"AIC-type" Information Criteria

Francis J. DiTraglia

March 3, 2014

1 Introduction

Akaike's Information Criterion (AIC) summarizes the quality of a model by trading fit, measured by the maximized log likelihood, against complexity, measured by the number of estimated parameters. But where does this complexity penalty come from? In these notes we'll take a closer look at the AIC along with two related information criteria: Takeuichi's Information Criterion (TIC), and the Corrected AIC (AIC$_c$) of Hurvich and Tsai (1989). All three are based on approximations to the **Kullback-Leibler Divergence**, a fundamental quantity from information theory that is inextricably linked to maximum likelihood information.

1.1 Kullback-Leibler Divergence

Suppose that $\mathbf{y}$ is a T-dimensional random vector drawn from a probability distribution G with density $g(\mathbf{y})$. This is the *true DGP* and is unknown to us. Since we don't know g, we attempt to approximate it using a parametric model $f(\mathbf{y}|\theta)$, where θ is a p-vector of parameters that we estimate via maximum likelihood.[1] Since f is not the true data density, a natural question is

---

[1]I've written the model without covariates to keep the notation from getting out of control, but you could just as well write $f(\mathbf{y}|X,\theta)$. Similarly, I will sometimes write $f(\mathbf{y})$ for $f(\mathbf{y}|\theta)$ to simplify the notation below.

how well does f approximate g? It turns out that for maximum likelihood estimation there is a particularly convenient way to answer this question using the **Kullback Leibler Divergence**.

Definition 1.1 (KL Divergence). *Let E_G denote expectation with respect to the true, unknown data density g. Then the Kullback-Leibler divergence from g to f is given by*

$$KL(g; f) = E_G \left[\log \left\{ \frac{g(\boldsymbol{y})}{f(\boldsymbol{y})} \right\} \right] = E_G \left[\log g(\boldsymbol{y}) \right] - E_G \left[\log f(\boldsymbol{y}) \right]$$

The quantity $E_G \left[\log f(\mathbf{y}) \right]$ is called the Expected Log-likelihood.

Key Features of the KL Divergence   There are several important features to note about the KL divergence:

1. It is *not* symmetric: $KL(g; f) \neq KL(f; g)$. Hence, the KL divergence is *not* a distance function (metric).

2. $KL(g; f) \geq 0$ with equality iff $f = g$. To see why, recall that, since log is a concave function, $-\log$ is convex. Thus

$$\begin{aligned}
KL(g; f) &= E_G \left[\log \left\{ \frac{g(\mathbf{y})}{f(\mathbf{y})} \right\} \right] = E_G \left[-\log \left\{ \frac{f(\mathbf{y})}{g(\mathbf{y})} \right\} \right] \\
&\geq -\log \left\{ E_G \left[\frac{f(\mathbf{y})}{g(\mathbf{y})} \right] \right\} = -\log \left(\int g(\mathbf{y}) \frac{f(\mathbf{y})}{g(\mathbf{y})} \, d\mathbf{y} \right) \\
&= -\log \left(\int f(\mathbf{y}) \, d\mathbf{y} \right) = -\log(1) = 0
\end{aligned}$$

by Jensen's Inequality. The inequality is strict only for a non-degenerate random variable and a strictly convex function. Since $-\log$ is strictly convex, the only way to make the inequality strict is for $f(\mathbf{y})/g(\mathbf{y})$ to be degenerate. This occurs precisely when $f = g$ almost everywhere.

3. Minimizing the KL divergence $KL(g; f)$ is *equivalent* to maximizing the Expected Log-Likelihood $E_G[\log f(\mathbf{y})]$. This is because the first term in the KL divergence is a constant: it in no way depends on the model $f(\mathbf{y})$. The expected Log-likelihood enters negatively:

$$KL(g; f) = E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y})\right]$$

Thus, if we can find a way to estimate the Expected Log-likelihood, we can use the KL divergence for model selection: the larger the Expected Log-likelihood, the smaller the KL divergence, and the better the model.

4. The KL divergence equals the negative of **Boltzmann's Entropy** from Statistical Mechanics. Accordingly, it represents the *information lost* when $g(\mathbf{y})$ is encoded by $f(\mathbf{y})$.

1.2   Relationship of MLE to KL

It turns out that the KL divergence is inextricably linked to maximum likelihood estimation. To make the points a little clearer, I'll assume from now on that $\mathbf{y}$ consists of iid observations y_t for $t = 1, \ldots, T$. This is not in fact necessary for any of the derivations that follow, but it simplifies the notation. Since the expected log likelihood $E_G\left[\log f(Y, \theta)\right]$ is unknown, we might try to approximate it using the sample analogue

$$E_{\widehat{G}}\left[\log f(\mathbf{y}, \theta)\right] = \frac{1}{T}\sum_{t=1}^{T} \log f(y_t, \theta) = \frac{1}{T}\ell(\theta)$$

where we have replaced G with the empirical distribution $\widehat{G}$. Now, by the Weak Law of Large Numbers for iid observations

$$\frac{1}{T}\ell(\theta) \xrightarrow{p} E_G\left[\log f(\mathbf{y}, \theta)\right]$$

Under the standard regularity conditions (see Newey and McFadden, 1994) we can strengthen this result to show that

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{T} \ell(\theta) \xrightarrow{p} \arg \max_{\theta \in \Theta} E_G \left[\log f(\mathbf{y}, \theta) \right]$$

Since minimizing the KL divergence is the same as maximizing the expected log-likelihood we have the following result:

Proposition 1.1. *The ML estimator $\hat{\theta}$ converges in probability to the value of θ that minimizes the KL divergence from unknown true density $g(\mathbf{y})$ to the parametric family $f(\mathbf{y}|\theta)$. When $g(\mathbf{y}) = f(\mathbf{y}|\theta)$ for some value of $\theta \in \Theta$, the divergence is minimized at zero.*

1.3   A Naïve Information Criterion

If $g(\mathbf{y})$ were known, we could choose between two parametric models $f(\mathbf{y}|\theta)$ and $h(\mathbf{y}|\gamma)$ by comparing maximized Log-likelihoods. Define

$$\begin{aligned} \theta_0 &= \arg \max_{\theta \in \Theta} E_G \left[\log f(\mathbf{y}, \theta) \right] \\ \gamma_0 &= \arg \max_{\gamma \in \Gamma} E_G \left[\log h(\mathbf{y}, \gamma) \right] \end{aligned}$$

If $E_G \left[\log f(\mathbf{y}, \theta_0) \right] > E_G \left[\log h(\mathbf{y}, \gamma_0) \right]$, then the KL divergence from $g(\mathbf{y})$ to the parametric family f_θ is smaller than that from $g(\mathbf{y})$ to h_γ. Now, we know from above that $\hat{\theta} \xrightarrow{p} \theta_0$. Further, $\frac{1}{T}\ell(\theta) \xrightarrow{p} E_G \left[\log f(Y, \theta) \right]$. Of course, T will be constant across models, so why not use the maximized sample likelihood $\ell(\hat{\theta})$ for model comparison? Unfortunately, $\ell(\hat{\theta})$ is a *biased estimator of the expected log likelihood* because is uses the data twice: first to estimate $\hat{\theta}$ and then directly in the sum $\sum_{t=1}^{T} \log f(y_t, \hat{\theta})$. Because $\hat{\theta}$ was chosen to conform to the idiosyncrasies of the data at hand, $\ell(\hat{\theta})$ is overly optimistic.

We can see this as follows. Since θ_0 is the population minimizer of the KL

divergence from g to f_θ, we have

$$
\begin{aligned}
KL\left[g(\mathbf{y}); f(\mathbf{y}, \theta)\right] &\geq KL\left[g(\mathbf{y}); f(\mathbf{y}, \theta_0)\right] \\
E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y}, \theta)\right] &\geq E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y}, \theta_0)\right] \\
E_G\left[\log f(\mathbf{y}, \theta)\right] &\leq E_G\left[\log f(\mathbf{y}, \theta_0)\right]
\end{aligned}
$$

for all $\theta \in \Theta$. Recall that $\frac{1}{T}\ell(\theta) = E_{\widehat{G}}\left[\log f(\mathbf{y}, \theta)\right]$. By the definition of the maximum likelihood estimate, $\ell(\hat{\theta}) \geq \ell(\theta_0)$. Thus,

$$
E_{\widehat{G}}\left[\log f(\mathbf{y}, \hat{\theta})\right] \geq E_{\widehat{G}}\left[\log f(\mathbf{y}, \theta_0)\right]
$$

In sample, the estimate $\hat{\theta}$ will show a higher maximized log-likelihood than the value of θ that maximizes the population log-likelihood. Ideally, what we want to do is evaluate the likelihood at a *new* set of observations not used to estimate $\hat{\theta}$.

2 The AIC and TIC

In the previous section we explained that using the KL divergence to do model selection is equivalent to maximizing the expected log-likelihood across models. Unfortunately, using the maximized log-likelihood, based on the estimated parameters, is a biased estimator of this quantity: it is systematically too high. Both the AIC and the TIC address this problem by using asymptotic theory to get an expression for the bias so that we can correct for it.

2.1 An Asymptotic Approximation to the Bias

First some notation:

- Make the sample size and the dependence of $\hat{\theta}$ on the data explicit by

writing

$$\log f\left(\mathbf{y}_n \middle| \hat{\theta}(\mathbf{y}_n)\right) = \sum_{t=1}^{T} \log f(y_t|\hat{\theta})$$

- Let $E_{G(\mathbf{y}_n)}$ denote expectation with respect to the joint distribution of $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)'$. The realizations of this random vector are $\mathbf{y}_n = y_1, \ldots, y_n$.

- Let $\mathbf{Z}^{(n)} = (Z_1, \ldots, Z_n)'$ be a collection of iid random variables with the same distribution as $\mathbf{Y}^{(n)}$ that were *not* used to estimate $\hat{\theta}$. (Assume that $\mathbf{Z}^{(n)}$ and $\mathbf{Y}^{(n)}$ are mutually independent.) Let $G(\mathbf{z}_n)$ denote expectation with respect to the joint distribution of $\mathbf{Z}^{(n)}$.

- The bias of the maximized Log-likelihood as an estimator of the expected Log-likelihood is given by

$$Bias = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - E_{G(\mathbf{z}_n)}\left\{\log f\left(\mathbf{Z}^{(n)} \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

Calculating the Bias: Since $Z_1, \ldots, Z_n$ are iid, the expectation with respect to $G(\mathbf{z}_n)$ simplifies, and we can write

$$Bias = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - nE_{G(z)}\left\{\log f\left(Z \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

where Z is any of the $Z_1, \ldots, Z_n$ and $G(z)$ is its *marginal* distribution. We expand the bias into $D_1 + D_2 + D_3$ where

$$D_1 = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - \log f\left(\mathbf{Y}^{(n)} \middle| \theta_0\right)\right]$$

$$D_2 = E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)} \middle| \theta_0\right) - nE_{G(z)}\left\{\log f\left(Z \middle| \theta_0\right)\right\}\right]$$

$$D_3 = E_{G(\mathbf{y}_n)}\left[nE_{G(z)}\left\{\log f\left(Z \middle| \theta_0\right)\right\} - nE_{G(z)}\left\{\log f\left(Z \middle| \hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right)\right\}\right]$$

Now consider each part separately.

Step 1: Calculation of D_2   This is the easiest part as it involves no estimators. Note first that $E_{G(z)}\{\log f\left(Z|\theta_0\right)\}$ is a constant, so that

$$
\begin{aligned}
D_2 &= E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)}|\theta_0\right)\right] - nE_{G(z)}\left[\log f\left(Z|\theta_0\right)\right]\\
&= nE_{G(y)}\left[\log f(Y|\theta_0)\right] - nE_{G(z)}\left[\log f\left(Z|\theta_0\right)\right]\\
&= 0
\end{aligned}
$$

where we have used the fact that $Y_1,\ldots,Y_n$ are iid and that Y and Z have the same distribution.

Step 2: Calculation of D_3   Define $\eta(\hat{\theta}) = E_{G(z)}\left[\log f\left(z|\hat{\theta}(\mathbf{Y}^{(n)})\right)\right]$ and Taylor expand around θ_0, the solution to the population moment condition

$$
E_{G(z)}\left[\frac{\partial}{\partial\theta}\log f(z|\theta)\right] = 0
$$

yielding

$$
\eta(\hat{\theta}) = \eta(\theta_0) + \left(\hat{\theta} - \theta_0\right)'\frac{\partial\eta(\theta_0)}{\partial\theta} + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)\frac{\partial^2\eta(\theta_0)}{\partial\theta\partial\theta'}\left(\hat{\theta} - \theta_0\right) + o_p(1)
$$

Now, under the appropriate regularity conditions (Lebesgue Dominated Convergence) we can exchange the order of expectation and differentiation, so that

$$
\frac{\partial\eta(\theta_0)}{\partial\theta} = \frac{\partial}{\partial\theta}E_{G(z)}\left[\log f(z|\theta_0)\right] = E_{G(z)}\left[\frac{\partial}{\partial\theta}\log f(z|\theta_0)\right] = 0
$$

by the definition of θ_0 as the solution to the population moment condition. Substituting this into the Taylor Expansion, and assuming further that we may exchange the *second* derivative and expectation with respect to $G(z)$, we have

$$
\begin{aligned}
\eta(\hat{\theta}) &= \eta(\theta_0) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)'\frac{\partial^2}{\partial\theta\partial\theta'}E_{G(z)}\left[\log f(z|\theta_0)\right]\left(\hat{\theta} - \theta_0\right) + o_p(1)\\
&= \eta(\theta_0) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)'E_{G(z)}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(z|\theta_0)\right]\left(\hat{\theta} - \theta_0\right) + o_p(1)
\end{aligned}
$$

Thus, defining

$$J(\theta_0) = -E_{G(z)}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(z|\theta_0)\right]$$

we see that

$$\eta(\hat{\theta}) = E_{G(z)}\left[\log f(z|\theta_0)\right] - \frac{1}{2}\left(\hat{\theta} - \theta_0\right)' J(\theta_0)\left(\hat{\theta} - \theta_0\right) + o_p(1)$$

Substituting this expansion for $E_{G(z)}\left[\log f\left(z|\hat{\theta}(\mathbf{Y}^{(n)})\right)\right]$ into the expression for D_3 cancels the leading first term $nE_{G(z)}\left\{\log f(Z|\theta_0)\right\}$ in the expectation so that (remember that $\hat{\theta}$ is a function of $\mathbf{y}_n$ and thus, so is $o_p(1)$)

$$
\begin{aligned}
D_3 &= E_{G(\mathbf{y}_n)}\left[\frac{n}{2}\left(\hat{\theta} - \theta_0\right)' J(\theta_0)\left(\hat{\theta} - \theta_0\right) + o_p(1)\right] \\
&= \frac{1}{2}E_{G(\mathbf{y}_n)}\left[\operatorname{trace}\left\{n\left(\hat{\theta} - \theta_0\right)' J(\theta_0)\left(\hat{\theta} - \theta_0\right)\right\}\right] + o_p(1) \\
&= \frac{1}{2}\operatorname{trace}\left\{J(\theta_0)E_{G(\mathbf{y}_n)}\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\sqrt{n}\left(\hat{\theta} - \theta_0\right)'\right]\right\} + o_p(1)
\end{aligned}
$$

where we have used the facts that: (1) a scalar equals its trace, (2) multiplication within the trace operator is commutative provided that the products remain conformable, (3) trace and expectation can be exchanged because both are linear operators, and (4) $J(\theta_0)$ is a constant with respect to $E_{G(\mathbf{y}_n)}$. We know that

$$E_{G(\mathbf{y}_n)}\left[\sqrt{n}\left(\hat{\theta} - \theta_0\right)\sqrt{n}\left(\hat{\theta} - \theta_0\right)'\right] \to J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)$$

as $n \to \infty$ where

$$
\begin{aligned}
I(\theta_0) &= E_{G(z)}\left[\frac{\partial \log f(Z|\theta)}{\partial\theta}\frac{\partial \log f(Z|\theta)}{\partial\theta'}\right] \\
J(\theta_0) &= -E_{G(z)}\left[\frac{\partial^2 \log f(Z|\theta)}{\partial\theta\partial\theta'}\right]
\end{aligned}
$$

8

Thus, assuming the necessary regularity conditions to ensure that the remainder is indeed of small order, as $n \to \infty$ we have

$$D_3 \to \frac{1}{2}\text{trace}\left\{J(\theta_0)J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}\right\} = \frac{1}{2}\text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}$$

Step 3: Calculation of D_1 Taylor expand $\ell(\theta_0) = \log f(\mathbf{y}_n|\theta_0)$ around $\hat{\theta}$, noting that by the definition of the MLE,

$$\frac{\partial \ell(\hat{\theta})}{\partial \theta} = 0$$

Thus,

$$\ell(\theta_0) = \ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\tilde{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right)$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ_0, so we have

$$\ell(\hat{\theta}) - \ell(\theta_0) = \frac{1}{2}\sqrt{n}\left(\theta - \hat{\theta}\right)'\left[-\frac{1}{n}\frac{\partial^2 \ell(\tilde{\theta})}{\partial\theta\partial\theta'}\right]\sqrt{n}\left(\theta - \hat{\theta}\right)$$

By the WLLN for iid observations, we know that for fixed θ

$$-\frac{1}{n}\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'} = -\frac{1}{n}\sum_{t=1}^{T}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(y_t|\theta) \xrightarrow{p} J(\theta)$$

Furthermore, by the standard consistency result for MLE $\hat{\theta} \to \theta_0$. Thus, by a Uniform WLLN,

$$-\frac{1}{n}\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \xrightarrow{p} J(\theta_0)$$

It follows that

$$
\begin{aligned}
D_1 &= E_{G(\mathbf{y}_n)}\left[\log f\left(\mathbf{Y}^{(n)}\big|\hat{\theta}\left(\mathbf{Y}^{(n)}\right)\right) - \log f\left(\mathbf{Y}^{(n)}|\theta_0\right)\right]\\
&= E_{G(\mathbf{y}_n)}\left[\ell(\hat{\theta}) - \ell(\theta_0)\right]\\
&\to \frac{1}{2}\text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}
\end{aligned}
$$

9

using the same arguments as in Step 2.

Finally The Bias: Combining the three steps, we have

$$Bias = D_1 + D_2 + D_3 = \text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\}$$

2.2 Correcting the Bias

Now that we have an asymptotic approximation to the bias of the maximized likelihood as an estimator of the expected log-likelihood, we can correct it. Suppose that we have consistent estimators $\hat{I}$ and $\hat{J}$ of $I(\theta_0)$ and $J(\theta_0)$. Then, by the Continuous Mapping Theorem $\hat{b} = \text{trace}(\hat{I}\hat{J}^{-1})$ is a consistent estimator of the bias term. Recall that the maximized log-likelihood is biased *upwards*, hence to correct it we need to subtract this quantity. This estimate yields **Takeuchi's Information Criterion** (TIC):

$$TIC = 2\left[\sum_{t=1}^{T}\log f\left(y_t|\hat{\theta}\right) - \text{trace}(\hat{I}\hat{J}^{-1})\right] = 2\left[\ell(\theta) - \text{trace}\left(\widehat{I}\widehat{J}^{-1}\right)\right]$$

The scaling factor of two is traditional, but has no effect on model comparisons.

Akaike's Information Criterion (AIC) uses a slightly different bias correction. Recall that if there exists a $\theta \in \Theta$ such that the true density $g(y)$ equals the model $f(y|\theta)$, then $I(\theta_0) = J(\theta_0)^{-1}$. This is the famous "Information Matrix Equality." In this case, the bias approximation becomes:

$$Bias = \text{trace}\left\{I(\theta_0)J(\theta_0)^{-1}\right\} = \text{trace}\left\{\mathbf{I}_p\right\} = p$$

that is, the dimension of the parameter vector. Using this as a bias correction yields **Akaike's Information Criterion**:

$$AIC = 2\left[\sum_{t=1}^{T}\log f\left(y_t|\hat{\theta}\right) - p\right] = 2\left[\ell(\theta) - p\right]$$

10

Again, it is traditional to multiply by two. Although the TIC and AIC are similar, there are several subtleties:

- The AIC is derived assuming that the model is *correctly specified*, while the TIC is not. In this sense the AIC is a special case of the TIC.

- Typically, the matrices $I(\theta_0)$ and $J(\theta_0)$ are large, meaning that the estimates will have high variance (we need to estimate $p^2 + p$ elements).

- In contrast, the AIC a has much smaller variance because the bias correction *does not depend on the data*.

- Thus, even if the model is mis-specified, it may be preferable to use AIC rather than TIC unless the sample size is large.

- It has been argued that for models where the Information Matrix is not satisfied, the AIC will still be close to the TIC. (The log-likelihood term should dominate the bias correction in such situations.)

2.3   Friends of AIC

Recall the idea behind the AIC:

1. Use the KL divergence to choose a model.

2. Suffices to compare Expected Log-likelihoods.

3. Sample analogue, maximized log-likelihood, requires a bias correction.

The AIC and TIC use an analytical bias correction. Other possibilities:

- Cross-Validation

- Bootstrap Information Criterion (EIC)

Maximum likelihood estimators can be unstable. If we want to use a different kind of estimator (e.g. maximum penalized likelihood) but are still willing to write down a likelihood, we can use the KL divergence for model selection via the Generalized Information Criterion (GIC) of Konishi and Kitagawa, or its bootstrap equivalent (EGIC).

3   Refinements to the AIC

In the derivations above we used asymptotic theory to derive an analytical bias correction. These approximations tend to work well as long n is fairly large relative to p but when this is not the case, they can break down. We'll now consider some alternatives that work better in such situations. The first uses *exact* small-sample theory to derive the appropriate bias correction, while the second uses the Bootstrap.

3.1   Corrected AIC

This is based on the derivation in Hurvich & Tsai (1989). Suppose that the true DGP is

$$\mathbf{y} = X\beta_0 + \boldsymbol{\epsilon}$$

where $\epsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T)$. Then $\mathbf{y}|X \sim N(X\beta_0, \sigma_0^2 \mathbf{I}_T)$ so the likelihood is

$$g(\mathbf{y}|X; \beta_0, \sigma_0^2) = \left(2\pi\sigma_0^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta_0)'(y - X\beta_0)\right\}$$

and the log-likelihood is

$$\log\left[g(\mathbf{y}|X; \beta_0, \sigma_0^2)\right] = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\left(\mathbf{y} - X\beta_0\right)'\left(\mathbf{y} - X\beta_0\right)$$

To keep the notation from getting out of control, we'll use the shorthand $\log[g(\mathbf{y})]$ for this quantity. Now suppose we evaluated the log-likelihood at some *other* parameter values β_1 and σ_1^2. The vector β_1 might, for example,

correspond to dropping some regressors from the model by setting their coefficients to zero, or perhaps adding in some additional regressors. We have

$$\log[f(\mathbf{y}|X; \beta_1, \sigma_1^2)] = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_1^2) - \frac{1}{2\sigma_1^2}(\mathbf{y} - X\beta_1)'(\mathbf{y} - X\beta_1)$$

$$KL(g; f) = \int \log[g(\mathbf{y})]g(\mathbf{y})\,d\mathbf{y} - \int \log[f(\mathbf{y})]g(\mathbf{y})\,d\mathbf{y} = A - B$$

$$KL(g; f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + \left(\frac{1}{2\sigma_1^2}\right)(\beta_0 - \beta_1)' X'X (\beta_0 - \beta_1)$$

$$
\begin{aligned}
A &= \int \log[g(\mathbf{y})]g(\mathbf{y})\,d\mathbf{y} \\
&= \int \left[-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}(\mathbf{y} - X\beta_0)'(\mathbf{y} - X\beta_0)\right]g(\mathbf{y})\,d\mathbf{y} \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_0^2)\right] - \frac{1}{2\sigma_0^2}E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_0)'(\mathbf{y} - X\beta_0)\right] \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_0^2)\right] - \frac{1}{2\sigma_0^2}\text{trace}\left\{E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_0)(\mathbf{y} - X\beta_0)'\right]\right\} \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_0^2)\right] - \frac{1}{2\sigma_0^2}\text{trace}\left\{Var(\mathbf{y}|X)\right\} \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_0^2)\right] - \frac{1}{2\sigma_0^2}\left(T\sigma_0^2\right) \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_0^2) + 1\right]
\end{aligned}
$$

$$
\begin{aligned}
B &= \int \log[f(\mathbf{y})]g(\mathbf{y})\, d\mathbf{y} \\
&= \int \left[-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_1^2) - \frac{1}{2\sigma_1^2}(\mathbf{y} - X\beta_1)'(\mathbf{y} - X\beta_1) \right] g(\mathbf{y})\, d\mathbf{y} \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_1^2)\right] - \frac{1}{2\sigma_1^2} E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_1)'(\mathbf{y} - X\beta_1)\right] \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_1^2)\right] - \left(\frac{1}{2\sigma_1^2}\right) C
\end{aligned}
$$

$$
\begin{aligned}
C &= E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_1)'(\mathbf{y} - X\beta_1)\right] \\
&= E_{\mathbf{y}|X}\left[\{(\mathbf{y} - X\beta_0) + X(\beta_0 - \beta_1)\}'\{(\mathbf{y} - X\beta_0) + X(\beta_0 - \beta_1)\}\right] \\
&= E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_0)'(\mathbf{y} - X\beta_0)\right] \\
&\qquad + E_{\mathbf{y}|X}\left[(\mathbf{y} - X\beta_0)'X(\beta_0 - \beta_1)\right] \\
&\qquad + E_{\mathbf{y}|X}\left\{[X(\beta_0 - \beta_1)]'(\mathbf{y} - X\beta_0)\right\} \\
&\qquad + E_{\mathbf{y}|X}\left[\{X(\beta_0 - \beta_1)\}'\{X(\beta_0 - \beta_1)\}\right] \\
&= Var(\mathbf{y}|X) \\
&\qquad + E_{\mathbf{y}|X}\left[\mathbf{y} - X\beta_0\right]'X(\beta_0 - \beta_1) \\
&\qquad + (\beta_0 - \beta_1)'X'E_{\mathbf{y}|X}\left[\mathbf{y} - X\beta_0\right] \\
&\qquad + (\beta_0 - \beta_1)X'X(\beta_0 - \beta_1) \\
&= T\sigma_0^2 + 0 + 0 + (\beta_0 - \beta_1)X'X(\beta_0 - \beta_1)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
B &= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_1^2)\right] - \left(\frac{1}{2\sigma_1^2}\right)\left[T\sigma_0^2 + (\beta_0 - \beta_1)X'X(\beta_0 - \beta_1)\right] \\
&= -\frac{T}{2}\left[\log(2\pi) + \log(\sigma_1^2) + \frac{\sigma_0^2}{\sigma_1^2}\right] - \left(\frac{1}{2\sigma_1^2}\right)(\beta_0 - \beta_1)X'X(\beta_0 - \beta_1)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
KL(g; f) &= A - B \\
&= \left\{ -\frac{T}{2} \left[\log(2\pi) + \log(\sigma_0^2) + 1 \right] \right\} \\
&\quad - \left\{ -\frac{T}{2} \left[\log(2\pi) + \log(\sigma_1^2) + \frac{\sigma_0^2}{\sigma_1^2} \right] - \left(\frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1) X' X (\beta_0 - \beta_1) \right\} \\
&= -\frac{T}{2} \left[\log(\sigma_0^2) + 1 - \log(\sigma_1^2) - \frac{\sigma_0^2}{\sigma_1^2} \right] + \left(\frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1) X' X (\beta_0 - \beta_1) \\
&= \frac{T}{2} \left[\frac{\sigma_0^2}{\sigma_1^2} - \log \left(\frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left(\frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1) X' X (\beta_0 - \beta_1)
\end{aligned}
$$

We need to estimate this quantity for it to be of any use in model selection. If let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the maximum likelihood estimators of β_1 and σ_1^2 and substitute them into the expression for the KL divergence, we have

$$
\widehat{KL}(g; f) = \frac{T}{2} \left[\frac{\sigma_0^2}{\widehat{\sigma}^2} - \log \left(\frac{\sigma_0^2}{\widehat{\sigma}^2} \right) - 1 \right] + \left(\frac{1}{2\widehat{\sigma}^2} \right) \left(\beta_0 - \widehat{\beta} \right) X' X \left(\beta_0 - \widehat{\beta} \right)
$$

We still have two problems. First, we haven't been entirely clear about what β_1 and σ_1 are. At the moment, they seem to be something like "pseudo-true" values. Second, and more importantly, we don't know β_0 and σ_0^2 so we can't use the preceding expression to compare models.

Hurvich and Tsai (1989) address both of these problems with the assumption that all models under consideration are *at least correctly specified*. That is, while they may include a regressor whose coefficient is in fact zero, they do not exclude any regressors with a non-zero coefficient. This is the same assumption that we used above to reduce TIC to AIC. Under this assumption, β_1 and σ_1^2 *are precisely the same* as β_0 and σ_0^2. More importantly, we can use all of the standard results for the exact finite sample distribution of regression estimators to help us. The idea is to construct an *unbiased* estimator of the

15

KL divergence. Taking expecations and rearranging slightly, we have

$$E\left[\widehat{KL}(g;f)\right] = \frac{T}{2}\left\{E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right\}$$
$$+ \frac{1}{2}E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)\right]$$

Now, under our assumptions $T\widehat{\sigma}^2/\sigma_0^2 \sim \chi_{T-k}^2$ where k is the number of estimated coefficients in $\widehat{\beta}$. Further, if $Z \sim \chi_\nu^2$ then $E[1/Z] = 1/(\nu-2)$. It follows that

$$E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] = E\left[\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right] = \frac{T}{T-k-2}$$

We can rewrite the final term similarly:

$$E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)\right] = E\left[\left(\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right)\frac{\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)}{\sigma_0^2}\right]$$

Under our assumptions the two terms in the product are independent, so we can break apart the expectation. First, we have

$$E\left[\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right] = \frac{T}{T-k-2}$$

as above. For the second part,

$$\frac{\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)}{\sigma_0^2} \sim \chi_k^2$$

and hence

$$E\left[\frac{\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)}{\sigma_0^2}\right] = k$$

Putting all the pieces together,

$$
\begin{aligned}
E\left[\widehat{KL}(g;f)\right] &= \frac{T}{2}\left\{E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] + \log(\sigma_0^2) - E\left[\log(\widehat{\sigma}^2)\right] - 1\right\} \\
&\quad + \frac{1}{2}E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta} - \beta_0\right)X'X\left(\widehat{\beta} - \beta_0\right)\right] \\
&= \frac{T}{2}\left(\frac{T}{T-k-2} - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right) + \frac{T}{2}\left(\frac{k}{T-k-2}\right) \\
&= \frac{T}{2}\left(\frac{T+k}{T-k-2} - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right)
\end{aligned}
$$

Since $\log(\widehat{\sigma}^2)$ is an unbiased estimator of $E[\log(\widehat{\sigma}^2)]$, substituting this give us an unbiased estimator of $E\left[\widehat{KL}(g;f)\right]$ as desired. The only terms that vary across candidate models are the first and the third. Moreover, the multiplicative factor of $T/2$ does not affect model selection. Hence, the criterion is

$$
AIC_c = \log(\widehat{\sigma}^2) + \frac{T+k}{T-k-2}
$$

So how does this compare to the plain-vanilla AIC for normal linear regression? The maximum likelihood estimators for this problem are

$$
\begin{aligned}
\widehat{\beta} &= (X'X)^{-1}X'\mathbf{y} \\
\widehat{\sigma}^2 &= \frac{(\mathbf{y} - X\widehat{\beta})'(\mathbf{y} - X\widehat{\beta})}{T}
\end{aligned}
$$

It follows that the maximized log-likehood is

$$
\begin{aligned}
\log\left[f(\mathbf{y}|X;\widehat{\theta})\right] &= -\frac{T}{2}\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}(y - X\widehat{\beta})'(y - X\widehat{\beta}) \\
&= -\frac{T}{2}\log(\widehat{\sigma}^2) - \frac{T}{2}
\end{aligned}
$$

by substituting $T\widehat{\sigma}^2$ for the numerator of the second term. Hence, the AIC for

this problem is

$$AIC = 2\left(\ell(\widehat{\theta}) - p\right) = -T\log(\widehat{\sigma}^2) - T - 2p$$

Check the convention for signs, but in units comparable to AIC_c.