# SELECTING REGRESSORS USING THE BAYESIAN INFORMATION CRITERION (BIC)

Abstract. In the context of linear regression and OLS, we discuss information-criteria-based approaches for selecting relevant regressors our of a list of potential regressors. We discuss consistency and oracle properties, and post-selection inference.

Contents



## 1. Selecting regressors

We will discuss the problem of selecting relevant regressors in the context of the linear regression model. However, the procedures discussed here can be generalized and similarly applied with nonlinear models such as logit, GMM, and etc. Consider a linear regression model with $k$ *potential regressors*:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{i,j} + U_i, \tag{1.1}$$

$$EX_{i,j}U_i = 0, \quad j = 1, \ldots, k.$$

For now we assume that the number of potential regressors is small: $k$ is fixed an does not depend on $n$.

Let the set $\mathcal{A}$ denote the list of regressors with non-zero coefficients:

$$\mathcal{A} = \{j : \beta_j \neq 0\}.$$

For example, $\mathcal{A} = \{1, 3, 7\}$ implies that only the regressors $X_{i,1}$, $X_{i,3}$, and $X_{i,7}$ have non-zero coefficients, and that the remaining regressors have coefficients equal to zero. We use $\mathcal{A}_0$ to denote the *true set of relevant regressors*: i.e. the true data generating process (DGP) for $Y_i$ only includes the regressors in $\mathcal{A}_0$:

$$Y_i = \sum_{j \in \mathcal{A}_0} \beta_j X_{i,j} + U_i.$$

Our goal is to estimate $\mathcal{A}_0$ using the data $\{(Y_i, X_i')', i = 1, \ldots, n\}$. We use $\hat{\mathcal{A}}_n$ to denote an estimated set of relevant regressors produced by a selection procedure. We say that the selection procedure is consistent if

$$P\left(\hat{\mathcal{A}}_n = \mathcal{A}_0\right) \to 1 \tag{1.2}$$

as $n \to \infty$.

Let $\beta = (\beta_1, \ldots, \beta_k)'$, and by $\beta_{\mathcal{A}}$ we denote the subvector of $\beta$ that includes only the coefficients in $\mathcal{A}$:

$$\beta_{\mathcal{A}} = (\beta_j : j \in \mathcal{A}).$$

We use $|\mathcal{A}|$ to denote the number of elements in $\mathcal{A}$, and hence $\beta_{\mathcal{A}}$ is a $|\mathcal{A}|$-subvector of the $k$-vector $\beta$.

Suppose a procedure produced a set of selected regressors $\hat{\mathcal{A}}_n$ and vector of estimates $\hat{\beta}_n = (\hat{\beta}_{n,1}, \ldots, \hat{\beta}_{n,k})'$. Note that it is reasonable to set $\hat{\beta}_{n,j} = 0$ for $j \notin \hat{\mathcal{A}}_n$. We say that the procedure is *oracle* if in addition to the consistency property in (1.2),

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}_0} - \beta_{\mathcal{A}_0}) \to_d N(0, V(\mathcal{A}_0)),$$

where $V(\mathcal{A}_0)$ is the best asymptotic variance one can obtain when the true model $\mathcal{A}_0$ is known. The oracle property means that not only the econometrician consistently selects the true regressors, but also the coefficients on the relevant regressors are estimated as precisely as when the set of the true relevant regressors in the DGP is known.

## 2. BIC

Recall that if the econometrician tries to select the regressors by minimizing the sample sum of squared residuals (SSR), or equivalently maximizing $R^2$, the procedure would result in overfitting: the SSR is monotone non-increasing in the number of included regressors. The idea behind BIC is to penalize the SSR for the model complexity.

Let $X_i = (X_{i,1}, \ldots, X_{i,k})'$, and define $X_{i,\mathcal{A}}$ as the subvector of $X_i$ that includes only the regressors in $\mathcal{A}$:

$$X_{i,\mathcal{A}} = (X_{i,j} : j \in \mathcal{A}).$$

Again, $X_{i,\mathcal{A}}$ is a $|\mathcal{A}|$-subvector of the $k$-vector $X_i$. The true DGP can now be written as

$$\begin{aligned} Y_i &= \sum_{j \in \mathcal{A}_0} \beta_j X_{i,j} + U_i \\ &= X'_{i,\mathcal{A}_0} \beta_{\mathcal{A}_0} + U_i. \end{aligned}$$

Let $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ denote the OLS estimator of $\beta_{\mathcal{A}}$ that only uses the regressors in $\mathcal{A}$:

$$\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) = \left( \sum_{i=1}^n X_{i,\mathcal{A}} X'_{i,\mathcal{A}} \right)^{-1} \sum_{i=1}^n X_{i,\mathcal{A}} Y_i.$$

We can set

$$\hat{\beta}_{n,\mathcal{A}^c}(\mathcal{A}) = 0,$$

and view $\hat{\beta}_n(\mathcal{A}) = (\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})', \hat{\beta}_{n,\mathcal{A}^c}(\mathcal{A})')'$ as the estimator *of* $\beta = (\beta'_{\mathcal{A}}, \beta'_{\mathcal{A}^c})'$ *under the model* $\mathcal{A}$. The corresponding SSR is given by

$$SSR_n(\mathcal{A}) = \sum_{i=1}^n \left( Y_i - X'_{i,\mathcal{A}} \hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) \right)^2.$$

The complexity of the model $\mathcal{A}$ can be measured by the number of included regressors, i.e. the number of elements in $\mathcal{A}$. BIC for the model $\mathcal{A}$ is defined as

$$BIC_n(\mathcal{A}) = SSR_n(\mathcal{A}) + |\mathcal{A}| \log n,$$

where the second term is a *penalty term*. Note that a model with more included regressors receives a larger penalty. A BIC-based selection procedure selects the regressors by minimizing BIC across all possible models:

$$\hat{\mathcal{A}}_n^{BIC} = \arg\min_{\mathcal{A}} BIC_n(\mathcal{A}).$$

We show below that BIC selects the relevant regressors consistently.

**Proposition 2.1.** *Suppose that data are iid, $EX_iX_i'$ and $EU_i^2X_iX_i'$ are finite and positive definite, and $EU_i^2 < \infty$. Then $P\left(\hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) \to 1$ as $n \to \infty$.*

*Proof.* It suffices to show that for all $\mathcal{A} \neq \mathcal{A}_0$

$$P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) \to 1, \tag{2.1}$$

i.e. the true model $\mathcal{A}_0$ minimizes BIC with probability approaching one.

First, consider the *average* SSR for the true model:

$$\begin{aligned}
n^{-1}SSR_n(\mathcal{A}_0) &= n^{-1}\sum_{i=1}^n \left(Y_i - X_{i,\mathcal{A}_0}'\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0)\right)^2 \\
&= n^{-1}\sum_{i=1}^n \left(U_i - X_{i,\mathcal{A}_0}'(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0})\right)^2 \\
&= n^{-1}\sum_{i=1}^n U_i^2 + (\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0})'\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}'\right)(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0}) \\
&\quad - 2\left(n^{-1}\sum_{i=1}^n X_{i,\mathcal{A}_0}U_i\right)(\hat{\beta}_{n,\mathcal{A}_0}(\mathcal{A}_0) - \beta_{\mathcal{A}_0}) \\
&= EU_i^2 + o_p(1),
\end{aligned}$$

where the $o_p(1)$ term in the last line is by the LLN and consistency of the OLS estimator under the true model:

$$n^{-1}\sum_{i=1}^n U_i^2 = EU_i^2 + o_p(1),$$

$$\hat{\beta}_{n,\mathcal{A}_0} = \beta_{\mathcal{A}_0} + o_p(1),$$

$$n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}' = EX_{i,\mathcal{A}_0}X_{i,\mathcal{A}_0}' + o_p(1),$$

$$n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}_0}U_i = o_p(1).$$

Suppose that a model$\mathcal{A}$ that omits some relevant regressors:

$$(\mathcal{A} \cap \mathcal{A}_0) \neq \mathcal{A}_0.$$

Since the OLS estimator is in general inconsistent when there are omitted relevant regressors,

$$\hat{\beta}_n(\mathcal{A}) - \beta \to_p \delta \neq 0,$$

where $\hat{\beta}_{n,j}(\mathcal{A})$ is the corresponding element of $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ for $j \in \mathcal{A}$, and $\hat{\beta}_n(\mathcal{A}) = 0$ for $j \notin \mathcal{A}$. We have:

$$\begin{aligned}
n^{-1}SSR_n(\mathcal{A}) &= n^{-1}\sum_{i=1}^{n}\left(Y_i - X_i'\hat{\beta}_n(\mathcal{A})\right)^2 \\
&= n^{-1}\sum_{i=1}^{n}\left(U_i - X_i'\left(\hat{\beta}_n(\mathcal{A}) - \beta\right)\right)^2 \\
&= n^{-1}\sum_{i=1}^{n} U_i^2 + (\hat{\beta}_n(\mathcal{A}) - \beta)'\left(n^{-1}\sum_{i=1}^{n} X_iX_i'\right)(\hat{\beta}_n(\mathcal{A}) - \beta) \\
&\quad - 2\left(n^{-1}\sum_{i=1}^{n} X_iU_i\right)(\hat{\beta}_n(\mathcal{A}) - \beta) \\
&= EU_i^2 + \delta' EX_iX_i'\delta + o_p(1).
\end{aligned}$$

Note also that

$$|\mathcal{A}|\frac{\log n}{n} = o(1).$$

Therefore, for such a model $\mathcal{A}$,

$$\begin{aligned}
P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) &= P\left(n^{-1}BIC_n(\mathcal{A}) > n^{-1}BIC_n(\mathcal{A}_0)\right) \\
&= P\left(n^{-1}SSR_n(\mathcal{A}) + |\mathcal{A}|\frac{\log n}{n} > n^{-1}SSR_n(\mathcal{A}_0) + |\mathcal{A}_0|\frac{\log n}{n}\right)
\end{aligned}$$

$$= P\left(\delta' E X_i X_i' \delta + o_p(1) + o(1) > 0\right)$$

$$\to 1,$$

where convergence in the last line holds because $\delta \neq 0$ and $EX_iX_i'$ is positive definite.

Next, consider a model $\mathcal{A}$ such that

$$\mathcal{A}_0 \subset \mathcal{A}.$$

In this case, $\mathcal{A}$ contains all the relevant regressors as well as some irrelevant. The OLS estimator $\hat{\beta}_{n,\mathcal{A}}(\mathcal{A})$ is consistent and asymptotically normal:

$$n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) \to_d \Psi_{\mathcal{A}},$$

where

$$\Psi_{\mathcal{A}} \sim N\left(0, V(A)\right),$$
$$V(\mathcal{A}) = \left(EX_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right)^{-1} EU_i^2 X_{i,\mathcal{A}}X_{i,\mathcal{A}}' \left(EX_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right)^{-1}.$$

The result follows from

$$n^{-1/2}\sum_{i=1}^{n} X_i U_i \to_d \Phi_{\mathcal{A}},$$

where

$$\Phi_{\mathcal{A}} \sim N\left(0, EU_i^2 X_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right).$$

We have:

$$\begin{aligned}
SSR_n(\mathcal{A}) - \sum_{i=1}^{n} U_i^2 &= \sum_{i=1}^{n}\left(U_i - X_{i,\mathcal{A}}'(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}})\right)^2 - \sum_{i=1}^{n} U_i^2 \\
&= n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}})' \left(n^{-1}\sum_{i=1}^{n} X_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right) n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) \\
&\quad - 2\left(n^{-1/2}\sum_{i=1}^{n} X_{i,\mathcal{A}}U_i\right) n^{1/2}(\hat{\beta}_{n,\mathcal{A}}(\mathcal{A}) - \beta_{\mathcal{A}}) \\
&\to_d \Psi_{\mathcal{A}}'\left(EX_{i,\mathcal{A}}X_{i,\mathcal{A}}'\right)\Psi_{\mathcal{A}} - 2\Phi_{\mathcal{A}}'\Psi_{\mathcal{A}}
\end{aligned}$$

$$= O_p(1).$$

By the same arguments,

$$SSR_n(\mathcal{A}_0) - \sum_{i=1}^{n} U_i^2 \to_d \Psi'_{\mathcal{A}_0} \left( EX_{i,\mathcal{A}_0} X'_{i,\mathcal{A}_0} \right) \Psi_{\mathcal{A}_0} - 2\Phi'_{\mathcal{A}_0} \Psi_{\mathcal{A}_0}$$

$$= O_p(1).$$

Lastly, when $\mathcal{A}_0 \subset \mathcal{A}$,

$$\begin{aligned} P\left(BIC_n(\mathcal{A}) > BIC_n(\mathcal{A}_0)\right) &= P\left(SSR_n(\mathcal{A}) - SSR_n(\mathcal{A}_0) > (|\mathcal{A}_0| - |\mathcal{A}|)\log n\right) \\ &= P\left(O_p(1) > (|\mathcal{A}_0| - |\mathcal{A}|)\log n\right) \\ &\to 1, \end{aligned}$$

where convergence in the last line holds since $|\mathcal{A}_0| < |\mathcal{A}|$, and therefore

$$(|\mathcal{A}_0| - |\mathcal{A}|)\log n \to -\infty.$$

□

## 3. Post BIC inference

Suppose the econometrician selects the true model using $\hat{\mathcal{A}}_n^{BIC}$ and conducts inference using $\hat{\beta}_n(\hat{\mathcal{A}}_n^{BIC})$. For $j \in \hat{\mathcal{A}}_n^{BIC}$, the distribution of the estimator of the $j$-the coefficient is given by

$$\begin{aligned} &P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \le u\right) \\ &= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \le u, \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) \\ &\quad + P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \le u, \hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) \\ &= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \le u, \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) + o(1) \\ &= P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \le u \mid \hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) P\left(\hat{\mathcal{A}}_n^{BIC} = \mathcal{A}_0\right) + o(1) \end{aligned}$$

$$= P\left(n^{1/2}(\hat{\beta}_{n,j}(\mathcal{A}_0) - \beta_j) \leq u\right)(1 + o(1)) + o(1)$$
$$= P\left(n^{1/2}(\hat{\beta}_{n,j}(\mathcal{A}_0) - \beta_j) \leq u\right) + o(1).$$

where the second equality holds by

$$P\left(n^{1/2}(\hat{\beta}_{n,j}(\hat{\mathcal{A}}_n^{BIC}) - \beta_j) \leq u, \hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) \leq P\left(\hat{\mathcal{A}}_n^{BIC} \neq \mathcal{A}_0\right) = o(1).$$

Hence, the BIC-based selection and estimation procedure is an oracle procedure.

## 4. Akaike Information Criterion (AIC)

AIC is another popular criterion for model selection (and actually precedes BIC). AIC for a model $\mathcal{A}$ is defined as

$$AIC_n(\mathcal{A}) = SSR_n(\mathcal{A}) + 2|\mathcal{A}|.$$

In comparison with BIC, AIC penalizes the model complexity less heavily and, therefore, tends to select a bigger model with more regressors than BIC.

By the same arguments as in the proof of Proposition 2.1, for a model $\mathcal{A}$ that omits some relevant regressors, i.e. $(\mathcal{A} \cap \mathcal{A}_0) \neq \mathcal{A}_0$,

$$P\left(AIC_n(\mathcal{A}) > AIC_n(\mathcal{A}_0)\right) \to 1.$$

However, because AIC penalty is not sufficiently strong, if $\mathcal{A}_0 \subset \mathcal{A}$,

$$P\left(AIC_n(\mathcal{A}) > AIC_n(\mathcal{A}_0)\right) \nrightarrow 1.$$

Hence, while AIC detects omitted regressors with probability approaching one, it is more likely to overfit by also including some irrelevant regressors than BIC.

## 5. Limitations

One should note several limitations of our arguments. First, we assumed that $k$ is small (fixed) and some of our arguments do not apply when the number of potential

regressors is comparable to the sample size. However, this technical issue can be resolved with somewhat different arguments.

More importantly, our analysis ignores the situation where some $\beta_j$, while non-zero, are very close to zero. It is unreasonable to expect that the BIC (or any other procedure) can detect such small coefficients with a probability approaching one. Thus, even in the limit, regressors with very small coefficients are likely to be omitted from the model, which can potentially create an omitted variable bias. This shortcoming can be addressed using a double selection procedure, which will be discussed later in the context of Lasso.

Lastly, the BIC procedure may be infeasible in practice if the number of potential regressors is very large. There are $2^k$ possible models $\mathcal{A}$, and if $k = 30$ one has to run and compare over 1 billion potential regressions. For $k = 40$, one has to run over 1 trillion models. For example, suppose that the econometrician considers flexible specifications that include quadratic terms as well as pairwise interaction terms of the right-hand side variables. In that case, 10 potential right-hand side variables generate 65 potential regressors.