Machine Learning in Engineering

Northeastern University, Fall 2024

Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.

1. (Total: 15 points)

   This exercise involves the Auto data set. Make sure that the missing values have been removed from the data.

   (a) (2 points) Which of the predictors are quantitative, and which are qualitative?

   - Quantitative predictors:
     - ['mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'year']
   - Qualitative predictors:
     - ['origin', 'name']

   'origin' values are 1, 2, 3. seems it has been coded

   (b) (2 points) What is the range (e.g., minimum and maximum) of each quantitative predictor?

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
min	9.0	3.0	68.0	46.0	1613.0	8.0	70.0
max	46.6	8.0	455.0	230.0	5140.0	24.8	82.0
range	37.6	5.0	387.0	184.0	3527.0	16.8	12.0

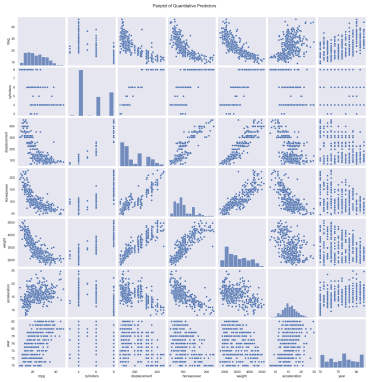
   (c) (2 points) What is the mean and standard deviation of each quantitative predictor?

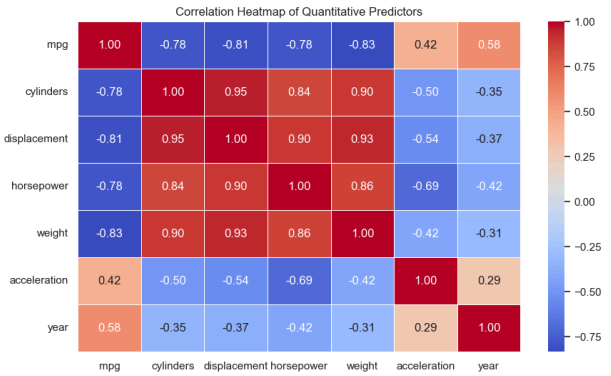
	mpg	cylinders	displacement	horsepower	weight	acceleration	year
mean	23.45	5.47	194.41	104.47	2977.58	15.54	75.98
std	7.81	1.71	104.64	38.49	849.40	2.76	3.68

(d) (3 points) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
range	35.60	5.00	387.00	184.00	3348.00	16.30	12.00
mean	24.40	5.37	187.24	100.72	2935.97	15.73	77.15
std	7.87	1.65	99.68	35.71	811.30	2.69	3.11

(e) (3 points) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Com ment on your findings.



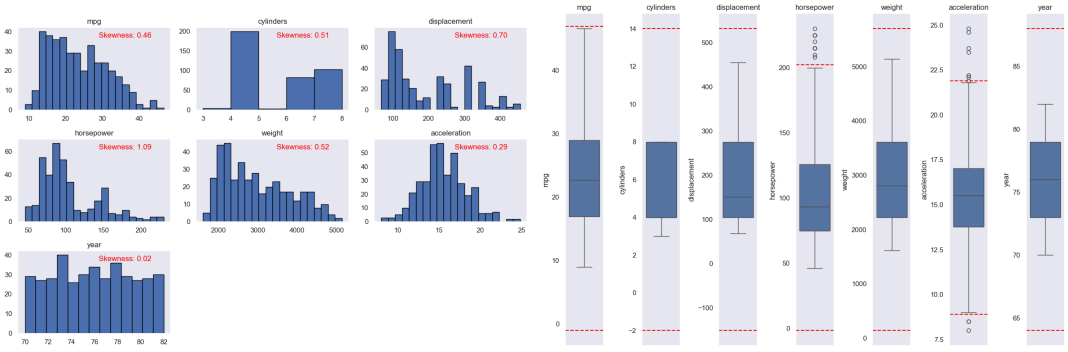


Quantitative predictors:
['mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'year']

- Largest absolute correlation coefficient value:
  - mpg and weight: strong negative correlation (r = -0.83)
- Smallest absolute correlation coefficient value:
  - acceleration and year: weak positive correlation (r = 0.29)

comparing these Quantitative predictors, and observer the correlation plots and correlation coefficients row by row:

It can be separate into 3 groups:
- First Group: mpg
- Second Group:cylinders, displacement, horsepower, weight
  - strong positive correlation cluster in-between the second group predictors (r = 0.7 ~ 1.0)
  - strong negative correlations with mpg (r = -0.7 ~ -1.0)
- Third Group: acceleration, year
  - moderate positive correlation with mpg (r = 0.4 ~ 0.7)
  - accleration has moderate correlation with other predictors, but with some exception:
    - acceleration and year: weak correlation (r = 0.29)
    - acceleration and horsepower: moderate correlation (r = -0.69)
  - year has weak correlation with other predictors, but with some exception:
    - year and mpg: moderate correlation (r = 0.58)
    - year and horsepower: moderate correlation (r = -0.42)



Distribution Insights from Histograms:
- mpg: symmetric distribution (skewness = 0.46)
- cylinders:right-skewed distribution (skewness = 0.51)
  a. few minor values at 3, 5
- displacement: right-skewed distribution (skewness = 0.70)
  a. 3 gaps in the whole dataset
- horsepower: right-skewed distribution (skewness = 1.09)

- weight: right-skewed distribution (skewness = 0.52)
- acceleration: symmetric distribution (skewness = 0.29)
- year: symmetric distribution (skewness = 0.02)

Outliers Detection from boxplots (1.5*IQR above Q3 and below Q1):

- No outliers detected in: mpg, cylinders, displacement, weight, year
- Outliers for horsepower:
  - Too High: 10 values above 202.50
- Outliers for acceleration:
  - Too Low: 3 values below 8.90
  - Too High: 8 values above 21.90

(f) (3 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- Strong Negative Correlation with Weight, Horsepower, Displacement, and Cylinders:

These variables have strong negative correlations with "mpg" (correlation coefficients between -0.7 and -1.0).

This suggests that as values in these predictors increase, "mpg" decreases, making them highly predictive for estimating gas mileage.

- Moderate Positive Correlation with Year and Acceleration:
  - "Year" has a moderate positive correlation with "mpg" (r = 0.58), indicating that newer cars tend to have better fuel efficiency.
  - "Acceleration" has a weaker but still moderate correlation, with some predictive potential, though less strong than the second group variables.

- Grouping Analysis:
  - Separating the predictors into three groups supports focusing on the second group (weight, horsepower, displacement, cylinders) for primary predictive value, given their internal positive correlation and strong negative associations with "mpg."

This analysis implies that the predictors in the second group are particularly useful for building a predictive model for "mpg," while "year" and "acceleration" could add moderate value to the model as secondary predictors.

2. (Total: 20 points) This exercise involves the Boston housing data set.

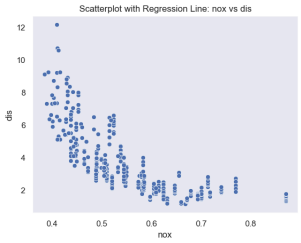
(a) (2 points) How many rows are in this data set? How many columns? What do the rows and columns represent?

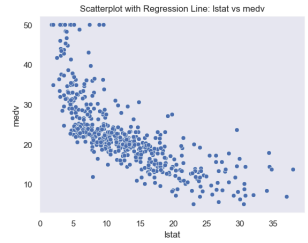
- Boston dataset has 506 rows and 15 columns.

- Rows: Each row represents an individual house or observation in the dataset
- Columns: Each column represents a specific variable or feature related to the houses.

- Neighborhood Characteristics:
  - crim: Per capita crime rate by town
  - zn: Proportion of residential land zoned for lots over 25,000 sq.ft.
  - indus: Proportion of non-retail business acres per town
  - chas: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  - nox: Nitric oxides concentration (parts per 10 million)
- Housing Attributes:
  - rm: Average number of rooms per dwelling
  - age: Proportion of owner-occupied units built prior to 1940
- Accessibility Measures:
  - dis: Weighted distances to five Boston employment centers
  - rad: Index of accessibility to radial highways
- Socioeconomic Factors:
  - tax: Full-value property tax rate per $10,000
  - ptratio: Pupil-teacher ratio by town
  - black: 1000(B - 0.63)^2 where B is the proportion of African American residents by town
  - lstat: Percentage of lower status of the population
- Target Variable:
  - medv: Median value of owner-occupied homes in $1000s

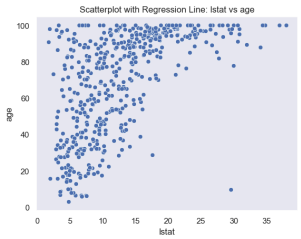
(b) (3 points) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



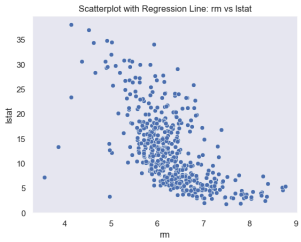


Scatterplot with Regression Line: nox vs dis: curve shape relation
Scatterplot with Regression Line: lstat vs medv: curve shape relation

Scatterplot with Regression Line: dis vs age: cone shape relation



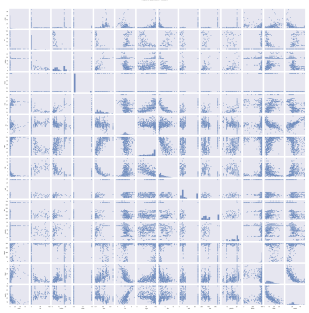
Scatterplot with Regression Line: lstat vs age: clustered in the top-left corner



Scatterplot with Regression Line: rm vs lstat: S-shape logit shape

Using pairwise scatterplots, I calculated the correlation coefficients and examined the predictors with the largest and smallest absolute values. High correlations (|r| > 0.7) were found between the following pairs:

1. nox and indus: r = 0.76
2. age and nox: r = 0.73
3. dis and indus: r = -0.71
4. dis and nox: r = -0.77
5. dis and age: r = -0.75
6. tax and indus: r = 0.72
7. tax and rad: r = 0.91
8. medv and lstat: r = -0.74



(c) (2 points) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Moderate Positive Correlations:

- rad (index of accessibility to radial highways): correlation of 0.63
- tax (full-value property tax rate): correlation of 0.58
- lstat (percentage of lower status population): correlation of 0.46

These variables show a moderate positive correlation with crim, suggesting that higher values in these predictors are associated with an increase in crime rate.

Weak Negative Correlations:

- dis (weighted distances to employment centers): correlation of -0.38
- black (proportion of Black residents): correlation of -0.39
- rm (average number of rooms per dwelling): correlation of -0.22

These predictors have weak negative correlations with crim, indicating that as distance from employment centers, the proportion of Black residents, and the average room count increase, crime rates tend to decrease, albeit weakly.

This may suggest that areas with higher housing quality and residential distance from business centers are less associated with crime.

(d) (3 points) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Suburbs with the highest crime rates:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
380	88.9762	0.0	18.1	0	0.671	6.968	91.9	1.4165	24	666	20.2	396.90	17.21	10.4
418	73.5341	0.0	18.1	0	0.679	5.957	100.0	1.8026	24	666	20.2	16.45	20.62	8.8
405	67.9208	0.0	18.1	0	0.693	5.683	100.0	1.4254	24	666	20.2	384.97	22.98	5.0
410	51.1358	0.0	18.1	0	0.597	5.757	100.0	1.4130	24	666	20.2	2.60	10.11	15.0
414	45.7461	0.0	18.1	0	0.693	4.519	100.0	1.6582	24	666	20.2	88.27	36.98	7.0

Suburbs with the highest tax rates:

	crim	tax	ptratio
492	0.11132	711	20.1
491	0.10574	711	20.1
490	0.20746	711	20.1
489	0.18337	711	20.1
488	0.15086	711	20.1

Suburbs with the highest pupil-teacher ratios:

	crim	tax	ptratio
354	0.04301	334	22.0
355	0.10659	334	22.0
135	0.55778	437	21.2
127	0.25915	437	21.2
136	0.32264	437	21.2

Range of each predictor:

	crim	tax	ptratio
min	0.00632	187.0	12.6
max	88.97620	711.0	22.0
range	88.97000	524.0	9.4

- Suburbs with Particularly High Crime Rates:

The highest crime rates in the dataset are represented by the crim values, with the maximum reaching 88.98.

Notable suburbs commonality with high crime rates include: zn of 0, indus of 18.1, chas of 0, rad of 24, tax rate of 666, ptratio of 20.2

- Suburbs with Particularly High Tax Rates:

The tax rate (tax) in the dataset ranges from a minimum of 187 to a maximum of 711. Suburbs with the highest tax rates of 711 have crime rates ranging from 0.11 to 0.21 and pupil-teacher ratios (ptratio) at 20.1.

- Suburbs with Particularly High Pupil-Teacher Ratios:

The ptratio ranges from 12.6 to 22.0, with the highest ratio indicating potentially more crowded schools.
Suburbs with the highest pupil-teacher ratio of 22.0 show lower crime rates (e.g., crim values of 0.04 to 0.11), which suggests that the pupil-teacher ratio might be influenced more by regional population density rather than crime.

- Range of Each Predictor:

Crime Rate (crim): Ranges from 0.0063 to 88.98, showing a very broad range across the dataset, with a difference of 88.97.
This suggests that crime rates vary significantly, likely reflecting both low-crime and high-crime areas.

- Tax Rate (tax):

Ranges from 187 to 711, with a range of 524. This variation could indicate differences in property value assessments or municipal services between different areas.

- Pupil-Teacher Ratio (ptratio):

Ranges from 12.6 to 22.0, with a narrower range of 9.4 compared to other predictors. This reflects less variance in educational resources across suburbs, but the higher values still indicate areas where school resources might be stretched.

Summary
Overall, several suburbs display high crime rates, tax rates, and pupil-teacher ratios, each varying widely in this dataset. The highest crime and tax rate neighborhoods often coincide, which may highlight regions with increased socioeconomic challenges or urban density.

(e) (2 points) How many of the suburbs in this data set bound the Charles river?

Number of suburbs that bound the Charles River: 35

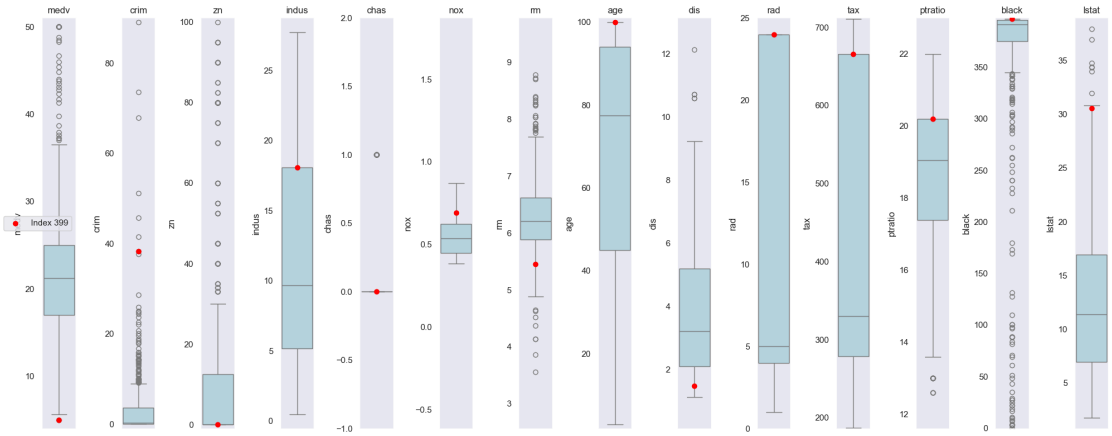
(f) (2 points) What is the median pupil-teacher ratio among the towns in this data set?

Median pupil-teacher ratio among the towns: 19.05

(g) (3 points) Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Unnamed: 0	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
399.0	38.3518	0.0	18.1	0.0	0.693	5.453	100.0	1.4896	24.0	666.0	20.2	396.9	30.59	5.0

No. 399 has the lowest median value of owner-occupied homes, and the values of the other predictors for that suburb is shown above



In summary, this suburb has relatively high crime, pollution, and industrial land usage, with smaller homes and a higher proportion of older housing. Its property tax rate is also high, and it has a higher percentage of lower-status individuals, along with a low median home value.

- Crime Rate (crim):

Suburb Value: 38.35; Overall Range: 0.01 ~ 88.98
Comment:
The crime rate for this suburb is relatively high compared to the overall range of crime rates across all suburbs. It is close to the upper end of the range, indicating this suburb experiences a significant amount of crime.

- Proportion of Residential Land Zoned for Large Lots (zn):

Suburb Value: 0.00; Overall Range: 0.00 ~ 100.00
Comment:
This suburb has no land zoned for residential areas with large lots, which is at the lowest end of the overall range. This may indicate more urbanized or densely populated areas.

- Proportion of Non-Retail Business Acres (indus):

Suburb Value: 18.10; Overall Range: 0.00 ~ 27.74
Comment:
This suburb has a relatively high proportion of land allocated to non-retail businesses, close to the maximum value. This suggests that it is likely an industrial or commercial area.

- Charles River Dummy Variable (chas):

Suburb Value: 0.00; Overall Range: 0.00 ~ 1.00
Comment:
This suburb does not bound the Charles River (chas = 0), which places it in the lower half of the range for this variable.

- Nitric Oxides Concentration (nox):

Suburb Value: 0.69; Overall Range: 0.38 ~ 0.87
Comment:
This value is near the higher end of the overall range, suggesting this suburb has relatively high levels of pollution from nitric oxides compared to other suburbs.

- Average Number of Rooms (rm):

Suburb Value: 5.45; Overall Range: 3.56 ~ 8.78

Comment:
The average number of rooms in this suburb is on the lower end of the overall range. This may indicate smaller dwellings in this area.

- Proportion of Units Built Before 1940 (age):

Suburb Value: 100.00; Overall Range: 2.90 ~ 100.00
Comment:
This suburb has a very high proportion of older homes, indicating that it is an older or more historic neighborhood. It is at the maximum end of the range.

- Weighted Distance to Employment Centers (dis):

Suburb Value: 1.48; Overall Range: 1.13 ~ 12.13
Comment:
The suburb's distance to employment centers is quite low, suggesting that it is well-connected and likely has easier access to jobs compared to other suburbs.

- Index of Accessibility to Radial Highways (rad):

Suburb Value: 24.00; Overall Range: 1.00 ~ 24.00
Comment:
This suburb has a high value for accessibility to radial highways, indicating good transport links and connectivity.

- Full-Value Property Tax Rate (tax):

Suburb Value: 666.00; Overall Range: 187.00 ~ 711.00
Comment:
The tax rate in this suburb is high, placing it near the upper end of the overall range. This could indicate higher property taxes in this area.

- Pupil-Teacher Ratio (ptratio):

Suburb Value: 20.20; Overall Range: 12.60 ~ 22.00
Comment:
The pupil-teacher ratio for this suburb is relatively high, indicating that class sizes are likely larger than in other areas, though it is within the typical range for Boston suburbs.

- Proportion of African American Residents (black):

Suburb Value: 396.90; Overall Range: 0.32 ~ 396.9
Comment:
The proportion of African American residents in this suburb is at the maximum for the dataset. This suggests a high concentration of African American residents relative to other suburbs.

- Percentage of Lower Status Population (lstat):

Suburb Value: 30.59; Overall Range: 1.73 ~ 37.97
Comment:
The proportion of lower status individuals in this suburb is relatively high compared to other suburbs, placing it near the upper end of the range.

- Median Value of Owner-Occupied Homes (medv):

Suburb Value: 5.00; Overall Range: 5.00 ~ 50.0
Comment:
The median home value in this suburb is at the very low end of the range. This is consistent with the suburb being one with lower property values.

(h) (3 points) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Number of suburbs with more than 7 rooms per dwelling: 64
Number of suburbs with more than 8 rooms per dwelling: 13

Suburbs with more than 8 rooms per dwelling are typically more affluent areas. These suburbs generally have larger homes, indicating more spacious living conditions. Most of them exhibit low crime rates, except for suburb 364, which stands out with a notably higher crime rate. Despite this, the crime rate is still not extreme.

Educational quality, as reflected in pupil-teacher ratios, is average across these suburbs, ranging from 14 to 20, indicating typical school environments. Property values are high, especially in suburbs like 163, 204, and 257, which have a median home value of 50.0, signaling wealthier areas. Tax rates vary but tend to be on the higher side, aligning with the affluence of these suburbs.

The key takeaway is that these suburbs represent desirable living areas with large homes, good educational resources, and low crime, though some have high tax rates and varying proximity to urban centers. Suburb 364's higher crime rate contrasts with the general trend, but overall, these are prosperous and spacious communities.

3. (Total: 24 points)

In this question, you should use the Carseats data set to predict the sales in a new store with Price=$120, Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.

(a) (3 points) Fit a multiple regression model to predict Sales using Price, Advertising Urban, and US. Write out the model in equation form, being careful to handle the qualitative variables properly.

Regression Equation:
Sales = 13.01 + (-0.05 * Price) + (0.12 * Advertising) + (-0.04 * Urban) + (0.06 * US)

1 for 'Yes', 0 for 'No' (Urban)
1 for 'Yes', 0 for 'No' (US)

Predicted Sales for the new store: 7.68
(Price=$120, Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.)
Advertising should be input as 10 into the model, as the Advertising column in the Carseats.csv dataset appears to be in units of $1000 USD.

(b) (3 points) Provide an interpretation of each coefficient in the model. Be careful–some of the vari ables in the model are qualitative!

- Intercept (const = 13.01):

This is the expected average Sales when all other variables are zero.
In this case, when Price, Advertising, Urban, and US are all at zero, the predicted Sales would be 13.01 units.

- Price (coef = -0.05):

For each one-unit increase in Price, Sales is expected to decrease by 0.05 units, holding all other variables constant. This negative coefficient suggests that higher prices are associated with lower sales, which aligns with typical consumer behavior.

- Advertising (coef = 0.12):

For each additional unit in Advertising, Sales is expected to increase by 0.12 units, assuming other variables are constant. This positive coefficient indicates that increased advertising spending is associated with increased sales.

- Urban (coef = -0.04):

Urban is a qualitative variable, where 1 represents "Yes" (the store is in an urban area) and 0 represents "No".
The coefficient of -0.04 means that, on average, stores in urban areas are expected to have Sales that are 0.04 units lower than stores in non-urban areas.
However, with a high p-value (0.883), this effect is not statistically significant, indicating no strong evidence that being in an urban area impacts sales.

- US (coef = 0.06):

US is also a qualitative variable, where 1 represents "Yes" and 0 represents "No".
The coefficient of 0.06 suggests that stores in the US are expected to have Sales that are 0.06 units higher than stores outside the US.
However, the high p-value (0.865) means this effect is also not statistically significant, suggesting weak evidence that being in the US has a meaningful impact on sales.

Summary
Price and Advertising are statistically significant predictors of Sales, with Price having a

negative effect and Advertising having a positive effect.
Urban and US are not statistically significant, implying that location does not have a significant effect on Sales in this model.

(c) (4 points) Using the model from (a), predict sales in the new store and calculate 68% and 95% confidence intervals.

Regression Equation: Sales = 13.00 + (-0.0546 * Price) + (0.1231 * Advertising)
Predicted Sales: 7.68
68% Confidence Interval: (7.51, 7.85)
95% Confidence Interval: (7.35, 8.01)

(d) (3 points) Using the model from (a), what is the probability that sales will be greater than 12000 units in the new store?

Probability that sales will be greater than 12000 units: 0.0000
Z-score for the threshold: 25.8020

The probability of sales being greater than 12000 units is very close to 0, which means that the predicted sales for the new store are far below 12000 units. Here's why:

- Predicted Sales: The regression model likely predicts a sales value significantly higher than 12000 units, based on the input values for Price, Advertising, Urban, and US.
- Z-score: The Z-score measures how far 12000 units is from the predicted sales in terms of standard deviations. If the Z-score is very large (i.e., the threshold of 12000 units is far below the predicted value), then the probability of sales exceeding 12000 units is very small.
- Normal Distribution: Since the probability is calculated using the cumulative distribution function (CDF), if the Z-score is large, the probability that sales will exceed 12000 units approaches zero.

It is extremely unlikely that sales will exceed 12 units. Therefore, the probability of sales being greater than 12 units is very close to zero.

(e) (3 points) Using the model from (a), what is the probability that sales will be between 6000 and 10000 units in the new store?

Probability that sales will be between 6000 and 10000 units: 1.0000
Z-scores for the lower and upper bounds: -10.0419, 13.8541

The probability of sales being between 6000 and 10000 units is almost 1, which means it's extremely likely that sales will fall within this range

- Z-scores:

The Z-score for 6000 units is -10.04, which means 6000 units is extremely far below the predicted sales of 7.68 units (more than 10 standard deviations away).
The Z-score for 10000 units is 13.85, indicating that 10000 units is very far above the predicted sales (more than 13 standard deviations away).

- 2. CDF Calculation:

The CDF for the Z-score of 6 units is almost 0 (since it's far below the predicted sales).
The CDF for the Z-score of 10 units is almost 1 (since it's far above the predicted sales).

- 3. Interpretation: The difference between the two CDF values (CDF for 10000 units minus CDF for 6000 units) is essentially 1, indicating that the probability of sales falling between 6 and 10000 units is almost 100%. This suggests that the range from 6000 to 10000 units is very much within the possible range of predicted sales, given the model.

This result implies that the model's predicted sales are so close to the middle of the 6000–10000 units range that almost all of the probability mass lies within this interval.

(f) (2 points) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.0113	0.633	20.544	0.000	11.766	14.256
Price	-0.0546	0.005	-10.710	0.000	-0.065	-0.045
Advertising	0.1203	0.025	4.845	0.000	0.072	0.169
Urban	-0.0388	0.264	-0.147	0.883	-0.558	0.481
US	0.0585	0.345	0.170	0.865	-0.620	0.737

- Interpretation of p-values:
  - Price: p-value = 0.000 (significant at the 0.05 level)
  - Advertising: p-value = 0.000 (significant at the 0.05 level)
  - Urban: p-value = 0.883 (not significant at the 0.05 level)
  - US: p-value = 0.865 (not significant at the 0.05 level)

Conclusion:
We reject the null hypothesis for Price and Advertising, because their p-values are both below 0.05. This means these predictors have a statistically significant relationship with sales.

Thus, Price and Advertising are the predictors for which can reject $H_0 : \beta_j = 0$ and conclude that they significantly influence sales in the new store.

(g) (4 points) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Using this model, predict sales in the new store and calculate 68% and 95% confidence intervals.

Regression Equation: Sales = 13.00 + (-0.0546 * Price) + (0.1231 * Advertising)
Predicted sales: 7.68
68% Confidence Interval: (7.41, 7.95)
95% Confidence Interval: (2.96, 12.41)

(h) (2 points) How well do the models in (a) and (g) fit the data?

- Both models explain about 28% of the variability in Sales (**R-squared = 0.282**), indicating limited explanatory power.
- Model (g), which includes only Price and Advertising, has a higher F-statistic (77.91 vs. 38.77 in model a), suggesting it fits the data better.
- Urban and US are non-significant predictors in model (a), so their exclusion in model (g) improves the model's fit without losing explanatory power.
- Model (g) is more efficient, as it retains only significant predictors, making it the better model for prediction.

4. (Total: 27 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

Price (in Dollars)

Age (in months)

Mileage

FuelType Fuel Type (diesel, petrol, CNG)

MetColor Metallic color (1=yes, 0=no)

Automatic Automatic transmission (1=yes, 0=no)

Displacement Engine displacement (in cu. inches)

Doors Number of doors

Weight (in pounds)

Horsepower Engine horsepower

(a) (2 points) Which of the predictors are quantitative, and which are qualitative?

Quantitative Predictors:
['Price', 'Age', 'Mileage', 'Horsepower', 'Displacement', 'Doors', 'Weight']

Qualitative Predictors:
['FuelType', 'MetColor', 'Automatic']

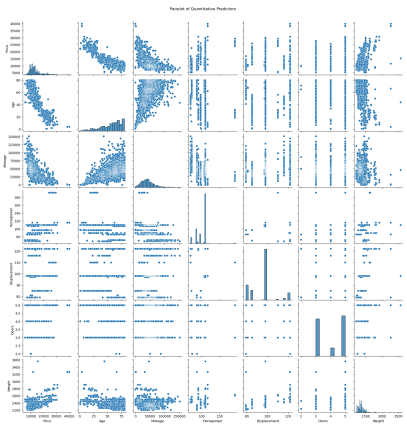
(b) (2 points) What is the range (i.e., min and max) of each quantitative predictor?

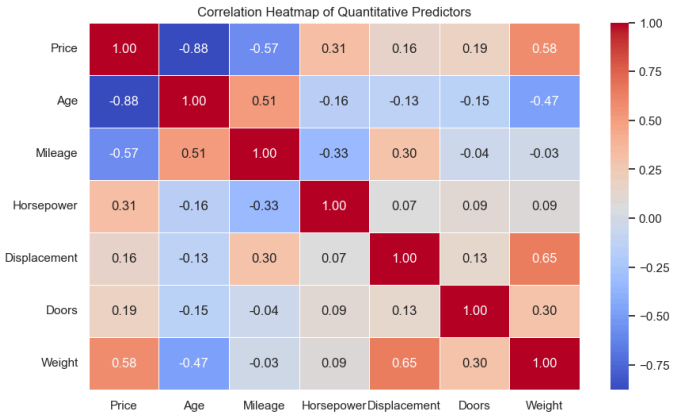
	Price	Age	Mileage	Horsepower	Displacement	Doors	Weight
Min	5351	1	1	69	79	2	2205
Max	39975	80	150993	192	122	5	3560
Range	34624	79	150992	123	43	3	1355

(c) (2 points) What is the mean and standard deviation of each quantitative predictor?

	Price	Age	Mileage	Horsepower	Displacement	Doors	Weight
Mean	13199.27	55.95	42584.59	101.50	95.72	4.03	2364.44
Standard Deviation	4461.16	18.60	23305.40	14.98	11.59	0.95	115.95

(d) (4 points) Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.





- Correlations:
  - Price and Age show a strong negative correlation (-0.88), meaning older cars tend to be priced lower.
  - Mileage and Price also have a moderate negative correlation (-0.57), indicating higher mileage usually results in a lower price.
  - Weight and Price are positively correlated (0.58), suggesting that heavier cars are generally more expensive.

- Key Insights:
  - Older cars and those with higher mileage tend to be priced lower.
  - Heavier cars generally have higher prices.
  - Most cars in the dataset have 4 doors and an average age of about 4.6 years.

(e) (4 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- Price (-0.57): A moderate negative correlation suggests that as the price of the car increases, mileage (mpg) tends to decrease. This could indicate that higher-priced cars may have larger engines or more luxury features, leading to lower fuel efficiency.
- Age (0.51): A moderate positive correlation means that older cars might have better fuel efficiency (higher mpg), though the relationship might not be as strong.
- Horsepower (-0.33): A moderate negative correlation suggests that cars with higher horsepower tend to have lower fuel efficiency, which aligns with the idea that more powerful engines often consume more fuel.
- Displacement (0.30): A weak positive correlation means that as engine displacement increases, mpg tends to improve slightly. However, the relationship is not very strong.
- Doors (-0.04): A very weak negative correlation indicates that the number of doors does not significantly affect mpg.
- Weight (-0.03): A very weak negative correlation suggests that vehicle weight doesn't have much of an effect on fuel efficiency, though heavier cars are typically less fuel-efficient.

From these correlations, the most promising predictors for gas mileage (mpg) are Price, Age, and Horsepower, with Price and Horsepower showing moderate negative correlations, suggesting that higher prices and more horsepower are likely associated with lower mpg.

The Age variable shows a moderate positive correlation, meaning older cars might be slightly more fuel-efficient. However, the correlations are not exceptionally strong, indicating that additional features or non-linear relationships might need to be explored further.

(f) (4 points) Fit a simple linear regression with Price as the response and Age as the predictor.

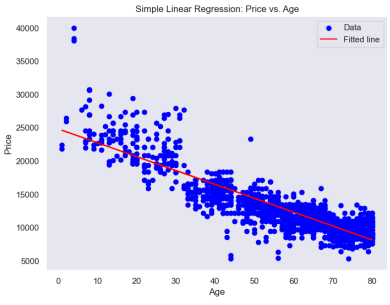
(i) Is there a relationship between the predictor and the response?

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.768
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	4758.
Date:	Sat, 09 Nov 2024	Prob (F-statistic):	0.00
Time:	16:41:14	Log-Likelihood:	-13054.
No. Observations:	1436	AIC:	2.611e+04
Df Residuals:	1434	BIC:	2.612e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.496e+04	179.699	138.910	0.000	2.46e+04	2.53e+04
Age	-210.2481	3.048	-68.978	0.000	-216.227	-204.269

Omnibus:	359.248	Durbin-Watson:	1.214
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2773.781
Skew:	0.946	Prob(JB):	0.00
Kurtosis:	9.540	Cond. No.	187.



Yes, there is a significant relationship between the predictor (Age) and the response (Price).

The p-value for Age is very small (0.000), which indicates that Age is a statistically significant predictor of Price.

(ii) How strong is the relationship between the predictor and the response?

The relationship between Age and Price is moderately strong.

The R-squared value is 0.768, meaning that approximately 76.8% of the variability in Price is explained by Age. This suggests a fairly strong linear relationship.

(iii) What is the predicted price associated for a car with an age of 48 months? What are the associated 95% confidence intervals?

Formula for the predicted price: Price = 24962.03 + (-210.25 * Age)

Predicted price for a car with Age = 48 months: 14870.12

95% confidence interval: (14583.36, 15156.88)

(g) (5 points) Fit a multiple linear regression with Price as the response and all other variables the predictors.

(i) Is there a relationship between the predictors and the response?

OLS Regression Results

Dep. Variable:	log_Price	R-squared:	0.848
Model:	OLS	Adj. R-squared:	0.847
Method:	Least Squares	F-statistic:	634.8
Date:	Sat, 09 Nov 2024	Prob (F-statistic):	0.00
Time:	18:17:36	Log-Likelihood:	848.67
No. Observations:	1148	AIC:	-1675.
Df Residuals:	1137	BIC:	-1620.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.4809	0.140	60.450	0.000	8.206	8.756
Age	-0.0103	0.000	-40.201	0.000	-0.011	-0.010
Mileage	-2.634e-06	2.1e-07	-12.550	0.000	-3.05e-06	-2.22e-06
Horsepower	0.0021	0.001	3.587	0.000	0.001	0.003
MetColor	0.0062	0.007	0.847	0.397	-0.008	0.021
Automatic	0.0306	0.016	1.970	0.049	0.000	0.061
Displacement	-0.0006	0.001	-0.674	0.500	-0.002	0.001
Doors	0.0054	0.004	1.362	0.173	-0.002	0.013
Weight	0.0006	6.05e-05	9.666	0.000	0.000	0.001
FuelType_Diesel	0.0561	0.049	1.141	0.254	-0.040	0.153
FuelType_Petrol	0.0871	0.030	2.932	0.003	0.029	0.145

Omnibus:	237.457	Durbin-Watson:	2.040
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1209.231
Skew:	-0.861	Prob(JB):	2.62e-263
Kurtosis:	7.724	Cond. No.	2.02e+06

Yes, there is a relationship between the predictors and the response.

The model has a significant F-statistic of 634.8, which suggests that the predictors collectively explain a significant portion of the variance in the response variable.

The p-values for several variables are below the significance level (0.05), indicating that they contribute to the prediction of the price.

(ii) How strong is the relationship between the predictors and the response?

The R-squared value of 0.848 indicates that the model explains approximately 84.8% of the variation in the log-transformed price.

This is a strong relationship, meaning that the predictors do a good job of explaining the price variability.

(iii) Which predictors appear to have a statistically significant relationship to the response?

The statistically significant predictors are:

- Age (p-value: 0.000)
- Mileage (p-value: 0.000)
- Horsepower (p-value: 0.000)
- Automatic (p-value: 0.049)
- FuelType_Petrol (p-value: 0.003)

These predictors have p-values less than 0.05, meaning they have a statistically significant relationship with the log-transformed price.

(iv) What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

- The coefficient for Age is -0.0103, which suggests that for each additional year of age, the log-transformed price decreases by approximately 1.03%.
- The t-statistic for age is -40.201, and the p-value is 0.000, indicating that the effect of age on price is statistically significant and very precise.

(v) What is the predicted price associated for a car with a mileage of 45000 miles, 48 months, diesel, automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepower of 90, and non-metallic color? What are the associated 95% confidence intervals?

Predicted Price: $14707.44
95% Confidence Interval: $14119.07 - $15320.33

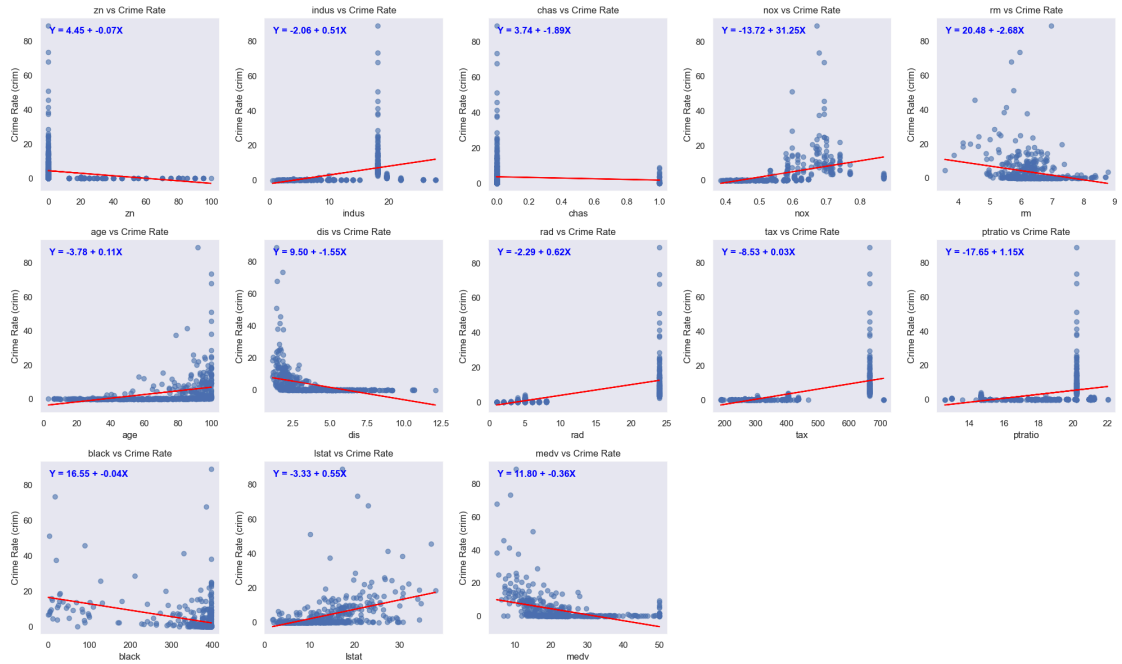
(h) (4 points) Which predictors predictors matter most for predicting the price for a car? (Find the first and the second most important variables)

- To determine which predictors matter the most, look at the absolute values of the coefficients in the linear regression model:
  - Age has a coefficient of -0.0103 (strong negative impact).
  - Mileage has a coefficient of -2.634e-06 (very small negative impact).
  - Horsepower has a coefficient of 0.0021 (positive impact).
  - Weight has a coefficient of 0.0006 (positive impact).
- The most important predictors are likely:
  - Mileage (strong negative relationship with the price)
  - Age (strong negative relationship)

5. (Total: 14 points)

This problem involves the Boston data set. We want 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) (3 points) For each predictor, fit a simple linear regression model to predict the response. Describe 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.



The predictors 'zn', 'indus', 'nox', 'rm', 'age', 'dis', 'rad', 'tax', 'ptratio', 'black', 'lstat', and 'medv' all show statistically significant associations with the response variable, with p-values less than 0.05, and they explain a significant portion of the variation in the crime rate, as indicated by their R-squared values.

The predictor 'chas', however, does not show a statistically significant association with the crime rate, with a p-value of 0.2094 and a low R-squared value, indicating it explains little of the variation in the response.

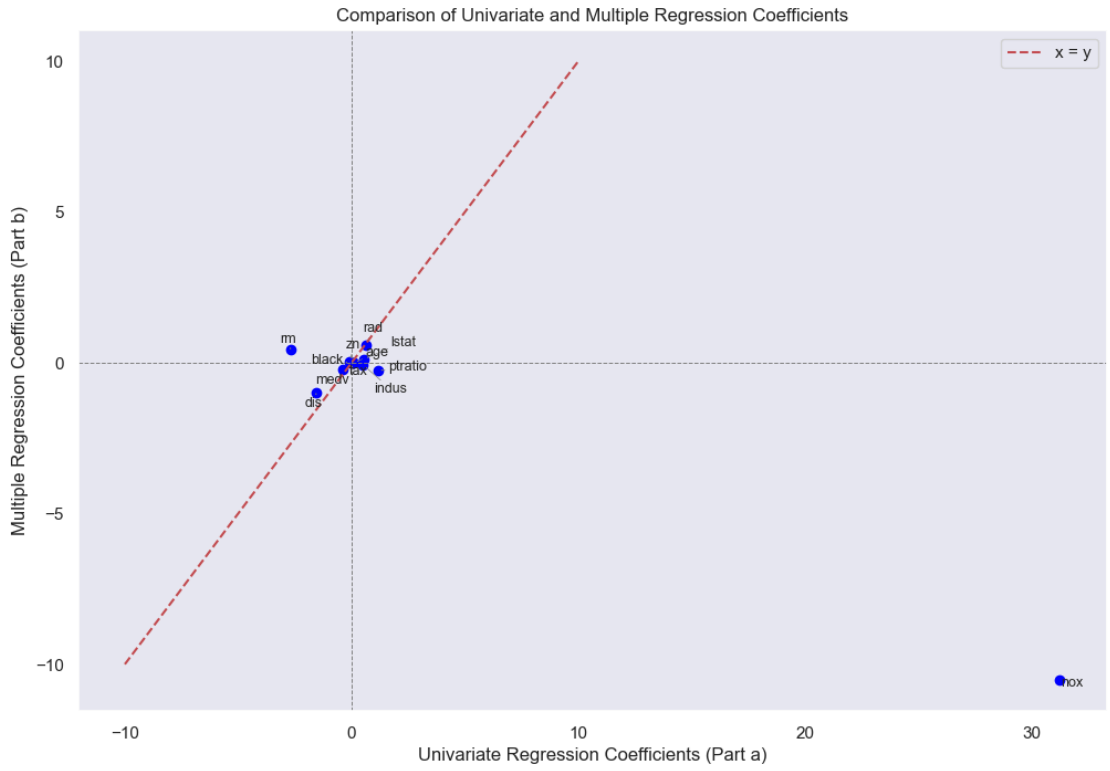
(b) (3 points) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

- The p-values for each predictor help determine whether we can reject the null hypothesis H0: βj = 0 (no effect). If the p-value is less than 0.05, we can reject the null hypothesis, suggesting a statistically significant relationship between the predictor and the response variable.
- We can reject the null hypothesis for the following predictors, as their p-values are less than 0.05:
    - zn (p = 0.015)
    - dis (p = 0.000)
    - rad (p = 0.000)
    - black (p = 0.042)
    - medv (p = 0.001)

- The following predictors have p-values greater than 0.05, meaning we cannot reject the null hypothesis for them, suggesting they do not have a statistically significant association with crime rates:
    - Unnamed: 0 (p = 0.555)
    - indus (p = 0.442)
    - chas (p = 0.536)
    - nox (p = 0.054)
    - rm (p = 0.458)
    - age (p = 0.987)
    - tax (p = 0.526)
    - ptratio (p = 0.149)
    - lstat (p = 0.102)

- Thus, the predictors zn, dis, rad, black, and medv have a statistically significant relationship with the crime rate.

(c) (4 points) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) 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.



Comparison of Univariate and Multiple Regression Coefficients:

	Predictor	Univariate Coefficient	Multiple Coefficient
0	zn	−0.0739	0.0450
1	indus	0.5098	−0.0692
2	nox	31.2485	−10.5274
3	rm	−2.6841	0.4400
4	age	0.1078	0.0009
5	dis	−1.5509	−0.9933
6	rad	0.6179	0.5843
7	tax	0.0297	−0.0035
8	ptratio	1.1520	−0.2660
9	black	−0.0363	−0.0076
10	lstat	0.5488	0.1260
11	medv	−0.3632	−0.2045

- Direction Changes: Some predictors have coefficients that change direction between univariate and multiple regression:
  - zn: Changes from -0.0739 (negative) to 0.0450 (positive).
  - indus: Changes from 0.5098 (positive) to -0.0692 (negative).
  - nox: Decreases substantially from 31.2485 (positive) to -10.5274 (negative).
  - rm: Changes from -2.6841 (negative) to 0.4400 (positive).
  - tax: Changes from 0.0297 (positive) to -0.0035 (negative).
  - ptratio: Changes from 1.1520 (positive) to -0.2660 (negative).

- Magnitude Differences: Several predictors have coefficients that differ significantly in magnitude:
  - nox has a high coefficient (31.2485) in the univariate model, which decreases to -10.5274 in the multiple regression, likely due to multicollinearity effects.

- lstat and medv show smaller reductions in magnitude between the univariate and multiple regression models.

- Stable Predictors: Some predictors, such as dis and rad, retain similar coefficient values across both models, indicating more stable relationships with crim regardless of other variables.

- Plot Insights: The plot should provide a visual way to see which coefficients differ significantly between univariate and multiple models. Points close to the line y=x represent predictors with similar coefficients in both models, while those further away indicate predictors whose effect size changes more when adjusting for other variables.

(d) (4 points) 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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 +$$

- Significant p-values for x^2 ($p < 0.05$):
  - indus: p-value for x^2 is 3.42e-10, which is highly significant, indicating a potential non-linear relationship.
  - nox: p-value for x^2 is 6.81e-15, which is also highly significant.
  - dis: p-value for x^2 is 4.94e-12, which is highly significant.
  - ptratio: p-value for x^2 is 4.12e-03, which is significant.

- Significant p-values for x^3 ($p < 0.05$):
  - indus: p-value for x^3 is 1.19e-12, which suggests that a cubic relationship may also be significant for this predictor.
  - nox: p-value for x^3 is 6.96e-16, which also suggests significance for a cubic term.
  - dis: p-value for x^3 is 1.09e-08, indicating a significant cubic relationship.
  - medv: p-value for x^3 is 1.05e-12, indicating a significant cubic relationship.

Conclusion:

- Yes, there is evidence of non-linear associations between several predictors and the response variable (crime rate). Specifically:
  - indus (industrial area), nox (nitrogen oxide concentration), and dis (weighted distance to employment centers) show significant evidence of non-linear relationships with the response, as indicated by the highly significant p-values for both the quadratic and cubic terms.
  - ptratio (pupil-teacher ratio) and medv (median value of owner-occupied homes) also show some non-linear trends, although the significance may be less strong than for the others.

- In summary, predictors like indus, nox, and dis exhibit strong evidence of non-linear relationships with the response variable. Other predictors, like ptratio and medv, also show some evidence but with less pronounced significance.