PHC 6091 (Spring 2021) Exam #2

Due: Thursday, 4/1, by 11:59pm.

Background: Dataset <mark>exam2.sas7dat</mark> is the data of an observational study designed to compare the effectiveness of a new drug, which is viewed by the investigators as the active treatment vs. a standard inhaler, for treating asthma. The patient follow-up period begins at the hospitalization at which treatment for asthma is initiated, using either the new drug or an inhaler. The follow-up period (since the initiation of the interventions) is one year. Variables in the study include:

Severity   = An index for asthma severity, with higher values indicating great severity.

QOL   = quality of life index

Hosp1yr   = a binary 0-1 variable indicating the presence of at least one hospitalization during the 1 year follow-up period

Treatment   = a binary 0-1 indicator variable where 1 indicates that a patient received the active experimental drug and 0 indicates the patient used the inhaler (standard of care).

QOLCat   = a 4-level ordered categorical variable, with higher scores representing a higher QOL

Num_hosp   = the number of the hospital, emergency department, or urgent care admissions during the 1-year follow-up period.

Age1   = Age in years

Male   = 0-1 indicator variable for male sex

Hispanic   = 0-1 indicator variable for Hispanic ethnicity (0=non-Hispanic; 1=Hispanic)

MaternalEd   = Maternal education in years

Urban   = 0-1 indicator for residence in an urban area (0=rural, 1= urban)

FamIncome   = Annual family income in $

Question 1: Regression model selections. Using <u>QOL</u> as the <u>outcome</u> and age1, treatment, and severity as the candidate <u>predictor variables</u>, conduct a stepwise model selection procedure (8 pts)

1. Using SLEntry = 0.01 and SLStay = 0.05, what are variables being selected? (2 pts)

2. Describe this model selection procedure from step 1 to the final step (2 pts)

3. What is the R squared and CP value for the final model? (2 pts)

4. Check if there are any collinearity problems for the final model? (2 pts)

Question 2: Examine the effect of the QOL category (QoLCat) on severity. Please answer the following questions (18 pts)

1. Using One-way analysis of variance (ANOVA) model to test if QoLCat is a statistically significant factor on severity (note: severity is the dependent variable/outcome). (13 pts)

(a) What is the degree of freedom for the QoLCat effect? (2 pts)

(b) What is the total sum of square (SS) for severity that is explained by QoLCat? (2 pts)

(c) What is the mean square (MS) for severity that is explained by QoLCat? (2 pts)

(d) What is the degree of freedom for the error term? (2 pts)

(e) What is the mean square error (MSE)? (2 pts)

(f) Conduct a hypothesis test to assess the effect of QoLCat on severity. (3 pts)

2. Obtain the estimated mean of severity for each of the four QoL levels. (2 pts)

3. Using the Tukey-Kramer method to conduct the pairwise comparisons between different levels of QoLCat that are adjusted for multiple comparisons. Which pairs have significant difference in the mean of severity? (Hint: The total number of unique pairwise comparisons is 6) (3 pts)

Question 3: Using Two-way ANOVA model to test if Treatment and QoLCat are statistically significant factors on severity (23 pts)

1. In the two-way ANOVA model, you first include the two main effect and the interaction between the two main effects. (10 pts)

(a) Conduct an overall F test to examine if this ANOVA model is significantly better than the null model. (2 pts)

(b) In this ANOVA model, what are the degrees of freedom for (i) Treatment, (ii) QoLCat, (iii) the interaction term, and (iv) the error term? (4 pts)

(c) In this ANOVA model, what are the type III sum of squares for (i) Treatment, (ii) QoLCat, (iii) the interaction term, and (iv) the error term? (4 pts)

2. Conduct a type III F test to examine if the interaction term is statistically significant (3 pts)

3. How many terms should be included in the <u>final</u> two-way ANONA model? What are they? (2 pts)

4. Conduct the type III F test for two main effects (Treatment, QolCat) in the final ANOVA model (3 pts)

5. Based on the final ANOVA model, what is the estimated mean for each combination of Treatment and QoLCat? (Hint: there are 8 combinations in total) (5 pts)

Question 4: Polynomial model of QOL

Now, using severity as the outcome variable, build the polynomial model using quality of life index (QOL) as the independent variable. (10 pts)

1. Start the polynomial model with the linear and quadratic terms of QOL. (6 pts)

(a) Are there any influential points? If so, remove them in the analyses below. If not, please use the entire dataset (you do not need to worry about the outliers here). (2 pts)

(b) What is the Pearson's correlation coefficient between the linear term (QOL) and the quadratic terms (QOL^2)? (2 pts)

(c) Using rule of thumb for variation inflation factor (VIF) to see if there is a collinearity issue. (2 pts)

2. If there is a severe collinearity issue, what is the remedy? Rerun your model using this method and then test significances of the corresponding linear and quadratic terms. (4 pts)

Question 5: Logistic regression model

Now, using Hosp1yr (a binary 0-1 variable indicating the presence of at least one hospitalization during the 1-year follow-up period) as the outcome variable and Treatment as the independent variable (exposure) to build up logistic regression models. (33 pts)

1. Conduct a simple logistic regression using Treatment as the only independent variable. (10 pts)

(a) Conduct a global test to examine if the model is significantly better than the null model. Please specify the name of the test and the test statistic that you are using (2 pts)

(b) What are the values of AIC and deviance (-2log likelihood) for the model? (1 pt)

(c) Conduct the Wald test for the effect of Treatment on hospitalization. (2 pts)

(d) Express the simple logistic regression model using the resultant parameter coefficients. (2 pt)

(e) What is the unadjusted odds ratio (OR) and its 95% confidence interval (CI) of hospitalization between patients treated with the experimental drug and the standard of care? (3 pts)

2. Test if gender is a confounder for the relationship between hospitalization and Treatment. (5 pts)

3. Test if gender is an effect modifier for the relationship between hospitalization and Treatment. (5 pts)

4. Conduct a multiple logistic regression model for the relationship between Treatment & hospitalization adjusting for the following covariates: age, Hispanic race, and residency (urban/rural). (5 pts)

(a) Conduct the partial Wald test for the treatment effect on hospitalization in this model. (2 pts)

(b) What is the adjusted OR of hospitalization comparing patients treated with the new drug and the standard of care? (2 pts)

Question 6: **Evaluate the predictive accuracy of the model in question 5 using the ROC curve analysis (8 pts)**

(a) Plot the ROC curve using the predicted probability based on the model you fit in question 5 (3 pts)

(b) What is the area under the curve (AUC)? (2 pt)

(c) Explain the AUC value (3 pts)