The Delta Method: Uncertainty for Derived Quantities

What it is, where it comes from, and why it works (an NRES 746 reference handout)

A fitted model hands you estimates of its parameters and, from the Hessian, their variances and covariances. The number you care about is often something other than the fitted parameters themselves: a prediction, a ratio of parameters, a back-transformed parameter (e.g., on the original response scale rather than the log-transformed response scale), a population growth rate built from several vital rates (often called a "derived quantity").

The **delta method** turns the uncertainty in the parameters into an approximate standard error for any smooth function of them (i.e., a function that has a derivative!). It needs two things: the variance-covariance matrix of the fitted parameters (usually computed from the Hessian at the MLE, which is returned from an optimization routine like `optim()`), and the derivatives of the function. Bolker covers it in Section 7.5.2 and Appendix A.7 of *Ecological Models and Data in R*; this handout adds more details, including a derivation and a geometric intuition.

This handout assumes you know about derivatives (including partial derivatives) and how to compute the variance of a sum. The example we'll use is the power-law fit to R's built-in `trees` data, timber volume V against diameter D for 31 black cherry trees:

$$V_i \sim \text{Normal}\left(aD_i^b,\ \sigma^2\right).$$

Fitting by maximum likelihood with `optim()`, estimating $\log\sigma$ so that σ stays positive, gives $\hat{a} = 0.087$, $\hat{b} = 2.24$, $\widehat{\log\sigma} = 1.157$, and, from the inverse of the Hessian, standard errors 0.021, 0.086, and 0.127. The correlation between $\hat{a}$ and $\hat{b}$ is -0.998.

1. Why we need the delta method

Two derived quantities arise with this model.

- **The residual SD.** We estimated $\log\sigma$, but we want to report $\sigma = e^{\log\sigma} = 3.18$ cubic feet, with a standard error. But… the Hessian only gives the standard error of $\log\sigma$.
- **A prediction.** The expected volume of a 13-inch tree is $\hat{a} \cdot 13^{\hat{b}} = 26.84$ cubic feet. It's easy to compute this prediction, but how uncertain is this prediction?

The parameters $\hat{a}$ and $\hat{b}$ are both uncertain, so the obvious solution is to combine their two variances. But if we treat the parameters as independent (calculated in Section 2) we get an estimated standard error of 8.73 cubic feet – a third of the prediction itself! That's absurd

for a curve that runs through the middle of 31 tightly scattered points. The missing piece is the correlation of -0.998: when the data push $\hat{a}$ up, they push $\hat{b}$ down (and vice versa), so for trees within the range of the data, the predicted volume barely moves. The delta method uses that covariance information and gives a much more reasonable estimate of 0.66 cubic feet.

Other places the same question comes up:

- back-transforming any parameter fit on a log, logit, or atanh scale (correlation parameters are commonly fitted on the atanh scale, and tanh is the back-transformation that puts them back on the correlation scale, -1 to 1);
- a ratio of curve parameters, such as the initial slope a/b of a Michaelis-Menten curve $ax/(b+x)$;
- the predictor value where a logistic curve crosses 0.5 (like we saw in the LD50 example in lab), which is $-\beta_0/\beta_1$;
- mean survival time from a Weibull fit (Bolker's example in 7.5.2; Weibull is a very common probability distribution in survival analysis);
- standard errors of GLM predictions on the response scale. `predict(..., type = "response", se.fit = TRUE)` uses the delta method behind the scenes.

2. The formula

Bolker's notation. For a function $f(a, b)$ of two estimates (Bolker's Eq. 7.5.4, with each derivative evaluated at the estimates):

$$\operatorname{Var}\Big(f(\hat{a},\hat{b})\Big) \approx \operatorname{Var}(\hat{a})\left(\frac{\partial f}{\partial a}\right)^2 + \operatorname{Var}(\hat{b})\left(\frac{\partial f}{\partial b}\right)^2 + 2\operatorname{Cov}(\hat{a},\hat{b})\,\frac{\partial f}{\partial a}\frac{\partial f}{\partial b}.$$

With one parameter, only the first term is left, so the standard error of f is the absolute derivative times the standard error of the estimate:

$$\operatorname{Var}(f(\hat{\theta})) \approx f'(\hat{\theta})^2 \operatorname{Var}(\hat{\theta}), \qquad \operatorname{SE}(f) \approx |f'(\hat{\theta})|\operatorname{SE}(\hat{\theta}).$$

Matrix notation. For k parameters, collect the partial derivatives of f into a column vector, the gradient of f, $\mathbf{g} = (\partial f/\partial\theta_1, \ldots, \partial f/\partial\theta_k)^T$, and write Σ for the variance-covariance matrix of the estimates (the inverse of the Hessian, `solve(fit$hessian)` after `optim()`). This is not the gradient you've seen in optimization, the slopes of the log-likelihood with respect to each parameter (which are zero at the MLE). Here $\mathbf{g}$ holds the slopes of the derived quantity f, and the log-likelihood's curvature enters only through Σ. Then

$$\operatorname{Var}\Big(f(\hat{\boldsymbol{\theta}})\Big) \approx \mathbf{g}^T \Sigma\, \mathbf{g} = \sum_{i=1}^{k}\sum_{j=1}^{k} g_i\, g_j\, \Sigma_{ij}.$$

The double sum has a squared derivative times a variance for each diagonal entry ($i = j$) and a cross term for each off-diagonal pair. With $k = 2$, the two off-diagonal entries are equal and give Bolker's $2\operatorname{Cov}$ term, so the two notations say the same thing.

Note that the matrix version is more compact and more general! This is one of the many reasons statisticians love matrix notation!! Note also that the format (vector transpose times matrix times vector) is called a **quadratic form**, and occurs frequently in statistics because in some sense it defines a multivariate "distance" or "size/magnitude". For example, the kernel of the multivariate normal distribution is a quadratic form representing the effective/standardized "distance" from the mean (called **Mahalanobis distance**), where probability density is lower as this standardized distance increases.

One parameter: the residual SD. With $f(\theta) = e^\theta$ and $\theta = \log\sigma$, the derivative is $e^\theta = \sigma$, so

$$\mathrm{Var}(\hat\sigma) \approx f'(\hat\theta)^2\,\mathrm{Var}(\hat\theta) = \hat\sigma^2\,\mathrm{Var}(\widehat{\log\sigma}) = 3.18^2 \times 0.0161 = 0.163.$$

Taking the square root gives the standard error:

$$\mathrm{SE}(\hat\sigma) \approx \sqrt{0.163} = 0.404,$$

which is the same as $\hat\sigma \cdot \mathrm{SE}(\widehat{\log\sigma}) = 3.18 \times 0.127$. Fitting σ directly and inverting the Hessian gives exactly the same number (at the MLE, the two curvatures differ by exactly the squared derivative, so estimating on the log scale costs nothing!).

Two parameters: the prediction at 13 inches. For $f(a, b) = a \cdot 13^b$, the derivative with respect to a is just 13^b. For b, the exponent is the variable, so first rewrite the power with the natural base (e): $13^b = e^{b\log 13}$ (log is the natural log throughout). The chain rule then gives

$$\frac{\partial}{\partial b}\,13^b = \frac{\partial}{\partial b}\,e^{b\log 13} = \log 13 \cdot e^{b\log 13} = 13^b \log 13,$$

and f carries an extra factor of a. At the estimates,

$$\frac{\partial f}{\partial a} = 13^b = 309.9, \qquad \frac{\partial f}{\partial b} = a\,13^b \log 13 = 68.8.$$

Then

$$\begin{aligned}\mathbf{g}^T\Sigma\,\mathbf{g} &= 309.9^2(4.29 \times 10^{-4}) + 68.8^2(0.00739) + 2(309.9)(68.8)(-0.00178)\\ &= 41.216 + 35.036 - 75.814 = 0.438,\end{aligned}$$

a standard error of 0.66. The two variance terms are large, and the negative covariance term cancels nearly all of their combined variance. Treating $\hat a$ and $\hat b$ as independent (erroneously!) would mean dropping that term:

$$41.216 + 35.036 = 76.252, \qquad \sqrt{76.252} = 8.73,$$

which results in the heavily inflated standard error we introduced in Section 1 (8.73).

In R, `emdbook::deltavar()` does the whole calculation. You give it the function itself (not its derivatives), the estimates, and Σ; it differentiates the function with R's symbolic differentiator `D()`, and switches to numerical derivatives if `D()` can't handle it. For the 13-inch prediction, `deltavar(a * 13^b, meanval = c(a = a_hat, b = b_hat), Sigma = Sigma_ab)` returns 0.438. To do the calculation yourself without working out the derivatives, approximate each one by a finite difference, such as $(f(a+h, b) - f(a, b))/h$ for a small h, and put the resulting gradient into $\mathbf{g}^T\Sigma\,\mathbf{g}$.

3. Where it comes from

The delta method can be derived from (1) a first-order Taylor expansion (to linearize the target function of the random estimator) followed by (2) an application of the formula for computing the variance of a linear function (which you already know).

One parameter. Near the true value θ_0, replace f with its linearization, where the difference between $f(\hat{\theta})$ and $f(\theta_0)$ is approximated by the first derivative at θ_0 multiplied by the difference between $\hat{\theta}$ and θ_0 (another way to say this is that the approximation lies on the **tangent line** to the function at the true value θ_0):

$$f(\hat{\theta}) \approx f(\theta_0) + f'(\theta_0)\,(\hat{\theta} - \theta_0).$$

The right side is a constant plus a constant times $\hat{\theta}$. Adding a constant doesn't change a variance, and multiplying by c multiplies it by c^2, so

$$\mathrm{Var}\Big(f(\hat{\theta})\Big) \approx f'(\theta_0)^2\,\mathrm{Var}(\hat{\theta}).$$

In practice, we don't know θ_0, so the derivative is evaluated at $\hat{\theta}$, which is close to it when the standard error is small.

Several parameters. The tangent line becomes a tangent *plane*:

$$f(\hat{\boldsymbol{\theta}}) \approx f(\boldsymbol{\theta}_0) + \sum_i g_i\,(\hat{\theta}_i - \theta_{0,i}) = f(\boldsymbol{\theta}_0) + \mathbf{g}^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

The variance of a weighted sum $\sum_i g_i X_i$ is $\sum_i \sum_j g_i g_j\,\mathrm{Cov}(X_i, X_j)$, which is $\mathbf{g}^T \Sigma\, \mathbf{g}$. For two parameters,

$$\mathrm{Var}(g_a X_a + g_b X_b) = g_a^2 \mathrm{Var}(X_a) + g_b^2 \mathrm{Var}(X_b) + 2 g_a g_b \mathrm{Cov}(X_a, X_b),$$

which is the 2-parameter formula Bolker presents.

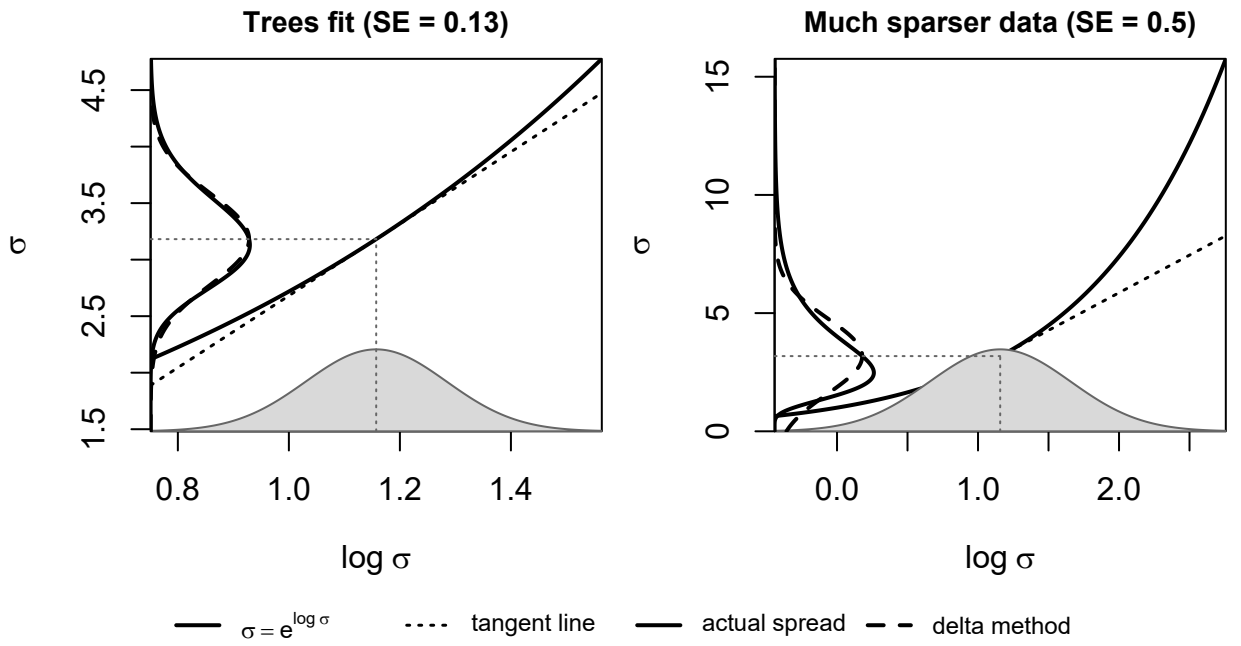
The shape of the distribution, too. The sampling distribution of the MLEs is approximately (with a large enough sample) multivariate normal (see the companion handout at the end), and a linear function of normal variables is also normal. That gives two approximations working together. If the estimates are close to normal (a large-sample result), and f is close to its tangent plane over the range where they fall (the delta method's assumption), then $f(\hat{\boldsymbol{\theta}})$ is close to a linear function of normal variables, and so close to normal itself:

$$f(\hat{\boldsymbol{\theta}}) \approx \mathrm{Normal}\Big(f(\boldsymbol{\theta}_0),\ \mathbf{g}^T \Sigma\, \mathbf{g}\Big),$$

and a 95% interval is $f(\hat{\boldsymbol{\theta}}) \pm 1.96\,\mathrm{SE}$. For the 13-inch tree, that's [25.5, 28.1] cubic feet.

Everything rests on one assumption: **over the range where the estimates plausibly fall, f is close to its tangent line (or plane).** The smaller the standard errors, the smaller that range, and the better the approximation.

4. The geometry with one parameter



Read each panel from the bottom up.

- **Bottom (grey):** the sampling distribution of $\widehat{\log \sigma}$, normal with the standard error from the Hessian.
- **The curve** carries each possible estimate up to a value of σ. The spread of the estimates along the bottom becomes a spread of $\hat{\sigma}$ along the side. The solid curve on the left edge is that actual spread.
- **The tangent line** at the estimate is the delta method's stand-in for the curve. A line stretches every distance by the same factor, its slope, so it turns a normal distribution into a normal distribution with an SD multiplied by $|f'|$. The dashed curve on the left edge is that result.

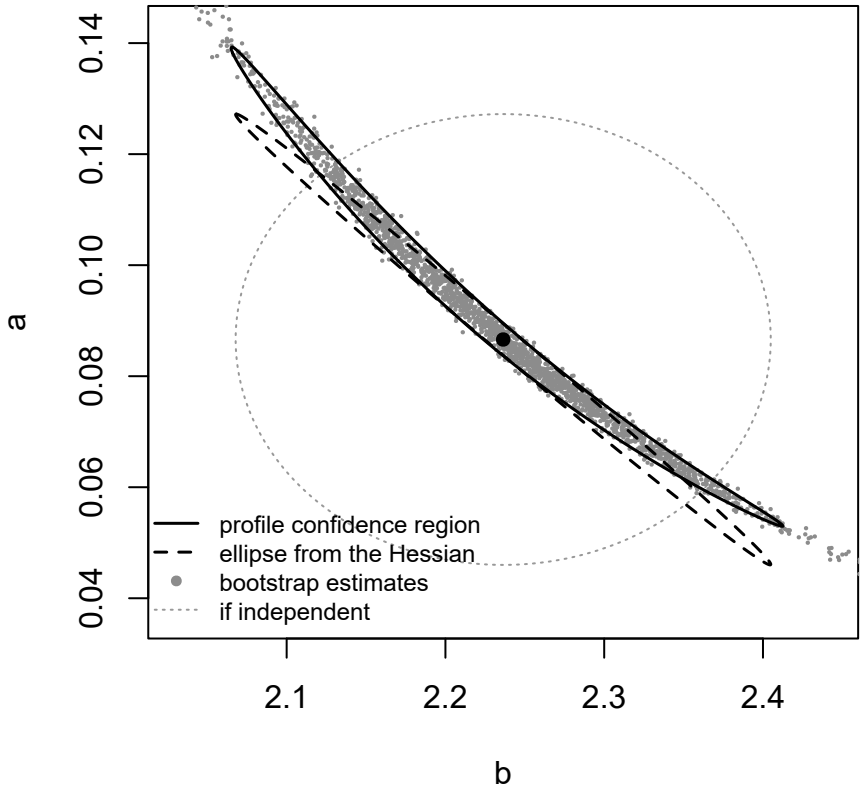
On the left, the estimates fall in a narrow band where the curve and its tangent are hard to tell apart, and the two side curves nearly coincide. On the right, the band is wide enough that the curve bends away from its tangent. The upper half of the band is stretched more than the lower half, so the actual spread is skewed to the right, and the symmetric delta-method normal misses it. The slope of f sets the width of the result; the bending of f over the band sets how wrong the delta method is.

5. The geometry with two parameters

With two parameters, the estimates live in the (b, a) plane, and the prediction is a landscape over that plane. Its contour lines connect parameter pairs that give the same prediction. The delta method makes two straight-line approximations here: it treats the region of plausible

estimates as an ellipse, and it treats the prediction's contours as straight, evenly spaced lines.

The region of plausible estimates.



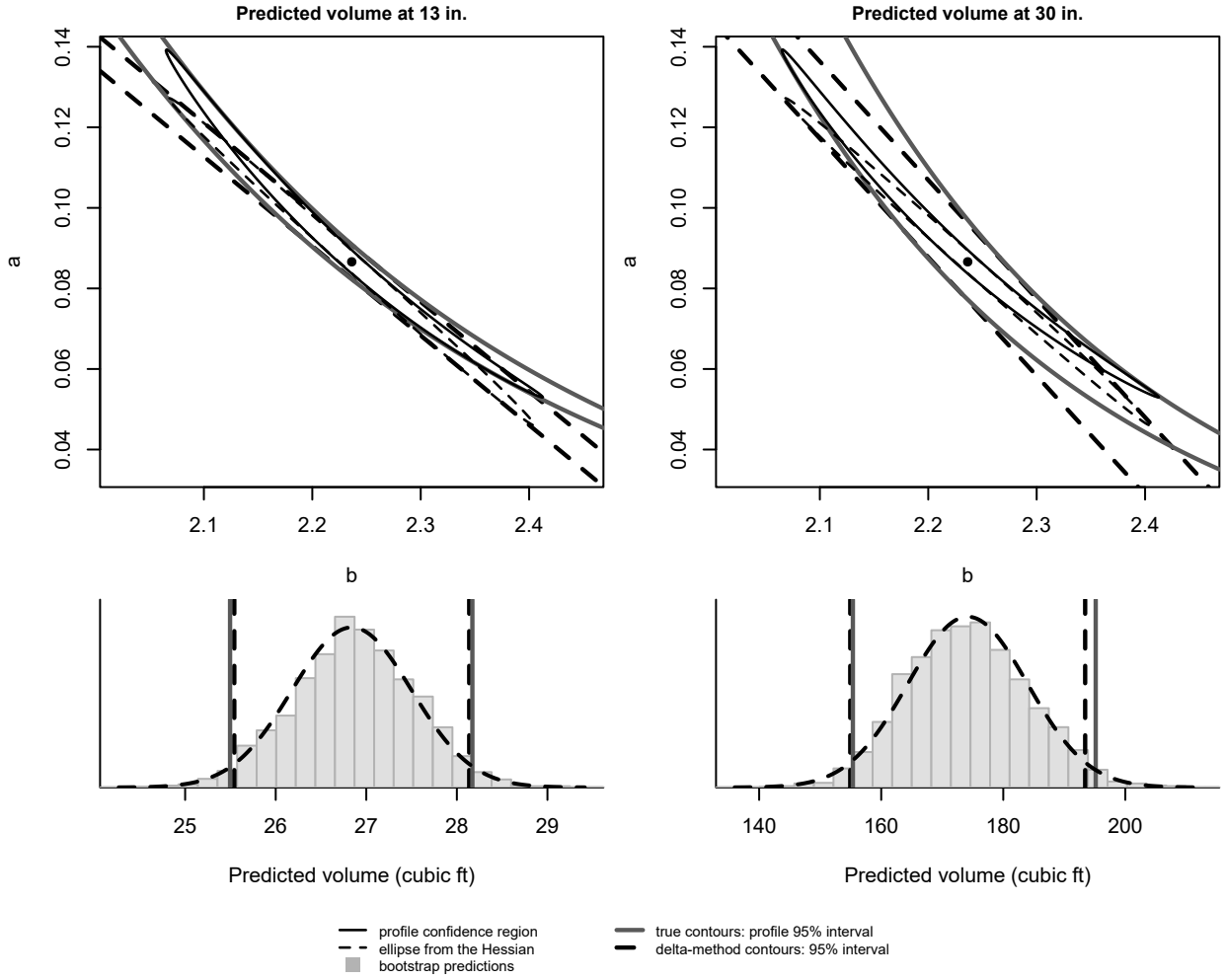
- **The solid curve** is the profile likelihood confidence region (the confidence region, for short): every (a, b) pair whose deviance, with σ profiled out, is within $3.84 = 1.96^2$ of the minimum. It is thin and curved. The data constrain the combination of a and b far better than either alone.
- **The dashed ellipse** is the Hessian's quadratic approximation to that region, the set of points where $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \Sigma^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1.96^2$. It is drawn at 1.96 rather than at the joint 95% radius (2.45) so that its shadow on any direction is a 95% interval for one quantity.
- **The grey dots** are 2,000 bootstrap estimates (see Section 6). They follow the curved confidence region, not the straight ellipse.
- **The faint dotted ellipse** is what you'd assume if $\hat{a}$ and $\hat{b}$ were independent: the same variances with no tilt. Its width across the confidence region is where the inflated SE of 8.73 in Section 1 comes from. The covariance is what tilts the ellipse to line up with the confidence region.

From the region to a prediction. In each top panel of the next figure, the two regions appear again, with two pairs of contours:

- **Straight dashed lines** are the delta method's contours at $\hat{f} \pm 1.96\,\mathrm{SE}$. They just touch the ellipse.
- **Curved grey lines** are the true prediction contours that just touch the confidence

region. The range of predictions across that region is the profile likelihood 95% interval for the prediction.

The strips under them put both intervals on the prediction scale, over a histogram of the bootstrap predictions and the delta method's normal curve (dashed). The table gives the same intervals, with the bootstrap percentile interval for comparison.



Diameter (in)	Delta method	Profile likelihood	Bootstrap
13	[25.54, 28.14]	[25.50, 28.17]	[25.57, 28.14]
30	[154.93, 193.42]	[155.36, 195.15]	[155.90, 193.53]

At 13 inches (the middle of the data), the contours run almost along the confidence region, so the region spans only a narrow range of predictions: SE 0.66 cubic feet, and the two pairs of lines nearly coincide. At 30 inches (beyond the largest tree, 20.6), the contours have rotated to cross the confidence region at an angle. The SE grows to 9.82 cubic feet, about 6% of the prediction instead of 2.5%. The true contours also bend across the confidence region, so

the profile interval shifts up (upper limit 195.15 vs. 193.42), the same right skew as in the Section 4 figure. Extrapolating points the prediction in a direction the data don't constrain.

Why the ellipse still works. The ellipse misses the confidence region at both ends, but that miss runs along the contours, so it barely changes the prediction. The delta method only uses the ellipse's width across the contours, and that width is close to right (and it is often close to right- that's why the method is so commonly used!). The gradient of f, $\mathbf{g}$, points straight across the straight contours, and its length says how close together they are. $\mathbf{g}^T \Sigma \, \mathbf{g}$ is then the variance of the ellipse in that one direction: squash the ellipse onto the gradient's direction, take the variance of that shadow, and scale by $|\mathbf{g}|^2$. An ellipse's shadow depends on its tilt, and the tilt is the covariance.

6. Checking it, and when it breaks

A simulation check. Simulate 2,000 new data sets from the fitted model, refit each one, and compute the prediction from each refit (a parametric bootstrap). The SD of those predictions is a check that doesn't use the delta method:

Diameter (in)	Prediction (cubic ft)	Delta-method SE	SE ignoring covariance	Bootstrap SD
8	9.1	0.56	2.71	0.57
13	26.8	0.66	8.73	0.67
21	78.4	2.24	27.82	2.26
30	174.2	9.82	65.80	9.87

The delta method matches the bootstrap closely across the data and at 30 inches. Leaving out the covariance overstates the uncertainty by a factor of 5 to 13.

When it breaks. Each case is a way for f to stop looking like its tangent over the plausible range of the estimates.

- **Large standard errors relative to the curvature of f.** Check out the right panel of the Section 4 figure: the real distribution of $f(\hat{\theta})$ is skewed, and $\pm 1.96\,\mathrm{SE}$ is symmetric. Bolker's Weibull example (7.5.5) shows the cost: for mean survival time, the delta-method interval is [7.4, 22.4] and the profile likelihood interval is [9.2, 27.0].
- **A derivative near zero.** At a peak or trough of f, the tangent is flat and the delta method reports almost no uncertainty, whatever the estimates do.
- **Ratios with an uncertain denominator.** For $-\beta_0/\beta_1$, if the interval for β_1 comes near 0, the ratio can be huge or change sign, and no straight line describes it.
- **Bad covariances.** The method is only as good as Σ. If the Hessian approximation is poor (small samples, boundaries, the wrong model), the delta method inherits the problem.

What to do instead.

- **Change the scale first.** For a positive quantity, apply the delta method to $\log f$, build the interval there, and exponentiate the endpoints. The result stays positive and is skewed in the right direction. For σ: $\exp(1.157 \pm 1.96 \times 0.127)$ gives [2.48, 4.08], compared with [2.39, 3.97] on the raw scale.
- **Profile likelihood** for f, by reparameterizing so f is a parameter (Bolker 7.5.1).
- **Simulate.** Use a parametric bootstrap, as above, or draw parameter sets from $\text{MVN}(\hat{\boldsymbol{\theta}}, \Sigma)$ and compute f for each (Bolker's population prediction intervals, 7.5.3). The second option inherits the normal approximation's shape for the parameters, so its answer depends on the scale they're on. Here, drawing (a, b) from the MVN gives an SD of 1.53 cubic feet at 13 inches, against the bootstrap's 0.67. The real sampling distribution of $(\hat{b}, \hat{a})$ bends along the curved contours (the bootstrap dots in the first Section 5 figure), and a straight ellipse can't follow it. On the $(\log a, b)$ scale, where those contours are straight lines, the draws give 0.66.
- **Go Bayesian.** With posterior draws, compute f for each draw and summarize; nothing needs to be linearized.

A note on Bolker's Appendix A.7. It lists the product of two independent variables as a case where the formula is exact. It isn't quite: the exact variance has an extra term,

$$\text{Var}(XY) = \mu_Y^2 \,\text{Var}(X) + \mu_X^2 \,\text{Var}(Y) + \text{Var}(X)\,\text{Var}(Y).$$

The delta method drops the last term, which is small when both coefficients of variation are small. The formula is exact for linear functions, as A.7 also says.

Summary

Idea	Picture	Formula
One parameter	tangent line stretches the spread by its slope	$\text{SE}(f) \approx \lvert f'(\hat{\theta})\rvert \,\text{SE}(\hat{\theta})$
Two parameters (Bolker)	spread of the ellipse across the contours	$\text{Var}(\hat{a}) f_a^2 + \text{Var}(\hat{b}) f_b^2 + 2\,\text{Cov}(\hat{a}, \hat{b}) f_a f_b$, with $f_a = \partial f / \partial a$
k parameters	shadow of the ellipse on the gradient direction	$\mathbf{g}^T \Sigma\, \mathbf{g}$
Where it comes from	first-order Taylor expansion	$f(\hat{\theta}) \approx f(\theta_0) + \mathbf{g}^T(\hat{\theta} - \theta_0)$
Interval	normal, from the normal MLE	$f(\hat{\theta}) \pm 1.96\,\text{SE}$
When it fails	f bends over the plausible range	use a log scale, profile, or simulate

The delta method assumes the function is straight over the range where the estimates fall, the same kind of assumption as the parabola behind Hessian intervals. A companion handout, *Why Likelihood Intervals Work*, explains where Σ comes from and why the MLE is approximately normal.

References

Bolker, B. M. 2008. *Ecological Models and Data in R.* Princeton University Press, Princeton, NJ. (Section 7.5 and Appendix A.7.)

Oehlert, G. W. 1992. A note on the delta method. *The American Statistician* 46: 27-29.

Ver Hoef, J. M. 2012. Who invented the delta method? *The American Statistician* 66: 124-127.