Transformations to Correct Model Inadequacies

In Chapter 4 you have seen mostly graphical methods to check whether the assumptions we place on a linear regression model

- linear relationship between each predictor and the response
- uncorrelated residuals with $N(0, \sigma^2)$ distribution

are satisfied. We have not yet discussed, what can be done if the model assumptions are *not* satisfied. Depending on which assumption(s) are violated, different corrective procedures may be employed. Some are quite simple, while others require more complicated procedures. One of the simplest adjustments to a regression model is to transform one or more of the predictor variables or the response.

Variance-Stabilizing Transformations

We can see in a residual plot, in which the residuals are plotted against the predicted values $\hat{y}$ or against a predictor, whether the residuals have constant variance or not. If the constant variance assumption is violated, then the cause is often that the response values y do not follow a Normal distribution with constant variance but instead some other distribution for which the variance is functionally related to the mean. In this case, transforming the response can be enough to satisfy the model assumptions. Which transformation is appropriate depends on the shape of the point cloud in the residual plot and on the problem itself. Functions that can be easily explained in the context the data is taken from are vastly preferable over more complicated functions, even if the latter would lead to "prettier" residual plots in the end.

Table of common transformations and when to use them:

Relationship of σ^2 to $E(y)$	Transformation	
$\sigma^2 \propto$ constant	$y' = y$	(do nothing)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$	(square root)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = sin^{-1}(y)$	(arcsine transformation)
$\sigma^2 \propto E(y)^2$	$y' = \ln(y)$	(log-transformation)
$\sigma^2 \propto E(y)^3$	$y' = 1/\sqrt{y}$	(reciprocal square root)
$\sigma^2 \propto E(y)^4$	$y' = 1/y$	(reciprocal)

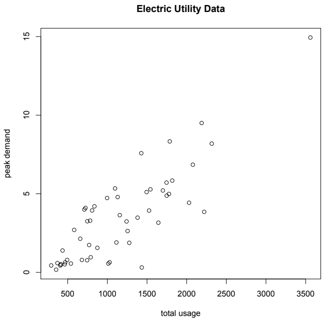
In Reality: transforming the response to adjust the model for non-constant residual variance is a trial-and-error procedure. You apply some of the common transformations to the response, refit the model with the transformed response and re-check the residual plots for each model. Weigh complicatedness of the transformation against "prettiness" of the residual plots. The simplest transformation that leads to acceptable residual plots should be preferred.

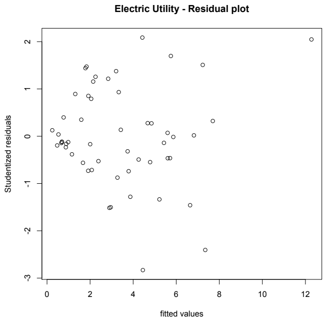
Note: If a variance stabilizing transformation has been applied to the response, then predicted values and prediction intervals or confidence intervals for the mean are now expressed in the new (transformed) units and not in the original units of y. The same is true for measures such as the MS_{res} or the PRESS statistic. It is usually possible to "translate" these values back into the original units.

Example: The Electric Utility Data

An electric utility company is interested in developing a model that relates peak-hour demand (y in KW) to total energy usage (x, in KWh) during the month. This relationship is important for the company, because they have to plan their generation system for the peak usage, while customers are charged for the total energy they use. The data set "ElectricUtility.txt" contains observations on 53 residential customers for the month of August.

Begin by fitting a simple linear regression model to the data:

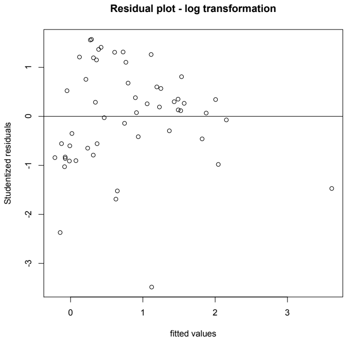

Electric Utility Data

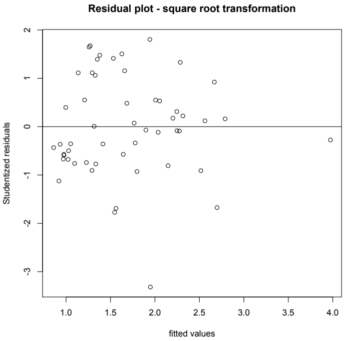

Electric Utility - Residual plot

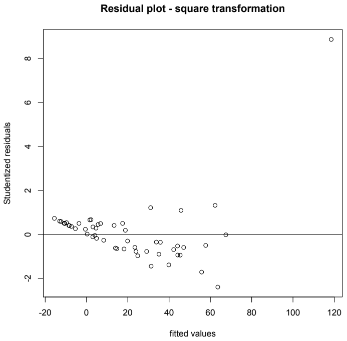
```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 302.63 302.633  121.66 4.106e-15 ***
Residuals 51 126.87   2.488
```

Even though one might be able to accept the appearance of the scatterplot (approximately linear relationship), and even though the test for significance of regression is very small ($p = 4\ 10^{-15}$), the residual plot shows a clear "funnel-shaped" pattern that suggests that the assumption of constant residual variance is not satisfied. For this reason, the p-value in the ANOVA table *cannot* be interpreted.

Let's try three different transformations on the response y and investigate their effects on the model residuals.


Residual plot - log transformation

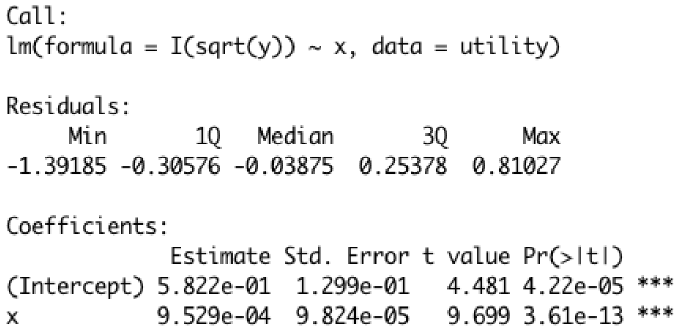

Residual plot - square root transformation


Residual plot - square transformation

Some transformations ($y' = y^2$) are clearly not an improvement. The best transformation is likely the $y' = \sqrt{y}$ transformation, since the residual plot for the transformed model has the least pattern. Thus, the transformed model is

$$\sqrt{y} = \beta_0 + \beta_1 x + \epsilon$$

The coefficients for the transformed model can be computed in R:



```
Call:
lm(formula = I(sqrt(y)) ~ x, data = utility)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.39185 -0.30576 -0.03875  0.25378  0.81027 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.822e-01  1.299e-01   4.481 4.22e-05 ***
x           9.529e-04  9.824e-05   9.699 3.61e-13 ***
```

Use the output shown above to predict the peak hour usage for a customer whose total energy usage during August is 2500 KWh.

Transformations to Linearize the Model

Nonlinearity can occur not only in the response but also in one or more predictors in a multiple regression model. Sometimes, a nonlinear relationship between a predictor and the response can be detected in the panel-scatter plot or with residual plots against the predictors. In many cases, the model can be improved by replacing the predictor that is causing the problem with a non-linear function of the same variable. Such nonlinear models are called intrinsically linear.

Example: If the scatterplot of y against x suggest an exponential relationship between x and y, then an appropriate model would be

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

This model is intrinsically linear, because it is equivalent to the linear model

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

or

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

Here, the usual model assumptions are placed on the transformed residuals ϵ' and *not* the raw residuals ϵ.

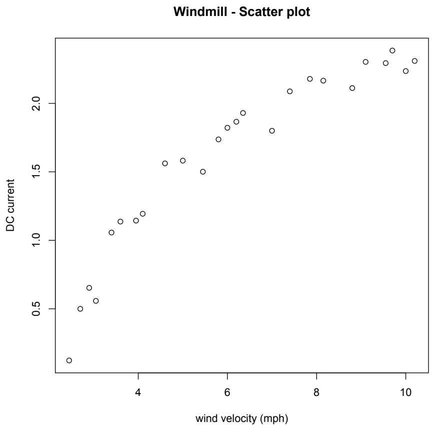
Other common transformations for regression models include:

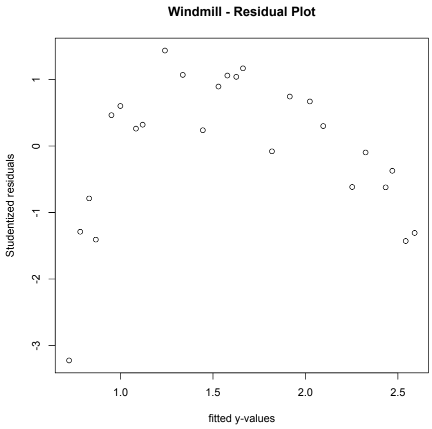
Linearizable Function	Transformation	Linear Form
$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Note: You should be familiar with the shape of the functions most commonly used for transformations $(e^x, \ln x, 1/x$ etc.$)$ and be able to recognize those shapes in scatter plots, for example.

Example: The Windmill Data

An engineer is using a windmill to generate electricity. She has collected data on the DC output of the windmill and the corresponding wind velocity (in mph). The data are available in the file "Windmill.txt" on the course website. A scatter plot of the data shows that the relationship between predictor and response is clearly not linear. The curve pattern is even more pronounced in the corresponding residual plot.





Some model characteristics of the (poorly fitting) linear model are:

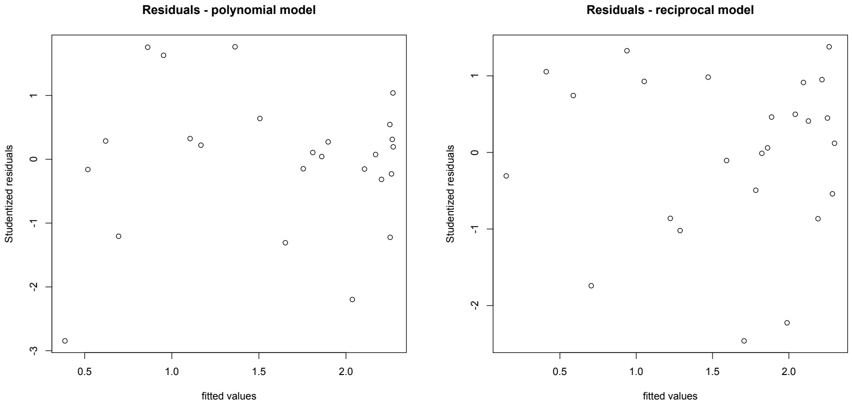
Model	R^2	MS_{Res}	F
$\hat{y} = \beta_0 + \beta_1 x$	0.8745	0.2361^2	160.3

Look closely at the scatterplot and suggest some possible functional relationships between x and y:

Repeat creating the residual plots and finding the model characteristics for all "candidate" models.

Model	R^2	MS_{Res}	F
$\hat{y} = \beta_0 + \beta_1 x$	0.8745	0.2361^2	160.3
$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$	0.9676	0.1227^2	328.3
$\hat{y} = \beta_0 + \beta_1 \frac{1}{x}$	0.98	0.09417^2	1128

Weigh all pieces of information you have against each other. You would ideally want a residual plot without any pattern (confetti in a box). You would want a high R^2, a low MS_{Res} and a large F-test statistic for significance of regression. In addition, any information that we have about the situation the data is taken from, should also be considered when deciding which the "best" model for this data is. For instance, we know that the DC current increases as the wind speed increases. In the quadratic model, the DC current would decrease again as the velocity increases - this may not make sense for this application.



Which model would you prefer?

Use the selected model to predict the DC current we would get from a windmill on a day where the wind speed is 8mph and provide a 95% confidence interval for your prediction.

The Box-Cox Method

In some problems, it is not obvious what the "best" transformation for a given data set would be. In 1964 George Box and David Cox came up with a method that would automatically select an optimal transformation of the response y. The method makes use of a family of possible power transformations.

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

The optimal transformation parameter λ and the parameters of the least squares regression model

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

are computed simultaneously with maximum likelihood techniques.

Note: This method works only, if the response $\mathbf{y}$ takes on only positive values. That's usually not a problem. If the response takes on some negative values, add an appropriately large constant to all observations as a "pre-transformation".

Basic Idea: Where does the family of functions come from that are used in the Box-Cox method? It would be nice, if we could just consider simple functions like y^λ for various values of λ. But this function makes regression impossible if $\lambda = 0$. So, consider instead the function $(y^\lambda - 1)/\lambda$ which, for large values of λ behaves just like a power function. As λ approaches zero, the limit of this function is the natural logarithm of y. Note, that λ can take on both positive and negative values.

In Practice: The optimal value of λ minimizes the residual sum of squares for the regression model of the transformed data. Usually, this value of λ is found empirically, by fitting models for a sequence of different values of λ and plotting $SS_{Res}(\lambda)$ for each model. Simple choices for values of λ are usually best. The following table shows what transforming with specific values of λ means in practice:

λ	y'
$\vdots$	$\vdots$
-2	$y' \propto 1/y^2$
-1	$y' \propto 1/y$
-0.5	$y' \propto 1/\sqrt{y}$
0	$y' \propto \ln y$
0.5	$y' \propto \sqrt{y}$
1	$y' \propto y$
2	$y' \propto y^2$
$\vdots$	$\vdots$

When software is used to efficiently compute the likelihood (or log-likelihood) for many different values of λ, the program often also reports a confidence interval for λ

together with the value of λ that maximizes the log-likelihood function. If one of the easily explainable transformations (e.g., $\lambda = 0.5$) is in the confidence interval, then this transformation may be preferable, even if a close-by value (e.g., $\lambda = 0.476$) has a higher log-likelihood value. If $\lambda = 1$ is in the confidence interval, then no transformation may be necessary.

The construction of this confidence interval is based on the fact that the log-likelihood function for λ can be written in terms of the residual sum of squares:

$$L(\lambda) = -\frac{1}{2} n \ln[SS_{Res}(\lambda)]$$

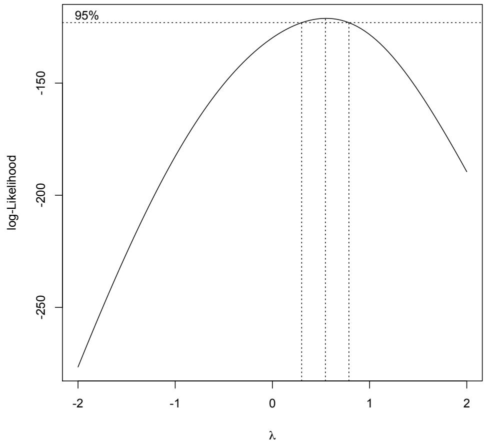
Let $\hat{\lambda}$ be the value that maximizes the log-likelihood function. Since maximizing the log-likelihood function is the same as minimizing the residual sum of squares and the residual sum of squares has a χ^2-distribution, a $(1-\alpha)100\%$ confidence interval for λ is obtained by drawing a horizontal line at height

$$L(\hat{\lambda}) - \frac{1}{2}\chi^2_{1-\alpha,1}$$

This line will "cut" the log-likelihood function at two points. These two points establish the upper and lower bounds of the confidence interval for λ.

Example: The R-package `MASS` contains the function `boxcox()` that computes log-likelihood values for a given sequence of λ-points.

Recall, that for the Electric Utility data we had selected a transformation of $y' = \sqrt{y}$ based on the appearance of the residual plots. If the Box-Cox method is applied to the same data set, the resulting log-likelihood function is graphed as a function of λ. The middle dashed line represents the ML-estimate of λ. The outside dashed lines represent the 95% confidence interval for λ.



Use R to find the ML estimate $\hat{\lambda}$ and approximate upper and lower bounds for the 95% confidence interval for λ.