AEC3_surname_name.ipynb

This notebook contains the resolution of the AEC3 test as per the instructions. All questions have been copied and answered in Markdown cells. Problems are answered with detailed justifications.

Name: Pablo Bas Genís

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

Answer: c) Dependency methods focus on determining the impact of independent variables on dependent variables.

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

Answer: b) One metric dependent variable, several metric or non-metric independent variables.

3. In the next equation, what are ei? $Yi = (\beta 0 + \beta 1X1i + \beta 2X2i + \cdots + \beta nXni) + ei$

Answer: a) The differences between the observed values and the estimated values of the model.

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

Answer: b) Dummy variables are the average percentage by which each level influences the dependent variable Y compared to the reference level of the predictor.

5. Why do we use adjusted R² in multiple linear models?

Answer: a) Adjusted R² penalizes the number of predictors included in the model based on sample size.

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

Answer: d) All of the others.

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

Answer: b) Model validation evaluates the performance of the model on new, unseen observations.

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

Answer: d) K-Fold Cross-Validation.

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

Answer: c) RMSE quantifies the goodness of fit of the model.

10. What is the primary purpose of logistic regression?

Answer: c) To predict the probability of an event (binary dependent variable) occurring based on independent variables.

11. Select the correct answer about the binomial distribution:

Answer: b) It represents situations with only two possible events.

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

Answer: b) It creates a sigmoidal function to model Bernoulli distributions.

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

Answer: d) It is the double or mixed procedure.

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

Answer: d) With the confusion matrix.

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

Answer: b) Constant variance when the mean changes.

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

Answer: a) It makes the interpretation of coefficients unreliable.

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

Answer: d) Adjust the probability cut-off point for classification decisions.

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

Answer: a) Time series data involves observations taken at multiple points in time for a single entity.

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

Answer: b) The analysis of time series and time series forecasting.

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

Answer: c) Data taken from multiple entities at multiple time points.

21. What is the difference between heteroscedasticity and homoscedasticity?

Answer: 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?

Answer: b) Seasonality.

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

Answer: c) They are stochastic and unpredictable.

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

Answer: b) To stabilize the mean of the time series.

25. Select the correct answer:

Answer: 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).

Problem 1

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

- **Intercept (20.5)**: This is the estimated per capita wager when all predictors are 0, although this value is often not meaningful by itself.
- **Population Density (0.03)**: A unit increase in population density is associated with a €0.03 increase in per capita wager, holding other variables constant. The effect is statistically significant (p = 0.005).
- **Average Income (1.25)**: For each additional unit (e.g., €1,000) in average income, the wager increases by €1.25 per capita. This is highly significant (p < 0.001), indicating a strong relationship.
- **% Urban Population (-0.40)**: More urbanized areas are associated with a decrease in wagering per capita, possibly due to access to alternative leisure activities.
- **R² (0.72)** indicates that 72% of the variability in wagering is explained by the model.
- **Adjusted R² (0.69)** slightly penalizes model complexity and still shows a strong explanatory power.

---

b) Justify if this model would be valid or how you would obtain a valid one.

Validity Assessment:

- **Linearity**: Check linearity assumption via residual plots. If patterns exist, transformation may be needed.
- **Homoscedasticity**: Residuals should have constant variance.

A valid prediction model would pass these diagnostics and have good performance in cross-validation (e.g., low RMSE).

Problem 2

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

A time series typically has four components:

- **Trend**: Long-term progression of the series. E.g., increasing lottery sales over years.
- **Seasonality**: Repeating short-term cycles, often linked to calendar periods (e.g., holiday peaks).
- **Cyclicality**: Medium- or long-term fluctuations not tied to season, like economic cycles.
- **Irregular component (noise)**: Random variation not explained by the above.

In the decomposition figure:

- If trend is increasing, the data has a positive overall progression.
- Seasonality appears as repeated peaks/dips annually.
- Noise appears in the residuals after removing trend/seasonality.

---

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

The plot likely shows a **differenced time series**.

- **Differencing** removes trend or seasonality by computing the difference between consecutive observations.
- If applied once: removes linear trend.
- If applied twice: removes quadratic trend.
- Purpose: **to achieve stationarity**.

---

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

Chosen model: SARIMA (Seasonal ARIMA)

- **Why?** Because:
  - It handles both non-seasonal and seasonal components.
  - Suitable for univariate time series with trend + seasonality.
  - Supports differencing to enforce stationarity.
  - Flexible and interpretable.