Infernce post-model-selection, simplicity, and the role of sample size

Hannes Leeb (University of Vienna)

CFE Workshop: Foundations for Ockham's Razor
CMU, June 22–24, 2012

Informal summary

We consider inference post-model-selection in two scenarios:

1. $d/n \approx 0$, where the candidate models are comparatively simple in relation to sample size.
2. $d/n \not\approx 0$, where the candidate models are comparatively complex.

In the first scenario, we show that certain aspects of inference post-model-selection are hard or ill-posed. In the second scenario, we show that prediction post-model-selection can be tackled successfully.

The findings for the second scenario are for Gaussian data and rely crucially on a property exhibited only by the Gaussian distribution: If the response and the explanatory variables are Gaussian, then every linear submodel is 'correct.' We show that a relaxed version of this property holds for a much larger class of distributions.

A rather simple linear model

Consider n independent observations (y_i, x_i, z_i), $i = 1, \ldots, n$ from the linear regression

$$y \quad = \quad \alpha x + \beta z + u. \tag{1}$$

The object of interest is α, and we consider two candidate models, namely the unrestricted m_2 model with two parameters, and the restricted model m_1 where $\beta = 0$. The (unknown) minimal true model depends on β an will be denoted by $m_*(\beta)$. A model-selection procedure is a data-driven rule $\hat{m}_*$ that takes on the values m_1 and m_2 and thus estimates $m_*(\beta)$.

Write $\hat{\alpha}(m_i)$ for the least-squares corresponding to the model m_i. The model-selection procedure $\hat{m}_*$ leads to the post-model-selection estimator $\hat{\alpha}(\hat{m}_*)$ for α.

Consistent model selection

Write $P_{n,\alpha,\beta}$ for the distribution of a sample of size n if the true parameters are α and β. We assume that the model selector $\hat{m}_*$ is consistent, in the sense that

$$P_{n,\alpha,\beta}\left(\hat{m}_* = m_*(\beta)\right) \quad \overset{n\to\infty}{\longrightarrow} \quad 1$$

holds for each pair (α, β). [E.g., BIC, MDL, consistent pre-testing, etc.]

For large n, this suggests that $\hat{\alpha}(\hat{m}_*)$ behaves like the infeasible 'estimator' $\hat{\alpha}(m_*(\beta))$.

An oracle property

Assume throughout that the errors in (1) are i.i.d. normal and that $n\mathrm{Var}(\hat{\alpha}(m_2), \hat{\beta}(m_2)) \to \Sigma > 0$ (can be relaxed; e.g. LAN). With this,

$$\sqrt{n}(\hat{\alpha}(m_2) - \alpha) \quad \xrightarrow{P_{n,\alpha,\beta}} \quad \Phi_{\sigma(m_2)}(\cdot)$$

and, if $\beta = 0$ or, equivalently, if $m_*(\beta) = 1$, also

$$\sqrt{n}(\hat{\alpha}(m_1) - \alpha) \quad \xrightarrow{P_{n,\alpha,0}} \quad \Phi_{\sigma(m_1)}(\cdot),$$

where $\sigma(m_1) \leq \sigma(m_2)$. Write $F_{n,\alpha,\beta}(t)$ for the c.d.f. of $\sqrt{n}(\hat{\alpha}(\hat{m}_*) - \alpha)$. Since $\hat{m}_*$ is consistent, we have, for each pair α, β, that

$$F_{n,\alpha,\beta}(t) - \Phi_{\sigma(m_*(\beta))}(t) \quad \xrightarrow{n\to\infty} \quad 0.$$

Theorem 1

If $\sigma(m_1) < \sigma(m_2)$, then

$$\sup_{|\beta|<1/\sqrt{n}} \left| F_{n,\alpha,\beta}(t) - \Phi_{\sigma(m_*(\beta))}(t) \right| \quad \xrightarrow{n\to\infty} \quad \delta \quad > \quad 0$$

for each α.

An oracle property

Assume throughout that the errors in (1) are i.i.d. normal and that $n\mathrm{Var}(\hat{\alpha}(m_2), \hat{\beta}(m_2)) \to \Sigma > 0$ (can be relaxed; e.g. LAN). With this,

$$\sqrt{n}(\hat{\alpha}(m_2) - \alpha) \quad \xrightarrow{P_{n,\alpha,\beta}} \quad \Phi_{\sigma(m_2)}(\cdot)$$

and, if $\beta = 0$ or, equivalently, if $m_*(\beta) = 1$, also

$$\sqrt{n}(\hat{\alpha}(m_1) - \alpha) \quad \xrightarrow{P_{n,\alpha,0}} \quad \Phi_{\sigma(m_1)}(\cdot),$$

where $\sigma(m_1) \leq \sigma(m_2)$. Write $F_{n,\alpha,\beta}(t)$ for the c.d.f. of $\sqrt{n}(\hat{\alpha}(\hat{m}_*) - \alpha)$. Since $\hat{m}_*$ is consistent, we have, for each pair α, β, that

$$F_{n,\alpha,\beta}(t) - \Phi_{\sigma(m_*(\beta))}(t) \quad \xrightarrow{n\to\infty} \quad 0.$$

Theorem 1

If $\sigma(m_1) < \sigma(m_2)$, then

$$\sup_{|\beta|<1/\sqrt{n}} \left| F_{n,\alpha,\beta}(t) - \Phi_{\sigma(m_*(\beta))}(t) \right| \quad \xrightarrow{n\to\infty} \quad \delta \quad > \quad 0$$

for each α.

Estimating the c.d.f. of the post-model-selection estimator

It is easy to construct a consistent estimator $\hat{F}_n(t)$ for $F_{n,\alpha,\beta}(t)$, i.e., an estimator satisfying

$$P_{n,\alpha,\beta}\left(\left|\hat{F}_n(t) - F_{n,\alpha,\beta}(t)\right| > \epsilon\right) \quad \xrightarrow{n\to\infty} \quad 0$$

for each $\epsilon > 0$ and each pair (α, β). But . . .

Theorem 2

If $\sigma(m_1) < \sigma(m_2)$, and if $\hat{F}_n(t)$ is a consistent estimator for $F_{n,\alpha,\beta}(t)$, then

$$\sup_{|\beta|<1/\sqrt{n}} P_{n,\alpha,\beta}\left(\left|\hat{F}_n(t) - F_{n,\alpha,\beta}(t)\right| > \epsilon_\circ\right) \quad \xrightarrow{n\to\infty} \quad 1$$

for some $\epsilon_\circ > 0$ and for each α.

Estimating the c.d.f. of the post-model-selection estimator

It is easy to construct a consistent estimator $\hat{F}_n(t)$ for $F_{n,\alpha,\beta}(t)$, i.e., an estimator satisfying

$$P_{n,\alpha,\beta}\left(\left|\hat{F}_n(t) - F_{n,\alpha,\beta}(t)\right| > \epsilon\right) \quad \overset{n\to\infty}{\longrightarrow} \quad 0$$

for each $\epsilon > 0$ and each pair (α, β). But ...

Theorem 2

If $\sigma(m_1) < \sigma(m_2)$, and if $\hat{F}_n(t)$ is a consistent estimator for $F_{n,\alpha,\beta}(t)$, then

$$\sup_{|\beta|<1/\sqrt{n}} P_{n,\alpha,\beta}\left(\left|\hat{F}_n(t) - F_{n,\alpha,\beta}(t)\right| > \epsilon_\circ\right) \quad \overset{n\to\infty}{\longrightarrow} \quad 1$$

for some $\epsilon_\circ > 0$ and for each α.

Remarks & Extensions

- Theorem 2 also holds for randomized estimators (e.g., subsampling or bootstrap).
- Theorem 2 continues to hold for arbitrary (not necessarily consistent) estimators, if the limit 1 is replaced by $1/2$.
- All the results discussed so far can be extended to also cover conservative model selectors like AIC, Cp, etc.
- All the results discussed so far continue to hold in multivariate linear models with a fixed number of explanatory variables.
- All the result discussed so far continue to hold if the c.d.f. of $\hat{\alpha}(\hat{m}_*)$ (scaled & centered) is replaced by its (scaled) risk. For consistent model selectors, the worst-case risk is unbounded!
- Results parallel to those discussed so far also hold for other estimators and other estimation targets like, e.g., the c.d.f. of the LASSO and related estimators, or the risk of the James-Stein estimator.

So what?

- If $d/n \approx 0$, the potential benefits of model selection for, e.g., prediction, are negligible.
- In many challenging contemporary problems, one faces many potentially important explanatory variables or factors, together with a comparatively small sample size.

A not so simple linear model

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y = \sum_{j=1}^{\infty} x_j \theta_j + u \qquad (2)$$

with $x_1 = 1$. Assume that the x_j's with $j > 1$ and u are jointly nondegenerate Gaussian; that u has mean zero and is uncorrelated with the x_j's; and that the sum converges in L_2.

A not so simple linear model

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y = \sum_{j=1}^{\infty} x_j \theta_j + u \tag{2}$$

with $x_1 = 1$. Assume that the x_j's with $j > 1$ and u are jointly nondegenerate Gaussian; that u has mean zero and is uncorrelated with the x_j's; and that the sum converges in L_2.

No further regularity conditions are imposed.

A not so simple linear model

Consider a response y that is related to a (possibly infinite) number of explanatory variables x_j, $j \geq 1$, by

$$y = \sum_{j=1}^{\infty} x_j \theta_j + u \qquad (2)$$

with $x_1 = 1$. Assume that the x_j's with $j > 1$ and u are jointly nondegenerate Gaussian; that u has mean zero and is uncorrelated with the x_j's; and that the sum converges in L_2.

No further regularity conditions are imposed.

In view of this, parameter estimation is infeasible. We focus on prediction instead.

The candidate models and predictors

Consider a sample (X, Y) of n independent realizations of (x, y) as in (2), and a collection $\mathcal{M}$ of candidate models. Each model $m \in \mathcal{M}$ is assumed to contain the intercept and to satisfy $|m| < n - 1$. Each model m is fit to the data by least-squares. Given a new set of explanatory variables $x^{(f)}$, the corresponding response $y^{(f)}$ is predicted by

$$\hat{y}^{(f)}(m) \quad = \quad \sum_{j=1}^{\infty} x_j^{(f)} \tilde{\theta}_j(m)$$

when using model m. Here, $x^{(f)}, y^{(f)}$ is another independent realization from (2), and $\tilde{\theta}(m)$ is the restricted least-squares estimator corresponding to m.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

(i) For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m) \right)^2 \middle\| X, Y \right].$$

(ii) For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \,\Big\|\, X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

Two goals

(i) Select a 'good' model from $\mathcal{M}$ for prediction out-of-sample, and (ii) conduct predictive inference based on the selected model, both conditional on the training sample.

Two Quantities of Interest

(i) For $m \in \mathcal{M}$, let $\rho^2(m)$ denote the conditional mean-squared error of the predictor $\hat{y}^{(f)}(m)$ given the training sample, i.e.,

$$\rho^2(m) \quad = \quad E\left[\left(y^{(f)} - \hat{y}^{(f)}(m) \right)^2 \middle\| X, Y \right].$$

(ii) For $m \in \mathcal{M}$, the conditional distribution of the prediction error $\hat{y}^{(f)}(m) - y^{(f)}$ given the training sample is

$$\hat{y}^{(f)}(m) - y^{(f)} \, \Big\| \, X, Y \quad \sim \quad N(\nu(m), \delta^2(m)) \quad \equiv \quad \mathbb{L}(m).$$

Note that $\rho^2(m) = \nu^2(m) + \delta^2(m)$.

Estimation of $\rho^2(m)$ and model selection

We will estimate $\rho^2(m)$ by

$$\hat{\rho}^2(m) \quad = \quad \sigma^2(m)\frac{n}{n+1-|m|},$$

which is closely related to GCV (Craven & Whaba, 1978) and to S_p (Tuckey, 1967).

Write m_* and $\hat{m}$ for the truly best and the empirically best candidate model, i.e.,

$$m_* = \operatorname{argmin}_{\mathcal{M}} \rho^2(m) \quad \text{and} \quad \hat{m} = \operatorname{argmin}_{\mathcal{M}} \hat{\rho}^2(m).$$

Performance of the model selector

Remember: The truly best model m_* minimizes $\rho^2(m)$ over $m \in \mathcal{M}$; the selected model $\hat{m}$ minimizes $\hat{\rho}^2(m)$ instead. Moreover, $\#\mathcal{M}$ is the number of candidate models and $|\mathcal{M}|$ is the number of parameters in the most complex candidate model.

Theorem 3

For each fixed sample size n and uniformly over all data-generating processes as in (2), we have

$$\begin{aligned} P\left(\log\frac{\rho^2(\hat{m})}{\rho^2(m_*)} \;>\; \epsilon\right) \quad &\leq \quad 6\exp\left[\log\#\mathcal{M} - \frac{n-|\mathcal{M}|}{16}\frac{\epsilon^2}{\epsilon+16}\right], \\ P\left(\left|\log\frac{\hat{\rho}^2(\hat{m})}{\rho^2(\hat{m})}\right| \;>\; \epsilon\right) \quad &\leq \quad 6\exp\left[\log\#\mathcal{M} - \frac{n-|\mathcal{M}|}{8}\frac{\epsilon^2}{\epsilon+8}\right], \end{aligned}$$

for each $\epsilon > 0$.

Predictive Inference based on the selected model

Idea: Estimate the conditional distribution of the prediction error, i.e., $\mathbb{L}(m) \equiv N(\nu(m), \delta^2(m))$, by

$$\hat{\mathbb{L}}(m) \quad \equiv \quad N(0, \hat{\delta}^2(m)),$$

where $\hat{\delta}^2(m)$ is defined as $\hat{\rho}^2(m)$ before.

Theorem 4

For each fixed sample size n and uniformly over all data-generating processes as in (2), we have

$$\begin{aligned} P\left(\left|\left| \hat{\mathbb{L}}(\hat{m}) - \mathbb{L}(\hat{m}) \right|\right|_{TV} > \frac{1}{\sqrt{n}} + \epsilon \right) & \\ \leq \quad 7 \exp\left[\log \#\mathcal{M} - \frac{n - |\mathcal{M}|}{2} \frac{\epsilon^2}{\epsilon + 2} \right] & \end{aligned}$$

for each ϵ with $0 < \epsilon < \log(2)$.

Predictive Inference based on the selected model

Idea: Estimate the conditional distribution of the prediction error, i.e., $\mathbb{L}(m) \equiv N(\nu(m), \delta^2(m))$, by

$$\hat{\mathbb{L}}(m) \quad \equiv \quad N(0, \hat{\delta}^2(m)),$$

where $\hat{\delta}^2(m)$ is defined as $\hat{\rho}^2(m)$ before.

Theorem 4

For each fixed sample size n and uniformly over all data-generating processes as in (2), we have

$$\begin{aligned} P\left(\left\| \hat{\mathbb{L}}(\hat{m}) - \mathbb{L}(\hat{m}) \right\|_{TV} > \frac{1}{\sqrt{n}} + \epsilon \right) \\ \leq \quad 7 \exp\left[\log \# \mathcal{M} - \frac{n - |\mathcal{M}|}{2} \, \frac{\epsilon^2}{\epsilon + 2} \right] \end{aligned}$$

for each ϵ with $0 < \epsilon < \log(2)$.

Note

These results rely on Gaussianity in two respects:

1. Tail behavior. For non-Gaussian models, we can here use modern empirical process theory and concentration inequalities. [See Beran (2007) and the references given there.]
2. The following property of the Gaussian: If (w, y) are jointly normal, then $\mathbb{E}[y\|w]$ is linear in w, and $\mathrm{Var}[y\|w]$ is constant in w. In particular, we can write

$$y = \beta' w + u$$

where the error u is uncorrelated with w, has zero mean and constant variance.

Low-dimensional projections of high-dimensional data

For each dimension d, consider a random d-vector Z that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also assume that the elements of Z are independent, have bounded moments of order up to 9, and bounded Lebesgue densities (can be relaxed).

Our results are asymptotic as $d \to \infty$.

Consider two projections of Z of the form $\alpha' Z$ and $\beta' Z$ for unit d-vectors α and β.

Low-dimensional projections of high-dimensional data

For each dimension d, consider a random d-vector Z that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also assume that the elements of Z are independent, have bounded moments of order up to 9, and bounded Lebesgue densities (can be relaxed).

Our results are asymptotic as $d \to \infty$.

Consider two projections of Z of the form $\alpha' Z$ and $\beta' Z$ for unit d-vectors α and β.

We study the conditional mean and with the conditional variance of $\alpha' Z$ given $\beta' Z$.

Low-dimensional projections of high-dimensional data

For each dimension d, consider a random d-vector Z that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also assume that the elements of Z are independent, have bounded moments of order up to 9, and bounded Lebesgue densities (can be relaxed).

Our results are asymptotic as $d \to \infty$.

Consider two projections of Z of the form $\alpha' Z$ and $\beta' Z$ for unit d-vectors α and β.

More precisely, we will study the following two conditions: The vector β is such that ...

(i) for each α, the conditional mean of $\alpha' Z$ given $\beta' Z = x$ is linear in $x \in \mathbb{R}$;

(ii) for each α, the conditional variance of $\alpha' Z$ given $\beta' Z = x$ is constant in $x \in \mathbb{R}$.

On conditions (i) and (ii)

If β is such that both (i) and (ii) hold, and if we set $y = \alpha' Z$ and $x = \beta' Z$, then the simple linear model, namely

$$y = ax + u,$$

with unknown parameter $a \in \mathbb{R}$ given by $a = \beta'\alpha$, and with u haven mean zero and constant variance given x, is ‘correct,’ irrespective of α. The true data-generating process is $y = \alpha' Z$.

On conditions (i) and (ii)

If β is such that both (i) and (ii) hold, and if we set $y = \alpha' Z$ and $x = \beta' Z$, then the simple linear model, namely

$$y \quad = \quad ax + u,$$

with unknown parameter $a \in \mathbb{R}$ given by $a = \beta'\alpha$, and with u haven mean zero and constant variance given x, is 'correct,' irrespective of α. The true data-generating process is $y = \alpha' Z$.

But (i) and (ii) are restrictive: If Z is such that (i) holds for <u>all unit-vectors β</u>, then the law of Z must be spherically symmetric (Eaton, 1986). And if both (i) and (ii) hold for <u>all unit-vectors β</u>, then Z is Gaussian (Bryc 1995).

On conditions (i) and (ii)

If β is such that both (i) and (ii) hold, and if we set $y = \alpha' Z$ and $x = \beta' Z$, then the simple linear model, namely

$$y \quad = \quad ax + u,$$

with unknown parameter $a \in \mathbb{R}$ given by $a = \beta' \alpha$, and with u haven mean zero and constant variance given x, is 'correct,' irrespective of α. The true data-generating process is $y = \alpha' Z$.

Next result (informally):

Both conditions (i) and (ii) are approximately satisfied for most unit-vectors β, namely on a set of β's whose size, as measured by the uniform distribution on the unit sphere in $\mathbb{R}^d$, goes to one as $d \to \infty$.

In that sense, most simple linear submodels are approximately 'correct', provided only that d is large.

Most linear submodels are approximately ‘correct’

The vector β satisfies conditions (i) and (ii) if and only if

$$\begin{aligned} \left\| \mathbb{E}[Z \| \beta' Z = x] \right\|^2 - x^2 &= 0 \quad \text{and} \\ \left\| \mathbb{E}[ZZ' \| \beta' Z = x] - (I_d + (x^2 - 1)\beta\beta' \right\|^2 &= 0 \end{aligned}$$

hold for each $x \in \mathbb{R}$.

Theorem 5

There are sets $B_d \subseteq \mathbb{R}^d$ satisfying $\upsilon(B_d) \xrightarrow{d \to \infty} 1$ so that, for each $\epsilon > 0$, both

$$\sup_{\beta \in B_d} \mathbb{P}\left(\left\| \mathbb{E}[Z \| \beta' Z] \right\|^2 - (\beta' Z)^2 > \epsilon \right) \quad \text{and}$$

$$\sup_{\beta \in B_d} \mathbb{P}\left(\left\| \mathbb{E}[ZZ' \| \beta' Z] - (I_d + ((\beta' Z)^2 - 1)\beta\beta' \right\|^2 > \epsilon \right)$$

converge to zero as $d \to \infty$.

Most linear submodels are approximately ‘correct’

The vector β satisfies conditions (i) and (ii) if and only if

$$\begin{aligned} \left\|\mathbb{E}[Z\|\beta' Z = x]\right\|^2 - x^2 &= 0 \quad \text{and} \\ \left\|\mathbb{E}[ZZ'\|\beta' Z = x] - (I_d + (x^2-1)\beta\beta'\right\|^2 &= 0 \end{aligned}$$

hold for each $x \in \mathbb{R}$.

Theorem 5

There are sets $B_d \subseteq \mathbb{R}^d$ satisfying $\upsilon(B_d) \xrightarrow{d\to\infty} 1$ so that, for each $\epsilon > 0$, both

$$\sup_{\beta \in B_d} \mathbb{P}\left(\left\|\mathbb{E}[Z\|\beta' Z]\right\|^2 - (\beta' Z)^2 \; > \; \epsilon \right) \quad \text{and}$$

$$\sup_{\beta \in B_d} \mathbb{P}\left(\left\|\mathbb{E}[ZZ'\|\beta' Z] - (I_d + ((\beta' Z)^2 - 1)\beta\beta'\right\|^2 \; > \; \epsilon \right)$$

converge to zero as $d \to \infty$.

Outlook

- Extension to several projections $\beta_1, \ldots, \beta_k$ (i.e., simple models with several explanatory variables); joint with Lukas Steinberger.
- Approximately valid prediction and inference when fitting simple linear submodels to complex data-generating processes; joint with Lukas Steinberger.
- Finite-d bounds on the error probabilities in Theorem 1, with applications to model selection and regularization; joint with Ivana Milovic.

References

- H. Leeb (2008): Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process, Bernoulli:**14**, 661–690.
- H. Leeb (2009): Conditional predictive inference after model selection, Ann. Statist.:**37**,2838–2876.
- H. Leeb (2012): On the conditional distributions of low-dimensional projections from high-dimensional data. Manuscript.
- H. Leeb and B.M. Pötscher (2003): The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations, Econometric Theory:**19**, 100–142.
- H. Leeb and B.M. Pötscher (2006): Can one estimate the conditional distribution of post-model-selection estimators? Ann. Statist.:**34**, 2554–2591.
- H. Leeb and B.M. Pötscher (2008): Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory:**24**, 338–376.
- B.M. Pötscher (1991): Effects of model selection on inference, Econometric Theory:**7**, 163–185.