CSC591: Foundations of Data Science
HW2: Probability distributions, Expectation, Maximum Likelihood Estimation, Sampling Distribution, Central Limit Theorem, Confidence Intervals, Hypothesis Testing.

Released: 10/9/18
Due: 10/22/18 (23:55 PM)

Student Name:
Student ID:

Notes

- Filename: Lastname_StudentID.pdf (only pdf).
- You can also submit scanned hand-written solution (should be legible, TA's interpretation is final).
- This h/w is worth **4**% of total grade.
- *Answer all questions. A subset of the questions will be graded by the TAs* (these questions will be decided by the instructor).
- All submission 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 or bonus; for regarding policies, refer to syllabus and 1st day lecture slides.
- You are encouraged to do research, study online materials; discuss with fellow students; BUT ANSWERS SHOULD BE YOUR OWN. Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
- All questions of this h/w require hand calculations using the formulas that you learned from the course materials. You can use any regular calculator. [Note: It is important to do it by hand as you are expected to do it in exam; so, this gives you practice; show all calculations in tabular form]

Q1. Simple Linear Regression

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

x	y
6	540
4	415
6	555
9	650
3	412
9	562
6	482
3	443
9	706
5	574
3	342
9	672

(a) Draw 2-d scatter plot (choose appropriate scaling for x and y axis).
(b) Describe the relationship between x and y by looking at the scatter plot.
(c) Compute correlation between x and y
(d) Compute the slope and intercept of the simple linear regression equation (show all computations. **Hint**: use tabular format to compute intermediate quantities)
(e) Draw resulting regression line on the 2-d scatter plot (you can copy initial plot from (a)).
(f) Compute the fitted values and residuals for each observation and verify that the residuals sum to zero (or approximately zero).
(g) How much of variation in y is explained by x?
(h) Compute the standard error of the estimate
(i) Test for significance of "r" (linear relationship) at $\alpha = 0.01$.
(j) Compute the prediction interval for same α

Q2. Regression Assumptions

(a) State 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).

(b) Pickup any one assumption from each group (i.e., simple, multiple, and logistic) and state what happens if that assumption is violated (These three assumptions should be different from each other).

Q3. Logistic regression

In logistic regression analysis, we mentioned that

(a) Errors can't be normally distributed. With proper analysis (or sound arguments) show why this is the case?

(b) Error variance is not constant. Show why?

Q4. General questions

(a) When the relationship between x and y is not linear, what can you do so that SLR can still be applied to the data?

(b) Prove that, if X1 and X2 are statistically independent, Cov(X1, X2) = 0.

Q5. Least Squares

(a) Using least square technique, derive the formulas for intercept and slope for simple linear regression equation.

(b) Show that $(\overline{x}, \overline{y})$ lies on the regression line.

Q6. Hypothesis Testing

(a) Test the claim that the standard deviation of the number of aircraft stolen each year in the United States is less than 15 if a sample of 12 years had a standard deviation of 9. Use $\alpha = 0.05$ and assume normal or approximately normal distribution.

(b) Average weight of last year CS graduate students is 162.5lb, with a standard deviation of 6.9 lb. A sample of 50 students from this year batch is 165.2lb with same standard deviation. Answer the following:

a. Is there a reason to believe that there is a change in average weight of current batch of students? State your conclusion using traditional hypothesis testing using critical value of 0.05;

b. For the claim: *there is a change in average weight of current batch of students*, is there evidence to support this claim at 0.05 significance level? Use the p-value method.