Suggested Solution for Final Exam

Statistics and Data Analysis, Fall 2015

1. (20 points; 5 points each)

(a) [162.8857, 168.5643]

(b) [163.3599, 168.0901]

(c) The 95% confidence interval is larger. We need a larger interval to cover 95% than 90%, so we are more confident that the population mean will be within this interval.

(d) Since the sample size is larger than 30, we may use the *z* distribution.

Since the population is normally distributed, we may use the *t* distribution.

2. (30 points; 10 points each)

(a) Class 1: 0.387548

Class 2: 0.030974

(b) Class 1: Since the p-value is larger than 0.05, we cannot reject H_0. (Cannot reject doesn't mean accept!)

Class 2: Since the p-value is smaller than 0.05, we reject H_0. We are 95% confident that instructor 2 is teaching well that the average scores of all her/his students are higher than 70.

(c) No, because we know nothing about the performance of instructor 1.

3. (10 points; 5 points each)

(a) $scores = 67.08 + 3.42class$. The p-value of *class* is 0.2558. R-squared is 0.0329. It seems that the variable is not significant enough and the model only explains about 32% of the variance.

(b) In a regression model, the hypothesis we want to test is H0: $\beta = 0$; H1: $\beta \neq 0$. The result shows that the true p-value of each level in *class* is 0.2558/2 which is 0.1279. Since that the p-value is too high, we cannot reject the null hypothesis that the coefficient equals to 0. Thus, we are not able to make any conclusions.

4. (20 points; 5 points each)

(a) $score = 70.22 - 9.75class_2 - 3.57class_3 - 10.3class_4$

R-squared is 0.34. The p-value of $class_2, class_3$ and $class_4$ are 7.447×10^{-6}, 6.25×10^{-2} and 4.568×10^{-7}. The model explains about 34% of the variance. Except for $class_3$, other variables are significant. Compared to $class_1$, *score* decrease 9.75 for $class_2$, $score$ decrease 3.57 for $class_3$, and $score$ decrease 10.3 for $class_4$.

(b) $score = 60.47 + 9.75class_1 + 6.18class_3 - 0.55class_4$

R-squared is 0.34. The p-value of $class_1, class_3$ and $class_4$ are 7.447×10^{-6}, 1.87×10^{-3} and 7.7×10^{-1} . The model explains about 34% of the variance. Except for $class_4$, other variables are significant. Compared to $class_2$, *score* increase 9.75 for $class_1$, $score$ increase 6.18 for $class_3$, and $score$ decrease 0.55 for $class_4$.

(c) $score = 65.7 - 14.02\frac{1}{credit}$

R-squared is 0.00197. The p-value of $\frac{1}{credit}$ is 0.69. The model can just explain nothing because of a small R-squared. The p-value of coefficient is large so that it is not significant.

(d) $score = 70.125 - 7.82class_2 - 5.45class_3 - 9.06class_4 + 0.175\ genderMale - 4.61class_2 * genderMale + 3.08class_3 * genderMale - 4.09class_4 * genderMale$

R-squared is 0.39. The p-value of $class_1, class_3, class_4, genderMale, class_2 * genderMale, class_3 * genderMale, class_4 * genderMale$ are 6.7×10^{-3}, $6.18 \times 10^{-2}, 6.32 \times 10^{-4}, 9.5 \times 10^{-1}, 2.58 \times 10^{-1}, 4.13 \times 10^{-1}$ and 2.93×10^{-1}. Interaction term and *gender* (reference level: F) are not significant. We may try some other interactions and transformations. (You may set the reference level for *gender* and *class* on your own)

5. (20 points; 2 points each)

(a) F

(b) F

(c) F

(d) T

(e) F

(f) F

(g) T

(h) T

(i) T

(j) F

6. (Bonus 10 points; 5 points each)

(a) Given $\beta_0 = -1 \ and \ \beta_1 = 1$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|0-2| + |1-4| + |3-3| + |5-6| + |7-9|) = 8.$

(b) Given $\beta_0 = -1 \ and \ \beta_1 = 0$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|-1-2| + |-1-4| + |-1-3| + |-1-6| + |-1-9|) = 29.$

Given $\beta_0 = -1 \ and \ \beta_1 = 0.5$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|-0.5-2| + |0-4| + |1-3| + |2-6| + |3-9|) = 18.5.$

Given $\beta_0 = 0 \ and \ \beta_1 = 0$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|0-2| + |0-4| + |0-3| + |0-6| + |0-9|) = 24.$

Given $\beta_0 = 0 \ and \ \beta_1 = 0.5$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|0.5-2| + |1-4| + |2-3| + |3-6| + |4-9|) = 13.5.$

Given $\beta_0 = 0 \ and \ \beta_1 = 1$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|1-2| + |2-4| + |4-3| + |6-6| + |8-9|) = 5.$

Given $\beta_0 = 1 \ and \ \beta_1 = 0$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|1-2| + |1-4| + |1-3| + |1-6| + |1-9|) = 19.$

Given $\beta_0 = 1 \ and \ \beta_1 = 0.5$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|1.5-2| + |2-4| + |3-3| + |3-6| + |4-9|) = 8.5.$

Given $\beta_0 = 1 \ and \ \beta_1 = 1$, SAE($\beta_0$, β_1) = $\sum_{i=1}^{5} |\hat{y}_i - y_i| = \sum(|2-2| + |3-4| + |5-3| + |7-6| + |9-9|) = 4.$

The wining combination that minimize SAE(β_0, β_1) is (β_0, β_1) = (1, 1). SAE(1, 1) = 4.