Teoria e Pratica dei Modelli Statistici

Homework 4

Edoardo Ricca

1. *In a sequence of nested models, increasing the number of model parameters always increases the fit (i.e., it reduces the sum of squared errors). Nonetheless, a model with more parameters is not necessarily preferable. Explain.*

There are multiple reasons behind the fact that more parameters are not necessarily preferable.

Let's consider a vector of n covariates $\mathbf{x}$ and its vector of estimated parameters $\hat{\boldsymbol{\beta}}$. Let's now partition $\mathbf{x}$ into different subsets $\mathbf{x}_k$ containing $k < n$ covariates and from these let's estimate their parameter vector $\tilde{\boldsymbol{\beta}}$. It can be shown that for each subset of $k < n$, when estimating the parameter β_j from the two datasets the following is true:

$$\mathbb{V}ar[\hat{\beta}_j] > \mathbb{V}ar[\tilde{\beta}_j]$$

which means that if the model contains irrelevant variables, even though it would diminish the sum of squared errors, the variance of the estimated parameters increases and, consequently, their precision decreases.

Now, considering the independent variable $\mathbf{y}$ with $\mathbb{E}[y_i] = \mu_i$ and an estimator $\hat{\mathbf{y}}_k = \mathbf{X}_k\tilde{\boldsymbol{\beta}}_k$, the following two formulas show another reason why increasing the number of parameters can be inefficient:

- **Sum of the mean squared errors**: measuring the sum of the expected square distance between the estimate $\hat{y_{ik}}$ and the true value μ_i, the formula can be decomposed into:

$$SMSE = |k|\sigma^2 + \sum_{i=1}^{n}(\mu_{ik} - \mu_i)^2$$

- **Sum of squared prediction error**: measuring the sum of the expected square distance between a future observation y_{n+1} and the estimate $\hat{y_{ik}}$, its formula is equal to:

$$SPSE = n\sigma^2 + |k|\sigma^2 + \sum_{i=1}^{n}(\mu_{ik} - \mu_i)^2 = n\sigma^2 + SMSE$$

In both cases, our objective is to minimize the results of the equation (whether it's the distance between the prediction and the mean or the future value and the prediction). From the formula we can see, being the variance an always positive term, that both functions are monotonically strictly increasing with respect to k. Therefore, the higher the number of parameters the higher both expected distances will be. In addition, it can be shown that:

$$\sum_{i=1}^{n}\mathbb{V}ar[\hat{y}_{ik}] = |k|\sigma^2$$

meaning that the sum of the variances of $\hat{y}_{ik}$ increases as the number of parameters increases; leading to more volatile predictions.

2.- 3. *Suppose the true model is a linear model with design matrix* $\mathbf{X} = [\mathbf{X}_1\mathbf{X}_2]$; *however, you fit a model with only* $\mathbf{X}_1$. *Write the expectation of the estimator and discuss the bias: in which cases does the bias vanish?.Following the previous point, write the formula for the bias in the special case where the full model has two covariates, say* z *and* w, *whereas the fitted model has* z *but* w *is omitted. In the above notation,* $\boldsymbol{X}_1$ *has a column of ones and a column with the values of* z, *while* $\boldsymbol{X}_2$ *has a single column with the values of* w. *[in classroom we considered an example where y=income, x=education, w=ses (socio-economic status of the parents)]*

A variable $\mathbf{y}$ with a true model as indicated would mean that:

$$\mathbf{y} = \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}$$

while the model we fit is:

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X}_1 + \boldsymbol{\varepsilon}$$

Using the OLS estimator for $\boldsymbol{\beta}$ of our restricted model we obtain:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

whose expected value is equal to:

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}] &= \mathbb{E}\left[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y}\right] \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbb{E}\left[\mathbf{y}\right] \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbb{E}\left[\boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}\right] \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2) \\
&= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2
\end{aligned}$$

which, in the bivariate case, when fully developing in vector notation, giving the name Z to the variable inside $\mathbf{X}_1$, and W to the variable inside $\mathbf{X}_2$ appears as:

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \begin{pmatrix} \beta_0 + \left(\bar{w} - \bar{z}\frac{\mathbb{C}ov(Z,W)}{\mathbb{V}ar[Z]}\right)\beta_2 \\ \beta_1 + \frac{\mathbb{C}ov(Z,W)}{\mathbb{V}ar[Z]}\beta_2 \end{pmatrix}$$

where the two different biases revolve around the covariance between the omitted variable and the variable fitted, the variance of the variable fitted and the parameter of the omitted variable.

From this, we can conclude that the expected value of both the slope and the intercept of the model is shifted by the parameter of the omitted variable and the relationship between the fitted and omitted variables. In fact, if there is no linear relationship between the Z and W ($\mathbb{C}ov(Z,W) = 0$ in the bivariate case and $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ in the multivariate case) the bias term vanishes (in both cases).

4. *Suppose you omit a relevant covariate, so that the estimator $\tilde{\beta}_j$ of the j-th regression coefficient is biased. Explain why in terms of $MSE(\tilde{\beta}_j)$ the biased estimator $\tilde{\beta}_j$ may be better than the unbiased estimator $\hat{\beta}_j$ based on the full model*

Using the decomposition formula of the MSE:

$$\begin{aligned} MSE(\tilde{\beta}_j) &= \mathbb{E}[(\tilde{\beta}_j - \beta_j)^2] = \\ &= \mathbb{E}[(\tilde{\beta}_j - \mathbb{E}[\beta_j] + \mathbb{E}[\beta_j] - \beta_j)^2] = \\ &= \mathbb{E}[(\tilde{\beta}_j - \mathbb{E}[\tilde{\beta}_j)^2] + \mathbb{E}[(\tilde{\beta}_j - \beta_j)]^2 + 2\mathbb{E}\left[\tilde{\beta}_j - \mathbb{E}[\tilde{\beta}_j]\right]\mathbb{E}[\tilde{\beta}_j - \beta_j] \\ &= \mathbb{E}[(\tilde{\beta}_j - \mathbb{E}[\tilde{\beta}_j)^2] + \mathbb{E}[(\tilde{\beta}_j - \beta_j)]^2 \\ &= \mathbb{V}ar[\tilde{\beta}_j] + \mathbb{B}[\tilde{\beta}_j]^2 \end{aligned}$$

indicating that the MSE can be seen as the sum of the variance and the bias squared. Therefore, an unbiased estimator will cancel the bias term diminishing the MSE. However, as mentioned previously, the omittance of a relevant covariate can reduce the variance of beta. There can exist, therefore, estimators that, while being biased due to the omittance of a covariate, have a lower MSE than other unbiased estimators thanks to their reduced variance.

5. *Write the definition of the sum of prediction squared errors (SPSE). Then write SPSE as the sum of three components and comment.*

The definition and formula of the SPSE were already presented in answer 1. Hereby, the three components of the equation:

1. **Irreducible Prediction error**: the term $n\sigma^2$ is generated by the variance of the errors and the number of observations. It is therefore impossible to reduce this term voluntarily hence is called irreducible.
2. **Sum of the variances**: as explained before the term $|k|\sigma^2$ represents the sum of the variances of $\hat{y_{ik}}$ where k is the cardinality of the subset induced by the partition and σ^2 is the variance of the errors. While, again, the variance of the errors can't be reduced, the term k can be reduced by decreasing the number of parameters in the model. As a consequence, the whole term will be reduced.
3. **Squared bias**: the last term $\sum_{i=1}^{n}(\mu_{ik} - \mu_i)^2$ can be interpreted as the bias since it measures the squared distance between the estimator μ_{ik} and the real value μ_i. This term can be decreased by increasing the number of parameters.

This inverse relationship with respect to the number of parameters between the second and third term of the SPSE is what generates the bias-variance trade-off.

6. *Suppose you take the sum of squared residuals as an estimator of SPSE: write the bias. Do you underestimate or overestimate SPSE? Is the bias more severe for a simple model or for a complex model?*

The SSR as an estimator for the SPSE is a biased estimator since:

$$\mathbb{E}\left[\sum_{i=1}^{n}(y_i - \hat{y}_{ik})^2\right] = SPSE - 2|k|\sigma^2$$

The estimator has consequently a bias of $-2|k|\sigma^2$ that, given the positive nature of both k and σ^2, leads to an underestimation of the true SPSE.
Given that k represents the cardinality of the subset chosen,i.e the number of covariates/parameters, the higher the number of covariates the larger the bias will be. In other words, the more complex the model is the larger the bias.

7. *Write the expressions of the AIC and BIC indexes and discuss similarities and differences. Then apply those indexes to choose between the following two models: model A has M=5 parameters and maximized log-likelihood l=300, whereas model B has M=9 parameters and maximized log-likelihood l=290. First suppose the dataset has size n=100. Then repeat assuming the dataset has size n=200. Comment.*

1. **AIC**:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2_{ML}) + 2(|k| + 1)$$

   assuming normally distributed errors:

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2_{ML}) + 2(|k| + 1)$$

2. **BIC**:

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2_{ML}) + \log(n)2(|k| + 1)$$

   again, assuming gaussian errors:

$$\text{BIC} = n \cdot \log(\hat{\sigma}^2_{ML}) + \log(n)\,2(|k| + 1)$$

The two indexes are very similar. They compute the (log)likelihood as a function of the estimate $\hat{\boldsymbol{\beta}}_k$ and σ^2_{ML}. The higher the likelihood, the higher the probability that the estimates are the parameters that generated $\mathbf{y}$ and the lower both indices will be. Therefore, the objective when using these two indexes is to minimize both. The other term of the addition represents the punishment for the number of parameters. The more complex a model is, the higher the index will be (keeping the likelihood constant). The only difference between the two is in this punishment term. The **BIC**, due to its $log(n)$ term, penalizes complex models more than the **AIC** does.

	AIC	BIC
Model A (n = 100)	-588	-544.74
Model A (n = 200)	-588	-536.42
Model B (n = 100)	-560	-487.90
Model B (n = 200)	-560	-474.03

Using the indexes on the models provided we obtain the following table above.

Observing the AIC, the higher likelihood and the lower number of parameters in model A are rewarded by the index while the number of observations is irrelevant. We can conclude that, according to the AIC, the best model is model A. Looking at the BIC, again, both the higher likelihood and the lower number of parameters indicate that model A is the correct one. In summary, we can conclude that both indexes indicate model A as the best choice.

Two things should be noted: first, the BIC, as expected, penalizes more the number of parameters resulting in a value always higher than the AIC. Second, the number of observations alters the BIC and since the term $\log(n)$ increases the index, the lower the number of observations the smaller the punishing term is.

8. *Explain why in automatic selection procedures (backward, forward etc.) it is better to use fit indexes like AIC and BIC rather than statistical tests.*

The main reasons behind the usage of the global model choice criterion instead of statistical tests are:

- comparability between the different models selected by the automatic selection procedures which becomes impossible with statistical tests.
- approaches such as the Efroymson may eliminate highly collinear covariates even if important while they can be included in AIC-based models.
- tests are not exact due to their construction.

Stata

Replicate "Example 3.17 Correlated Covariates" of the textbook of Fahrmeir, Kneib, Lang Marx. Note: choose a value for the seed of the pseudo-random generator (e.g. in Stata type set seed), then change the seed until you get results similar (not necessarily equal) to those in tables 3.3 and 3.4. Hint: to see how to randomly generate a dataset, look in Moodle at the files 'Testing linear effects.do' and 'overfitting.do'.

Following the instructions on example 3.17 we set the seed at 2000 and obtain the following results:

Variable	Coefficient	s.e	t-value	p-value	LB 95%	UB 95%
x_1	.1245788	.0844309	1.48	**0.142**	-.0422857	.2914434
x_2	.0726307	.0550593	1.32	**0.189**	-.0361856	.181447
x_3	.1961866	.0561838	3.49	0.001	.085148	.3072252
intercept	-1.016586	.0498118	-20.41	0.000	-1.115032	-.918140

Table 1: Table for the regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Variable	Coefficient	s.e	t-value	p-value	LB 95%	UB 95%
x_1	.2052714	.0583417	3.52	0.001	.0899746	.3205682
x_3	.1972728	.056319	3.50	0.001	.0859733	.3085724
intercept	-.9846838	.0436564	-22.56	0.000	-1.070959	-.8984084

Table 2: Table for the regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_3$

Application: model selection

Use the data con test scores in California (`caschool.dta`); see the description in `californiatestscores.pdf`. Discard the following variables: `observat dist_cod county district read_scr math_scr` (don't use them in the models). The response variable is `testscr` (the average test score for students in the district). All other variables are potential covariates. Then

1. Perform model selection using the 'vselect' command of Stata with forward, backward, and best (the option to search across all models with the 'leaps and bounds' algorithm). First use AIC, then use BIC. Compare the selected models.

After dropping the requested variables, we create a dummy variable for the grade span of district. Then we perform all the requested model selection algorithms leading to the following table, indicating the global choice criteria of the models indicated by the procedures:

Model (procedure)	df	AIC	BIC
M1 (backward AIC)	8	2978.593	3010.915
M2 (forward AIC)	8	2978.593	3010.915
M3 (backward BIC)	6	2980.378	3004.62
M4 (forward BIC)	6	2980.378	3004.62

In this case, forward selection and backward elimination lead to the same choice of model for criterion. The difference between the AIC models and the BIC ones is that the latter discarded the variables `calw_pct` and `expn_stu` leading, as a consequence, to higher BICs.
In this situation, with the information provided until here, I would probably choose the model given by the BIC-based algorithms.

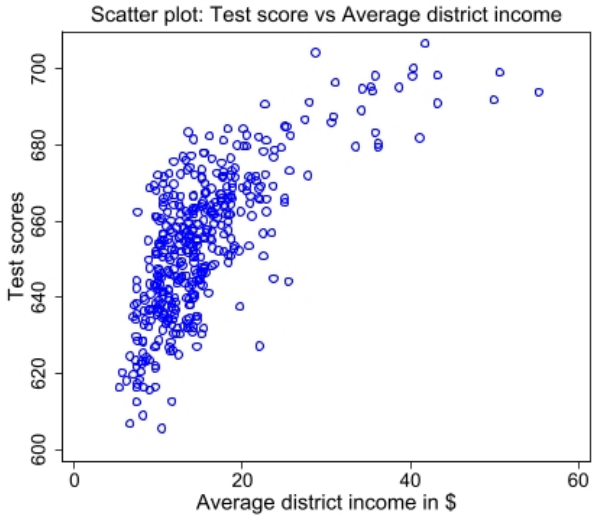
Using the leaps and bounds algorithm, instead, we obtain the following:

#Predictors	R^2 Adj.	C	AIC	AICC	BIC
1	.7541781	120.6655	3080.265	3080.323	3088.346
2	.7798245	65.52566	3034.983	3035.079	3047.104
3	.8010543	20.24966	2993.389	2993.534	3009.55
4	.8053508	11.8982	2985.209	2985.412	3005.41
5	.8080282	7.092237	2980.378	2980.65	**3004.62**
6	.8091415	5.68829	2978.92	2979.27	3007.202
7	**.8097352**	5.413953	**2978.593**	**2979.032**	3010.915
8	.8094918	6.942667	2980.109	2980.647	3016.472
9	.8092851	8.390232	2981.542	2982.189	3021.944
10	.8090008	10.00153	2983.142	2983.908	3027.584
11	.8085334	12	2985.14	2986.037	3033.623

This time, the procedure (while still indicating the model with 5 parameters suggested by the BIC-based procedures) suggests that the best model is the one indicated by the AIC algorithms; in light of the additional higher adjusted R^2 and the lower complexity parameter, it seems that the model with 7 parameters to be the best choice.
However, looking at the table, we can see that the model with 6 parameters can be seen as a compromise between the two decreasing the BIC at the cost of a slightly higher AIC and C. In my opinion, this model could also be considered an optimal choice. In the end, by choosing the model suggested by the leap and bound algorithm we obtain:

$$\begin{aligned}\widehat{test\_score} =& 658.4919 - 0.3619524 \cdot meal\_pct - 0.2166673 \cdot el\_pct + 0.6181897 \cdot avginc \\ & - 3.477994 \cdot gr\_span\_dum + 0.0015359 \cdot expn\_stu + 12.83214 \cdot comp\_stu - 0.085822 \cdot calw\_pct\end{aligned}$$

Although this model seems a good fit, I noticed a particular relationship between average income and test scores.



The relationship seems to be non-linear. As a consequence, I decided to use alternative covariates

for the average district income. At first glance, it would seem that a log-transformation would be a good choice but, after performing model selection and cross-validation, I concluded that this is true only when modelling test scores with average district income as its only covariate.

Consequently, I decided to use orthogonal polynomials of `avginc` until the third degree and reiterate the model selection procedure. The results were the following:

Model (procedure)	df	AIC	BIC
M5 (backward AIC)	8	2973.608	3009.97
M6 (forward AIC)	8	2973.608	3009.97
M7 (backward BIC)	6	2978.413	3002.655
M8 (forward BIC)	6	2978.413	3002.655

Again, the procedures choose the same model with respect to their criterion. This time both AIC and BIC of the selected models are lower compared to the previous ones. Using the best procedure instead we obtain the table below

Predictors	R^2 Adj.	C	AIC	AICC	BIC
1	.7541781	129.1423	3080.265	3080.323	3088.346
2	.7798245	73.09996	3034.983	3035.079	3047.104
3	.8010543	27.07723	2993.389	2993.534	3009.55
4	.8053508	18.56226	2985.209	2985.412	3005.41
5	.8089243	11.68061	2978.413	2978.685	**3002.655**
6	.8109848	8.152019	2974.844	2975.194	3003.125
7	.8117878	7.394209	2974.037	2974.476	3006.359
8	.8124181	7.021252	**2973.608**	**2974.146**	3009.97
9	**.8126641**	7.490851	2974.033	2974.68	3014.436
10	.8124615	8.936634	2975.462	2976.228	3019.905
11	.8122899	10.31309	2976.818	2977.714	3025.301
12	.8119693	12.00963	2978.504	2979.541	3031.027
13	.8115106	14	2980.494	2981.682	3037.058

This time, it becomes harder to decide which model is the best since basically any model with a number of predictors between 9 and 5 has at least one criterion that is similar to or better than the others. Therefore I decided to test them with cross-validation to see which one has the smallest mean squared error of prediction.

2. Take your preferred model and compute the mean squared error of prediction using cross-validation using 5 partitions, 10 partitions, 20 partitions and n partitions (i.e. leave-one-out). For example, in Stata you can use `cv_kfold` and `cv_regress`. Discuss you findings.

Using the functions aforementioned I initially implemented them with a number of repetitions of 100.

However, noticing that the models gave a very similar mean squared error for each partition and the small sample generated enough randomness to change the conclusions at each iteration of the test I decided to increase the number of repetitions to 1000. The results are in the table below:

# Parameters	k(5)	k(10)	k(20)	Leave-one-out
9	8.36778	**8.34938**	**8.34537**	**8.3401**
8	8.37005	8.35677	8.35074	8.3452
7	**8.36199**	8.35082	8.34565	8.3411
6	8.36845	8.35774	8.35443	8.3492
5	8.40040	8.39264	8.38822	8.3850

I decided to test also the model with 6 parameters given the fact that, in the regression with 7, the p-values of expenses per student and computers per student were slightly high. Suspecting collinearity, I tested for correlation and the two seem to have almost 30% of correlation. In the end, however, the model with 7 parameters gave better predictions in the cross-validation procedure therefore (considering that the p-values were not too extreme) I ruled out model 6.

The two models that performed best in the cross-validation for each k-fold were model 7 and model 9. Given the small differences in the cross-validation (each iteration of the test led to a different conclusion due to randomness), all the other global choice criteria and the p-values in some parameters of model 9, I concluded that the model:

$$\begin{aligned}\widehat{test\_score} =& 670.1228 - 0.3967566 \cdot meal\_pct - 0.2143334 \cdot el\_pct + 4.396992 \cdot avginc \\ & - 3.526186 \cdot gr\_span\_dum - 1.086443 \cdot avginc^3 + 0.0012145 \cdot expn\_stu + 12.59799 \cdot comp\_stu\end{aligned}$$

is the best model available.

I also tested the models with linear covariate for the district average income and, as suspected, they performed worse than the non-linear models. In addition, the model I indicated as optimal compared to the one highlighted by the best procedure seems to have a smaller mean squared error.