Introduction to model selection and model averaging

Svetlana Bryzgalova

sabryzgalova@gmail.com

September 22, 2013

Model risk

- Every time you do an empirical project, you have to pick a model specification to estimate
- (recall, even Gauss-Markov theorem requires your model to be a correct one!)
- The best case scenario: your choice is driven by theory
- Quite often, it is not an option
- Estimating different models often yields different results
- What to do then?

How to pick a model specification? How to draw conclusions from several models at the same time? This is the topic of today's session.

There are 2 types of selection

- testing 2 nested model specifications against each other
- choosing the best among a range various models

Depending on your goal, you will need different tools

- statistical tests
- maximizing a certain criterion

The main requirement for any model-selection procedure is its **consistency**: ability to choose the true data-generating process with probability approaching 1 as the sample size goes to infinity.

All good tests and criteria have to satisfy this requirement.

R-squared and such do not. Never maximize R-squared.

Nested models

When 1 model is a particular case of another, you can test this as a restriction on parameters (linear or nonlinear)

- Keeping regressors or not: t-test, F-test
- In most cases Likelihood Ratio works:

$$LR = -2log\frac{Likelihood(model\ 1)}{Likelihood(model\ 2)}$$

This covers restrictions on included regressors, comparing pooled regression with panel data models, estimating time seties model (ARMA, GARCH, etc)

LASSO: least absolute shrinkage and selection operator

- What if you could choose variables and estimate model parameters at the same time?
- LASSO is designed to do both simultaneously
- Penalized least squares:

$$min \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x_j\beta_j)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

- Penalty discourages the use of too many parameters, having weak effect on y
- λ is a tuning parameter, that needs to be specifically chosen for this procedure to work well
- other penalties work as well (a whole zoo of them!) for all sorts of models and data features
- many of them, like adaptive LASSO, have consistency
- perform especially well when there are many factors: Fan and Li (2001)

One of many

- Information criteria are designed to evaluate the probability that the data comes from a particular model.
- They work for two non-nested models
- They are also good for picking the best out of many
- Choosing the best ARMA-GARCH or comparing several regressions
- Main idea: pick a specification with high likelihood and few parameters that drive it

Two most widespread information criteria:

- Akaike (AIC): for choosing the order of ARMA

$$AIC = 2k - 2log(Likelihood)$$

- Schwarz (Bayesian Information Criterion, BIC): for choosing the rest

$$BIC = log(T)k - 2log(Likelihood)$$

where k is the number of parameters, T is the number of observations

Selecting regressors

For the regression setup, AIC is not consistent, but BIC is. It selects a parsimonious, true DGP with high probability.
Why does it happen? BIC directly approximates the probability the data comes from a model. Example:

$$BIC \approx Prob\left(y = x\beta + \epsilon \,|Data\right)$$

We do not fix a particular value of β here, we evaluate the probability of coming from this structure in general, by averaging the likelihood over all possible parameter values.

BIC is an estimate of the **integrated likelihood**. Integrated over a whole range of β's.

Contrast it with the usual maximum likelihood, fixed at the optimal parameter values that maximize it.

Distribution-free measures

- Similar to AIC and BIC, there exist other criteria that are designed for all sorts of models.
- When the model is formulated by moment conditions, likelihood is not available.
- However, there exist analogues to AIC and BIC: Andrews (1999)

Sometimes it's not easy to choose between models, as they are so close, but also so different.

What to do? Take the best of both worlds!

Hoeting et al(1999): *Implementing Bayesian Model Averaging*

Box: "All models are wrong, but some are useful"

- No linear regression is the true data generating process, but some could be closer than others
- Study the weighted average from different models, for example

$$Prob(\beta = \beta_0 \,|Data) =$$
$$= Prob(\beta = \beta_0 \,|Data, Model\ 1)Prob(Model\ 1; |Data)+$$
$$+Prob(\beta = \beta_0 \,|Data, Model\ 2)Prob(Model\ 2; |Data)$$

We already have nearly everything to do it:

- $Prob(\beta = \beta_0 \,|Data, Model\ 1)$ comes from the usual Gauss-Markov theorem, etc
- $Prob(Model\ 1; |Data)$ is approximated by BIC
- some other weights can be used as well, e.g. as in Hansen (2007)

Sala-i-Martin, Doppelhofer, Miller(2004): *Determinants of long-term economic growth*

- There exist many variables that seem to explain a significant propertion of economic growth in different countries
- Are there any common factors?
- A comprehensive study of data on 88 countries and all sorts of factors (67) from different papers
- Looking for something robust **across all the specifications**

Q: Which factors would you include?

Rank	Variable	Description and Source
	Average Growth Rate of GDP per capita 1960-96	Growth of GDP per capita at Purchasing Power Parities between 1960 and 1996. From Heston, Summers and Aten (2001).
1	East Asian Dummy	Dummy for East Asian countries.
2	Primary Schooling in 1960	Enrollment rate in primary education in 1960. Barro and Lee (1993).
3	Investment Price	Average investment price level between 1960 and 1964 on purchasing power parity basis. From Heston, Summers and Aten (2001).
4	GDP in 1960 (log)	Logarithm of GDP per capita in 1960. From Heston, Summers and Aten (2001).
5	Fraction of Tropical Area	Proportion of country's land area within geographical tropics. From Gallup, Mellinger and Sachs (2001).
6	Population Density Coastal in 1960s	Coastal (within 100km of coastline) population per coastal area in 1965. From Gallup, Mellinger and Sachs (2001).
7	Malaria Prevalence in 1960s	Index of malaria prevalence in 1966. From Gallup, Mellinger and Sachs (2001).
8	Life Expectancy in 1960	Life Expectancy in 1960. Barro and Lee (1993).
9	Fraction Confucius	Fraction of population Confucian. Barro (1999).
10	African Dummy	Dummy for sub-saharan African countries.
11	Latin American Dummy	Dummy for Latin American countries.
12	Fraction GDP in Mining	Fraction of GDP in Mining. From Hall and Jones (1999).
13	Spanish Colony	Dummy variable for fomer Spanish colonies. Barro (1999).
14	Years Open 1950-94	Number of years economy has been open between 1950 and 1994. From Sachs and Warner (1995).
15	Fraction Muslim	Fraction of population Muslim in 1960. Barro (1999).
16	Fraction Buddhist	Fraction of population Buddhist in 1960. Barro (1999).
17	Ethnolinguistic Fractionalization	Average of five different indices of ethnolinguistic fractionalization which is the probability of two random people in a country not speaking the same language. From Easterly and Levine (1997).
18	Gov. Consumption Share 1960s	Share of expenditures on government consumption to GDP in 1961. Barro and Lee (1993).
19	Population Density 1960	Population per area in 1960. Barro and Lee (1993).
20	Real Exchange Rate Distortions	Real Exchange Rate Distortions. Levine and Renelt (1992).
21	Fraction Speaking Foreign Language	Fraction of population speaking foreign language. Hall and Jones (1999).
22	Openess measure 1965-74	Ratio of exports plus imports to GDP, averaged over 1965 to 1974. This variable was provided by Robert Barro.
23	Political Rights	Political rights index. From Barro (1999).
24	Government Share of GDP in 1960s	Average share government spending to GDP between 1960-64. From Heston, Summers and Aten (2001).
25	Higher Education 1960	Enrollment rates in higher education. Barro and Lee (1993).
26	Fraction Population In Tropics	Proportion of country's population living in geographical tropics. From Gallup, Mellinger and Sachs (2001).
27	Primary Exports 1970	Fraction of primary exports in total exports in 1970. From Sachs and Warner (1997).
28	Public Investment Share	Average share of expenditures on public investment as fraction of GDP between 1960 and 1965. Barro and Lee (1993).
29	Fraction Protestants	Fraction of population Protestant in 1960. Barro (1999).
30	Fraction Hindus	Fraction of the population Hindu in 1960. Barro (1999).

Source: Sala-i-Martin, Doppelhofer, Miller(2004)

18 factors robust across specifications

Rank	Variable	Posterior Inclusion Probabiliy	Posterior Mean Conditional on Inclusion	Posterior s.d. Conditional on Inclusion	Sign Certainty Probability	Fraction of Regressions with |tstat|>2
		(1)	(2)	(3)	(4)	(5)
1	East Asian	0.823	0.021805	0.006118	0.999	0.99
2	Primary Schooling 1960	0.796	0.026852	0.007977	0.999	0.96
3	Investment Price	0.774	-0.000084	0.000025	0.999	0.99
4	GDP 1960 (log)	0.685	-0.008538	0.002888	0.999	0.30
5	Fraction of Tropical Area (or people)	0.563	-0.014757	0.004227	0.997	0.59
6	Pop. Density Coastal 1960s	0.428	0.000009	0.000003	0.996	0.85
7	Malaria Prevalence in 1960s	0.252	-0.015702	0.006177	0.990	0.84
8	Life Expectancy in 1960	0.209	0.000808	0.000354	0.986	0.79
9	Fraction Confucious	0.206	0.054429	0.022426	0.988	0.97
10	African Dummy	0.154	-0.014706	0.006866	0.980	0.90
11	Latin American Dummy	0.149	-0.012758	0.005834	0.969	0.30
12	Fraction GDP in Mining	0.124	0.038823	0.019255	0.978	0.07
13	Spanish Colony	0.123	-0.010720	0.005041	0.972	0.24
14	Years Open	0.119	0.012209	0.006287	0.977	0.98
15	Fraction Muslim	0.114	0.012629	0.006257	0.973	0.11
16	Fraction Buddhist	0.108	0.021667	0.010722	0.974	0.90
17	Ethnolinguistic Fractionaliz.	0.105	-0.011281	0.005835	0.974	0.52
18	Gov. Consumption Share 60s	0.104	-0.044171	0.025383	0.975	0.77
19	Population Density 1960s	0.086	0.000013	0.000007	0.965	0.01
20	Real Exc. Rate Distortions	0.082	-0.000079	0.000043	0.966	0.92
21	Fraction Speaking Foreign Language	0.080	0.007006	0.003960	0.962	0.43

Source: Sala-i-Martin, Doppelhofer, Miller(2004)

Avramov (2002): *Stock return predictability and model uncertainty*

- Model averaging can be used not only for inference, but in prediction as well
- Various models for stock market predictability has identified various significant factors
- Model averaging allows to minimize extreme swings, and accounts for the model risk
- Quite often using MA diminishes "apparent" predictability, especially in-sample

Avramov (2002) builds a structural framework, where model averaging is part of the investor's optimal decision

Monthly and quarterly data (1953-1998), 6 portfolios of stocks formed by their characteristics.

Stock market factors

- dividend yield on the value-weighted NYSE index,
- book-to-market on the Standard & Poors Industrials,
- earnings yield on the Standard & Poors Composite,
- the winners-minus-losers one-year momentum in stock returns,
- default risk spread, formed as the difference in annualized yields of Moodys Baa and Aaa rated bonds,
- monthlyrate of a three-month Treasury bill,
- excess return on the CRSP value-weighted index with dividends,
- default risk premium, formed as the difference between return on long-term corporate bonds and return on long-term government bond,
- term premium, formed as the difference between the monthly return on longterm government bond and the one-month Treasurybill rate,
- January Dummy,
- monthly inflation rate,
- size premium,
- value premium,
- term spread, formed as the difference in annualized yield of ten-year and one year Treasuries.

Findings

- Financial returns are slighly predictable, both in-sample and out-of-sample
- Model averaging substantially improves and robustifies prediction
- The best predictors are term and market premium
- Model uncertainty is even more important for prediction, than the estimation error from running regressions

Further questions. What about

- bond or currency pricing?
- volatility effects? (is GARCH any good?)
- influence on portfolio optimisation and market microstructure?

Fereira, Santa-Clara (2012): *Forecasting stock market return: when the sum is greater than the whole*

Sometimes combining predictions from several models can improve the overall performance, because different factors have different properties.

Consider a typical stock return decomposition into capital gain (CG) and dividend gain (DG):

$$1 + R_{t+1} = 1 + CG_{t+1} + DG_{t+1} = \frac{P_{t+1}}{P_t} + \frac{D_{t+1}}{P_t}$$

The capital gain can be expressed through the price/earnings ratio, etc:

$$1 + CG_{t+1} = \frac{P_{t+1}}{P_t} = \frac{P_{t+1}/E_{t+1}}{P_t/E_t}\frac{E_{t+1}}{E_t} = (1 + GPE_{t+1})(1 + GE_{t+1})$$

where GPE is growth in P/E ratio, and GE - in earnings.

Combinings predictors

What of dividend yield?

$$1 + DY_{t+1} = \frac{D_{t+1}}{P_t} = \frac{D_{t+1}}{D_t}\frac{P_{t+1}}{P_t} = DP_{t+1}(1 + GPE_{t+1})(1 + GE_{t+1})$$

where DP_{t+1} is the dividend-price ratio.

Substitute and sum everything together:

$$1 + R_{t+1} = \frac{P_{t+1}}{P_t} + \frac{D_{t+1}}{P_t} = (1 + GPE_{t+1})(1 + GE_{t+1})(1 + DP_{t+1})$$

Take logs:

$$r_{t+1} = log\,(1 + R_{t+1}) = gpe_{t+1} + ge_{t+1} + dp_{t+1}$$

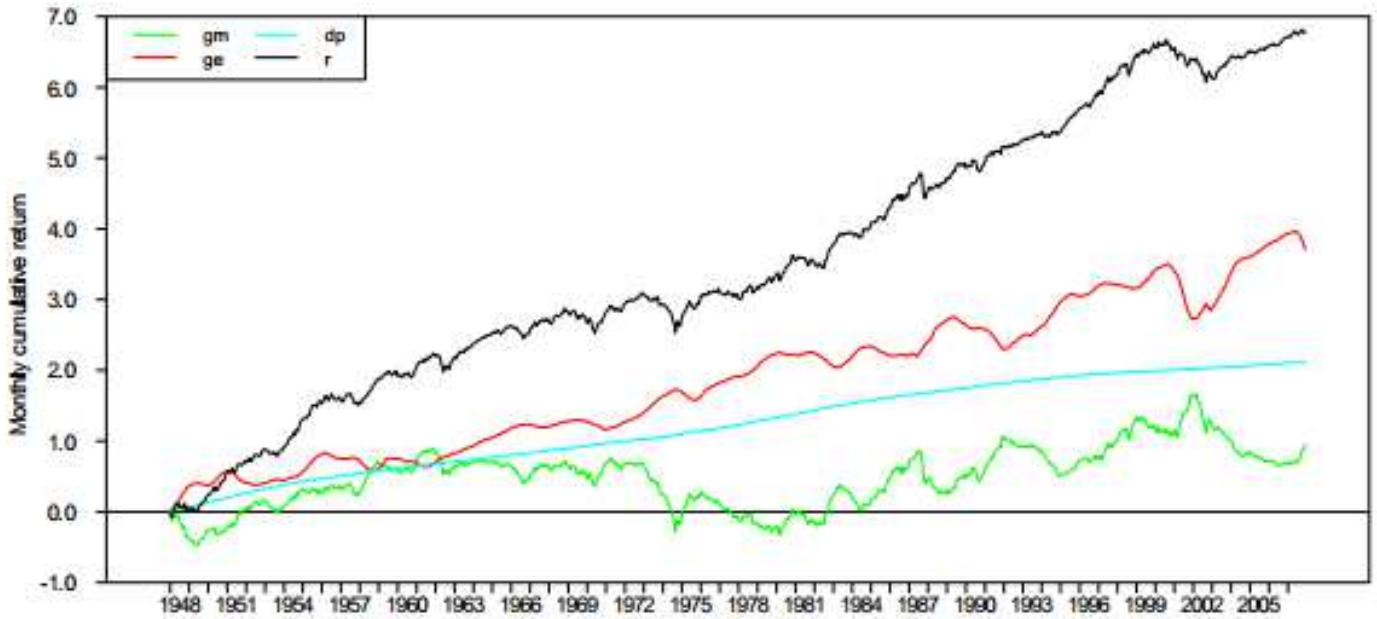
where lower-case letter stand for log-rates.

Sum-of-the-parts method

Instead of trying to forecast returns per se, forecast each component separately, and then sum up.

Taking advantage of the fact that different factors have different predictability: some are more persistent than others (e.g.earnings growth).

These figures show monthly cumulative realized price-earnings ratio growth (gm), earnings growth (ge), dividend price (dp), and stock market return (r).



Source: Fereira, Santa-Clara (2012)

Results

- Campbell and Thomas(2008): usual predictive regressions perform rather poorly, because their parameters change over time
- Different factors can be forecasted using accounting and market-wide data separately
- Combined procedure yields stable predictability over different subperiods (still, not much - only 1-2% monthly out-of-sample)
- Results are better than those from the factors from Goyal and Welsh (2008) (another extensive list of variables)
- A trading strategy, formed using combined prediction, earns a Sharpe ratio of over 0.4 compared to the usual mean forecast

What have we learned

- Model uncertainty is a large component or estimation risk
- Disregarding it can influence many results in empirical work
- There are 2 ways to deal with it: either wisely choose the best model, or try to make inference from several of them
- Bayesian Model Averaging allows to weight the contribution of various specifications
- Combining several forecasts together may improve the fit, because it takes into account individual predictability features