**Data Analysis Project 2**

Spring 2017

**Before you begin:**

* You may use any reference materials you wish (textbook, class notes, lectures, labs, homework, practice exercises, etc.), including the computer/internet. However, you must work on this project **by yourself** and you **may not discuss this project with anyone**. Emails to the professor for clarification of questions are the only exception to the “no discussing” rule.

**General Instructions:**

* You are required to use Stata for all data analysis.
* Your final product will be a **written report** split into five parts (plus an appendix).
* A template for this report is provided separately (**Project2\_TEMPLATE.docx**). Please use the template as the file into which you write your answers.

**Submission Instructions:**

* The project is due by **Thursday April 27, 2017, 11:59pm.**
* **No late assignments will be accepted.**
* To submit, upload your project to the “Data Analysis Project 2” Dropbox on Carmen. Please submit your project as **one file**.
* Be sure to “sign” the Honor Code pledge in your submission.

Good luck!

**Honor Code Signature:**

Joshua Wu

I pledge on my honor that I have **neither given nor received unauthorized aid on this project**. Typing my name on the line above serves as a signature of this pledge.

**Part 1: Summarizing the Sample**

*Write-up:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std. Dev** |  |
| Age | 135 | 57.95 | 13.29 |  |
| BMI | 135 | 27.10 | 5.84 |  |
|  |  | **Male** | **Female** |  |
| Sex | 135 | 65 | 70 |  |
|  |  | **Never** | **Current** | **Former** |
| Smoking Status | 134 | 74 | 10 | 50 |
|  |  | **Nodular** | **Superficial** | **Mixed Type** |
| Type of BCC Tumor | 135 | 47 | 33 | 55 |

One data entry for smoking status is missing.

**Part 2: Modeling the gene expression z-score**

*Write-up:*

There are differences in gene expression z-score by BCC tumor. Using ANOVA, we have obtained a p-value of 0.0096. This tells us that there are indeed differences, but it doesn’t tell us which ones are different from each other. Following the ANOVA with pairwise comparisons, while using the Bonferroni adjustment to decrease Type 1 error. By doing that, I have obtained significant difference between 2 vs 1 (Superficial vs Nodular), and 3 vs 1 (Mixed Type vs Nodular) with p-values 0.099 and 0.010 respectively (Both p-values are less than 0.0167 [0.05/3]). The estimated difference in gene expression z-score mean between superficial and nodular is 0.448, and the estimated difference in gene expression z-score mean between mixed type and nodular is 0.543.

**Part 3: Modeling one specific mRNA marker (CD68)**

*Write-up:*

The model did not pass the normality assumption in the beginning. By log-transforming the Y variable, the normality assumption has now been passed. The p-value for age is 0.091, for sex it is 0.030, and for bmi it is 0.365. In this case, age and sex are significant predictors of ln\_cd68. For every year increase in age, there is an estimated geometric mean of cd89 decrease by 1.45 percent, controlling for sex and bmi. The estimated geometric mean of cd68 for men is 66.90 percent larger than women, controlling for age and bmi.

**Part 4: Modeling a high level of one specific mRNA marker**

*Write-up:*

From the Chi2 model we have, we can see that having a high level of cd25 is associated with bcc type (p-value = 0.002). By using the logistic regression, we can see that there were significant odds ratios between superficial vs nodular, and mixed type and nodular (p-value 0.028 and 0.001 respectively). The estimated odds of high cd25 for people with superficial BCC type is 2.83 times greater than the odds for people with nodular BCC type. The estimated odds of high cd25 for people with mixed type BCC type is 4.125 times the odds of people with nodular BCC type.

**Part 5: Assessing time to tumor recurrence**

*Write-up:*

****

From the plot above, we see that the survival estimate is better in the “all markers high” category, the other two categories look to be the same. By doing the hypothesis test, we have determined that there are differences in time to recurrence by gene expressions category (p-value = 0.0171). The estimates for the 25th percentile of survival for “No markers high” is 9.83 months; “some markers high” is 7.69 months, and “all markers high” is 30.48 months. The survival probability for 24 months for “no markers high” is 61.70%, for “some markers high” it is 59.86%, for “all markers high” it is 83.00%.

**APPENDIX**

*All requested Stata code here (properly formatted so readable):*

**Part 2:**

****

****

**Part 3:**

****

**Part 4:**

****

****

**Part 5:**

****

****

****

****

****

****