Lecture 2: Various Model Selection Criteria

Francis J. DiTraglia

March 8, 2014

1 The Corrected AIC

To derive the TIC and AIC we used asymptotic theory to construct an analytical bias correction. Such approximations tend to work as long as T is fairly large relative to p but when this is not the case, they can break down. We'll now consider an alternative that makes stronger assumptions and relies on *exact* small-sample theory rather than asymptotics: the "Corrected" AIC, or AIC_c, of Hurvich and Tsai (1989). Suppose that the true DGP is a linear regression model:

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

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

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

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

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

Since we've specified the density from which the data were generated as well as the density of the approximating model, we can *directly calculate* the KL divergence rather than trying to find a reasonable large sample approximation. It turns out that for this example

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

You will demonstrate this on the problem set. 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

KL divergence. Taking expecations and rearranging slightly, we have

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

Now, under our assumptions $T\widehat{\sigma}^2/\sigma_0^2 \sim \chi^2_{T-k}$ where k is the number of estimated coefficients in $\widehat{\beta}$. Further, if $Z \sim \chi^2_\nu$ 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,

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

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

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

In its broad strokes, this makes perfect sense. The residual error variance $\widehat{\sigma}^2$ measures in-sample fit. But since in-sample fit is a mis-leading guide to out-of-sample fit, we add a complexity penalty. Note that the way this expression is written, *smaller* values indicate a better model.

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

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

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

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

But this way of writing things uses the *opposite* sign convention from AIC_c. It's important to keep track of this, since different authors use different sign conventions for information criteria. To make the AIC comparable with our scaling of the AIC_c, we multiply through by $-1/T$ yielding

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

where *smaller* values now indicate a better model.

2 Cross-Validation

In the last lecture, we learned that choosing a model by minimizing the KL divergence is equivalent to choosing a model by maximizing the expected log likelihood. We also learned that the sample analogue

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

provides a biased estimator of this quantity. Intuitively, the problem is that it uses the data twice: first to estimate $\widehat{\theta}$ and then to approximate the integral

$$\int g(y)\log f(y|\widehat{\theta})\,dy = E_G\left[\log f(Y_{new}|\widehat{\theta})\right]$$

using the empirical CDF constructed from the sample observations. If the problem is that we've used the data twice, then an obvious idea is to find some way to use *different* data for estimating θ than we use to approximate the integral. This is the idea behind cross-validation. We split the data into

two parts, use one for estimation and the *other* for model evaluation. To avoid "wasting data" we repeat this process sucessively for *different* splits, so each observation has a chance to be used for for estimation and evaluation but *never for both at the same time*.

To make this more concrete, we'll consider what is perhaps the most common form of cross-validation: "leave-one-out" cross-validation. For simplicity, suppose we have iid observations $Y_1, \ldots, Y_T$ and let $\widehat{\theta}_{(t)}$ denote the ML estimator for θ using all observations *except* Y_t. The leave-one-out cross-validation estimator of the expected log likelihood is

$$CV(1) = \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\widehat{\theta}_{(t)})$$

The idea is that, since the data are iid, $\widehat{\theta}_{(t)}$ is *independent* of Y_t. Accordingly, the cross-validation estimate of the expected log-likelihood should *not* be subject to the over-optimism problem that plagues the maximized log-likelihood. To use cross-validation for model selection, we simply calculate $CV(1)$ for the various models under consideration, and choose the one with the *highest* value.

As it turns out, leave-one-out cross-validation is intimately connected with the TIC. In fact the two are *asymptotically equivalent* as we'll now show. To begin note that, by a first-order Taylor Expansion of the leave-one-out estimator around the full-sample MLE we have

$$\begin{aligned} CV(1) &= \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\widehat{\theta}_{(t)}) \\ &= \frac{1}{T}\sum_{t=1}^{T} \left[\log f(y_t|\widehat{\theta}) + \frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'}\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right)\right] + o_p(1) \\ &= \frac{\ell(\widehat{\theta})}{T} + \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'}\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right) + o_p(1) \end{aligned}$$

so we simply need to show that

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'}\left(\widehat{\theta}_{(t)}-\widehat{\theta}\right)=-\frac{1}{T}\text{trace}\left(\widehat{J}^{-1}\widehat{K}\right)+o_p(1)$$

Everything that follows amounts to proving this statement.

To understand the preceding assertion, we'll need to take a slight detour and talk about *influence functions*, an idea from the robust estimation literature. Let $\mathbb{T}=\mathbb{T}(G)$ be a functional and G be some probability distribution. Then the influence function of $\mathbb{T}$ at a point y is defined as

$$\text{infl}(G,y)=\lim_{\epsilon\to 0}\frac{\mathbb{T}\left((1-\epsilon)\,G+\epsilon\delta_y\right)-\mathbb{T}(G)}{\epsilon}$$

where δ_y is a *point mass* at y, that is

$$\delta_y(a)\begin{cases}0,\ a<y\\ 1,\ a\geq y\end{cases}$$

All kinds of quantities that we know and love can be viewed as functionals of a distribution, for example the mean and variance.[1] Here we're going to be concerned with a particular functional that should look familiar from our last lecture:

$$\theta_0=\mathbb{T}(G)=\arg\min_{\theta\in\Theta}\ E_G\left[\log\left\{\frac{g(Y)}{f(Y|\theta)}\right\}\right]$$

What this says is that we can view θ_0 as the result of applying an *operator* $\mathbb{T}$ to the distribution G. In this case θ_0 is simply the pseudo-true value: the probability limit of the maximum likelihood estimator of θ based on $f(y|\theta)$. Clearly the pseudo-true value depends on the DGP, namely G. Different distributions G would yield different pseudo-true values for the *same* likelihood f. If we evaulate $\mathbb{T}$ at the *empirical* distribution $\widehat{G}$ we get the maximum likelihood estimator $\widehat{\theta}$ rather than the pseudo-true value θ_0.

[1] If you're interested in learning more about this, the book "Information Criteria and Statistical Modeling" by Konishi and Kitagawa (2008) provides a good overview.

The influence function is in fact a *functional derivative*. It allows us to evaulate, for example, how the pseudo-true value θ_0 would change if we *slightly* changed the distribution G that generated the data by "polluting" it with a tiny mass point located at y. We could also consider how the maximum likelihood estimator, $\widehat{\theta}$, would change if we slightly changed the dataset, represented by empirical distribution function. Now, since the empirical distribution is given by

$$\widehat{G}(a) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{y_t \leq a\right\} = \frac{1}{T}\sum_{t=1}^{T} \delta_{y_t}(a)$$

we can re-write it as

$$\widehat{G} = (1 - 1/T)\widehat{G}_{(t)} + \delta_{y_t}/T$$

where $\widehat{G}_{(t)}$ is the empirical distribution with y_t excluded from the dataset. Applying $\mathbb{T}$ to both sides, subtracting $\mathbb{T}(\widehat{G}_{(t)})$ and multiplying the right-hand-side by T/T, we have

$$\mathbb{T}(\widehat{G}) - \mathbb{T}(\widehat{G}_{(t)}) = \frac{1}{T}\left[\frac{\mathbb{T}\left((1 - 1/T)\widehat{G}_{(t)} + \delta_{y_t}/T\right) - \mathbb{T}(\widehat{G}_{(t)})}{1/T}\right]$$

If we take $\epsilon = 1/T$, the term in square brackets is *exactly* the expression whose limit is defined as the influence function. Hence, for T large we have the approximation

$$\mathbb{T}(\widehat{G}) - \mathbb{T}(\widehat{G}_{(t)}) = \frac{1}{T}\text{infl}\left(\widehat{G}_{(t)}, y_t\right) + o_p(1)$$

Now, since $\mathbb{T}(\widehat{G}) = \widehat{\theta}$ and $\mathbb{T}(\widehat{G}_{(t)}) = \widehat{\theta}_{(t)}$, we have the following expression for the leave-one-out estimator

$$\begin{aligned}\widehat{\theta}_{(t)} &= \widehat{\theta} - \frac{1}{T}\text{infl}\left(\widehat{G}_{(t)}, y_t\right) + o_p(1) \\ &= \widehat{\theta} - \frac{1}{T}\text{infl}\left(\widehat{G}, y_t\right) + o_p(1)\end{aligned}$$

The second expression indicates that dropping one observation is asymptotically negligible in its effect, through the empirical CDF, on the influence function. Now, it can be shown that the influence function for maximum likelihood estimation is

$$\text{infl}(G, y) = J^{-1}\left(\frac{\partial \log f(y|\theta_0)}{\partial \theta}\right)$$

where $\theta_0 = \mathbb{T}(G)$ is the pseudo-true value.[2] Hence, evaluating this expression at $\widehat{G}$ and y_t and substituting into our expression for $\widehat{\theta}_{(t)}$

$$\widehat{\theta}_{(t)} = \widehat{\theta} - \frac{1}{T}\widehat{J}^{-1}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right) + o_p(1)$$

since $\mathbb{T}(\widehat{G}) = \widehat{\theta}$. This gives us an asymptotic expansion for $\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right)$, namely

$$\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right) = -\frac{1}{T}\widehat{J}^{-1}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right) + o_p(1)$$

which is exactly what we need. Finally, substituting this back into the expression we initially set out to prove,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta'}\left(\widehat{\theta}_{(t)} - \widehat{\theta}\right) &= -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)'\frac{\widehat{J}^{-1}}{T}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right) + o_p(1) \\
&= -\frac{1}{T}\text{trace}\left\{\widehat{J}^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)\left(\frac{\partial \log f(y_t|\widehat{\theta})}{\partial \theta}\right)'\right]\right\} \\
&\quad + o_p(1) \\
&= -\frac{1}{T}\text{trace}\left\{\widehat{J}^{-1}\widehat{K}\right\}
\end{aligned}$$

Although we've used cross-validation to evaulate the KL divergence here, it's actually a very general idea and can be used with *any* loss function. For example, we could use a zero-one loss function for binary predictions, or a

[2]For details, see the Appendix to this section.

quadratic loss function for continuous predictions. The idea is the same in any case: we creat a "pseudo-out-of-sample observations" by witholding one observation at a time, and use these to evaluate the loss. While cross-validation can be computationally intensive, it turns out that there's a shortcut for models that can be expressed as linear smoothers. You'll examine a simple case of this on the problem set. It's also worth pointing out that there are varieties of cross-validation other than leave-one-out, but these have different properties.

Appendix: Deriving the Influence Function for MLE

In the preceding argument I claimed that the influence function for MLE is given by

$$\text{infl}(G, y) = J^{-1}\left(\frac{\partial \log f(y|\theta_0)}{\partial \theta}\right)$$

Here's a proof of this statement, following Chapter 5 of Konishi & Kitagawa (2008).[3] First, note that the functional $\mathbb{T}$ for MLE can be (implicitly) defined in terms of the score by

$$\int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} dG\left(z\right) = 0$$

where $\mathbb{T}\left(G\right) = \theta_0$. Now, to calculate the influence function, we need to evaulate $\mathbb{T}$ not at G but at $\left(1-\epsilon\right)G + \epsilon\delta_y$. Substituting, we have

$$\int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}((1-\epsilon)G+\epsilon\delta_y)} d\left(\left(1-\epsilon\right)G\left(z\right) + \epsilon\delta_y\left(z\right)\right) = 0$$

Note that the pseudo-true value has changed to $\theta^* = \mathbb{T}(\left(1-\epsilon\right)G + \epsilon\delta_y) \neq \theta_0$ since we're evaluating the functional at a different distribution than G. In fact, the preceding expression gives θ^* as an *implicit function* of ϵ.

Now, if we differentiate both sides of the preceding expression with respect

[3] Many thanks to Laura Liu for working out the details of this argument and typing it up for me!

to ϵ and evaulate the result at $\epsilon = 0$, we have

$$\int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} d\left(\delta_y\left(z\right) - G\left(z\right)\right)$$
$$+ \int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta \partial \theta'}\right|_{\theta=\mathbb{T}(G)} dG\left(z\right) \cdot \frac{\partial}{\partial \epsilon}\left[\mathbb{T}\left(\left(1-\epsilon\right)G + \epsilon\delta_y\right)\right] = 0 \qquad (1)$$

To ensure the uniqueness of $\frac{\partial}{\partial \epsilon}\left[\mathbb{T}\left(\left(1-\epsilon\right)G + \epsilon\delta_y\right)\right]$, we need the following condition (equation 5.17 in the book)

$$\int \frac{\partial}{\partial \epsilon}\left[\mathbb{T}\left(\left(1-\epsilon\right)G + \epsilon\delta_y\right)\right] dG\left(z\right) = 0$$

So the first term in equation (1) becomes

$$\int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} d\left(\delta_y\left(z\right) - G\left(z\right)\right) = \int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} d\delta_y\left(z\right)$$
$$= \left.\frac{\partial \log f\left(y|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} \qquad (2)$$

Thus,

$$\begin{aligned}
\text{infl}\left(\text{G}, \text{y}\right) &= \lim_{\epsilon \to 0} \frac{\mathbb{T}\left(\left(1-\epsilon\right)G + \epsilon\delta_y\right) - \mathbb{T}\left(G\right)}{\epsilon} \\
&= \frac{\partial}{\partial \epsilon}\left[\mathbb{T}\left(\left(1-\epsilon\right)G + \epsilon\delta_y\right)\right] \\
&= \left\{ - \int \left.\frac{\partial \log f\left(z|\theta\right)}{\partial \theta \partial \theta'}\right|_{\theta=\mathbb{T}(G)} dG\left(z\right) \right\}^{-1} \left.\frac{\partial \log f\left(y|\theta\right)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} \\
&= J^{-1} \frac{\partial \log f\left(y|\theta_0\right)}{\partial \theta}
\end{aligned}$$

where the third line is given by equations (1) and (2), and the fourth line follows the definitions of J and θ_0.

3 Mallow's C_p

Suppose that we want to predict y from $\mathbf{x}$ using a linear regression model:

$$\underset{(T\times 1)}{\mathbf{y}} = \underset{(T\times K)}{X}\underset{(K\times 1)}{\beta} + \boldsymbol{\epsilon}$$

Where $E[\boldsymbol{\epsilon}|X] = 0$ and $Var(\boldsymbol{\epsilon}|X) = \sigma^2\mathbf{I}$. We know that the conditional mean is the minimum mean-squared error predictor. This means that if β were *known*, we could never improve upon simply using all the regressors for prediction. But since β must be *estimated* from the data, a bias-variance tradeoff arises. In particular, we might be better off *excluding* a regressor with small coefficient, since it adds very little predictive power but introduces additional estimation uncertainty. Mallow's C_p is a model selection criterion that trys to capture this idea by approximating the *predictive mean squared error* of each model, relative to the infeasible optimum where β is known.

We'll now consider using *subsets* of X rather than the full data matrix. Let X_M denote a design matrix that possibly excludes some columns of X. The index M refers to a particular *model*. Accordingly, let $\widehat{\beta}_M$ be the least-squares estimator based on the design matrix X_M. We'll adopt the convention that $\widehat{\beta}_M$ is padded out with zeros for the elements of β that are *not estimated* under model M. This way we can write

$$X\widehat{\beta}_M = X_{(-M)}\mathbf{0} + X_M(X_M'X_M)^{-1}X_M'\mathbf{y} = P_M\mathbf{y}$$

Now, suppose we want to compare the predictive power of the competing estimators $\widehat{\beta}_M$ using mean-squared error. A naïve idea would be to use in-sample prediction error to compare models:

$$RSS(M) = (\mathbf{y} - X\widehat{\beta}_M)'(\mathbf{y} - X\widehat{\beta}_M)$$

As is well-known, however, the residual sum of squares can *never* decrease even as we add irrelevant predictors to our model. In constrast, out-of-sample

predictive ability can easily decrease when we add more predictors: there's a bias-variance tradeoff that arises from estimation uncertainty. Somehow or other we need to develop a criterion to take this into account.

We'll start off by calculating the predictive mean-squared error of $X\widehat{\beta}_M$ relative to the infeasible optimum, namely $X\beta$. Let $P_M = X_M(X_M'X_M)^{-1}X_M$. Then we have

$$\begin{aligned}
\left|\left|X\widehat{\beta}_M - X\beta\right|\right|^2 &= (P_M\mathbf{y} - X\beta)'(P_M\mathbf{y} - X\beta) \\
&= \left\{P_M(Y - X\beta) - (\mathbf{I} - P_M)X\beta\right\}'\left\{P_M(Y - X\beta) - (\mathbf{I} - P_M)X\beta\right\} \\
&= \left\{P_M\boldsymbol{\epsilon} - (\mathbf{I} - P_M)X\beta\right\}'\left\{P_M\boldsymbol{\epsilon} + (\mathbf{I} - P_M)X\beta\right\} \\
&= \boldsymbol{\epsilon}'P_M'P_M\boldsymbol{\epsilon} - \beta'X'(\mathbf{I} - P_M)'P_M\boldsymbol{\epsilon} \\
&\qquad - \boldsymbol{\epsilon}'P_M'(\mathbf{I} - P_M)X\beta + \beta'X'(\mathbf{I} - P_M)(\mathbf{I} - P_M)X\beta \\
&= \boldsymbol{\epsilon}'P_M\boldsymbol{\epsilon} + \beta'X'(\mathbf{I} - P_M)X\beta
\end{aligned}$$

since P_M and $(\mathbf{I} - P_M)$ are both symmetric and idempotent and their product in any order is zero. Thus, evaluating the predictive mean-squared error conditional on X, we have

$$\begin{aligned}
\text{MSE}(M|X) &= E\left[(X\widehat{\beta}_M - X\beta)'(X\widehat{\beta}_M - X\beta)|X\right] \\
&= E\left[\boldsymbol{\epsilon}'P_M\boldsymbol{\epsilon}|X\right] + E\left[\beta'X'(\mathbf{I} - P_M)X\beta|X\right] \\
&= E\left[\text{trace}\left\{\boldsymbol{\epsilon}'P_M\boldsymbol{\epsilon}\right\}|X\right] + \beta'X'(\mathbf{I} - P_M)X\beta \\
&= \text{trace}\left\{E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|X]P_M\right\} + \beta'X'(\mathbf{I} - P_M)X\beta \\
&= \text{trace}\left\{\sigma^2 P_M\right\} + \beta'X'(\mathbf{I} - P_M)X\beta \\
&= \sigma^2 k_M + \beta'X'(\mathbf{I} - P_M)X\beta
\end{aligned}$$

where k_M denotes the number of regressors included in X_M and we have used the fact that the trace of a projection matrix equals its dimension.

So far, so good: we have derived an expression of the predictive mean-squared error of each model M. The problem is that it's infeasible: it depends on the unknown quantities σ^2 and β. To get around this, Mallow's C_p con-

structs an *unbiased* estimator of MSE. We proceed as follows. First, let $\widehat{\beta}$ be the estimator based on the full set of regressors, i.e. $\widehat{\beta} = (X'X)^{-1}X'\mathbf{y}$ and let P_X be the corresponding projection matrix so that we have

$$X\widehat{\theta} = X(X'X)^{-1}X'\mathbf{y} = P_X\mathbf{y}$$

Using the fact that $P_M P_X = P_X P_M = P_M$,

$$\begin{aligned}
E\left[\widehat{\beta}'X'(\mathbf{I}-P_M)X\widehat{\beta}|X\right] &= E\left[\mathbf{y}'P_X'(\mathbf{I}-P_M)P_X\mathbf{y}|X\right] \\
&= E\left[\mathbf{y}'(P_X'P_X - P_X'P_MP_X)\mathbf{y}|X\right] \\
&= E\left[\mathbf{y}'(P_X - P_M)\mathbf{y}|X\right] \\
&\vdots \\
&\boxed{\textbf{You'll fill in the details on the problem set}} \\
&\vdots \\
&= \beta'X'(\mathbf{I}-P_M)X\beta + \sigma^2(K-k_M)
\end{aligned}$$

In other words, substituting the estimator $\widehat{\beta}$ from the full model in order to estaimate $\beta'X(\mathbf{I}-P_M)X\beta$ *doesn't work*. The estimator $\widehat{\beta}'X'(\mathbf{I}-P_M)X\widehat{\beta}$ is *biased upwards*. However, we have an explicit expression for the bias, namely $\sigma^2(K-k_M)$. This means that if we can find an unbiased estimator of σ^2, we can *correct* the bias in our estimator of $\beta'X(\mathbf{I}-P_M)X\beta$. Fortunately there is an obvious unbiased estimator of σ^2: we simply use the residuals from the full model:

$$\widehat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{I}-P_X)\mathbf{y}}{(T-K)}$$

Thus,

$$E[\widehat{\beta}'X'(\mathbf{I}-P_M)X\widehat{\beta} - \widehat{\sigma}^2(K-k_M)|X] = \beta'X(\mathbf{I}-P_M)X\beta$$

Now we've managed to construct an unbiased estimator of the *second* term of the MSE. What about the first? This one is easy. We already have an

unbiased estimator of $\widehat{\sigma}^2$ and k_M is a known constant: it's simply the number of regressors in model M. Therefore, collecting terms

$$\begin{aligned} \text{MC}_M &= \widehat{\sigma}^2 k_M + \left[\widehat{\beta}' X'(\mathbf{I} - P_M)X\widehat{\beta} - \widehat{\sigma}^2(K - k_M)\right] \\ &= \widehat{\beta}' X'(\mathbf{I} - P_M)X\widehat{\beta} + \widehat{\sigma}^2(k_M - K) \end{aligned}$$

is an unbiased estimator of $\text{MSE}(M|X)$. This criterion contains exactly the same information as Mallow's C_p but expressed in a slightly different way. To get the formula from the textbooks, we need to do some more algebra. First,

$$\begin{aligned} MC_M - 2\widehat{\sigma}^2 k_M &= \widehat{\beta}' X'(\mathbf{I} - P_M)X\widehat{\beta} - K\widehat{\sigma}^2 \\ &= \mathbf{y}'(P_X - P_M)\mathbf{y} - K\widehat{\sigma}^2 \\ &\vdots \\ &\boxed{\textbf{You'll fill in the details on the problem set}} \\ &\vdots \\ &= \mathbf{y}'(\mathbf{I} - P_M)\mathbf{y} - T\widehat{\sigma}^2 \\ &= RSS(M) - T\widehat{\sigma^2} \end{aligned}$$

Substituting this into the expression for MC_M we see that

$$\text{MC}_M = RSS(M) + \widehat{\sigma}^2(2k_M - T)$$

which is much easier to interpret than the formula we have before. Finally, dividing through by $\widehat{\sigma}^2$ gives Mallow's C_p

$$C_p(M) = \frac{RSS(M)}{\widehat{\sigma}^2} + 2k_M - T$$

This expression tells us how we need to *adjust* the residual sum of squares to account for the fact that in-sample fit is a misleading guide to out-of-sample predictive performance.

4 Bayesian Information Criterion

Since Frank2 talked about this in his part of the course, I won't discuss this derivation in class but I wanted to provide the details for completeness. As in our derivation of TIC and AIC, we'll consider a setting with an iid sample of scalar random variables $Y_1, \ldots, Y_T$. The results still hold in the more general case, but this simplifies the notation.

4.1 Overview of the BIC

Despite its name, the BIC is *not* a Bayesian procedure. It is a large-sample Frequentist *approximation* to Bayesian model selection:

1. Begin with a uniform prior on the set of candidate models so that it suffices to maximize the Marginal Likelihood.

2. The BIC is a large sample approximation to the Marginal Likelihood:

$$\int \pi(\beta_i) f_i(\mathbf{y}|\beta_i) d\beta_i$$

   where i indexes models M_i in a set $\mathcal{M}$.

3. As usual when Bayesian procedures are subjected to Frequentist asymptotics, the priors on parameters vanish in the limit.

4. We proceed by a *Laplace Approximation* to the Marginal Likelihood

4.2 Laplace Approximation

For the moment simplify the notation by suppressing dependence on M_i. We want to approximate:

$$\int \pi(\beta) f(\mathbf{y}|\beta) d\beta$$

This is actually a common problem in applications of Bayesian inference:

- Notice that $\pi(\beta)f(\mathbf{y}|\beta)$ is the *kernel* of some probability density, i.e. the density without its normalizing constant.
- *How do we know this?* By Bayes' Rule

$$\pi(\beta|\mathbf{y}) = \frac{\pi(\beta)f(\mathbf{y}|\beta)}{\int \pi(\beta)f(\mathbf{y}|\beta)d\beta}$$

  is a proper probability density and the denominator is *constant* with respect to β. (The parameter has been "integrated out.")
- In Bayesian inference, we specify $\pi(\beta)$ and $f(\mathbf{y}|\beta)$, so $\pi(\beta)f(\mathbf{y}|\beta)$ is known. But to calculate the posterior we need to *integrate* to find the normalizing constant.
- Only in special cases (e.g. conjugate families) can we find the exact normalizing constant. Typically some kind of approximation is needed:
  - Importance Sampling
  - Markov-Chain Monte Carlo (MCMC)
  - *Laplace Approximation*

The Laplace Approximation is an *analytical approximation* based on Taylor Expansion arguments. In Bayesian applications, the expansion is carried out around the posterior mode, i.e. the mode of $\pi(\beta)f(\mathbf{y}|\beta)$, but we will expand around the Maximum likelihood estimator.

Proposition 4.1 (Laplace Approximation)**.**

$$\int \pi(\beta)f(\mathbf{y}|\beta)d\beta \approx \frac{\exp\left\{\ell(\hat{\beta})\right\}\pi(\hat{\beta})(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\beta})\right|^{1/2}}$$

Where $\hat{\beta}$ *is the* maximum likelihood estimator*,* p *the dimension of* β *and*

$$J(\hat{\beta}) = -\frac{1}{n}\frac{\partial^2 \log f(\mathbf{y}|\hat{\beta})}{\partial\beta\partial\beta'}$$

Proof. A rigorous proof of this result is complicated. The following is a sketch. First write $\ell(\beta)$ for $\log f(\mathbf{y}|\beta)$ so that

$$\pi(\beta)f(\mathbf{y}|\beta) = \pi(\beta)\exp\left\{\log f(\mathbf{y}|\beta)\right\} = \pi(\beta)\exp\left\{\log \ell(\beta)\right\}$$

By a second-order Taylor Expansion around the MLE $\hat{\beta}$

$$\ell(\beta) = \ell(\hat{\beta}) + \frac{1}{2}\left(\beta - \hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\ell \tag{3}$$

since the derivative of $\ell(\beta)$ is zero at $\hat{\beta}$ by the definition of MLE. A first-order expansion is sufficient for $\pi(\beta)$ because the derivative does not vanish at $\hat{\beta}$

$$\pi(\beta) = \pi(\hat{\beta}) + \frac{\partial\pi(\hat{\beta})}{\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\pi \tag{4}$$

Substituting Equations 3 and 4,

$$\begin{aligned}
\int \pi(\beta)f(\mathbf{y}|\beta)d\beta &= \int \exp\left\{\ell(\hat{\beta}) + \frac{1}{2}\left(\beta - \hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\ell\right\} \\
&\qquad \times \left[\pi(\hat{\beta}) + \left(\beta - \hat{\beta}\right)' \frac{\partial\pi(\hat{\beta})}{\partial\beta} + R_\pi\right] d\beta \\
&= \exp\left\{\ell(\hat{\beta})\right\}(I_1 + I_2 + I_3)
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \pi(\hat{\beta}) \int \exp\left\{\frac{1}{2}\left(\beta - \hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\ell\right\} d\beta \\
I_2 &= \frac{\partial\pi(\hat{\beta})}{\partial\beta'} \int \left(\beta - \hat{\beta}\right) \exp\left\{\frac{1}{2}\left(\beta - \hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\ell\right\} d\beta \\
I_3 &= \int R_\pi \exp\left\{\frac{1}{2}\left(\beta - \hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta - \hat{\beta}\right) + R_\ell\right\} d\beta
\end{aligned}$$

Under certain regularity conditions (not the standard ones!) we can treat R_ℓ and R_π as approximately equal to zero for large n uniformly in β, so that

$$\begin{aligned} I_1 &\approx \pi(\hat{\beta}) \int \exp\left\{\frac{1}{2}\left(\beta-\hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta-\hat{\beta}\right)\right\} d\beta \\ I_2 &\approx \frac{\partial \pi(\hat{\beta})}{\partial\beta'} \int \left(\beta-\hat{\beta}\right) \exp\left\{\frac{1}{2}\left(\beta-\hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta-\hat{\beta}\right)\right\} d\beta \\ I_3 &\approx 0 \end{aligned}$$

Because $\hat{\beta}$ is the MLE,

$$\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}$$

must be negative definite, so

$$-\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}$$

is positive definite. It follows that

$$\exp\left\{\frac{1}{2}\left(\beta-\hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta-\hat{\beta}\right)\right\} = \exp\left\{-\frac{1}{2}\left(\beta-\hat{\beta}\right)' \left[\left(-\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\right)^{-1}\right]^{-1}\left(\beta-\hat{\beta}\right)\right\}$$

can be viewed as the kernel of a Normal distribution with mean $\hat{\beta}$ and variance matrix

$$\left(-\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\right)^{-1}$$

Thus,

$$\int \exp\left\{\frac{1}{2}\left(\beta-\hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta-\hat{\beta}\right)\right\} d\beta = (2\pi)^{p/2} \left|\left(-\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\right)^{-1}\right|^{1/2}$$

and

$$\int \left(\beta-\hat{\beta}\right) \exp\left\{\frac{1}{2}\left(\beta-\hat{\beta}\right)' \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'}\left(\beta-\hat{\beta}\right)\right\} d\beta = 0$$

Therefore,

$$\begin{aligned}
\int \pi(\beta) f(\mathbf{y}|\beta) d\beta &\approx \exp\left\{\ell(\hat{\beta})\right\} \pi(\hat{\beta}) (2\pi)^{p/2} \left| \left(-\frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'} \right)^{-1} \right|^{1/2} \\
&= \exp\left\{\ell(\hat{\beta})\right\} \pi(\hat{\beta}) (2\pi)^{p/2} \left| n \left(-\frac{1}{n} \frac{\partial^2 \ell(\hat{\beta})}{\partial\beta\partial\beta'} \right) \right|^{-1/2} \\
&= \frac{\exp\left\{\ell(\hat{\beta})\right\} \pi(\hat{\beta}) (2\pi)^{p/2}}{n^{p/2} \left| J(\hat{\beta}) \right|^{1/2}}
\end{aligned}$$

□

4.3 Finally the BIC

Now we re-introduce the dependence on the model M_i. Taking logs of the Laplace Approximation and multiplying by two (again, this is traditional but has no effect on model comparisons)

$$\begin{aligned}
2 \log f(y|M_i) &= 2 \log \left\{ \int f_i(y|\beta_i) \pi(\beta_i) d\beta_i \right\} \\
&\approx 2\ell(\hat{\beta}_i) - p \log(n) + p \log(2\pi) - \pi(\hat{\beta}_i) - \log \left| J(\hat{\beta}_i) \right|
\end{aligned}$$

The first two terms are $O_p(n)$ and $O_p(\log n)$, while the last three are $O_p(1)$, hence negligible as $n \to \infty$. This gives us Schwarz's BIC

$$BIC(M_i) = 2 \log f_i(\mathbf{y}|\hat{\beta}_i) - p \log n$$

We choose the model M_i for which $BIC(M_i)$ is largest. Notice that the prior on the parameter, $\pi(\beta)$, drops out in the limit, and recall that we began by putting a uniform prior on the *models* under consideration.