

The ‘Delta method’ ...

Suppose you have done a study, over 4 years, which yields 3 estimates of survival (say, $\hat{\phi}_1$, $\hat{\phi}_2$, and $\hat{\phi}_3$). But, suppose what you are really interested in is the estimate of the product of the three survival values (i.e., the probability of surviving from the beginning of the study to the end of the study)? While it is easy enough to derive an estimate of this product (as $[\hat{\phi}_1 \times \hat{\phi}_2 \times \hat{\phi}_3]$), how do you derive an estimate of the variance of the product? In other words, how do you derive an estimate of the variance of a transformation of one or more random variables (in this case, we transform the three random variables - $\hat{\phi}_i$ - by considering their product)?

One commonly used approach which is easily implemented, not computer-intensive, and can be robustly applied in many (but not all) situations is the so-called *Delta method* (also known as the method of propagation of errors). In this appendix, we briefly introduce the underlying background theory, and the implementation of the Delta method, to fairly typical scenarios.

B.1. Mean and variance of random variables

Our primary interest here is developing a method that will allow us to estimate the mean and variance for functions of random variables. Let's start by considering the formal approach for deriving these values explicitly, based on the *method of moments*. For continuous random variables, consider a continuous function $f(x)$ on the interval $[-\infty, +\infty]$. The first three moments of $f(x)$ can be written as

$$M_0 = \int_{-\infty}^{+\infty} f(x) dx$$

$$M_1 = \int_{-\infty}^{+\infty} x f(x) dx$$

$$M_2 = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

In the particular case that the function is a probability density (as for a continuous random variable), then $M_0 = 1$ (i.e., the area under the PDF must equal 1).

For example, consider the uniform distribution on the finite interval $[a, b]$. A uniform distribution (sometimes also known as a rectangular distribution), is a distribution that has constant probability

over the interval. The probability density function (pdf) for a continuous uniform distribution on the finite interval $[a, b]$ is

$$P(x) = \begin{cases} 0 & \text{for } x < a \\ 1/(b-a) & \text{for } a < x < b \\ 0 & \text{for } x > b \end{cases}$$

Integrating the pdf, for $p(x) = 1/(b-a)$,

$$\begin{aligned} M_0 &= \int_a^b p(x) dx \\ &= \int_a^b \frac{1}{b-a} dx = 1 \end{aligned} \quad (\text{B.1})$$

$$\begin{aligned} M_1 &= \int_a^b x p(x) dx \\ &= \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2} \end{aligned} \quad (\text{B.2})$$

$$\begin{aligned} M_2 &= \int_a^b x^2 p(x) dx \\ &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3} (a^2 + ab + b^2) \end{aligned} \quad (\text{B.3})$$

We see clearly that M_1 is the mean of the distribution. What about the variance? Where does the second moment M_2 come in? Recall that the variance is defined as the average value of the fundamental quantity [distance from mean]². The squaring of the distance is so the values to either side of the mean don't cancel out. Standard deviation is simply the square-root of the variance.

Given some discrete random variable x_i , with probability p_i , and mean μ , we define the variance as

$$\text{Var} = \sum (x_i - \mu)^2 p_i$$

Note we don't have to divide by the number of values of x because the sum of the discrete probability distribution is 1 (i.e., $\sum p_i = 1$). For a continuous probability distribution, with mean μ , we define the variance as

$$\text{Var} = \int_a^b (x - \mu)^2 p(x) dx$$

Given our moment equations, we can then write

$$\begin{aligned} \text{Var} &= \int_a^b (x - \mu)^2 p(x) dx \\ &= \int_a^b (x^2 - 2\mu x + \mu^2) p(x) dx \\ &= \int_a^b x^2 p(x) dx - \int_a^b 2\mu x p(x) dx + \int_a^b \mu^2 p(x) dx \\ &= \int_a^b x^2 p(x) dx - 2\mu \int_a^b x p(x) dx + \mu^2 \int_a^b p(x) dx \end{aligned}$$

Now, if we look closely at the last line, we see that in fact the terms represent the different moments

of the distribution. Thus we can write

$$\begin{aligned} \text{Var} &= \int_a^b (x - \mu)^2 p(x) dx \\ &= \int_a^b x^2 p(x) dx - 2\mu \int_a^b x p(x) dx + \mu^2 \int_a^b p(x) dx \\ &= M_2 - 2\mu (M_1) + \mu^2 (M_0) \end{aligned}$$

Since $M_1 = \mu$, and $M_0 = 1$ then

$$\begin{aligned} \text{Var} &= M_2 - 2\mu (M_1) + \mu^2 (M_0) \\ &= M_2 - 2\mu(\mu) + \mu^2(1) \\ &= M_2 - 2\mu^2 + \mu^2 \\ &= M_2 - \mu^2 \\ &= M_2 - (M_1)^2 \end{aligned}$$

In other words, the variance for the pdf is simply the second moment (M_2) minus the square of the first moment ($(M_1)^2$). Thus, for a continuous uniform random variable x on the interval $[a, b]$,

$$\begin{aligned} \text{Var} &= M_2 - (M_1)^2 \\ &= \frac{(a - b)^2}{12} \end{aligned}$$

B.2. Transformations of random variables and the Delta method

OK - that's fine. If the pdf is specified, we can use the method of moments to formally derive the mean and variance of the distribution. But, what about functions of random variables having poorly specified or unspecified distributions? Or, situations where the pdf is not easily defined?

In such cases, we may need other approaches. We'll introduce one such approach (the Delta method) here, by considering the case of a simple linear transformation of a random normal distribution.

Let

$$X_1, X_2, \dots \sim N(10, \sigma^2 = 2)$$

In other words, random deviates drawn from a normal distribution with a mean of 10, and a variance of 2. Consider some transformations of these random values. You might recall from some earlier statistics or probability class that linearly transformed normal random variables are themselves normally distributed. Consider for example, $X_i \sim N(10, 2)$ - which we then linearly transform to Y_i , such that $Y_i = 4X_i + 3$.

Now, recall that for real scalar constants a and b we can show that

- i. $E(a) = a, E(aX + b) = aE(X) + b$
- ii. $\text{var}(a) = 0, \text{var}(aX + b) = a^2\text{var}(X)$

Thus, given $X_i \sim N(10, 2)$ and the linear transformation $Y_i = 4X_i + 3$, we can write

$$Y \sim N(4(10) + 3 = 43, (4^2)2) = N(43, 32)$$

Now, an important point to note is that some transformations of the normal distribution are close to normal (i.e., are linear) and some are not. Since linear transformations of random normal values are normal, it seems reasonable to conclude that approximately linear transformations (over some range) of random normal data should also be approximately normal.

OK, to continue. Let $X \sim N(\mu, \sigma^2)$, and let $Y = g(X)$, where g is some transformation of X (in the previous example, $g(X) = 4X + 3$). It is hopefully relatively intuitive that the closer $g(X)$ is to linear over the likely range of X (i.e., within 3 or so standard deviations of μ), the closer $Y = g(X)$ will be to normally distributed. From calculus, we recall that if you look at any differentiable function over a narrow enough region, the function appears approximately linear. The approximating line is the tangent line to the curve, and its slope is the derivative of the function.

Since most of the mass (i.e., most of the random values) of X is concentrated around μ , let's figure out the tangent line at μ , using two different methods. First, we know that the tangent line passes through $(\mu, g(\mu))$, and that its slope is $g'(\mu)$ (we use the ' g' ' notation to indicate the first derivative of the function g). Thus, the equation of the tangent line is $Y = g'(\mu)X + b$ for some b . Replacing (X, Y) with the known point $(\mu, g(\mu))$, we find $g(\mu) = g'(\mu)\mu + b$ and so $b = g(\mu) - g'(\mu)\mu$. Thus, the equation of the tangent line is $Y = g'(\mu)X + g(\mu) - g'(\mu)\mu$.

Now for the big step – we can derive an approximation to the same tangent line by using a *Taylor series expansion* of $g(x)$ (to first order) around $X = \mu$

$$\begin{aligned} Y &= g(X) \\ &\approx g(\mu) + g'(\mu)(X - \mu) + \epsilon \\ &= g'(\mu)X + g(\mu) - g'(\mu)\mu + \epsilon \end{aligned}$$

OK, at this point you might be asking yourself 'so what?'. Well, suppose that $X \sim N(\mu, \sigma^2)$ and $Y = g(X)$, where $g'(\mu) \neq 0$. Then, whenever the tangent line (derived earlier) is approximately correct over the likely range of X (i.e., if the transformed function is approximately linear over the likely range of X), then the transformation $Y = g(X)$ will have an approximate normal distribution. That approximate normal distribution may be found using the usual rules for linear transformations of normals.

Thus, to first order,

$$E(Y) = g'(\mu)\mu + g(\mu) - g'(\mu)\mu = g(\mu)$$

$$\begin{aligned} \text{var}(Y) &= \text{var}(g(X)) = (g(X) - g(\mu))^2 \\ &= (g'(\mu)(X - \mu))^2 \\ &= (g'(\mu))^2(X - \mu)^2 \\ &= (g'(\mu))^2\text{var}(X) \end{aligned}$$

In other words, we take the derivative of the transformed function with respect to the parameter, square it, and multiply it by the estimated variance of the untransformed parameter.

These first-order approximations to the variance of a transformed parameter are usually referred to as the *Delta method*.

begin sidebar

Taylor series expansions?

A very important, and frequently used tool. If you have no familiarity at all with series expansions, here is a (very) short introduction. Briefly, the *Taylor series* is a power series expansion of an infinitely differentiable real (or complex) function defined on an open interval around some specified point. For example, a one-dimensional Taylor series is an expansion of a real function $f(x)$ about a point $x = a$ over the interval $(a - r, a + r)$, is given as:

$$f(x) \approx f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \dots$$

where $f'(a)$ is the first derivative of f with respect to a , $f''(a)$ is the second derivative of f with respect to a , and so on.

For example, suppose the function is $f(x) = e^x$. The convenient fact about this function is that all its derivatives are equal to e^x as well (i.e., $f(x) = e^x, f'(x) = e^x, f''(x) = e^x, \dots$). In particular, $f^{(n)}(0) = e^0 = 1$ so that $f^{(n)}(0) = 1$. This means that the coefficients of the Taylor series are given by

$$a_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{n!}$$

and so the Taylor series is given by

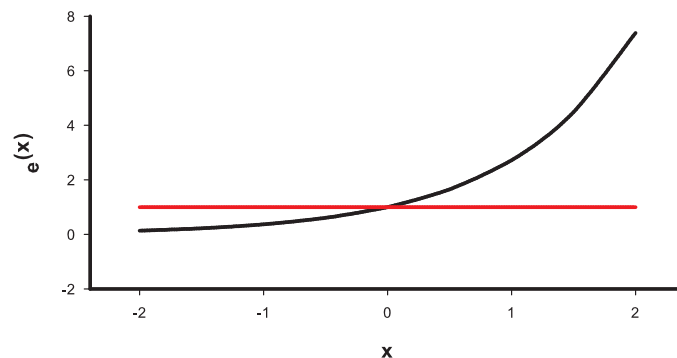
$$1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots + \frac{x^n}{n!} + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

The primary utility of such a power series in simple application is that differentiation and integration of power series can be performed term by term and is hence particularly (or, at least relatively) easy. In addition, the (truncated) series can be used to compute function values approximately.

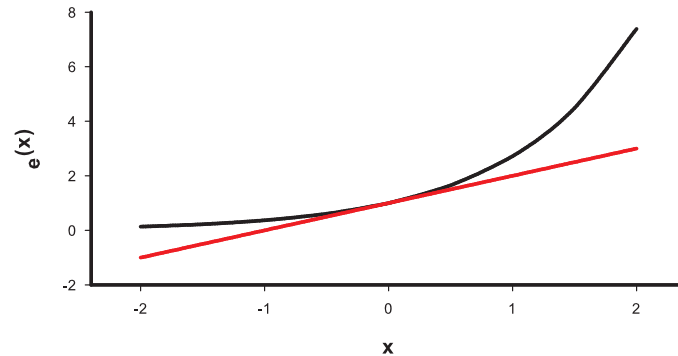
Now, let's look at an example of the "fit" of a Taylor series to a familiar function, given a certain number of terms in the series. For our example, we'll expand the function $f(x) = e^x$, at $x = a = 0$, on the interval $(a - 2, a + 2)$, for $n = 0, n = 1, n = 2, \dots$ (where n is the number of terms in the series). For $n = 0$, the Taylor expansion is a scalar constant (1):

$$f(x) \approx 1$$

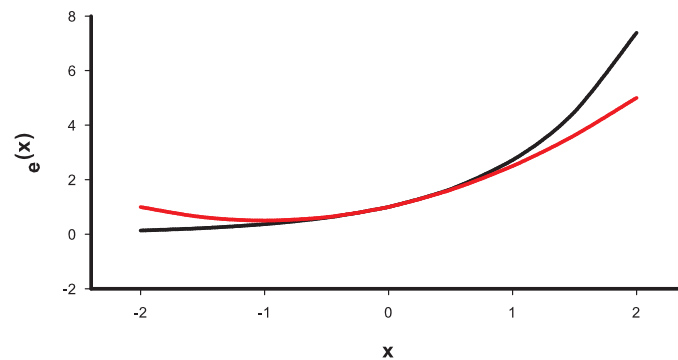
which is obviously a poor approximation to the function $f(x) = e^x$ at any point. This is shown clearly in the following figure - the black line in the figure is the function $f(x) = e^x$, evaluated over the interval $(-2, 2)$, and the red line is the Taylor series approximation for $n = a = 0$.



What happens when we add higher order terms? Here is the plot of the Taylor series for $n = 1$.



Hmmm...a bit better. What about $n = 2$?



We see that when we add more terms (i.e., use a higher-order series), the fit gets progressively better. Often, for 'nice, smooth' functions (i.e., those nearly linear at the point of interest), we don't need many terms at all. For this example, $n = 4$ yields a near-perfect fit (over the interval $(-2, 2)$).

Another example - suppose the function of interest is $f(x) = (x)^{1/3}$ (i.e., $f(x) = \sqrt[3]{x}$). Suppose we're interested in $f(x) = (x)^{1/3}$ where $x = 27$ (i.e., $f(27) = \sqrt[3]{27}$). Now, it is straightforward to show that $f(27) = \sqrt[3]{27} = 3$. But suppose we want to know $f(25) = \sqrt[3]{25}$, using a Taylor series approximation? We recall that to first order,

$$f(x) = f(a) + f'(a)(x - a)$$

where in this case, $a = 25$ and $x = 27$. The derivative of f with respect to x for this function $f(a) = (a)^{1/3}$ is

$$f'(a) = \frac{a^{-2/3}}{3} = \frac{1}{3\sqrt[3]{a^2}}$$

Thus, using the first-order Taylor series, we write

$$\begin{aligned} f(25) &\approx f(27) + f'(27)(25 - 27) \\ &= 3 + (0.037037)(-2) \\ &= 3 - 0.0740741 \\ &= 2.926 \end{aligned}$$

which is very close to the true value of $f(25) = \sqrt[3]{25} = 2.924$. In other words, the first-order Taylor approximation works pretty well for this function.

end sidebar

B.3. Transformations of one variable

OK, enough background for now. Let's see some applications. Let's check the Delta method out in a case where we know the answer. Assume we have an estimate of density \widehat{D} and its conditional sampling variance, $\widehat{\text{var}}(D_s)$. We want to multiply this by some constant c to make it comparable with other values from the literature. Thus, we want $\widehat{D}_s = g(D) = c\widehat{D}$ and $\widehat{\text{var}}D_s$.

The Delta method gives

$$\begin{aligned} \widehat{\text{var}}(D_s) &= (g'(D))^2 \hat{\sigma}_D^2 \\ &= \left(\frac{\partial D_s}{\partial \widehat{D}} \right)^2 \cdot \widehat{\text{var}}(D) \\ &= c^2 \cdot \widehat{\text{var}}(\widehat{D}) \end{aligned}$$

which we know to be true for the variance of a random variable multiplied by a real constant.

Another example of the same thing – consider a known number of harvested fish and an average weight ($\widehat{\mu}_w$) and its variance. If you want an estimate of total biomass (B), then $\widehat{B} = N \cdot \widehat{\mu}_w$ and the variance of \widehat{B} is $N^2 \cdot \widehat{\text{var}}(\widehat{\mu}_w)$.

Still another example - you have some parameter θ , which you transform by dividing it by some constant c . Thus, by the Delta method,

$$\widehat{\text{var}}\left(\frac{\widehat{\theta}}{c}\right) = \left(\frac{1}{c}\right)^2 \cdot \widehat{\text{var}}(\widehat{\theta})$$

B.3.1. A potential complication - violation of assumptions

A final - and important - example for transformations of single variables. The importance lies in the demonstration that the Delta method does not always work - remember, it assumes that the transformation is approximately linear over the expected range of the parameter. Suppose one has an MLE for the mean and estimated variance for some parameter θ which is bounded random uniform on the interval $[0, 2]$. Suppose you want to transform this parameter such that

$$\psi = e^{(\theta)}$$

(Recall that this is a convenient transformation since the derivative of e^x is e^x , making the calculations very simple). Now, based on the Delta method, the variance for ψ would be estimated as

$$\begin{aligned}\widehat{\text{var}}(\widehat{\psi}) &= \left(\frac{\partial \widehat{\psi}}{\partial \widehat{\theta}}\right)^2 \cdot \widehat{\text{var}}(\widehat{\theta}) \\ &= (e^{\widehat{\theta}})^2 \cdot \widehat{\text{var}}(\widehat{\theta})\end{aligned}$$

Now, suppose that $\widehat{\theta} = 1.0$, and $\widehat{\text{var}}(\widehat{\theta}) = 0.33\dot{3}$. Then, from the Delta method,

$$\begin{aligned}\widehat{\text{var}}(\widehat{\psi}) &= (e^{\widehat{\theta}})^2 \cdot \widehat{\text{var}}(\widehat{\theta}) \\ &= (7.38906)(0.33\dot{3}) \\ &= 2.46302\end{aligned}$$

OK, so what's the problem? Well, let's derive the variance of ψ using the method of moments. To do this, we need to integrate the pdf (uniform, in this case) over some range. Since the variance of a uniform distribution is $(b-a)^2/12$, and if b and a are symmetric around the mean (1.0), then we can show by algebra that given a variance of $0.33\dot{3}$, then $a = 0$ and $b = 2$.

Given a uniform distribution, the pdf is $p(\theta) = 1/(b-a)$. Thus, by the method of moments,

$$\begin{aligned}M_1 &= \int_a^b g(x)p(x)dx \\ &= \int_a^b \frac{g(x)}{b-a}dx \\ &= -\frac{e^b - e^a}{a-b}\end{aligned}$$

$$\begin{aligned}M_2 &= \int_a^b \frac{g(x)^2}{b-a}dx \\ &= \frac{1}{2} \frac{e^{2a} - e^{2b}}{a-b}\end{aligned}$$

Thus, by moments, $\widehat{\text{var}}(E(\widehat{\psi}))$ is

$$\widehat{\text{var}}(E(\widehat{\psi})) = M_2 - (M_1)^2 = \frac{1}{2} \frac{-e^{2b} + e^{2a}}{-b+a} - \frac{(e^b - e^a)^2}{(a-b)^2}$$

If $a = 0$ and $b = 2$, then

$$\widehat{\text{var}}(E(\widehat{\psi})) = M_2 - (M_1)^2 = \frac{1}{2} \frac{-e^{2b} + e^{2a}}{-b+a} - \frac{(e^b - e^a)^2}{(a-b)^2} = 3.19453$$

which is not particularly close to the value estimated by the Delta method (2.46302).

Why the discrepancy? As discussed earlier, the Delta method rests on the assumption the first-order Taylor expansion around the parameter value is effectively linear over the range of values likely

to be encountered. Since in this example we're using a uniform pdf, then all values between a and b are equally likely. Thus, we might anticipate that as the interval between a and b gets smaller, then the approximation to the variance (which will clearly decrease) will get better and better (since the smaller the interval, the more likely it is that the function is approximately linear over that range). For example, if $a = 0.5$ and $b = 1.5$ (same mean of 1.0), then the true variance of θ will be 0.083. Thus, by the Delta method, the estimated variance of ψ will be 0.61575, while by the method of moments (which is exact), the variance will be 0.65792. Clearly, the proportional difference between the two values has declined markedly. But, we achieved this 'improvement' by artificially reducing the true variance of the untransformed variable θ . Obviously, we can't do this in general practice.

So, what are the practical options? Well, one possible solution is to use a higher-order Taylor series approximation - by including higher-order terms, we can achieve a better 'fit' to the function (see the preceding sidebar). If we used a second-order TSE,

$$E(g(x)) \approx g(\mu) + \frac{1}{2}g''(\mu)\sigma^2 \quad (\text{B.4})$$

$$\text{Var}(g(x)) \approx g'(\mu)^2\sigma^2 + \frac{1}{4}(g''(\mu))^2(\text{Var}(x^2) - 4\mu^2\sigma^2) \quad (\text{B.5})$$

we should do a bit better. For the variance estimate, we need to know $\text{var}(x^2)$, which for a continuous uniform distribution by the method of moments is

$$\left(\frac{1}{5} \cdot \frac{b^5 - a^5}{b - a}\right) - \left(\frac{1}{9} \cdot \frac{(b^3 - a^3)^2}{(b - a)^2}\right)$$

Thus, from the second-order approximation, and again assuming $a = 0$ and $b = 2$, then $\widehat{\text{var}}(\psi)$ is

$$\begin{aligned} \widehat{\text{var}}(\psi) &\approx (e^\theta)^2 \cdot \text{var}(\theta) + \frac{1}{4}(e^\theta)^2 \cdot \text{var}(\theta^2) - 4\mu^2\text{var}(\theta) \\ &= 3.756316 \end{aligned}$$

which is closer (proportionately) to the true variance (3.19453) than was the estimate using only the first-order TSE (2.46302). The reason that even a second-order approximation isn't 'much closer' is because the transformation is very non-linear over the range of the data (uniform $[0, 2]$ in this case), such that the second-order approximation doesn't 'fit' particularly well over this range.

So, we see that the classical Delta method, which is based on a first-order Taylor series expansion of the transformed function, may not do particularly well if the function is highly non-linear over the range of values being examined.

Of course, it would be fair to note that the preceding example made the assumption that the distribution was random uniform over the interval. For most of our work with **MARK**, the interval is likely to have a symmetric mass around the estimate, typically β . As such, most of data, and thus the transformed data, will actually fall closer to the parameter value in question (the mean in this example) than we've demonstrated here. So much so, that the discrepancy between the first order 'Delta' approximation to the variance and the true value of the variance will likely be significantly smaller than shown here, even for a strongly non-linear transformation. We leave it to you as an exercise to prove this for yourself. But, this point notwithstanding, it is important to be aware of the assumptions underlying the Delta method - if your transformation is non-linear, and there is considerable variation in your data, the first-order approximation may not be particular good. Fortunately, use of second order Taylor series approximations is not heroically difficult - the challenge is usually coming up with $\widehat{\text{var}}(X^2)$. If the pdf for the untransformed data is specified (which

is essentially equivalent to assuming an informative prior), then you can derive $\widehat{\text{var}(X^2)}$ fairly easily using the method of moments.

B.4. Transformations of two or more variables

Clearly, we are often interested in transformations involving more than one variable. Fortunately, there are also multivariate generalizations of the Delta method.

Suppose you've estimated p different random variables X_1, X_2, \dots, X_p . In matrix notation, these variables would constitute a $(p \times 1)$ random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

which has a mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_p \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

and the $(p \times p)$ variance-covariance matrix is

$$\begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_p) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \dots & \text{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_p, X_1) & \text{cov}(X_p, X_2) & \dots & \text{var}(X_p) \end{pmatrix}$$

Note that if the variables are independent, then the off-diagonal elements (i.e., the covariance terms) are all zero.

Then, for a $(k \times p)$ matrix of constants $\mathbf{A} = a_{ij}$, the expectation of a random vector $\mathbf{Y} = \mathbf{A}\mathbf{X}$ is given as

$$\begin{pmatrix} EY_1 \\ EY_2 \\ \vdots \\ EY_p \end{pmatrix} = \mathbf{A}\boldsymbol{\mu}$$

with a variance-covariance matrix

$$\text{cov}(\mathbf{Y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

Now, using the same logic we first considered for developing the Delta method for a single variable, for each x_i near μ_i , we can write

$$y = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_p(x) \end{pmatrix} \approx \begin{pmatrix} g_1(\mu) \\ g_2(\mu) \\ \vdots \\ g_p(\mu) \end{pmatrix} + \mathbf{D}(x - \mu)$$

where \mathbf{D} is the matrix of partial derivatives of g_i with respect to x_j , evaluated at $(x - \mu)$.

As with the single-variable Delta method, if the variances of the X_i are small (so that with high probability Y is near μ , such that the linear approximation is usually valid), then to first-order we can write

$$\begin{pmatrix} EY_1 \\ EY_2 \\ \vdots \\ EY_p \end{pmatrix} = \begin{pmatrix} g_1(\mu) \\ g_2(\mu) \\ \vdots \\ g_p(\mu) \end{pmatrix} \quad \text{var}(\mathbf{Y}) \approx \mathbf{D}\Sigma\mathbf{D}^T$$

In other words, to approximate the variance of some multi-variable function \mathbf{Y} , we (i) take the vector of partial derivatives of the function with respect to each parameter in turn, \mathbf{D} , (ii) right-multiply this vector by the variance-covariance matrix, Σ , and (iii) right-multiply the resulting product by the transpose of the original vector of partial derivatives, \mathbf{D}^T .*

Example (1) - variance of product of survival probabilities

Let's consider the application of the Delta method in estimating sampling variances of a fairly common function - the product of several parameter estimates.

Now, from the preceding, we see that

$$\text{var}(\mathbf{Y}) \approx \mathbf{D}\Sigma\mathbf{D}^T = \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right) \cdot \hat{\Sigma} \cdot \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right)^T$$

where Y is some linear or nonlinear function of the parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \dots$. The first term on the RHS of the variance expression is a row vector containing partial derivatives of Y with respect

* There are alternative formulations of this expression which may be more convenient to implement in some instances. When the variables $\theta_1, \theta_2, \dots, \theta_k$ (in the function, \mathbf{Y}) are independent, then

$$\begin{aligned} \text{var}(\mathbf{Y}) &\approx \text{var}(f(\theta_1, \theta_2, \dots, \theta_k)) \\ &= \sum_{i=1}^k \text{var}(\theta_i) \left(\frac{\partial f}{\partial \theta_i} \right)^2 \end{aligned}$$

where $\partial f / \partial \theta_i$ is the partial derivative of \mathbf{Y} with respect to θ_i .

When the variables $\theta_1, \theta_2, \dots, \theta_k$ (in the function, \mathbf{Y}) are **not** independent, then the covariance structure among the variables must be accounted for:

$$\begin{aligned} \text{var}(\mathbf{Y}) &\approx \text{var}(f(\theta_1, \theta_2, \dots, \theta_k)) \\ &= \sum_{i=1}^k \text{var}(\theta_i) \left(\frac{\partial f}{\partial \theta_i} \right)^2 + 2 \sum_{i=1}^k \sum_{j=1}^k \text{cov}(\theta_i, \theta_j) \left(\frac{\partial f}{\partial \theta_i} \right) \left(\frac{\partial f}{\partial \theta_j} \right) \end{aligned}$$

to each of these parameters ($\hat{\theta}_1, \hat{\theta}_2, \dots$). The right-most term of the RHS of the variance expression is simply a transpose of this row vector (i.e., a column vector). The middle-term is simply the estimated variance-covariance matrix for the parameters.

OK, let's try an example - let's use estimates from the male European dipper data set (yes, again). We'll fit model $\{\phi_i p.\}$ to these data. Suppose we're interested in the probability of surviving from the start of the first interval to the end of the third interval. Well, the point-estimate of this probability is easy enough - it's simply $(\hat{\phi}_1 \times \hat{\phi}_2 \times \hat{\phi}_3) = (0.6109350 \times 0.458263 \times 0.4960239) = 0.138871$. So, the probability of a male Dipper surviving over the first three intervals is $\sim 14\%$ (again, assuming that our time-dependent survival model is a valid model).

To derive the estimate of the variance of the product, we will also need the **variance-covariance matrix** for the survival estimates. You can generate the matrix easily in **MARK** by selecting 'Output | Specific Model Output | Variance-Covariance Matrices | Real Estimates'.

Here is the variance-covariance matrix for the male Dipper data, generated from model $\{\phi_i p.\}$:

```

male dippers
Real Parameter Estimates Variances and Covariances
{phi(t)p(.)}

Variance-Covariance matrix of estimates on diagonal and below,
Correlation matrix of estimates above diagonal.
-----
      1      2      3      4      5      6
-----
1 | 0.02243 | -0.02638 | 0.00513 | 0.00735 | 0.00516 | 0.02379
  | -0.09253 |          |          |          |          |          |
2 | -0.00039 | 0.00997  | -0.02779 | 0.00545 | 0.00383 | 0.01765
  | -0.06865 |          |          |          |          |          |
3 | 0.00007  | -0.00024 | 0.00724  | -0.03332 | 0.00309 | 0.01427
  | -0.05549 |          |          |          |          |          |
4 | 0.00009  | 0.00004  | -0.00023 | 0.00661  | -0.04175 | 0.02042
  | -0.07941 |          |          |          |          |          |
5 | 0.00006  | 0.00003  | 0.00002  | -0.00026 | 0.00581  | -0.02857
  | -0.05572 |          |          |          |          |          |
6 | 0.00028  | 0.00014  | 0.00010  | 0.00013  | -0.00017 | 0.00634
  | -0.25711 |          |          |          |          |          |
7 | -0.00053 | -0.00026 | -0.00018 | -0.00025 | -0.00016 | -0.00078
  | 0.00146  |          |          |          |          |          |

```

In the Notepad output from **MARK**, the variance-covariance values are *below* the diagonal, whereas the standardized correlation values are *above* the diagonal. The variances are given *along* the diagonal.

However, it is **very important** to note that the V-C matrix that **MARK** outputs to the Notepad is *rounded* to 5 significant digits. For the actual calculations, we need to use the full precision values. To get those, you need to either (i) output the V-C matrix into a dBase file (which you could then open with dBase, or Excel), or (ii) copy the V-C matrix into the Windows clipboard, and then paste it into some other application. Failure to use the full precision V-C matrix will often (almost always, in fact) lead to 'rounding errors'. The 'full precision' V-C matrix for the survival values is shown at the top of the next page.

$$\begin{aligned}\widehat{\text{cov}}(Y) &= \begin{pmatrix} \text{var}(\phi_1) & \text{cov}(\phi_1, \phi_2) & \text{cov}(\phi_1, \phi_3) \\ \text{cov}(\phi_2, \phi_1) & \text{var}(\phi_2) & \text{cov}(\phi_2, \phi_3) \\ \text{cov}(\phi_3, \phi_1) & \text{cov}(\phi_3, \phi_2) & \text{var}(\phi_3) \end{pmatrix} \\ &= \begin{pmatrix} 0.0224330125 & -0.0003945405 & 0.0000654469 \\ -0.0003945405 & 0.0099722201 & -0.0002361998 \\ 0.0000654469 & -0.0002361998 & 0.0072418858 \end{pmatrix}\end{aligned}$$

Now what? First, we need to identify the transformation we're applying to our estimates ($\hat{\phi}_1$, $\hat{\phi}_2$, and $\hat{\phi}_3$). In this case, the transformation (which we'll call Y) is simple - it is the product of the three estimated survival rates. Conveniently, this makes differentiating the transformation straightforward.

So, here is the variance estimator, in full:

$$\widehat{\text{var}}(Y) \approx \begin{bmatrix} \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_1}\right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_2}\right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_3}\right) \end{bmatrix} \cdot \widehat{\Sigma} \cdot \begin{bmatrix} \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_1}\right) \\ \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_2}\right) \\ \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_3}\right) \end{bmatrix}$$

Each of the partial derivatives is easy enough for this example. Since $\hat{Y} = \hat{\phi}_1\hat{\phi}_2\hat{\phi}_3$, then $\partial\hat{Y}/\partial\hat{\phi}_1 = \hat{\phi}_2\hat{\phi}_3$. And so on.

So,

$$\begin{aligned}\widehat{\text{var}}(Y) &\approx \begin{bmatrix} \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_1}\right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_2}\right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_3}\right) \end{bmatrix} \cdot \widehat{\Sigma} \cdot \begin{bmatrix} \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_1}\right) \\ \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_2}\right) \\ \left(\frac{\partial(\hat{Y})}{\partial\hat{\phi}_3}\right) \end{bmatrix} \\ &= \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix} \cdot \widehat{\Sigma} \cdot \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix}\end{aligned}$$

OK, what about the variance-covariance matrix? Well, from the preceding we see that

$$\widehat{\text{cov}}(Y) = \begin{pmatrix} \text{var}(\phi_1) & \text{cov}(\phi_1, \phi_2) & \text{cov}(\phi_1, \phi_3) \\ \text{cov}(\phi_1, \phi_1) & \text{var}(\phi_2) & \text{cov}(\phi_2, \phi_3) \\ \text{cov}(\phi_3, \phi_1) & \text{cov}(\phi_3, \phi_2) & \text{var}(\phi_3) \end{pmatrix}$$

$$= \begin{pmatrix} 0.0224330125 & -0.0003945405 & 0.0000654469 \\ -0.0003945405 & 0.0099722201 & -0.0002361998 \\ 0.0000654469 & -0.0002361998 & 0.0072418858 \end{pmatrix}$$

Thus,

$$\begin{aligned} \widehat{\text{var}}(Y) &\approx \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix} \cdot \hat{\Sigma} \cdot \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix} \\ &= \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_3) & (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix} \cdot \begin{pmatrix} \text{var}(\phi_1) & \text{cov}(\phi_1, \phi_2) & \text{cov}(\phi_1, \phi_3) \\ \text{cov}(\phi_1, \phi_1) & \text{var}(\phi_2) & \text{cov}(\phi_2, \phi_3) \\ \text{cov}(\phi_3, \phi_1) & \text{cov}(\phi_3, \phi_2) & \text{var}(\phi_3) \end{pmatrix} \cdot \begin{bmatrix} (\hat{\phi}_2\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_3) \\ (\hat{\phi}_1\hat{\phi}_2) \end{bmatrix} \end{aligned}$$

Clearly, this expression is getting more and more 'impressive' as we progress. Here is the resulting expression (written in piecewise fashion to make it easier to see the basic pattern):

$$\begin{aligned} \widehat{\text{var}}(Y) &\approx \hat{\phi}_2^2 \hat{\phi}_3^2 (\widehat{\text{var}}_1) \\ &\quad + 2\hat{\phi}_2 \hat{\phi}_3^2 \hat{\phi}_1 (\widehat{\text{cov}}_{1,2}) \\ &\quad + 2\hat{\phi}_2^2 \hat{\phi}_3 \hat{\phi}_1 (\widehat{\text{cov}}_{1,3}) \\ &\quad + \hat{\phi}_1^2 \hat{\phi}_3^2 (\widehat{\text{var}}_2) \\ &\quad + 2\hat{\phi}_1^2 \hat{\phi}_3 \hat{\phi}_2 (\widehat{\text{cov}}_{2,3}) \\ &\quad + \hat{\phi}_1^2 \hat{\phi}_2^2 (\widehat{\text{var}}_3) \end{aligned}$$

Whew - a lot of work (and if you think this equation looks 'impressive', try it using a second-order Taylor series approximation!). But, under some assumptions, the Delta method does rather well in allowing you to derive an estimate of the sampling variance for functions of random variables (or, as we've described, functions of estimated parameters). So, after substituting in our estimates for ϕ_i and the variances and covariances, our estimate for the sampling variance of the product $\hat{Y} = (\hat{\phi}_1\hat{\phi}_2\hat{\phi}_3)$ is (approximately) 0.0025565.

Example (2) - variance of estimate of reporting rate

In some cases animals are tagged or banded to estimate a "reporting rate" - the proportion of banded animals reported, given that they were killed and retrieved by a hunter or angler (see chapter 9 for more details). Thus, N_c animals are tagged with normal (control) tags and, of these, R_c are recovered the first year following release. The recovery rate of control animals is merely R_c/N_c and we denote this as f_c .

Another group of animals, of size N_r , are tagged with reward tags; these tags indicate that some amount of money (say, \$50) will be given to people reporting these special tags. It is assumed that \$50

is sufficient to ensure that all such tags will be reported, thus these serve as a basis for comparison and the estimation of a reporting rate. The recovery probability for the reward tagged animals is merely R_r / N_r , where R_r is the number of recoveries of reward-tagged animals the first year following release. We denote this recovery probability as f_r .

The estimator of the *reporting rate* is a ratio of the *recovery rates* and we denote this as λ . Thus,

$$\hat{\lambda} = \frac{\hat{f}_c}{\hat{f}_r}$$

Now, note that both recovery probabilities are binomials. Thus,

$$\widehat{\text{var}}(f_c) = \frac{\hat{f}_c(1 - \hat{f}_c)}{N_c} \quad \widehat{\text{var}}(f_r) = \frac{\hat{f}_r(1 - \hat{f}_r)}{N_r}$$

In this case, the samples are independent, thus $\text{cov}(f_c, f_r)$ and the sampling variance-covariance matrix is diagonal:

$$\begin{pmatrix} \widehat{\text{var}}(f_c) & 0 \\ 0 & \widehat{\text{var}}(f_r) \end{pmatrix}$$

Next, we need the derivatives of λ with respect to f_c and f_r :

$$\frac{\partial \hat{\lambda}}{\partial \hat{f}_c} = \frac{1}{\hat{f}_r} \quad \frac{\partial \hat{\lambda}}{\partial \hat{f}_r} = -\frac{\hat{f}_c}{\hat{f}_r^2}$$

Thus,

$$\widehat{\text{var}}(\lambda) \approx \begin{pmatrix} \frac{1}{\hat{f}_r} & -\frac{\hat{f}_c}{\hat{f}_r^2} \\ \frac{1}{\hat{f}_r} & -\frac{\hat{f}_c}{\hat{f}_r^2} \end{pmatrix} \begin{pmatrix} \widehat{\text{var}}(f_c) & 0 \\ 0 & \widehat{\text{var}}(f_r) \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{f}_r} \\ -\frac{\hat{f}_c}{\hat{f}_r^2} \end{pmatrix}$$

Example (3) - variance of back-transformed estimates - simple

The basic idea behind this worked example was introduced back in Chapter 6 - in that chapter, we demonstrated how we can 'back-transform' from the estimate of β on the logit scale to an estimate of some parameter θ (e.g., ϕ or p) on the probability scale (which is bounded $[0, 1]$). But, we're clearly also interested in an estimate of the variance (precision) of our estimate, on both scales. Your first thought might be to simply back-transform from the link function (in our example, the logit link), to the probability scale, just as we did above. But, as discussed in chapter 6, this does not work.

For example, consider the male Dipper data. Using the logit link, we fit model $\{\phi, p.\}$ to the data - no time-dependence for either parameter. Let's consider only the estimate for $\hat{\phi}$. The estimate for β for ϕ is 0.2648275. Thus, our estimate of $\hat{\phi}$ on the probability scale is

$$\hat{\phi} = \frac{e^{0.2648275}}{1 + e^{0.2648275}} = \frac{1.303206}{2.303206} = 0.5658226$$

which is exactly what **MARK** reports (to within rounding error).

But, what about the variance? Well, if we look at the β estimates, **MARK** reports that the standard error for the estimate of β corresponding to survival is 0.1446688. If we simply back-transform this from the logit scale to the probability scale, we get

$$\begin{aligned}\widehat{\text{SE}} &= \frac{e^{0.1446688}}{1 + e^{0.1446688}} \\ &= \frac{1.155657}{2.155657} = 0.5361043\end{aligned}$$

However, **MARK** reports the estimated standard error for ϕ as 0.0355404, which isn't even remotely close to our back-transformed value of 0.5361043.

What has happened? Well, hopefully you now realize that you're 'transforming' the estimate from one scale (logit) to another (probability). And, since you're working with a 'transformation', you need to use the Delta method to estimate the variance of the back-transformed parameter. Since

$$\widehat{\phi} = \frac{e^{\widehat{\beta}}}{1 + e^{\widehat{\beta}}}$$

then

$$\begin{aligned}\widehat{\text{var}}(\widehat{\phi}) &\approx \left(\frac{\partial \widehat{\phi}}{\partial \widehat{\beta}}\right)^2 \times \widehat{\text{var}}(\widehat{\beta}) \\ &= \left(\frac{e^{\widehat{\beta}}}{1 + e^{\widehat{\beta}}} - \frac{(e^{\widehat{\beta}})^2}{1 + (e^{\widehat{\beta}})^2}\right)^2 \times \widehat{\text{var}}(\widehat{\beta}) \\ &= \left(\frac{e^{\widehat{\beta}}}{(1 + e^{\widehat{\beta}})^2}\right)^2 \times \widehat{\text{var}}(\widehat{\beta})\end{aligned}$$

It is worth noting that if

$$\widehat{\phi} = \frac{e^{\widehat{\beta}}}{1 + e^{\widehat{\beta}}}$$

then it can be easily shown that

$$\widehat{\phi}(1 - \widehat{\phi}) = \frac{e^{\widehat{\beta}}}{(1 + e^{\widehat{\beta}})^2}$$

which is the derivative of ϕ with respect to β . So, we could rewrite our expression for the variance of $\widehat{\phi}$ conveniently as

$$\widehat{\text{var}}(\widehat{\phi}) \approx \left(\frac{e^{\widehat{\beta}}}{(1 + e^{\widehat{\beta}})^2}\right)^2 \times \widehat{\text{var}}(\widehat{\beta}) = (\widehat{\phi}(1 - \widehat{\phi}))^2 \times \widehat{\text{var}}(\widehat{\beta})$$

From **MARK**, the estimate of the SE for $\widehat{\beta}$ was 0.1446688. Thus, the estimate of $\widehat{\text{var}}(\widehat{\beta})$ is $0.1446688^2 = 0.02092906$. Given the estimate of $\widehat{\beta}$ of 0.2648275, we substitute into the preceding expression, which

yields

$$\begin{aligned}\widehat{\text{var}}(\hat{\phi}) &\approx \left(\frac{e^{\hat{\beta}}}{(1 + e^{\hat{\beta}})^2} \right)^2 \times \widehat{\text{var}}(\hat{\beta}) \\ &= 0.0603525 \times 0.02092906 = 0.001263\end{aligned}$$

So, the estimated SE for $\hat{\phi}$ is $\sqrt{0.001263} = 0.0355404$, which is what is reported by **MARK** (again, within rounding error).

begin sidebar

SE and 95% CI

The standard approach to calculating 95% confidence limits for some parameter θ is $\theta \pm (1.96 \times \text{SE})$. Is this how **MARK** calculates the 95% CI on the real probability scale? Well, take the example we just considered - the estimated SE for $\hat{\phi} = 0.5658226$ was $\sqrt{0.001263} = 0.0355404$. So, you might assume that the 95% CI on the real probability scale would be $0.5658226 \pm (2 \times 0.0355404)$ - [0.4947418, 0.6369034].

However, this is not what is reported by **MARK** - [0.4953193, 0.6337593], which is quite close, but not exactly the same. Why the difference? The difference is because **MARK** first calculated the 95% CI on the logit scale, before back-transforming to the real probability scale. So, for our estimate of $\hat{\phi}$, the 95% CI on the logit scale for $\hat{\beta} = 0.2648275$ is [-0.0187234, 0.5483785], which, when back-transformed to the real probability scale is [0.4953193, 0.6337593], which is what is reported by **MARK**. In this case, the very small difference between the two CI's is because the parameter estimate was quite close to 0.5. In such cases, not only will the 95% CI be nearly the same (for estimates of 0.5, it will be identical), but they will also be symmetrical.

However, because the logit transform is not linear, the *reconstituted* 95% CI will not be symmetrical around the parameter estimate, especially for parameters estimated near the [0, 1] boundaries. For example, consider the estimate for $\hat{p} = 0.9231757$. On the logit scale, the 95% CI for the β corresponding to p (SE=0.5120845) is [1.4826128, 3.4899840]. The back-transformed CI is [0.8149669, 0.9704014]. This CI is clearly **not** symmetric around $\hat{p} = 0.9231757$. Essentially the degree of asymmetry is a function of how close the estimated parameter is to either the 0 or 1 boundary. Further, the estimated variance for \hat{p}

$$\begin{aligned}\widehat{\text{var}}(\hat{p}) &\approx (\hat{p}(1 - \hat{p}))^2 \times \widehat{\text{var}}(\hat{\beta}) \\ &= (0.9231757(1 - 0.9231757))^2 \times 0.262231 = 0.001319\end{aligned}$$

yields an estimated SE of 0.036318 on the normal probability scale (which is what is reported by **MARK**).

Estimating the 95% CI on the normal probability scale simply as $0.9231757 \pm (2 \times 0.036318)$ yields [0.85054, 0.99581], which is clearly quite a bit different, and more symmetrical, than what is reported by **MARK** (from above, [0.8149669, 0.9704014]).

MARK uses the back-transformed CI to ensure that the reported CI is bounded [0, 1]. As the estimated parameter approaches either the 0 or 1 boundary, the degree of asymmetry in the back-transformed 95% CI that **MARK** reports will increase.

end sidebar

Got it? Well, as a final test, consider the following, more difficult, example of back-transforming the CI from a model fit using individual covariates.

Example (3) - variance of back-transformed estimates - somewhat harder

In Chapter 6 we considered the analysis of variation in the apparent survival of the European Dipper, as a function of whether or not there was a flood in the sampling area. Here, we will consider just the male Dipper data (the encounter data are contained in `ed_males.inp`). Recall that for these data, there are 7 sampling occasions (6 intervals), and that a flood occurred during the second and third intervals. For present purposes, we'll assume that encounter probability was constant over time, and that survival is a linear function of 'flood' or 'non-flood'. Using a logit link function, where 'flood' years were coded using a '1', and non-flood years were coded using a '0', the linear model for survival on the logit scale is

$$\text{logit}(\phi) = 0.4267863 - 0.5066372(\text{flood})$$

So, in a flood year,

$$\begin{aligned} \text{logit}(\hat{\phi}_{\text{flood}}) &= 0.4267863 - 0.5066372(\text{flood}) \\ &= 0.4267863 - 0.5066372(1) \\ &= -0.0798509 \end{aligned}$$

Back-transforming onto the real probability scale,

$$\hat{\phi}_{\text{flood}} = \frac{e^{-0.0798509}}{1 + e^{-0.0798509}} = 0.48005$$

which is precisely what is reported by **MARK**.

Now, what about the estimated variance for ϕ_{flood} ? First, what is our 'transformation function' (Y)? Simple – it is the 'back-transform' of the linear equation on the logit scale. Given that

$$\begin{aligned} \text{logit}(\hat{\phi}) &= \beta_0 + \beta_1(\text{flood}) \\ &= 0.4267863 - 0.5066372(\text{flood}) \end{aligned}$$

then the back-transform function Y is

$$Y = \frac{e^{0.4267863 - 0.5066372(\text{flood})}}{1 + e^{0.4267863 - 0.5066372(\text{flood})}}$$

Second, since our transformation clearly involves multiple parameters (β_0, β_1), the estimate of the variance is given to first-order by

$$\begin{aligned} \text{var}(\mathbf{Y}) &\approx \mathbf{D}\Sigma\mathbf{D}^T \\ &= \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right) \cdot \hat{\Sigma} \cdot \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right)^T \end{aligned}$$

Given our linear (transformation) equation, then the vector of partial derivatives is (we've transposed it to make it easily fit on the page):

$$\begin{aligned} & \left[\left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_0} \right) \quad \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_1} \right) \right]^T \\ &= \left[\begin{array}{c} \frac{e^{\beta_0 + \beta_1(\text{flood})}}{1 + e^{\beta_0 + \beta_1(\text{flood})}} - \frac{(e^{\beta_0 + \beta_1(\text{flood})})^2}{(1 + e^{\beta_0 + \beta_1(\text{flood})})^2} \\ \frac{\text{flood} \times e^{\beta_0 + \beta_1(\text{flood})}}{1 + e^{\beta_0 + \beta_1(\text{flood})}} - \frac{\text{flood} \times (e^{\beta_0 + \beta_1(\text{flood})})^2}{(1 + e^{\beta_0 + \beta_1(\text{flood})})^2} \end{array} \right] \end{aligned}$$

While this is fairly 'ugly' looking, the structure is quite straightforward - the only difference between the 2 elements of the vector is that the numerator of both terms (on either side of the minus sign) are multiplied by 1, and flood, respectively. Where do these scalar multipliers come from? They're simply the partial derivatives of the linear model (we'll call it Y) on the logit scale

$$Y = \text{logit}(\hat{\phi}) = \beta_0 + \beta_1(\text{flood})$$

with respect to each of the parameters (β_i) in turn. In other words, $\partial Y / \partial \beta_0 = 1$, and $\partial Y / \partial \beta_1 = \text{flood}$.

Substituting in our estimates for $\hat{\beta}_0 = 0.4267863$ and $\hat{\beta}_1 = -0.5066372$, and setting flood=1 (to indicate a 'flood year') yields

$$\begin{aligned} & \left[\left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_0} \right) \quad \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_1} \right) \right] \\ &= \left[\begin{array}{cc} 0.249602 & 0.249602 \end{array} \right] \end{aligned}$$

From the **MARK** output (after exporting to a dBase file - and not to the Notepad - in order to get full precision), the full V-C matrix for the parameters β_0 and β_1 is

$$\begin{pmatrix} 0.0321405326 & -0.0321581167 \\ -0.0321581167 & 0.0975720877 \end{pmatrix}$$

So,

$$\begin{aligned} \widehat{\text{var}}(\hat{Y}) &\approx \left[\begin{array}{cc} 0.249602 & 0.249602 \end{array} \right] \times \begin{pmatrix} 0.0321405326 & -0.0321581167 \\ -0.0321581167 & 0.0975720877 \end{pmatrix} \times \left[\begin{array}{c} 0.249602 \\ 0.249602 \end{array} \right] \\ &= 0.0040742678 \end{aligned}$$

So, the estimated SE for $\widehat{\text{var}}$ for the reconstituted value of survival for an individual during a 'flood year' is $\sqrt{0.0040742678} = 0.0638300$, which is what is reported by **MARK** (to within rounding error).

Example (4) - variance of back-transformed estimates - a bit harder still

Recall that in Chapter 11, we considered analysis of the effect of various functions of mass (specifically, mass, and mass²) on the survival of a hypothetical species of bird (the simulated data are in file `indcov1.inp`). The linear function relating survival to mass and mass2, on the logit scale, is

$$\text{logit}(\phi) = 0.256732 + 1.1750358(\text{mass}_s) - 1.0554864(\text{mass}_s^2)$$

Note that for the two mass terms, there is a small subscript 's' - reflecting the fact that these are 'standardized' masses. Recall that we standardized the covariates by subtracting the mean of the covariate, and dividing by the standard deviation (the use of standardized or non-standardized covariates is discussed at length in Chapter 11).

Thus, for each individual in the sample, the estimated survival probability (on the logit scale) for that individual, given its mass, is given by

$$\text{logit}(\phi) = 0.256732 + 1.1750358 \left(\frac{m - \bar{m}}{SD_m} \right) - 1.0554864 \left(\frac{m^2 - \bar{m}^2}{SD_{m^2}} \right)$$

In this expression, m refers to mass and m^2 refers to mass2. The output from **MARK** (preceding page) actually gives you the mean and standard deviations for both covariates: for mass, mean = 109.97, and SD = 24.79, while for mass2, the mean = 12707.46, and the SD = 5532.03. The 'value' column shows the standardized values for mass and mass2 (0.803 and 0.752) for the first individual in the data file.

Let's look at an example. Suppose the mass of the bird was 110 units. Thus $\text{mass} = 110$, $\text{mass}^2 = 110^2 = 12100$. Thus,

$$\text{logit}(\phi) = 0.2567 + 1.17504 \left(\frac{(110 - 109.97)}{24.79} \right) - 1.0555 \left(\frac{(12100 - 12707.46)}{5532.03} \right) = 0.374.$$

So, if $\text{logit}(\phi) = 0.374$, then the reconstituted estimate of ϕ , transformed back from the logit scale is

$$\frac{e^{0.374}}{1 + e^{0.374}} = 0.592$$

Thus, for an individual weighing 110 units, the expected annual survival probability is approximately 0.5925 (which is what **MARK** reports if you use the 'User specify covariate' option).

OK, but what about the variance (and corresponding SE) for this estimate? First, what is our 'transformation function' (Y)? Easy - it is the 'back-transform' of the linear equation on the logit scale. Given that

$$\begin{aligned} \text{logit}(\hat{\phi}) &= \beta_0 + \beta_1(\text{mass}_s) + \beta_2(\text{mass}_s^2) \\ &= 0.2567 + 1.17505(\text{mass}_s) - 1.0555(\text{mass}_s^2) \end{aligned}$$

then the back-transform function Y is

$$Y = \frac{e^{0.2567 + 1.17505(\text{mass}_s) - 1.0555(\text{mass}_s^2)}}{1 + e^{0.2567 + 1.17505(\text{mass}_s) - 1.0555(\text{mass}_s^2)}}$$

As in the preceding example, since our transformation clearly involves multiple parameters ($\beta_0, \beta_1, \beta_2$), the estimate of the variance is given by

$$\begin{aligned} \text{var}(\mathbf{Y}) &\approx \mathbf{D}\Sigma\mathbf{D}^T \\ &= \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right) \cdot \hat{\Sigma} \cdot \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})} \right)^T \end{aligned}$$

Given our linear (transformation) equation (from above) then the vector of partial derivatives is (we've substituted m for mass and m_2 for mass2, and transposed it to make it easily fit on the page):

$$\begin{aligned} &\left[\left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_0} \right) \quad \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_1} \right) \quad \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_2} \right) \right]^T \\ &= \begin{bmatrix} \frac{e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}}{1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}} - \frac{(e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2}{(1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2} \\ \frac{m \times e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}}{1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}} - \frac{m \times (e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2}{(1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2} \\ \frac{m_2 \times e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}}{1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)}} - \frac{m_2 \times (e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2}{(1 + e^{\beta_0 + \beta_1(m) + \beta_2(m_2)})^2} \end{bmatrix} \end{aligned}$$

Again, while this is fairly 'ugly' looking (even more so than the previous example), the structure is again quite straightforward – the only difference between the 3 elements of the vector is that the numerator of both terms (on either side of the minus sign) are multiplied by 1, m , and m_2 , respectively, which are simply the partial derivatives of the linear model (we'll call it Y) on the logit scale

$$Y = \text{logit}(\hat{\phi}) = \beta_0 + \beta_1(m_s) + \beta_2(m_s^2)$$

with respect to each of the parameters (β_i) in turn. In other words, $\partial Y / \partial \beta_0 = 1$, $\partial Y / \partial \beta_1 = m$, and $\partial Y / \partial \beta_2 = m_2$.

So, now that we have our vectors of partial derivatives of the transformation function with respect to each of the parameters, we can simplify things considerably by substituting in the standardized values for m and m_2 , and the estimated parameter values ($\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$). For a mass of 110 g, the standardized values for mass and mass2 are

$$\text{mass}_s = \left(\frac{110 - 109.97}{24.79} \right) = 0.0012102 \quad \text{mass2}_s = \left(\frac{12100 - 12707.46}{5532.03} \right) = -0.109808$$

The estimates for $\hat{\beta}_i$ we read directly from **MARK**: $\hat{\beta}_0 = 0.2567333, \hat{\beta}_1 = 1.1750545, \hat{\beta}_2 = -1.0554864$.

Substituting in these estimates for $\hat{\beta}_i$ and the standardized m and m_2 values (from the previous page) into our vector of partial derivatives (above) yields

$$\left[\begin{array}{ccc} \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_0} \right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_1} \right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_2} \right) \end{array} \right]^T = \begin{bmatrix} 0.24145 \\ 0.00029 \\ -0.02651 \end{bmatrix}$$

From the **MARK** output (after exporting to a dBase file - and **not** to the Notepad - in order to get full precision), the full V-C matrix for the β parameters is

$$\begin{pmatrix} 0.0009006921 & -0.0004109710 & 0.0003662359 \\ -0.0004109710 & 0.0373887267 & -0.0364250288 \\ 0.0003662359 & -0.0364250288 & 0.0362776933 \end{pmatrix}$$

So,

$$\begin{aligned} \widehat{\text{var}}(\hat{Y}) &\approx \begin{bmatrix} 0.24145 & 0.00029 & -0.02651 \end{bmatrix} \\ &\times \begin{pmatrix} 0.0009006921 & -0.0004109710 & 0.0003662359 \\ -0.0004109710 & 0.0373887267 & -0.0364250288 \\ 0.0003662359 & -0.0364250288 & 0.0362776933 \end{pmatrix} \times \begin{bmatrix} 0.24145 \\ 0.00029 \\ -0.02651 \end{bmatrix} \\ &= 0.00007387 \end{aligned}$$

So, the estimated SE for $\widehat{\text{var}}$ for the reconstituted value of survival for an individual weighing 110 g is $\sqrt{0.00007387} = 0.00860$, which is what is reported by **MARK** (again, to within rounding error).

It is important to remember that the estimated variance will vary depending on the mass you use - the estimate of the variance for a 110 g individual (0.00007387) will differ from the estimated variance for a (say) 120 g individual. For a 120 g individual, the standardized values of mass and mass^2 are 0.4045568999 and 0.3059519429, respectively. Based on these values, then

$$\left[\begin{array}{ccc} \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_0} \right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_1} \right) & \left(\frac{\partial(\hat{Y})}{\partial\hat{\beta}_2} \right) \end{array} \right]^T = \begin{bmatrix} 0.23982 \\ 0.08871 \\ 0.07337 \end{bmatrix}$$

Given the variance covariance-matrix for this model (shown above), then

$$\widehat{\text{var}}(\hat{Y}) \approx \mathbf{D}\Sigma\mathbf{D}^T = 0.000074214$$

Thus, the estimated SE for $\widehat{\text{var}}$ for the reconstituted value of survival for an individual weighing 120 g is $\sqrt{0.000074214} = 0.008615$, which is what is reported by **MARK** (again, within rounding error).

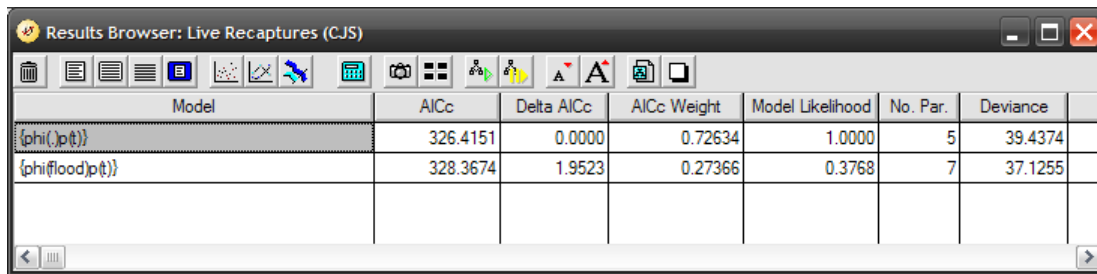
Note that this value for the SE for a 120 g individual (0.008615) differs from the SE estimated for a 110 g individual (0.008600), albeit not by much (the small difference here is because this is a very large simulated data set based on a deterministic model - see Chapter 11 for details). Since each weight would have it's own estimated survival, and associated estimated variance and SE, to generate a curve showing the reconstituted values and their SE, you'd need to iteratively calculate $\mathbf{D}\Sigma\mathbf{D}^T$ over a range of weights. We'll leave it to you to figure out how to handle the programming if you want to

do this on your own. For the less ambitious, **MARK** now has the capacity to do much of this for you - you can output the 95% CI 'data' over a range of individual covariate values to a spreadsheet (see section 11.5 in Chapter 11).

B.5. Delta method and model averaging

In the preceding examples, we focused on the application of the Delta method to transformations of parameter estimates from a single model. However, as introduced in Chapter 4 - and emphasized throughout the remainder of this book - we're often interested in accounting for model selection uncertainty by using model-averaged values. There is no major complication for application of the Delta method to model-averaged parameter values - you simply need to make sure you use model-averaged values for each element of the calculations.

We'll demonstrate this using analysis of the male dipper data (`ed_male.inp`). Suppose that we fit 2 candidate models to these data: $\{\phi.p_t\}$ and $\{\phi_{flood}p_t\}$. In other words, a model where survival is constant over time, and a model where survival is constrained to be a function of a binary 'flood' variable' (see section 6.4 of Chapter 6). Here are the results of fitting these 2 models to the data:



Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance
$\{\phi(.p_t)\}$	326.4151	0.0000	0.72634	1.0000	5	39.4374
$\{\phi(flood)p_t\}$	328.3674	1.9523	0.27366	0.3768	7	37.1255

As expected (based on the analysis of these data presented in Chapter 6), we see that there is some evidence of model selection uncertainty - the model where survival is constant over time has roughly 2-3 times the weight as does a model where survival is constrained to be a function of the binary 'flood' variable.

The model averaged values for each interval are shown below:

	1	2	3	4	5	6
<i>estimate</i>	0.5673	0.5332	0.5332	0.5673	0.5673	0.5673
<i>SE</i>	0.0441	0.0581	0.0581	0.0441	0.0441	0.0441

Now, suppose we want to derive the best estimate of the probability of survival over (say) the first 3 intervals. Clearly, all we need to do is take the product of the 3 model-averaged values corresponding to the first 3 intervals:

$$(0.5673 \times 0.5332 \times 0.5332) = 0.1613$$

In other words, our best estimate of the probability that a male dipper would survive from the start of the time series to the end of the third interval is 0.1613.

What about the standard error of this product? Here, we use the Delta method. Recall that

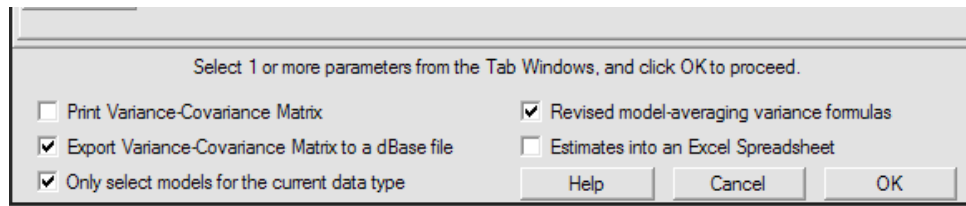
$$\text{var}(\mathbf{Y}) \approx \mathbf{D}\mathbf{\Sigma}\mathbf{D}^T$$

which we write out more fully as

$$\begin{aligned}\text{var}(\mathbf{Y}) &\approx \mathbf{D}\Sigma\mathbf{D}^T \\ &= \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})}\right) \cdot \hat{\Sigma} \cdot \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})}\right)^T\end{aligned}$$

where Y is some linear or nonlinear function of the parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \dots$. For this example, Y is the product of the survival estimates.

So, the first thing we need to do is to generate the estimated variance-covariance matrix for the model averaged survival estimates. This is easy enough to do - in the 'Model Averaging Parameter Selection' window, you simply need to 'Export Variance-Covariance Matrix to a dBase file' - you do this by checking the appropriate check box (lower-left, as shown at the top of the next page):



The 'rounded' values which would be output to the Notepad (or whatever editor you've specified) are shown at the top of the next page. Recall that the variance-covariance matrix of estimates is given on the diagonal and below (whereas the correlation matrix of the estimates is shown above the diagonal). (*Note*: remember that for the actual calculations you need the full precision variance-covariance matrix from the exported dBase file).

Unconditional Variance-Covariance Matrix of Model Averaged Estimates
 Variance-Covariance matrix of estimates on diagonal and below,
 Correlation matrix of estimates above diagonal.

	1	2	3
1	0.00194	0.04923	0.04923
2	0.00013	0.00337	1.00000
3	0.00013	0.00337	0.00337

All that remains is to substitute our model-averaged estimates for (i) $\hat{\phi}$ and (ii) the variance-covariance matrix, into $\text{var}(\mathbf{Y}) \approx \mathbf{D}\Sigma\mathbf{D}^T$.

Thus,

$$\begin{aligned}\text{var}(\mathbf{Y}) &\approx \mathbf{D}\Sigma\mathbf{D}^T \\ &= \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})}\right) \cdot \hat{\Sigma} \cdot \left(\frac{\partial(\hat{Y})}{\partial(\hat{\theta})}\right)^T\end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} (\bar{\hat{\phi}}_2 \bar{\hat{\phi}}_3) & (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_3) & (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_2) \end{bmatrix} \cdot \hat{\Sigma} \cdot \begin{bmatrix} (\bar{\hat{\phi}}_2 \bar{\hat{\phi}}_3) \\ (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_3) \\ (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_2) \end{bmatrix} \\
&= \begin{bmatrix} (\bar{\hat{\phi}}_2 \bar{\hat{\phi}}_3) & (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_3) & (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_2) \end{bmatrix} \cdot \begin{pmatrix} \text{var}(\hat{\phi}_1) & \text{cov}(\hat{\phi}_1, \hat{\phi}_2) & \text{cov}(\hat{\phi}_1, \hat{\phi}_3) \\ \text{cov}(\hat{\phi}_1, \hat{\phi}_2) & \text{var}(\hat{\phi}_2) & \text{cov}(\hat{\phi}_2, \hat{\phi}_3) \\ \text{cov}(\hat{\phi}_3, \hat{\phi}_1) & \text{cov}(\hat{\phi}_3, \hat{\phi}_2) & \text{var}(\hat{\phi}_3) \end{pmatrix} \cdot \begin{bmatrix} (\bar{\hat{\phi}}_2 \bar{\hat{\phi}}_3) \\ (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_3) \\ (\bar{\hat{\phi}}_1 \bar{\hat{\phi}}_2) \end{bmatrix} \\
&= \begin{bmatrix} 0.284303069 & 0.3024783390 & 0.3024783390 \end{bmatrix} \\
&\quad \times \begin{pmatrix} 0.0019410083 & 0.0001259569 & 0.0001259569 \\ 0.0001259569 & 0.0033727452 & 0.0033727423 \\ 0.0001259569 & 0.0033727423 & 0.0033727452 \end{pmatrix} \times \begin{bmatrix} 0.284303069 \\ 0.3024783390 \\ 0.3024783390 \end{bmatrix} \\
&= 0.001435
\end{aligned}$$

B.6. Summary

In this appendix, we've briefly introduced a convenient, fairly straightforward method for deriving an estimate of the sampling variance for transformations of one or more variables. Such transformations are quite commonly encountered when using **MARK**, and having a method to derive estimates of the sampling variances is convenient. The most straightforward method – based on a first-order Taylor series expansion – is known generally as the 'Delta method'. However, as we saw, the first-order Taylor series approximation may not always be appropriate, especially if the transformation is highly non-linear, and if there is significant variation in the data. In such case, you may have to resort to higher-order approximations, or numerically intensive bootstrapping approaches.