DS3010 Midterm 1 (Practice)

Total: 100 Points

Name: ____

Part 1: Conceptual Questions (50 points)

A. Single Choice (5 points each)

1. In multiple regression, what does β_j represent?

   A. The correlation between X_j and Y

   B. The change in Y for one-unit increase in X_j, holding other predictors constant

   C. The predicted value of Y

   D. The variance of X_j

2. As model flexibility increases, what generally happens?

   A. Bias increases, variance decreases

   B. Bias decreases, variance increases

   C. Both increase

   D. Both decrease

3. What is the purpose of the overall F-test?

   A. Test if $\beta_0 = 0$

   B. Test if exactly one predictor is significant

   C. Test if at least one predictor has a non-zero coefficient

   D. Identify the most important predictor

B. Multiple Choice (7 points each)

1. Suppose the true relationship between X and Y is linear. You compare a simple linear regression model to a highly flexible polynomial model (degree 10). Which of the following statements are generally true?

   A. The polynomial model will have training MSE less than or equal to the linear model.

   B. The polynomial model will always have lower test MSE than the linear model.

   C. The polynomial model is more likely to have higher variance than the linear model.

D. The linear model is more likely to have higher bias than the polynomial model.
E. Increasing model flexibility always decreases both bias and variance.

2. Which are components of reducible error?

A. Bias
B. Variance
C. Irreducible noise
D. Estimation error

C. True / False (3 points each)

1. True or False: A small p-value means the null hypothesis is probably true.
2. True or False: LOOCV produces identical results on repeated runs with the same dataset.
3. True or False: Given the same X values, confidence intervals are wider than prediction intervals.
4. True or False: Increasing the sample size generally reduces the variance of a regression model.

D. Short Answer (9 points)

1. Suppose Model A (linear) has higher training MSE but lower test MSE than Model B (flexible non-linear). Explain what this tells you about bias and variance.

DS3010 Midterm 1 (Practice)

Total: 100 Points

Name: ____

Part 2: Model Complexity and Cross-Validation (50 points)

Suppose the true model is:

$$Y = 1 + 1X_1 + 2X_1^2 + 3X_1^3 + \epsilon, \quad \epsilon \sim N(0, 1)$$

1. (10 points) Generate a dataset with 500 observations. Use the following code to generate the dataset:
```
set.seed(123)
X1 < - seq(0, 5, length.out = n)
```

2. **(40 points)** Use LOOCV to estimate the test MSE for polynomial degrees 1–5.

   (a) (25 points) Plot *or* report the LOOCV test MSE for each degree.

   (b) (5 points) Which polynomial degree performs best (i.e., has the smallest LOOCV test MSE)?

   (c) (10 points) Provide a brief justification for your choice

3. (**Extra credit: 10 points**) Apply 10-fold cross-validation to estimate the test MSE for polynomial degrees 1-5. Report the 10-fold test MSE, which model has the smallest expected test MSE?