3 Model Selection

You are given N variables $x_{1,t}, ..., x_{N,t}$ observed at different discrete points in time t. Our goal is to explain a variable y_t using linear regression, that is:

$$y_t = a + \sum_{k=1}^{N} I_k \beta_k x_{k,t} + \epsilon_t \tag{17}$$

where fit is an error term, I_k is an indicator variable that equals 1 if the variable x_k, is included in the model, zero otherwise, β_k are the regression coefficients, α is a constant. Assume properties are such that ordinary least squares can be applied to estimate the coefficients β_k.

3.1 How many different models (i.e., combinations of variables) can you construct using the N variables $x_1, ..., x_N$?

It will be possible to construct as many different models as subsets we can find in the set of N elements $\{x_1, ..., x_N\}$. In fact, as described in the statement of the problem, it is possible think of this set of N numbered elements as a binary string, such that we have a 1 if the X_i element is chosen for the regression model, and a 0 if not. As an example, let us assume we have N=4 and choose $I_1 = I_4 = 1$ and $I_2 = I_3 = 0$. The equivalent binary string would be: $\{1, 0, 0, 1\}$, and the resulting model $y_t = \alpha + \beta_1 x_{1,t} + \beta_4 x_{4,t} + \epsilon_t$.

The total number of combinations of N elements will thus be equal to the total possible combinations per string (only two: 0 and 1), to the power of the number of elements in the string (n). As consequence, the **total number of combinations** will be equal to: $\mathbf{2^n - 1}$. The (-1) appears because we must eliminate one combination, given that we do not consider the possibility of a null combination of parameters (at least one element must be selected in order to perform linear regression).

3.2 How would you perform model selection? That is, which regression model would you select as the "best" model among the linear models in this family?

The selection of the “best” regression model must be performed according to two different steps: i) first, we must define what constitutes a good regression model (characteristics it must meet), and ii) second, we must define the procedure to select such model among all the other models. The following sub-sections detail these two processes.

3.2.1 “Best” Model Definition: characteristics of the ideal model

An "ideal" regression model must have 3 characteristics:

1. **Correct Model Specification.** A model has a correct specification if the variables chosen to define it are unbiased and precise. This means that the number of variables chosen is neither too high (overspecified models tend to be biased), neither too low (underspecified models tend to be imprecise).

   We can also include under a correct model specification, that the observations of variables chosen to describe the dependent variable must meet all the assumptions of linear regression:

   - **Linearity Assumption.** The relationship between independent and dependent variables is linear.
   - **Homoskedasticity Assumption**. The variance of the error term of the regression is the same for all observations.
   - **Independence Assumption**. The error term of the regression is uncorrelated across observations.
   - **Normality Assumption**. The error terms of the regression are normally distributed.
   - **Unbiasedness Assumption**. The error terms of the regression have an expected value of zero (unbiased).

2. **Good Regression Fit.** The level of regression fit explains whether the selected model is accurate representing the variation of the dependent variable. Although there can be several parameters that can be used in order to assess whether one regression is better than other, a robust method to compare performance between different regressions is the **Analysis of Variance (ANOVA)**, which helps quantify the usefulness of the independent variable(s) in explaining variation in the dependent variable. The components of ANOVA are: i) the Sum of Squared Errors (SSE) (unexplained variation in dependent variable), ii) the Regression Sum of Squares (RSS) (explained variation in dependent variable), and iii) the Total Sum of Squares (SST) (total variation in independent variable, which is sum of the previous two).

   Thanks to ANOVA and the specification of the SSE, RSS and the SST, it is possible to make use of 3 metrics that constitute a good framework for regression performance comparison:

   - **Standard Error of Estimate (SEE).** Measures the standard deviation of the error terms of the regression (equals the square root of the mean SSE).
   - **F-statistic.** Ratio that measures how well the regression equation explains the variation in the dependent variable (Mean RSS/Mean SSE).
   - $\boldsymbol{R^2}$**.** Measures the fraction of the total variation in the dependent variable that is explained by the independent variable (RSS/SST).

3. **Statistical Significance.** The **p-value** reflects the statistical significance of the regression. If the p-value is low, the regression results will indicate high statistical significance (few outcomes are more extreme than the case of study).

3.2.2 "Best" Model Selection: procedure to find the ideal model

Once we know what constitutes a good model, we can start designing an approach in order to find the ideal model, from the $2^n - 1$ possible combinations. From worst to best, I propose three possible approaches:

1. **Naive approach: Analyzing all $2^n - 1$ combinations.**

   In this approach, I would try all of the possible combinations, according to the following procedure.

   I will first start from the null model (only the intercept coefficient). Then, I will calculate the regression coefficients for all the models with only one independent variable (n possible combinations). Among all of these models, I will filter out those which fail to satisfy the 1^{st} condition of the ideal model (correct specification), or the 3^{rd} condition (those whose p-value is higher than a threshold). Among the remaining models, I will select the one which which presents higher R^2. In case of very similar results, I will give preference to the model with higher F-statistic.

$$\begin{array}{ll}
\text{Subset 1:} & \text{Subset 2:} \\
\quad Y_i = b_0 + b_1X_{1_i} & \quad Y_i = b_0 + b_1X_{1_i} + b_2X_{2_i}\ldots \\
\quad \boldsymbol{Y_i = b_0 + b_2X_{2_i}} & \quad Y_i = b_0 + b_2X_{2_i} + b_3X_{3_i}\ldots \\
\quad Y_i = b_0 + b_nX_{n_i} & \quad \boldsymbol{Y_i = b_0 + b_4X_{4_i} + b_nX_{n_i}}\ldots
\end{array}$$

   Once done this process for all the models of 1 variable, I will repeat the same procedure for all the models with 2, 3,... and n variables. Finally, I will compare the winners of each subset among them, in order to find the best overall model.

2. **Average approach: Stepwise Selection.**

   The disadvantage of the naive approach is the computational complexity needed to analyze the $2^n - 1$ combinations of possible regressions. The stepwise selection approach, partially addresses this problem and only analyzes n! combinations of independent variables.

   The procedure in this case is very similar to the previous one. We start first analyzing all the models with only one independent variable, and choose the best one following the 3 rules of the ideal model, as detailed before. However, in this case instead of analyzing again all the possible combinations with 2 independent variables, we only select those that include the "best" (e.g. if we concluded that the best 1 variable regression model was for X_2, we will now analyze regressions that combine X_2 plus any other second variable).

$$\begin{array}{ll}
\text{Subset 1:} & \text{Subset 2:} \\
\quad Y_i = b_0 + b_1X_{1_i} & \quad Y_i = b_0 + b_2X_{2_i} + b_2X_{2_i}\ldots \\
\quad \boldsymbol{Y_i = b_0 + b_2X_{2_i}} & \quad Y_i = b_0 + b_2X_{2_i} + b_3X_{3_i}\ldots \\
\quad Y_i = b_0 + b_nX_{n_i} & \quad \boldsymbol{Y_i = b_0 + b_2X_{2_i} + b_nX_{n_i}}\ldots
\end{array}$$

This process is repeated until in aggregate a model with n variables, and we will select the best regression among all the subset winners.

The selection process described correspond to the FORWARD stepwise selection (starting from the null combination and adding variables). The same procedure is applicable for the BACKWARD stepwise selection (starting from the complete combination and eliminating variables).

3. **Best-practices approach: Variable Stepwise Selection using AIC/BIC**

The disadvantage of the average approach, is that forward/backward selection methods usually are overused, and often lead to the solution of models that are only locally optimal. This disadvantage can be addressed with the use of variable stepwise selection, combined with information theory based metrics, that allow us to compare the amount of information loss between different combination of variables, and select the most statistically significant combination of variables all the time. This approach will allow us to find a very good compromise between computational time complexity and optimality of the solution.

The procedure to implement **VARIABLE stepwise selection** is similar to the forward/backward stepwise process. The idea is to start with either the null combination or the full combination, and then investigate i) the effect of removing the least informational variable (for instance, the one with the lowest p-value), and ii) the effect of adding the most informational variable (the non-selected variable that produces the highest p-value on the regression). The removing and adding steps will be alternatively performed until we find that adding/removing variables no longer produce an improvement in the informational efficiency of the regression (the p-value is the lowest).

Additionally, instead of using p-value as only metric of statistical relevance, it is possible to improve the optimality of the solution by making use of informational theory based metrics, like the **AIC** (Akaike Information Criterion) or the **BIC** (Bayesian Information Criterion), which offer an estimate of the relative information lost when a given model is used to represent the process that generated the data. They are both similar, with the only difference being that BIC penalizes more regressions with a high number of independent variables:

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2(p+1) \tag{18}$$

$$BIC = n \ln\left(\frac{RSS}{n}\right) + (p+1)\ln n \tag{19}$$

All in all, we can conclude that applying the variable stepwise selection algorithm with either AIC/BIC, will maximize the trade-off between computational complexity and quality of the solution, and thus this is the procedure I would use to select the "best" model out of all the $2^n - 1$ combinations.

Additional tools that I would use to verify that this is indeed the best model would include PCA (Principal Component Analysis), in order to explain which variables contribute more to explaining the model.

3.3 Assume that a member of our team developed a statistical test that you can apply to your models. The test creates a test statistics for each model and the true distribution of the test statistics is known. The null hypothesis is that the model is “not good”, the alternative is that the model is “good”. You apply the test to all of the K models that you constructed and choose the confidence interval $1 - \alpha$ equal to 95%.

I) If K is very large, how many models do you expect to be considered “good” by the test?

II) Assume 100% of models that you test are “not good”. How many models will be considered ”good” in this case?

I) If K is very large, how many models do you expect to be considered “good” by the test?

If the probability distribution of the test statistic is known, then it would be possible to perform a hypothesis test to verify how extreme the obtained outcome is, and compare it with the confidence interval of 95%. If the test statistic is lower than 0.05, then the result can be considered as significantly abnormal and we could reject the null hypothesis, concluding that the model is “good”.

To answer the question of how many models are expected to be considered as “good” by the test, is important to clarify one statement of the problem: if the distribution known is the test statistic probability distribution, understanding this as all the frequencies of obtaining a given result in the test statistic, by definition we can expect only a 5% of the models to be considered as “good” (the 5% most favorable results). This is because out of the 100% test results, we know from the probability distribution that 95% of the outcomes will be considered as “not good”. Therefore, the 5% remaining with the most extreme test statistic will be classified as “good”.

Total models to be considered “good” by the test”:

$$\text{N}^{\underline{\text{o}}} \text{ of “good” models} = 0.05 \cdot (2^n - 1) \tag{20}$$

*Again, this is supposing that the probability distribution of the test statistics obtain a measure for the test (cardinal scale), although a binomial distribution with only two different cases (“good” and “not good”) could also concern here, case in which a logistic regression should be applied.

II) Assume 100% of models that you test are “not good”. How many models will be considered ”good” in this case?

This question depends on the number of models that have actually been tested. Out of a population of $2^n - 1$ possible combinations of models, if we test all of them as suggested in the naive approach of section 3.2, it would be impossible by definition to obtain 0% of "good" models (because the probability distribution indicates that at least a 5% of "good" models must exist at 95% confidence interval.

If only a portion of the $2^n - 1$ possible combinations of models have been tested, it is possible to obtain a 0% of "good" models. In this case, we can take two approaches:

- **Absolutist approach.** If we are still concerned about obtaining a model which ranks within the best 5% of models of the probability distribution, we CANNOT consider any of the tested models as "good". We will have two options to continue: either i) test more models until we find a "good" one, or ii) lower the confidence interval to obtain a "good" model. This last path however will increase the probabilities of Type I error (consider as "good" a model that is actually "not good").
- **Relativist approach.** If we have run out of resources to run additional tests, and none of the obtained tests are good, we can consider all the tested models as a sample of the population (although it would be a biased sample, because it counts with no "good" models), and consider the 5% of models within our tested sample as "good" models. Alternatively, we could again make use of the previous approach of lowering the confidence interval, assuming a higher risk of obtaining a Type I error.