|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PM592: Regression Analysis for Data Science** | | | Name: |  |  |
| **HW8** |  |  |  |  |  |
| *Logistic Regression II* | | | | |  |

**Instructions**

* Answer questions directly within this document.
* Upload to Blackboard by the due date & time.
* Clearly indicate your answers to all questions.
* If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
* For the purpose of this assignment, statistical evidencerefers to a test statistic and associated p-value.
* If a question requires a table or figure, you must present these in a professionally formatted way (e.g., <https://www.researchgate.net/figure/Results-of-Logistic-Regression-Analysis-Unadjusted-and-Adjusted-Odds-Ratios-of_tbl2_47335480>)
* If a question requires a conclusion, it must be phrased professionally and coherently.
* There are 2 questions and 30 points possible.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 1** |  |  |  |  | [19 points] |  |

Shanahan et al. (2016) studied the effect of exposure to nature and health outcomes in an urban Australian population. A questionnaire was administered to assess health outcomes and the amount of time each participant typically spends in green spaces (e.g., parks). The data is located in **green.csv**.

depres = Subject was classified as having depression

hibp = Subject was classified as having high blood pressure

dur = Duration (minutes) of each visit to green spaces   
(Note that response options included 0, 1-15, 16-30, 31-45, …, 106-120, >120 minutes. To approximate these categorical responses as continuous, the midpoint of each interval was taken as the value for “dur.” Individuals responding >120 minutes had a recorded “dur” value of 128.5.)

|  |  |
| --- | --- |
|  | 1a. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the grouped smooth approach. |

|  |  |
| --- | --- |
|  | 1b. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the LOESS smoothing approach. |

|  |  |
| --- | --- |
|  | 1c. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the fractional polynomials approach. |

|  |  |
| --- | --- |
|  | 1d. [3 points] Write a short (3-5 sentences) methods section describing how you evaluated the assumption of linearity in these models and providing your rationale for modeling the variables in the way you did. |

|  |  |
| --- | --- |
|  | 1e. [4 points] Write a short (3-5 sentences) results/conclusion section describing your modeling approach and the parameter estimates (with 95% CI and p-values) of interest. In your conclusion, make specific recommendations about how much time residents should spend in green spaces in order to improve these two health outcomes. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 2** |  |  |  |  | [11 points] |  |

Read the article by Kalligeros et al. (2020) on Blackboard. The authors performed a logistic regression to determine risk factors for two outcomes: 1) being admitted to the ICU and 2) being intubated within 10 days of hospital admission for COVID-19.

|  |  |
| --- | --- |
|  | 2a. [3 points] Do the authors assess the linearity assumption for BMI? Why do you think they ended up treating BMI as categories instead of as a linear variable? |

|  |  |
| --- | --- |
|  | 2b. [3 points] The authors believed that severely obese patients may be at greater risk for ICU admission due to having a higher risk of heart disease. What did the authors do to ensure that the effect of severe obesity on ICU admission was due to COVID and not other comorbidities? |

|  |  |
| --- | --- |
|  | 2c. [3 points] The confidence intervals for some parameter estimates in Tables 2 & 3 are quite wide (e.g., 1.39 – 71.69). Does this concern you? What was the authors’ explanation for the size of the confidence intervals? |

|  |  |
| --- | --- |
|  | 2d. [2 points] Suppose you are in charge of conducting a new study to replicate these findings. How would you change the study design to prevent these large confidence intervals? |