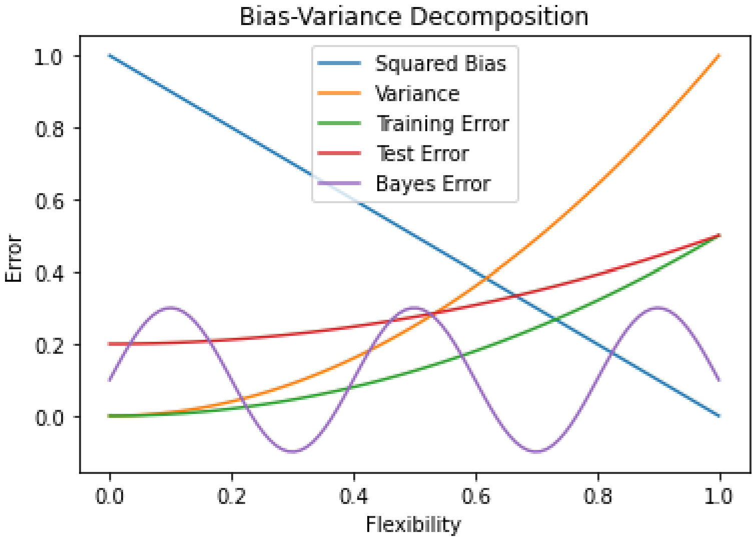
1. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(b) Explain why each of the five curves has the shape displayed in question (1a).

Squared Bias:

- Initially high because simpler models have insufficient complexity to capture the underlying relationships in the data.
- Decreases with flexibility as more complex models can better fit the data.
- Stabilizes because increasing model complexity beyond a certain point doesn't significantly reduce bias and may introduce noise.

Variance:

- Begins low because simpler models have fewer parameters and less capacity to vary.
- Increases with flexibility as more complex models can capture more variance in the data.
- Stabilizes or slightly decreases at very high flexibility due to overfitting, where the model starts capturing noise instead of the underlying patterns.

Training Error:

- Initially high because simpler models struggle to capture the complexity of the data.

  - Decreases with flexibility as more complex models fit the data better.

Test Error:

  - High initially due to bias, decreases as flexibility increases because of reduced bias and better fit to the data.
  - May start increasing due to overfitting as the model becomes too flexible and starts capturing noise instead of true patterns.

Bayes Error:

  - Remains constant regardless of flexibility because it represents the inherent noise in the data, which cannot be reduced by any model.
  - It serves as a theoretical lower bound for the achievable error.

2. (ISLP: Chapter 2, Question 3) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Advantages of a very flexible approach:

- Better Fit: Very flexible models can capture complex relationships in the data, potentially leading to lower bias and better predictive accuracy.
- Higher Performance: In cases where the underlying relationship between features and target is highly nonlinear or complex, a flexible model may outperform less flexible alternatives.
- Adaptability: Flexible models can adapt well to different types of data and may perform better across a wide range of datasets.

Disadvantages of a very flexible approach:

- Overfitting: There's a higher risk of overfitting, where the model learns noise in the training data rather than true patterns, leading to poor generalization to unseen data.
- Computational Complexity: Very flexible models often require more computational resources and time for training and inference.
- Interpretability: Highly flexible models are often more complex, making them harder to interpret and understand compared to simpler models.

Circumstances favoring a more flexible approach:

- Complex Relationships: When the underlying relationship between features and target is highly nonlinear or intricate, a flexible model can capture these complexities more effectively.

- Large Datasets: With large datasets, flexible models can better utilize the available information to capture intricate patterns without overfitting.

Circumstances favoring a less flexible approach:

- Interpretability: When interpretability of the model is crucial (e.g., in medical or legal contexts), simpler models are preferred because they are easier to interpret and explain to stakeholders.
- Limited Data: With limited data, simpler models are less prone to overfitting and may generalize better to unseen instances.
- Computational Constraints: In situations where computational resources are limited, less flexible models may be preferred due to their lower computational complexity and faster training and inference times.

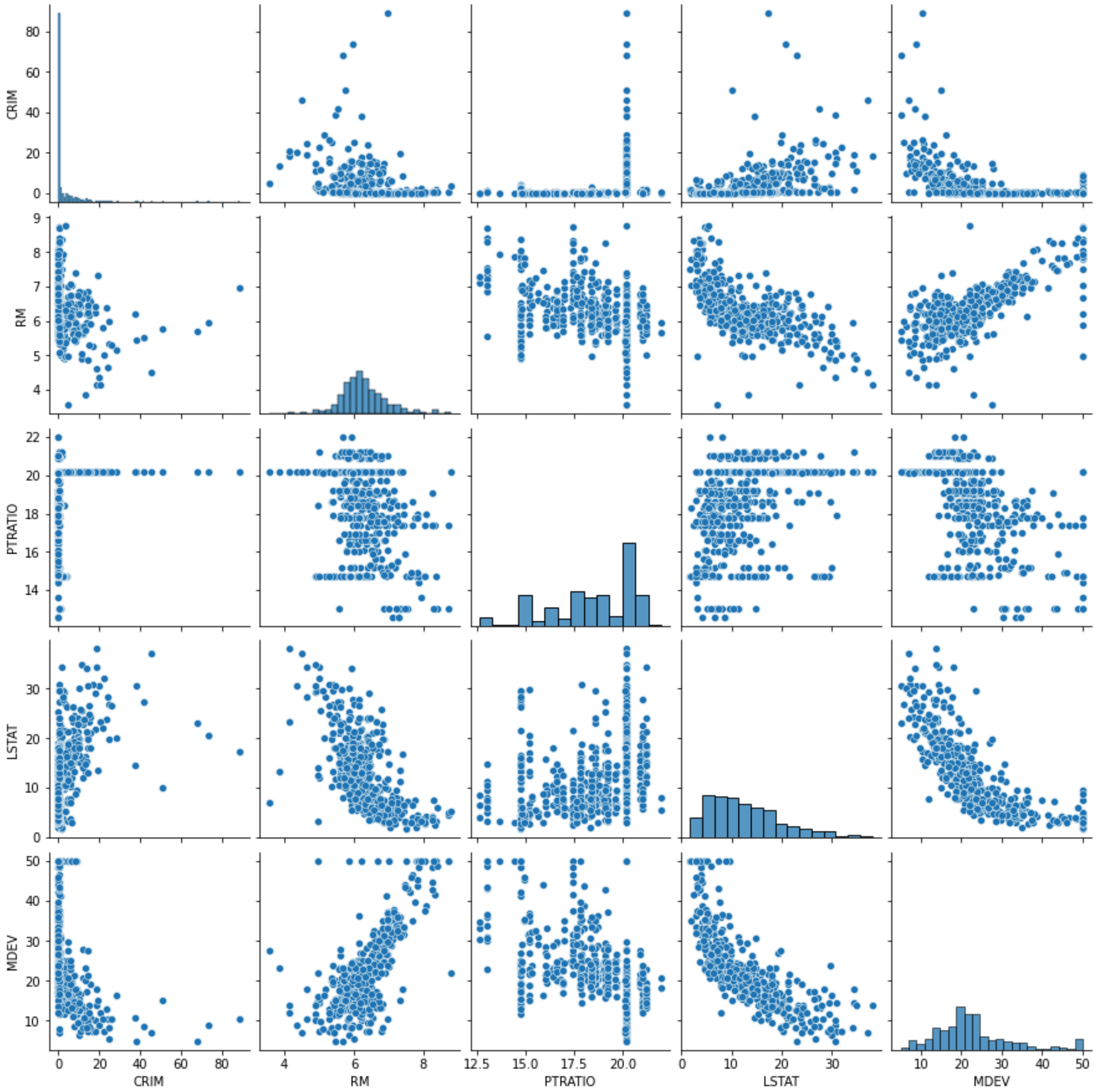
3. (ISLP: Chapter 2, Question 10) This exercise involves the Boston housing data set.4

(a) To begin, load in the Boston data set, which is available on Canvas and/or can be loaded as part of the ISLP library.

(b) How many rows are in this data set? How many columns? What do the rows and columns represent?

Number of rows: 506
Number of columns: 14

(c) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



The observed pattern in the CRIM vs. MEDV scatterplot suggests a negative correlation: as the per capita crime rate (CRIM) increases, the median value of owner-occupied homes (MEDV) tends to decrease. This aligns with expectations, indicating that areas with higher crime rates typically have lower home values.

In the RM vs. MEDV scatterplot, with the average number of rooms (RM) on the x-axis and the median value of owner-occupied homes (MEDV) on the y-axis, most points cluster around an RM value of 6. As RM increases, MEDV tends to increase as well.

In the LSTAT vs. MEDV scatterplot, with the percentage of lower-status population (LSTAT) on the x-axis and MEDV on the y-axis, lower LSTAT values correspond to higher MEDV values.

(d) Are any of the predictors associated with per capita crime rate? If so, explain the Relationship.

```
CRIM       1.000000
RAD        0.622029
TAX        0.579564
LSTAT      0.452220
NOX        0.417521
INDUS      0.404471
AGE        0.350784
PTRATIO    0.288250
CHAS      -0.055295
ZN        -0.199458
RM        -0.219940
B         -0.377365
DIS       -0.377904
MDEV      -0.385832
Name: CRIM, dtype: float64
```

The per capita crime rate (CRIM) is positively correlated with variables such as RAD (index of accessibility to highways), TAX (property-tax rate), LSTAT (% lower status of the population), NOX (nitric oxides concentration), INDUS (proportion of non-retail business acres), and AGE (proportion of owner-occupied units built before 1940). This suggests that areas with higher rates of highway accessibility, higher property-tax rates, and lower socioeconomic status tend to have higher crime rates. Additionally, factors such as the pupil-teacher ratio (PTRATIO) and location not bordering the Charles River (CHAS) also show moderate positive correlations with crime rates. Conversely, variables like RM (average number of rooms) and MEDV (median home value) exhibit weak negative correlations with CRIM, indicating that higher property values and larger homes may be associated with lower crime rates.

(e) Do any of the Boston census tracts appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
Range of Crime Rates (CRIM): 88.96988
Range of Tax Rates (TAX): 524.0
Range of Pupil-Teacher Ratios (PTRATIO): 9.4
```

The range of crime rates (CRIM) across Boston census tracts is approximately 88.97 per capita, while the range of tax rates (TAX) is 524 per $10,000. Pupil-teacher ratios (PTRATIO) vary by up to 9.4 across census tracts. These wide ranges indicate notable differences in crime rates, tax rates, and educational resources among different areas in Boston.

(f) How many of the census tracts in this data set bound the Charles river?

Number of census tracts that bound the Charles River: 35.0

(g) What is the median pupil-teacher ratio among the census tracts in this data set?

Median pupil-teacher ratio among census tracts: 19.05

(h) Which Boston census tract has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Values of other predictors for the census tract with the lowest MDEV:

```
CRIM        38.3518
ZN           0.0000
INDUS       18.1000
CHAS         0.0000
NOX          0.6930
RM           5.4530
AGE        100.0000
DIS          1.4896
RAD         24.0000
TAX        666.0000
PTRATIO     20.2000
B          396.9000
LSTAT       30.5900
Name: 398, dtype: float64
```

Overall ranges for the predictors:
Range of CRIM: 0.00632 - 88.9762
Range of ZN: 0.0 - 100.0

Comment:
The census tract with the lowest MDEV has the following values for the predictors:

```
CRIM       38.3518
ZN          0.0000
INDUS      18.1000
CHAS        0.0000
NOX         0.6930
RM          5.4530
AGE       100.0000
DIS         1.4896
RAD        24.0000
TAX       666.0000
PTRATIO    20.2000
B         396.9000
LSTAT      30.5900
Name: 398, dtype: float64
```

These values can be compared to the overall ranges for those predictors to assess how they deviate from typical values.

(i) In this data set, how many of the census tract average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

Number of census tracts that average more than seven rooms per dwelling: 64
Number of census tracts that average more than eight rooms per dwelling: 13

Comment:
Census tracts that average more than eight rooms per dwelling represent areas with larger, potentially more upscale homes. These areas may be characterized by higher socioeconomic status and property values.

4. (ISLP: Chapter 3, Question 3) Suppose we have a data set with five predictors, $X1$ =GPA, $X2$ =IQ, $X3$ =Level (1 for College and 0 for High School), $X4$ = Interaction between GPA and IQ, and $X5$ = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get

$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, and $\hat{\beta}_5 = -10$.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 Level + \hat{\beta}_4 (GPA \times IQ) + \hat{\beta}_5 (GPA \times Level)$$

(a) Which answer is correct, and why?
i. For a fixed value of IQ and GPA, high school graduates earn more, on average,

than college graduates.
ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates. This is because the coefficient of the level predictor (β3) is positive (35), indicating that college graduates earn more than high school graduates on average after controlling for other factors like GPA and IQ. The interaction term coefficient (β5) is negative (-10), but it does not change the interpretation of the main effect of the level predictor.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

Salary = β0 + β1 * GPA + β2 * IQ + β3 * Level + β4 * (GPA * IQ) + β5 * (GPA * Level)

Plugging in the values:

Salary = 50 + 20 * 4.0 + 0.07 * 110 + 35 + 0.01 * (4.0 * 110) + (-10) * (4.0 * 1)

Salary = 50 + 80 + 7.7 + 35 + 0.01 * 440 + (-10) * 4

Salary ≈ 50 + 80 + 7.7 + 35 + 4.4 - 40

Salary ≈ 137.1

So, the predicted salary of a college graduate with an IQ of 110 and a GPA of 4.0 is approximately $137,100.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

True. The coefficient for the GPA/IQ interaction term (β4) being very small (0.01) suggests that the interaction effect between GPA and IQ is minimal. In statistical terms, when the coefficient for an interaction term is small, it indicates that the relationship between the predictors (GPA and IQ, in this case) does not significantly affect the response variable (salary). Therefore, there is indeed very little evidence of an interaction effect between GPA and IQ in predicting starting salary.

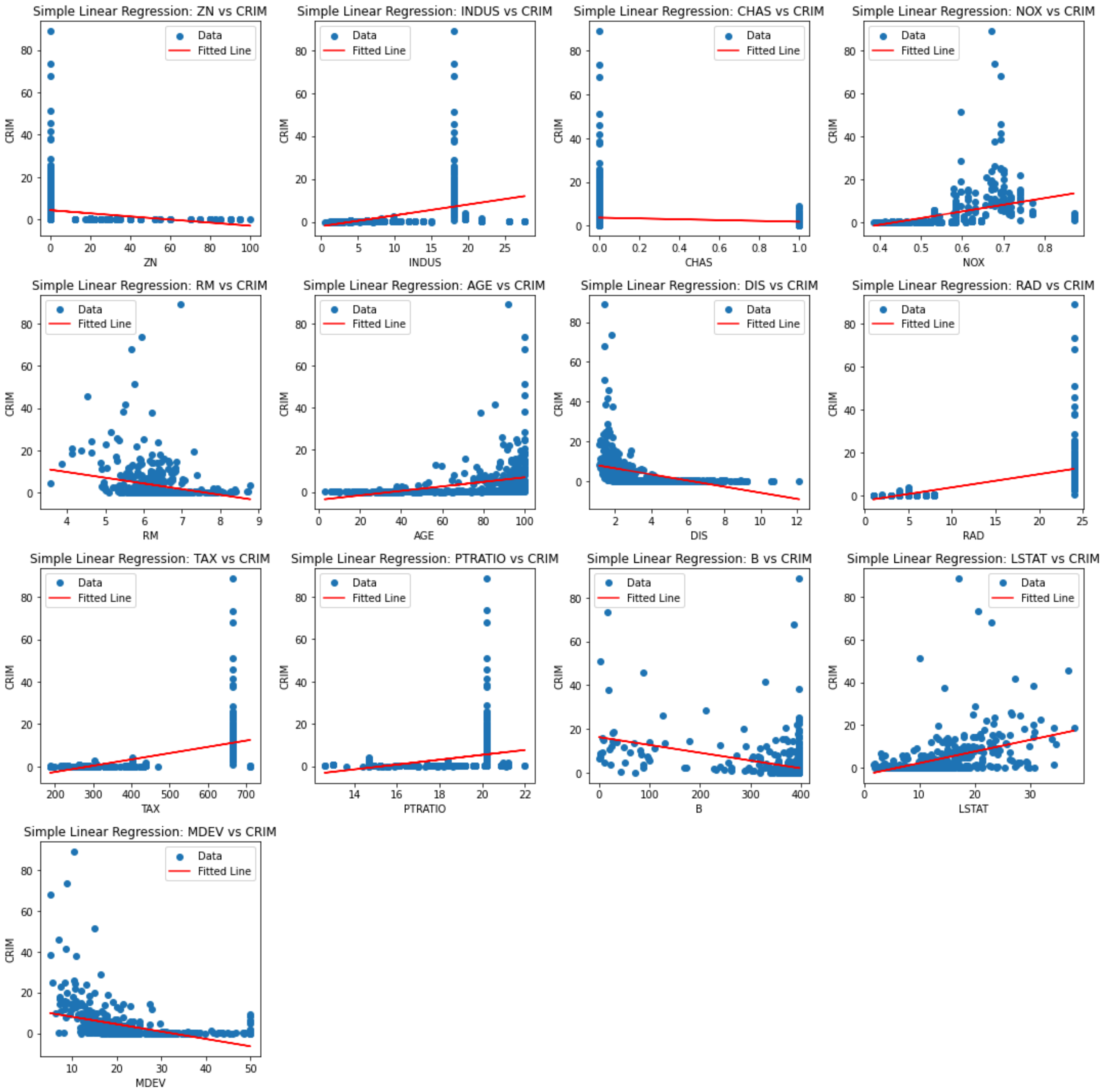
5. (ISLP: Chapter 3, Question 15) This problem involves the Boston data set (which is used as an example in the Python lab at the end of Chapter 3 in the textbook). We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. De scribe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Based on the results:

- Predictors with statistically significant associations (p-value < 0.05) with per capita crime rate (CRIM):
  - ZN, INDUS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MDEV
- Predictor with a non-significant association (p-value > 0.05) with CRIM:
  - CHAS

This suggests that most predictors have a significant impact on the per capita crime rate, except for CHAS



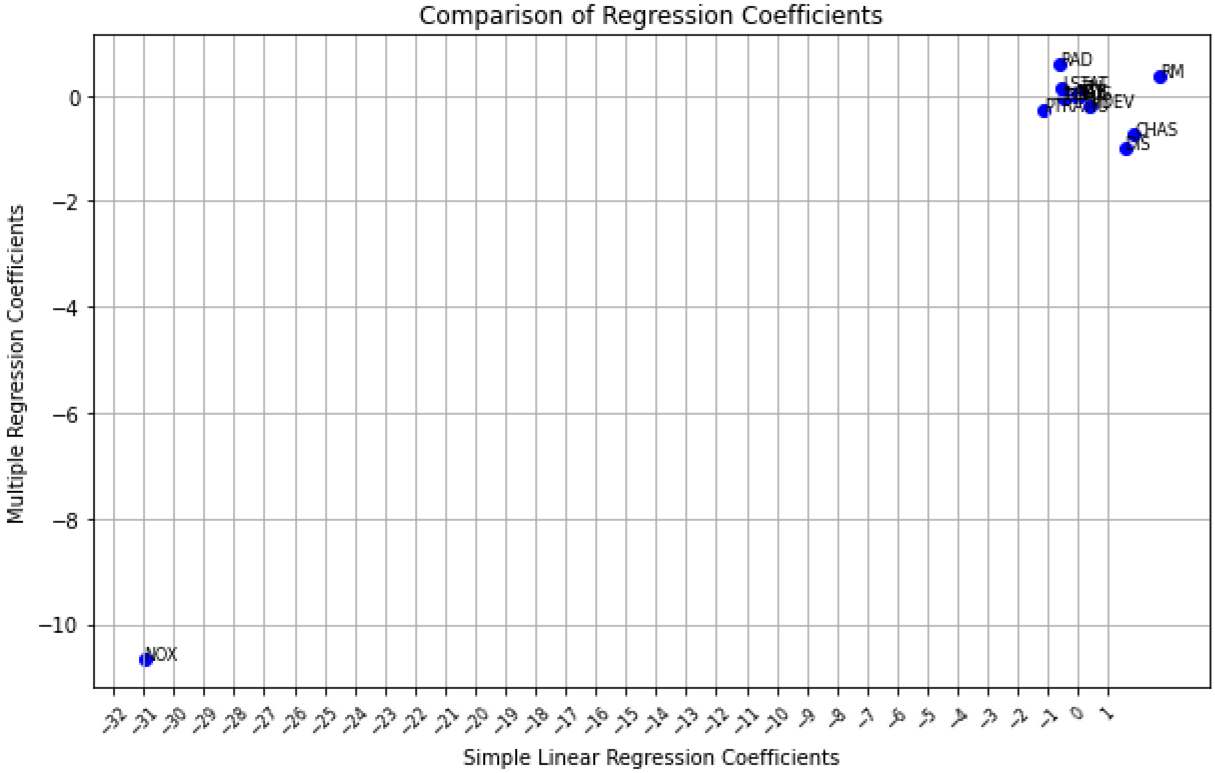
(b) Fit a multiple regression model to predict the response using all of the predictors. De scribe your results. For which predictors can we reject the null hypothesis $H0 : \beta j = 0$?

In the multiple regression model:

- The overall model is statistically significant (p < 0.05), indicating that the predictors collectively have a significant association with the per capita crime rate.

- Predictors ZN, NOX, DIS, RAD, and MDEV reject the null hypothesis ($p < 0.05$), suggesting significant associations with the per capita crime rate.
- Predictors INDUS, CHAS, RM, AGE, TAX, PTRATIO, B, and LSTAT do not reject the null hypothesis ($p > 0.05$), indicating no significant associations with the per capita crime rate at the chosen significance level.

(c) How do your results from question (5a) compare to your results from question (5b)? Create a plot displaying the univariate regression coefficients from question (5a) on the x-axis, and the multiple regression coefficients from question (5b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.



(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form $Y =$

$$\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

Predictor: ZN
R-squared: 0.058
Significant predictors: ZN
Notes: ZN^2 and ZN^3 are not significant.

Predictor: INDUS
R-squared: 0.257
Significant predictors: INDUS, INDUS^2, INDUS^3

Predictor: CHAS
R-squared: 0.003
No significant predictors.

Predictor: NOX
R-squared: 0.292
Significant predictors: NOX, NOX^2, NOX^3

Predictor: RM
R-squared: 0.068
No significant predictors.

Predictor: AGE
R-squared: 0.172
Significant predictors: AGE^2, AGE^3

Predictor: DIS
R-squared: 0.276
Significant predictors: DIS, DIS^2, DIS^3

Predictor: RAD
R-squared: 0.396
No significant predictors.

Predictor: TAX
R-squared: 0.365
No significant predictors.

```
Predictor: PTRATIO
R-squared: 0.112
Significant predictors: PTRATIO, PTRATIO^2, PTRATIO^3

Predictor: B
R-squared: 0.144
No significant predictors.

Predictor: LSTAT
R-squared: 0.214
Significant predictors: LSTAT^2

Predictor: MDEV
R-squared: 0.416
Significant predictors: MDEV, MDEV^2, MDEV^3
```

```
# 1

import numpy as np
import matplotlib.pyplot as plt

flexibility = np.linspace(0, 1, 100)
squared_bias = 1 - flexibility
variance = flexibility ** 2
training_error = 0.5 * flexibility ** 2
test_error = 0.5 * flexibility ** 2 + 0.2 * (1 - flexibility ** 2)  # Modified test error curve
bayes_error = 0.1 + 0.2 * np.sin(5 * np.pi * flexibility)  # Modified Bayes error curve

# Plotting
plt.plot(flexibility, squared_bias, label='Squared Bias')
plt.plot(flexibility, variance, label='Variance')
plt.plot(flexibility, training_error, label='Training Error')
plt.plot(flexibility, test_error, label='Test Error')
plt.plot(flexibility, bayes_error, label='Bayes Error')

plt.xlabel('Flexibility')
plt.ylabel('Error')
plt.title('Bias-Variance Decomposition')
plt.legend()


plt.show()
```

```
# 3

import pandas as pd


# ?a?
file_path = "/Users/rouren/Desktop/24S ML/HW/Boston/Boston.csv"
data = pd.read_csv(file_path)

# ?b?
num_rows, num_columns = data.shape

print("Number of rows:", num_rows)
print("Number of columns:", num_columns)

print("\nColumn names and their meanings:")
print(data.columns)

# (c)
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

important_vars = ['CRIM', 'RM', 'PTRATIO', 'LSTAT', 'MDEV']
selected_data = data[important_vars]
```

```
# Pairwise scatterplots of selected variables
sns.pairplot(selected_data)
plt.show()

# (d)
correlation = data.corr()['CRIM'].sort_values(ascending=False)
print(correlation)

# (e)
range_crime_rate = data['CRIM'].max() - data['CRIM'].min()
range_tax_rate = data['TAX'].max() - data['TAX'].min()
range_ptratio = data['PTRATIO'].max() - data['PTRATIO'].min()

print("Range of Crime Rates (CRIM):", range_crime_rate)
print("Range of Tax Rates (TAX):", range_tax_rate)
print("Range of Pupil-Teacher Ratios (PTRATIO):", range_ptratio)

# (f)
num_bound_charles_river = data['CHAS'].sum()
print("Number of census tracts that bound the Charles River:", num_bound_charles_river)

# (g)
median_ptratio = data['PTRATIO'].median()
print("Median pupil-teacher ratio among census tracts:", median_ptratio)
```

```
# (h)

# Find the Boston census tract with the lowest median value of owner-occupied homes
lowest_medv_index = data['MDEV'].idxmin()
lowest_medv_tract = data.loc[lowest_medv_index]

print("Values of other predictors for the census tract with the lowest MDEV:")
print(lowest_medv_tract.drop('MDEV'))

print("\nOverall ranges for the predictors:")
print("Range of CRIM:", data['CRIM'].min(), "-", data['CRIM'].max())
print("Range of ZN:", data['ZN'].min(), "-", data['ZN'].max())

print("\nComment:")
print("The census tract with the lowest MDEV has the following values for the predictors:")
print(lowest_medv_tract.drop('MDEV'))
print("These values can be compared to the overall ranges for those predictors to assess how they deviate

# (i)
num_more_than_seven_rooms = (data['RM'] > 7).sum()
num_more_than_eight_rooms = (data['RM'] > 8).sum()

print("Number of census tracts that average more than seven rooms per dwelling:", num_more_than_seven
print("Number of census tracts that average more than eight rooms per dwelling:", num_more_than_eight_

print("\nComment:")
print("Census tracts that average more than eight rooms per dwelling represent areas with larger, potential


# 5

# (a)


import statsmodels.api as sm

import matplotlib.pyplot as plt


plt.figure(figsize=(15, 15))


response_variable = 'CRIM'

predictor_variables = ['ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT


results = pd.DataFrame(columns=['Predictor', 'Coefficient', 'P-value'])


# Loop through each predictor variable

for i, predictor in enumerate(predictor_variables, 1):

    X = sm.add_constant(data[predictor])

    y = data[response_variable]


    model = sm.OLS(y, X).fit()


    coefficient = model.params[predictor]

    p_value = model.pvalues[predictor]


    results = results.append({'Predictor': predictor, 'Coefficient': coefficient, 'P-value': p_value}, ignore_index
    # Create a subplot

    plt.subplot(4, 4, i)

    plt.scatter(data[predictor], y, label='Data')

    plt.plot(data[predictor], model.predict(X), color='red', label='Fitted Line')

    plt.xlabel(predictor)

    plt.ylabel(response_variable)

    plt.title(f"Simple Linear Regression: {predictor} vs {response_variable}")

    plt.legend()


plt.tight_layout()


print(results)

plt.show()


print(results)


# (b)

# Add a constant term to the predictor variables

X = sm.add_constant(data[predictor_variables])


model = sm.OLS(data[response_variable], X).fit()


print(model.summary())


# (c)
```

```
import matplotlib.pyplot as plt

# Coefficients from simple linear regression models (question 5a)
simple_regression_coefficients = [0.073521, -0.506847, 1.871545, -30.975259, 2.691045, -0.107131, 1.54

# Coefficients from the multiple regression model (question 5b)
multiple_regression_coefficients = [0.0449, -0.0616, -0.7414, -10.6455, 0.3811, 0.0020, -0.9950, 0.5888, -0

predictor_variables = ['ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT

# Plotting
plt.figure(figsize=(10, 6))
plt.scatter(simple_regression_coefficients, multiple_regression_coefficients, color='blue')
plt.xlabel('Simple Linear Regression Coefficients')
plt.ylabel('Multiple Regression Coefficients')
plt.title('Comparison of Regression Coefficients')
plt.xticks(fontsize=8, rotation=45)  # Set x-axis tick fontsize to 8 and rotate the labels
plt.yticks(fontsize=10)  # Set y-axis tick fontsize to 10
plt.grid(True)

plt.xticks(range(-32, 2))  # Adjust the range as needed

for i, predictor in enumerate(predictor_variables):
    plt.annotate(predictor, (simple_regression_coefficients[i], multiple_regression_coefficients[i]), fontsize=8)

plt.show()
```

```
# (d)

import statsmodels.api as sm

predictors = ['ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MDEV

for predictor in predictors:
    # Extract predictor and response variables
    X = data[predictor]
    y = data['CRIM']

    # Add polynomial features (up to degree 3)
    X_poly = sm.add_constant(X)
    X_poly[predictor+'^2'] = X ** 2
    X_poly[predictor+'^3'] = X ** 3

    model = sm.OLS(y, X_poly).fit()

    print(f"Predictor: {predictor}")
    print(model.summary())
    print("\n")
```