|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PM592: Regression Analysis for Data Science** | | | Name: |  |  |
| **HW2** |  |  |  |  |  |
| *Bivariate relationships, simple linear regression* | | | | |  |

**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.
* There are 4 questions and 30 points possible.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 1** |  |  |  |  | [7 points] |  |

Use the WCGS data you saved from HW1.

|  |  |
| --- | --- |
|  | 1a. [1 point]. Provide a scatter plot of the relationship between SBP (Y) and BMI (X). Does anything in this scatter plot give you concern? |

|  |  |
| --- | --- |
|  | 1b. [1 point] Provide the null and alternative hypotheses that would test whether SBP is related to BMI. |

|  |  |
| --- | --- |
|  | 1c. [1 point] Provide the linear regression equation for the relationship of SBP regressed on mean-centeredBMI. |

|  |  |
| --- | --- |
|  | 1d. [2 points] Based on your regression results, what are your decisions for the hypotheses in (1b)? Provide the test statistic and p-value. |

|  |  |
| --- | --- |
|  | 1e. [1 point] How much is SBP expected to change if BMI increases by 5 units? |

|  |  |
| --- | --- |
|  | 1f. [1 point] Add the confidence interval and prediction interval to your plot in (1a). What is the interpretation of each of these intervals? |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 2** |  |  |  |  | [8 points] |  |

Continue to use your WCGS file.

|  |  |
| --- | --- |
|  | 2a. [3 points] Fill in the table below by running 3 separate regressions of SBP on NCIGS (# cigarettes smoked per day). For each regression, provide the estimate of the intercept (& 95% CI), the estimate of the slope (& 95% CI), and the R2 value. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Predictor | Intercept (95% CI) | Slope (95% CI) | R2 |
| 1 | NCIGS |  |  |  |
| 2 | (NCIGS – 10) |  |  |  |
| 3 | (NCIGS – mean(NCIGS)) |  |  |  |

|  |  |
| --- | --- |
|  | 2b. [2 points] For each of the 3 models, interpret the estimates of the intercept in a meaningful way. |

|  |  |
| --- | --- |
|  | 2c. [1 point] Comment on the respective widths of the 95% confidence intervals. |

|  |  |
| --- | --- |
|  | 2d. [1 point] Comment on the slopes and explain why they are similar or different. |

|  |  |
| --- | --- |
|  | 2e. [1 point] Use the predict() function to obtain the estimate of SBP for somebody who smokes 19 cigarettes per day. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 3** |  |  |  |  | [8 points] |  |

Read the following web page:  
<https://www.visualcapitalist.com/chart-money-can-buy-happiness-after-all/>

|  |  |
| --- | --- |
|  | 3a. [2 points] Download or manually enter in the data from the section “The Results”. You should have 3 variables: annual income, well-being (experienced), and well-being (evaluative). Produce a scatter plot of each well-being variable vs. annual income. |

|  |  |
| --- | --- |
|  | 3b. [2 points] Does well-being appear linearly related to income? The website reports “a recent study found that happiness increases linearly with reported income (logarithmic).” Does performing this transformation appear to better satisfy the linearity assumption? |

|  |  |
| --- | --- |
|  | 3c. [2 points] The website reports the well-being variables are “measured in standard deviations from the mean”. What is another name for this type of variable? Why do you think the authors did this transformation? |

|  |  |
| --- | --- |
|  | 3d. [2 points] Report what you believe is the best regression relationship between each well-being variable and income (however you choose to transform it, if at all). |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question 4** |  |  |  |  | [8 points] |  |

Use the article by Sarafidis et al. on Blackboard to answer the following questions.

|  |  |
| --- | --- |
|  | 4a. [2points] Write a brief summary describing the hypotheses being tested, how the sample was obtained, the sample sizes in each group, and the statistical procedures used to analyze the data. |

|  |  |
| --- | --- |
|  | 4b. [2 points] It is not explicitly stated, but which statistical test do you think the authors used to obtain the p-values in Table 1? |

|  |  |
| --- | --- |
|  | 4c. [2 points] Write a small R program that can input a mean, standard deviation, and sample size for each of two groups and compute this test comparing the two groups. Can you re-create the p-values obtained by the authors using this method? |

|  |  |
| --- | --- |
|  | 4d. [1 point] Did the authors report whether they tested any of the assumptions of linear regression? |