Summative Quiz 1 (Simple Linear Regression) Solutions:

1. An article on the American Journal of Public Health reported on a study to investigate the relationship between school start time and adolescent sleep patterns.[1]

   The researchers used *linear regression* methods, including simple linear regression to analyze these data. Here are some adapted linear regression results from this study:

   $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, where y = typical weeknight sleep duration in minutes, and x_1 = school start time, measured as the number of minutes after 7 AM that the school starts. (So for example for 7:30, x_1= 30; for 8:15, x_1 = 75 etc..) The estimates and standard errors for the regression slope is:

   $\hat{\beta}_1 = 0.47;\ SE(\hat{\beta}_1) = 0.12\ ;\ \hat{\beta}_0 = 433.9;\ SE(\hat{\beta}_0)$

   What is the estimated average nightly sleep duration for students with a start time of 8:15 AM? (rounded to the nearest integer value)

   Answer: 469 minutes.

   Reasoning: The simple linear regression equation is $\hat{y} = 433.9 + 0.47x_1$, and 8:15 AM is 75 minutes after 7:00 AM. This equation evaluated at x_1=75 give an estimated mean sleep duration of $\hat{y} = 433.9 + 0.47(75) = 469.15 \approx 469$ minutes.

2. (this item references the same linear regression results as item #1)

   Is the relationship between *y* and x_1 statistically significant (α=0.05)?

   Answer: yes

   Reasoning: The estimated 95% CI for the slope β_1 is given by $\hat{\beta}_1 \pm 2\widehat{SE}(\hat{\beta}_1) \rightarrow 0.47 \pm 2(0.12) \rightarrow (0.23, 0.71)$. The null value for a slope (from any regression type, not just linear) is 0, and this value is not included in the resulting confidence interval.

[1] Paksarian D, Rudolph KE, He JP, and Merikangas KR. School start times and adolescent sleep patterns: results from the US National Comorbidity Survey—Adolescent Supplement. *American Journal of Public Health* (2015). 105 (7)

3. (this item references the same linear regression results as item #1)

   What is the average difference in nightly sleep duration for students who start at 8:15 AM compared to those who start at 7:30 AM?

   Answer: 21.2 minutes

   Reasoningl: This is a mean difference in the outcome of sleep duration for two groups who differ by 45minutes in the predictor, start time. As the slope of 0.47 is the estimated mean difference in $\hat{y}$ for two groups of students who have a 1 minute difference in start time, 0.47*45 = 21.15 ≈ 21.2 minutes is the estimated mean difference for two groups of students whose schools' start time differ by 45 minutes.

4. (this item references the same linear regression results as item #1)

   What is the 95% CI for average difference in nightly sleep duration for students who start at 8:15 AM compared to those who start at 7:30 AM?

   Answer: (10.4, 32.0) minutes

   Reasoning: $(0.47 * 45) \pm 2 * (0.47)(0.12) = (10.4, 32.0)$ minutes

5. (this item references the same linear regression results as item #1)

   What information would you need to see to assess whether the association between sleep duration and school start time is approximately linear?

   Answer: A scatterplot of the individual students sleep time and school start time values.

6. The 2010 National Hospital Ambulatory Medical Care Survey (NHAMCS) is a national (United States) sample survey of visits to hospital outpatient and emergency departments. This survey was conducted by the National Center for Health Statistics. In this exercise, simple linear regression will be used to examine factors associated with patient waiting time (*in minutes*) of persons admitted to the Emergency Departments (EDs) of participating hospitals in 2010. The average waiting time reported by the over 27,000 survey participants is 56.3 minutes (s= 78.7 minutes, with a range of 0 to 1,335 minutes).

   To start, a simple linear regression was performed to relate average ED waiting times to subject's sex, with x_1 coded as 1 for male and 0 for female:

   $\hat{y} = 57.4 + -2.5x_1$, and the estimated standard error of the slope for sex is 0.95. The R^2 for this regression is < .01.

   What is the mean waiting for females in the sample?

   Answer: 57.4 minutes

   Reasoning: x_1 = 0 for females, so the estimate mean for this group is just the intercept from this simple linear regression model: $\hat{y} = 57.4 + -2.5(0) = 57.4\ minutes.$

7. (This item references the same linear regression results as item #6)

   What is the mean waiting time for males in the sample?

   Answer: 54.9 minutes

   Reasoning: x_1 = 1 for males, so the estimate mean for this group is given by: $\hat{y} = 57.4 + -2.5(1) = 54.9\ minutes.$

8. (This item references the same linear regression results as item #6)

   Suppose x_1 had instead been coded as 1 for females, and 0 for males. What would the intercept value be in the regression of waiting time on x with this new coding schema (1 for females, 0 for males)?

   Answer: 54.9 minutes

   Reasoning: With this alternate coding schema for x_1, males now represent the reference group, and the intercept for the corresponding regression model is now an estimate of the mean waiting time for this reference group, i.e. males.

9. Also of interest is whether there are any differences in the waiting time distributions by self –reported race of the survey participants. Participants classified themselves into 1 of 3 racial categories: while, black, other (did not identify as either black or white). Below are the results from a simple regression performed to relate waiting time to race:

   $\hat{y} = 51.5 + 19.3x_1 + 2.6x_2$ where x_1 = 1 if the respondent identifies as black (and 0 if not), and x_2=1 if the participant identifies as other (0 if not). The standard error of the slope for x_1 is 1.2, and for x_2 is 2.5. The R^2 is 0.01.

   Answer: 70.8 minutes

   Reasoning: For this group, x_1=1, x_2 =0, and hence the estimated mean waiting time is given by:

$$\hat{y} = 51.5 + 19.3(1) + 2.6(0) = 70.8\, minutes.$$

10. (This item references the same linear regression results as item #9)

    Based on these regression results, what is the estimated mean difference in waiting times (and 95% CI) for Blacks compared to Whites?

    Answer: 19.3 minutes; (16.9, 21.7) minutes

    Reasoning: The slope estimate for the indicator x_1, which is an indicator of whether the subject identifies as black is an estimate of the mean difference in waiting times between those who identify as black, and the reference group (those who identify as white): this slope estimate is 19.3 minute. The corresponding 95% CI is estimated via 19.3±2*1.2 →(16.9, 21.7)

11. (This item references the same linear regression results as item #9)

    Based on these regression results, what is the estimated mean difference in waiting times for Black compared to Other? (Black - Other)?

    Answer: 16.7 minutes

    Reasoning: Based on the regression results, the estimated mean difference in waiting time between the black and white groups is the slope for x_1, 19.3 minutes. Similarly,

the estimated mean difference in waiting time between the other and white groups is the slope for x_1, 2.6 minutes. The estimated mean difference between the black and other groups can be expressed as (black- white) – (other -white) = 19.3 – 2.6 = 16.7 minutes.

12. (This item references the same linear regression results as item #9)

    What is the interpretation of the R^2 value?

    Answer: 1% of the variation in waiting times is explained by race.