Generalized Linear Models

Lecture 4. Models with normally distributed response

Formulation of the problem

Assumptions:

- Observations y_i are realizations of (conditional) r.v. Y_i
- $Y_i \sim N(\mu_i, \sigma^2)$
- Independence: $cov(Y_i, Y_j) = 0, \; i \neq j$

R.v.-s Y_i constitute r.v. $\mathbf{Y} = (Y_1, \dots, Y_n)^T$

$$\Rightarrow \qquad \mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$$

Sample $\boldsymbol{y}$ is a random realization of n observations from $\mathbf{Y}$, $\boldsymbol{y} = (y_1, \dots, y_n)^T$
Design matrix $\mathbf{X}$

Classical linear model:

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

Link function: identity $g(\mu_i) = \mu_i$
Depending on the type of arguments we reach different classical models

Advantages of classical linear model

Models with normal response are simpler as compared to other members of exponential family:

- canonical link is identity
- variance function does not depend on the mean
- all cumulants except for first two are equal to 0
- in case of multivariate normal setup, the dependency structure is determined by covariance or correlation matrix

In case of other distributions, situation is not as simple nor clear

Assessing the normality assumption

Question

How important is the assumption of normality?

- important if n is small
- if $n \to \infty$, asymptotic normality follows from the central limit theorem

Central limit theorem assumes homogenous (constant) variance!

$\Rightarrow$ outliers may violate this assumption and void the convergence to normal distribution even if $n \to \infty$

Thus, we consider models where the response has constant variance

Estimation of β (fixed σ^2), 1

Consider the model $\mu_i = \boldsymbol{x}_i^T \beta$

Question

How to estimate the parameters β (model parameter) and σ^2 (parameter of dist.)?

In case of independent observations, the sample log-likelihood is

$$\ln L(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum \frac{(y_i - \mu_i)^2}{\sigma^2}$$

where $\mu_i = \boldsymbol{x}_i^T \beta$ (and assume that σ^2 is fixed)
NB! Maximizing the log-likelikood is equivalent to minimizing the residual sum of squares:

$$RSS(\beta) = \sum (y_i - \mu_i)^2 = (\boldsymbol{y} - \mathbf{X}\beta)^T (\boldsymbol{y} - \mathbf{X}\beta)$$

Derivative w.r.t. β leads us to normal equations:

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \boldsymbol{y}$$

Estimation of β (fixed σ^2), 2

If $\mathbf{X}$ has full rank, so has $\mathbf{X}^T\mathbf{X}$, which implies that $\exists(\mathbf{X}^T\mathbf{X})^{-1}$ so that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

If the inverse matrix does not exist, generalized inverse can be used (but the solution is not unique!)

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\boldsymbol{y}$$

Estimation of parameter β. Algorithmic solutions

Main difficulty: estimation of $(\mathbf{X}^T\mathbf{X})^{-1}$

- **Gauss elimination method**. Beaton (1964)
  SWEEP-operator technique
- **Cholesky decomposition**
  Main idea is to find a triangular matrix $\mathbf{L}$ such that $\mathbf{X}^T\mathbf{X} = \mathbf{L}\mathbf{L}^T$, which implies $(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{L}^{-1})^T\mathbf{L}^{-1}$
- **QR decomposition** (*Gram-Schmidt orthogonalization*)
  Matrix $\mathbf{X}$ is decomposed as a product $\mathbf{X} = \mathbf{Q}\mathbf{R}$,
  where $\mathbf{Q}$ is a $n \times n$ orthogonal matrix, i.e $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$
  $\mathbf{R}$ – $n \times p$ (upper) triangular matrix such that $\mathbf{R}^T\mathbf{R} = \mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R} = \mathbf{X}^T\mathbf{X}$
  $\mathbf{Q}, \mathbf{R}$ can be found using different methods (Householder's method, Givens rotation, and more)

Properties of the ordinary least squares (OLS) estimator

By Gauss-Markov theorem (provided that the assumptions hold)

- OLS estimator is unbiased: $\mathbf{E}\hat{\beta} = \boldsymbol{\beta}$
- OLS estimator is effective (has minimal variance)

i.e. OLS estimate is *BLUE – best linear unbiased estimate*

Assumptions:

- $\mathbf{E}\varepsilon_i = 0,\ \mathbf{D}\varepsilon_i = \sigma^2, \forall i$
- $cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ then OLS estimate is also ML estimate and

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Estimation of σ^2

Log-likelihood of a sample: $\ln L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum \frac{(y_i - \mu_i)^2}{\sigma^2}$

where $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
Now, substitute the obtained estimate $\hat{\boldsymbol{\beta}}$ to the equation

$$\ln L(\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{RSS(\hat{\beta})}{\sigma^2}$$

to get so-called **profile likelihood** for σ^2
As usual, take the derivative by σ^2, equate it to zero to obtain the following (biased!) estimate

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n}$$

Unbiased estimate is given by:

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n-p}$$

Hypotesis testing. Wald test

A. To test a single parameter $H_0: \ \beta_j = 0$

$$t = \frac{\hat{\beta}_j}{\sqrt{\sigma^2_{\hat{\beta}_j}}}$$

If σ^2 **estimated** then $t \sim t_{n-p}$; If σ^2 **known** then $t \sim N(0,1)$
In case of big samples $(n \to \infty)$ $t \overset{a}{\sim} N(0,1)$

B. To test more than one parameter $H_0: \ \boldsymbol{\beta}_2 = 0$
$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $(p_1 + p_2)$-dimensional

$$w = \hat{\boldsymbol{\beta}}_2^T \Sigma^{-2}_{\hat{\boldsymbol{\beta}}_2} \hat{\boldsymbol{\beta}}_2$$

Under the normality assumption, $w \sim \chi^2_{p_2}$, if σ^2 is **known**
If σ^2 is **estimated** then $\frac{w}{p_2} \sim F_{p_2, n-p}$, $p = p_1 + p_2$
If $n \to \infty$ then $n - p \to \infty$ and (scaled!) F-distribution $\to \chi^2_{p_2}$

Hypothesis testing. Likelihood ratio test

To test more than one parameter $H_0 : \beta_2 = 0$
$\beta = (\beta_1^T, \beta_2^T)^T$, $(p_1 + p_2)$-dimensional

$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is divided into two parts (p_1 and p_2 parameters)

Compare the models:
$M = M(\mathbf{X})$ (upper model, all arguments)
$M_1 = M(\mathbf{X}_1)$ (lower model, p_1 parameters, $k_1 = p_1 - 1$ arguments)
Compare the corresponding log-likelihoods (σ^2 **known**)
$\max \ln L(\beta_1) = C - \frac{1}{2}\frac{RSS(\mathbf{X}_1)}{\sigma^2}$, where $C = -\frac{n}{2}\ln(2\pi\sigma^2)$ does not depend on β
$\max \ln L(\beta) = C - \frac{1}{2}\frac{RSS(\mathbf{X}_1+\mathbf{X}_2)}{\sigma^2}$

Likelihood ratio statistic (λ)

$$-2\ln\lambda = \frac{RSS(\mathbf{X}_1) - RSS(\mathbf{X}_1 + \mathbf{X}_2)}{\sigma^2}$$

If σ^2 is **not known**, it will be estimated from the upper model:
$\hat{\sigma}^2 = RSS(\mathbf{X}_1 + \mathbf{X}_2)/(n-p)$
In case of big samples $-2\ln\lambda \sim \chi^2_{p_2}$

Regression diagnostics. Residual analysis

Model $\boldsymbol{y} = \mathbf{X}\beta + \varepsilon$

Model residuals $\hat{\varepsilon}$ (or $\boldsymbol{e}$) are the estimates of random error ε
$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}, \quad \hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the "hat" matrix $\hat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y}$
$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}, \quad \mathbf{D}\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}$

Variance of i-th residual is thus $\sigma^2_{\hat{\varepsilon}_i} = (1 - h_{ii})\sigma^2$

$\Rightarrow$ **residuals may have different variances even if the observations have constant variance (σ^2), since the estimates also depend on the arguments!**

Standarized/Studentized residuals

Standardized residuals (also *internally studentized*)

$$e_{iS} = \frac{e_i}{\sqrt{1 - h_{ii}}\hat{\sigma}}$$

Studentized residuals (also *externally studentized, studentized deleted*)

$$e_{iT} = \frac{e_i}{\sqrt{1 - h_{ii}}\hat{\sigma}_{(i)}}$$

Standardized/Studentized residual is too big if it is ≈ 3 (already > 2 can be considered)

Leverage and influence

Leverage is the diagonal element h_{ii} of hat matrix H (*Hat diag*)

$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, $rank(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = k+1 \Rightarrow \frac{k+1}{n}$
Leverage is too big: $h_{ii} > \frac{2(k+1)}{n}$

Influence is the observation's effect on parameters (prediction, parameters' variance)

Observation's influence is estimated by Cook's statistic (`cooks.distance`, in R package `stats`)
Observation's influence to a particular parameter estimate: dfbetas (*Difference of Betas*, in R package `stats`)

$$\texttt{dfbetas(model)}_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{\hat{\sigma}_{(i)}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}$$

Empirical estimate: influence is too big if `dfbetas` $> \frac{2}{\sqrt{n}}$

Transformations

Transformations are used to transform non-symmetric distributions close to normal and also to stabilize the variance

George Edward Pelham Box (b. 1919), Sir David Roxbee Cox (b. 1924)

- Box-Cox (1964) family of power-transformations
- Yeo-Johnson (2000) family of power-transformations

Box-Cox transforms are modified, because

1. Not all data can be transformed to be close to normal
2. Initial restriction $y > 0$
3. Work well if the transformation is applied to a unimodal non-symmetric distribution
4. Do not work well in case of U-shaped distributions

Box-Cox family of transformations

Box and Cox (1964) – there exist non-symmetric distributions that can be transformed quite close to a normal distribution

General form of the transformation:

$$y(\lambda) = \left\{ \begin{array}{ll} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{array} \right\}$$

$y > 0$, λ – parameter of the transformation, usually $\lambda \in (-2, 2)$
The transformation is simplified to y^λ if $\lambda \neq 0$ (Cleveland, 1993)

Known transformations:

$$\lambda = -1 \Rightarrow \frac{1}{y}$$

$$\lambda = 0 \Rightarrow \ln y$$

$$\lambda = 0.5 \Rightarrow \sqrt{y}$$

$$\lambda = 1 \Rightarrow y$$

$$\lambda = 2 \Rightarrow y^2$$

Box-Cox transformation. General schema

Assume that $\exists \lambda$, such that the transformed data is normal:

$$Y_i(\lambda) \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)$$

Estimation (using ML):

1. fix λ, estimate $\boldsymbol{\beta}$, σ^2
2. substitute the obtained estimates to ML expression to get the function $pL(\lambda)$

$pL(\lambda)$ – **profile likelihood** of parameter λ

Box-Cox transformation (1)

NB! Don't forget the Jacobian $J(\lambda, y)$ while transforming $y \to y(\lambda)$

$\lambda \neq 0,\ y(\lambda) = \frac{y^\lambda - 1}{\lambda}$

$$f(y_i|\lambda, \mu_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ y_i^{\lambda-1} \exp[-\frac{1}{2\sigma^2}\{\frac{y_i^\lambda - 1}{\lambda} - \mu_i\}^2]$$

$\lambda = 0,\ y(\lambda) = \ln y$

$$f(y_i|0, \mu_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ y_i^{-1} \exp[-\frac{(\ln y_i - \mu_i)^2}{2\sigma^2}]$$

$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, thus $\boldsymbol{\mu} = \mu(\boldsymbol{\beta})$

Box-Cox transformation (2)

Main steps:

1. Find the log-likelihood of the sample
2. Fix λ, find the partial derivatives of the log-likelihood by $\boldsymbol{\beta}$ and σ^2
3. Equate the derivatives to 0, obtain the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$
4. Substitute the estimates to the expression of likelihood, obtain the profile log-likelihood for λ:

$$pl(\lambda) = -\frac{n}{2} \ln RSS(\lambda) + (\lambda - 1) \sum \ln y_i$$

5. Maximizing over λ-s gives the optimal λ

R: function `boxcox` (package `MASS`), more advanced version: `function boxCox` (package `car`), SAS: `proc TRANSREG`

Box-Cox transform. Example 1

Data: distance (in km) and fuel consumption (in litres), $n = 107$
Simple regression model: y – distance, x – fuel consumption
Box-Cox transform was used

Results:

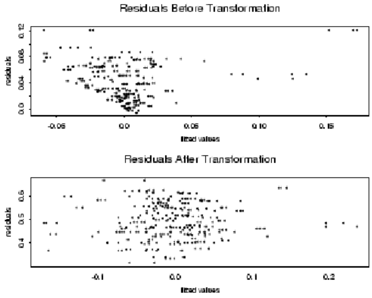
- model parameters: intercept $\hat{\beta}_0 = -636.9$, $\hat{\beta}_1 = 211.9$, $R^2 = 0.49$
- estimated $\lambda = 1.5$ 95% CI: (0.7; 2.4)

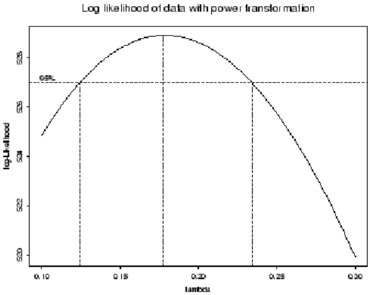
Can you write down the corresponding model?

NB! Box-Cox method gives a suggestion about the range of transformations
NB! The transformation changes the scale, thus it is also important to consider the interpretability of the model!

Source: Chen, Lockhart, Stephens (2002)

Box-Cox transform. Example 2





Left figure shows the residuals before and after transform
Right figure shows the log-likelihood of data under different λ-s, maximum is obtained if $\lambda = 0.2$, i.e. the transformation is $\sqrt[5]{y}$

The necessity of a transform. Atkinson scores

Question

Is the Box-Cox transformation necessary at all?

To test that, an additional term will be added to the model:

$$a_i = y_i(\ln \frac{y_i}{\tilde{y}} - 1),$$

where $\tilde{y}$ is the geometric mean of $\boldsymbol{y}$

Let us denote the coefficent of the extra term a_i by γ
If the extra term is significant then the Box-Cox transform is necessary and

$$\hat{\lambda} \approx 1 - \hat{\gamma},$$

where $\hat{\gamma}$ is the estimate of γ from the model

Source: Atkinson (1985)

Argument transforms

Box, Tidwell (1962): similar approach as with Atkinson scores

Question

Is an argument transform necessary?

To test if, in case of a continuous argument x, it is necessary to add x^λ to a model (if x already is included), an extra term $a = x \ln x$ is used so that the model contains x (coefficient β) and $x \ln x$ (coefficient γ)

If the extra term is significant, then the transform is necessary and $\hat{\lambda} \approx \frac{\hat{\gamma}}{\hat{\beta}} + 1$, where $\hat{\gamma}$ is the estimated coefficient of the extra term,
$\hat{\beta}$ is the coefficient of argument x from the original model (without $x \ln x$)

Both Atkinson and Box-Tidwell method are based on the Taylor series expansion. Assume that the correct model is $y = \alpha + \beta x^\lambda + \epsilon$, using Taylor expansion x^λ at $\lambda = 1$ yields $x^\lambda \approx x + (\lambda - 1)x \ln x$
Substitute this into the model, get $y = \alpha + \beta x + \beta(\lambda - 1)x \ln x + \epsilon$ and denote $\gamma = \beta(\lambda - 1)$

R: `function boxTidwell` (package `car`)

Yeo-Johnson family of power-transformations

Box-Cox: restriction $y > 0$

Idea: find a transform that minimizes Kullback-Leibler information and transforms a skewed distribution to symmetric
New concepts: *relative skewness* (Zwet, 1964), *more right-skewed, more left-skewed*

Yeo-Johnson family of power-transformations

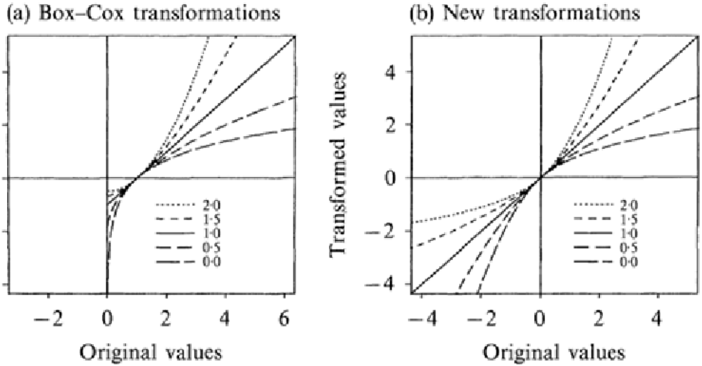
$$\psi(y,\lambda) = \begin{cases} ((y+1)^{\lambda} - 1)/\lambda, \quad \lambda \neq 0, y \geq 0 \\ \ln(y+1), \quad \lambda = 0, y \geq 0 \\ -((-y+1)^{2-\lambda} - 1)/(2-\lambda), \quad \lambda \neq 2, y < 0 \\ -\ln(-y+1), \quad \lambda = 2, y < 0 \end{cases}$$

If case $y > 0$, this construction is equivalent to Box-Cox transformation

R: `function boxCox` with parameter family="yjPower" (package `car`)

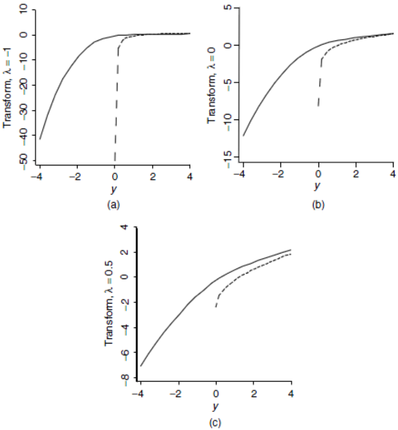
Yeo, I.-K., Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87,4,954–959

Comparison of transformations (1)



Comparison of Box-Cox transformations and new (Yeo-Johnson) transformations under different values of λ

Comparison of transformations (2)



Comparison of Box-Cox transformations and new (Yeo-Johnson) transformations if $y \to 0$

Comments about transformations

- Box-Cox method gives a suggestion about the range of transformations. The transformation changes the scale, thus it is also important to consider the interpretability of the model.
- Box-Cox transforms are empirical, based on data.
  There are also transforms for stabilizing the variance that are based on theoretical considerations
- John Tukey, Fred Mosteller (1977) '*bulging rule*' – two-dimensional graphs show which transformation to use

Bulging rule

Transformation depending on data

Figure 4.6 from Fox (1997)