Bar Code

اسم الطالب

رقم الطالب

Cairo University
Faculty of Computers and Artificial Intelligence

Final Exam
Model (A)

Department: Operations Research and Decision Support

Course Title: Data Analytics
Course Code: DS342
Semester: Fall 2022-2023
Instructor: Dr. Marwa Sabry
Date: 19th January 2023
Exam Duration: 2 Hours

تعليمات هامة

- حيازة التليفون المحمول مفتوحا داخل لجنة الإمتحان يعتبر حالة غش تستوجب العقاب وإذا كان ضرورى الدخول بالمحمول فيوضع مغلق فى الحقائب.
- لا يسمح بدخول سماعة الأذن أو البلوتوث.
- لايسمح بدخول أي كتب أو ملازم أو أوراق داخل اللجنة والمخالفة تعتبر حالة غش.

	60

Question	Mark	Signature
One		
Two		
Three		
Four		
Five		
Six		
Seven		
Eight		
Nine		
Ten		
Total Marks		

Total Marks in Writing: _____

Part I – True or False ***[Each 0.5 mark - Total 10 marks]***

***In the "MCQ / T or F" section of the bubble sheet*, indicate whether the below statements are true or false.**

1. In multiple regression, if there is multicollinearity between independent variables, the t-tests of the individual coefficients may indicate that some variables are not linearly related to the dependent variable, when in fact, they are. T
2. A relatively new aspect of business analytics is big data, which typically implies the analysis of the very large data sets that companies currently encounter. T
3. In testing the overall fit of a multiple regression model in which there are three explanatory variables, the null hypothesis is H_0: $\beta_1 = \beta_2 = \beta_3$. F
4. To help explain or predict the response variable in every regression study, we use one or more explanatory variables. These variables are also called response variables or independent variables. F
5. If we use a value close to 1 for the smoothing constant α in a simple exponential smoothing model, then we expect the model to respond very slowly to changes in the level. F
6. In a simple linear regression problem, suppose that $\sum e_i^2 = 12.48$ and $\sum(Y_i - \overline{Y})^2 = 124.8$. Then R^2 = 0.90. T
7. Cross-sectional data are usually data gathered across different periods of time from a population. F
8. A regression analysis between X = sales (in $1000s) and Y = advertising (in $) resulted in the following least squares line: Y = 32 + 8X. This implies that an increase of $1 in advertising is expected to result in an increase of $8 in sales. F
9. Scatterplots are used for identifying outliers and indicating what you should do about the outliers you may find. F
10. In regression analysis, the total variation in the dependent variable Y, measured by $\sum(Y_i - \overline{Y})^2$ and referred to as SST, can be decomposed into two parts: the explained variation, measured by SSR, and the unexplained variation, measured by SSE. T
11. If a time series exhibits an exponential trend, then a plot of its logarithm should be approximately linear. T

12. A useful graph in almost any regression analysis is a scatterplot of residuals (on the vertical axis) versus fitted values (on the horizontal axis), where a "good" fit not only has small residuals, but it has residuals scattered randomly around zero with no apparent pattern. T
13. Data analysis includes data description, data visualization, data inference, and the search for relationships in data. T
14. To calculate the five-period moving average for a time series, we average the values in the two preceding periods, and the values in the three following time periods. F
15. Holt's method is an exponential smoothing method, which is appropriate for a series with seasonality and possibly a trend. F
16. The two primary objectives of regression analysis are to study relationships between variables and to use those relationships to make predictions. T
17. The most common form of autocorrelation is positive autocorrelation, where large observations tend to follow large observations and small observations tend to follow small observations. T
18. Heteroscedasticity means that the variability of Y values is larger for some X values than for others. T
19. Spreadsheet modeling is the process of entering the outputs into a spreadsheet and then relating them appropriately, by means of formulas, to obtain the decision variables. F
20. In a simple linear regression problem, if = 0.95, this means that 95% of the variation in the explanatory variable X can be explained by the regression. F

Part II – Multiple Choice Questions (MCQ) ***[Each 0.5 mark - Total 10 marks]***
In the "MCQ / T or F" section of the bubble sheet, **fill in the letter of the choice that best completes each sentence.**

21. A "fan" shape in a scatterplot indicates
    a. **unequal variance.** b. a nonlinear relationship.
    c. the absence of outliers. d. sampling error.

22. Suppose you try to create a pivot table from multiple tables stored in a Data Model and, when you check a particular field to be placed in the pivot table, you get a warning about a missing relationship. Which of the following does this not imply?
    a. **You are trying to use fields from unrelatable tables, so no matter what you do, your pivot table results will be wrong.**
    b. You can open the Power Pivot window, create the missing relationships, and try building the pivot table again.
    c. You can click the Auto-Detect button in the PivotTable Fields pane, and the chances are that Excel will correctly create the missing relationship.
    d. You can click the CREATE button in the PivotTable Fields pane, which allows you to manually create the missing relationship.

23. In linear regression, the fitted value is
    a. **the predicted value of the dependent variable.**
    b. the predicted value of the independent value.
    c. the predicted value of the slope.
    d. the predicted value of the intercept.

24. The standard error of the estimate (s_e) is essentially the
    a. mean of the residuals. b. **standard deviation of the residuals.**
    c. mean of the explanatory variable. d. standard deviation of the explanatory variable.

25. The forecast error is the difference between
    a. this period's value and the next period's value.
    b. the average value and the expected value of the response variable.
    c. the explanatory variable value and the response variable value.
    d. **the actual value and the forecast value.**

26. Categorizing an age variable as "young," "middle-aged," and "elderly" is an example of
    a. counting. b. ordering. c. quantifying. d. **binning**.

27. Which of the following is ***not*** one of the assumptions of regression?
    a. There is a population regression line that joins the means of the dependent variable for all values of the explanatory variables.
    b. The response variable is normally distributed.
    c. **The standard deviation of the response variable increases as the explanatory variables increase.**
    d. The errors are probabilistically independent.

28. Which summary measure for forecast errors does not depend on the units of the forecast variable?
    a. MAE (mean absolute error) b. MFE (mean forecast error)
    c. RMSE (root mean square error) d. **MAPE (mean absolute percentage error)**

29. The data below represents sales for a particular product. If you were to use the moving average method with a span of 3 periods, what would be your forecast for ***period* 5**?

Period	Sales (in units)
1	90
2	120
3	110
4	100

    a. 90 b. 100 c.105 d.**110**

30. In a simple linear regression analysis, the following sums of squares are produced:

$$\sum (Y-\bar{Y})^2 = 400,\ \sum (Y-\hat{Y})^2 = 80,\ \sum (\hat{Y}-\bar{Y})^2 = 320$$

    The proportion of the variation in *Y* that is explained by the variation in *X* is
    a. 20%. b. **80%.** c. 25%. d. 50%.

31. Which of the following is *not* a method for dealing with seasonality in data?
    a. Winter's exponential smoothing model
    b. De-seasonalizing the data, using any forecasting model, then re-seasonalizing the data
    c. **Multiple regression with lags for the seasons**
    d. Multiple regression with dummy variables for the seasons

32. When you run a query to import external data into Excel, there are several options as to where to store the data. Which of the following is not one of those options?
    a. The data can be stored into an Excel table.
    b. The data can be stored into a pivot table (or pivot chart) report.

c. The data can be stored as a Data Model.
d. **The data can be stored as a Power Model.**

33. The adjusted R^2 adjusts R^2 for
a. non-linearity. b. outliers. c. low correlation.
d. **the number of explanatory variables in a multiple regression model.**

34. Suppose that a simple exponential smoothing model is used (with $\alpha = 0.40$) to forecast monthly sandwich sales at a local sandwich shop. The forecasted demand for September was 1560 and the actual demand was 1480 sandwiches. Given this information, what would be the forecast number of sandwiches for October?
a. 1480 b. **1528** c. 1560 d. 1592

35. Winters' model differs from Holt's model and simple exponential smoothing in that it includes an index for
a. **seasonality**. b. trend. c. residuals. d. cyclical fluctuations.

36. Which of the following assumptions do we make while deriving linear regression parameters?
1. The true relationship between dependent ***y*** and predictor ***x*** is linear
2. The model errors are statistically independent
3. The errors are normally distributed with a 0 mean and constant standard deviation
4. The predictor x is non-stochastic and is measured error-free

a. 1,2 and 3. b. 1,3 and 4. c. 1 and 3. d. **All of above.**

37. Which of the following is the relevant sampling distribution for regression coefficients?
a. Normal distribution
b. *t*-distribution with *n-1* degrees of freedom
c. ***t*-distribution with *n-1-k* degrees of freedom**
d. *F*-distribution with *n-1-k* degrees of freedom

38. The ANOVA table splits the total variation into two parts. They are the _____ variation.
a. acceptable and unacceptable b. adequate and inadequate
c. resolved and unresolved d. **explained and unexplained**

39. A researcher can check whether the errors are normally distributed by using:
a. a frequency distribution b. the Durbin-Watson statistic
c. a t-test or an F-test d. **a histogram or a Q-Q plot**

40. Suppose you run a regression of a person's height on his/her right and left foot sizes, and you suspect that there may be multicollinearity between the foot sizes. What types of problems might you see if your suspicions are true?
a. "Wrong" values for the coefficients for the left and right foot size
b. Small *p*-values for the coefficients for the left and right foot size
c. **Small *t*-values for the coefficients for the left and right foot size**
d. Large *t*-values for the coefficients for the left and right foot size

Part III – Problems – Multiple Choice Questions (MCQ)

For the next four problems, refer to the folder located on your desktop named "DS342 - Final Exam Data". ***In the "MCQ / T or F" section of the bubble sheet, fill in the letter of the choice that provides the solution to the statement.***

Problem 1 ***[Each 2 mark - Total 10 marks]***

Refer to the data file named "Company's Sales 2018.xlsx", it is an Excel file with four related tables: Orders, Orders Details, Products, Product Categories.

These tables are not currently related, but they are relatable through key "ID" fields. Create pivot tables and check the option to add this to a Data Model. Then add the corresponding fields from all the tables to the pivot table, creating relationships when necessary.

You are requested to answer the following:

41. How many **Quantity Ordered**, for the **Product Name** "StaDry Cycling Pants" of the **Clothing** category that took place in **September**?
    a. 34 **b. 8** c. 67 d. None of previous

42. What is the Sum of **Quantity Ordered**, for the same product mentioned in the previous question all over the 6 months?
    a. 116 b. 303 **c. 258** d. None of previous

43. How many products are there in the "Wheels" **Category**?
    a. 9 **b. 3** c. 40 d. None of previous

44. What is the sum of the **Retail Price** for the **Product Name** "Turbo Twin Tires" in the "Wheels" **Category**?
    a. 87 **b. 29** c. 92 d. None of previous

45. How many Orders have been placed during the third quarter of 2018.
    a. 2019 b. 2198 c. 4231 d. None of previous

Problem 2 ***[Each 3 mark - Total 15 marks]***

Refer to the Excel workbook data file named "Data – 2023", go to the first sheet named "Performance Ratings". You are requested to answer the following:

When potential workers apply for a job that requires extensive manual assembly of small intricate parts, they are initially given three different tests to measure their manual dexterity. The ones who are hired are then periodically given a performance rating on a 0 to 100 scale that combines their speed and accuracy in performing the required assembly operations. The file lists the test scores and performance ratings for a randomly selected group of employees. It also lists their seniority (months with the company) at the time of the performance rating. If you need to regress the performance rating (Y) versus the seniority and the three tests (X_i), respectively.

46. Generate a matrix of correlations. What is the correlation between "*Test3*" and "*Test2*"?
    a. 0.595 b. **0.796** c. 0.658 d. 0.522

47. Is there any evidence (from the correlation matrix) that multicollinearity will be a problem?
    a. Yes, as the performance rating is highly correlated with seniority and the three tests.
    b. Yes, as the three tests are highly correlated.
    c. No, as the seniority is not correlated with the three tests.
    d. None of the above.

48. Run the regression of Performance Rating versus all four explanatory variables. What is the regression equation?
    a. $Y = 6.187 + 0.155X_1 + 0.112X_2 + 0.135X_3 + 0.154X_4$
    b. $Y = 1.060 + 5.171X_1 + 2.693X_2 + 0.640X_3 + 2.638X_4$
    c. $Y = 6.557 + 0.801X_1 + 0.300X_2 + 0.0862X_3 + 0.407X_4$
    d. None of the above.

49. Referring to the equation you got in question (**48)**, if a worker (outside of the 80 in the sample) has 15 months of seniority and test scores of 57, 71, and 63, What is his prediction and an approximate 90% prediction interval for this worker's Performance Rating score.
    a. 67.46, and (63.57, 71.35) b. 34.15, and (52.9, 82.02)
    c. 67.46, and (55.29, 79.63) d. 67.46, and (64.20, 70.72)

50. Are all of the explanatory variables significant?
    a. Yes, all of the *t*-values are positive. b. No, some of the *t*-values are less than 5.
    c. No, one of the *t*-values is less than one. d. None of the above.

Problem 3 ***[Each 2.5 mark - Total 5 marks]***

Refer to the Excel workbook data file named "Data – 2023", go to the second sheet named "Nike Revenues". You are requested to answer the following:

The file lists annual revenues (in millions of dollars) for Nike from 1984 to 2017. Create a time series graph of these data. Then superimpose a trend line with Excel's Trendline option.

51. Which of the possible Trendline options seems to provide the best fit?
    a. Linear b. Exponential c. Logarithmic **d. Polynomial with order 2**

52. Based on your choice in part (a) what is your forecast for the next year?
    a. 52,116.33 **b. 34,831.23** c. 117,453,345.64 d. 28,768.70

Problem 4 ***[Each 2.5 mark - Total 5 marks]***

Refer to the Excel workbook data file named "Data – 2023", go to the third sheet named "Sales". You are requested to answer the following:

The file contains five years of monthly data for a company. The first variable is the Time (1 to 60). The second variable, Sales1, contains data on sales of a certain product.

53. Forecast this series with the moving average method having a span of 4. What is your forecast for the next month?
    a. 2696.00 **b. 2649.25** c. 2617.75 d. 2679.00

54. Forecast the series with the simple exponential smoothing method having a smoothing constant of 0.3. What is your forecast for the next month?
    a. 2623.38 b. 2592.25 c. 2696.00 d. 2681.86

Problem 5 ***[Each 2.5 mark - Total 5 marks]***

In the "MCQ / T or F" section of the bubble sheet, fill in the letter of the choice that provides the solution to the statement.

You have been assigned to forecast the number of aircraft engines ordered each month from an engine manufacturing company. At the end of February, the forecast is that 100 engines will be ordered during April. Then during March, 120 engines are actually ordered. **(<u>Hint:</u>** ***Refer to the Excel workbook data file named "Data – 2023", go to the fourth sheet named "Problem 5"*****)**

55. Using the simple exponential smoothing method with $\alpha = 0.3$, What is the forecast (at the end of March) for the number of orders placed during April?
    a. 114 b. 100 **c. 106** d. None of previous

56. Suppose that *MAE* = 16 at the end of March. At the end of March, the company can be 68% sure that April orders will be between what two values, assuming normally distributed forecast errors? (***Hint***: It can be shown that the standard deviation of forecast errors is approximately 1.25 times MAE.)
    a. (94, 134) **b. (86, 126)** c. (80, 120) d. (95.12, 116.88)

Good Luck

Dr. Marwa Sabry