Math 455 **Midterm 1** **Spring 2026**

No books, no notes, no calculators.

Name: _____

Page	Points	Score
2	13	
3	16	
4	12	
5	18	
6	20	
7	16	
8	19	
9	14	
10	20	
11	12	
Total:	160	

1. The standard simple linear regression model assumes an approximate linear relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where the ϵ_i are normal random variables with mean 0 and variance σ^2

(a) (2 points) What else, if anything, do the standard assumptions of linear regression say about the ϵ_i?

(b) (6 points) The sum $\epsilon_1 + \cdots + \epsilon_n$ is a random variable. What can be said about the distribution of this random variable? (Give the name of the distribution and the values of any relevant parameters that define it.)

2. Consider the following regression summary output from a study examining factors affecting the sale price of houses in a specific neighborhood. The variables considered are:

- price, the sale price of the house in *thousands* of dollars,
- size, the square footage of the house in *hundreds* of square feet, and
- age, the age of the house in years.

```
Call:
lm(formula = price ~ size + age, data = housing)

Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     120.000      12.000   10.000   2.1e-10 ***
size             15.000       3.000    5.000   1.5e-05 ***
age              -2.000       0.500   -4.000   0.00032 ***
```

(a) (5 points) What is the predicted sale price (in *dollars*) for a house that is 2,000 square feet and is 10 years old? (Hint: note the units of measurement in the regression description above.)

(b) (8 points) Write a sentence interpreting the coefficient of `age` in this regression, comprehensible to a non-statistician.

3. We fit a regression model with two independent variables `X1` and `X2` in `R`, and we have only $n = 6$ rows of data.

```
> lmod1 <- lm(Y ~ X1 + X2, data = df)
> summary(lmod1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0357     3.6093   1.949    0.146
X1            0.3420     0.4562   0.750    0.508
X2            1.6580     1.0599   1.564    0.216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2557 on 3 degrees of freedom
Multiple R-squared:  0.9965,Adjusted R-squared:  0.9942
F-statistic: 432.3 on 2 and 3 DF,  p-value: 0.0002033
```

(a) (4 points) In terms of n, the number of rows of data, and p, the number of parameters in the regression model, how many numerator and denominator degrees of freedom does the F-statistic have (in general; your answer should be in terms of n and p). Hint: be sure to check that your answer is consistent with the 2 and 3 in the output above.

(b) (4 points) The p-value `0.216` associated with the coefficient of `X1` is the result of a hypothesis test. What are the null and alternative hypotheses of this test?

(c) (4 points) The coefficient of `X2` is not statistically significant (its p-value is 0.216), yet the overall F-test is highly significant (p-value = 0.0002). A student says: "This is a contradiction — if the model is significant overall, all of its predictors should be significant individually." Is the student correct? Explain why or why not.

(d) (6 points) Consider the following sequence of steps that `R` performs to produce the entry `0.216` in the row for `X2`:

1. Compute the estimate `1.6580`.
2. Compute the standard error `1.0599`.
3. Compute the t-value `1.564`.
4. Compute the p-value `0.216`.
5. Declare the result not statistically significant at level $\alpha = 0.10$.

Which, if any, is the first of these steps to require the assumption that the errors ϵ_i are normally distributed, and how is that assumption used?

4. Suppose we have a linear regression model with 2 independent variables U, V and n data points, so the defining equation is

$$Y_i = \beta_0 + \beta_1 U_i + \beta_2 V_i + \epsilon_i$$

with the usual assumptions on the errors ϵ_i. The textbook tells us that linear regression is an orthogonal projection, which is a linear map.

(a) (2 points) What is the domain of this linear map?

(b) (4 points) What is the codomain (also sometimes called the target or range) of this linear map?

(c) (6 points) How can we describe the matrix of this linear map? (Hint: note that it is not sufficient to write a formula with X in it, because no X is defined in this problem!)

5. (4 points) You obtain a list of 4 samples y_1, y_2, y_3, y_4 from a random variable Y which has either the t-distribution or the F-distribution (degrees of freedom unknown).

   This list turns out to be $0.34, 4.36, -0.93, 1.37$. Is the random variable Y more likely to have the t- or F-distribution? Why?

6. Suppose we define two matrices in `R` by

```
A <- matrix(-1, nrow = 2, ncol = 2)
B <- matrix(1, nrow = 2, ncol = 2)
```

   (a) (2 points) What is the matrix `A` (that is, the output of `print(A)`)?

   (b) (2 points) What is the matrix `B` (that is, the output of `print(B)`)?

(c) (4 points) What will the output of `print( A * B )` be?

(d) (4 points) What will the output of `print( A %*% B)` be?

7. (4 points) Fill in the blanks in the following definition of level (of significance) and type I error:

   A type I error is made if _____ is _____ when H_0 is _____.

   The _____ (assuming that H_0 is true) is denoted by α and is called the level of the test.

8. Let X_1, X_2, X_3, X_4 be independent and identically distributed random variables which are all normal with mean 0 and variance 1. Let $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$. Let $W = \sum_{i=1}^{4}(X_i - \bar{X})^2$.

   (a) (4 points) What is the distribution of W? (Give the values of any parameters that are relevant.)

   (b) (4 points) Notice that $\bar{X}$ appears in the formula defining W. Are $\bar{X}$ and W independent random variables? How do you know?

9. A study was conducted to investigate factors affecting the daily number of steps walked by adults. Thirty individuals were tracked over a 2-week period using a fitness monitor, and the average number of steps walked per day, Y, was recorded for each. Four additional variables were recorded for each individual: x_1 = age, x_2 = body mass index (BMI), x_3 = hours of sedentary work per day, and x_4 = self-reported motivation score (on a scale of 1–10). Consider the three models given below:

$$\begin{aligned} \text{Model I:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \\ \text{Model II:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \\ \text{Model III:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3 x_4 + \epsilon \end{aligned}$$

Indicate whether each of the following statements is true or false, and give a short reason for each answer.

(a) (4 points) Models I and II can be compared using an F-test.

(b) (4 points) After fitting Models I and II and computing their sums of squared errors, $\text{SSE}_{\text{I}} \geq \text{SSE}_{\text{II}}$.

(c) (4 points) Models II and III can be compared using an F-test.

(d) (4 points) After fitting Models II and III and computing their R^2 values, $R^2_{\text{III}} \geq R^2_{\text{II}}$.

10. Suppose X has the uniform distribution on $[-1, 1]$ and $Y = X^2$.

(a) (1 point) Find $E[X]$

(b) (2 points) Find $E[Y]$

(c) (2 points) Are X and Y independent?

(d) (4 points) Find $\text{Cov}(X, Y)$.

(e) (6 points) Suppose we fit a simple regression $Y = \beta_0 + \beta_1 X + \epsilon$ using $n = 9999$ i.i.d. draws from this joint distribution. What approximate values do you expect for $\hat{\beta}_0$ and $\hat{\beta}_1$?

(f) (4 points) What approximate value would you expect for R^2 in this regression?

11. Suppose we have 3 variables `Y`, `X`, and `Z` defined in `R`. We consider 3 linear models defined by

```
lmod1 <- lm(Y ~ X + Z)
lmod2 <- lm(Y ~ -1 + I(X+Z))
lmod3 <- lm(Y ~ X + offset(2*Z))
```

The equation that defines the model `lmod1` is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

(a) (4 points) What equation defines `lmod2`?

(b) (4 points) What equation defines `lmod3`?

(c) (6 points) If we execute the command `anova(lmod1, lmod2)` we see the output:

```
> anova(lmod1,lmod2)
Analysis of Variance Table

Model 1: Y ~ X + Z
Model 2: Y ~ -1 + I(X + Z)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1      7 13.719
2      9 17.967 -2   -4.2486 1.084  0.389
```

We see a p-value (`0.389`) in this output, which comes from a statistical test. What are the null hypothesis and the alternative hypothesis of this statistical test?

Multiple Choice: Circle the letter corresponding to the best response.

12. (4 points) Which of the following best describes the Hill criterion of *strength of association*?

    (A) The association between the exposure and outcome must be statistically significant at the $\alpha = .05$ level.

    (B) A large effect size makes it less likely that the association is due to confounding or bias alone.

    (C) The association must be replicated in at least two independent studies before it can be considered causal.

    (D) The exposure must precede the outcome in time.

13. (4 points) A study finds that individuals who smoke more cigarettes per day have a higher risk of lung cancer, and that those who quit smoking see their risk decrease over time. Which Hill criterion does this most directly address?

    (A) Consistency

    (B) Gradient

    (C) Specificity

    (D) Coherence

14. (4 points) A student argues: "Our regression model shows a highly significant association between X and Y with a very small p-value, so by the Hill criterion of strength of association, we have strong evidence of a causal relationship." What is the most important flaw in this argument?

    (A) A small p-value indicates a large effect size, which is not the same as a causal relationship.

    (B) Statistical significance reflects sample size as much as effect size, and in any case no single Hill criterion is sufficient to establish causation on its own.

    (C) The student should have used a confidence interval rather than a p-value to assess strength of association.

    (D) Regression models cannot be used to assess the Hill criteria because they are designed for prediction, not causal inference.

15. (4 points) In a least-squares regression model with an intercept, expressed in matrix notation as $Y = X\beta + \epsilon$, the residual vector e is always orthogonal to which of the following?

    (A) The response vector Y

    (B) The fitted values vector $\hat{Y}$

    (C) The column space of the design matrix X

    (D) The null space of $X^T X$

16. (4 points) The hat matrix H in a regression is defined as

$$H = X(X^T X)^{-1} X^T.$$

    What are the properties of H?

(A) It is symmetric and idempotent.

(B) It is diagonal.

(C) It is always invertible.

(D) It has eigenvalues strictly greater than zero.

17. (4 points) Which of the following best explains why a prediction interval is wider than a confidence interval?

(A) A prediction interval accounts for both the uncertainty in the estimated regression function and the natural variability in future observations.

(B) A prediction interval uses a smaller significance level than a confidence interval.

(C) A confidence interval assumes that future observations will be close to the mean response.

(D) A prediction interval is based on a different test statistic than a confidence interval.

18. (4 points) What is the primary source of additional uncertainty in a prediction interval compared to a confidence interval?

(A) Sampling variability in estimating $\hat{\beta}$

(B) The residual variance, which accounts for random variation in individual observations

(C) The presence of collinearity in the predictors

(D) The need to estimate the variance of the independent variables

19. (4 points) In a linear regression analysis with n rows of data, with the usual assumptions, which of the following quantities can never be zero?

A) The estimated intercept, $\hat{\beta}_0$

B) The first residual e_1

C) The last fitted value, $\hat{Y}_n$

D) All of the above quantities can be zero.