Where AIC's 2k Penalty Comes From

NRES 746 reference notes: the derivation behind Bolker's footnote in Section 6.6.2

The Akaike Information Criterion, AIC $= -2 \log \mathcal{L} + 2k$, penalizes a model by twice its number of parameters, k. Where does that $2k$ come from? Bolker's *Ecological Models and Data in R* answers in a one-paragraph footnote (Section 6.6.2; printed p. 210 in the published book, p. 277 in the free draft). The footnote is compressed, and as written it comes up short: it does one Taylor expansion, which yields $k/2$, then drops a factor of $\frac{1}{2}$ to reach k. The full argument needs **two** expansions, each contributing $k/2$.

These notes do the derivation in full:

1. what AIC is trying to estimate (Section 1);
2. a one-parameter example where every step can be done with ordinary algebra and the answer comes out exact (Section 2);
3. the matrix tools needed for k parameters (Section 3);
4. the two halves of the penalty with matrices (Sections 4 and 5);
5. a simulation check (Section 6), what can break (Section 7), and a step-by-step reading of Bolker's footnote (Section 8).

The background assumed is a first course in calculus (derivatives and Taylor expansions) and in statistics (expected values and variances). For the main idea, Sections 1, 2, and 6 are enough; Section 2 needs only algebra. Sections 3 to 5 make the same argument for any number of parameters, using matrices, and Section 3 introduces the matrix tools from scratch.

1. What AIC is trying to estimate

Notation

Bolker's footnote uses x for both the data and the parameters, which is most of why it's hard to read. Here:

Symbol	Meaning
f	the true process that generates data (unknown)
$g(y \mid \theta)$	the candidate model: probability of data y given parameters θ
k	number of parameters; θ is a column vector of length k
y	the observed dataset, used to fit the model
y^*	an imaginary **new** dataset, same size, drawn independently from f

Symbol	Meaning
$\mathrm{NLL}_y(\theta)$	negative log-likelihood computed from dataset y: $-\log g(y \mid \theta)$
$\hat{\theta}$	the MLE from the observed data; it minimizes $\mathrm{NLL}_y(\theta)$
θ_0	the "best" parameter value for this model (defined below)

The NLL is written as a function of the parameter alone, with a subscript marking which dataset it's computed from.

Kullback-Leibler distance, minus the part we can't compute

The K-L distance from the truth f to the model at parameter θ is the expected log ratio of their predictions, averaged over the true distribution of the data:

$$D(\theta) = E_f\left[\log \frac{f(y^*)}{g(y^* \mid \theta)}\right] = \underbrace{E_f[\log f(y^*)]}_{\text{same for every model}} + E_f[-\log g(y^* \mid \theta)].$$

The first term involves only the truth, so it's the same constant for every candidate model and drops out of any comparison. What's left is the expected NLL on new data:

$$Q(\theta) = E_{y^*}\Big[\mathrm{NLL}_{y^*}(\theta)\Big].$$

Smaller Q means closer to the truth. Define θ_0 as the value that minimizes Q: the best this model family can do. If the model family contains the truth, θ_0 is the true parameter value; otherwise it's the closest approximation (sometimes called the "pseudo-true" value). It is not an MLE. It's a fixed, unknown target, just like a true parameter value.

The target, and why the obvious estimate is biased

We never get to use θ_0. We use $\hat{\theta}$, fitted from y. So the fair measure of a fitted model is how well $\hat{\theta}$ predicts **new** data, averaged over the datasets we might have collected:

$$T = E_y\Big[\, Q(\hat{\theta}) \,\Big] = E_y\Big[\, E_{y^*}\Big[\mathrm{NLL}_{y^*}(\hat{\theta})\Big]\Big].$$

Two expectations appear because two things are random: the data used to fit (y, which determines $\hat{\theta}$), and the data used to judge (y^*). Bolker's "E[E[...]]" is this double expectation.

The obvious estimate of T is the observed minimum, $\mathrm{NLL}_y(\hat{\theta})$. But it uses the same data twice, once to choose $\hat{\theta}$ and once to score it. $\hat{\theta}$ was chosen to make NLL_y as small as possible, so it fits y better than it will fit y^*, and $\mathrm{NLL}_y(\hat{\theta})$ is too small on average. The more parameters, the more freedom to fit noise in y, and the larger this optimism.

The whole result: the average optimism is about k.

$$T \approx E_y\Big[\mathrm{NLL}_y(\hat{\theta})\Big] + k \qquad \Longrightarrow \qquad 2T \approx E\Big[\underbrace{2\,\mathrm{NLL}_y(\hat{\theta}) + 2k}_{\text{AIC}}\Big].$$

Multiplying by 2 just puts things on the deviance scale. So AIC is an (approximately unbiased) estimate of twice the expected out-of-sample NLL, which is the model-dependent part of the expected K-L distance.

The bridge between the observed $\mathrm{NLL}_y(\hat{\theta})$ and the target $Q(\hat{\theta})$ goes through θ_0, in two steps:

$$\underbrace{Q(\hat{\theta}) - Q(\theta_0)}_{\text{Half 1: cost of using } \hat{\theta}} \qquad \text{and} \qquad \underbrace{Q(\theta_0) - \mathrm{NLL}_y(\hat{\theta})}_{\text{Half 2: optimism of the in-sample fit}} .$$

Each averages about $k/2$.

2. Warm-up: one parameter, no matrices, exact answer

Suppose we measure snout-vent length (mm) of $n = 10$ lizards, each $y_i \sim \mathrm{Normal}(\mu, \sigma^2)$, and pretend $\sigma = 4$ is known (say, from a much larger study), so $k = 1$. The truth is $\mu_0 = 60$.

The observed NLL

$$\mathrm{NLL}_y(\mu) = n \log\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

The sum of squares splits around the sample mean $\bar{y}$. Add and subtract $\bar{y}$ inside the square, then expand $(a+b)^2 = a^2 + 2ab + b^2$ with $a = y_i - \bar{y}$ and $b = \bar{y} - \mu$:

$$\sum(y_i - \mu)^2 = \sum\Big[(y_i - \bar{y}) + (\bar{y} - \mu)\Big]^2 = \sum(y_i - \bar{y})^2 + 2\sum(y_i - \bar{y})(\bar{y} - \mu) + \sum(\bar{y} - \mu)^2.$$

The last term has no i in it, so summing it n times gives $n(\bar{y} - \mu)^2$. In the middle (cross) term, $(\bar{y} - \mu)$ also doesn't depend on i, so it factors out of the sum:

$$2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - \mu) = 2(\bar{y} - \mu)\sum_{i=1}^{n}(y_i - \bar{y}).$$

The remaining sum is zero, because deviations from the mean always cancel:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} = n\bar{y} - n\bar{y} = 0,$$

using $\sum y_i = n\bar{y}$ (the definition of the mean). So the cross term drops out, leaving

$$\sum(y_i - \mu)^2 = \sum(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2.$$

Substituting into the NLL:

$$\mathrm{NLL}_y(\mu) = C + \frac{S}{2\sigma^2} + \frac{n(\bar{y} - \mu)^2}{2\sigma^2}, \qquad C = n\log\left(\sigma\sqrt{2\pi}\right),\ S = \sum(y_i - \bar{y})^2.$$

This is a parabola in μ with its minimum at $\hat{\mu} = \bar{y}$.

The expected NLL on new data

For one new observation, $E[(y_i^* - \mu)^2] = \mathrm{Var}(y_i^*) + (E[y_i^*] - \mu)^2 = \sigma^2 + (\mu_0 - \mu)^2$. Summing over n observations:

$$Q(\mu) = C + \frac{n\left[\sigma^2 + (\mu - \mu_0)^2\right]}{2\sigma^2} = C + \frac{n}{2} + \frac{n(\mu - \mu_0)^2}{2\sigma^2}.$$

Another parabola, with the same curvature, but its minimum is at the truth, μ_0.

Half 1: the cost of using $\hat{\mu}$ instead of μ_0

Plug $\hat{\mu} = \bar{y}$ into Q and subtract $Q(\mu_0)$:

$$Q(\hat{\mu}) - Q(\mu_0) = \frac{n(\bar{y} - \mu_0)^2}{2\sigma^2} = \frac{1}{2}\left(\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right)^2 = \frac{1}{2}Z^2,$$

where Z is a standard normal (since $\bar{y} \sim \mathrm{Normal}(\mu_0, \sigma^2/n)$). $E[Z^2] = \mathrm{Var}(Z) = 1$, so this half averages $\frac{1}{2}$.

Half 2: the optimism of the in-sample fit

Evaluate the observed NLL at the truth and at the MLE:

$$\mathrm{NLL}_y(\mu_0) - \mathrm{NLL}_y(\hat{\mu}) = \frac{n(\bar{y} - \mu_0)^2}{2\sigma^2} = \frac{1}{2}Z^2.$$

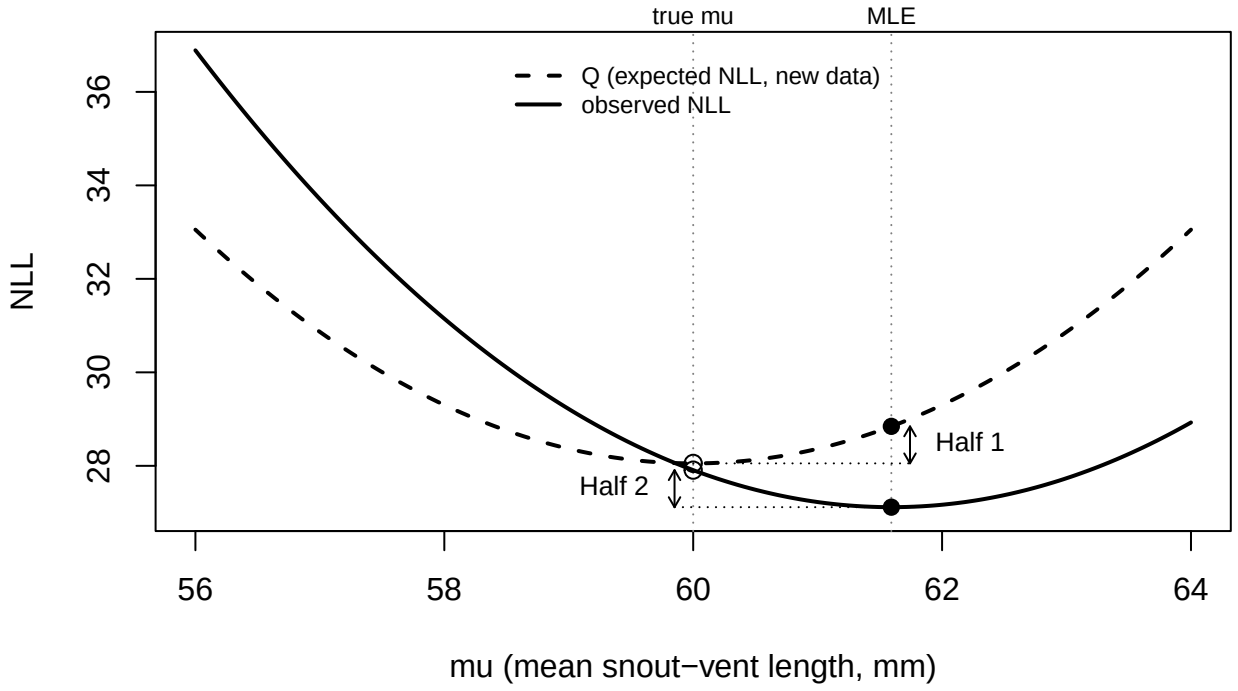
The same quantity again, averaging $\frac{1}{2}$. (This is also half the likelihood ratio statistic, which Wilks' theorem says is χ^2_1 with mean 1.) To connect it to $Q(\mu_0)$, take the expectation of $\mathrm{NLL}_y(\mu_0)$ directly: μ_0 is a fixed number, so this is just $Q(\mu_0)$, since y and y^* have the same distribution. Therefore

$$E\left[\mathrm{NLL}_y(\hat{\mu})\right] = Q(\mu_0) - \tfrac{1}{2}.$$

Putting them together

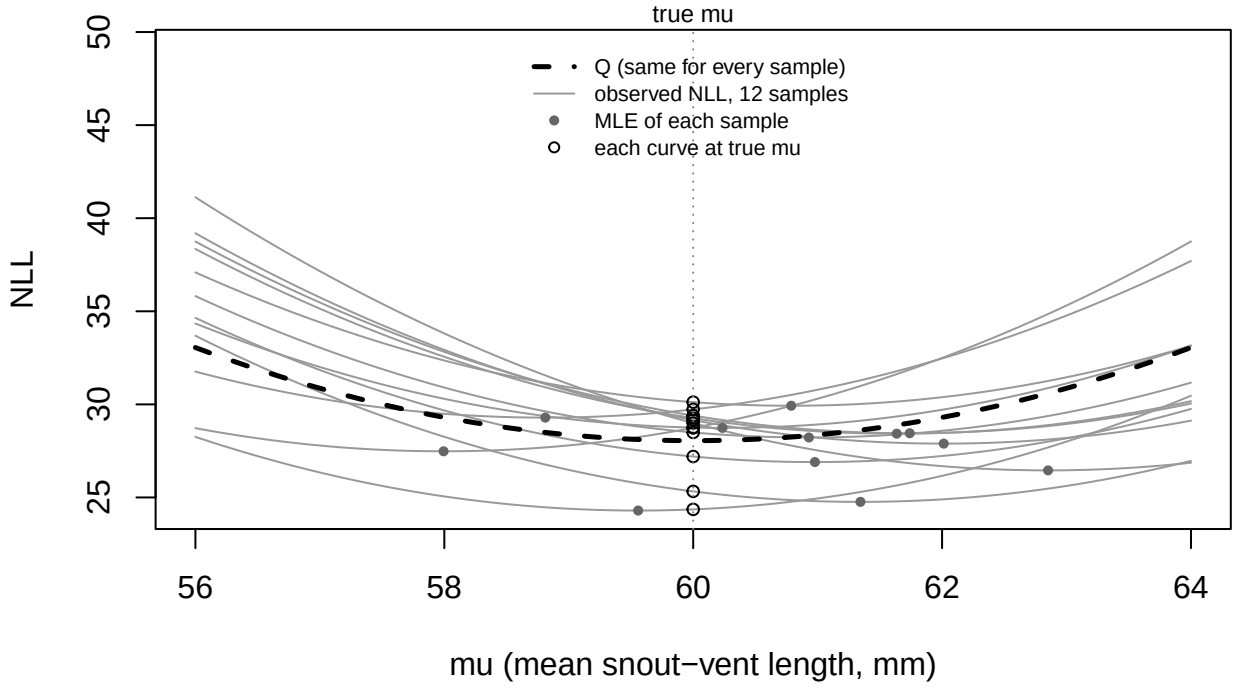
$$T = E[Q(\hat{\mu})] = Q(\mu_0) + \tfrac{1}{2} = E\left[\mathrm{NLL}_y(\hat{\mu})\right] + \tfrac{1}{2} + \tfrac{1}{2} = E\left[\mathrm{NLL}_y(\hat{\mu})\right] + 1.$$

With $k = 1$ the optimism is 1, and $2 \times 1 = 2k$. In this example there's no approximation anywhere: both NLL curves are parabolas, so the Taylor expansions in Sections 4 and 5 are exact.



The figure shows one simulated sample ($\bar{y} = 61.59$). The dashed curve Q is centered on the truth; the solid observed curve is centered on the MLE. Half 1 is how far up Q's wall the MLE sits. Half 2 is how far up the observed curve's wall the truth sits. In this one-parameter normal case they're the same height, $\frac{1}{2}Z^2$.

This sample was chosen so the two curves nearly meet at μ_0, which makes both halves easy to see. Nothing forces that. The figure below draws the observed curve for 12 more simulated samples around the same Q.



What stays put and what moves:

- Q **doesn't depend on the data.** It's the same dashed curve for every sample.
- **Each observed curve has the same shape as** Q (curvature n/σ^2 in this example), but its position is random. It slides sideways with $\bar{y}$; that sideways shift is what both halves measure. It also slides up and down, because the height at μ_0 depends on $\sum(y_i - \mu_0)^2$.
- **The vertical gap at** μ_0 **is large.** It equals $\left[\sum(y_i - \mu_0)^2/\sigma^2 - n\right]/2 = (\chi^2_n - n)/2$, with mean 0 and standard deviation $\sqrt{n/2} \approx 2.2$ NLL units (2.25 in 20,000 simulated samples). That's much larger than either half, which averages 0.5.
- **On average the observed curves sit on** Q**.** Averaging the observed NLL over all those simulated samples at each μ reproduces Q (largest difference along the curve: 0.04). This is the definition of Q, and it's why Half 2 only needs the expectation $E_y[\mathrm{NLL}_y(\mu_0)] = Q(\mu_0)$, never a match in any single sample.

3. Matrix tools for k parameters

With k parameters, the slopes and curvatures of Section 2 become a vector and a matrix. Everything below treats vectors as columns, and ^{T} means transpose (turn a column into a row).

Gradient and Hessian

For a function $h(\theta)$ of $\theta = (\theta_1, \ldots, \theta_k)^T$:

- The **gradient** ∇h is the column of first partial derivatives, $(\partial h/\partial\theta_1, \ldots, \partial h/\partial\theta_k)^T$. At a minimum, every entry is zero.
- The **Hessian** H is the $k \times k$ matrix of second partial derivatives, $H_{ij} = \partial^2 h/\partial\theta_i\,\partial\theta_j$. It's symmetric ($H_{ij} = H_{ji}$). Diagonal entries are the curvature along each parameter axis; off-diagonal entries describe how the surface twists (how the slope in one parameter changes as another changes).

Quadratic forms

For a vector Δ and a symmetric matrix A, the **quadratic form** $\Delta^T A \Delta$ is a single number:

$$\Delta^T A \Delta = \sum_{i=1}^{k} \sum_{j=1}^{k} \Delta_i\, A_{ij}\, \Delta_j.$$

For $k = 2$, written out:

$$\Delta^T A \Delta = A_{11}\Delta_1^2 + 2A_{12}\Delta_1\Delta_2 + A_{22}\Delta_2^2.$$

With $k = 1$ it's $A\Delta^2$, which is what appeared in Section 2.

Taylor expansion in several variables

The one-variable second-order expansion $h(\theta_0 + \Delta) \approx h(\theta_0) + h'(\theta_0)\Delta + \frac{1}{2}h''(\theta_0)\Delta^2$ becomes

$$h(\theta_0 + \Delta) \approx h(\theta_0) + \nabla h(\theta_0)^T \Delta + \tfrac{1}{2}\, \Delta^T H(\theta_0)\, \Delta.$$

The middle term, $\nabla h^T \Delta = \sum_i (\partial h/\partial \theta_i)\Delta_i$, is a single number (a row times a column). If θ_0 is a minimum, it's zero.

Inner and outer products

For a column vector Δ there are two different products with itself, and the difference matters for one step in Bolker's footnote (Section 8):

- **Inner product** $\Delta^T\Delta$ (row times column): a single number, $\sum_i \Delta_i^2$.
- **Outer product** $\Delta\Delta^T$ (column times row): a $k \times k$ matrix whose (i, j) entry is $\Delta_i\Delta_j$.

If $\Delta = \hat{\theta} - \theta_0$ and $E[\Delta] \approx 0$, the expected outer product is the **variance-covariance matrix** of the estimates:

$$\Sigma = E[\Delta\Delta^T], \qquad \Sigma_{ii} = E[\Delta_i^2] = \mathrm{Var}(\hat{\theta}_i), \qquad \Sigma_{ij} = E[\Delta_i\Delta_j] = \mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j).$$

The expected inner product is a single number, the sum of the variances, so it can't be Σ.

Trace, and the trace trick

The **trace** of a square matrix is the sum of its diagonal: $\mathrm{tr}(M) = \sum_i M_{ii}$. Two facts:

- **Trace of a product.** The i-th diagonal entry of AB is $\sum_j A_{ij}B_{ji}$, so $\mathrm{tr}(AB) = \sum_i \sum_j A_{ij}B_{ji}$. Swapping the names of i and j shows $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
- **Trace of the identity.** The $k \times k$ identity matrix I_k has ones on the diagonal, so $\mathrm{tr}(I_k) = k$. **This is where the k in AIC comes from.**

The **trace trick** turns the expectation of a quadratic form into a trace. Start from the double sum and regroup:

$$\Delta^T A \Delta = \sum_i \sum_j A_{ij}\, \Delta_j \Delta_i = \sum_i \sum_j A_{ij}\, (\Delta\Delta^T)_{ji} = \mathrm{tr}\Big(A\, \Delta\Delta^T\Big).$$

The last step is the trace-of-a-product formula with $B = \Delta\Delta^T$. Expectation is linear and A is fixed, so it passes inside the sums:

$$\boxed{E\Big[\Delta^T A \Delta\Big] = \mathrm{tr}\Big(A\, E[\Delta\Delta^T]\Big) = \mathrm{tr}(A\, \Sigma).}$$

(This holds for any random vector. If $E[\Delta] = b \neq 0$, then $E[\Delta\Delta^T] = \Sigma + bb^T$, but for an MLE the bias b is small in large samples and is ignored here.)

Check with numbers. Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$. Then

$$\operatorname{tr}(A\Sigma) = (2)(0.5) + (1)(0.1) + (1)(0.1) + (3)(0.2) = 1.8.$$

Writing out the $k = 2$ quadratic form instead: $E[2\Delta_1^2 + 2\Delta_1\Delta_2 + 3\Delta_2^2] = 2(0.5) + 2(0.1) + 3(0.2) = 1.8$, the same thing. Averaging $\Delta^T A \Delta$ over 200,000 simulated normal vectors with covariance Σ gives 1.805.

Information and the variance of the MLE

The **Fisher information matrix** for the whole dataset is the expected Hessian of the NLL at θ_0:

$$\mathcal{I} = E\Big[\nabla^2 \operatorname{NLL}_y(\theta_0)\Big],$$

the matrix version of the one-parameter curvature I. For one parameter, when the model is correct, $\operatorname{Var}(\hat{\theta}) \approx 1/I$, so the standard error is $1/\sqrt{I}$ (see the companion handout *Why Likelihood Intervals Work* for why). The matrix version is

$$\Sigma \approx \mathcal{I}^{-1},$$

where $\mathcal{I}^{-1}$ is the matrix inverse ($\mathcal{I}\mathcal{I}^{-1} = I_k$). This is the variance-covariance matrix R reports as the inverse of the Hessian from `optim(..., hessian = TRUE)`.

Sign conventions trip people up here. The Hessian of the NLL is positive at a minimum, and its expectation is the information. The Hessian of the **log-likelihood** is the negative of that. Bolker's V is the Hessian of the log-likelihood, so his $-V^{-1} \approx \Sigma$ is the same statement as $\mathcal{I}^{-1} \approx \Sigma$.

4. Half 1: the cost of using $\hat{\theta}$ instead of θ_0

Goal: show $E_y[Q(\hat{\theta})] \approx Q(\theta_0) + k/2$.

Step 1: expand Q around θ_0. Let $\Delta = \hat{\theta} - \theta_0$. Using the multivariate Taylor expansion from Section 3,

$$Q(\hat{\theta}) \approx Q(\theta_0) + \nabla Q(\theta_0)^T \Delta + \tfrac{1}{2}\, \Delta^T\, \nabla^2 Q(\theta_0)\, \Delta.$$

Step 2: the linear term vanishes. θ_0 was defined as the minimum of Q, so $\nabla Q(\theta_0) = 0$. This is where "the likelihood is flat" is used in this half: the **expected** NLL is flat at θ_0. (The observed NLL is not flat at θ_0; it's flat at $\hat{\theta}$.)

Step 3: the curvature is the information. Q is the expected NLL, and the derivative of an expectation is the expectation of the derivative:

$$\nabla^2 Q(\theta_0) = E\Big[\nabla^2 \operatorname{NLL}_{y^*}(\theta_0)\Big] = \mathcal{I}.$$

So

$$Q(\hat{\theta}) \approx Q(\theta_0) + \tfrac{1}{2}\, \Delta^T \mathcal{I}\, \Delta.$$

A note on what's random. $Q(\hat{\theta})$ means: take the fitted value $\hat{\theta}$ from y, hold it fixed, and average its NLL over new datasets y^*. The only thing still random is $\hat{\theta}$ (through y), which is why only Δ is random in this expression. This works because y^* is independent of y.

Step 4: average over y with the trace trick. $\mathcal{I}$ is a fixed matrix, so

$$E_y\Big[\Delta^T \mathcal{I}\, \Delta\Big] = \mathrm{tr}\Big(\mathcal{I}\, \Sigma\Big) \approx \mathrm{tr}\Big(\mathcal{I}\, \mathcal{I}^{-1}\Big) = \mathrm{tr}(I_k) = k.$$

Result:

$$\boxed{E_y\Big[Q(\hat{\theta})\Big] \approx Q(\theta_0) + \frac{k}{2}.}$$

Every parameter that has to be estimated adds, on average, half a unit of NLL on new data, regardless of whether it's a slope, an intercept, or a variance. That independence from the parameter's units is what the trace trick buys: $\mathcal{I}$ and Σ both depend on the units, but their product doesn't.

5. Half 2: the optimism of the in-sample fit

Goal: show $E_y[\mathrm{NLL}_y(\hat{\theta})] \approx Q(\theta_0) - k/2$.

Step 1: expand the observed NLL around $\hat{\theta}$, and evaluate it at θ_0. Now $\theta_0 = \hat{\theta} - \Delta$, so the step away from the expansion point is $-\Delta$:

$$\mathrm{NLL}_y(\theta_0) \approx \mathrm{NLL}_y(\hat{\theta}) + \nabla \mathrm{NLL}_y(\hat{\theta})^T(-\Delta) + \tfrac{1}{2}\,(-\Delta)^T\, \hat{\mathcal{I}}\, (-\Delta),$$

where $\hat{\mathcal{I}} = \nabla^2\, \mathrm{NLL}_y(\hat{\theta})$ is the **observed** information (the Hessian R reports from `optim`).

Step 2: the linear term vanishes. $\hat{\theta}$ minimizes the observed NLL, so $\nabla \mathrm{NLL}_y(\hat{\theta}) = 0$. This is the place where "the likelihood is flat at $\hat{\theta}$" is the right reason. The two minus signs in the quadratic term cancel:

$$\mathrm{NLL}_y(\theta_0) \approx \mathrm{NLL}_y(\hat{\theta}) + \tfrac{1}{2}\, \Delta^T \hat{\mathcal{I}}\, \Delta.$$

Step 3: average over y. The left side: θ_0 is a fixed number, not fitted from y, so $E_y[\mathrm{NLL}_y(\theta_0)] = Q(\theta_0)$ (the observed dataset is just one more draw from f, like y^*). The right side: in large samples $\hat{\mathcal{I}} \approx \mathcal{I}$ (the observed curvature is close to the expected curvature), so the trace trick gives

$$E_y\Big[\Delta^T \hat{\mathcal{I}}\, \Delta\Big] \approx \mathrm{tr}(\mathcal{I}\, \Sigma) \approx k.$$

Result:

$$Q(\theta_0) \approx E_y\Big[\mathrm{NLL}_y(\hat{\theta})\Big] + \frac{k}{2} \qquad \Longrightarrow \qquad \boxed{E_y\Big[\mathrm{NLL}_y(\hat{\theta})\Big] \approx Q(\theta_0) - \frac{k}{2}.}$$

The in-sample minimum sits, on average, $k/2$ below what even the best possible parameter value achieves on new data. This half has a second derivation that doesn't need the trace: $2[\mathrm{NLL}_y(\theta_0) - \mathrm{NLL}_y(\hat{\theta})]$ is the likelihood ratio statistic for testing $\theta = \theta_0$, which by Wilks' theorem is approximately χ^2_k, with mean k.

Combining the halves

$$T = E_y\Big[Q(\hat{\theta})\Big] \approx Q(\theta_0) + \frac{k}{2} \approx \left(E_y\Big[\mathrm{NLL}_y(\hat{\theta})\Big] + \frac{k}{2}\right) + \frac{k}{2} = E_y\Big[\mathrm{NLL}_y(\hat{\theta})\Big] + k.$$

So $\mathrm{NLL}_y(\hat{\theta}) + k$ is an approximately unbiased estimate of T, and doubling it gives

$$\mathrm{AIC} = 2\,\mathrm{NLL}_y(\hat{\theta}) + 2k = -2\log\mathcal{L}(\hat{\theta}) + 2k.$$

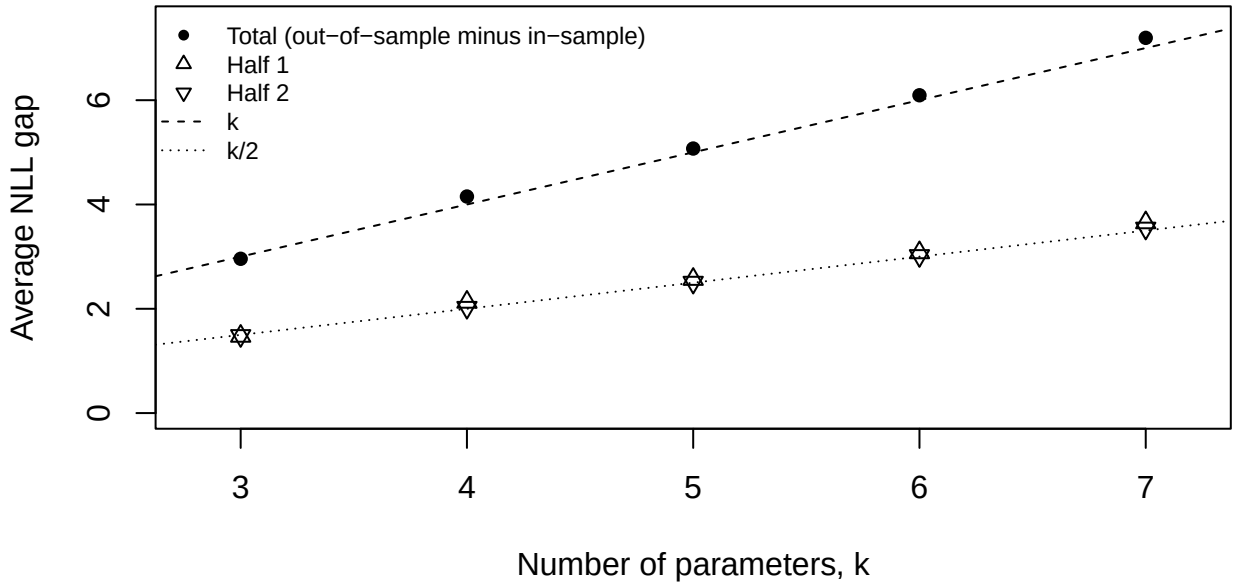
The two halves aren't double counting. They measure two different distances: Half 1 is how far the fitted model is from the best model when judged on new data; Half 2 is how much the fitted model flatters itself on the data it was fitted to.

6. Simulation check

To check the result outside the tidy normal case, simulate counts of some organism in $n = 50$ plots with a Poisson log-linear model: an intercept and two real covariates ($k = 3$ true parameters). Then fit models with $k = 3, 4, \ldots, 7$ parameters, where the extra covariates are pure noise (their true coefficients are zero, so each fitted model still contains the truth). For each of 4,000 simulated datasets per k, fit to y, then compute:

- **Half 1:** NLL of the fitted means on a new dataset y^*, minus NLL of the true means on y^*. (Using the same y^* for both cancels most of the noise; its expectation is $Q(\hat{\theta}) - Q(\theta_0)$.)
- **Half 2:** NLL of the true means on y, minus NLL of the fitted means on y.

k	Half 1	Half 2	Total (estimate of k)
3	1.48	1.48	2.96
4	2.13	2.02	4.15
5	2.57	2.51	5.07
6	3.08	3.02	6.09
7	3.65	3.54	7.20



Each half tracks $k/2$ (dotted line) and the total tracks k (dashed). Monte Carlo standard errors for the halves are about 0.05 or smaller. The total creeps above k as k grows, by more than simulation noise explains: with $n = 50$, n/k is getting small, and the large-sample approximations start to understate the optimism (Section 7).

7. What the derivation assumed

Each "$\approx$" above is an assumption, and each one can fail.

- **Large n.** The quadratic (parabola) approximation, $E[\Delta] \approx 0$, and $\hat{\mathcal{I}} \approx \mathcal{I}$ all need reasonably large samples relative to k. When n/k is small, the optimism is larger than k. AIC_c corrects for this; its correction term, $2k(k+1)/(n-k-1)$, comes from doing the two halves without approximation for the normal linear regression model (with σ estimated), so it's exact there and a reasonable patch elsewhere.
- **The model family contains the truth, or nearly so.** The step $\Sigma \approx \mathcal{I}^{-1}$ relies on the information identity (variance of the score equals its slope; Section 4 of *Why Likelihood Intervals Work*). If the model is badly wrong, the variance of the score, call it K, no longer equals $\mathcal{I}$, and the sandwich formula $\Sigma \approx \mathcal{I}^{-1} K \mathcal{I}^{-1}$ applies instead. Then the trace becomes

  $$\text{tr}(\mathcal{I}\,\Sigma) \approx \text{tr}(\mathcal{I}\,\mathcal{I}^{-1} K\,\mathcal{I}^{-1}) = \text{tr}(K\,\mathcal{I}^{-1}),$$

  which isn't k in general. Using this in place of k gives Takeuchi's information criterion (TIC). It's rarely used because K is hard to estimate well; AIC is the version that assumes the candidate models are at least roughly adequate. (This is one reason to check goodness of fit before comparing models with AIC: a model that fits badly can still "win.")
- **Regular parameters.** Parameters on a boundary (a variance component at zero, a probability at 0 or 1) break the parabola approximation. Mixed and hierarchical models

also make "number of parameters" ambiguous, which is why they need other criteria (e.g. DIC, WAIC).

8. Reading Bolker's footnote

Bolker's footnote sketches the argument above in a few lines. For readers working through it, here is how each step maps onto these notes, and where it needs a correction. In the notation above, his x_0 is θ_0, his $\hat{x}$ is $\hat{\theta}$, and his $g(x_0)$ is $g(y \mid \theta_0)$.

1. **Setup.** The K-L distance and dropping the $\int f \log f$ constant: correct, and matches Section 1. Calling AIC "the expected value of the K-L distance" is loose: AIC estimates twice the expected *relative* K-L distance (the part that differs between models).
2. **"the true MLE x_0."** θ_0 is not an MLE. It's the K-L-best parameter value, which equals the true value only when the model is correct.
3. **Using x for both data and parameters.** The densities $g(x_0)$ and $g(\hat{x})$ are functions of the data, evaluated at a parameter value. The footnote never distinguishes the data used to fit from the data used to judge (y versus y^*), and that distinction is the core of the argument.
4. **"Taylor expanding $-\log g(\hat{x})$ around x_0."** The expression that follows has $-\log g(\hat{x})$ as its constant term, which is an expansion *around* $\hat{x}$ (our Half 2), not around x_0 (our Half 1). The quadratic term also has a bare x where x_0 is meant.
5. **"the likelihood is flat at $\hat{x}$."** True for the observed likelihood (Half 2). In Half 1 the reason is different: the *expected* NLL is flat at θ_0.
6. **Calling V "the information matrix."** V is the Hessian of the log-likelihood, which is the *negative* of the information. His next statement, $-V^{-1} \approx \Sigma$, is correct.
7. **"By definition, $E[(x-\hat{x})^T(x-\hat{x})]$ also equals Σ."** That's the inner product, a single number. The covariance matrix is the outer product, $E[\Delta\Delta^T]$, and it equals Σ only approximately. The step that connects it to the quadratic form $\Delta^T V \Delta$ is the trace trick (Section 3), which the footnote skips ("after more math").
8. **The missing factor of 2.** His expansion carries a $\frac{1}{2}$, so its expectation is $\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma) = k/2$, not k. Doubling NLL + $k/2$ would give a penalty of k. The footnote reaches $+k$ by dropping the $\frac{1}{2}$. The real source of the other $k/2$ is the second expansion: one expansion handles Half 2 (in-sample optimism), and Half 1 (the out-of-sample cost of estimating θ) is never derived.

Burnham and Anderson (2002, Chapter 7) give the full derivation, in more general form (including TIC).

Summary

Step	What it says	Size
Target	expected NLL of the fitted model on **new** data, $E_y[Q(\hat{\theta})]$	
Half 1	fitted model vs. best model, judged on new data	$+k/2$
Half 2	best model on its own data vs. fitted model on its own data	$+k/2$
Key tool	$E[\Delta^T A \Delta] = \mathrm{tr}(A\Sigma)$, with $\Sigma \approx \mathcal{I}^{-1}$	$\mathrm{tr}(I_k) = k$
AIC	$2\,[\mathrm{NLL}_y(\hat{\theta}) + k]$	$-2\log\mathcal{L} + 2k$

In one sentence: **a model fitted and scored on the same data looks better than it is by about k NLL units, half because the fitted parameters miss the best ones and half because they chase the noise in the data they were fitted to, and AIC adds that k back (doubled).**

References

Bolker, B. M. 2008. *Ecological Models and Data in R.* Princeton University Press, Princeton, NJ.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd edition. Springer, New York.