Multiple linear regression

Prepared by: Prof. Yasmina Guechari

1. What is Multiple linear regression?

I. **Definition:** Multiple linear regression refers to a statistical technique that is used to predict one variable based on the value of two or more variables.

- The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

1. What is Multiple linear regression?

II. Multiple Linear Regression Formula

- Linear regression formula is as follow:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots \beta_p X_p + \varepsilon$$

- β_0: is the y-intercept, i.e., the value of y when all X_i are 0.
- Y is the dependent or predicted variable.
- β_i is the slope coefficient for (or the regression coefficients associateted with) each independent variable X_i.
- ε : is the model's random error (residual) term.

1. What is Multiple linear regression?

III. Assumptions of Multiple Linear Regression

- Multiple linear regression is based on the following assumptions:

i. **A linear relationship between the dependent and independent variables**

ii. **The independent variables are not correlated with each other:** The data should not show multicollinearity, which occurs when the independent variables are highly correlated.

1. What is Multiple linear regression?

iii. **The variance of the residuals is constant:** This is known as homoscedasticity.

iv. **Independence of observation**: this means that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic.

2. Estimates of the Model Parameters

I. Estimates of β coefficients

- The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample it is calculated using **the least square method**. The exact formula for this coefficient is given by a matrix notation.

Estimates of the Model Parameters

II. Interpretation of the Model coefficients

- Each β coefficient represents the change in the dependent variable *y*, per unit increase in the associated independent variable when all the other predictors are held constant.
- For example, β_1 represents the change in the dependent variable *y*, per unit increase or decrease in x_1 when x_2, x_3, ..., x_p are held constant.
- The intercept term, β_0, represents the dependent, *y*, when all the predictors or independent variable x_1, x_2, ..., x_p, are all zero

2. Estimates of the Model Parameters

Example:

- Suppose we fit a multiple linear regression model using the predictor variables *hours studied* and *prep exams taken* and a response variable *exam score*.
- The following table shows what the multiple linear regression output might look like for this model:

	Coefficient	Std. Error	t-Statistic	Prob.
Constant	67,67352554	2,815802228	24,03347965	1,45819E-14
hours	5,555748295	0,899191699	6,178602739	1,01069E-05
prep_exam	-0,601686805	0,914385031	24,03347965	0,519335226

2. Estimates of the Model Parameters

- From the model output, the coefficients allow us to form an estimated multiple linear regression model:
- **Exam score = 67.67 + 5.56*(hours) – 0.60*(prep exams)**
- The way to interpret the coefficients are as follows:
- Each additional one unit increase in hours studied is associated with an average increase of **5.56** points in exam score, *assuming prep exams is held constant.*
- Each additional one unit increase in prep exams taken is associated with an average decrease of **0.60** points in exam score, *assuming hours studied is held constant.*

2. Estimates of the Model Parameters

- We can also use this model to find the expected exam score a student will receive based on their total hours studied and prep exams taken. For example, a student who studies for 4 hours and takes 1 prep exam is expected to get a score of **89.31** on the exam:
- Exam score = 67.67 + 5.56*(4) -0.60*(1) = **89.31**

3. Fitted Values and Residuals

- A **residual** (**error**) term is calculated as $\varepsilon = Y_i - \hat{Y}_i$, the difference between an actual and a predicted value of y.
- Variation of the dependent variable ***unexplained*** by the independent variables.
- SSE= $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$

4. ANOVA Table

Source	SS	MS	F
Regression	$SSR = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2$	MSR = SSR / k	MSR / MSE
Error	SSE= $\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$	MSE = SSE / (n – (k+1))	
Total	$SST = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$		

regression statistic	
Coefficient of determination	0,856753884
Coefficient of determination R-squared	0,734027217
Coefficient of détermination adjusted R-squared	0,702736301
Standard Error	5,365703261
Observations	20

ANOVA				
	Degree of freedom	Sum squared	F	Valeur critique de F
Regression SSR	2	1350,756885	23,45815717	1,29156E-05
Residual SSE	17	489,4431152		
Total. SST	19	1840,2		

5. Coefficient of Determination, R-squared, and Adjusted R-squared

- As in simple linear regression $(R)^2$ measures the proportion of the total variation of the dependent variable that is explained by the independent variables.
- R^2 =SSR/SST = $1 - SSE/SST$
- If we start with a simple linear regression model with one predictor variable, X_1, then add a second predictor variable, X_2, *SSE* will decrease (or stay the same), *SSR will increase (or stay the same),* while *SST* remains constant, and so, R^2 will increase (or stay the same).
- In other words, R^2 always increases (or stays the same) as more predictors are added to a multiple linear regression model.

5. Coefficient of Determination, R-squared, and Adjusted R-squared

- an alternative measure, adjusted R^2, does not necessarily increase as more predictors are added, and can be used to help us identify which predictors should be included in a model and which should be excluded.
- Adjusted $R^2 = 1 - (\frac{(n-1)}{(n-(k+1))})(1-R^2)$
- while it has no practical interpretation, is useful for such model building purposes.
- Simply stated, when comparing two models used to predict the same response variable, we generally prefer the model with the higher value of adjusted R2

5. Coefficient of Determination, R-squared, and Adjusted R-squared

- Continue with our example:
- $R^2 = \frac{1350.75}{1840.20} = 0.734 = 73.4\%$
- In this example, 73.4% of the variation in the exam scores can be explained by the number of hours studied and the number of prep exams taken.
- $R^2 = 1 - \frac{489.44}{1840.20}$=0.734=73.4%

$Adjusted\ R^2 = 1 - (\left(\frac{19}{(20-(2+1))}\right)(1 - 0.734))$= 0.703

6. Significance Testing of Each Variable

- Within a multiple regression model, we may want to know whether a particular *x*–variable is making a useful contribution to the model.
- That is, given the presence of the other *x*–variables in the model, does a particular *x*–variable help us predict or explain the *y*–variable?
- For instance, suppose that we have two *x*–variables in the model. The general structure of the model could be
- $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$

6. Significance Testing of Each Variable

- As an example, to determine whether variable x_1 is a useful predictor variable in this model, we could test
- $\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$
- Hypothesis above were the case (is accepted), then a change in the value of x_1 would not change y, so y and x_1 are not linearly related. Also, we would still be left with variables x_2 being present in the model.

6. Significance Testing of Each Variable

- To test the the hypothesis we have to check the t-statistics or the p-value of the relevent coefficient
- **Coefficient P-values.** The individual p-values tell us whether or not each explanatory variable is statistically significant.
- In our example
- $\begin{cases} H_0 : \text{There is no relationship between hour studied and exam score } \beta_1 = 0 \\ H_a : \text{There is a significant relationship between hour studied and exam score. } \beta_1 \neq 0 \end{cases}$
- $\begin{cases} H_0 : \text{There is no relationship between prep}-\text{examp and exam score } \beta_2 = 0 \\ H_a : \text{There is a significant relationship between prep}-\text{exam and exam score. } \beta_2 \neq 0 \end{cases}$

	Coefficient	Std. Error	t-Statistic	Prob.
Constant	67,67352554	2,815802228	24,03347965	1,45819E-14
hours	5,555748295	0,899191699	6,178602739	1,01069E-05
prep_exam	-0,601686805	0,914385031	24,03347965	0,519335226

- We can see that hours studied is statistically significant (p = 0.00) while prep exams taken (p = 0.52) is not statistically significant at α = 0.05.
- So we reject the null hypothesis and we confirm the alternative hypothesis, for the hour studied variable, this means that each 1 unit increase (decrease) in hour studied the exam score move up (move down) by 5.55 unit.
- In the other hand according to t-statistic and p value the prep_exam has no significant relationship with the exam score as we cannot reject the null hypothesis.

7. Testing the overall significance of the model

- The F-statistic is used to test the overall significance of a regression model, and it compares the fit of the full model (with predictors) to the fit of a null model (without predictors). The formula for the F-statistic is:
- $F = \frac{\frac{SST-SSE}{k}}{\frac{SSE}{n-k-1}}$
- In other words, it tells us if the two explanatory variables combined have a statistically significant association with the response variable.
- $\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots 0 \\ H_a: \beta_1 \neq 0 and\ \beta_2 \neq 0 \end{cases}$

7. Testing the overall significance of the model

- The F-statistic helps assess whether adding independent variables to the model improves its overall fit.
- **Significance of the Model:** A significant F-statistic indicates that the model, with at least one predictor variable, is statistically different from a model with no predictors.

In our example we find f-statistic and its associated p-value as follow:

F	Valeur critique de F
23,45815717	1,29156E-05

- In this case the p-value is less than 0.05, which indicates that the explanatory variables hours studied and prep exams taken combined have a statistically significant association with exam score.

8. Testing the contribution of an additional explanatory variable

- Here's what happens to the F-statistic when you add more independent variables:

1. If the added variables contribute significantly to explaining the variance in the dependent variable:

The numerator ((SST−SSE)/k) may increase because the model with predictors explains more variance in the dependent variable compared to the null model.This leads to an increase in the F-statistic.

2. If the added variables do not contribute significantly:

The increase in the numerator may be small or negligible, and the F-statistic may not increase significantly.

8. Testing the contribution of an additional explanatory variable

- From the previous example we can calculate the F-statistic
- **With only one independent variable Hour studied: F=(SSR/k)/(SSE/(n-(K+1)) =(1338.29/1)/(501/17)= 47.99**
- With **two independent variables Hour studied and prep-exam**

F= (1350.76/2)/(489.44/17)=23.46

The F-statistic is a measure of whether the overall model (including all predictors) is statistically different from a model with no predictors.

In our example we find that the additional predictors do not improve the model fit, This suggests that the model with two predictors does not perform significantly better than a model with only one predictor.

8. Testing the contribution of an additional explanatory variable

- The F-statistic tends to decrease after adding another explanatory variable,this means that the prep-exam do not contribute meaningfully to explaining the variance in the dependent variable, this additional predictors do not improve the model fit
- Here's what happens to the R-squared when you add more independent variables:

8. Testing the contribution of an additional explanatory variable

- The R-squared for the model with two explanatory variable

regression statistic	
Coefficient of determination	0,856753884
Coefficient of determination R-squared	0,734027217
Coefficient de détermination adjusted R-squared	0,702736301
Standard Error	5,365703261
Observations	20

- The Coeffiient of determination for the model with only one independent variable

regression statistic	
Coefficient of determination	0,85279119
Coefficient of determination R-squared	0,727252814
Coefficient de détermination adjusted R-squared	0,712100192
Standard Error	5,280516451
Observations	20

8. Testing the contribution of an additional explanatory variable

- Simply stated, when comparing two models used to predict the same response variable, we generally prefer the model with the higher value of adjusted R2
- In our example **the model with only one independent variable explain better the variation in the dependent variable that the model with two independent variables**.

9. How to Check for multicollinearity?

1. **Multicollinearity** refers to the presence of high correlations among independent variables in a regression model.

- It can cause issues in the estimation of regression coefficients and their interpretation.
- Here the most important methods to test for multicollinearity:

i. Variance Inflation Factor (VIF):

- **VIF measures how much the variance of an estimated regression coefficient increases if the predictors are correlated.**
- For each independent variable, calculate its VIF using the formula: $VIF = \frac{1}{1-R^2}$ where R^2 is the coefficient of determination from regressing one independent variable against all the other independent variables.

9. How to Check for multicollinearity?

- Generally, a VIF value above 5 is often considered an indicator of problematic multicollinearity.

  When R_i^2 is equal to 0, and therefore, when VIF is equal to 1, the i^{th} independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.

- In general terms,
- If the $VIF \leq 1$ the variables are not correlated (no multicollinearity)
- If $1 < VIF \leq 5$ variables are moderately correlated (there is small degree multicollinearity)
- If $VIF > 5$ variables are highly correlated (thereis multicollinearity)

9. How to Check for multicollinearity?

- The higher the VIF, the higher the possibility that multicollinearity exists. When VIF is higher than 5, there is significant multicollinearity that needs to be corrected.
- We can remove the multicollinearity by using the following techniques:

i. Identify pairs of independent variables with high correlation and consider removing one of them.

ii. If possible, combine highly correlated variables into a single variable. For example, if you have variables measuring similar concepts in different units, consider creating an index or using principal component analysis.

9. How to Check for multicollinearity?

iii. Transform variables, such as taking the logarithm or square root, to reduce the impact of extreme values and potentially alleviate multicollinearity.

iv. Collect more data to increase the sample size. A larger sample size can help stabilize estimates and reduce the impact of multicollinearity.

v. Centring variables: Centering involves subtracting the mean of a variable from all observations. Centering can sometimes help reduce multicollinearity, especially if variables have different scales.

9. How to Check for multicollinearity?

vi. Use Stepwise Regression: Perform stepwise regression to iteratively add or remove variables based on statistical criteria. This can help select a subset of variables that minimizes multicollinearity.

10. How to Check for the heteroscedasticity?

- **The variance of the residuals is constant:** Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. This scenario is known as homoscedasticity.
- **Heteroscedasticity** refers to the situation where the variance of the errors (residuals) in a regression model is not constant across all levels of the independent variable(s).
- In other words, the spread of the residuals varies systematically as a function of the predictors.

ii. **The Breusch-Pagan test** is used to detect heteroscedasticity in a regression model by examining whether the variance of the residuals is constant across different levels of the independent variables.

- The test involves fitting an auxiliary regression of the squared residuals on the independent variables and testing the significance of the coefficients.
- $$\begin{cases} H_0: Constant\ variance\ of\ errors\ (homoscedasticity) \\ H_a: Non-constant\ variance\ of\ errors\ (heteroscedasticity) \end{cases}$$

10. How to Check for the heteroscedasticity?

- A small p-value suggests rejection of the null hypothesis of homoscedasticity.
- If the F-statistic is significantly different from the critical value or if the p-value is below the chosen significance level (5% in our case), you may reject the null hypothesis of homoscedasticity, suggesting the presence of heteroscedasticity.

10. How to Check for the heteroscedasticity?

- If heteroscedasticity is detected, consider using techniques such as:

i. Weighted least squares regression (WLS),

ii. Transforming the dependent variable,

- Transforming the dependent variable in the context of addressing heteroscedasticity means applying a mathematical transformation to the variable itself. The goal is to stabilize the variance of the residuals and make it more constant across different levels of the independent variable(s). This can help meet one of the assumptions of linear regression, which is homoscedasticity (constant variance of residuals).

10. How to Check for the heteroscedasticity?

- **Logarithmic Transformation:** Taking the natural logarithm of the dependent variable (*Y*) is a common transformation. Transformed Y=ln(Y).
- **Square Root Transformation:**The square root transformation is another option to stabilize the variance.
- Transformed $Y = \sqrt{y}$
- This transformation is appropriate when the variance of *Y* increases with the level of *X*, but not exponentially.
- It's essential to address heteroscedasticity to obtain valid statistical inferences from your regression model.

11. How to Check for the Autocorrelation?

iii. **Autocorrelation**, also known as serial correlation, refers to the correlation of a time series with its own past or future values. It is commonly checked in time series data to identify patterns or dependencies over time.

- Simply put, the model assumes that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic.

11. How to Check for the Autocorrelation?

- The null and alternative hypotheses for the Durbin-Watson test are as follows:
- **Null Hypothesis (*H*0):There is no first-order autocorrelation in the residuals.**
- **Alternative Hypothesis (**H1**):There is first-order autocorrelation in the residuals.**

11. How to Check for the Autocorrelation?

- The Durbin-Watson test statistic is computed using the following formula:
- $d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t{}^2}$

ε_t is the residual at time t,*n* is the number of observations.

In this formula, the numerator represents the sum of squared differences between consecutive residuals, and the denominator represents the sum of squared residuals. The test statistic *d* ranges from 0 to 4.

11. How to Check for the Autocorrelation?

- The test statistic, denoted as *d*, ranges between 0 and 4.

a. Value of *d* close to 2 indicates no first-order autocorrelation (null hypothesis not rejected).

b. Values significantly less than 2 suggest positive autocorrelation,

c. Values significantly greater than 2 suggest negative autocorrelation.

- $d \approx 2$: No evidence of first-order autocorrelation.
- $d < 2$: Evidence of positive autocorrelation.
- $d > 2$: Evidence of negative autocorrelation.

11. How to Check for the Autocorrelation?

- If autocorrelation is detected, consider these technique to correct it:

i. Incorporating **lagged values** into your model;

ii. Differencing data involves taking the first difference of the dependent variable or the independent variables. This is often done in time series analysis. Differencing can help remove the trend and make the data more stationary, reducing autocorrelation.

- If autocorrelation is not addressed, it can affect the reliability of statistical inferences.

12. Example

Time period	GDP growth rate (%)	Unemployment rate (%)	Inflation rate (%)
2010	2,1	6,5	2,5
2011	2,4	6,2	3
2012	2,7	5,8	2,8
2013	2,9	5,5	2,2
2014	2,6	5	2,5
2015	2,5	4,5	2
2016	3	4,9	2,5
2017	2,8	4,2	2,2
2018	2,2	4,1	1,8
2019	2,5	3,9	2
2020	3,1	4	1,5
2021	3,6	3,5	2,8
2022	3,2	3	2,5
2023	2,8	3,2	2

12. Example

- GDP Growth Rate" represents the percentage change in Gross Domestic Product (GDP) from the previous year.
- "Unemployment Rate" represents the percentage of the labor force that is unemployed and actively seeking employment.
- "Inflation Rate" represents the percentage change in the general price level of goods and services.