How to obtain those nasty standard errors from transformed data — and why they should not be used

Erik Jørgensen & Asger Roer Pedersen

Biometry Research Unit
Department of Biometry and Informatics
Danish Institute of Agricultural Sciences
P.O. Box 50, DK-8830 Tjele, Denmark

1   Introduction

When statistical analyses are carried out, it is often neccessary to transform the measured value to make the assumptions used in the statistical analyses valid.

When results are reported, the estimated values of e.g. treatment means are often transformed back to the measured or traditional scale. The advantage is that the results become readily comparable with results in similar studies.

It has become accepted practice in many scientific journals to indicate the precision of such estimates by the estimated standard error of the estimate. For several reasons a more appropriate approach is to calculate the confidence interval of the estimate, e.g. it often includes the uncertainty in the estimate of the standard error. The argument for use of the standard error is, that it is more compact, i.e. only one value need to be reported, and if the estimates follow the normal distribution, confidence intervals can easily be calculated by means of the standard error. For example a 95 percent confidence interval is approximately equal to the mean $\pm$ twice the standard error for the normal distribution. For non-normal distributions, however, symmetric confidence intervals based on the mean and the standard error can be severely misleading.

When transformations are used, the standard error calculated in standard statistical program packages is the standard error on the transformed scale and not on the original (backtransformed) scale. If the reporting of the results must conform to the practice, we therefore need to transform these standard errors to the original scale. This note present guidelines for the calculation of approximate standard errors, and techniques for evaluation of the validity of the approximation. The note is *not* concerned with when and why variables should be transformed. On these issues we refer to standard statistical textbooks such as Seber & Wild (1989) and McCullagh & Nelder (1989) for interested readers.

The proposed technique is based on some general approximate formulas for the calculation of the expectation and the variance of the transformation of a random variable or vector. In section 2 we introduce these formulas and in section 3 we apply them to the situation outlined above. The generality of the transformation formulas implies that they can be applied in many other situations. Section 4 contains some examples of this. Finally section 5 introduces more advanced techniques and provides some discussion of the proposed techniques.

2   Approximate transformation formulas

The principle used in the methods presented here is that if we know the expectation and variance of a random variable, it is easy to calculate the variance of a linear transformation of the variable. For example if $Y = a + bX$ and we know $\mathbb{E}(X)$, the expectation of X, and $\mathbb{V}(X)$, the variance of X, then $\mathbb{E}(Y) = a + b\mathbb{E}(X)$ and $\mathbb{V}(Y) = b^2\mathbb{V}(X)$.

Even though the transformations that we use in the statistical analysis are often not linear, we can approximate any (differentiable) function by a linear function in the neighbourhood of a point a by the socalled 1st order Taylor expansion, see e.g. Seber & Wild (1989):

$$Y = g(X) \approx g(a) + g'(a)(X - a) \tag{1}$$

Choosing $a = \mathbb{E}(X)$ we can calculate the expectation and variance of Y based on this approximation:

$$\mathbb{E}(Y) \approx g(\mathbb{E}(X)) \tag{2}$$
$$\mathbb{V}(Y) \approx g'(\mathbb{E}(X))^2 \mathbb{V}(X) \tag{3}$$

Since the accuracy of the approximation is highest when X is close to $a = \mathbb{E}(X)$, the accuracy of (2) and (3) is highest for small values of $\mathbb{V}(X)$. Formulas (2)–(3) assume that X and Y are one-dimensional. When X and Y are multi-dimensional, e.g. X is p-dimensional and Y is q-dimensional, then the corresponding versions of (2)–(3) are:

$$\mathbb{E}(Y) \approx g(\mathbb{E}(X)) \tag{4}$$
$$\mathbb{V}(Y) \approx g'(\mathbb{E}(X))\mathbb{V}(X)g'(\mathbb{E}(X))^\top \tag{5}$$

Here $^\top$ denotes matrixtransposition, e.g.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}^\top = (x, y, z)$$

and g' denotes the $q \times p$-matrix of partial derivatives of g, i.e.

$$g'(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \cdots & \frac{\partial g_1}{\partial x_p}(x) \\ \vdots & & \vdots \\ \frac{\partial g_i}{\partial x_1}(x) & \cdots & \frac{\partial g_i}{\partial x_p}(x) \\ \vdots & & \vdots \\ \frac{\partial g_q}{\partial x_1}(x) & \cdots & \frac{\partial g_q}{\partial x_p}(x) \end{pmatrix}$$

where

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_q(x) \end{pmatrix}$$

3 Analysis of transformed data

Suppose some statistical analysis is performed on a transformed scale where the data has been transformed by the function g. On this transformed scale let Y denote an estimator of some quantity of interest, θ, and let us suppose that Y is unbiased, i.e. $\mathbb{E}(Y) = \theta$. Here θ may for instance be some treatment level possibly plus some average level for other effects, and Y may correspondingly be some mean or least squares mean of transformed data. An actual calculated value of Y will be denoted by $\hat{Y}$, and similarly, an estimated value of $\sigma_Y^2 = \mathbb{V}(Y)$ will be denoted by $\hat{\sigma}_Y^2$. A $(1-\alpha)$-confidence interval for θ is given by two random variables, Y_{low} and Y_{high}, for which the probability that θ belongs to the interval $[Y_{low}, Y_{high}]$ is $100(1-\alpha)\%$. Again actual calculated values of Y_{low} and Y_{high} are denoted by $\hat{Y}_{low}$ and $\hat{Y}_{high}$, respectively. In many classical linear and linear mixed normal models confidence limits with $100(1-\alpha)\%$ coverage, e.g. $\alpha = 0.05$ gives 95% coverage, are calculated as

$$\begin{aligned} \hat{Y}_{low} &= \hat{Y} - t_{(1-\alpha)/2}(d)\hat{\sigma}_Y \\ \hat{Y}_{high} &= \hat{Y} + t_{(1-\alpha)/2}(d)\hat{\sigma}_Y \end{aligned}$$

where $t_{(1-\alpha)/2}(d)$ denotes the $(1-\alpha)/2$-quantile of the t-distribution with d degrees of freedom, and $\hat{\sigma}_Y$ is a residual mean square type estimate of σ_Y based on d degrees of freedom.

Now it is common practice to backtransform Y, i.e. $X = g^{-1}(Y)$, where g^{-1} denotes the inverse function[1] of g. Recognizing the random variation in Y, this may not always be the most natural thing to do. The subject is further discussed in section 5, but for now we shall merely follow the practice.

In the following we need to 'reverse' the formulas (2)–(3), i.e. we need to apply the formulas to the backtransformation $X = g^{-1}(Y)$. The resulting formulas are:

$$\begin{aligned} \mathbb{E}(X) &\approx g^{-1}(\mathbb{E}(Y)) & (6) \\ \mathbb{V}(X) &\approx \frac{\mathbb{V}(Y)}{[g'(g^{-1}(\mathbb{E}(Y)))]^2} & (7) \end{aligned}$$

According to formula (6) we have that $\mathbb{E}(X) \approx g^{-1}(\theta)$. Hence X may be regarded as an (in general biased) estimator of $g^{-1}(\theta)$. Actual calculated values of X will be denoted by $\tilde{X}$, i.e. $\tilde{X} = g^{-1}(\hat{Y})$. The confidence limits on the transformed scale may also be backtransformed, i.e. if g is strictly increasing then

$$\begin{aligned} \hat{X}_{low} &= g^{-1}(\hat{Y}_{low}) \\ \hat{X}_{high} &= g^{-1}(\hat{Y}_{high}) \end{aligned}$$

---

[1]The inverse function of g exists if and only if g is either strictly increasing or strictly decreasing. In such cases g^{-1} is characterized by: $y = g(x)$ if and only if $x = g^{-1}(y)$. In particular, $g(g^{-1}(y)) = y$ and $g^{-1}(g(x)) = x$. The inverse function should not be confused with the reciprocal function $g(x)^{-1} = 1/g(x)$.

and if g is strictly decreasing then

$$\hat{X}_{low} = g^{-1}(\hat{Y}_{high})$$
$$\hat{X}_{high} = g^{-1}(\hat{Y}_{low})$$

If $[\hat{Y}_{low}, \hat{Y}_{high}]$ is a $(1 - \alpha)$-confidence interval for θ then $[\hat{X}_{low}, \hat{X}_{high}]$ is a $(1 - \alpha)$-confidence interval for $g^{-1}(\theta)$. Thus backtransformed confidence intervals maintain exactly the coverage percentage chosen on the transformed scale.

According to formula (7) we have that $\sigma_X^2 = \mathbb{V}(X) \approx \sigma_Y^2 / [g'(g^{-1}(\theta))]^2$. Inserting the estimates of θ and σ_Y^2 we find that the standard error of $\tilde{X}$ may be approximated by $\tilde{\sigma}_X = \hat{\sigma}_Y / |g'(\tilde{X})|$.

In summary, on the original (backtransformed) scale the estimate and the corresponding standard error may be approximated by:

$$\tilde{X} = g^{-1}(\hat{Y}) \tag{8}$$
$$\tilde{\sigma}_X = \frac{\hat{\sigma}_Y}{|g'(\tilde{X})|} \tag{9}$$

If a $(1 - \alpha)$-confidence interval, $[\hat{Y}_{low}, \hat{Y}_{high}]$, for θ has been calculated as

$$\hat{Y}_{low} = \hat{Y} - t_{(1-\alpha)/2}(d)\hat{\sigma}_Y$$
$$\hat{Y}_{high} = \hat{Y} + t_{(1-\alpha)/2}(d)\hat{\sigma}_Y$$

one might be tempted to calculate confidence limits for $g^{-1}(\theta)$ as

$$\tilde{X}_{low} = \tilde{X} - t_{(1-\alpha)/2}(d)\tilde{\sigma}_X$$
$$\tilde{X}_{high} = \tilde{X} + t_{(1-\alpha)/2}(d)\tilde{\sigma}_X$$

In this way estimation results may be presented more compactly in that only $\tilde{X}$ and $\tilde{\sigma}_X$ need to be reported. However, as far as the authors know there is no theoretical justification for doing so. The coverage percentage of the confidence interval $[\tilde{X}_{low}, \tilde{X}_{high}]$ is unknown, and in fact it is not clear at all if $t_{(1-\alpha)/2}(d)$ is a reasonable scaling constant. In general one only knows[2] that if $t_{(1-\alpha)/2}(d)$ is replaced by 2 then the coverage percentage is at least 75%. If $\hat{Y}$ is the maximum likelihood estimate (MLE) of θ then $g^{-1}(\hat{Y})$ is (usually) the MLE of $g^{-1}(\theta)$. This means that for large sample sizes $t_{(1-\alpha)/2}(d)$ may be replaced by 1.96 giving a coverage of approximately 95 %. Finally, the confidence interval $[\tilde{X}_{low}, \tilde{X}_{high}]$ is symmetric, in contrast to $[\hat{X}_{low}, \hat{X}_{high}]$ that may be severely non-symmetric for highly nonlinear transformations, g.

If, however, the standard deviation of Y, i.e. σ_Y, is sufficiently small and/or the transformation g is approximately linear, then the confidence intervals $[\tilde{X}_{low}, \tilde{X}_{high}]$ and $[\hat{X}_{low}, \hat{X}_{high}]$ may be very similar, and in such cases one can of course report any of the two. The easiest way to check if this is the case is of course to calculate both and compare. If the results differs widely, it cannot be recommended to report values using standard errors on the backtransformed scale.

---

[2]This general statement follows from Markov's inequality, but in fact it is only approximately true since it presumes that $\tilde{\sigma}_X$ is the true standard deviation of $\tilde{X}$.

Instead one should use the confidence intervals based on the transformed scale, i.e. $[\hat{X}_{low}, \hat{X}_{high}]$.
If the journal referees do not accept this, it should be possible to convince them by showing
them both confidence intervals.

As an illustration consider figure 1 where the logit transform is shown together with two linear
approximation functions, one around 0.85 on the logit-scale (i.e. corresponding to Y) or $X = 0.70$
and another one around 2.95 on the logit (Y) scale corresponding to $X = 0.95$.
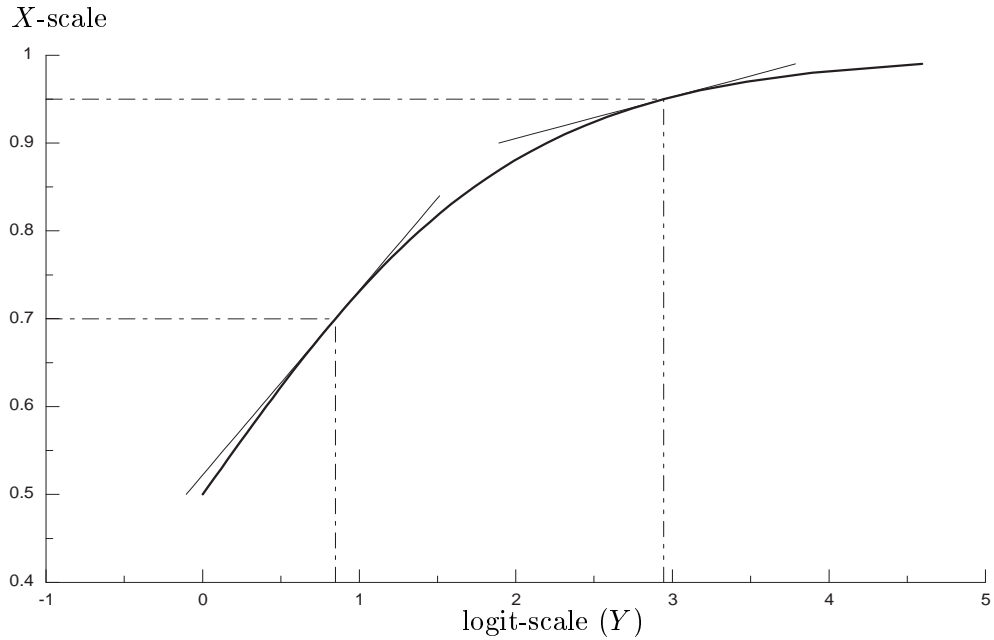

Figure 1: The 1st order Taylor approximation applied to the logit transform.

As the figure illustrates the accuracy of the linear approximation depends on the curvature of
the transformation function and on the distance to the point around which the linearization is
done. Around 0.85 minor differences in the logit scale will be reflected in large differences on
the back-transformed scale, whereas around 2.95 the same differences will result in much lower
differences on the backtransformed scale.

If we consider an example with relatively low standard error (0.20 on the logit scale), we obtain
the (unnormalized) distributions shown in figure 2 on the logit scale and on the original scale.
The back transformations results in widely different distributions on the original scale.
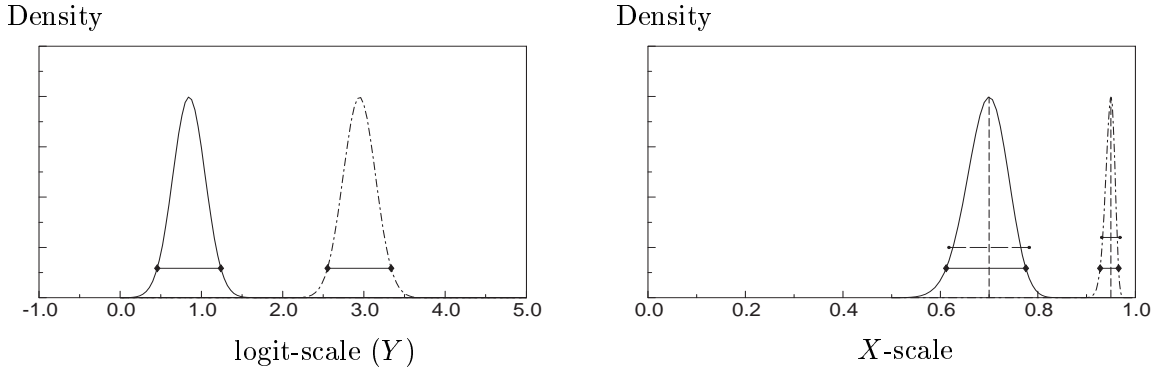
Figure 2: Distribution (unnormalized) on logit and backtransformed scale, respectively. Low standard error. The horisontal lines indicates the confidence intervals. ——— Exact interval $[\hat{X}_{low}, \hat{X}_{high}]$, − − − − Approximate interval, $[\tilde{X}_{low}, \tilde{X}_{high}]$.

If we calculate the 95%-confidence interval for the low mean (0.85), we obtain $[\hat{X}_{low}, \hat{X}_{high}] = [0.612, 0.775]$ based on the calculations on the transformed (logit) scale vs. $[\tilde{X}_{low}, \tilde{X}_{high}] = [0.618, 0.782]$ based on the calculations using the approximate standard error. Correspondingly for the high mean (2.94), we obtain $[\hat{X}_{low}, \hat{X}_{high}] = [0.928, 0.969]$ vs. $[\tilde{X}_{low}, \tilde{X}_{high}] = [0.931, 0.969]$, i.e. the approximations have performed adequately.

If we turn to an example with high standard error (0.50) on the logit-scale we obtain the results shown in figure 3.
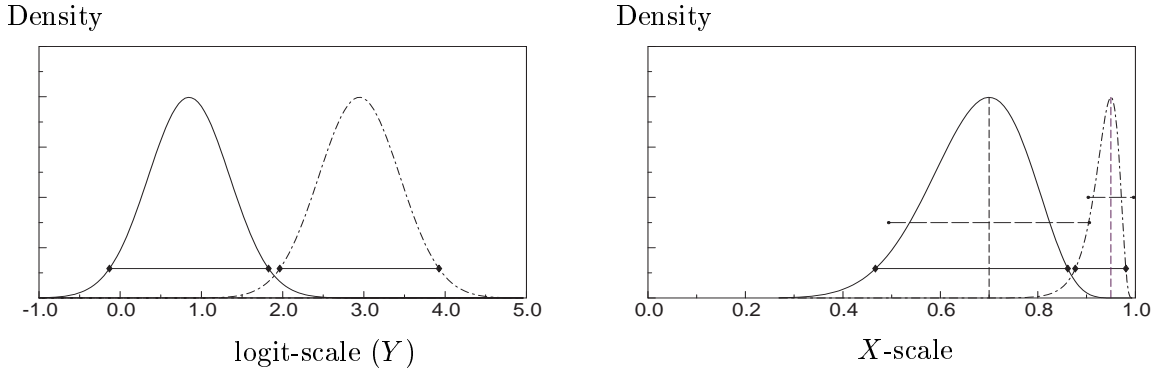

Figure 3: Distribution (unnormalized) on logit and backtransformed scale, respectively. High standard error. The horisontal lines indicates the confidence intervals. ——— Exact interval $[\hat{X}_{low}, \hat{X}_{high}]$, − − − − Approximate interval, $[\tilde{X}_{low}, \tilde{X}_{high}]$.

Again, the distributions differs markedly. The exact confidence interval at the low level is $[\hat{X}_{low}, \hat{X}_{high}] = [0.467, 0.861]$ vs. $[\tilde{X}_{low}, \tilde{X}_{high}] = [0.494, 0.906]$ by the approximate method. At the high level $[\hat{X}_{low}, \hat{X}_{high}] = [0.877, 0.981]$ vs. $[\tilde{X}_{low}, \tilde{X}_{high}] = [0.903, 0.997]$. Note that the approximated intervals differs markedly from the exact. The main reason is that the back-transformed distributions are skewed, and not symmetrical around the mode. The length of the confidence intervals (and the standard error) are almost identical. The high limit in the approximate confidence interval is very close to 1, and may exceed 1, despite the fact that the values on the X-scale are restricted to the interval $]0, 1[$.

6

3.1 Examples

Above we considered a general approach to backtransformations. In the present section we will present examples on how to proceed with backtransformations for different concrete transformation functions. An overviev over these transformations is given in table 1. The first example, the logarithmic transform will be treated in detail, while the others will only be shown summarily.

Table 1: Overview over transforms

Name	Domain	$g(x)$	$g^{-1}(y)$	$g'(x)$
logarithmic (natural)	$x > 0$	$\ln(x)$	$\exp(y)$	$1/x$
logarithmic (base 10)	$x > 0$	$\log_{10}(x)$	$\exp(y\ln(10))$	$1/(x\ln(10))$
logit	$0 < x < 1$	$\ln(x/(1-x))$	$\exp(y)/(1+\exp(y))$	$1/(x(1-x))$
square root	$x > 0$	$\sqrt{x}$	y^2	$1/(2\sqrt{x})$
reciprocal	$x \neq 0$	$1/x$	$1/y$	$-1/x^2$
inverse squared sine	$0 < x < 1$	$\sin^{-1}(\sqrt{x})$	$(\sin(y))^2$	$1/(2\sqrt{x}\cos(\sin^{-1}(\sqrt{x})))$
power(a)	$x > 0, a \neq 0$	x^a	$y^{1/a}$	ax^{a-1}

3.1.1 Logarithmic (natural) transformation

One of the most frequent transformations used, is the logarithmic transformation, i.e.

$$Y_i = \ln(X_i)$$

From table 1 we see that if we differentiate $\ln(x)$ we obtain:

$$\ln'(x) = \frac{d\ln(x)}{dx} = \frac{1}{x}$$

and from formulas (8)–(9) and we see that

$$\begin{aligned} \tilde{X} &= \exp(\hat{Y}) \\ \tilde{\sigma}_X &= \tilde{X}\hat{\sigma}_Y \end{aligned}$$

In a (fictive) experiment comparing four treatments the least squares mean for treatment 1 is 1.08 with standard error 0.042. These values may for instance be obtained using a `LSMEANS TREAT /STDERR` statement in the SAS procedure GLM, see SAS Institute Inc. (1989). Using the formulas above we obtain

$$\begin{aligned} \tilde{X} &= \exp(1.08) = 2.95 \\ \tilde{\sigma}_X &= 2.95 \times 0.042 = 0.125 \end{aligned}$$

The results for the other combinations are shown in table 2.

If we want to evaluate the approximation we proceed by calculating both types of confidence intervals. On the transformed scale:

$$\begin{aligned} \hat{Y}_{low} &= \hat{Y} - t_{d,\alpha/2}\hat{\sigma}_Y = 1.08 - 2.03 \times 0.042 = 0.995 \\ \hat{Y}_{high} &= \hat{Y} + t_{d,\alpha/2}\hat{\sigma}_Y = 1.08 + 2.03 \times 0.042 = 1.165 \end{aligned}$$

7

Table 2: Examples of results from an experiment

Treat-ment	N	Transformed scale		Original scale	
		LSmean $(\hat{Y})$	Std.Err.$(\hat{\sigma}_Y)$	$\tilde{X}$	$\tilde{\sigma}_X$
1	10	1.08	0.042	2.94	0.125
2	9	1.51	0.045	4.53	0.202
3	10	1.27	0.042	3.56	0.151
4	8	1.03	0.047	2.80	0.133

(The degrees of freedom in this experiment with $(10+9+10+8) = 37$ experimental units is the degress of freedom in the estimation of the residual mean square, i.e. 33, so the t-value $t_{33,0.975} = 2.03$ has been used. This corresponds to $\left[\exp(\hat{Y}_{low}), \exp(\hat{Y}_{high})\right] = [\exp(0.995), \exp(1.165)] = [2.705, 3.205]$ on the original scale. The approximate method yields

$$\tilde{X}_{low} = \tilde{X} - t_{d,\alpha/2}\tilde{\sigma}_X = 2.94 - 2.03 \times 0.125 = 2.686$$
$$\tilde{X}_{high} = \tilde{X} + t_{d,\alpha/2}\tilde{\sigma}_X = 2.94 + 2.03 \times 0.125 = 3.194$$

The two confidence intervals are thus in close correspondance, and the approximation can be used without problems. In table 3 the results of the comparison between confidence intervals are shown for all four treatments. All confidence intervals are comparable.

Table 3: Comparison between confidence intervals for the results in table 2

Treat-ment	Transformed scale		Original scale, exact		Original scale, approx.	
	$\hat{Y}_{low}$	$\hat{Y}_{high}$	$\hat{X}_{low}$	$\hat{X}_{high}$	$\tilde{X}_{low}$	$\tilde{X}_{high}$
1	0.995	1.165	2.704	3.207	2.686	3.194
2	1.419	1.601	4.132	4.960	4.120	4.940
3	1.185	1.355	3.270	3.878	3.253	3.867
4	0.935	1.125	2.546	3.081	2.530	3.070

If the standard errors of the treatment means were larger we would obtain different results. In table 4 and 5 results are shown with the standard errors 10 times as high. Note that the confidence intervals in table 5 differs much between the exact and the approximate method. In this case the approximate method should not be used and the exact confidence intervals should be reported.

Table 4: Examples of results from an experiment, as table 2 but with higher standard errors

Treat-ment	N	Transformed scale		Original scale	
		LSmean $(\hat{Y})$	Std.Err.$(\hat{\sigma}_Y)$	$\tilde{X}$	$\tilde{\sigma}_X$
1	10	1.08	0.424	2.945	1.248
2	9	1.51	0.447	4.527	2.022
3	10	1.27	0.424	3.561	1.509
4	8	1.03	0.478	2.801	1.327

Table 5: Comparison between confidence intervals for the results in table 4, i.e. high standard error

Treat- ment	Transformed scale		Original scale, exact		Original scale, approx.	
	$\hat{Y}_{low}$	$\hat{Y}_{high}$	$\hat{X}_{low}$	$\hat{X}_{high}$	$\tilde{X}_{low}$	$\tilde{X}_{high}$
1	0.219	1.941	1.245	6.964	0.412	5.478
2	0.602	2.417	1.827	11.217	0.422	8.632
3	0.409	2.131	1.506	8.421	0.498	6.624
4	0.060	2.000	1.061	7.392	0.107	5.495

3.1.2   Logarithmic (base 10) transformation

This is essentially the same as the natural logarithmic transformation since

$$\log_{10}(x) = \ln(x)/\ln(10)$$

and from formulas (8)–(9) and we see that

$$\tilde{X} = \exp(\hat{Y}\ln(10))$$
$$\tilde{\sigma}_X = \ln(10)\tilde{X}\hat{\sigma}_Y$$

3.1.3   Logit transformation

The logit transformation is much used when analysing variables that are proportions, i.e. the observed values are between 0 and 1. The logit transform of such data, i.e.

$$Y_i = \text{logit}(X_i) = \ln\{X_i/(1 - X_i)\}$$

produces data that may take on any real value, and so they may possibly be analyzed by means of the normal distribution (classical statistical analysis). From table 1 we see that if we differentiate $\text{logit}(x)$ we obtain:

$$\text{logit}'(x) = \frac{1}{x(1 - x)}$$

and from formulas (8)–(9) we see that

$$\tilde{X} = \frac{\exp(\hat{Y})}{1 + \exp(\hat{Y})}$$
$$\tilde{\sigma}_X = \tilde{X}(1 - \tilde{X})\hat{\sigma}_Y$$

3.1.4   Square root transformation

The classsical use of the square root transformation is for variables that follow the Poisson distribution, e.g. counts of number of events in a period, where the events occur with relatively low frequency. Applying the square root transformation to such data, i.e.

$$Y_i = \sqrt{X_i}$$

may imply that the transformed data can be analyzed by means of the normal distribution. From table 1 we see that if we differentiate $\sqrt{x}$ we obtain:

$$(\sqrt{x})' = \frac{1}{2\sqrt{x}}$$

and from formulas (8)–(9) we see that

$$\tilde{X} = Y^2$$
$$\tilde{\sigma}_X = 2\sqrt{\tilde{X}}\hat{\sigma}_Y$$

3.1.5 Reciprocal transformation

In some cases the reciprocal can be used as a variance stabilization transformation. In other cases the reciprocal transformation may be used due to interpretation. For some of the traditional measures used in agriculture it can be discussed whether to use the original value or its reciprocal transforms, e.g. *daily growth rate* or *days to harvest/slaughter*, efficiency of feed utilisation measured either as *feed units per kg gain* or as *kg gain per feed unit*. The reciprocal transformation is

$$Y_i = 1/X_i$$

From table 1 we see that if we differentiate $1/x$ we obtain:

$$(\frac{1}{x})' = -\frac{1}{x^2}$$

and from formulas (8)–(9) we see that

$$\tilde{X} = \frac{1}{\hat{Y}}$$
$$\tilde{\sigma}_X = \tilde{X}^2 \hat{\sigma}_Y$$

Note that the absolute vale of $g'(x)$ is used.

3.1.6 Inverse squared sine transformation

The inverse squared sine transformation is often used as a variance stabilization transformation for relative frequency data, e.g. the number of successes divided by the number of trials. It is often called the arcsin transformation. The transformation is

$$Y_i = \sin^{-1}(\sqrt{X_i})$$

The inverse sine of a number is an angle *or* minus the same angle. Here $\sin^{-1}$ denotes the positive angle. From table 1 we see that if we differentiate $\sin^{-1}(\sqrt{x})$ we obtain:

$$(\sin^{-1}(\sqrt{x}))' = 1/(2\sqrt{x}\cos(\sin^{-1}(\sqrt{x})))$$

and from formulas (8)–(9) we see that

$$\tilde{X} = (\sin(\hat{Y}))^2$$
$$\tilde{\sigma}_X = 2\hat{\sigma}_Y\sqrt{\tilde{X}}\cos(\sin^{-1}(\sqrt{\tilde{X}}))$$

3.1.7 The power transformation

The power transformation is actually a general description of a series of different transformations, including the inverse and square root transformations mentioned in the previous sections. The difference is the value of a in the expression:

$$Y_i = X_i^a$$

The power transform is often used as an variance stabilization transformation, in which case the power transformations are also called Box-Cox transformations, see also section 5 and Seber & Wild (1989). From table 1 we see that if we differentiate x^a we obtain:

$$(x^a)' = ax^{a-1}$$

and from formulas (8)–(9) we see that

$$
\begin{aligned}
\tilde{X} &= \hat{Y}^{1/a} \\
\tilde{\sigma}_X &= \hat{\sigma}_Y \frac{1}{a\tilde{X}^{a-1}}
\end{aligned}
$$

4 Applications of the approximate transformation formulas

The general nature of the approximate formulas (4)–(5) implies that they can be applied in many contexts. In this section we briefly introduce a few of these with particular emphasis on the problem of calculating confidence intervals for the root of a first order polynomial.

4.1 The root of a first order polynomial

In many statistical applications the root of a first order polynomial is of interest. If for instance a linear relationship

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

has been established between some controllable covariate x and some observable response Y, the 'dose' or value of x for which the response is zero, i.e. the root of the first order polynomial $\alpha + \beta x$, may have some interesting scientific interpretation. In the equation above the ε's are random error terms reflecting the random variation in the response variable. Another example is when an experiment has been performed using increasing levels of an input factor x. The response value Y is measured and a quadratic regression model fitted to the data, i.e.

$$Y_i = f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where again the ε's represent random errors. If we differentiate f we see that

$$f'(x) = \beta_1 + 2\beta_2 x = 0$$

Assuming that $\beta_2 > 0$ we see that the optimal input level is the solution to the equation

$$\beta_1 + 2\beta_2 x = 0$$

i.e. the root of the first order polynomial $\beta_1 + 2\beta_2 x$.

The root of the polynomial $\alpha + \beta x$ is of course given by $-\alpha/\beta$, provided $\beta \neq 0$, but, as in both of the examples above, the values of α and β are not known exactly but with uncertainty, i.e. they have been estimated from data. This means that the root is also only known with uncertainty, and the problem is to calculate the standard error of the (estimated) root or maybe to calculate a confidence interval for the root. Two solutions to this problem are given in the following.

4.1.1   An approximate solution

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the (maximum likelihood) estimates of α and β, respectively, and let

$$\hat{\Sigma} = \left(\begin{array}{cc} \hat{\sigma}_\alpha^2 & \hat{\sigma}_{\alpha\beta} \\ \hat{\sigma}_{\alpha\beta} & \hat{\sigma}_\beta^2 \end{array} \right)$$

denote the corresponding estimated (2 by 2) covariance matrix for the estimators. Then $\hat{x} = -\hat{\alpha}/\hat{\beta}$ is the corresponding (maximum likelihood) estimate of the root of the equation: $\alpha + \beta x = 0$. Hence $\hat{x}$ can be written as $\hat{x} = g(\hat{\alpha}, \hat{\beta})$, where the function g is given by $g(\alpha, \beta) = -\alpha/\beta$. Applying formulas (4)–(5) we see that the standard error of $\hat{x}$ can be approximated by

$$\tilde{\sigma}_x \approx \sqrt{\frac{1}{\hat{\beta}^2} \left(\hat{\sigma}_\alpha^2 - 2\hat{\sigma}_{\alpha\beta} \frac{\hat{\alpha}}{\hat{\beta}} + \hat{\sigma}_\beta^2 \frac{\hat{\alpha}^2}{\hat{\beta}^2} \right)}$$

Since $\hat{x}$ is a maximum likelihood estimate[3] it follows from standard asymptotic theory, that for large sample sizes that the limits

$$\hat{x} \pm 1.96 \tilde{\sigma}_x$$

constitute an approximate 95%-confidence interval for the root.

4.1.2   An exact method

Under certain normality assumptions an exact confidence interval for the root of the polynomial $\alpha + \beta x$ can be calculated. Suppose that

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \ i = 1, \ldots, n$$

where the ε_i's are independent and $\mathcal{N}(0, \sigma^2)$-distributed random error terms. Then α and β can be estimated by standard techniques (maximum likelihood estimation), see e.g. McCullagh & Nelder (1989), and the root can be estimated by $\hat{x} = -\hat{\alpha}/\hat{\beta}$. According to standard methods for linear normal models, an exact 95%-confidence interval for the root is given by those values of x_0 for which

$$(n-2) \left(\frac{S_Y^2}{n} + \frac{SPD_{xY}^2}{SSD_x} - \frac{(SP_{xY} - x_0 S_Y)^2}{SS_x + n x_0^2 - 2 x_0 S_x} \right) \Big/ \left(SSD_Y - \frac{SPD_{xY}^2}{SSD_x} \right) < F_{95\%}(1, n-2) \tag{10}$$

---

[3]This is almost always the case when standard statistical software packages are used.

where $F_{95\%}(1, n-2)$ denotes the 95%-quantile of the $F(1, n-2)$-distribution, and

$$
\begin{aligned}
S_Y &= \sum_{i=1}^{n} Y_i \\
SS_Y &= \sum_{i=1}^{n} Y_i^2 \\
SSD_Y &= SS_Y - \frac{S_Y^2}{n} \\[2mm]
S_x &= \sum_{i=1}^{n} x_i \\
SS_x &= \sum_{i=1}^{n} x_i^2 \\
SSD_x &= SS_x - \frac{S_x^2}{n} \\[2mm]
SP_{xY} &= \sum_{i=1}^{n} x_i Y_i \\
SPD_{xY} &= SP_{xY} - \frac{S_x S_Y}{n}
\end{aligned}
$$

To obtain a $(1-\alpha)$-confidence interval, $F_{95\%}(1, n-2)$ must be replaced by $F_{100(1-\alpha)\%}(1, n-2)$ in inequality (10). Now let

$$
\begin{aligned}
F &= F_{95\%}(1, n-2) \\
N &= \left(SSD_Y - \frac{SPD_{xY}^2}{SSD_x} \right) \Big/ (n-2) \\
T &= \frac{S_Y^2}{n} + \frac{SPD_{xY}^2}{SSD_x}
\end{aligned}
$$

and

$$
\begin{aligned}
A &= n(T - FN) - S_Y^2 \\
B &= 2\left[S_Y SP_{xY} - S_x(T - FN) \right] \\
C &= (T - FN)SS_x - SP_{xY}^2
\end{aligned}
$$

Then the 95%-confidence interval characterized by inequality (10) can equivalently be characterized as consisting of those values of x_0 for which

$$
A x_0^2 + B x_0 + C < 0 \tag{11}
$$

The 95%-confidence interval, CI, defined by (11) depends significantly on the values of A, B and C. To summarize the various forms that CI may have put $D = B^2 - 4AC$. Furthermore, when $A \neq 0$ and $D \geq 0$ let x_* denote the smaller and let x^* denote the larger of the two roots:

$$
\frac{-B - \sqrt{D}}{2A}, \ \frac{-B + \sqrt{D}}{2A}
$$

The various forms that CI can have are summarized in table 6, where $\emptyset$ denotes the empty set and R denotes the set of all real values.

13

Table 6: The various forms of the 95%-confidence interval defined by inequality (11)

	$D \leq 0$	$D > 0$
$A < 0$	$CI = \begin{cases} \varnothing & , \quad \text{if } C > 0 \\ R & , \quad \text{if } C \leq 0 \end{cases}$	$CI = (-\infty, x_*] \cup [x^*, +\infty)$
$A = 0$	$CI = \begin{cases} \varnothing & , \quad \text{if } C \geq 0 \\ R & , \quad \text{if } C < 0 \end{cases}$	$CI = \begin{cases} (-\infty, -C/B] & , \quad \text{if } B > 0 \\ [-C/B, +\infty) & , \quad \text{if } B < 0 \end{cases}$
$A > 0$	$CI = \begin{cases} \varnothing & , \quad \text{if } C \geq 0 \\ R & , \quad \text{if } C < 0 \end{cases}$	$CI = [x_*, x^*]$

4.2   Nonlinear regression

Suppose that the relationship between a covariate x and a response Y is modeled by some nonlinear function f, i.e.

$$Y_i = f(x_i; \theta) + \varepsilon_i$$

where the ε's are random error terms and θ denotes a vector of unknown parameters that must be estimated from data. Suppose also that θ can be estimated by $\hat{\theta}$ with estimated covariance matrix $\hat{\Sigma}$. Then $f(x; \theta)$ may be estimated by $f(x; \hat{\theta})$ for any value of x of interest and the corresponding standard error may be approximated by

$$\sqrt{\mathbb{V}(f(x; \hat{\theta}))} \approx f'(x; \hat{\theta}) \hat{\Sigma} f'(x; \hat{\theta})^\top$$

4.3   In general

In any statistical model where an unknown vector of parameters θ can be estimated by some estimator $\hat{\theta}$ for which an estimator $\hat{\Sigma}$ of its covariance matrix is available, we may according to formulas (4)–(5) estimate $g(\theta)$ by $g(\hat{\theta})$ for any differentiable function g, and furthermore we may approximate the standard error of $g(\hat{\theta})$ by

$$\sqrt{\mathbb{V}(g(\hat{\theta}))} \approx g'(\hat{\theta}) \hat{\Sigma} g'(\hat{\theta})^\top$$

An example of this general character is given in Pedersen (1997).

5 Further remarks and related topics

This section contains more details on some of the topics treated in previous sections but it also contains some new and related topics.

5.1 More on backtransformations

In many cases when a statistical analysis has been performed on transformed data it is the estimated level, i.e. the mathematical expectation, of some treament that is backtransformed. If this level was known exactly that would be a trivial matter, but recognizing the uncertainty in the estimated level on the transformed scale implies that direct backtransformation as described in section 3 may not be the optimal thing to do. We will briefly discuss this problem below by considering an example.

Let $X_1, \ldots, X_n$ represent n independent measurements of some response on experimental units that have all received the same treatment. Suppose that $X_1, \ldots, X_n$ follow the log-normal distribution, i.e.

$$Y_i = \ln(X_i) \sim \mathcal{N}(\alpha, \sigma^2)$$

Then we analyze the data on the transformed scale, i.e. we have transformed the data by the transformation $g(x) = \ln(x)$, which leads to the usual estimators

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_.)^2$$

where $\bar{Y}_. = \frac{1}{n} \sum_{i=1}^{n} Y_i$. Hence $\hat{\alpha}$ is the estimator of $\alpha = \mathbb{E}(Y_i)$, i.e. the estimator of the level on the transformed scale. Furthermore, the distribution of $\hat{\alpha}$ is known: $\hat{\alpha} \sim \mathcal{N}(\alpha, \sigma^2/n)$. In particular, $\hat{\alpha}$ is unbiased, i.e. $\mathbb{E}(\hat{\alpha}) = \alpha$. This means that the estimator is correct 'on the average'.

Now we want to estimate the level on the original scale, i.e. $\mu = \mathbb{E}(X_i)$. The exact mathematical expectation on the original scale is given by

$$\mu = \exp(\alpha + \frac{1}{2}\sigma^2)$$

Hence a natural estimator of μ would be

$$\hat{\mu} = \exp(\hat{\alpha} + \frac{1}{2}s^2)$$

The estimator achieved by direct backtransformation is however given by

$$\tilde{\mu} = \exp(\hat{\alpha})$$

which is strictly smaller than $\hat{\mu}$. If there is no random variation in the data, i.e. $\sigma^2 = 0$, the two estimators are identical and equal to the true value, which in that extreme case is $\exp(\alpha)$. However, when there is random variation in the data, i.e. $\sigma^2 > 0$, then $\tilde{\mu}$ will be smaller then $\hat{\mu}$. In fact, for large sample sizes, n, $\hat{\mu}$ will be close to the true value $\mu = \exp(\alpha + \frac{1}{2}\sigma^2)$, whereas

$\tilde{\mu}$ will be close to $\exp(\alpha) < \exp(\alpha + \frac{1}{2}\sigma^2)$. In mathematical terms, $\hat{\mu}$ is consistent whereas $\tilde{\mu}$ is inconsistent. In other words, the estimator $\tilde{\mu}$ can not necessarily be improved by acquiring more observations. For all sample sizes such that $n - 1 > \sigma^2$ we furthermore have that

$$\mathbb{E}(\frac{\tilde{\mu} - \mu}{\mu}) = \exp(\frac{\sigma^2(1-n)}{2n}) - 1$$

$$\mathbb{E}(\frac{\hat{\mu} - \mu}{\mu}) = (\frac{n-1}{n-1-\sigma^2})^{\frac{n-1}{2}} \exp(\frac{\sigma^2(1-n)}{2n}) - 1$$

Since $\mathbb{E}(\tilde{\mu} - \mu) < 0$ we see that $\tilde{\mu}$ underestimates μ on the average, whereas the bias in $\hat{\mu}$ can be either positive or negative. The mean squared errors are

$$\mathbb{E}(\tilde{\mu} - \mu)^2 = \exp(2\alpha + 2\frac{\sigma^2}{n}) + \exp(2\alpha + \sigma^2) - 2\exp(2\alpha + \frac{1}{2}\sigma^2(1 + \frac{1}{n}))$$

$$\mathbb{E}(\hat{\mu} - \mu)^2 = \exp(2\alpha + 2\frac{\sigma^2}{n})(\frac{n-1}{n-1-2\sigma^2})^{\frac{n-1}{2}} + \exp(2\alpha + \sigma^2)$$
$$- 2\exp(2\alpha + \frac{1}{2}\sigma^2(1 + \frac{1}{n}))(\frac{n-1}{n-1-\sigma^2})^{\frac{n-1}{2}}$$

5.2 Higher order approximations

As illustrated in figure 1 the accuracy of the 1st order Taylor approximation depends on the distance to the point around which the linearization is done and on the curvature of the function. The first source of inaccuracy is translated into the fact that the accuracy of the formulas (4)–(5) depend on the smallness of the variance of the random variable that is transformed. This source of inaccuracy can not be improved upon but in most cases the second source of inaccuracy can be improved by choosing higher order Taylor approximations, in which case the function is no longer approximated by a linear function but by a polynomial. Below we present the corresponding formulas for the one-dimensional case where X and Y are (one-dimensional) random variables. The n'th order Taylor approximation of a function g in the neighbourhood of a point a is given by

$$g(x) \approx \sum_{k=0}^{n} \frac{g^{(k)}(a)}{k!}(x - a)^k$$

where $g^{(k)}$ denotes the k'th derivative of g $(g^{(0)} = g, g^{(1)} = g', g^{(2)} = g'', \dots)$ and $k!$ denotes the faculty of k, i.e. $k! = 1 \cdot 2 \cdots k$. Before applying this formula to the expectation and the variance of $Y = g(X)$ we need to introduce some notation. For $k = 0, 1, \dots$ put

$$\mu_k = \mathbb{E}(X - \mathbb{E}(X))^k$$

i.e. μ_k denotes the k'th central moment of X. With this notation we can write the n'th order versions of the formulas (2)–(3) as:

$$\mathbb{E}(Y) \approx g(\mathbb{E}(X)) + \sum_{k=1}^{n} \frac{g^{(k)}(\mathbb{E}(X))}{k!}\mu_k$$

$$\mathbb{V}(Y) \approx \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{g^{(k)}(\mathbb{E}(X))g^{(l)}(\mathbb{E}(X))}{k! \cdot l!}(\mu_{k+l} - \mu_k\mu_l)$$

16

The price for the higher accuracy is of course that the formulas require that one knows both the n first derivatives of g and the $2n$ first central moments of X. These quantities are however known in many cases. If for instance $X \sim \mathcal{N}(\alpha, \sigma^2)$ and $g(x) = \exp(x)$ then $g^{(k)}(x) = g(x)$ for all k and

$$\mu_k = \begin{cases} \sigma^k \cdot (k-1)!! & , \quad \text{if } k \text{ is even} \\ 0 & , \quad \text{if } k \text{ is odd} \end{cases}$$

where $x!! = 1 \cdot 3 \cdot 5 \cdots x$ for any odd integer x.

5.3 Box-Cox transformations and generalized linear models

The class of (power-) transformations defined by

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

is sometimes called the class of Box-Cox transformations, see Seber & Wild (1989). The Box-Cox transformations apply to positive data and are usually used as variance stabilization transformations. Suppose for a given data set the variance seems to increase with the level, e.g. the standard deviation seems to increase proportionally to the level, or more generally

$$\mathbb{V}(X_i) = \sigma^2 \mathbb{E}(X_i)^\beta$$

where β is some unknown constant. Then a Box-Cox transformation of the data might stabilize the variance. If for instance we put $Y_i = g_\lambda(X_i)$ for some λ, then according to formula (3) we have that

$$\mathbb{V}(Y_i) \approx \sigma^2 \mathbb{E}(X_i)^{2\lambda - 2 + \beta}$$

Hence if λ is chosen such that $2\lambda - 2 + \beta = 0$, i.e. $\lambda = 1 - \frac{\beta}{2}$, then

$$\mathbb{V}(Y_i) \approx \sigma^2$$

This means that choosing the right value of λ, i.e. the right Box-Cox transformation, may stabilize the variance of the transformed data. The idea is then to apply the Box-Cox transformations to the data for a lot of different values of λ and pick the one, if any, that makes the variance independent of the level on the transformed scale.

Another reason for data transformation may be to obtain a linear relationsship between some covariate and a response variable. Consider for instance the (fictive) data set in the left most plot in figure 4. In this case one might be tempted to transform the data by e.g. the ln-transform in order to obtain a linear relationship between the response and the covariate. This is done in the rightmost plot in figure 4. Obviously the relationsship between the covariate and response is now closer to being linear. Notice, however, that the random variation around the exponential curve on the original scale seems to have constant variance over the whole range, whereas this is certainly not the case on the ln-transformed scale, where the variance decreases with increasing values of the covariate thus violating the variance homogeneity assumption that is often necessary in linear regression analyses. The reason for this is of course that the data transformation changes the variance, cf. the discussion above. In such cases one might apply the theory for generalized linear models, see e.g. McCullagh & Nelder (1989), where it is possible to specify transformations of the model for the expexted values without affecting the variance or
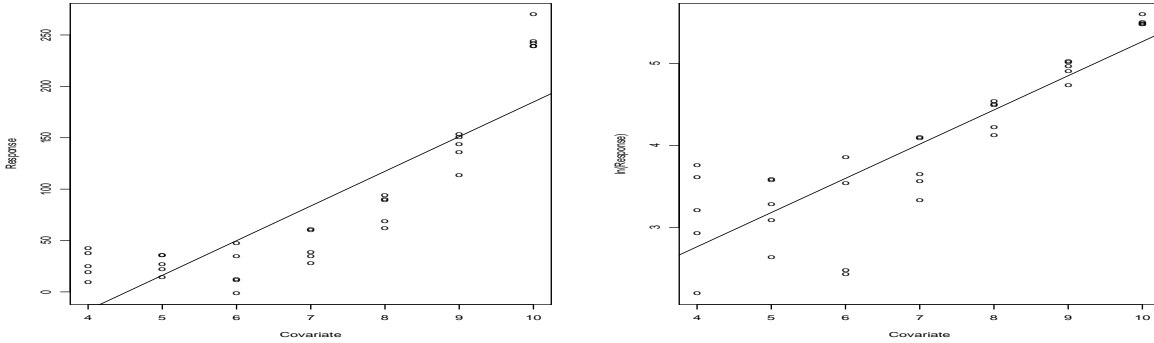
Figure 4: Fictive data set on the original and the ln-transformed scale

more generally the type of distribution of the data. In this concrete example an initial model for the original data might be that

$$X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

where μ_i depends linearly on some covariate t_i on a ln-scale, and the X_i's have constant variance $\mathbb{E}(X_i) = \sigma^2$. Transforming like $Y_i = \ln(X_i)$ 'destroys' the variance heterogeneity, but in the framework of generalized linear models it is possible to specify a model of the form

$$X_i \sim \mathcal{N}(\ln(\mu_i) = \alpha + \beta t_i, \sigma^2)$$

hence maintaining the variance homogeneity while modeling the expectations linearly on a ln-scale.

References

McCullagh, P. & Nelder, J. (1989). *Generalized linear models*. Chapman & Hall, London.

Pedersen, A.R. (1997). Analyse af flerdimensionale kovarianskomponentmodeller ved hjœlp af SAS-proceduren PROC MIXED. *Internal report, Biometry Research Unit, Danish Institute of Agricultural Sciences* **1**, 1–45.

SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6, Fourth Edition, volume 2*. SAS Institute Inc., Cary, NC.

Seber, G. & Wild, C. (1989). *Non-linear regression*. John Wiley & Sons, New York.

A Example of SAS program

The following SAS-program illustrates how the backtransformed values and confidence intervals may be calculated. The program has been tested, but not thoroughly, and it is not guaranteed to work. In other words, we assume no responsibilities for results from calculations made by the progam. The program may be downloaded from the www at the URL http://www.sp.dk/~ej/SASsamples/

```
data a ;

input trnstype $ treatno ymean ystde ;

** The following variables should be specified in the input data;
** YMean     : The mean on the transformed scale;
** YStde     : The standard error on the transformed scale ;
** trnstype : The type of the transformation, either: ;
**           : LOGIT LOG LOG10 SQRT INVERSE ARCSIN            ;

** The following variables are calculated;
** XMean     : The mean on the original scale;
** XStde     : The app. standard error on the original scale ;
** Ylow      : Low limit in confidence interval on the transformed scale;
** Yhigh     : High limit in confidence interval on the transformed scale;
** Xlow      : Low limit in confidence interval on the original scale;
** Xhigh     : High limit in confidence interval on the original scale;
** Xlowa     : App. low limit in confidence interval on the original scale;
** Xhigha    : App. high limit in confidence interval on the original scale;

** Note, if the degrees of freedom in the estimate of the standard error;
** are low, the 1.96 in the formulas below should be changed to the ;
** appropiate value from the t-distribution;

IF trnstype='LOGIT' then do ;
  Xmean=exp(Ymean)/(1+exp(Ymean)) ;
  Xstde=Ystde*(1-Xmean)*Xmean ;
  Xlow=exp(Ymean-1.96*Ystde)/(1+exp(Ymean-1.96*Ystde)) ;
  Xhigh=exp(Ymean+1.96*Ystde)/(1+exp(Ymean+1.96*Ystde));
  Xlowa=Xmean-1.96*Xstde ;
  Xhigha=Xmean+1.96*Xstde ;
  end ;

IF trnstype='LOG' then do ;
  Xmean=exp(Ymean) ;
  Xstde=Ystde*Xmean ;
  Xlow=exp(Ymean-1.96*Ystde) ;
  Xhigh=exp(Ymean+1.96*Ystde);
  Xlowa=Xmean-1.96*Xstde ;
```

```
    Xhigha=Xmean+1.96*Xstde ;
    end ;

IF trnstype='LOG10' then do ;
  Xmean=exp(Ymean)*log(10) ;
  Xstde=Ystde*Xmean*log(10) ;
  Xlow=exp(Ymean-1.96*Ystde) ;
  Xhigh=exp(Ymean+1.96*Ystde);
  Xlowa=Xmean-1.96*Xstde ;
  Xhigha=Xmean+1.96*Xstde ;
  end ;

IF trnstype='SQRT' then do ;
  Xmean=Ymean*Ymean ;
  Xstde=Ystde*0.5*sqrt(Xmean) ;
  Xlow=(Ymean-1.96*Ystde)*(Ymean-1.96*Ystde) ;
  Xhigh=(Ymean+1.96*Ystde)*(Ymean+1.96*Ystde);
  Xlowa=Xmean-1.96*Xstde ;
  Xhigha=Xmean+1.96*Xstde ;
  end ;

IF trnstype='INVERSE' then do ;
  Xmean=1/Ymean ;
  Xstde=Ystde*Xmean*Xmean ; * Absolute value of g'(x) ;
  Xlow=1/(Ymean-1.96*Ystde) ;
  Xhigh=1/(Ymean+1.96*Ystde);
  Xlowa=Xmean-1.96*Xstde ;
  Xhigha=Xmean+1.96*Xstde ;
  end ;

IF trnstype='ARCSIN' then do ;
  Xmean=Sin(Ymean)*sin(Ymean) ;

  Xstde=2*sqrt(Xmean)*cos(YMean)*Ystde ;   ;
  Xlow=sin(Ymean-1.96*Ystde)*sin(Ymean-1.96*Ystde) ;
  Xhigh=sin(Ymean+1.96*Ystde)*sin(Ymean+1.96*Ystde);
  Xlowa=Xmean-1.96*Xstde ;
  Xhigha=Xmean+1.96*Xstde ;
  end ;
cards ;
LOG 1 1.08 0.042
LOG 2 1.51 0.045
LOG 3 1.27 0.042
LOG 4 1.03 0.047
proc print ;
run;
```