Artificial Intelligence in Finance

Supervised learning - continuous outcome part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 9, 2024

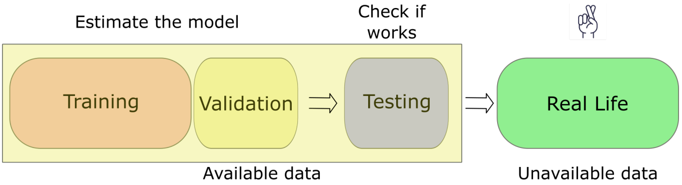
Outline for Section 1

Introduction

- In **supervised** learning, we **primarily** wish to understand how a set of features is **related** to some future **outcome(s)**.
- Understanding **which features** (and how) are **related** to the **outcome** is useful but now **secondary**.

Introduction

- We have **at least one** model that we **train** using a training or/and **validation** sample.
- **Testing** sample is used to **evaluate** (competing) models. Recall:



- **Training** and **validation** (calibration) sample is used **to estimate** (training) and allow models **to learn** (validation). Most of the time they both are referred to as a training sample.

Introduction

We will learn how to go through that process:

- Key estimation models:
  - OLS.
  - LASSO, RIDGE, Elastic Net.
  - Complete subset linear regression.
  - Random forest.
  - Boosted regression trees.
- How to select and evaluate a model:
  - Standard model selection criteria (model fit, AIC, BIC, ...).
  - ML driven methods - model confidence set of Hansen et al., [5].
- How to present results.

Outline for Section 2

Multivariate linear regression

For an overview of linear regression see Green [4]. In an **Econometrics course** the goal is to answer questions like:

- Is there a relationship between apartment area and price?
- How strong is that relationship?
- Is the relationship linear?
- How accurately can we predict price of an apartment?

In an **AI course**, we are **primarily** interested in the forecasting accuracy of one or more models. **Secondarily**, which variables are more or less important (variable importance & e**X**pl**AI**nability).

Recall

Let $Y_i \in R, i = 1, 2, ..., N$, be an observed realization of an outcome variable and $X_{i,k}, k = 1, 2, ..., K$ a specific k^{th} feature:

$$
\begin{aligned}
E(Y|X_i) &= f(X_{i,1}, X_{i,2}, ..., X_{i,K}) \\
E(Y|X_i) &= \beta_0 + \sum_{k=1}^{K} \beta_k X_{i,k}
\end{aligned}
\tag{1}
$$

In **linear** regression, $f(.)$ is assumed to be a **linear function**; linear in parameters β_0, β_k. Here, β_0 is referred to as an **intercept** and β_k as a **slope**.

Recall

In reality, the linear combination does not fit the data (why?):

$$Y_i \neq \beta_0 + \sum_{k=1}^{K} \beta_k X_{i,k} \tag{2}$$

In order to maintain equality we need to introduce a **random term** ϵ_i, with $E(\epsilon_i) = 0$, which satisfies:

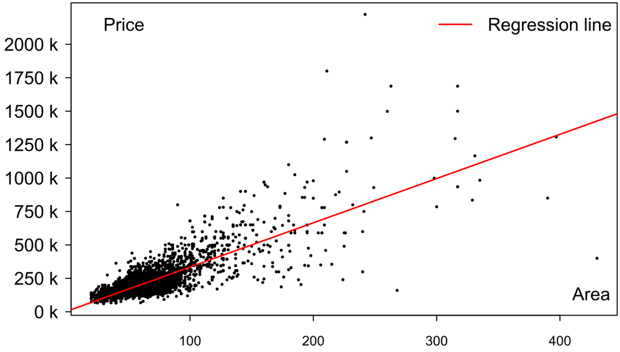
$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{i,k} + \epsilon_i \tag{3}$$

- Why the random term ϵ_i exists?
- What does the random term represent?

Interpretation

Let $P_i, i = 1, \ldots$ be the price of the i^{th} apartment and A_i apartment's area. **Interpret**:

$$P_i = 3702.5 + 3308.5A_i + \hat{\epsilon}_i \tag{4}$$

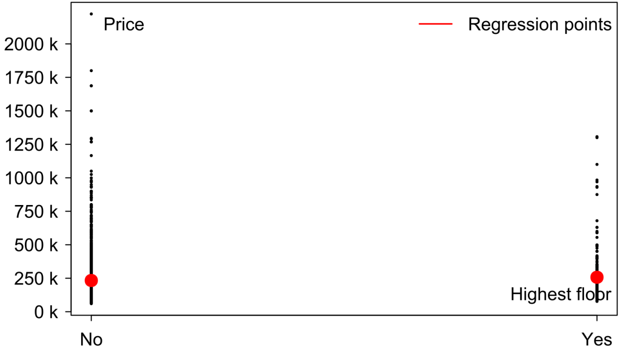


Yes, there is considerable heteroscedasticity of residuals, but point-estimates of the prediction are not affected.

Interpretation

Let $P_i, i = 1, \ldots$ be the price of the i^{th} apartment and F_i returning 1 if apartment is at the highest floor, 0 otherwise. **Interpret**:

$$P_i = 233221 + 24109.5F_i + \hat{\epsilon}_i \tag{5}$$



Interpretation

Compare more models:

$$P_i = 3703 + 3309A_i + \hat{\epsilon}_i$$
$$P_i = 233221 + 24110F_i + \hat{\epsilon}_i$$
$$P_i = 4202 + 3319A_i - 14727F_i + \hat{\epsilon}_i$$
$$P_i = -1751 + 3324A_i - 15758F_i + 12545L_i + \hat{\epsilon}_i$$
$$P_i = 7193 + 3197A_i + 57548F_i - 960F_iA_i - 31067L_i + 639L_iA_i + \hat{\epsilon}_i$$

Here, L_i returns 1 if the apartment is in a building with a lift or 0 otherwise. How do you make **predictions** from such models?

Predictions

Predict the price of an apartment given the model:

$$P_i = 7193 + 3197A_i + 57548F_i - 960F_iA_i - 31067L_i + 639L_iA_i + \hat{\epsilon}_i \quad (6)$$

The apartment has:

- area of $A_i = 50m^2$,
- is not on the highest floor $F_i = 0$,
- the building has a lift $L_i = 1$.

Predictions

The estimated model (within the training sample):

$$P_i = 7193 + 3197A_i + 57548F_i - 960F_iA_i - 31067L_i + 639L_iA_i + \hat{\epsilon}_i$$

The prediction:

$$\hat{P}_i = 7193 + 3197 \times 50 + 57548 \times 0 - 960 \times 0 \times 50 - 31067 \times 1 + 639 \times 1 \times 50$$

$$\hat{P}_i = 7193 + 3197 \times 50 - 31067 + 639 \times 50$$

$$\hat{P}_i = 167926$$

Estimation

Ordinary **L**east **S**quares $\rightarrow$ to estimate parameters, where $\hat{\beta}_0, \hat{\beta}_k, k = 1, 2, ...$ leads to lowest sum of squared residuals:

$$\min_{\hat{\beta}} \rightarrow \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \sum_{k=1} \hat{\beta}_k X_{i,k})^2$$

- How **do we find the minimum** of this function?
- There are **other possible criteria** - not just squared errors.
  - Absolute least squares.
  - Weighted least squares with various **weighting schemes**.

$$\min_{\hat{\beta}} \rightarrow \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} w_i (Y_i - \hat{\beta}_0 - \sum_{k=1} \hat{\beta}_k X_{i,k})^2$$

    - You can be **quite creative** about weights you employ.

Estimation

We estimate β's, but also the standard error (variation of) β, $var(\beta)$:

- It allows us to assess the **accuracy** of the parameter estimate.
- It allows us to test (under certain assumptions) **statistical hypotheses** about β, e.g. $H_0 : \beta_1 = 0$.
- We can assess what influences the precision of our parameter estimates.
- Given initial assumptions about the model (see later) we have many approaches to estimate $var(\beta)$.
- Opposed to **econometrics**, in machine learning, we **do not care so much about the variance of parameter estimates**, instead **we care about** the variance and bias of our **predictions**. Although the two are related.

Model assumptions

Model should work best (including inferences) after satisfying a series of assumptions/rules (see Greene [4]):

1. Model is **linear in parameters** β.
2. Independent variables are not stochastic.
3. $E(\epsilon_i|X_i) = 0 \rightarrow E(Y_i|X_i) = \beta_0 + \sum_{k=1} \beta_k X_{i,k}$.
4. **Homoscedasticity** of residuals $var(\epsilon_i|X_i) = \sigma^2$.
5. For $i \neq j$, residuals ϵ_i and ϵ_j are **uncorrelated** (time-series, spatial,...).
6. There is no covariance between ϵ_i and X_i.

Model assumptions

Model should work best (including inferences) after satisfying a series of assumptions/rules (see Greene [4]):

7. Number of observations needs to be at least the number of parameters, $K \leq N$.
8. Variance of X_k should be finite and positive.
9. Regression model should be **well-specified**.

Assumption No. 9 is quite important and the reason, why artificial intelligence methods (statistical learning in general) will likely get the upper-hand in the future.

Outline for Section 3

Model selection

'All models are wrong, but some are useful.'
Box, G. E. P. (1979, [3])

Standard econometric assumption suggests that we have a well-specified regression.

- That is **almost surely not true**.
- In Finance (observational studies), we can be pretty sure we do not have a well-specified model.

Instead of assuming that one model represents the 'correct specification', an **AI (data)-driven approach** wants to **learn** which model(s) tend to perform better.

In-Sample approach

In an in-sample approach, we do not learn from new data, but select a model (or a variable) using the **training dataset** (not validation) **only**. We will consider following scenarios:

- Select between M-competing models.
  - R^2 and adjusted R^2.
  - Model confidence set of Hansen et al., (2011, [5]).
  - Akaike and Schwartz Information Criteria.
- Variable selection approach.
  - Backward selection.
  - Forward selection.
  - Step-wise selection.

Selection between M-competing models

Example:

Let $R(1)_i, R(2)_i, ..., R(5)_i$ denote the number of rooms of i^{th} apartment and FR_i, PR_i, NW_i a full- and partial-reconstruction and a new apartment. We have $M = 4$ competing models:

$$P_i = \beta_0 + \beta_1 A_i + \epsilon_i$$

$$P_i = \beta_0 + \beta_1 A_i + \sum_{k=2} \beta_k R(k)_i + \epsilon_i$$

$$P_i = \beta_0 + \beta_1 A_i + \beta_2 FR_i + \beta_3 PR_i + \beta_4 NW_i + \epsilon_i$$

$$P_i = \beta_0 + \beta_1 A_i + \beta_2 FR_i + \beta_3 PR_i + \beta_4 NW_i + \sum_{k=5} \beta_k R(k)_i + \epsilon_i$$

Selection between M-competing models

Simple procedure for each of the measures:

1. **Estimate** all M models using data from the training sample.
2. For each model, **calculate** adjusted R^2 (AIC and/or BIC).
3. **Select** the model that has highest R^2 (lowest AIC and/or BIC).

Model fit - R^2

Let $Y_i, i = 1, 2, ..., N$ be the outcome variable and let TSS denote the **T**otal **S**um of **S**quares, i.e. the variability we want to explain in the first place:

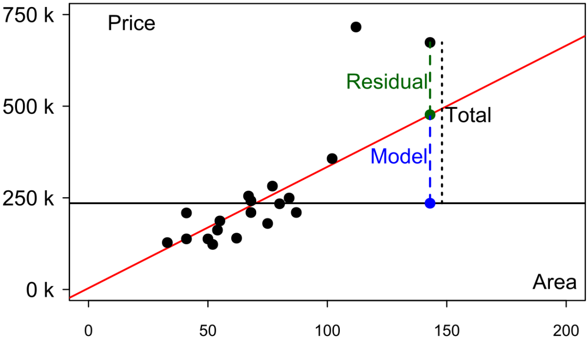
$$TSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{7}$$

We can decompose this variability into the part that we can explain via our model, the **E**xplained **S**um of **S**quares and into the part that the model was not able to explain, the **R**esidual **S**um of **S**quares:

$$TSS = ESS + RSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{8}$$

Model fit - R^2

Visual representation of the **components of** the R^2:



$$TSS = ESS + RSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \quad (9)$$

Model fit - R^2

Recall:

$$TSS = ESS + RSS$$

TSS is given by the outcome variable (Y_i). We can influence ESS (RSS) by using a suitable model. The lower the RSS, the better the model, which leads us to the following, **coefficient of determination**:

$$0 \leqq R^2 = 1 - \frac{RSS}{TSS} \leqq 1 \tag{10}$$

R^2 measures the ratio of explained variability of the outcome variable. It is one of the most popular measures of a goodnes of a model; in general, not just for linear regressions.

Selection between M-competing models

- With the increase of features in the linear regression model, the R^2 almost surely increases.
- It is still **very informative** as it tell us how one model (feature) increases our understanding about the variation of the outcome variable.
- We know that having too much parameters leads to over-fitting $\rightarrow$ we should penalize more complex models; **principle of parsimony**.

Let p denote the number of features, the adjusted R^2 is given as:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right) \leq 1 \tag{11}$$

Selection between M-competing models

Let $LL(m)$ be the log-likelihood (see Econometrics course) of a model $\rightarrow$ **how good the model fits the data**. The Akaike (1974, section V in [1]) and Schwartz (1978, [6]) information criteria are given as:

$$\begin{aligned} AIC &= -2LL(m) + 2p \\ BIC &= -2LL(m) + ln(N)p \end{aligned} \tag{12}$$

Specifically if comparing models from linear regressions, assuming $\epsilon_i \sim N(\mu, \sigma^2)$, the term $-2LL(m)$ is replaced with:

$$Nln\left(\frac{RSS_m}{N}\right) + c \tag{13}$$

and constant c can be ignored.

Selection between M-competing models

Let's turn back to our example:

Comparing models				
Criteria	OLS-1	OLS-2	OLS-3	OLS-4
R^2	61.92%	63.51%	64.89%	66.13%
R_a^2	61.91%	63.45%	64.85%	**66.05**%
AIC	89154	89016	88879	**88763**
BIC	89173	89065	88915	**88831**

$OLS-4$ seems to be the preferred according to all three criteria.

Variable selection approach

Backward elimination:

1. Select a threshold $p-$value (say 0.10).
2. Estimate regression with all features.
3. Estimate $p-$value ($H_0 : \beta = 0$, and use appropriate technique to estimate coefficient standard errors, e.g. bootstrapping, HC, HAC).
4. Find the feature with the highest $p-$value and if it is above the threshold, remove the feature from the specification.
5. Estimate the model with new specification.
6. Repeat step 3 to 5 until the highest $p-$value is less than the threshold $p-$value.

Variable (and model) selection approach

Backward elimination: Looking back at our example, the selected model (OLS-B henceforth) is:

$$\begin{aligned} P_i = &\beta_0 + \beta_1 A_i + \beta_2 R(2)_i + \beta_3 R(3)_i + \\ &\beta_4 R(5)_i + \beta_5 FR_i + \beta_6 NW_i + \epsilon_i \end{aligned}$$

with parameter estimates given by:

$$\begin{aligned} P_i = &1266 + 3123 A_i - 10016 R(2)_i - 30866 R(3)_i + \\ &72638 R(5)_i + 31793 FR_i + 63849 NW_i + \epsilon_i \end{aligned}$$

Compared to $m = 4$, the $R_a^2 = 66.06\%$, $AIC = 88758$ and $BIC = 88808$ are now improved.

Variable selection approach

Forward elimination:

1. Select a threshold $p-$value (say 0.10).
2. Estimate a regression with an intercept.
3. Add one feature which leads to the lowest $p-$value on the corresponding regression coefficient ($H_0 : \beta = 0$).
4. Repeat step 3 until you are **unable to find** a feature that would have a $p-$value below the threshold defined in step 1.

Variable selection approach

Forward elimination: Looking back at our example, the selected model (OLS-F henceforth) is:

$$\begin{aligned} P_i = & \beta_0 + \beta_1 A_i + \beta_2 R(1)_i + \beta_3 R(2)_i + \\ & \beta_4 R(3)_i + \beta_5 SR_i + \beta_6 NW_i + \beta_7 FR_i + \beta_8 R(5)_i + \epsilon_i \end{aligned}$$

with parameter estimates given by:

$$\begin{aligned} P_i = & 5094 + 3094 A_i - 7282 R(1)_i - 13536 R(2)_i - \\ & 33735 R(3)_i + 1856 SR_i + 65418 NW_i + 33055 FR_i + 72705 R(5)_i + \epsilon_i \end{aligned}$$

The $R_a^2 = 66.05\%$, $AIC = 88761$ and $BIC = 88823$ are **worse** as for $m = 5$ (OLS-B).

Variable selection approach

Bi-directional selection

1. Select a threshold p-value (say 0.10).
2. Estimate a regression with an intercept.
3. Add one feature to the model that leads to the lowest p-value.
4. Remove feature(s) that have a p-value above the threshold.
5. Repeat step 3 to 4 until adding a new feature does not lead to feature's p-value below the threshold p-value.

Some variables may be added at one iteration, removed in another and added back to the set of preferred features latter on!

Variable selection approach

Bi-directional elimination: Looking back at our example, the selected model is:

$$P_i = \beta_0 + \beta_1 A_i + \beta_2 R(2)_i + \beta_3 R(3)_i + \\ \beta_4 NW_i + \beta_5 FR_i + \beta_6 R(5)_i + \epsilon_i$$

with parameter estimates given by:

$$P_i = 1266 + 3123 A_i - 10016 R(2)_i - 30866 R(3)_i + \\ 63849 NW_i + 31793 FR_i + 72638 R(5)_i + \epsilon_i$$

The model is the same as $m = 5$!

Loss functions

Forecasts errors lead to costs:

- **Economic losses** are very specific and depend on the application of the forecast.
- **Statistical losses** are more general. For continuous variables most common are:
  - Square Error.
  - Absolute Error.

We use these functions to **optimize** (parameters) our models and to evaluate model forecasts - to **rank** models.

Loss functions

Let $Y_i, i = 1, 2, ...N$, be the i^{th} observed outcome variable and $\hat{Y}_{i,m}, m = 1, 2, \ldots$ the corresponding forecast from model m.
The two common loss (cost) functions are:

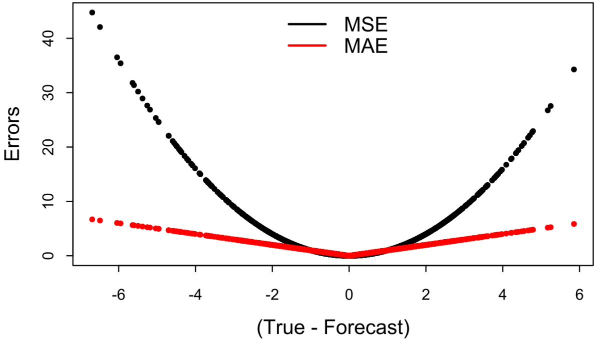
- Mean Square Error.

$$MSE_m = N^{-1} \sum_{i=1}^{N} (Y_i - \hat{Y}_{i,m})^2 \tag{14}$$

- Mean Absolute Error.

$$MAE_m = N^{-1} \sum_{i=1}^{N} |Y_i - \hat{Y}_{i,m}| \tag{15}$$

Loss functions



- Different preference for larger losses.
- Symmetric behavior for under/over predictions.
- You can specify your own losses (e.g. asymmetric losses).

Loss functions

Let's continue with out example. Using models from the **training** sample only (not validation) we predict prices in the **testing** sample: **training** $\rightarrow$ **testing**.

Comparing models		
Models	MSE $\times 10^{-9}$	MAE $\times 10^{-4}$
OLS-1	12.31	6.07
OLS-2	11.99	5.93
OLS-3	11.37	5.64
OLS-4	**11.16**	**5.56**
OLS-B	11.18	**5.56**
OLS-F	11.17	**5.56**

Model confidence set

In the previous example, differences between models $m = 4, 5, 6$ seem to be negligible. Perhaps the three models lead to similarly accurate forecast.

Can we perform a statistical test to compare models?

Currently popular is the **Model Confidence Set** (MCS henceforth) approach of Hansen et al., (2011, [5]). We will follow the exposition of Bernardi and Catania (2018, [2]).

Model confidence set

Let $\hat{M}^0$ denote an initial set of q models ($m = 1, 2, ..., q$). The goal is **for a given confidence** level of $1 - \alpha$ **arrive at a smaller set** of **superior** models denoted as $\hat{M}^*_{1-\alpha}$ with $q^* \leq q$. The general algorithm [2]:

1. Set $M = M_0$, i.e. the initial set consists of all models.
2. Test a null hypothesis that all models lead to equal predictive accuracy (EPA).
   - If the hypothesis is not rejected, the algorithm is terminated.
   - If the hypothesis is rejected, determine the worst model and remove it from M.
3. Repeat step 2 until the EPA hypothesis is not rejected.

Model confidence set

The **EPA hypothesis**. Let $l_{m,i}$ be a loss (e.g. squared error) of the m^{th} model $m = 1, 2, ..., q$ at predicted observation i. The **loss differential** is:

$$d_{m,r,i} = l_{m,i} - l_{r,i}, ;\ m \neq r \tag{16}$$

The EPA hypothesis for a given set of models M can be defined as:

$$\begin{aligned} H_{0,M} &: E(d_{m,r}) = 0, \quad \forall m, r = 1, 2, ..., q \\ H_{1,M} &: E(d_{m,r}) \neq 0, \quad \text{for some } m, r = 1, 2, ..., q \end{aligned} \tag{17}$$

Model confidence set

For comparing only two models, the **test statistics** is given by:

$$t_{m,r} = \frac{\bar{d}_{m,r}}{\sqrt{\widehat{var}(\bar{d}_{m,r})}} \tag{18}$$

where $\bar{d}_{m,r}$ is the average loss between models m and r. The challenge is to calculate $\widehat{var}(\bar{d}_{m,r})$. Hansen et al (2011, [5]) recommends block bootstrap procedure (can be adapted for cross-sections).
Comparing multiple models, the **test statistics** becomes:

$$T_{RM} = max_{m,r \in M} |t_{m,r}| \tag{19}$$

Model confidence set

Let's continue with out example. Using models from the **training** sample only (not validation) we predict prices in the **testing** sample: **training** $\rightarrow$ **testing**.

Comparing models				
Models	$MSE \times 10^{-9}$		$MAE \times 10^{-4}$	
$OLS-1$	12.31		6.07	
$OLS-2$	11.99		5.93	
$OLS-3$	**11.37**	†	5.64	
$OLS-4$	**11.16**	†	**5.56**	†
$OLS-B$	11.18	†	**5.56**	†
$OLS-F$	11.17	†	**5.56**	†

[1] Hirotugu Akaike. “A new look at the statistical model identification”. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.

[2] Mauro Bernardi and Leopoldo Catania. “The model confidence set package for R”. In: *International Journal of Computational Economics and Econometrics* 8 (2 2018), pp. 144–158.

[3] George EP Box. “Robustness in the strategy of scientific model building”. In: *Robustness in statistics*. Elsevier, 1979, pp. 201–236.

[4] William H Greene. *Econometric analysis*. Pearson Education India, 2003.

[5] Peter R Hansen, Asger Lunde, and James M Nason. “The model confidence set”. In: *Econometrica* 79.2 (2011), pp. 453–497.

[6] Gideon Schwarz. “Estimating the dimension of a model”. In: *The annals of statistics* (1978), pp. 461–464.

MUNI ECON

Artificial Intelligence in Finance

Supervised learning - continuous outcome part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 9, 2024

MASARYK
UNIVERSITY