Review Test Submission: Spring 2024 Term II Midterm Exam

User	Kohei Nishitani
Course	2024GSP_ANA_510_02 Statistical Modeling
Test	Spring 2024 Term II Midterm Exam
Started	4/14/24 1:02 AM
Submitted	4/15/24 2:25 AM
Due Date	4/15/24 8:00 AM
Status	Completed
Attempt Score	209 out of 290 points
Time Elapsed	25 hours, 22 minutes
Instructions	This Midterm Exam has three parts. Part I is a series of (simple) questions that should not take long to answer. Except for the last few questions, these questions do not require computation. The last few questions use the major league baseball data file, mlb1.gdt. Don't spend too much time and over-think the answers for Part I! Part II is intended to evaluate your knowledge from our review of ANA 500 Foundations of Data Analytics and focuses on multivariable regression using the whiteWines.gdt data file. Part III is intended to evaluate your understanding of logistic regression and uses the (German) credit data file creditData.gdt. In addition to the data files, I am attaching a Word doc with comments and information related to the Midterm Exam. I am also uploading a series of scripts for you to use if you want to. These include midtermPartIComputeF.inp for Part I, midtermPartIIWhiteWines.inp for Part II, and midtermPartIIILogisticRegression.inp for Part III. I have used these extensively. They should all work as written!

Question 1

1 out of 1 points

Consider a simple linear regression model, $y = \beta_0 + \beta_1 x + u$, . What does the zero conditional mean assumption imply?

Question 2

0 out of 1 points

The explained sum of squares for the regression function, $y_i = \beta_0 + \beta_1 x_1 + u_1$, is defined as _____.

Question 3

1 out of 1 points

If the residual sum of squares (SSR) in a regression analysis is 40.5 and the total sum of squares (SST) is equal to 90, what is the value of the coefficient of determination?

Question 4

1 out of 1 points

If x_i and y_i are positively correlated in the sample then the estimated slope is _____.

Question 5

0 out of 1 points

In a regression equation, changing the units of measurement of only the independent variable does not affect the _____.

Question 6

1 out of 1 points

The error term in a regression equation is said to exhibit homoskedasticty if _____.

Question 7

0 out of 1 points

Simple regression is an analysis of correlation between two variables.

Question 8

1 out of 1 points

R^2 is the ratio of the explained variation compared to the total variation.

Question 9

1 out of 1 points

A normal variable is standardized by:

Question 10

1 out of 1 points

Consider the equation, $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$. A null hypothesis, H_0: $\beta_2 = 0$ states that:

Question 11

0 out of 1 points

If the calculated value of the t statistic is greater than the critical value, the null hypothesis, H_0 is rejected in favor of the alternative hypothesis, H_1.

Question 12

1 out of 1 points

The normality assumption implies that:

Question 13

0 out of 1 points

The significance level of a test is:

Question 14

1 out of 1 points

Which of the following is a statistic that can be used to test hypotheses about a single population parameter?

Question 15

1 out of 1 points

Which of the following statements is true?

Question 16

1 out of 1 points

Which of the following statements is true?

Question 17

1 out of 1 points

Which of the following tools is used to test multiple linear restrictions?

Question 18

1 out of 1 points

A variable is standardized in the sample:

Question 19

1 out of 1 points

If the R-squared value is low, then using OLS equation is very easy to predict individual future outcomes on y given a set of values for the explanatory variables.

Question 20

1 out of 1 points

In the following equation, gdp refers to gross domestic product, and FDI refers to foreign direct investment.

$$\log(gdp) = 2.65 + 0.527\log(bankcredit) + 0.222FDI$$

(0.13) (0.022) (0.017)

Which of the following statements is then true?

Question 21

0 out of 1 points

In the following equation, gdp refers to gross domestic product, and FDI refers to foreign direct investment.
$$\log(gdp) = 2.65 + 0.527\log(bankcredit) + 0.222FDI$$

(0.13) (0.022) (0.017)

Which of the following statements is then true?

Question 22

0 out of 1 points

One popular measure to describe the relationship between the dependent variable y and each explanatory variable is the:

Question 23

0 out of 1 points

To make predictions of logarithmic dependent variables, they first have to be converted to their level forms.

Question 24

1 out of 1 points

Which of the following correctly identifies a limitation of logarithmic transformation of variables?

Question 25

0 out of 1 points

Which of the following correctly identifies an advantage of using adjusted R^2 over R^2?

Question 26

1 out of 1 points

A problem that often arises in policy and program evaluation is that individuals (or firms or cities) choose whether or not to participate in certain behaviors or programs.

Question 27

1 out of 1 points

Consider the following regression equation: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

In which of the following cases, the dependent variable is binary?

Question 28

1 out of 1 points

Consider the following regression equation: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

In which of the following cases, is '*y*' a discrete variable?

Question 29

1 out of 1 points

Consider the following regression equation: $graduate = \beta_0 + \beta_1 female + \beta_2 score + u$ where *graduate* is a dummy variable (1 if the person graduated from college, and 0 otherwise), *female* is a dummy variable (1 if the person is female, and 0 otherwise), and score is the college admission test score.

What does β_1 measure?

Question 30

0 out of 1 points

Consider the model: log $(wage) = \beta_0 + \beta_1 female + \beta_2 graduate + \beta_3 female * graduate + u$, where *graduate* is a dummy variable (1 if the person has graduated from college, and 0 otherwise), and *female* is a dummy variable (1 if the person is female, and 0 otherwise). Which of the following measures the return of graduating from college for men?

Question 31

1 out of 1 points

Consider the model: $\log(wage) = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 female * exper + u$, where *exper* is the years of work experience, and *female* is a dummy variable (1 if the person is female, and 0 otherwise). Which of the following measures the difference in the return of experience between men and women?

Question 32

1 out of 1 points

In the following regression equation, y is a binary variable:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u$$

In this case, the estimated slope coefficient, $\overline{\beta}_1$ measures _____.

Question 33

1 out of 1 points

The following simple model is used to determine the annual savings of an individual on the basis of his annual income and education.

Savings = $\beta_0 + \delta_0$ Edu + β_1Inc + u

The variable 'Edu' takes a value of 1 if the person is educated and the variable 'Inc' measures the income of the individual.

Refer to the model above. The benchmark group in this model is _____.

Question 34

1 out of 1 points

The following simple model is used to determine the annual savings of an individual on the basis of his annual income and education.
Savings = β_0 + δ_0 Edu + β_1Inc + u

The variable 'Edu' takes a value of 1 if the person is educated and the variable 'Inc' measures the income of the individual.

Refer to the above model. If $\delta_0 > 0$, _____.

Question 35

1 out of 1 points

Which of the following is true of dependent variables?

Question 36

1 out of 1 points

We have used the term "simple linear regression" to mean a regression model that involves only one dependent and one independent variable. However, we can go a little deeper into the meaning of simple "kinear regression". Select the best choice below to define "linear regression". That is, "linear regression" means that _____.

Question 37

1 out of 1 points

The model below is a simple linear model.

$$y = \beta_0 + \beta_1 Log(x_1) + e$$

Question 38

1 out of 1 points

The model below is linear.

$$\log(y) = \beta_0 + \beta_1 x_1 + e$$

Question 39

1 out of 1 points

The model below is linear.

$$\ln(y) = \beta_0 + \beta_1 \sqrt{x_1} + e$$

Question 40

1 out of 1 points

There are actually two assumptions in this statement with respect to ordinary least squares regression," variances must be evenly distributed and centered about 0 (zero)". Evenness of variances (or residuals) means that the expected value $E(e|x) = \sigma^2$ where the square root of σ^2 is the standard deviation, i.e. homoskedasticity is present. The fact that the variance must be centered about 0 (zero) referes to having "unbiased" coefficients. Further, the fact is that homoskedasticity plays no role in whether or not coefficients are unbiased. Mathematically this is written as:

$$E(y|x) = \beta_0 + \beta_1 x_1$$

$$Var(y|x) = \sigma^2$$

Question 41

1 out of 1 points

Consider the model below:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + e$$

where salary is a players total salary, gamesyr is the average games played per year,

and where bavg is the career batting average, hrunsyr is the home runs per year, and rbisyr is runs batted in per year are performance statistics.

To test whether or not the performance statistics have any effect on salary we would use the hypothesis _____ .

Question 42

1 out of 1 points

A model including all terms for all independent variables is called an "unrestricted" model. Whereas, a model that has some terms for some variables omitted is called a "restricted" model.

Question 43

0 out of 1 points

A t statistic can be used to test this model (below) and determine whether or not the terms involving the variables bavg, hrunsyr, and rbisyr are "individually" significant. That is, a simple t-test can be used to determine whether or not performance statistics should be included in the model.

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + e$$

Question 44

1 out of 1 points

The model shown below where terms related to the performance statistics have been removed is a "restricted model".

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + e$$

Question 45

1 out of 1 points

Take my word for it. The correct answer to this question is "True". This is important because it provides a much easier way to compute the F-statistic fpr more complicated, more real-world problems.

We can develop an F statistic (*or F ratio*) by using an unrestricted and a restricted model with the formula:

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where SSR_{ur} is the sum of squared residuals from the unrestricted model and SSR_r is the sum of squared residuals from the restricted model. q is the number of restrictions imposed moving from the unrestricted to the restricted model (e.g. if we remove 3 terms then q=3). And, where $n =$ and $k = the\ number\ of\ independent\ variables$.

Question 46

1 out of 1 points

Using the data file mlb1and the script provided enter the value of the sum of squares error for the unrestricted model. Be sure to round your answer to two decimal places.

Question 47

1 out of 1 points

Using the data file mlb1 and the script provided what is the value of the sum of squares error for the restricted model. Be sure to enter your value rounded to two decimals.

Question 48

1 out of 1 points

Last, using the data file mlb1 and the script provided what is the F statistic computed using the sum of squares error for the unrestricted and the restricted models. Be sure to round your answer to two decimal places.

Question 49

1 out of 1 points

Based on what you have learned so far in our course and your program of study, the data file mlb1 and the script provided, consider the following statement.

There is no way to obtain the critical value F without a computer and long, onerous computations.

Question 50

1 out of 1 points

Consider the following based on the data file mlb1 and output from the script provided.

Based on the critical value F you obtained and the F-statistic you computed you determine that the null hypothesis must be rejected. There is sufficient evidence that the performance statistics terms should remain in the model.

Question 51

0 out of 10 points

Be sure to refer back to your Word doc. This question starts the work (questions) from Part II Evaluating Refression Models on the White Wines dataset.

[A} is the dependent variable in this dataset.

Question 52

0 out of 10 points

What is the data type of this variable?

Question 53

10 out of 10 points

Is the distribution for this variable normal?

Question 54

10 out of 10 points

Does multicollinearity exist?

Question 55

10 out of 10 points

If multicollinearity exists, which variables are highly correlated. Be sure to check all the variables that are highly correlated. (Remember that I use 0.6 and above as a rule of thumb for high correlation. And, don’t worry about pairing up these variables right now. Just check the variable if it is highly correlated with another variable.)

Question 56

10 out of 10 points

Generate an OLS model for the dataset using all independent variables as parameters in the model. Are the intercept and all coefficients statistically significant?

Question 57

10 out of 10 points

Based on the R-squared value for this model, does the model explain most of the variation in the data?

Question 58

10 out of 10 points

Check the variable or variables you found that are highly correlated and are NOT statistically significant.

Question 59

0 out of 10 points

Build a new OLS model for this dataset and include Volatile Acidity, Residual Sugar, Sulphates, pH, Free Sulfur Dioxide, Density, and Alcohol as independent variables. Now are the intercept and all coefficients statistically significant?

Question 60

10 out of 10 points

Did the R-squared value improve for the last model you built?

Question 61

10 out of 10 points

Using some information published on the Internet to conduct feature selection (to choose which variables to use as independent variables), build an OLS regression model (Model 2) using Residual Sugar, Chlorides, Total Sulfur Dioxide, pH, and alcohol as the independent variables. Based on the value of R-squared, did the model improve in terms of explaining the variation in the data?

Question 62

10 out of 10 points

It appears that some of the variables still included in the model do not help explain the variation in the data. Or, as I like to say, "something hinky is going on..." Sample the dataset to find the observations that make the most difference to the quality of the wine. That is, first create a subset of the dataset for observations where quality is greater than 7; and, second create another subset where quality is less than 5. This should give you some idea of the variables differences between the worst and best wines. Which variables appear to make the most difference to the quality of white wines? Using Volatile Acidity, Citric Acid, Residual Sugar, and Free Sulfur Dioxide as independent variables build an OLS model. Based on the value of R-squared, does the model explain more of the variation in the data? Or in other words, did the R-squared value show this model improved?

Question 63

0 out of 10 points

Using Volatile Acidity, Citric Acid, Residual Sugar, and Free Sulfur Dioxide as independent variables and the formula you used in Part I of this exam to compute the F statistic, evaluate whether or not all the variables in the unrestricted model should be included. Do all variables need to remain in this model?

Question 64

10 out of 10 points

Again, be sure to refer back to your Word doc. This question begins the work (questions) in Part III Logistic Regression on the credit dataset.

What is the best type model for these data?

Question 65

10 out of 10 points

Build a logistic regression model and use its output to answer the related questions. Are the intercept and all the coefficients statistically significant?

Question 66

10 out of 10 points

Out of 1000 observations, what are the number of cases “correctly predicted” by Model1? This is an exception to the usual procedure. Since we are looking at the number of cases, be sure to enter your answer as a whole number (integer).

Question 67

0 out of 10 points

One of the problems with current credit rating systems is that they tend to not be very reliable in terms of predicting defaults, a Type II error. Output from Model2 indicates that _____ instances were predicted to be “good” or “no default” when in fact those instances had defaulted.

Question 68

0 out of 10 points

The problems with or number of Type II errors can be mitigated by decreasing the number of observations.

Question 69

10 out of 10 points

Type II errors can be decreased by increasing the level of significance. Of course, then the number of Type I errors will increase.

Question 70

10 out of 10 points

Given the output from Model2, the coefficients for the variables AccountBalance, ValueSavingsStocks, and ConcurrentCredits indicate that as a loan applicant’s account status increases, the value of his/her savings increases, and the less he/she currently owes on other credit accounts indicates his/her credit worthiness will decrease.

Question 71

10 out of 10 points

Now look at the average marginal effects (AME) and some probabilities, i.e. the probability that a debtor will default given some very basic data to consider. (Note that the fact a value is output for the AME with respect to the intercept (b1) is meaningless.) Since we will reuse some of the functions from the script from PS3 and add a bunch more functions, I’ll try to keep this simple. The dependent variable Creditability will be a function of the intercept and the independent variables AccountBalance, CreditAmount, ValueSavingsStocks, and ConcurrentCredits. The logic is that the more ability a debtor has to repay the less trouble he/she would have repaying a loan. And, the more the debtor borrows the more trouble he/she would have repaying that amount, etc.

Using the script provided, build a simpler logistic regression model, Model2, using the variables listed above. What is the average marginal effect of a loan applicant’s current bank account balance, i.e. AccountBalance?

Question 72

10 out of 10 points

The average marginal effect for AccountBalance means that a unit increase in a loan applicants accounts status will result in a 0.11 increase in the applicant’s credit worthiness or Creditability. (This is where the miscoding gets really sticky. This result is consistent with my discussion about potential miscoding above and I think consistent with results by other researchers.)

Question 73

0 out of 10 points

Considering average marginal effects again, if a loan applicant requests a higher loan amount that will have a nearly negligible effect on an applicant's credit worthiness or Creditability.

Question 74

10 out of 10 points

Now, considering the 95% confidence intervals, we can be 95% certain that within the true population a loan applicant will likely not currently have a checking or savings account.

Monday, April 15, 2024 2:25:06 AM EDT

← OK