THE UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON
SCHOOL OF PUBLIC HEALTH
PH1700 INTERMEDIATE BIOSTATISTICS

Individual Data Analysis Project

Multivariate Linear Regression and Logistic Regression Application

Due date: Friday December 7, 2018 before 11:59 PM CDT

This data analysis project is to be completed by each student.
The project will be graded and will be 15% of your final grade.
Page 2 of this document contains a summary of the study that you will analyze. Our TA Helen Z Engle will deliver you the ".csv" file which can be imported into STATA for your analysis (though you can use SAS, R or any statistical software). **Your** dataset name is **FirstName_LastName_EQOL.cvs.** This dataset is a subsample of the original dataset but, for the purposes of this final project, analyze it as it were all the data.

The last pages of this document contain the annotated data collection forms that were used in this study. The annotation are the variable names that are contained in the .csv file. You will need to refer to the data collection forms to understand variable coding for analysis.

Answer the study questions below using the data provided and prepare a written report containing your written summary along with supporting statistical analysis. Your results should be formatted as outlined below. In your write-up, bring in relevant information from the output. Be sure to provide statistical detail – indicate the analyses performed by mentioning specific statistical tests, provide relevant descriptive information, report test statistics and degrees of freedom along with p-values and provide relevant conclusions. The focus of the report should be on interpretation of the results for a reader who is not quantitatively-minded. Your report will be graded for clarity of interpretation (writing), conciseness, as well as appropriateness of statistical analysis.

We expect that you will analyze the data on your own, and that you write the brief report that you submit to us to evaluate and grade. This is an individual project, and it is **NOT a collaboration exercise.** We expect to do your own work. In other words, you must work on your own on this project and you are not allowed to share your project, results or electronic files or documentation with anyone. You are expected to erase the dataset provided to you after the final grade has been posted for this project. This project is subject to the University conduct and discipline code (Please read full description at http://legal.uth.tmc.edu/hoop/06/6_03.html). **The use of this dataset is for teaching purposes only and there can be no publication from the dataset even as an example of an approach to a particular type of analysis.**

Your individual dataset and project description appears with your own name in CAVAS under the INDIVIDUAL PROJECT section in the ASSIGNMENTS Content Area. All analyses are to be done in Stata, R, SAS or other stat software. The dataset you are

assigned is in a comma delimited text file (.csv) format. You will need to import this file into Stata using the import command: File > Import > Text data (delimited, *.csv, …).

Data description.
Each student receives a dataset. This dataset is a subset of the original dataset with the information of n=400 patients. Every student has a different dataset. The results may differ.
Each row in the dataset corresponds to one individual.
The last table of this document is a codebook with the column names in the dataset.
The columns indicate:
Demographic variables:

- Gender: Please Code female as 0 and male as 1.
- Age: Age in years at the time of injury
- race_ethn: Ethnicity. Use White as reference group.
- marital: Marital status 0 not married, 1 married

Injury Characteristics:

- TFC: Time to follow commands; duration of time in days from the onset of traumatic brain injury to the point in time during which a patient can comprehend and accurately follow spoken or written direction or command

Economic variables:

- EQOL: Economic quality of life score. Treat it as a continuous variable, the higher the more confident the person FEELS about their economic situation. The Economic QOL is a patient-reported outcome measure of economic quality of life specifically developed for individuals with disabilities. Its items pertain to one's ability to afford to pay bills and live comfortably. Respondents are instructed to indicate how they have felt lately on a scale of 1 (never) to 5 (always). Examples of items include the following: "I have enough income to live the life I want", "I can afford to feed myself and my family", and "I can afford to pay my bills on time." Items were developed based on semistructured interviews and focus groups conducted with individuals with disabilities; questions were informed by a comprehensive review of the literature.
- pre.employed: 1 if the person was pre-employed 0 other wise.
- post.employed: 1 if the person was post-employed 0 other wise.
- preinjury.retired: 1 if the person was retired at the time of injury, 0 otherwise. If this variable is equal to 1 the patient **must not** be included in the analysis where the response is post.employed (see introduction below).
- YearsEduc: Years of education at the moment of injury.

Your report need will have a total of 5 different sections: Introduction (that is given), Methods, Results, Discussion, and at the end "Appendix: Statistical Analysis".

Introduction: For almost 20 years, there has been a growing interest in gender differences following traumatic brain injury (TBI), yet there has also been a lack of systematic research in this area, meaning that long-term health and quality of life outcomes for girls and women with TBI remain relatively unknown. Early research in the area of gender differences in outcomes following TBI revealed that women had better predicted outcomes in terms of their capacity for returning to their pre-injury work environment. This early research did not control for preinjury employment (preemployment) of the participants. The objective of this study is to assess the role of gender in two quality of life outcomes: post-injury employment status (postemployment) and economic quality of life (EQOL), while controlling for preemployment. Two independent, one per outcome, analyses are required.

Hypothesis 1:
Gender has an impact on postemployment after controlling for the preemployment and other covariates.
Hypothesis 2: Gender has an impact on EQOL after controlling for preemployment, postemployment and other covariates.

Methods: This section should describe what steps and statistical methods you did to analyze the data and how you applied them to solve the questions asked. You also need to provide a description of what statistical methods were used and the rationale or purpose for it. Please describe any statistical methods used for testing assumptions of the test if needed. If you created new variables for your analyses, you need provide the rationale for creating the variable and describe the method you used to create the new variable. Add a sentence referencing the software, in this case Stata, SAS or R, you used for all your analyses, just as you are expected to do for any peer review publication. See below for the statistical analysis this project requires. **Do not provide results**.

Results: The results section needs to mimic a peer review publication, so it needs to include the following (in the order that favors your narrative) elements:

- Identify the variables used in the comparison and create a **summary table** that describes your sample. These descriptive statistics are based on the original data, rather than any new variables you create for your analyses. The data most summarize the predictor variable or covariate values and check if the distribution of this variables is the same in female and male subjects. This is, for example, if the distribution of Age, age in years at the time of injury, is the same for females and males. You need to report this comparison for each one of the covariates and the p-values must be computed using the tests according to the type of values the covariate takes. At a significance level

0.05, determine which covariates are unbalanced in women and men. Fill out the following table. The highlighted results are examples of how to fill out the data but they may not match the results using your dataset.

Summary Table

Continuous variables	Female (n= 127) Mean (SD)	Male (n=377) Mean(SD)	Test used	pvalue
Age			Two sample t-test or maybe wilcoxon test	
So on…				
Discrete variables	Female Percentage (n)	Male Percentages (n)	…	…
Race/Ethnicity White Black Hispanic Other	 52% (66) 29.1% (37) 15.7% (20) 3.1% (4)	 41.6% (157) 40.1% (151) 14.6% (55) 3.7% (14)	Chi sqare test or maybe Fisher exact test	
So on…				

\

For hypothesis 1 you will be using logistic models:

0. Remove the patients who were retired at the moment of data collection out of this analysis. This is remove the patients with preinjury.retired =1

1. Fit the logistic model with postemployment as outcome and the covariates in table below as predictors and interpret the results in the text.

Logistic Regression Model Estimating Effects of Gender and Other Variables on Post-Injury Employment

	Estimate	SE	Z Value	*p*	95% CI
Male Gender	1.04	.52	2.03	.04	.03, 2.05
Age					
Black Race					
Hispanic Ethnicity					
Other Race/Ethnicity					
Education					
Pre-Injury Employment					

Marital Status
TFC

- *Note. Estimate is the estimate of the regression coefficient.* Racial/ethnic groups were compared to a White reference group; TFC = time to follow commands; CI = Confidence Interval. Sample size n=XYZ

2. Remove from the logistic model the unnecessary predictors. Do not remove the predictor Gender.

- Fit the new model. This is, provide a table as the one above but only with the predictors that you decide to keep in the model. Specify how the predictor variables that appear in your final model were selected.

3. Summarize the key results from your final logistic regression model in the text, keeping in mind hypothesis 1, compare the odds ratios of postemployment between male and females. Check if which other predictors have an impact in the response.

For hypothesis 2:

0. Make sure to include all the patients, even the retired ones.
1. Run the regression model with all covariates in table below and interpret the results.

Multivariable Regression Model Estimating Effects of Gender and Other Variables on Post-Injury Economic Quality of Life

	Estimate	SE	T Value	*p*	95% CI
Male Gender	2.59	1.09	2.38	.02	.45, -4.73
Age					
Black Race					
Hispanic Ethnicity					
Other Race/Ethnicity					
Marital Status					
Education					
Pre-Injury Employment					
Post-Injury Employment					
TFC					

Racial/ethnic groups were compared to a White reference group; TFC = time to follow commands; CI = Confidence Interval. Sample size n=XYZ

2. Fit the new model. This is, provide a table as the one above but only with the predictors that you decide to keep in the model. Specify how the predictor variables that appear in your final model were selected.

3. Summarize the key results from your final regression model in the text, keeping in mind hypothesis 1, compare the odds ratios of postemployment between male and females.

Discussion: In this section you need to describe what the results mean in the context of the scientific question integrating all the questions asked for the project. Besides answering hypotheses 1 and 2 you can emphasize any interesting findings.
You can also discuss the some limitations of the analysis.

Direct any questions about the project to your TAs assigned to your class.

"Appendix: Statistical Analysis"
Provide with any other analysis or figures. For example intermediate models to go from the full model in step 1 to the model in step 2, goodness of fit of your models such as residual analysis, etc.

SUBMISSION

Your complete report **should be no longer than three pages in total** (three single-sided pages, not including tables or figures which can be attached to the report- a maximum of two additional single-sided pages may be included for tables and graphs).
Once you have completed all your analyses and written your report, you will upload the report and a do file that documents all your analyses and analytic decisions to CANVAS.
Upload the following files to the "Final Project" section in the Assignment Content area in CANVAS

1. Save your answers under the name **LastName_FirstName_ProjectReport.docx.**
2. Save your do file under **LastName_FirstName_Project.do**
3. Upload your completed assignment and do file no later than **11:59 PM (CDT), Friday, December 7, 2018.**

PROBLEMS AND GRADING: The following table outlines the content of your report and the points assigned to each section.

Section on Report	Points
Introduction	0 (It is given)
Methods	20
Summary Table (Table and text describing it)	10
Full Logistic model and interpretation	10
Final logistic model and interpretation	15
Full or initial Regression Model and interpretation	10
Final Regression Model and interpretation	15
Discussion	20
Total	**100**

CODEBOOK

Below is the description of the variables that you may have received in your dataset.

LIST OF VARIABLES:

Variable	Description
ID	Patient ID. Irrelevant for the analysis
Gender	Female and Male
Age	Age in years at time of injury
Marital	1=married, 0=single
race_ethn	race ethnicity: white, black Hispanic and other
YearsEduc	Years of education at time of injury
pre.employed	0 preUNemployed, 1 preemployed
preinjury.retired	0 not preinjury retired, 1 preinjury retired
post.employed	0 postUNemployed, 1 postemployed
TFC	Time to follow command in minutes
EQOL	Economic quality of life score