Regression diagnostics Part II
Variable transformations

Biometry 755

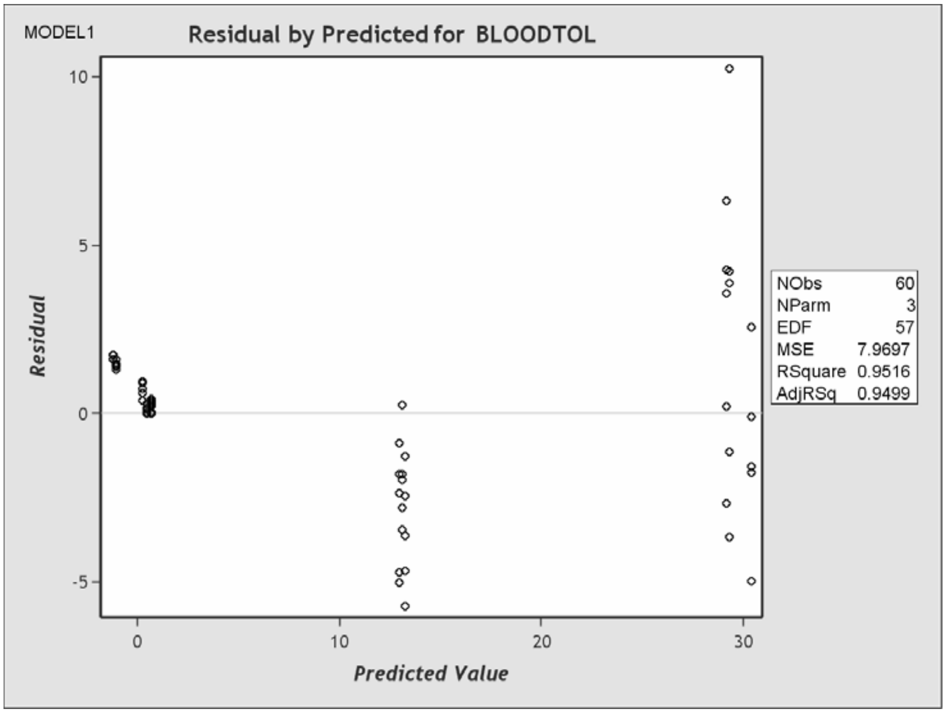
Spring 2009

Transformations

Transformations of either the response or a predictor or both can help alleviate assumption violations. There are three main reasons for transforming the data.

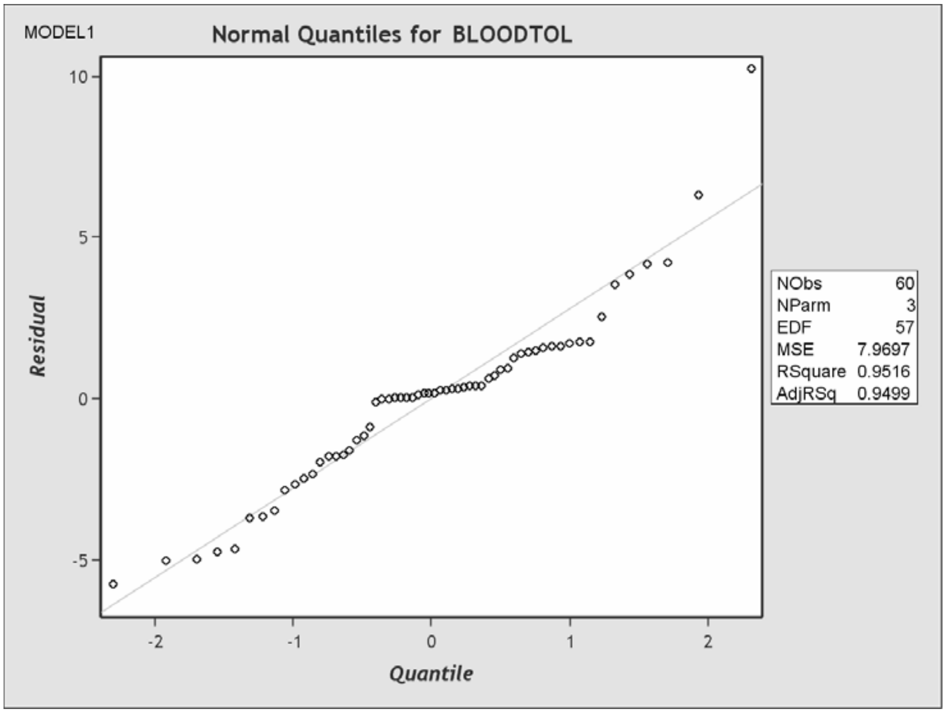
1. To stabilize variance (address violation of homoscedasticity assumption)
2. To normalize the dependent variable (address violation of normality assumption)
3. To linearize the regression model (address violation of linearity assumption)

Heterogeneous variance



NObs	60
NParm	3
EDF	57
MSE	7.9697
RSquare	0.9516
AdjRSq	0.9499

Normality violation



NObs	60
NParm	3
EDF	57
MSE	7.9697
RSquare	0.9516
AdjRSq	0.9499

Transformations on Y

Nonnormality and unequal variance often appear together. Frequently, these violations take the form of skewness and increasing variability of the distribution of the error terms as the mean response increases (i.e. variance is proportional to the mean). We observed this pattern on Slide 3. Note the normality violation for the same model shown on Slide 4. The usual transformations of Y that remedy these departures are

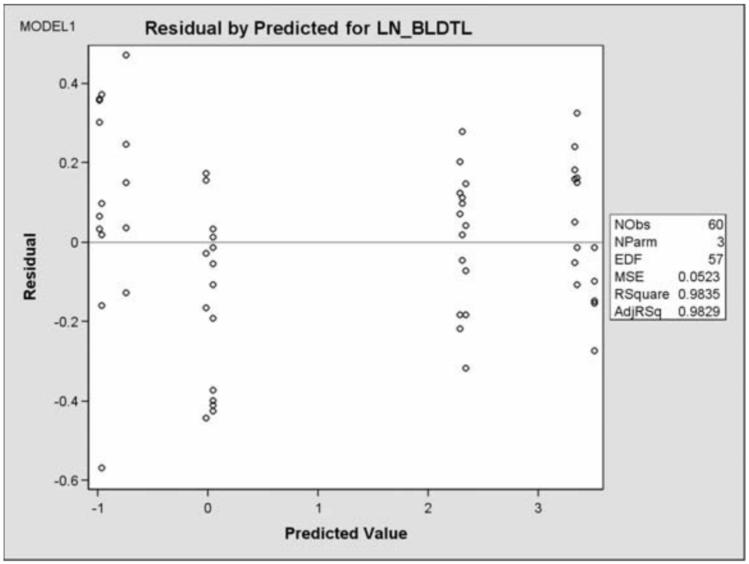
1. $\sqrt{Y}$
2. $\log_{10} Y$
3. $\frac{1}{Y}$

Transformations on Y (cont.)

Often there is an established transformation in the literature (e.g. $\log_{10}($viral load$)$). In the absence of an established functional form of the response, try various transformations to see which one best alleviates the violation. Note that a simultaneous transformation of X is often helpful or necessary.

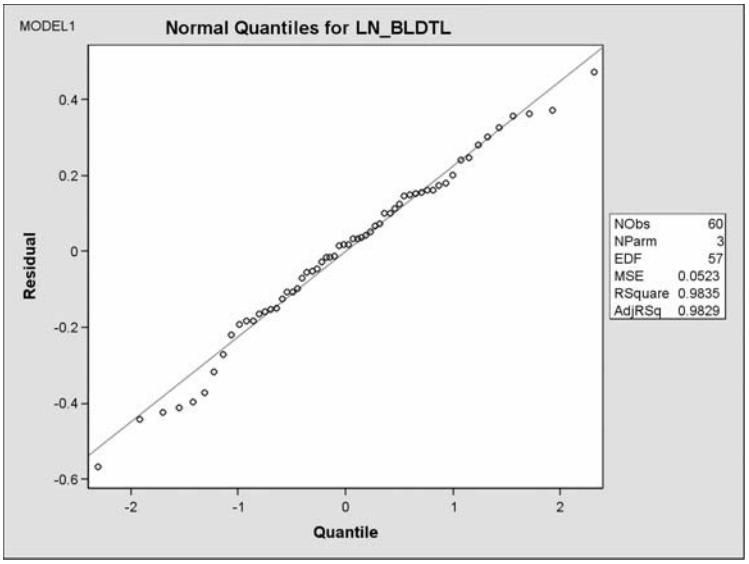
Transformations on Y (cont.)

This is a plot of the residuals against the predicted values for the same data as depicted in Slide 3. Here, both the response variable and the predictor have had a logarithmic transformation (base e).



Transformations on Y (cont.)

Note that this transformation has also alleviated the normality violation. Compare this normal qq plot of the residuals to that in Slide 4.



Box-Cox transformations of Y

(*Summarized from SAS documentation.*)

An automated approach to finding an appropriate transformation on Y is the Box-Cox (1964) transformation. This family of transformations of the positive dependent variable Y is controlled by the parameter λ. The transformation takes the form

$$\begin{array}{ll} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0. \end{array}$$

Box-Cox transformations of Y (cont.)

More generally, Box-Cox transformations take the form

$$\begin{array}{ll} ((y+c)^\lambda - 1)/(\lambda g) & \lambda \neq 0 \\ \log(y+c)/g & \lambda = 0. \end{array}$$

The parameter c can be used to rescale Y so that it is strictly positive. By default, $g = 1$. Alternatively, g can be $\dot{y}^{\lambda-1}$ where $\dot{y}$ is the geometric mean of Y.

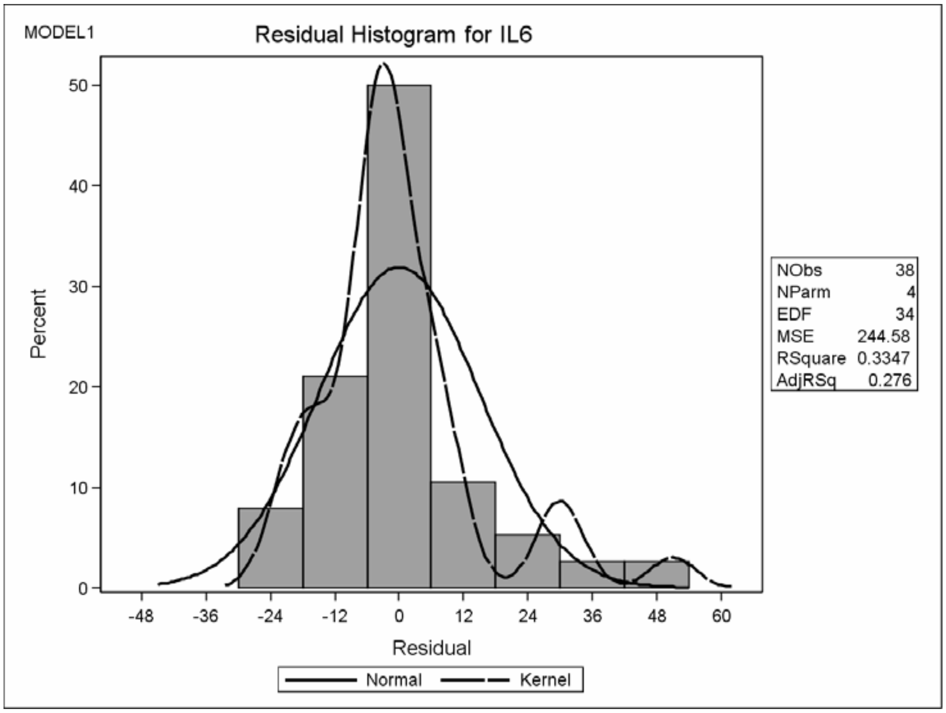
HPV/HNCa cytokine data

```
Obs       IL6      group       Obs       IL6      group       Obs       IL6      gr
  1     4.570        1         19     8.070        2         37     7.470       4
  2    43.320        1         20     8.670        2         38     7.090       4
  3     2.400        1         21    13.570        2
  4    22.340        1         22    81.240        2
  5    11.380        1         23    17.110        2
  6     5.040        1         24    38.130        2
  7     7.250        1         25    10.840        3
  8     8.510        1         26     1.350        3
  9     5.990        1         27     3.060        3
 10    15.050        1         28     0.675        3
 11     4.870        1         29     0.675        3
 12     7.500        1         30     3.480        3
 13    13.070        2         31     2.980        3
 14    58.840        2         32     3.860        3
 15    12.240        2         33     8.210        3
 16    15.100        2         34    14.100        3
 17    37.420        2         35     3.530        4
 18    59.770        2         36    11.760        4
```

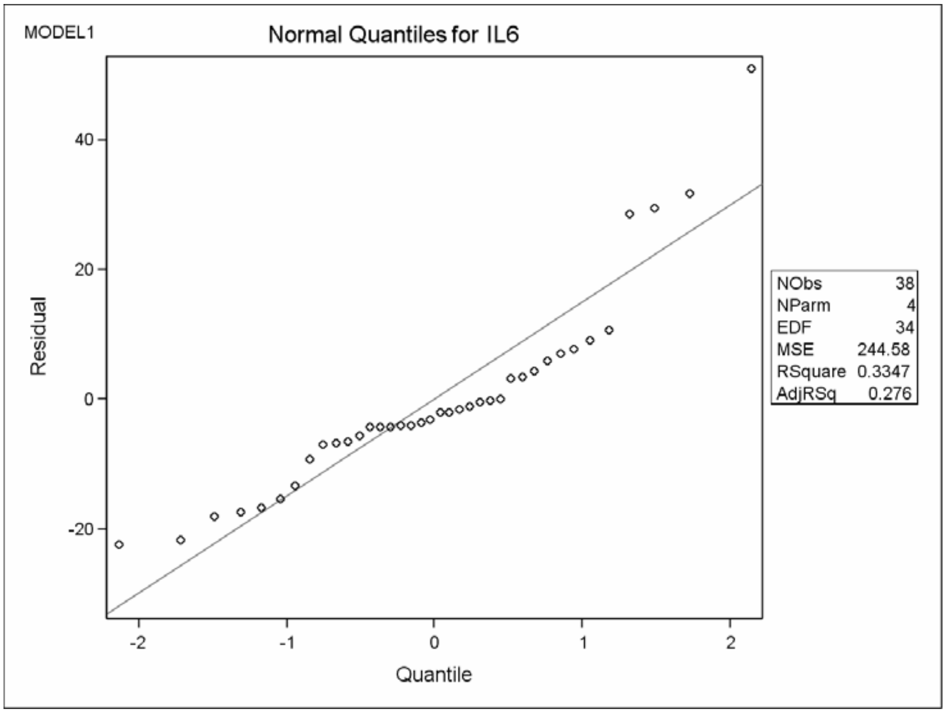
Regression of IL6 conc. on group

```
data one;
    set cytokine;
    if group = 1 then gpind1 = 1; else gpind1 = 0;
    if group = 2 then gpind2 = 1; else gpind2 = 0;
    if group = 3 then gpind3 = 1; else gpind3 = 0;
run;

ods html style = Journal;
ods graphics on;
ods select ResidualHistogram;
ods select QQPlot;
ods select ResidualByPredicted;

proc reg data = one plots(unpack);
    model IL6 = gpind1 gpind2 gpind3;
run; quit;

ods graphics off;
ods html close;
```

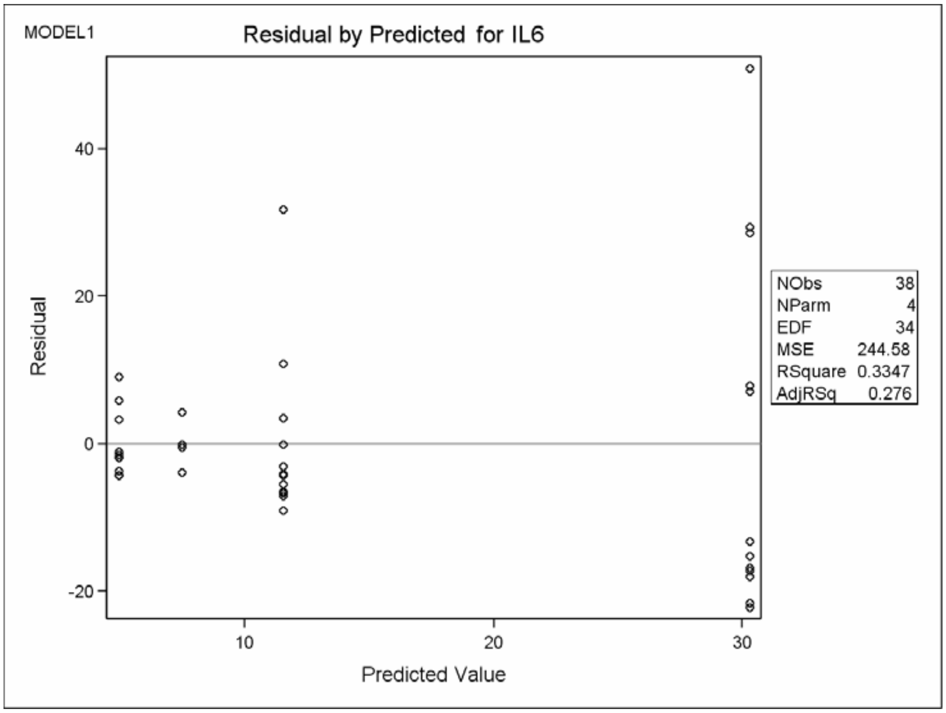
HPV/HNCa cytokine example



HPV/HNCa cytokine example (cont.)



HPV/HNCa cytokine example (cont.)



Box-Cox in SAS

```
proc transreg data = one;
    model boxcox(IL6) = identity(gpind1 gpind2 gpind3)
run;
```

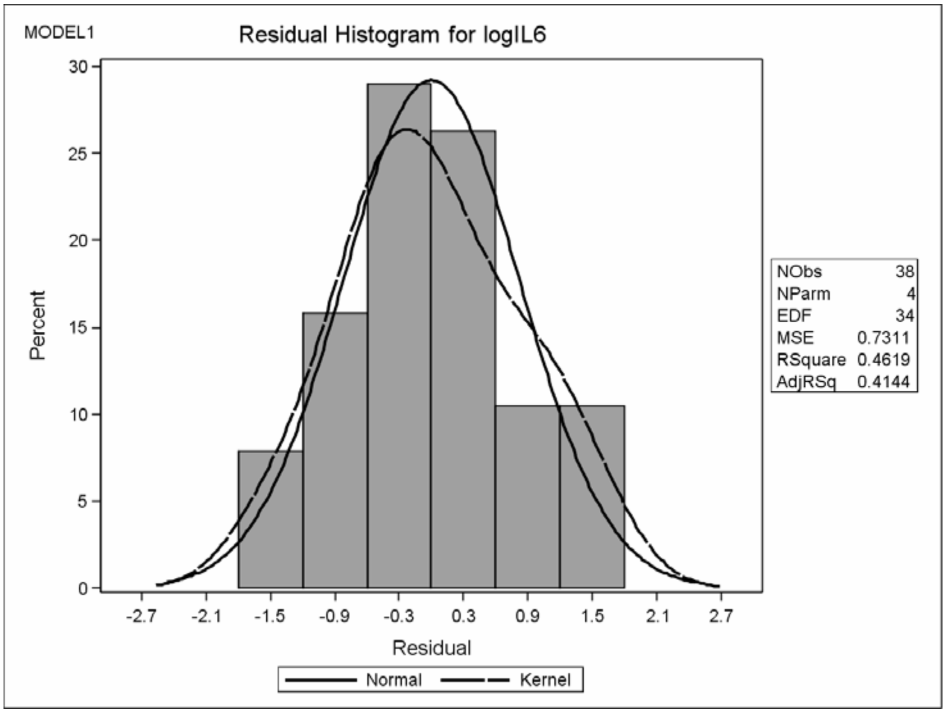
Box-Cox in SAS - output

```
     Transformation Information for BoxCox(IL6)
  Lambda      R-Square    Log Like     Lambda      R-Square    Log Like

   -3.00          0.18    -272.397      0.50           0.41     -83.921
   -2.75          0.19    -251.336      0.75           0.37     -93.169
   -2.50          0.19    -230.554      1.00           0.33    -104.491
   -2.25          0.20    -210.104      1.25           0.30    -117.272
   -2.00          0.21    -190.054      1.50           0.27    -131.151
   -1.75          0.23    -170.498      1.75           0.25    -145.908
   -1.50          0.25    -151.570      2.00           0.23    -161.395
   -1.25          0.28    -133.473      2.25           0.21    -177.498
   -1.00          0.31    -116.527      2.50           0.20    -194.129
   -0.75          0.36    -101.267      2.75           0.18    -211.214
   -0.50          0.41     -88.578      3.00           0.17    -228.689
   -0.25          0.45     -79.752
    0.00 +        0.46     -76.054 <
    0.25          0.45     -77.766 *

< - Best Lambda  * - Confidence Interval  + - Convenient Lambda
```

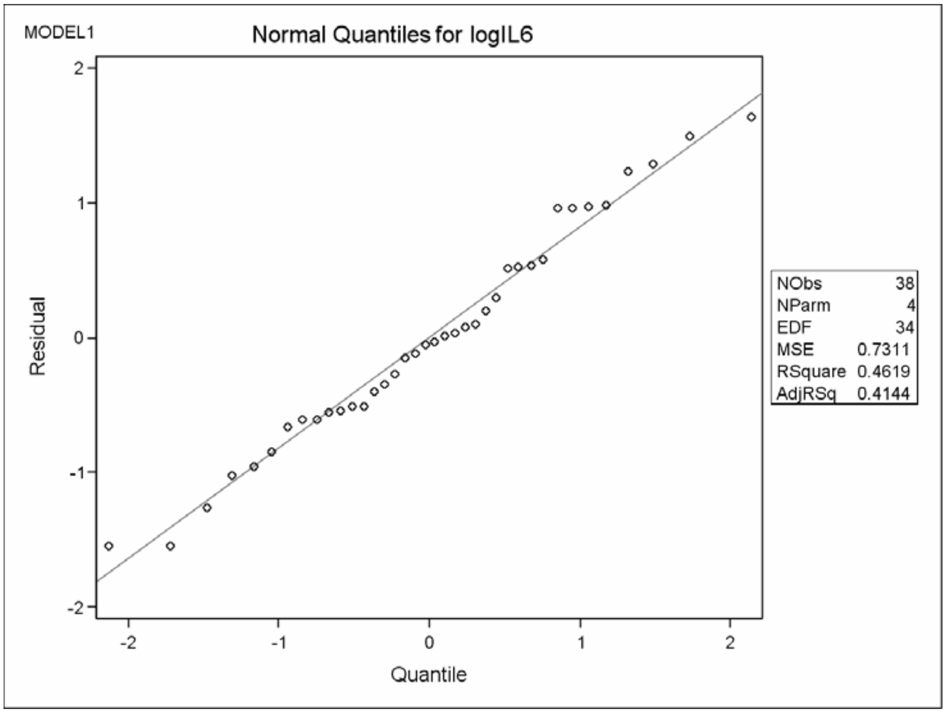
HPV/HNCa cytokine example - transformed

```
data two;
    set one;
    logIL6 = log(IL6);
run;

ods html style = Journal;
ods graphics on;
ods select ResidualHistogram;
ods select QQPlot;
ods select ResidualByPredicted;

proc reg data = two plots(unpack);
    model logIL6 = gpind1 gpind2 gpind3;
run;
quit;

ods graphics off;
ods html close;
```

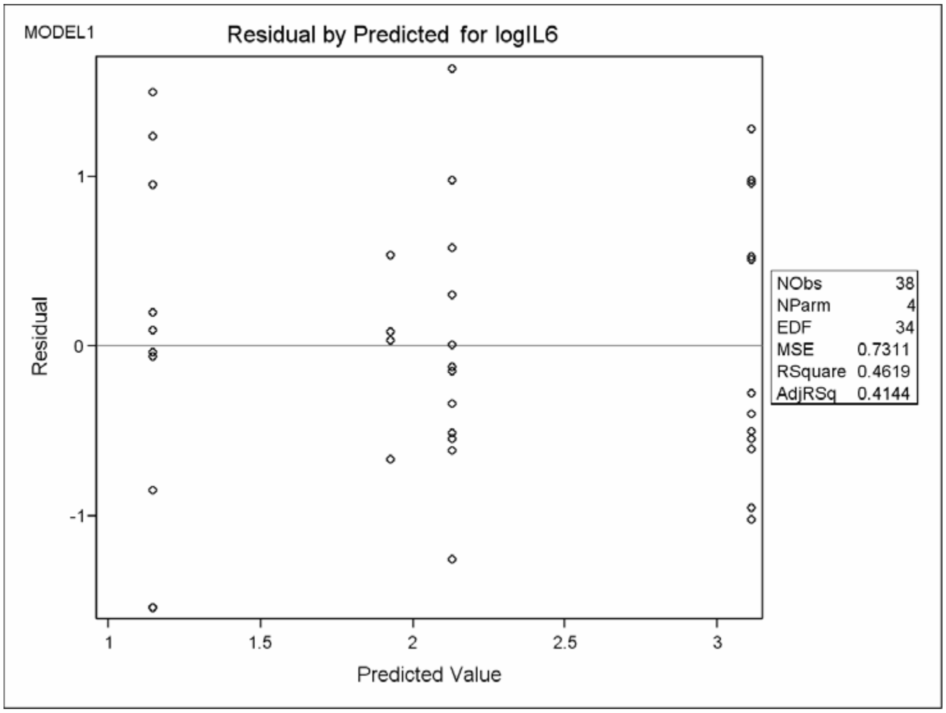
HPV/HNCa cytokine - transformed



HPV/HNCa cytokine - transformed



HPV/HNCa cytokine - transformed



Transformations on X

If you detect curvilinear behavior in a graphical assessment of Y versus X, but there are no obvious violations of homoscedasticity, then a transformation on the independent variable X may induce linearity. The graphical methods to assess that relationship are:

- scatterplots
- partial regression plots
- LOESS smoothed plots

Assessing functional form of model covariates

You can assess the functional form of model covariates by looking at *partial residual plots*. These plots give you a visual assessment of the 'shape' of the relationship between the response and an independent variable, while controlling for all other covariates in the model. (Note: In the case of SLR, just look at a scatterplot of Y versus X.)

Assessing functional form of covariates (cont.)

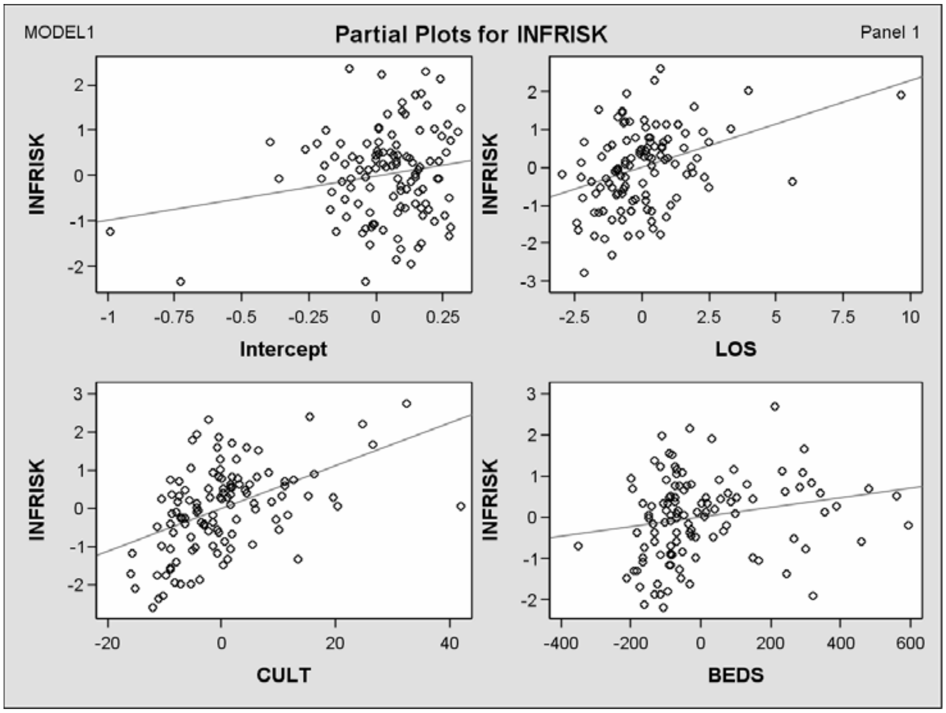
By way of example, suppose we want to assess the shape of the relationship between Y and X_1 in the presence of X_2 and X_3. Partial residual plots are constructed as follows:

1. Regress Y on X_2 and X_3. Let $\mathbf{r} = \{r_1, \ldots, r_n\}$ be the residuals from this regression. These residuals capture variation in Y after the effects of X_2 and X_3 have been removed.
2. Regress X_1 on X_2 and X_3. Let $\mathbf{r}^* = \{r_1^*, \ldots, r_n^*\}$ be the residuals from this regression. These residuals capture variation in X_1 after the effects of X_2 and X_3 have been removed.
3. Construct the scatterplot of $\mathbf{r}$ versus $\mathbf{r}^*$. This plot graphically illustrates the relationship between Y and X_1 after the effects of X_2 and X_3 have been removed.

Partial regression residual plots

```
ods html style = statistical;
ods graphics on;
ods select PartialPlotPanel1;

proc reg data = one plots(unpack);
    model infrisk = los cult beds/partial;
run;
quit;

ods graphics off;
ods html close;
```

Partial regression residual plots



LOESS

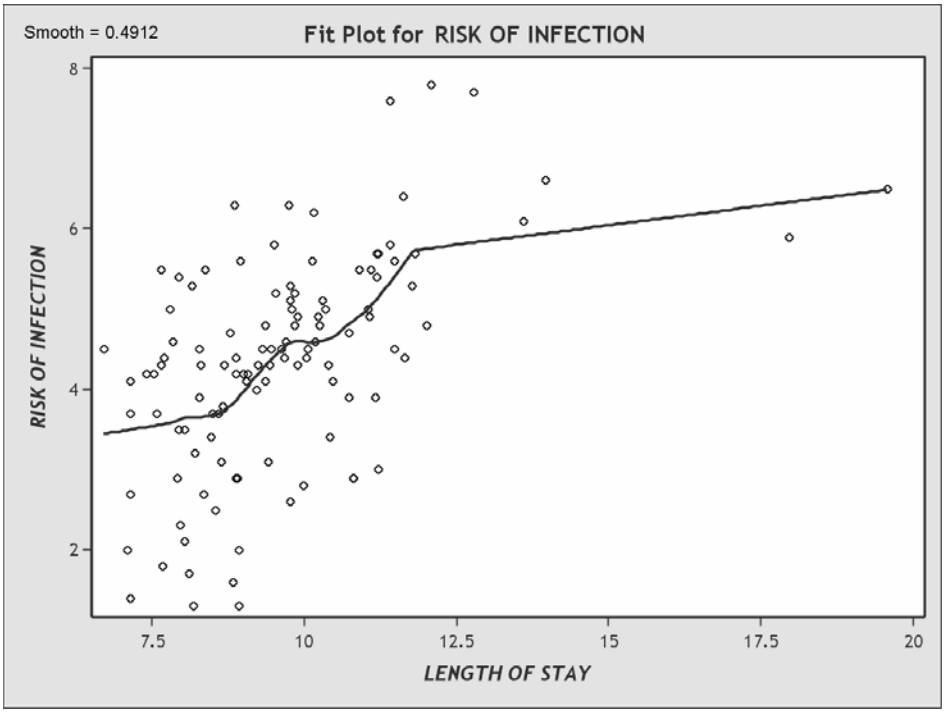
LOESS is a nonparametric smoother that helps determine the shape of the relationship between a predictor and the response. No adjustment is made for other covariates of interest. The name LOESS comes from the method by which the smoothed trend is fit - namely *locally weighted least squares*.

Segue to LOESS applets.

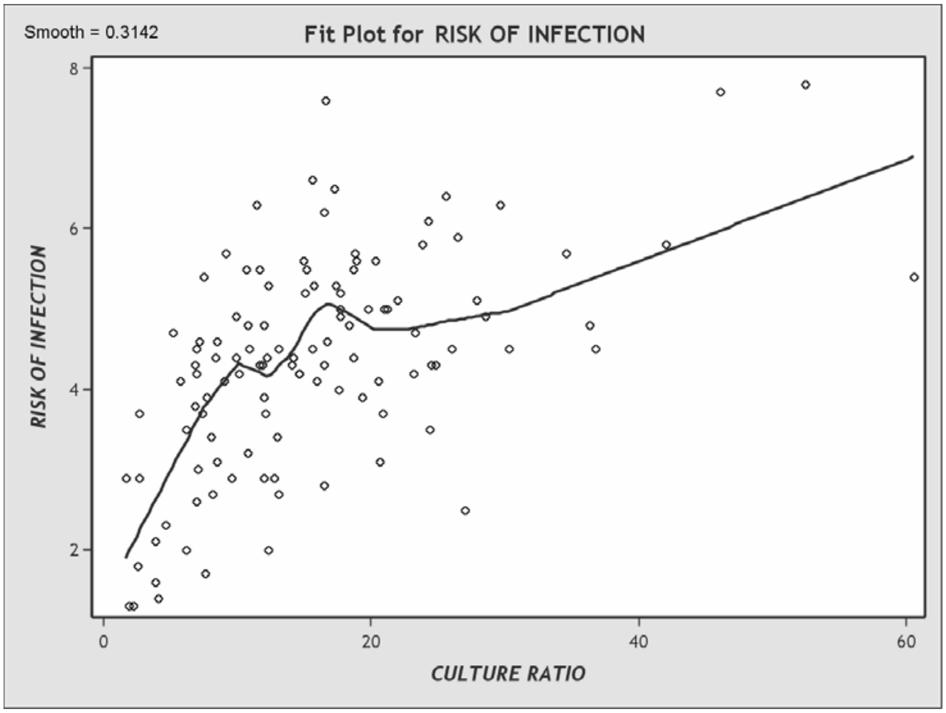
LOESS in SAS

```
ods rtf style = Journal;
ods graphics on;

proc loess data = one;
    model infrisk = los;
    *model infrisk = cult;
    *model infrisk = beds;
run;

ods graphics off;
ods rtf close;
```

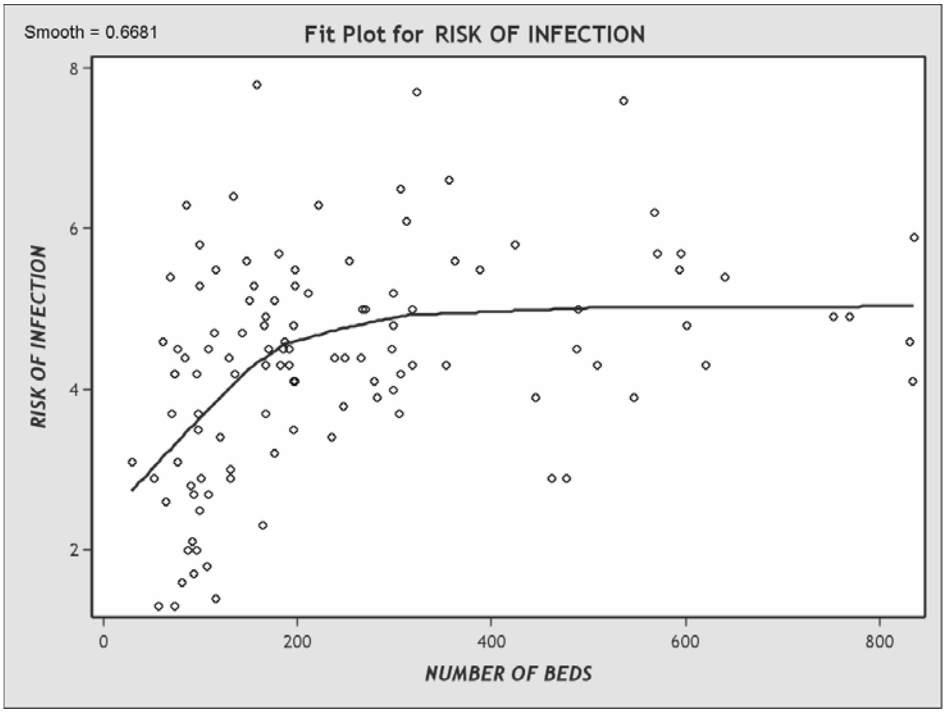
LOESS in SAS (cont.)



LOESS in SAS (cont.)



LOESS in SAS (cont.)



Fractional Polynomials

Royston and Altman (1994) proposed an ‘automated’ approach to finding flexible and interpretable transformations of covariates in regression models. We’ll begin with a SLR example. The goal is to rewrite

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (\textit{linear model})$$

as either

$$y = \beta_0 + \beta_1 F_1(x) + \varepsilon \qquad (\textit{1st order FP})$$

or

$$y = \beta_0 + \beta_1 F_1(x) + \beta_2 F_2(x) + \varepsilon \qquad (\textit{2nd order FP})$$

Fractional Polynomials (cont.)

Here are the rules that define $F_1(x)$ and $F_2(x)$:

$$F_1(x) = \begin{cases} x^{p_1} & p_1 \neq 0 \\ \log(x) & p_1 = 0 \end{cases}$$

and

$$F_2(x) = \begin{cases} x^{p_2} & p_2 \neq p_1, p_2 \neq 0 \\ \log(x) & p_2 \neq p_1, p_2 = 0 \\ F_1(x)\log(x) & p_2 = p_1 \end{cases}$$

p_1 and p_2 are selected from the set

$$\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}.$$

Fractional Polynomials examples

1. 1st order FP with $p_1 = 0$
2. 1st order FP with $p_1 = -0.5$
3. 2nd order FP with $p_1 = 0$ and $p_2 = -0.5$
4. 2nd order FP with $p_1 = 2$ and $p_2 = 2$

A small aside ... the *Likelihood*

The *likelihood* is a mathematical function that expresses the probability of the observed data as a function of the unknown parameters in a model. Fitting a statistical model to data (i.e. obtaining estimates of model parameters - βs) is done by finding the values of the parameters that maximize the likelihood. (In practice, the natural logarithm of the likelihood function is maximized or the negative of the log likelihood is minimized.) Therefore, the estimated parameters are those that maximize the probability of the observed data. Parameter estimates obtained by this method are referred to as *maximum likelihood estimates* or simply, *MLEs*.

It is easy to show that the LS solutions are equivalent to the MLEs.

Deviance and the likelihood ratio test

The deviance is defined as

$$d = -2 \text{ log-likelihood}.$$

It is a measure of how far the proposed model *deviates* from a model that fits the data perfectly (also known as a *saturated model*). The smaller the deviance, the better the fit.

Let model$_1$ and model$_2$ be nested models with model$_1$ nested within model$_2$. Let G = d(model$_1$) - d(model$_2$). Then

$$G \sim \chi^2_\nu$$

where ν is the number of parameters being tested. G is the likelihood-ratio statistic and a test constructed from G is called a likelihood ratio test.

Implementing Fractional Polynomials

1. Let $d(1)$ be the deviance for the linear model.
2. Let $d(p_1)$ be the deviance for the best fitting 1st order FP.
3. Let $d(p_1, p_2)$ be the deviance for the best fitting 2nd order FP.

Royston and Altman show that each FP contributes 2 df - one for the parameter and one for the exponent. We conduct *partial* likelihood ratio tests based on these deviances as shown on pages 5 and 6 of the SAS MACRO documentation.

Other implementations

Hosmer and Lemeshow (2000, pp. 100 - 103) suggest the following method, which is equivalent to the Royston and Altman method with no option for excluding the covariate:

1. Conduct a 3-df partial likelihood ratio test of the best 2nd order FP model versus the linear model.
   - If not significant, then model the covariate as linear.
   - If significant, then proceed to next step.
2. Conduct 2-df partial likelihood ratio test of the best 2nd order FP model versus the best 1st order FP model.
   - If not significant, then use best 1st order FP model model.
   - If significant, then choose best 2nd order FP model.

Additional precautions

Hosmer and Lemeshow (2000, p. 103) make the following cautionary statement:

“In an applied setting, we recommend that if a more complicated model is selected for use then it should provide a statistically significant improvement over the linear model, and it is vital that the transformation make clinical sense.”

Better than the linear model?

- The test between the 2nd order FP model and the linear model is a 3-df partial likelihood ratio test
- The test between the 1nd order FP model and the linear model is a 1-df partial likelihood ratio test

For the transformation of the variable BEDS, the algorithm selected the 1st order FP model with $p_1 = -0.05$, and a corresponding deviance of 305.498. The deviance of the linear model is 313.469. Then the difference in the deviances is 313.469 - 305.498 = 7.971. We compare this to a chi-square distribution with 1-df to obtain a p-value. The p-value = 0.005.

```
data test;
    pval = 1 - probchi(7.971,1);
run;
```