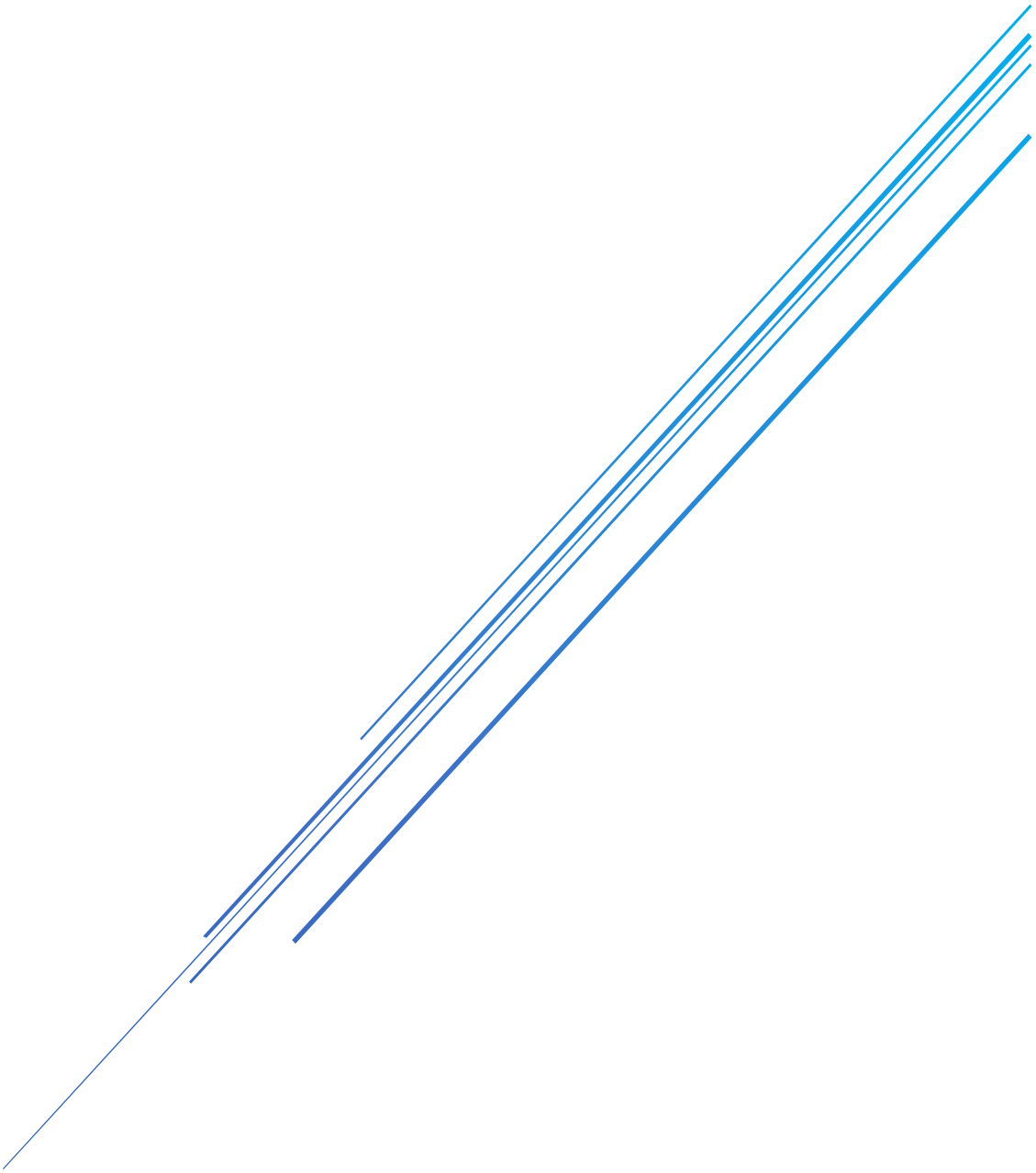
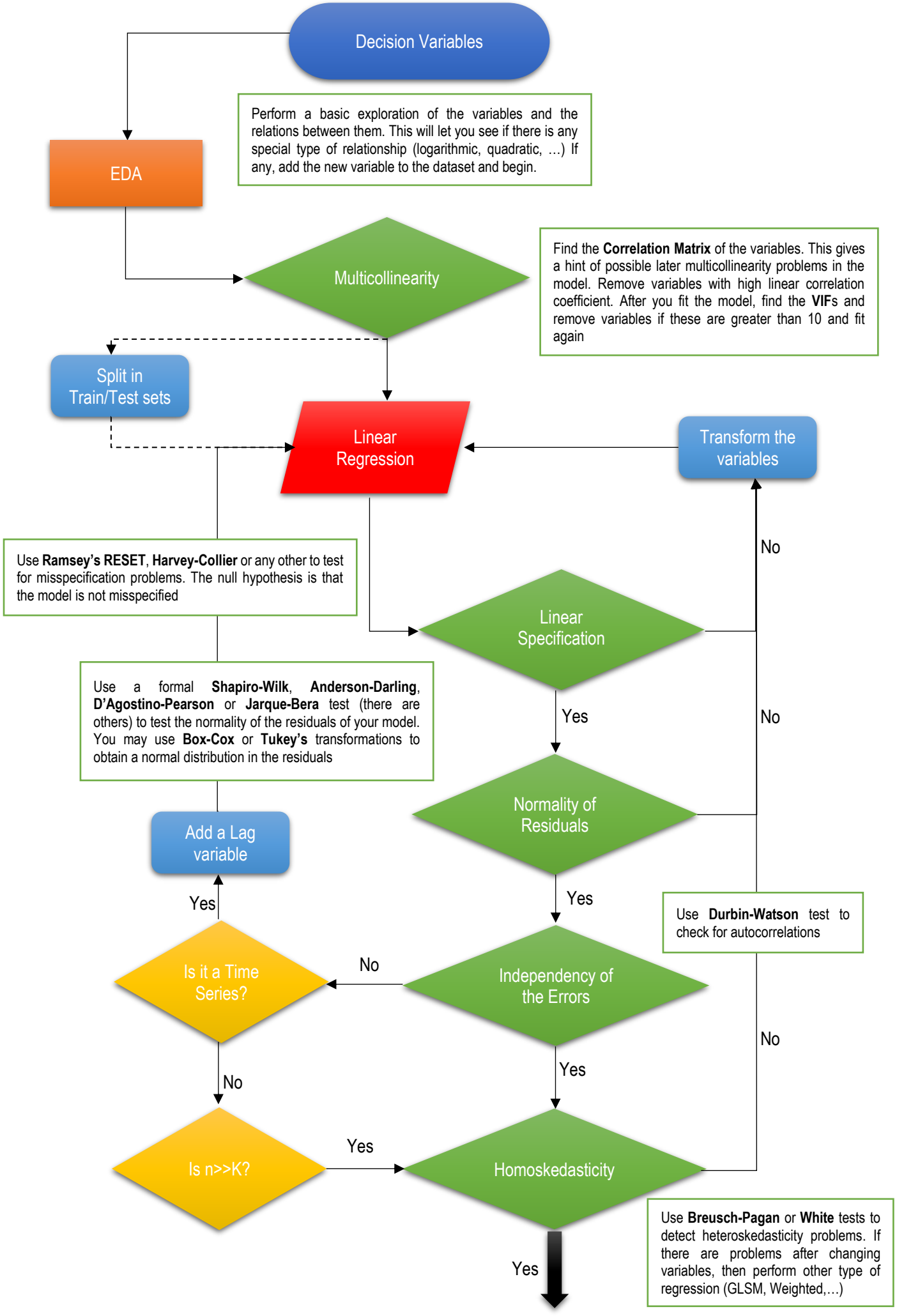
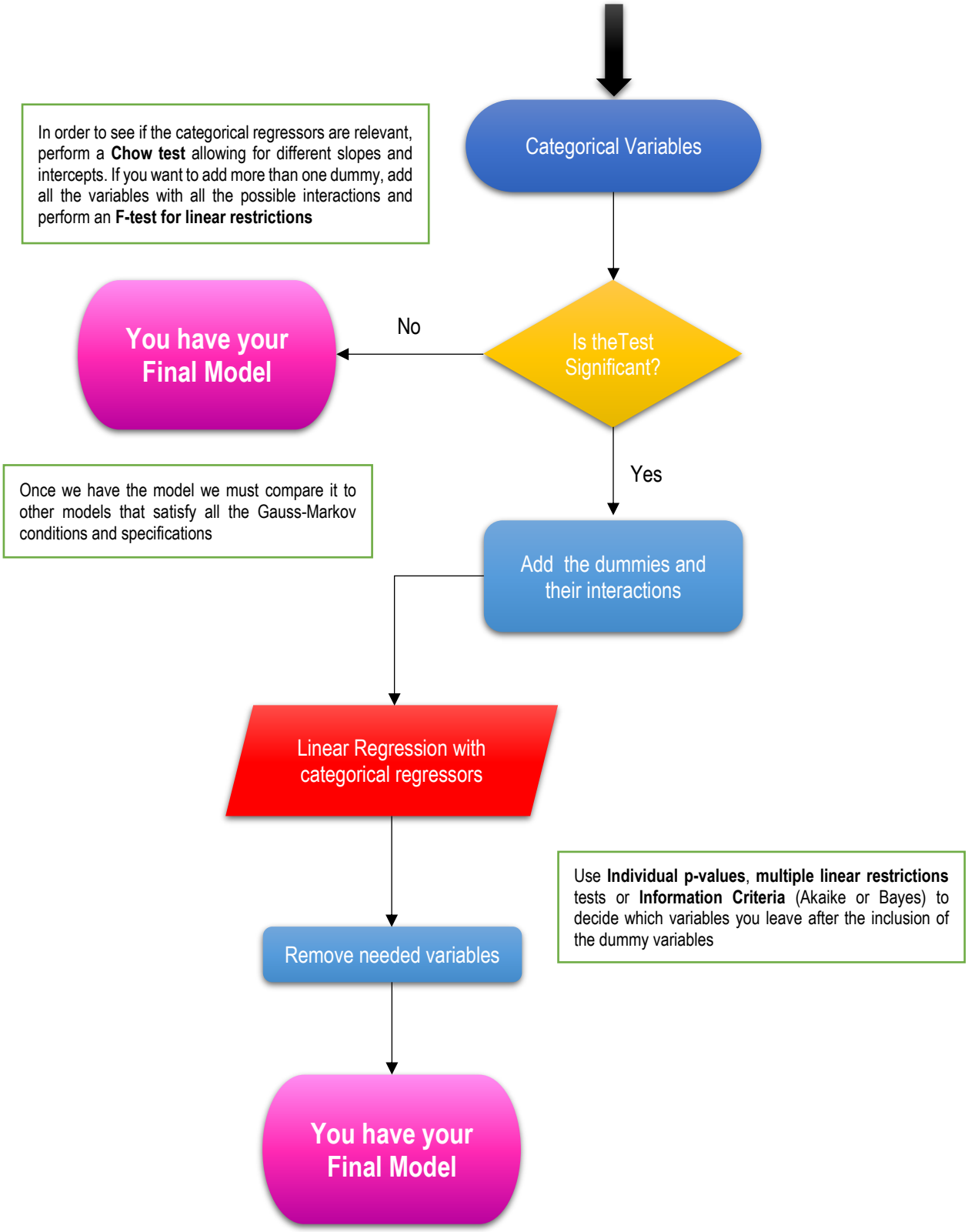
LINEAR REGRESSION WORKFLOW

Prof. Dr. Juanjo Manjarín



IE Exponential Learning
Part Time Bootcamp, 2019


Decision Variables
Perform a basic exploration of the variables and the relations between them. This will let you see if there is any special type of relationship (logarithmic, quadratic, …) If any, add the new variable to the dataset and begin.
EDA
Multicollinearity
Find the **Correlation Matrix** of the variables. This gives a hint of possible later multicollinearity problems in the model. Remove variables with high linear correlation coefficient. After you fit the model, find the **VIF**s and remove variables if these are greater than 10 and fit again
Split in Train/Test sets
Linear Regression
Transform the variables
No
Use **Ramsey's RESET**, **Harvey-Collier** or any other to test for misspecification problems. The null hypothesis is that the model is not misspecified
Linear Specification
Use a formal **Shapiro-Wilk**, **Anderson-Darling**, **D'Agostino-Pearson** or **Jarque-Bera** test (there are others) to test the normality of the residuals of your model. You may use **Box-Cox** or **Tukey's** transformations to obtain a normal distribution in the residuals
Yes
No
Normality of Residuals
Add a Lag variable
Yes
Yes
Use **Durbin-Watson** test to check for autocorrelations
Is it a Time Series?
No
Independency of the Errors
No
No
Yes
Is n>>K?
Yes
Homoskedasticity
Use **Breusch-Pagan** or **White** tests to detect heteroskedasticity problems. If there are problems after changing variables, then perform other type of regression (GLSM, Weighted,…)
Yes


Categorical Variables
In order to see if the categorical regressors are relevant, perform a **Chow test** allowing for different slopes and intercepts. If you want to add more than one dummy, add all the variables with all the possible interactions and perform an **F-test for linear restrictions**
Is theTest Significant?
No
You have your Final Model
Yes
Once we have the model we must compare it to other models that satisfy all the Gauss-Markov conditions and specifications
Add the dummies and their interactions
Linear Regression with categorical regressors
Use **Individual p-values**, **multiple linear restrictions** tests or **Information Criteria** (Akaike or Bayes) to decide which variables you leave after the inclusion of the dummy variables
Remove needed variables
You have your Final Model
Once you have the final model, and probably some others, you may need to perform tests as **Davidson-MacKinnon** or **White-Davidson-MacKinnon** in order to see if a model with logarithmic variables is better than other with linear variables.