

Godfrey Mawulizo (100901), Shema Hugor (100763), Shyaka Kevin
(100915),
Niyonsenga Jean Paul (100888), Nyirmanzi Jean Claude (100882)

Lecturer: Dr. Pacifique Nizeyimana

June 22, 2025



(Data-mining-group1)

- 1 Exercise 2.4, Question 1
- 2 Exercise 2.4, Question 10
- 3 Exercise 3.7, Question 1
- 4 Exercise 3.7, Question 6
- 5 Exercise 3.7, Question 11

For each of parts (a) through (d), indicate whether we would generally expect the performance of a **flexible** statistical learning method to be **better** or **worse** than an **inflexible** method. Justify your answer.

Condition: The sample size n is extremely large, and the number of predictors p is small.

Answer: A flexible method would perform **better**.

Justification: With a large sample size and a small number of predictors, flexible methods can effectively learn complex relationships without overfitting. The low variance due to a large n allows the model to take advantage of its flexibility.

Condition: The number of predictors p is extremely large, and the number of observations n is small.

Answer: A flexible method would perform **worse**.

Justification: In high-dimensional settings where $p \gg n$, flexible models tend to overfit, capturing noise instead of the true signal. Inflexible methods are more regularized and better suited to such situations.

Condition: The relationship between the predictors and the response is highly non-linear.

Answer: A flexible method would perform **better**.

Justification: Flexible models can adapt to non-linear structures in the data, capturing complex patterns. Inflexible models, such as linear regression, would have high bias and fail to represent the non-linear relationship accurately.

Condition: The variance of the error terms, i.e., $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high.

Answer: A flexible method would perform **worse**.

Justification: When the noise variance is high, flexible methods may overfit to the noise, resulting in high prediction variance. Inflexible methods tend to smooth over the noise, leading to more stable predictions.

Comprehensive analysis of the Boston housing dataset.

The Boston dataset was loaded successfully.

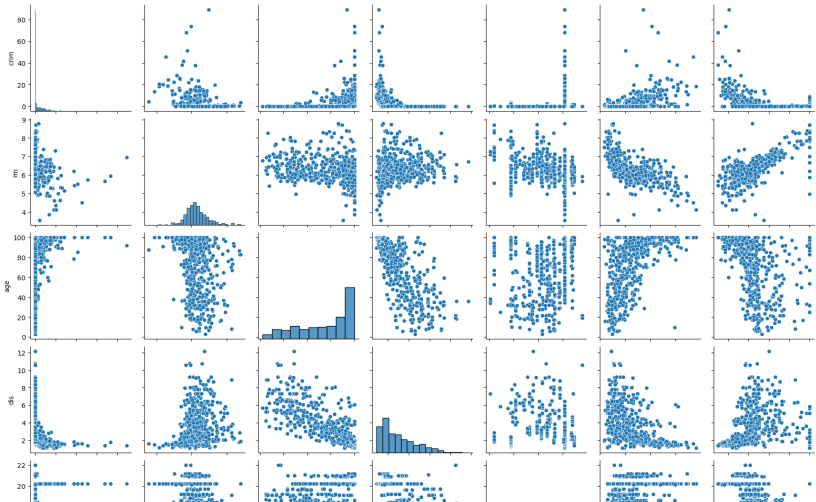
- Number of observations (rows): 506
- Number of predictors (columns): 13
- Each row represents a suburb in the Boston metropolitan area.
- Each column represents a different predictor variable (e.g., crime rate, tax rate).

Data Types:

float64 (10), int64 (3)

A subset of pairwise scatterplots was generated for selected predictors to visualize their relationships:

Pairwise Scatterplots of Selected Boston Predictors



- **medv vs. lstat**: Strong negative non-linear relationship.
- **medv vs. rm**: Positive linear trend; more rooms \Rightarrow higher home values.
- **dis vs. nox**: Negative correlation; farther suburbs have lower pollution.
- **crim**: Highly skewed, with few high-value outliers. Positively correlated with rad, tax, lstat; negatively with medv, dis.

- **Strong Positive Correlations:** rad (0.63), tax (0.58), lstat (0.46), nox (0.42), indus (0.41)
- **Strong Negative Correlations:** medv (-0.39), dis (-0.38), zn (-0.20)

Top 5 Suburbs with Highest Crime Rate:

lightblue Index	crim	tax	ptratio	medv
380	88.9762	666	20.2	10.4
418	73.5341	666	20.2	8.8
405	67.9208	666	20.2	5.0
410	51.1358	666	20.2	15.0
414	45.7461	666	20.2	7.0

Top 5 Suburbs with Highest Tax Rate:

lightblue Index	tax	medv
492	711	20.1
491	711	13.6
490	711	8.1
489	711	7.0
488	711	15.2

Top 5 Suburbs with Highest Pupil-Teacher Ratio:

lightblue Index	ptratio	medv
354	22.0	18.2
355	22.0	20.6
135	21.2	18.1
127	21.2	16.2
136	21.2	17.4

Ranges for Predictors:

crim: 0.006 - 88.976 tax: 187 - 711 ptratio: 12.6 - 22.0 medv:
5.0 - 50.0 (truncated at upper bound)

Number of suburbs bordering the Charles River ($\text{chas} = 1$): **35**

Median pupil-teacher ratio: **19.05**

- Index: 398
- Median value (medv): \$5,000
- High crim, lstat, tax, age, and rad
- Low rm, dis, zn, and chas
- Reflects socio-economic distress and possibly environmental pollution.

- $rm > 7$: 64 suburbs
- $rm > 8$: 13 suburbs
- These suburbs have:
 - Very high medv (often capped at 50.0)
 - Very low crim, lstat, nox
 - Indicate affluent, clean, and safe neighborhoods

Example (Index 267):

`rm = 8.297, medv = 50.0, crim = 0.57834, lstat = 7.44, nox = 0.`

The p-values in Table 3.4 test the null hypotheses that each advertising medium—TV, radio, and newspaper—has no effect on sales:

- H_0^{TV} : TV advertising budget has no impact on sales.
- H_0^{radio} : Radio advertising budget has no impact on sales.
- $H_0^{\text{newspaper}}$: Newspaper advertising budget has no impact on sales.

lightblue Predictor	Coefficient	Std. Error	t-Statistic	p-value	Regression
Intercept	2.939	0.3119	9.42	< 0.0001	
TV	0.046	0.0014	32.81	< 0.0001	
Radio	0.189	0.0086	21.89	< 0.0001	
Newspaper	-0.001	0.0059	-0.18	0.8599	

results for predicting Sales using TV, Radio, and Newspaper advertising budgets.

Conclusion in terms of Sales and Media Spend:

- **TV Advertising:** There is strong evidence that increasing TV advertising spend is associated with higher sales. The relationship is statistically significant with a very low p-value (< 0.0001).
- **Radio Advertising:** Similarly, radio advertising expenditure shows a significant positive association with sales. Increasing the radio budget is likely to increase sales, as indicated by the small p-value (< 0.0001).
- **Newspaper Advertising:** There is no statistical evidence that spending on newspaper advertising affects sales. The high p-value (0.8599) suggests that changes in newspaper ad budgets are not meaningfully associated with changes in sales, after accounting for TV and radio effects.

We aim to show that the least squares regression line from simple linear regression always passes through the point (\bar{x}, \bar{y}) .

The least squares regression line is defined as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof that the Line Passes Through (\bar{x}, \bar{y}) : Substitute $x = \bar{x}$ into the regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Replace $\hat{\beta}_0$ with $\bar{y} - \hat{\beta}_1 \bar{x}$:

$$\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

Thus, when $x = \bar{x}$, the predicted value $\hat{y} = \bar{y}$. Therefore, the least squares regression line always passes through the point (\bar{x}, \bar{y}) .

Analysis of regressions without intercept for y onto x and x onto y .

- Coefficient estimate: $\hat{\beta}_{yx} = 1.9762$
- Standard error: ≈ 0.117
- **t-statistic:** ≈ 16.898
- p-value: < 0.0001
- R^2 (uncentered): 0.743

- Coefficient estimate: $\hat{\beta}_{xy} = 0.3757$
- Standard error: ≈ 0.022
- **t-statistic:** ≈ 16.898
- p-value: < 0.0001
- R^2 (uncentered): 0.743

- The t-statistics are **identical** in both regressions: $t \approx 16.898$.
- The slope coefficients are not reciprocals, due to differences in variance between x and y .
- This symmetry in t-statistics is expected in regressions **without intercept**.

In a simple linear regression without intercept, the slope estimate is:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

The standard error of $\hat{\beta}$ is:

$$SE(\hat{\beta}) = \sqrt{\frac{\sum (y_i - x_i \hat{\beta})^2}{(n-1) \sum x_i^2}}$$

The t-statistic for testing $H_0 : \beta = 0$ is:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

An alternative algebraic form of the t-statistic is:

$$t = \frac{\sqrt{n-1} \sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n x_i y_i)^2}}$$

This expression shows the symmetry between x and y .

From the expression above, it is clear that the t-statistic depends only on the sums:

$$\sum x_i y_i, \quad \sum x_i^2, \quad \sum y_i^2$$

This symmetry implies that the t-statistic for the regression of y on x is the same as for the regression of x on y , when both are performed **without an intercept**. Our computations confirm this: both regressions yielded $t \approx 16.898$.

When an intercept is included, the symmetry breaks. The regression line no longer passes through the origin, and the standard error and estimated variance change. As a result:

- The t-statistics for y on x and x on y **differ**.
- The R^2 and coefficient values also generally change.

lightblue Case	Slope ($\hat{\beta}$)	t-statistic	Same t?
$y \sim x$ (no intercept)	1.9762	16.898	Yes
$x \sim y$ (no intercept)	0.3757	16.898	Yes
$y \sim x$ (with intercept)	≈ 2.0	$\neq 16.898$	No
$x \sim y$ (with intercept)	≈ 0.5	$\neq 16.898$	No