AEC3 OF DECISION-MAKING MODELS

MULTIPLE LINEAR REGRESSION, LOGISTIC REGRESSION & TIME SERIES

Name:

Instructions

- The test part consists of 25 multiple-choice questions.
- The right questions add up to 1 point.
- Incorrect questions subtract 0.33.
- Blank questions neither add nor subtract.
- The student must work in a .ipynb (Jupyter Notebook) file.
- Copy all the original questions and answer them in Markdown cells within the notebook.
- To submit the work, the student must send both the .ipynb file and the PDF generated directly from Visual Studio Code. To generate the PDF: click the three-dot button in the top right corner of the notebook, select "Export", and then choose "PDF".
- File naming convention: Name both files using the following format: AEC2_surname_name.ipynb and AEC3_surname_name.pdf (Replace surname and name with your actual last and first names.)

Multivariate Analysis and Multiple Linear Regression

1. Which of the following statements accurately describes dependency methods?

   a) Dependency methods analyze variables based on their interrelationships and determine their causality.
   b) Dependency methods assume that all variables are independent of each other.
   c) Dependency methods focus on determining the impact of independent variables on dependent variables.
   d) Dependency methods only consider the influence of dependent variables on independent variables.

2. Fill out the blanks in the next sentence: Multiple Linear Regression analyses the relationship between _____ and _____.

   a) One non-metric dependent variable, several metric independent variables.
   b) One metric dependent variable, several metric or non-metric independent variables.
   c) Several metric dependent variables, one metric independent variable.
   d) Several metric or non-metric dependent variables, one metric independent variable.

3. In the next equation, what are e_i?

$$Yi = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i$$

   a) The differences between the observed values and the estimated values of the model.
   b) The average effect of the increase in one unit of the predictor “i” on the variable “Y”.
   c) The ordinate at the origin, the value of the variable Y when all predictors are zero.
   d) None of the others.

4. What is the purpose of using dummy variables for categorical predictors when they have more than two levels in regression analysis?

   a) Dummy variables represent categorical predictors numerically, converting each class to a number inside the dataset.

b) Dummy variables are the average percentage by which each level influences the dependent variable Y compared to the reference level of the predictor.
c) Dummy variables help eliminate the influence of categorical predictors on the dependent variable because they can’t be included in the model if they are non-metric.
d) Dummy variables are used to create interactions between categorical predictors.

5. Why do we use adjusted R^2 in multiple linear models?

a) Adjusted R2 penalizes the number of predictors included in the model based on sample size.
b) Adjusted R2 increases as the number of predictors in the model increases.
c) Adjusted R2 is used to directly compare models with different numbers of predictors.
d) Adjusted R2 helps find the model that explains the variability of the dependent variable with the least number of predictors.

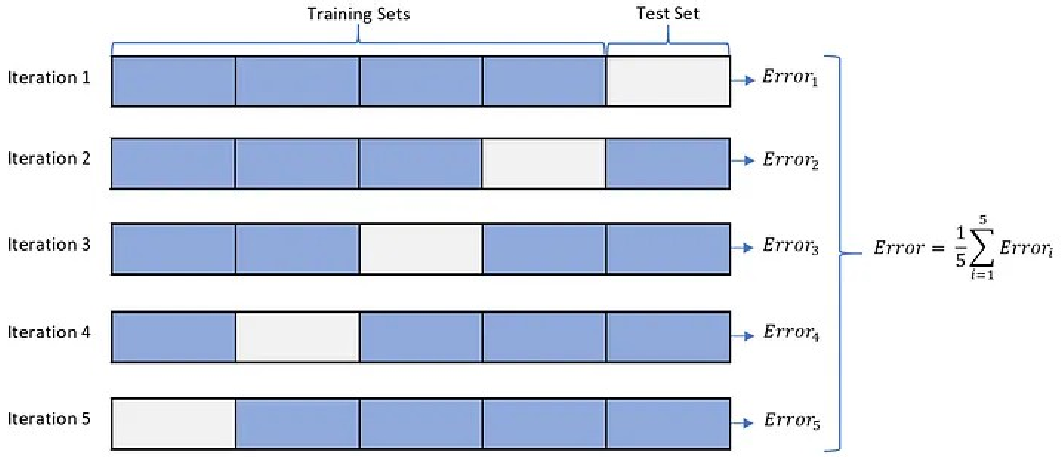
6. Which of the following methods can be used for predictor selection in regression analysis?

   a) Hierarchical method.
   b) Forced input method.
   c) Stepwise method.
   d) All of the others.

7. What is the use of model validation in regression analysis?

   a) Model validation ensures that the selected model has the highest possible complexity.
   b) Model validation evaluates the performance of the model on new, unseen observations.
   c) Model validation adjusts the model using all observations from the dataset.
   d) Model validation determines the optimal number of predictors to include in the model.

8. Which model validation is the one represented in the following picture?



   a) Simple Validation.
   b) Monte Carlo Cross-Validation.
   c) Leave One Out Cross-Validation.
   d) K-Fold Cross-Validation.

9. Why do we use Root Mean Square Error (RMSE) in model validation?

   a) RMSE measures the complexity of the model.
   b) RMSE evaluates the generalization ability of the model.
   c) RMSE quantifies the goodness of fit of the model.
   d) RMSE determines the optimal number of predictors in the model.

Logistic Regression

10. What is the primary purpose of logistic regression?

    a) To estimate the average value of a continuous dependent variable.
    b) To analyze the relationship between two continuous variables.
    c) To predict the probability of an event (binary dependent variable) occurring based on independent variables.
    d) To identify outliers in a dataset.

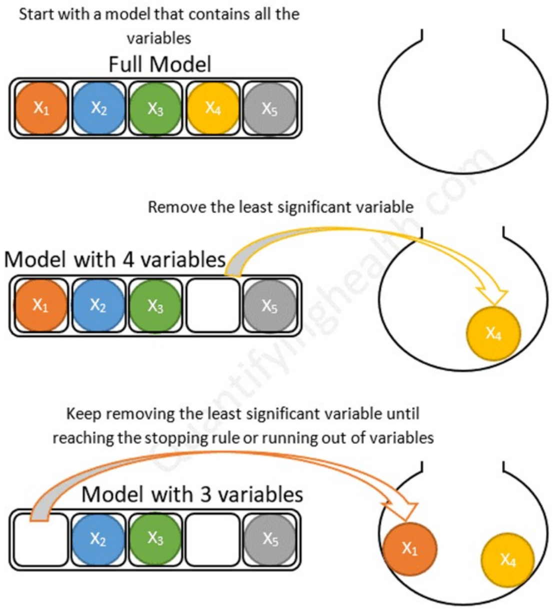
11. Select the correct answer about the binomial distribution:

    a) It has an infinite number of possible outcomes.
    b) It represents situations with only two possible events.
    c) It is used to study continuous variables.
    d) It requires a constant probability across all observations.

12. Which of the following statements accurately describes the use of the logit transformation in logistic regression?

    a) It converts continuous variables into categorical variables.
    b) It creates a sigmoidal function to model Bernoulli distributions.
    c) The logistic regression model doesn't use the logit transformation.
    d) It standardizes the coefficients of the independent variables.

13. We use a procedure illustrated in the following picture to select the predictors of a logistic regression. What is its name?


Start with a model that contains all the variables
Full Model
X_1
X_2
X_3
X_4
X_5
Remove the least significant variable
Model with 4 variables
X_1
X_2
X_3
X_5
X_4
Keep removing the least significant variable until reaching the stopping rule or running out of variables
Model with 3 variables
X_2
X_3
X_5
X_1
X_4

a) It is the backward procedure.
b) It is the forward procedure.
c) It is the hierarchical procedure.
d) It is the double or mixed procedure.

14. How do we measure the goodness of fit in the logistic regression?

a) With the R^2.
b) With the accuracy.
c) With the adjusted R^2.
d) With the confusion matrix.

15. Which one is NOT a condition of applicability of the logistic regression?

a) Normality of the Pearson residuals.
b) Constant variance when the mean changes.
c) Presence of influential values.
d) Balanced samples.

16. How does multicollinearity impact the interpretation of coefficients in statistical models?

a) It makes the interpretation of coefficients unreliable.
b) It enhances the interpretability of coefficients.
c) It has no effect on the interpretation of coefficients.
d) It reduces the need for interpreting coefficients.

17. What is one approach to address the problem of unbalanced samples in classification tasks?

a) Increase the number of successes in the sample.
b) Decrease the number of non-successes in the sample.
c) Ignore the imbalance and proceed with the original cut-off point.
d) Adjust the probability cut-off point for classification decisions.

<u>Time Series</u>

18. What distinguishes time series data from cross-sectional data?

a) Time series data involves observations taken at multiple points in time for a single entity.
b) Time series data consists of qualitative observations about multiple entities taken at a single point in time.
c) Time series data is used to study the characteristics of the response variable concerning other independent variables.
d) Time series analysis is only used for forecasting future events and cannot be applied to historical data.

19. What are the two main tasks concerning time series data in data mining or machine learning processes?

a) The analysis of time series and the analysis of cross-sectional data.
b) The analysis of time series and time series forecasting.
c) The analysis of time series and data visualization techniques.
d) The analysis of time series and hypothesis testing.

20. What is panel data, also known as longitudinal data?

a) Data taken from multiple individuals or instances at a single time point.
b) Data taken from an individual entity at multiple points in time.
c) Data taken from multiple entities at multiple time points.
d) Data taken from multiple individuals or instances without considering the time variable.

21. What is the difference between heteroscedasticity and homoscedasticity?

a) Heteroscedasticity refers to a constant error variance on the independent variable, while homoscedasticity refers to a varying error variance.
b) Heteroscedasticity refers to a linear relationship between the independent and dependent variables, while homoscedasticity refers to a non-linear relationship.
c) Heteroscedasticity refers to a constant mean of the dependent variable, while homoscedasticity refers to a varying mean.
d) Heteroscedasticity refers to a varying error variance on the independent variable, while homoscedasticity refers to a constant error variance.

22. Which time series component involves regular or fixed interval shifts within the dataset?

a) Trend.
b) Seasonality.
c) Cyclical.
d) Irregularity.

23. What characterizes the fluctuations caused by random or irregular variations in time series data?

a) They are predictable and controllable.
b) They follow a regular pattern.

c) They are stochastic and unpredictable.
d) They can be accurately modeled mathematically.

24. What is the purpose of differencing in time series analysis?

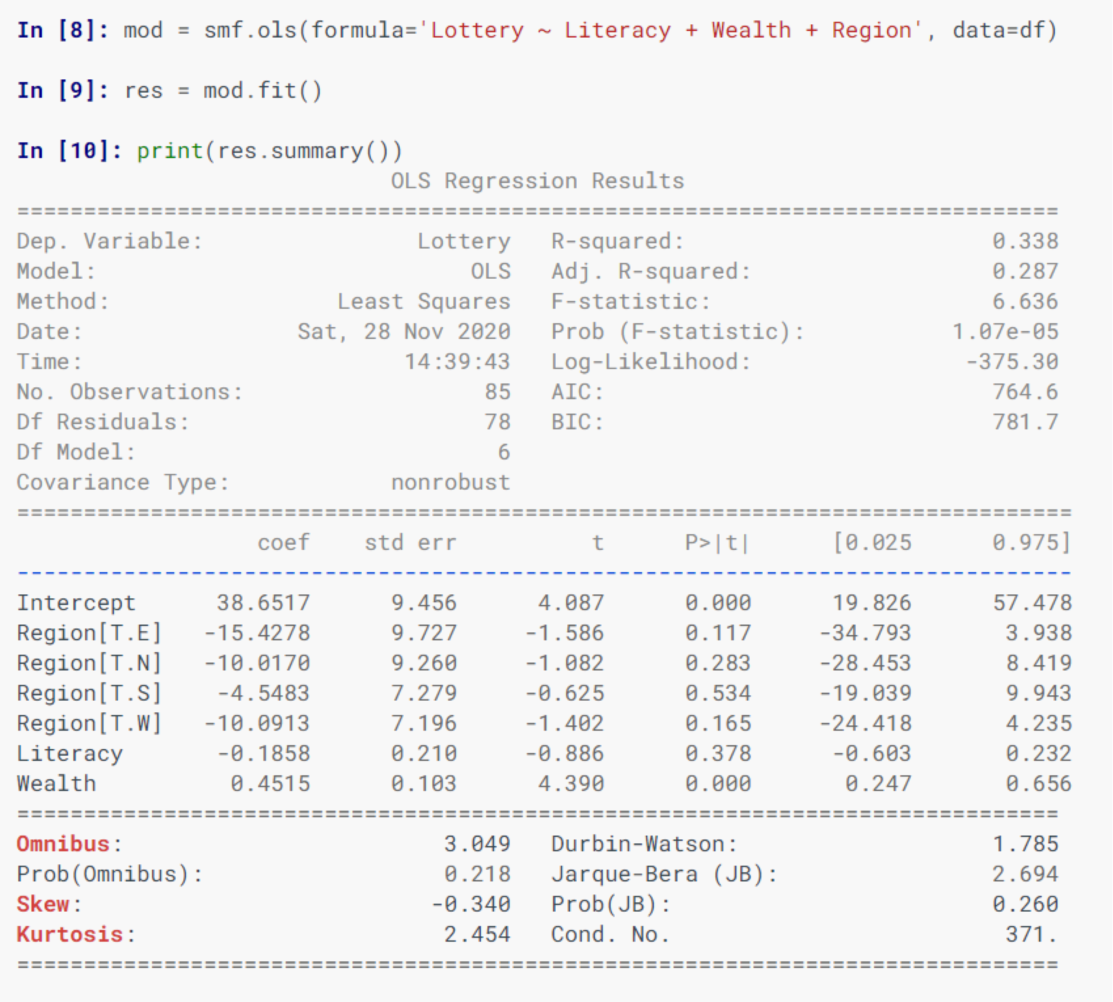
a) To introduce time dependence in the series.
b) To stabilize the mean of the time series.
c) To increase the trend and seasonality in the series.
d) To remove random fluctuations in the series.

25. Select the correct answer:

a) A pure autoregressive model (AR) is a model that depends on solely from its lags. To know if a lag is needed, we use the Auto-correlation Function (ACF).
b) A pure autoregressive model (AR) is a model that depends solely on the lags of the prediction errors. To know if a lag is needed, we use the Partial Correlations Function (PACF).
c) A pure autoregressive model (AR) is a model that depends solely on the lags of the prediction errors. To know if a lag is needed, we use the Auto-correlation Function (ACF).
d) A pure autoregressive model (AR) is a model that depends on solely from its lags. To know if a lag is needed, we use the Partial Correlations Function (PACF).

PROBLEMS (4 points)

1. **(2 points)** We have performed a multiple linear regression to determine per capita wager in the Royal Lottery of 1830's France using a few characteristics:

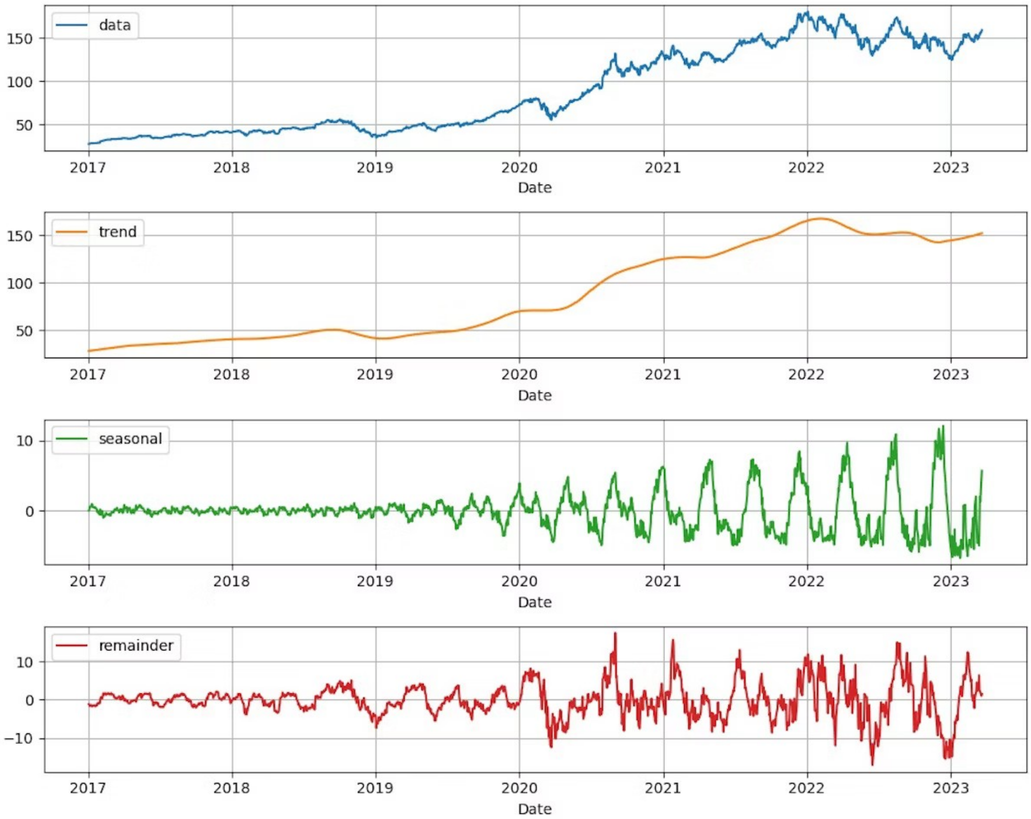


```
In [8]: mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)

In [9]: res = mod.fit()

In [10]: print(res.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     6.636
Date:                Sat, 28 Nov 2020   Prob (F-statistic):           1.07e-05
Time:                        14:39:43   Log-Likelihood:                -375.30
No. Observations:                  85   AIC:                             764.6
Df Residuals:                      78   BIC:                             781.7
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      38.6517      9.456      4.087      0.000      19.826      57.478
Region[T.E]   -15.4278      9.727     -1.586      0.117     -34.793       3.938
Region[T.N]   -10.0170      9.260     -1.082      0.283     -28.453       8.419
Region[T.S]    -4.5483      7.279     -0.625      0.534     -19.039       9.943
Region[T.W]   -10.0913      7.196     -1.402      0.165     -24.418       4.235
Literacy       -0.1858      0.210     -0.886      0.378      -0.603       0.232
Wealth          0.4515      0.103      4.390      0.000       0.247       0.656
==============================================================================
Omnibus:                        3.049   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.694
Skew:                          -0.340   Prob(JB):                        0.260
Kurtosis:                       2.454   Cond. No.                         371.
==============================================================================
```

a) (1.5 points) Interpret these results and all you have learnt from them.

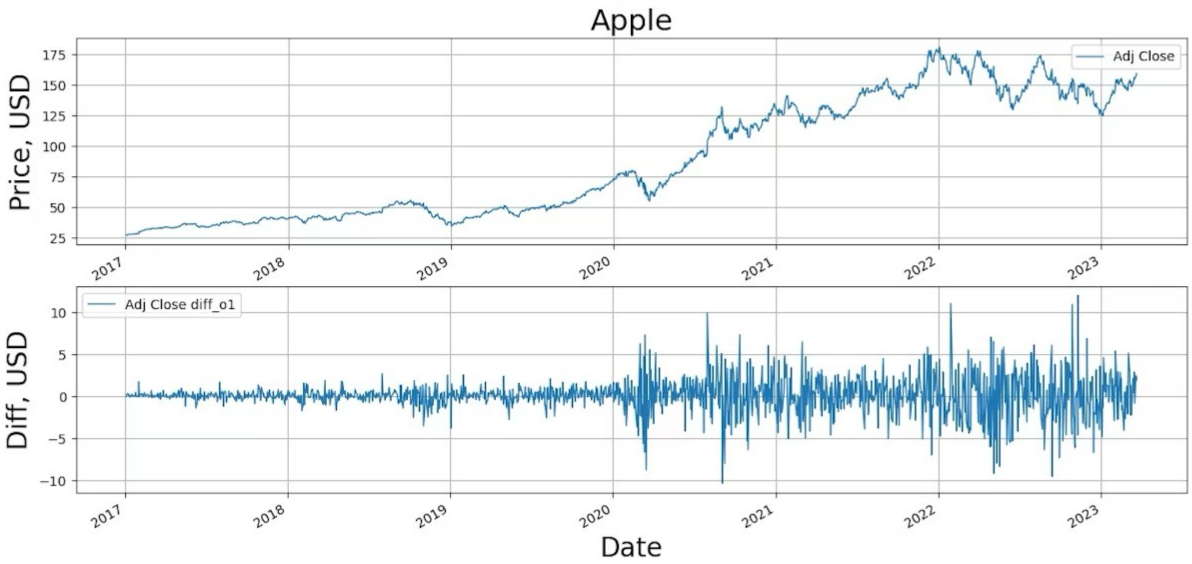
b) (0.5 points) If your objective is to create a prediction model, justify why this model would be valid or, if not, explain how you would obtain a valid one.

2. **(2 points)** Analyzing a time series data, we have decomposed it into the following components:



a) Explain these components and describe how this specific time series behaves in regards of them.

b) Explain which transformation has been applied in the following graphic, in what consists and what is the purpose of it:



c) If your purpose is to forecast this univariate time series data, which model of the ones seen at class would you choose to use and why?

.