Quiz 1

BSTA 512, Winter 2024

Jan 29, 2024

Name: Answer Key

Instructions

There are **7 total pages** in the exam and **12 questions** (9 multiple choice and 3 free response). Please make sure you have all of the pages!

1. I have written a "30 minute" quiz. However, you have 50 minutes from 2:00 - 2:50pm.
2. The quiz is open book and open notes. You may use books other than the class textbook, you may use anything on our course webpage, and you may use reference websites (like Wikipedia, Googling expected value of specific distribution, etc.).
3. No cheating will be tolerated. If one person is caught cheating, I will need to reconsider how to administer Quiz 2. Cheating includes:
    - Using ChatGPT
    - Using question and answer threads typically seen on sites like StackExchange, WikiHow, Quora, Reddit, StackOverflow, Chegg, etc.
    - Asking other students in the room or looking at other students' quiz work.
4. Each multiple choice question is worth 3 points. The free response questions are labelled with their point value.
5. You may use headphones during the quiz.

Grading

For Nicky to fill out:

Question	Points	Potential Points
Multiple Choice Questions		27
Free Response Questions (#8-10)		13
Total		40

Questions

1. The population equation for simple linear regression is:

   a. $Y = \beta_0 + \beta_1 X$

   b. $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

   [✓ c.] $Y = \beta_0 + \beta_1 X + \epsilon$

   d. $Y = \beta_1 + \beta_0 X + \epsilon$

2. What are residuals in the context of simple linear regression?

   a. The expected values of the outcome (Y)

   [✓ b.] The differences between the observed outcome and fitted values of the outcome Y given X

   c. The coefficient estimates in the regression line

   d. The differences between the observed outcome and mean of the outcome Y

3. If we wanted to create a variable that is the difference between two other columns, what R command would we use?

   a. select()

   b. pivot_longer()

   c. filter()

   [✓ d.] mutate()

4. Which fo the following is the output of the `R` function `pnorm()` with its default settings?

   a. The critical value, using the Normal distribution, corresponding to the probability value you input

   b. A list of sampled values from a Normal distribution that has the mean a standard deviation you input

   c. The probability, using the Normal distribution, of getting a value greater than the value (quantile) you input

   [✓ d.] The probability, using the Normal distribution, of getting a value less than the value (quantile) you input

5. What is the primary objective for ordinary least squares in simple linear regression?

   a. To maximize the correlation coefficient between the dependent and independent variables.

   [✓ b.] To minimize the sum of squared differences between the observed outcome and expected values of the outcome Y given X.

   c. To evaluate if linear regression is appropriate for fitting the data.

   d. To derive the equations for the population slope and intercept that fit the data best.

6. What does the symbol $\hat{\sigma}^2$ represent within linear regression?

   a. The estimate of the population slope of the regression line

   [✓ b.] The estimate of the population variance of the residuals

   c. The variance of the population model residuals

   d. The estimate of the population standard deviation of the residuals

Questions 7-12 will reference the below scatterplot that includes the data points and best-fit line (shown with the red line). More information on the red line is given below the scatterplot. We are looking at the relationship between peak exercise heart rate (bpm) and age (years). This is simulated data for individuals 40 to 85 years old.

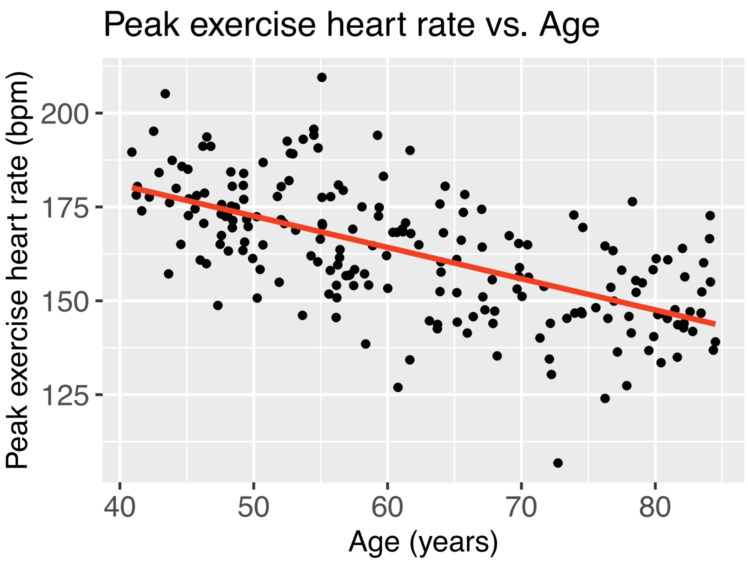

Figure 1: Scatterplot of simulated data with best-fit regression line

Here is regression table for the best fit (red) line as well:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	214.233	4.724	45.353	0	204.918	223.548
~~x~~ Ag	-0.834	0.075	-11.098	0	-0.982	-0.685

The SSE of the red line is: 22054.38

7. Which of the following statements is true about the value -0.834 in our regression table?

   a. It is the estimate of the sample intercept

   b. It is the estimate of the population intercept

   c. It is the estimate of the sample slope

   [✓ d. It is the estimate of the population slope]

8. (3 points) Please write out the equation for the fitted, red line. Please include the ~~estimated intercept and slope~~ values from the regression table.

$$\hat{Y} = 214.233 - 0.834X$$

- OR -

$$Y = 214.233 - 0.834X + \hat{\varepsilon}$$

$\hat{Y}$ can be replaced by:

- $\widehat{HR}$
- $\hat{E}(Y|X)$
- $\widehat{E(Y|X)}$
- $\widehat{\text{peak HR}}$
- or something similar

0.5 pt : X no hat   0.5 pt: correct values

0.5 pt: (Y has hat + no ε) or (Y no hat + ε w/hat)   1pt: X & Y in right spots

9. (5 points) Please interpret the **intercept** from the red line for the relationship between peak exercise hear rate and age (include the 95% confidence interval).

For someone 0 years old, the expected/estimated/average peak exercise heart rate is 214.233 beats per minute (95% CI: 204.918, 223.548).

1pt   1pt   0.5   1pt   0.5 pt   1pt

10. (5 points) Please interpret the **slope** from the red line for the relationship between peak exercise hear rate and age (include the 95% confidence interval).

For every one year increase in age, the expected/estimated/average peak exercise heart rate decreases, on average, by 0.834 beats per minute (95% CI: -0.982, -0.685).

0.5pt   0.5pt   1 pt   0.5pt   1pt   0.5pt   1pt

11. Using the confidence intervals and above interpretation of the **slope**, what would be the conclusion of a hypothesis test for $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$?

[✓ a.] Reject the null because the confidence interval does not include 0.

b. Reject the null because the confidence interval includes 0.

c. Fail to reject the null because the confidence interval does not include 0.

d. Fail to reject the null because the confidence interval includes 0.

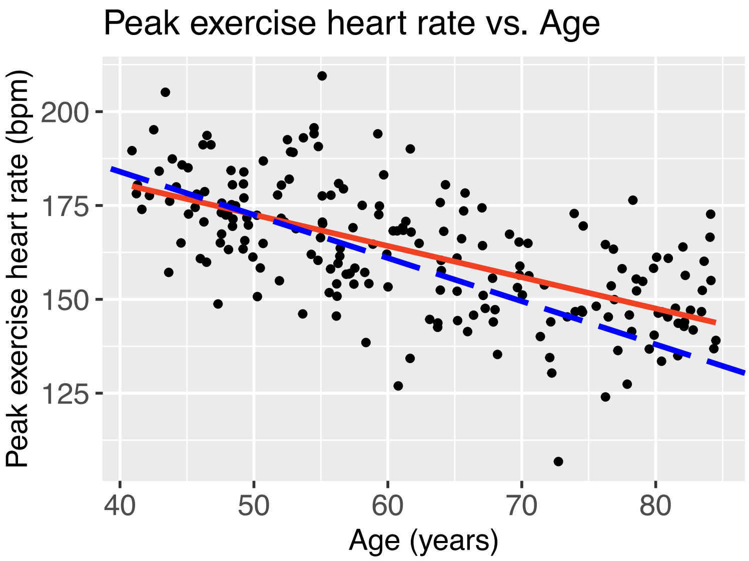

Figure 2: Added blue dashed line for question 12

12. If I decide to throw out the red line (still pictured), and draw the blue dashed line in Figure 2 above. What do I know abut the blue dashed line?

    a. Every individual observed data point has a larger residual value with the blue line (compared to the red line).

    b. The blue dashed line is now the best-fit line.

    [✓ c.] The sum of square errors is greater than 22054.38.

    d. The blue dashed line is fit with ordinary least squares.