{if } z > c \\ 0 & \text{if } z \leq c \end{cases}.$$

Model equations

Instead of predicting Y as in linear regression, predict $P(Y = 1)$.

log-Odds : $$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

Odds : $$\frac{p}{1-p} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}$$

Probabilities : $$p = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)}}$$

where *log* is the logarithm with basis e.

Regression coefficients

- Determine estimates of the regression coefficients β_j using the maximum likelihood method.

- Likelihood function

$$L = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-z_i}} \right)^{y_i} \cdot \left(1 - \frac{1}{1+e^{-z_i}} \right)^{1-y_i}$$

- Maximize the likelihood function iteratively, eg using the Newton-Raphson algorithm or Fisher Scoring method (used in the `glm()` function).

- Hypothesis tests of the regression coefficients $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.

Logistic regression in R

- Compute regression: `fit <- glm(Y ~ X1 + X2 + X3, family = binomial(link = "logit"), data = dat)`
- Results: `summary(fit)`
- Odds ratio of the predictor effects: `exp(coef(fit))`
- Nice formatting with confidence intervals of odds ratio: `jtools::summ(model1, exp = T, confint = T, model.fit = F, digits = 3)`

Logistic regression in R

Important:

- In R, the `glm` function always predicts the highest level of factor variables per default ($Y = 1$).
- For factors with $k > 2$ levels, `glm` aggregates the highest $k - 1$ categories and computes the odds ratio in comparison to the lowest level.

Measures of model fit

Deviance & likelihood ratio test

- Deviance (perfect model fit: Deviance $=0$, smaller $=$ better), is returned directly with `summary(glm())` and can be extracted with `deviance(glm())`.
- AIC (smaller $=$ better), is returned directly with `summary(glm())` and can be extracted with `AIC(glm())`.
- Deviance χ^2-test tests H_0 : Deviance $=0$.
- Likelihood ratio test: Compare the deviance of a model with the deviance of the null (or other smaller nested) model. This can be done with `anova(glm(model1), glm(model2))`.

Measures of model fit

Pseudo R^2 statistics

- McFadden's R^2: between 0 and 1, is based on the log-likelihood.
- Cox & Snell R^2: between 0 and 0.75, is based on the likelihood, weighted by sample size.
- Nagelkerke R^2: with correction of Cox & Snell R^2 so that it is between 0-1.
- Can be computed with `DescTools::PseudoR2(glm())`.
- More information and discussion: Mittlbock & Schemper (1996), Menard (2000).

Measures of model fit

Classification results

- Use `predict(glm(), type = "response")` to predict the probability of an event based on the regression model.
- Determine a cut-off for the probability to classify the participants in $\widehat{Y} = 1$ or $\widehat{Y} = 0$.
- Compare these with the actual values of Y.

Assumptions

- Y is binary (rather requirement instead of assumption).
- Linear relationship between predictors and log-odds. Check: scatter plot of the predicted log-odds from `predict(glm())` vs. predictor, check if you can see a linear relationship.
- Observations are independent (if not: mixed logistic regression model).
- No multicollinearity (cf. linear regression).
- Logistic regression models are generally less robust compared to linear regression, eg if binary predictors have a very unbalanced distribution.

Conclusions

Conclusion 1

- Y is normally distributed $\longrightarrow$ linear regression
- Y is binary $\longrightarrow$ logistic regression
- Y categorical $\longrightarrow$ multinomial regression
- Y ordinal $\longrightarrow$ ordinal regression
- Y are counts $\longrightarrow$ Poisson or Negative Binomial regression

Conclusion 2

In comparison of linear and logistic regression:

- Interpretation of regression coefficients and their standard errors as well as of hypothesis tests is similar (but is for different outcome).
- R^2 and residuals are different.

Exercise 1

In the KiGGS data:

- Investigate if the amount and frequency of eating pancakes (`fq41` and `fq41a`) are associated with shortsightedness (`e0251`), using logistic regression.
- What do we predict: incidence or prevalence?
- Interpret the results.
- Add further predictors (eg `age2`, `sex`, `bmiB`) and interpret the results.
- Use the function `predict(glm(), type = "response")` to predict the probability of shortsightedness by the model. Also check the model fit with respect to the classification of shortsightedness. Conclusion?
- See `R_10_exercise1.Rmd`.

Poisson (loglinear) regression - regression for count outcomes

Overview

Aim

- Model the number (Y) or the rate (relative frequency Y/T, T is the observation period) of occurences of an event.

Examples

- Predict the incidence of breast cancer dependent on the number of days exposed to magnet fields at work.
- Predict the number of visits to the children's doctor or the number of measles vaccinations by characteristics of the child and of the environment.

Model equation

Model

- $log(\mu) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$
  for the mean number of events (ie mean absolute frequency) μ

- $log(\mu) = log(T) + \beta_0 + \sum_{j=1}^{k} \beta_j x_j$
  for mean rate (ie mean relative frequency) $\lambda = \frac{\mu}{T}$, with offset $log(T)$

Model components

Components

- $Y \sim Poisson(\mu) \longrightarrow f(y) = \frac{e^{-\mu}\mu^{y}}{y!}$ where
  - $f(y)$ is the probability that y events occur in one observation unit
  - eg: probability that 2 patients in station B of hospital A die within 1 year due to complications of gastroscopy.
  - μ is the expected number of events per observation unit (eg in T observation periods): $\mu = \lambda \cdot T$ with the event rate λ).
- $E(Y) = \mu, \qquad Var(Y) = E(Y) = \mu$
- Link function: $g = log$: $g(\mu) = log(\mu) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$

Note: $f(y)$ is sometimes also written $P(Y = y)$, since the Poisson distribution is discrete.

Likelihood function

Likelihood function

$$L = \prod_{i=1}^{n} \frac{e^{-\mu} \cdot \mu^{y_i}}{y_i!}$$

Assumptions

Assumptions

- Independent events, max. 1 event per time point.
- Predicted mean and variance of the event rate are equal (no overdispersion).
- Predicted event rate is the same in all observation units (eg in all hospitals & years).
- I.e. in general: the number of predicted events based on the Poisson distribution and Poisson regression model matches well the number of actual events.

If assumptions are violated

- Use corrected estimates of the standard error
- Use another distribution

Note:
(i) "Predicted" means the predicted values in regression model based on the predictors.

Poisson Regression in R

- Compute model: `fit <- glm(Y ~ X1 + X2 + X3, family = poisson(link = "log"), data = dat)`, where Y has to be numeric.
- With offset: `fit <- glm(Y ~ X1 + X2 + X3, family = poisson(link = "log"), offset = T, data = dat)` or `fit <- glm(Y ~ X1 + X2 + X3 + offset(T), family = poisson(link = "log"), data = dat)`
- Results: `summary(fit)`
- Predicted number of events: `predict(fit, type = "response")`
- Predictor effects: `exp(coef(fit))` or with `jtools::summ` function

Interpretation

Interpretation of the regression coefficients:

- Model with absolute frequencies: (adjusted) log-ratio of mean frequencies
- Model with offset total-n: (adjusted) log incidence ratio
- Model with offset person-years: (adjusted) log incidence rate ratio (IRR)

Exercise 2

In the KiGGS data:

- Predict the amount of visits to the children's doctor (`E085z01`) by the number of siblings (`e006B1`), the sex and age of the children (`sex`, age2), place of residence (`STALA`, `OW`) and the monthly household income (`e093`), using a Poisson regression model.
- Interpret the results. Which variables are associated with the outcome? Is the model a good fit to the data?
- See `R_10_exercise2.Rmd`.

Generalized linear model

Overview

What is the generalized linear model?

- Theoretical statistical model/framework, which extends the linear regression model to non-normally distributed Y.
- Includes the linear regression model, analysis of variance, analysis of covariance, logistic regression, Poisson regression, and many more regression models as special cases.

Overview

Why is the abstract description helpful?

Any statistical theory/property that can be shown in general, can be used for all special cases.

But:

Despite the same underlying theory, there are differences in the practical applications, e.g. meaning of R^2, residuals, robustness of model etc.

GLM components

(1) Random component

Describes the outcome variable Y and its probability density function.

(2) Systematic component

Contains the linear combination of the predictors.

(3) Link function

Function, which links the two components ($E(Y)$ and the systematic component).

GLM components in more detail

(1) Random component

Variable Y with n independent observations, which has a distribution from the expontential familiy, i.e. density function

$$f(y, \theta) = a(\theta) \cdot b(y) \cdot \exp(y \cdot Q(\theta)).$$

(2) Systematic component

Linear predictor of k variables: $\beta_0 + \sum_{j=1}^{k} \beta_j x_j$.

(3) Link function

Function g to link (1) in form of $E(Y) = \mu$ with (2):

$$g(E(Y)) = g(\mu) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

Note:
(i) The linear predictor in (2) is often denoted by η: $\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$.
(ii) $\beta_0 + \sum_{j=1}^{k} \beta_j x_j$ is often written as $\sum_{j=0}^{k} \beta_j x_j$ with $X_0 = 1$.

GLM components in summary

The following things have to be specified for a GLM (i.e. to compute it in R):

- Which variable Y will be predicted, and what is its distribution?
- (What is the (theoretical) expected value of Y?)
- With which variables X_j will Y be predicted?
- With which link function will $E(Y)$ be linked to the X_j, i.e. which exact quantity will be predicted by the X_j?

GLM components: examples

- What are the components of linear regression?
  - $Y \sim N(\mu, \sigma^2)$
  - $E(Y) = \mu$
  - g is the identity: $g(\mu) = \mu \longrightarrow E(Y) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$

- What are the components of logistic regression?
  - $Y \sim Bernoulli(p) \quad$ with $P(Y = 1) = p$
  - $E(Y) = p$
  - g is logit: $g(p) = logit(p) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$

How often are these models used in epidemiological/medical research?

Number of studies which mention the model:

Int J Epidemiol

- "Logistic regression": 2060
- "Multinomial regression": 16
- "Ordinal regression": 17
- "Poisson regression": 339
- "Negative binomial regression": 42

JAMA

- "Logistic regression": 11517
- "Multinomial regression": 60
- "Ordinal regression": 58
- "Poisson regression": 1035
- "Negative binomial regression": 261

Search results as of November 8, 2019

Further regression models

Multinomial regression - overview

Aim

- Predict categorical variable with >2 categories that doesn't have an ordinal scale (or whose ordinal scale will not be modeled).

Examples

- Predict risk of complications A, B, C of diabetes (given age, sex, therapy, BMI).
- Model risk of different breast cancer types (which are based on estrogen receptors, progesterone receptors, HER2).

Idea

- Compare categories in logit models to a reference category.

Multinomial regression - theory

Model

- The multinomial regression model consists of $L-1$ binary logit models:

$$log\left(\frac{P(Y=l)}{P(Y=0)}\right)=\beta_{0l}+\sum_{j=1}^{k}\beta_{lj}x_j, \quad l=1,\ldots,L-1$$

- where $Y=0$ is the reference category (can be specified) and L is the number of categories of Y.
- I.e. each predictor can have a different effect on the different category comparisons of Y:

$$\begin{aligned} log\left(\tfrac{P(Y=1)}{P(Y=0)}\right) &= \beta_{0,0}+\textstyle\sum_{j=1}^{k}\beta_{0,j}x_j \\ log\left(\tfrac{P(Y=2)}{P(Y=0)}\right) &= \beta_{1,0}+\textstyle\sum_{j=1}^{k}\beta_{1,j}x_j \\ &\vdots \\ log\left(\tfrac{P(Y=L-1)}{P(Y=0)}\right) &= \beta_{L-1,0}+\textstyle\sum_{j=1}^{k}\beta_{L-1,j}x_j \end{aligned}$$

Multinomial regression - theory

Components

- $Y \sim Multinomial(n, p_1, \ldots, p_L) \longrightarrow f(y) = \frac{n!}{y_1! \cdots y_L!} p_1^{y_1} \cdots p_L^{y_L}$
  - where n is the sample size,
  - $p_1, \ldots, p_L$ is the event probability of the L categories of Y,
  - and $y = (y_1, \ldots, y_L)$ is the frequency of the L categories of Y.
- Link function: generalized logit.

Multinomial regression - in R

- Load `nnet` package in R.
- If necessary, redefine reference category: `Y <- relevel(Y, ref = "RefCat")`
- Compute regression model: `fit <- nnet::multinom(Y ~ X1 + X2 + X3, data = dat)`, where Y has to be a factor or numeric variable.
- Results: `summary(fit)`
- p-values of testing the regression coefficients (Wald tests):

  1.
     ```
     W <- summary(fit)$coefficients/
               summary(fit)$standard.errors
     ```
  2. `(1 - pnorm(abs(W), 0, 1)) * 2`
- Odds ratio of the predictor effects: `exp(summary(fit)$coefficients)`
- Predicted probabilities: `fitted(fit)`

Ordinal (logistic) regression - overview

Aim

- Predict ordinal variable with > 2 different values.

Examples

- Predict the improvement of headache (no, a little, moderate, high) after taking any of 3 different drugs, dependent on age and sex.
- Predict obesity categories (instead of eg BMI as a quantitative variable) by age, sex, biomarker xy.

Idea

- Use logistic regression model and incorporate the different values of Y in p.

Ordinal regression - theory

Model

- The ordinal (logistic) regression model consists of $L-1$ binary logit (logistic regression) models:

$$logit(p) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j, \quad p = P(Y \leq l), \quad l = 1 \ldots L$$

- where L is the number of distinct values of Y. I.e.:

$$\begin{gathered} logit(P(Y \leq 0)) = \beta_0 + \textstyle\sum_{j=1}^{k} \beta_j x_j \\ logit(P(Y \leq 1)) = \beta_0 + \textstyle\sum_{j=1}^{k} \beta_j x_j \\ \vdots \\ logit(P(Y \leq L-1)) = \beta_0 + \textstyle\sum_{j=1}^{k} \beta_j x_j \end{gathered}$$

Ordinal regression - application

Assumption

- The effect of the predictors is constant over all values of Y ("proportional odds models").

Extensions

- Using other link functions, it can also be modeled that eg the effect of the predictors is not constant over all values of Y, and eg that lower/higher values are more likely.

Ordinal regression - in R

- Load `MASS` package in R.
- Compute regression model: `fit <- MASS::polr(Y ~ X1 + X2 + X3, method = "logistic", Hess = TRUE, data = dat)`, where Y has to be a factor.
- Results: `summary(fit)`
- Compute p-values of testing the regression coefficients (Wald tests): `pnorm(abs(summary(fit)$coefficients[, 3]), lower.tail = FALSE) * 2`
- Odds ratio of the predictor effects: `exp(summary(fit)$coefficients[, 1])`

Exercise 3

Exercise 3

In the KiGGS data:

- Predict the birth weight of children (`e017a.k`, in categories) by the BMI of their parents (`Mbmi.k`, `Vbmi.k`), their sex (`sex`), and the birth week (`e016z`).
- For this, use a multinomial regression and ordinal regression model, and compare the results to pairwise logistic regression models.
- Are the results similar? What are the advantages/disadvantages of each approach?
- See `R_10_exercise3.Rmd`.

Negative binomial regression - overview

Aim

- Predict the number/rate of occurences of an event, like in Poisson regression.
- But: The variance of the event occurences can be unequal to the mean, i.e. this is a generalization of Poisson regression.

Negative binomial regression - theory

Model

- Just like Poisson regression but with the assumption that Y follows a Negative binomial distribution.

Components

- $Y \sim NB(D, \mu) \longrightarrow f(y) = \binom{y+D-1}{D-1} \left(\frac{D}{\mu+D}\right)^D \left(1 - \frac{D}{\mu+D}\right)^y$
  - where D^{-1} is the dispersion parameter (estimated from the data)
  - and μ is the expected number of events.
- $E(Y) = \mu, \quad Var(Y) = \mu + \frac{\mu^2}{D}$
- Link function: $g = log$

Negative binomial regression - in R

- Load `MASS` package in R.
- Compute regression model: `fit <- MASS::glm.nb(Y ~ X1 + X2 + X3, data = dat)`, where Y has to be numeric.
- With offset: `fit <- MASS::glm.nb(Y ~ X1 + X2 + X3 + offset(T), data = dat)`
- Everything else like Poisson regression.

Questions?

References

- Menard (2000) Coefficients of determination for multiple logistic regression analysis. The American Statistician 54: 17-24.
- Mittlbock, Schemper (1996) Explained variation in logistic regression. Statistics in Medicine 15: 1987-1997.
- Agresti (2012) Categorical data analysis. Wiley.

Homework

Homework

See file R_10_homework.Rmd.