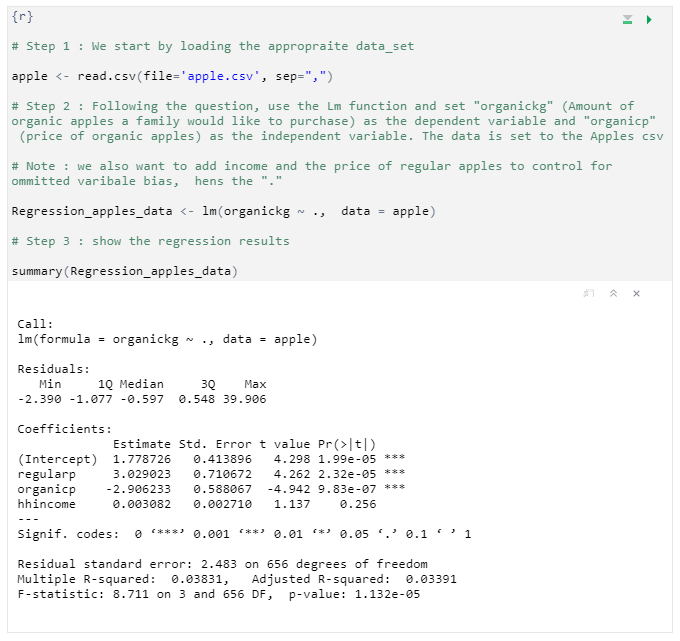
MSIN0094

Assignment Answer Sheet

Assessment 3:

Exercice 1 :





1. Write down your equation

Organickg = b0 + b1 Organicp + b2 Hincome + b3 regularp

**b0** = Intercept / Constant **b1** = Coefficient of the price of Organic apples **b2** = Coefficient of the Houshold income **b3** = Coefficient of regular income **e** = Disturbance

1. justify the model specification:

For our purposes, we have chosen the Ordinary least square estimation model. As our outcome variable is a quantity (continuous variable) and not a binary variable, we can disguard the Linear probability model, the Probit model and the logistic regression. To ensure that we can choose the OLS model however, we must also ensure the following:

**No Multicollinearity:** there is no correlation between predictor variables

**Homoscedacity:** Given our case we can assume that the error term remains constant there is no Heteroskedacity

**Exogeneity:** We can assume that this assumption is held. In our case as the independent variable is not dependent on the outcome variable. The amount of apples purchased won’t affect the price of the organic apples, price of regular apples and household income.

**Given these assumptions are met, we can use the OLS model in our case**

1. Regression results

**Organickg** = 1.779 + organicp \* ( -2.91) + hhincome \* (0.003) + regularp \* (3.03) + e

1. Interpretation:

Assuming all of the other variables are fixed at the same level ceteris paribus. The increase of  
one unit of the price of organic apples will decrease the purchase amount of number of apples by **2.91 KG**. Our regression output suggests that the result is very significant at a 0.01 level.

Exercice 2 :

a) best subset selection

Using the best subset selection method, we are going to determine the best model. The benefit of this method is that is tests each possible model despite being computational expensive.

**Step 1** : Estimate the best BIC for all models with **1** independent variable, testing for each available variable.

* The best results come from X 3 with 673,4 BIC

**Step 2:** Estimate the best BIC for all models with **2** independent variable , testing for each available variable.

* The best results come from X2, X 3 with 602,8 BIC

**Step 3:** Estimate the best BIC for all models with **3** independent variable, testing for each available variable.

* The best results come from X2, X 3, X4 with 597,3 BIC

**Step 4:** Estimate the best BIC for all models with **4** independent variable, testing for each available variable.

* The best results come from X1, X2, X3, X4 with 599,5 BIC

**Final step**: After estimating the best model for every number of variables, the next step is to compare them with each other. The best one being **X2, X3, X4** with 597,3 BIC

B) forward stepwise selection

Using the forward step-wise selection, is less computational expensive than the best subset selection method however doesn't explore every model.

**Step 1** : Estimate the best BIC for all models with **1** independent variable, testing for each available variable.

* The best results come from X3 with 673,4 BIC

**Step 2** : use the findings in Step one That X3 is the best model with one independent variable. Compare only the three models where X3 is paired up with X1, X2, and X4

* The best results come from X2, X3 with 602,8 BIC

**Step 3:** use the findings in Step two That X2, X3 is the model with two independent variable. Compare only the three models where X3, X3 is paired up with X1, X2, and X4

* The best results come from X2, X3, X4 with 597,3 BIC

**Step 4:** The only variable left that can be added is X1 resulting BIC is 599 fo X1, X2, X3, X4

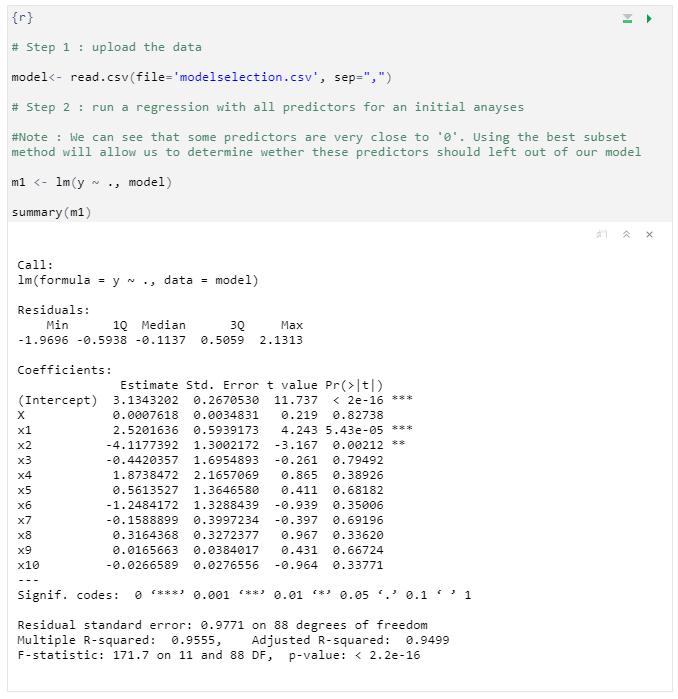
**Final step :** the final step is to compare the results of every model for each number of independent variables

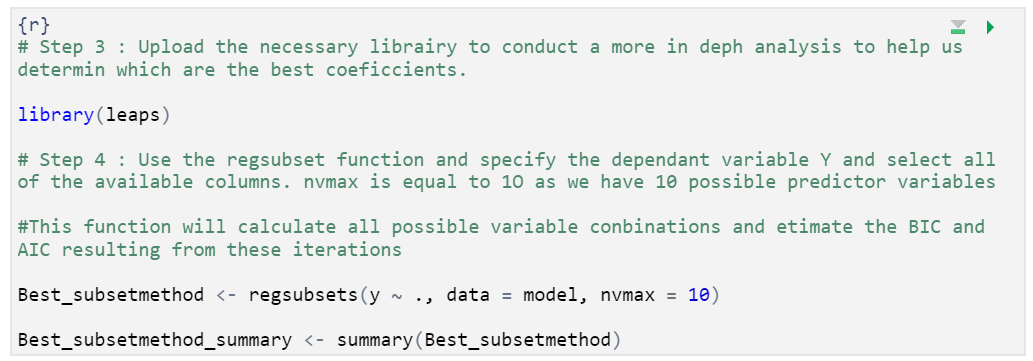
* X3 with 673,4 BIC
* X2, X3 with 602,8 BIC
* X2, X3, X4 with 597,3 BIC
* X1, X2, X3, X4 with 597,3 BIC

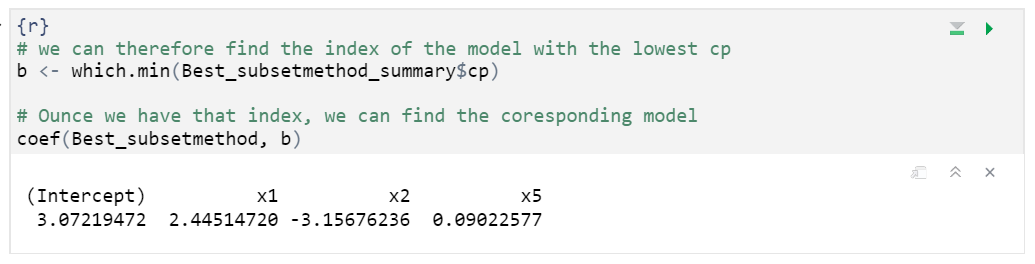
The model with the lowest BIC is therefore X2, X3, X4. This model is estimated to result in the least testing error.

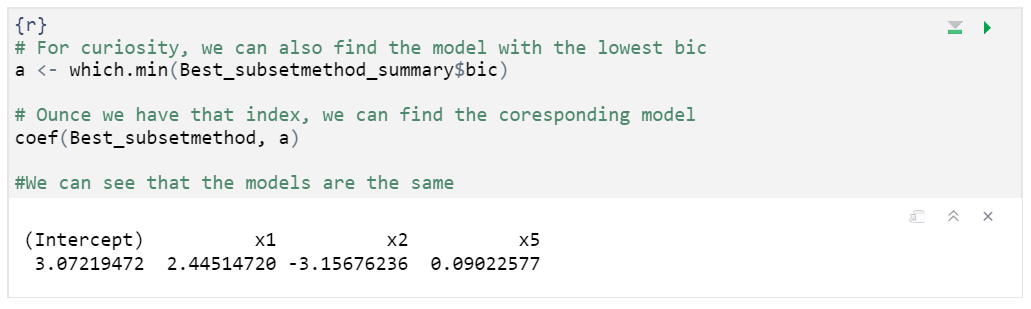
Exercice 3 :

* 1. Best subset method







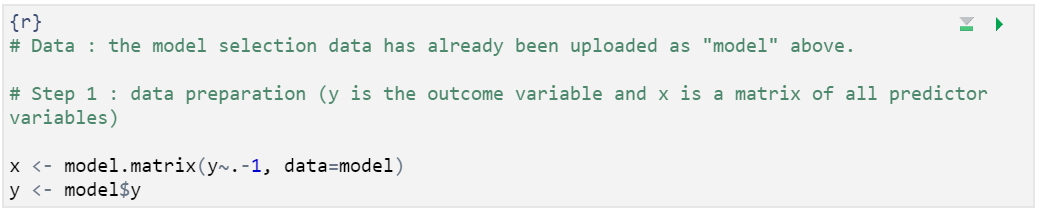


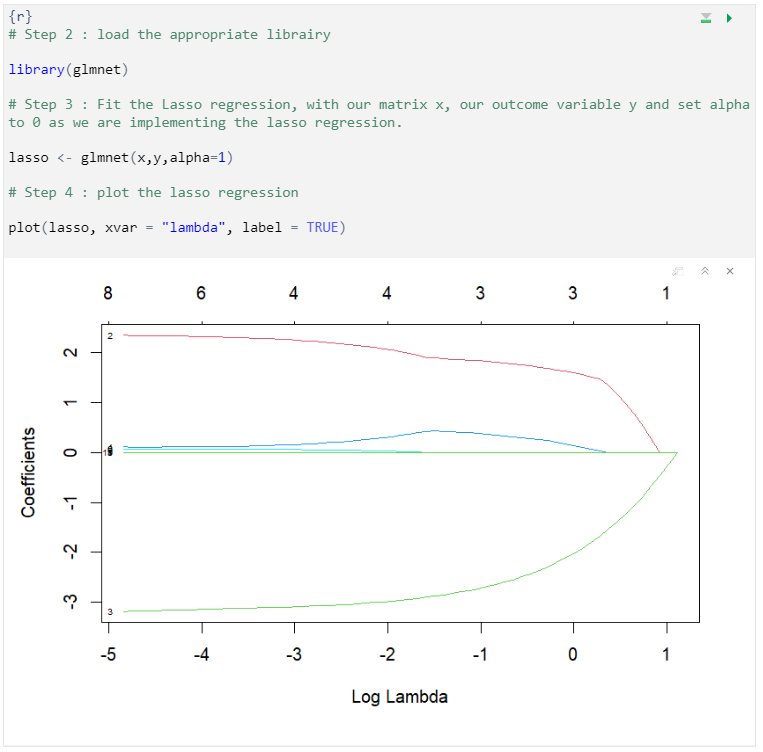
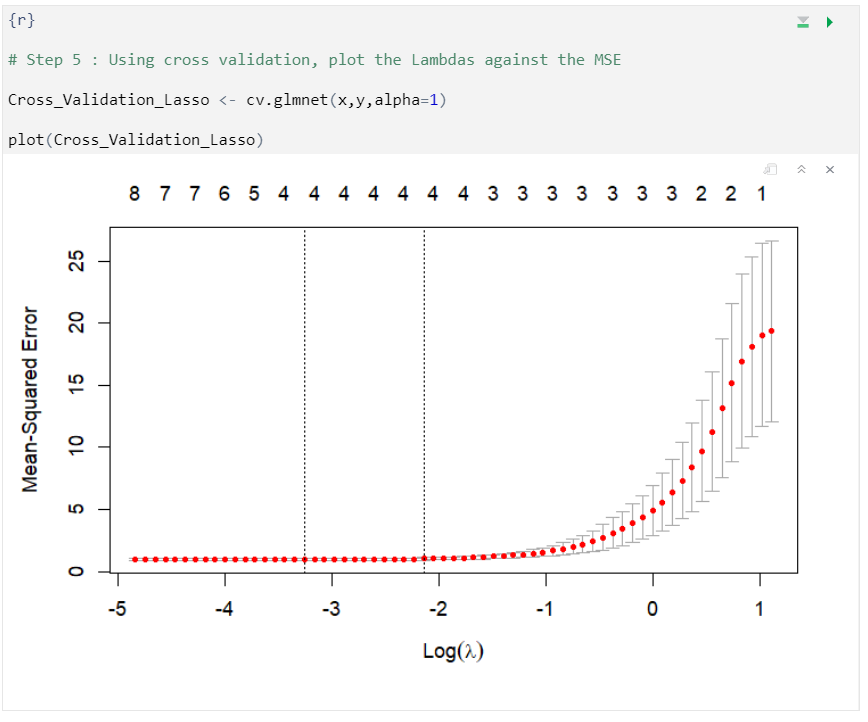
We can see that using the BIC and CP results in the same prediction. Using the best subset method confirms that the most appropriate model for our data is:



* 1. shrinkage method: Lasso Regression

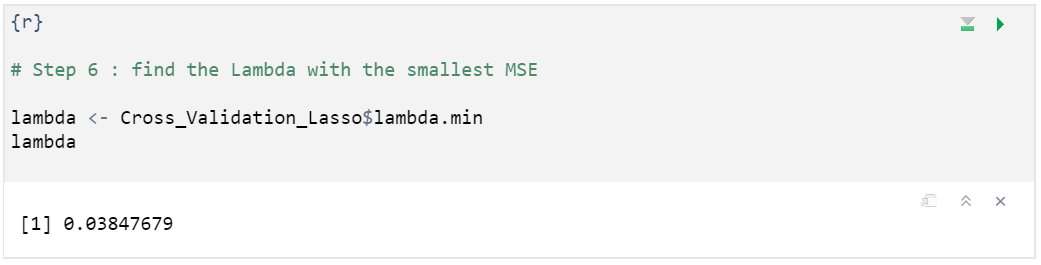
When we ran the Lm function on our data, we noticed that there were 5 variables that we very close to 0 (0.5 to 0). We therefore choose to use the Lasso regression as we have 10 explanatory variables and wish to create a more interpretable model that shrinks the insignificant variables to 0. The ridge method would have kept all predictors in the model decreasing interpretability.

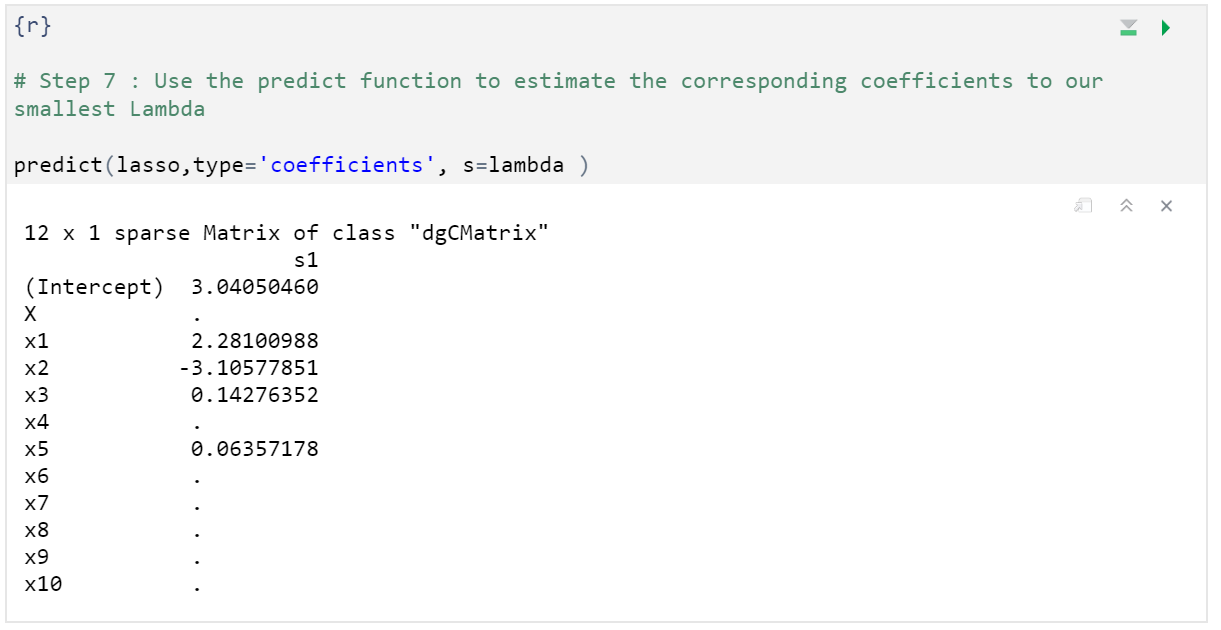
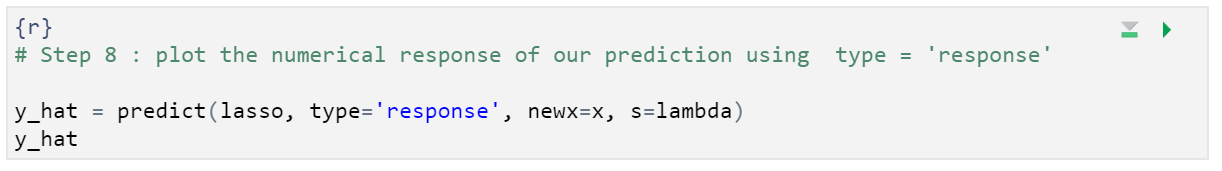




In the Lasso regression, the higher the lambda the more our coefficients get pushed to 0. In order to find which lambda results in the less bias we have to run tests to estimate the MSE for each lambda. we therefore use cross validation, a resampling procedure that allows us to estimate the bias using the entire data, the test set and the training set. Above is shown the plot of all of the possible Lambda on the x axis and the corresponding MSE on the Y axis.

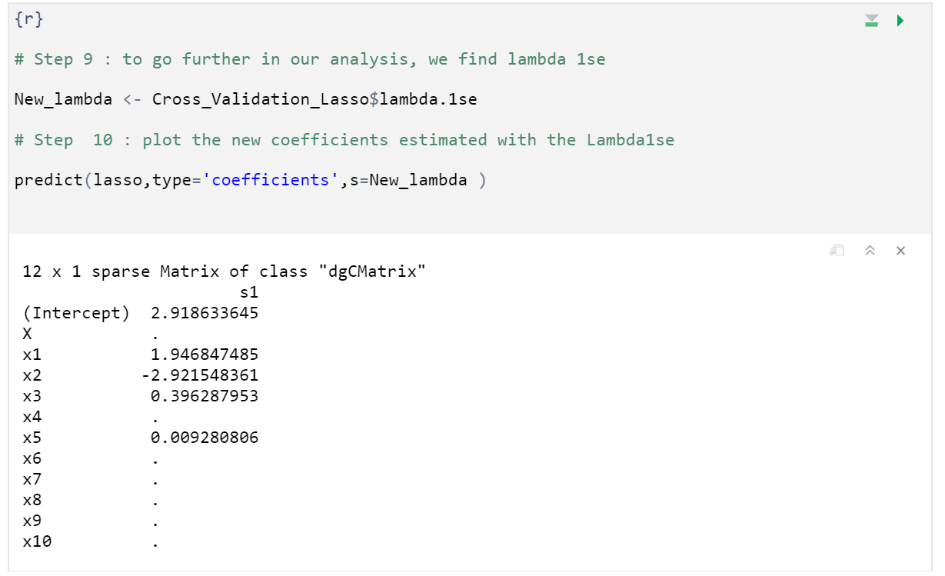
* + **First** we estimate the model with the lowest MSE

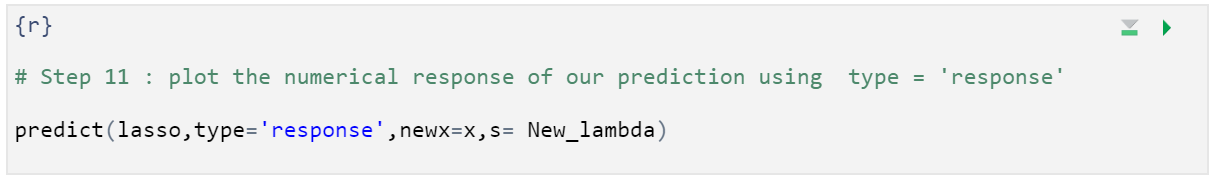




* + **Second**, we estimate the model using lambda 1se instead of the min Lambda

**lambda 1se**: Is the largest value of Lambda which outputs the model where the error of the cross-validation method is within one standard error of Lambdamin. The metric takes into account that the curve might be estimated with error and helps mitigate possible overfitting





Exercice 4:

**4a)**

We are given a data set with 5 observations (x1, x2, x3, x4, x5)

With the bootstrapping method, the probability of x1 appearing in x1\* is **1/5**

With Five observations in every bootstrap sample, the probability of x1 not appearing in any of the bootstrap observations is therefore **4/5 \*4/5 \*4/5 \*4/5 \*4/5 = 0,3277**

Similarly, with the bootstrapping method, the probability of x3 appearing in x1\* is **1/5**

With Five observations in every bootstrap sample, the probability of x3 not appearing in any of the bootstrap observations is therefore **4/5 \*4/5 \*4/5 \*4/5 \*4/5 = 0,3277**

**4B)**

We are given a data set with 100 observations (x1, x2, … x100)

With the bootstrapping method, the probability of x1 appearing in x1\* is **1/100**

With Five observations in every bootstrap sample, the probability of x1 not appearing in any of the bootstrap observations is therefore (**99/100) ^ 100 = 0.366**

Using our results, we can calculate