**CSC591**: Foundations of Data Science  
HW3: Regression Analysis

Released: 10/21/15  
Due: **10/27/15 (23:55pm);** (One day late: -25%; -100% after that).

Student Name:   
Student ID:

**Notes**

* Submit single zip file containing: (1) all solutions as single pdf file (Filename: Lastname\_StudentID.pdf); (2) separate R code file for Q4-Q5 with appropriately named readme files.
* You can also submit scanned hand written solution (should be legible, TA’s interpretation is final).
* This h/w is worth 5% of total grade + **Bonus questions account for 3% of grade**
* You can discuss with your friends, but solution should be yours.
* Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
* All submissions must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
* No makeups; for regarding policies, refer to syllabus and 1st day lecture slides.

|  |  |  |
| --- | --- | --- |
| Q# | Max Points | Your Score |
| 1 | 55 |  |
| 2 | 15 |  |
| 3 | 1% of grade (Bonus) |  |
| 4 | 30 (R mini project) |  |
| 5 | 2% of grade (Bonus) |  |

**Note**:

1. All questions and sub-parts (Q1-Q3) of this h/w require hand calculations using the formulas that you learned from the course materials. You can use any calculator.
2. For all project/programming questions, you need implementations in R and submit codes and plots (as requested). Please provide comments (in the code), and separate readme.txt file describing steps to run the code.

**Q1**. Simple Linear Regression (**55 points**)

Following table 1 show the data required to answer this question.

|  |  |
| --- | --- |
| **x** | **y** |
| 6 | 520 |
| 4 | 421 |
| 6 | 581 |
| 9 | 630 |
| 3 | 412 |
| 9 | 562 |
| 6 | 434 |
| 3 | 443 |
| 9 | 590 |
| 5 | 570 |
| 3 | 348 |
| 9 | 672 |

1. Draw 2-d scatter plot (choose appropriate scaling for x and y axis). (5 points)
2. Compute the slope and intercept of the simple linear regression equation (show all computations. **Hint**: use tabular format to compute intermediate quantities)   
   (10 points)
3. Draw resulting regression line on the 2-d scatter plot (you can copy initial plot from (a)). (5 points)
4. Compute the fitted values and residuals for each observation and verify that the residuals sum to zero (or approximately zero). (5 points)
5. How much of variation in y is explained by x? (5 points)
6. Compute the standard error of the estimate (5 points)
7. What are the predicted values for x = 2, 4, 6, 7, 10. (5 points)
8. Test for significance of “r” (linear relationship) at α = 0.05. (10 points)
9. Compute the prediction interval for same α(5 points)

**Q2**. (20 points)

1. Sate the assumptions of Simple, Multiple, and Logistic regression (if you can’t find answer in slides, you should look at any standard book and/or online; clearly cite the reference). (3x4 = 12 points).
2. Pickup any one assumption from each group (i.e., simple, multiple, and logistic) and state what happens if that assumption is violated (3x2 = 6 points). (These three assumptions should be different from each other).

**Q3.** Bonus Question (1% of grade)

In logistic regression analysis (Slide 8), we mentioned that

1. Errors can’t be normally distributed. With proper analysis (or sound arguments) show why this is the case? (5 points)
2. Error variance is not constant. Show why? (10 points)

**Mini R project (30 points; 10 points each for Q4-Q6)**

**Note:** You should provide answers (along with inline R code) in the pdf file. In addition submit R scripts.

**Q4**. Using “Advertising” data answer (a) and (b)

**(a)** Fit simple linear regression (separately) for each covariate. Provide scatter plots with fitted regression line. Which covariate provides best prediction?

**(b)** Fit multiple linear regression model for the data. Show resulting equation. How do you compare the β’s obtained with this model with corresponding β’s found in (a).

**(c)**. Fit logistic regression model for the dataset (hw3-q4c.txt). Note that this dataset contain 3 covariates, therefore you should use multiple logistic regression which is straight forward generalization of simple logistic regression (simply replace β0+β1X with β0+β1X1 + … + βnXn)

**Q5**. Apply your data science skills to **improve** the model fitted in (4.c). In what sense your improved model is **better** than the model found in (4.c). [Note the term “improve”; that is you still have to use multiple logistic model only). Show your work. (Worth **2% of final grade points**)