Model Diagnostics: residual plots, transformations

Emilija Perković

Dept. of Statistics
University of Washington

Checking assumptions

We assume:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.
Is this model reasonable? Are the assumptions on the errors $\underline{\epsilon}$ satisfied?
For all $i, j \in \{1, \dots, n\}$, $i \neq j$, we are assuming:

- $\mathrm{E}[\epsilon_i] = 0$,
- $\mathrm{Var}[\epsilon_i] = \sigma^2$,
- $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$,
- if $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

However, we do not observe ϵ_is, so there is no way to directly check our assumptions. Instead, we will be checking whether the residuals $\hat{\epsilon}_i$ satisfy the above assumptions.
We will mostly be relying on graphical checks of assumptions.
For now we will only consider how to check these assumptions. Remedial measures will be discussed later.

Mean zero and constant variance assumptions

Some ideas for graphical checks of $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$:

- plot residuals against each predictor x_i, $i = 1, \ldots, p$,
- plot residuals against the fitted values $\hat{y}$ (a linear combination of predictors) - also called the Tukey-Anscombe plot,
- plot the residuals against some other arbitrary variable.

If $\mathrm{E}[\epsilon_i] = 0$ is satisfied, all of the above plots should show a flat scatter around 0. Curvature or steps in these plots indicate presence of non-linearity, or of an omitted important predictor.

If $\mathrm{Var}[\epsilon_i] = \sigma^2$, is satisfied, all of the above plots should show a constant width of the points scatter around 0 for all values on the horizontal axis.

Residuals when the model is (not) correct

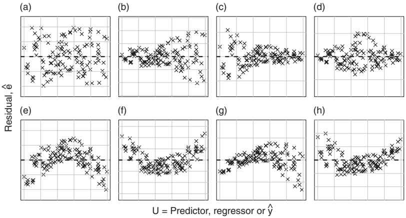


Figure 9.2 Residual plots: (a) null plot; (b) right-opening megaphone; (c) left-opening megaphone; (d) double outward bow; (e) and (f) nonlinearity; (g) and (h) combinations of nonlinearity and nonconstant variance function.

Figure 9.2 From Weisberg. See R script for more examples.

Example: Savings

Consider `savings` data from R package `faraway`.
Savings rates from 50 countries, averaged over the period 1960-1970.

- sr - personal saving divided by disposable income
- pop15 - percent population under age of 15
- pop75 - percent population over age of 75
- dpi - per-capita disposable income in dollars
- ddpi - percent growth rate of dpi

We fit a multiple linear regression with `sr` as response and check the assumptions using residual plots.

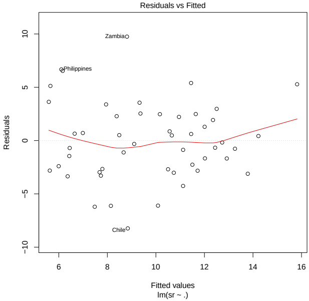
Example: Savings

```
lm(formula = sr ~ ., data = savings)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334
pop15       -0.4611931  0.1446422  -3.189 0.002603
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471
---
Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

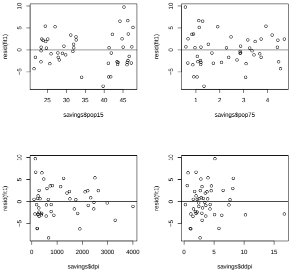
Residuals against fitted

Are the assumptions satisfied?



Residuals against fitted

Are the assumptions satisfied?



Another plot for checking constant variance

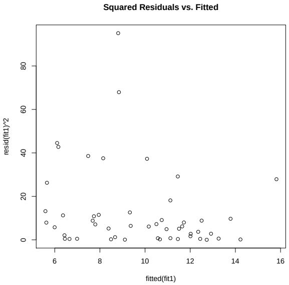
Since $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$, it follows that $\mathrm{E}[\epsilon_i^2] = \sigma^2$.

- Plot squared residuals against each of the predictors,
- Plot squared residuals against the fitted values.

Again, if $\mathrm{Var}[\epsilon_i] = \sigma^2$ is satisfied, the above plots should show a constant width of the scatter for all values on the horizontal axis.

Example: Savings

Is the assumption of constant variance satisfied?



Example: Savings

Another possible check (not very formal, similar to considering the smoother in the TA plot):

Fit a linear model with squared residuals as the response and fitted values from the multiple regression as the independent variable.

Is the slope significant?

```
lm(formula = resid(fit1)^2 ~ fitted(fit1))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.726     10.235   2.709  0.00933
fitted(fit1)   -1.521      1.023  -1.488  0.14337
```

Checking for zero correlation

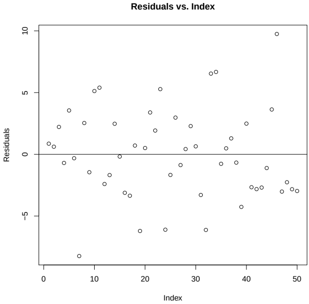
In order to check $\text{Cov}[\epsilon_i, \epsilon_j] = 0$:

- Plot residuals against the index.
- If the observations are dated, time-stamped or spatially located, plot the residuals as a function of time or make a map.
- If there is a meaningful order in the observations plot residuals from successive observations.

If $\text{Cov}[\epsilon_i, \epsilon_j] = 0$ is satisfied, we again expect to see a flat scatter around 0 and a general lack of structure in the plot.

Example: Savings

Is $\text{Cov}[\epsilon_i, \epsilon_j] = 0$ satisfied?



Residuals vs. Errors

Note that we are checking the assumptions made on the errors $\underline{\epsilon}$, using the residuals $\hat{\underline{\epsilon}}$.

However, from previous lectures we know that even if $\underline{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, then

$$\hat{\underline{\epsilon}} \sim \mathcal{N}(0, \sigma^2 (I_n - H)),$$

where $H = X(X'X)^{-1}X'$, also called the hat matrix, since $\hat{\underline{y}} = H\underline{y}$.
Properties of the hat matrix:

- H is an orthogonal projection (see Linear model handout):
  - $H' = H$ - symmetric,
  - $H^2 = H$ - idempotent,
- $HX = X$,
- $\sum_{i=1}^{n} h_{ii} = p + 1$, h_{ii} is the (i, i) element of H.

Residuals vs. Errors

Note that since,

$$\underline{\hat{\epsilon}} \sim \mathcal{N}(0, \sigma^2(I_n - H)),$$

We have for each $i, j \in \{1, \dots, n\}$:

- $\mathrm{E}[\hat{\epsilon}_i] = 0$,
- $\mathrm{Var}[\hat{\epsilon}_i] = \sigma^2(1 - h_{ii})$,
- $\mathrm{Cov}[\hat{\epsilon}_i, \hat{\epsilon}_j] = -\sigma^2 h_{ij}$.

So even if the assumptions on the errors are satisfied, we **expect to have some non-constant variance** in the residuals and **some non-zero correlation of residuals**.

How to check the assumptions?

Transforming the residuals

One idea is to replace the raw residuals with **standardized residuals**.

Note that since $\mathrm{Var}[\hat{\epsilon}_i] = \sigma^2(1 - h_{ii})$,

$$SE(\hat{\epsilon}_i) = \hat{\sigma}\sqrt{1 - h_{ii}}.$$

Standardized residuals r_i are defined as:

$$r_i = \frac{\hat{\epsilon}_i}{SE(\hat{\epsilon}_i)}.$$

Also, known as *internally studentized residuals*.

The distribution of $r_i^2/(n-p-1)$ is $\beta(1/2, (n-p-2)/2)$ (Seber and Lee, 2003, p. 267), so these residuals should have a constant variance.

Additionally, the distribution of r_i approaches a Gaussian for $n \to \infty$ and a fixed p.

Transforming the residuals

We can obtain a further refinement on the raw residuals.

Let $\underline{\hat{y}}_{(i)}$ denote the fitted values of an OLS regression which was fit after excluding the i-th data point from the original data set.

Assuming $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$, let

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{SE(y_i - \hat{y}_{i(i)})},$$

then

$$t_i \sim t_{n-p-2}.$$

t_i, $i \in \{1, \ldots n\}$ are referred to as the studentized residuals, or the leave-one-out residuals.Will be discussed further when we talk about outliers.

Transforming the residuals

There is a connection between the standardized and the studentized residuals:

$$t_i = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{1/2},$$

where

- r_i is the standardized residual,
- t_i is the studentized residual.

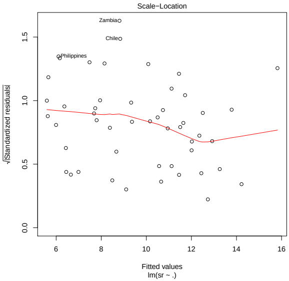
These residuals can be obtained in R with

- `rstandard()` - for standardized residuals
- `rstudent()` - for studentized residuals

All plots for checking constant variance and lack of correlation on the raw residuals can also be used with standardized or studentized residuals.

Example: Savings

In R `plot(fit, which=3)` gives the following plot (uses studentized residuals):



Is the assumption of constant variance satisfied?

The normality assumption

Two standard graphical ways of assessing normality are:

- **Histogram:** Make a histogram of $\hat{\epsilon}_i$'s.
  - This should look approximately bell-shaped if the population is really normal **and** there are enough observations.
  - If there are enough observations, graphically compare the histogram to a $N(0, \hat{\sigma}^2)$ distribution. (Or use standardized residuals and compare with $\mathcal{N}(0, 1)$.)
  - In small samples, the histograms need not look particularly bell-shaped.
- **Normal probability, or qq-plot:**
  If $\epsilon_i \sim N(0, \sigma^2)$ then the ordered residuals $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$ should correspond linearly with quantiles of a standard normal distribution.

Obtaining a normal QQ plot

Idea: Form the **order statistics**:

$$\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(n)},$$

Compare these *sample quantiles* to *theoretical quantiles*:

$$z_1 \le z_2 \le \cdots \le z_n.$$

Let Φ be the cdf of $N(0,1)$. Then

$$z_i = \Phi^{-1}\left(\frac{i-1/2}{n}\right), \quad i = 1, \dots, n,$$

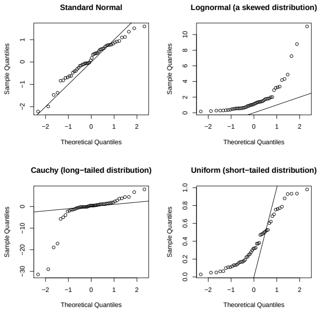
approximates the expectation of the i-th order statistic of a $N(0,1)$-sample. (Subtract 1/2 so as to not have $\Phi^{-1}(1) = \infty$, for $i = n$.)

If data are normal, the relationship should be approximately linear. Thus we plot the pairs

$$(z_1, \hat{\epsilon}_{(1)}), \dots, (z_n, \hat{\epsilon}_{(n)})$$

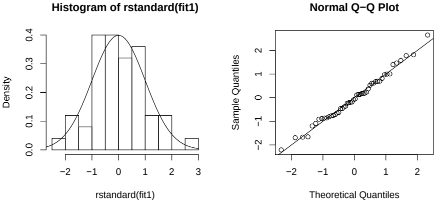
and assess the linearity of the relationship. (In R this is implemented in `qqplot()` with ordered standardized residuals.)

Example of QQ plots



Example: Savings

Is the normality assumption satisfied?



Remedial measures

What if one of the diagnostic plots shows some assumption violations?

Variable transformations.

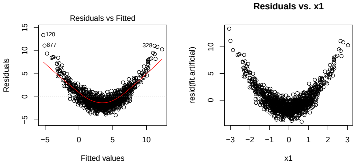
Ideally, would perform these before fitting the model.

Example: Artificial data (see R script). The true relationship between y and x_1 is:

$$y = 2 + 3x_1 + 1.5x_1^2 + \epsilon, \epsilon \sim \mathcal{N}(0,1).$$

What happens to the residual plots if we do not include the quadratic term?

Residual plots

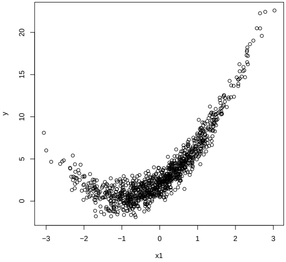


A curvature of the mean is present.
These plots seem to indicate a missing quadratic term of x_1.

Can we detect this prior to performing a regression?

Scatter plot

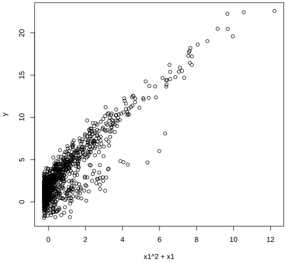
Consider the scatter plot of y vs. x_1:



Plot indicates a non-linear (quadratic) relationship of y with x_1.

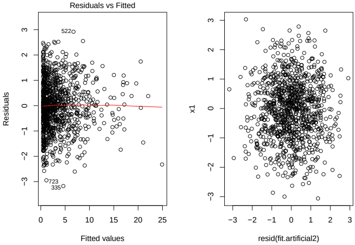
Scatter plot

Consider the scatter plot of y vs. $x_1 + x_1^2$:



This plot seems “more” linear!

Residual plots after including the quadratic term



The assumptions appear to be satisfied.

Guidelines for including polynomials

- Consider the pairwise scatter plot of y and each of the predictors and try to “guess” the relationship. (Not a very satisfying answer.)
- Non-linear least squares (will not be covered in this course, requires that you know the functional form, and are estimating the βs.)
- Use non-parametric regression methods such as smoothing (see e.g., `smooth.spline()` in R, we will also not cover this in the course)

A word of caution:

- Do not eliminate lower order terms from the model even if they are not significant after adding a higher order term. Why? Because we want our model to not be dependent on the additive changes in scale.

For example, suppose you fit the model:

$$y = \beta_0 + \beta_1 x_1^2 + \epsilon$$

and, then decide to transform x_1 as $x_1' = x_1 + a$,and model

$$\begin{aligned} y &= \beta_0 + \beta_1 {x_1'}^2 + \epsilon \\ &= (\beta_0 + \beta_1 a^2) + 2\beta_1 a x_1 + \beta_1 x_1^2 + \epsilon \end{aligned}$$

This model now contains a linear dependence on x_1!

Example: Brains

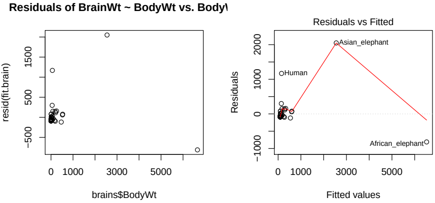
What about other non-linear relationships of the predictors and response?

Example: `brains` data set from `alr4`.

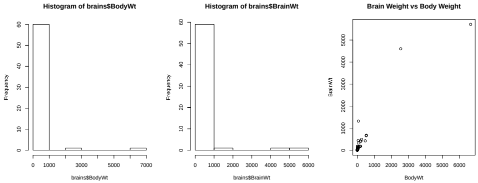
How does the brain weight of a mammal depend on its body weight?
$n = 62$ measurements of various mammals’ body weights and brain weights

- BodyWt - body weight in grams,
- BrainWt - brain weight in grams.

Residual plots

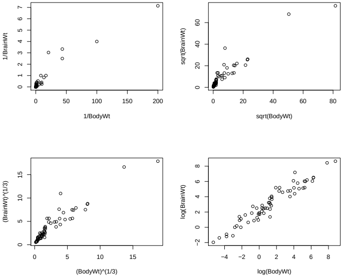


Consider the histograms and the scatterplot



Our data is very (right) skewed! In these cases, possible transformations: square root, cube root, inverse, log.

Some transformations



Seems that log transforming the predictor and response results in a linear relationship.

Example: Brains

Based on the scatter plot observations we could conclude that the linear model:

$$\underline{\texttt{BrainWt}} = \beta_0 + \beta_1 \underline{\texttt{BodyWt}} + \underline{\epsilon}, \tag{1}$$

with $\underline{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ makes little sense. However, the model:

$$\texttt{log}(\underline{\texttt{BrainWt}}) = \beta_0 + \beta_1 \texttt{log}(\underline{\texttt{BodyWt}}) + \underline{\epsilon}, \tag{2}$$

$\underline{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ appears to be more reasonable.
What does transforming the response mean for our assumption on the errors?We are now assuming that:

$$\begin{aligned} \underline{\texttt{BrainWt}} &= e^{\beta_0 + \beta_1 \texttt{log}(\underline{\texttt{BodyWt}})} e^{\underline{\epsilon}} \\ &= e^{\beta_0} e^{\texttt{log}(\underline{\texttt{BodyWt}})^{\beta_1}} \underline{\xi} \\ &= \beta_0' \underline{\texttt{BodyWt}}^{\beta_1} \underline{\xi}, \end{aligned}$$

where $\underline{\xi}$ now follows a multivariate log-normal distribution with parameters 0 and $\sigma^2 I_n$. Note that $\mathrm{E}[\xi_i] = \sigma^2/2$. Errors enter multiplicatively → non-constant variance.

Predicted value of y

For a simple prediction of the y-value on the original scale (**fitted value** of y), we can exponentiate to invert the log-transformation: $\hat{y} = exp(\hat{y}') = e^{\hat{y}'}$.

- Caution: this is an estimate for the median of the conditional distribution, but not of the conditional mean
- If we require unbiased fitted values on the original scale, applying a correction factor is required!
- We can either use :

  $$\hat{y} = exp(\hat{y}' + \hat{\sigma}^2/2),$$

  which is motivated by the link between Gaussian and lognormal distribution, or
- the smearing estimator proposed by Duan (1983):

  $$\hat{y} = exp(\hat{y}')\frac{1}{n}\sum_{i=1}^{n} exp(\hat{\epsilon}_i).$$

- These corrected values $\hat{y}$ are the estimates of the conditional mean (expectation) of y.

Variance stabilizing transformations

Transformations of the response when the residual vs. fitted plot indicates non-constant variance. Weisberg offers some general heuristics: (require y to be non-negative).

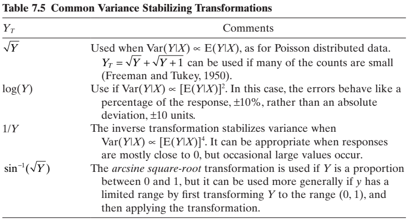


Table 7.5 Common Variance Stabilizing Transformations

Y_T	Comments
$\sqrt{Y}$	Used when Var(Y|X) ∝ E(Y|X), as for Poisson distributed data. $Y_T = \sqrt{Y} + \sqrt{Y+1}$ can be used if many of the counts are small (Freeman and Tukey, 1950).
log(Y)	Use if Var(Y|X) ∝ [E(Y|X)]^2. In this case, the errors behave like a percentage of the response, ±10%, rather than an absolute deviation, ±10 units.
1/Y	The inverse transformation stabilizes variance when Var(Y|X) ∝ [E(Y|X)]^4. It can be appropriate when responses are mostly close to 0, but occasional large values occur.
$\sin^{-1}(\sqrt{Y})$	The *arcsine square-root* transformation is used if Y is a proportion between 0 and 1, but it can be used more generally if y has a limited range by first transforming Y to the range (0, 1), and then applying the transformation.

These heuristics are also often applied to the predictors as well.

Box-Cox transformation

It is difficult to determine which transformation of the response is most appropriate by just relying on diagnostic plots.

The Box-Cox procedure (1964) offers a way to automatically identify a transformation from the family of power transformations on the response. We transform the response y as $g_\lambda(y)$ where

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

with λ a parameter estimated from data.Recall (by L'Hopital's rule):

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$$

Box-Cox is a popular way to find an appropriate transformation for a **strictly positive** response.
If the response is non-negative, but can take value zero, then usually a constant small value $\delta > 0$ is added to y in order to be able to perform a Box-Cox transformation.

Box-Cox transformation

How to choose λ?

- The estimation process for λ is based on maximizing a likelihood of the data when errors are assumed to be Normally distributed. That is, we assume:

  $$g_\lambda(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

  where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and estimate λ by maximizing the normal log-likelihood of the errors (see Box and Cox, 1964 for details).
- The Box-Cox family is appealing because it allows for easy interpretation (power transformation);
- it includes important special cases: untransformed, inverse, logarithm, square root and cube root.

This procedure is implemented in R function `boxcox()` in the package `MASS` and `boxCox()` in the package `car` (and probably some others).

Example: Savings

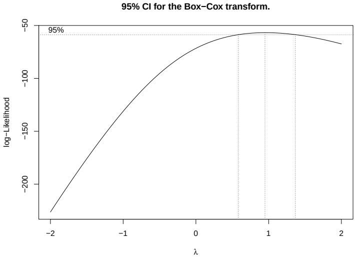
Consider again `savings` data from R package `faraway`.
Recall: Savings rates from 50 countries, averaged over the period 1960-1970.

- sr - personal saving divided by disposable income
- pop15 - percent population under age of 15
- pop75 - percent population over age of 75
- dpi - per-capita disposable income in dollars
- ddpi - percent growth rate of dpi

Let’s check whether `sr` needs a transformation.

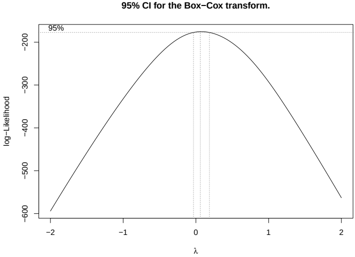
Example: Savings

```
boxcox(fit1,plotit=T)
title(main = "95% CI for the Box-Cox transform.")
```



What about the brains data?

```
boxcox(fit.brains,plotit=T)
title(main = "95% CI for the Box-Cox transform.")
```



Box-Cox: Practical considerations

- Estimate the best value $\widehat{\lambda}$ of λ.

- Given $\widehat{\lambda}$ use the transformation $y^{\widehat{\lambda}}$ for fitting regression models.(Can also use $(y^{\widehat{\lambda}} - 1)/\widehat{\lambda}$, but it's more conventional to use $y^{\widehat{\lambda}}$)

- We can test if a transformation is necessary by calculating a confidence interval for testing the null hypothesis $H_0 : \lambda = \lambda_*$ (e.g. $\lambda_* = 0$).

- Do not be overzealous with applying the precise $\hat{\lambda}$ transformation recommended by the `boxcox()` function.
  Consider the 95% confidence interval. If the $\hat{\lambda} = 0.67$, but the confidence interval includes .5, use the square root transform.

- To keep the interpretation task simpler, only transform the response when absolutely necessary.

Box-Cox: Practical considerations

- Keep in mind that the choice of λ is affected by outliers.
- There is a question whether estimation of λ should count as a parameter in the df of the model. This is a nontrivial issue, since λ is not a linear parameter and is not part of the least squares fit.
- Yeo-Johnson is a family of transformations that shares many good properties of Box-Cox family but can be applicable to variables with ranges over positive and negative values.

Yeo-Johnson transformation

A parametric family of transformations that can be used without restrictions on y, and has many good properties of the Box-Cox family.

We transform the response y as $\psi_\lambda(y)$ where

$$\psi_\lambda(y) = \left\{ \begin{array}{ll} \frac{(y+1)^\lambda - 1}{\lambda}, & \text{for } \lambda \neq 0 \text{ and } y \geq 0 \\ \log(y+1), & \text{for } \lambda = 0 \text{ and } y \geq 0 \\ \frac{1-(1-y)^{2-\lambda}}{2-\lambda}, & \text{for } \lambda \neq 2 \text{ and } y < 0 \\ -\log(1-y), & \text{for } \lambda = 2 \text{ and } y < 0 \end{array} \right\}$$

where λ is again a parameter to be determined from data (same as for the Box-Cox transformation, see Yeo and Johnson, 2000).

Yeo-Johnson transformation

- If y is strictly positive, then the Yeo-Johnson transformation is the same as the Box-Cox transformation of $1+y$, that is $g_{\lambda}(1+y)$.
- If y is strictly negative, then the Yeo-Johnson transformation is $-g_{2-\lambda}(1-y)$.
- With both negative and positive values of y, the transformation is a mixture of these two.

In R, you can estimate λ for the Yeo-Johnson transformation with function `boxCox(., family="yjPower")` in the package `car`.
See also function `yjPower()` in package `car`.

General guidelines

- When it comes to transformations keep in mind what they mean for the interpretability of your model. Is your primary goal predictive performance?
- Consider whether transforming the response is something you would consider.
- If yes, find an appropriate transformation using the Box-Cox, Yeo-Johnson procedure or some other method/heuristic.
- They consider scatter plots and the relationship of the transformed response and the predictor.
- Decide on potential predictor transformations.
- Fit the model and **perform the residual analysis.**