Math6450_Assignment1

September 3, 2025

Part 1: Data Exploration and Preparation

Task 1.1: Data Loading and Initial Exploration

Load the Boston Housing dataset and perform initial exploratory data analysis.

1.1 What are the dimensions of the dataset (number of observations and variables)?

```
============================================================
BOSTON HOUSING DATASET ANALYSIS
============================================================

1.1 DATASET DIMENSIONS
------------------------------
Number of observations (rows): 506
Number of variables (columns): 14
Dataset shape: (506, 14)

Column names: ['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad',
'tax', 'ptratio', 'b', 'lstat', 'medv']
```

1.2 Provide descriptive statistics for the target variable (medv) and primary feature (lstat).

```

1.2 DESCRIPTIVE STATISTICS
------------------------------

Descriptive statistics for TARGET VARIABLE (medv):
count   506.000
mean     22.533
std       9.197
min       5.000
25%      17.025
50%      21.200
75%      25.000
max      50.000
Name: medv, dtype: float64

Descriptive statistics for PRIMARY FEATURE (lstat):
count   506.000
mean     12.653
```

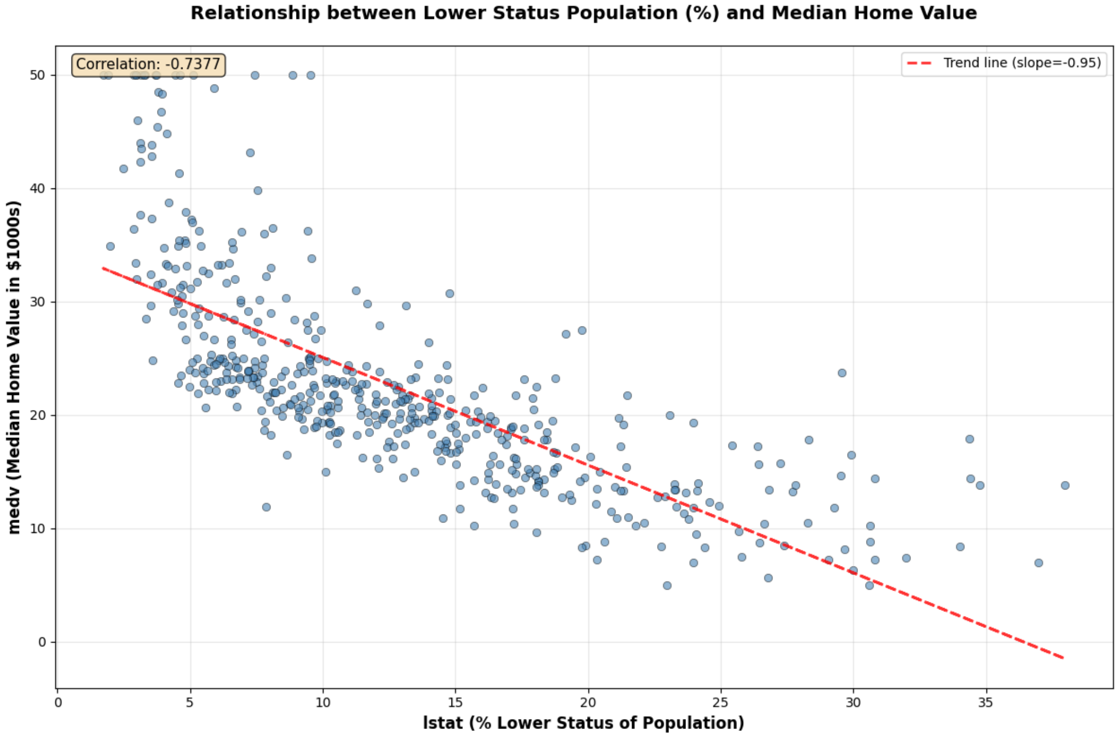
```
std       7.141
min       1.730
25%       6.950
50%      11.360
75%      16.955
max      37.970
Name: lstat, dtype: float64

Additional statistics for medv:
Variance: 84.5867
Standard deviation: 9.1971
Skewness: 1.1081
Kurtosis: 1.4952

Additional statistics for lstat:
Variance: 50.9948
Standard deviation: 7.1411
Skewness: 0.9065
Kurtosis: 0.4932
```

1.3 What is the correlation coefficient between medv and lstat? Interpret this value.

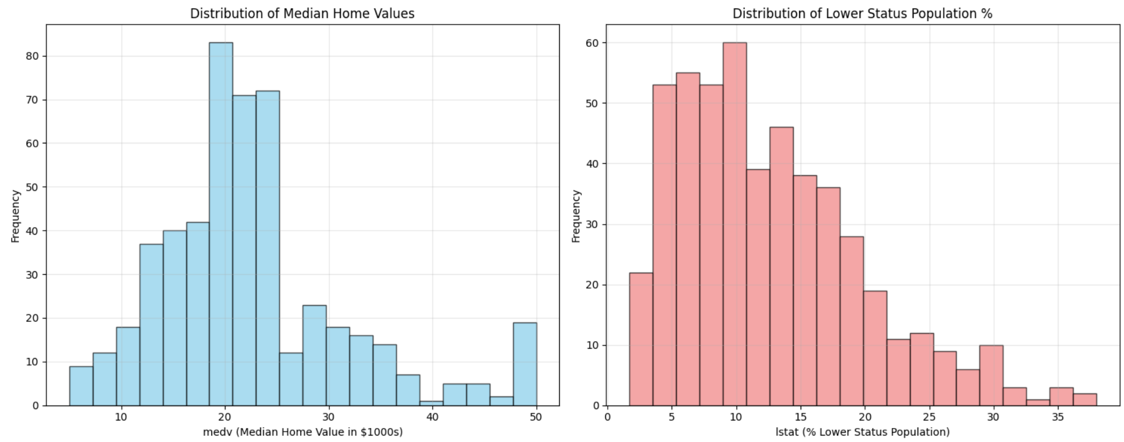
```
1.3 CORRELATION ANALYSIS
------------------------------
Correlation coefficient between medv and lstat: -0.7377

INTERPRETA1TION:
- The correlation coefficient of -0.7377 indicates a strong negative
relationship
- This means that as lstat (% lower status population) increases, medv (median
home value) tends to decrease
- The relationship explains approximately 54.4% of the variance (R^2 = 0.5441)
- Statistical significance: p-value = 5.08e-88
- The correlation is statistically significant at  = 0.05
```

1.4 Create a scatter plot showing the relationship between lstat (x-axis) and medv (y-axis). Describe the pattern you observe.

```
1.4 SCATTER PLOT ANALYSIS
------------------------------
```



PATTERN OBSERVED IN SCATTER PLOT:
- The scatter plot reveals a clear negative relationship between lstat and medv
- As the percentage of lower status population increases, median home values tend to decrease
- The relationship appears to be non-linear, showing a curved pattern rather than a straight line
- There's more variability in home values at lower lstat percentages
- The relationship seems stronger (steeper decline) at lower lstat values and levels off at higher lstat values
- There are some potential outliers, particularly homes with high values despite higher lstat percentages
- The data points form a characteristic negative exponential or power-law pattern




```
SUMMARY:
- Dataset contains 506 observations and 14 variables
- Strong negative correlation (-0.7377) between lstat and medv
- Non-linear relationship visible in scatter plot
- Both variables show reasonable distributions for regression analysis
```

Part 2: Linear Regression Model Fitting

Task 2.1: Model Estimation

Fit a simple linear regression model using lstat to predict medv using statsmodels.

2.1 Write the estimated regression equation in the form:

medv $= \hat{\beta}_0 + \hat{\beta}_1 \times$ lstat

$$\text{medv} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{lstat}$$

```
======================================================================
BOSTON HOUSING LINEAR REGRESSION ANALYSIS - PART 2
======================================================================

LINEAR REGRESSION RESULTS:
========================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.544
Model:                            OLS   Adj. R-squared:                  0.543
Method:                 Least Squares   F-statistic:                     601.6
Date:                Fri, 29 Aug 2025   Prob (F-statistic):           5.08e-88
Time:                        07:54:45   Log-Likelihood:                -1641.5
No. Observations:                 506   AIC:                             3287.
Df Residuals:                     504   BIC:                             3295.
Df Model:                           1
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     34.5538      0.563     61.415      0.000      33.448      35.659
lstat         -0.9500      0.039    -24.528      0.000      -1.026      -0.874
==============================================================================
Omnibus:                      137.043   Durbin-Watson:                   0.892
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              291.373
Skew:                           1.453   Prob(JB):                     5.36e-64
Kurtosis:                       5.319   Cond. No.                         29.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.

COEFFICIENTS:
Intercept ( ): 34.5538
Slope ( ): -0.9500

2.1 ESTIMATED REGRESSION EQUATION
----------------------------------------
medv = 34.5538 + (-0.9500) × lstat
medv = 34.5538 - 0.9500 × lstat

Alternative notation:
ŷ = 34.5538 + (-0.9500)x
where ŷ = predicted median home value and x = lstat
```

2.2 What is the interpretation of the intercept in the context of this problem? Does it have practical meaning?

$\hat{\beta}_0$

```
2.2 INTERPRETATION OF INTERCEPT ( )
----------------------------------------
Intercept value: 34.5538

INTERPRETATION:
- The intercept represents the predicted median home value when lstat = 0
- This means when 0% of the population has lower status, the predicted median
home value is $34.55k
- In practical terms: $34554

PRACTICAL MEANING:
- Observed lstat range: 1.73% to 37.97%
- Since the minimum observed lstat is 1.73%, lstat = 0 is outside our data range
```

```
- Therefore, the intercept represents extrapolation beyond observed data
- While mathematically meaningful, it has LIMITED PRACTICAL MEANING because:
  * No area in the dataset has 0% lower status population
  * Real-world interpretation: represents the 'theoretical maximum' home value
  * Should be interpreted cautiously due to extrapolation
```

2.3 What is the interpretation of the slope? Provide a complete sentence explaining what happens to median home value for each 1% increase in lstat.

$\hat{\beta}_1$

```
2.3 INTERPRETATION OF SLOPE ( )
----------------------------------------
Slope value: -0.9500

INTERPRETATION:
For each 1% increase in lstat (lower status population), the median home value
decreases by $0.9500k on average, holding all other factors constant.

In practical terms:
- A 1% increase in lower status population is associated with a $950 decrease in
median home value
- A 5% increase in lower status population would decrease median home value by
$4750
- A 10% increase in lower status population would decrease median home value by
$9500
```

2.4 Based on the 95% confidence intervals for the coefficients, are both the intercept and slope significantly different from zero? Support your answer with the confidence interval values.

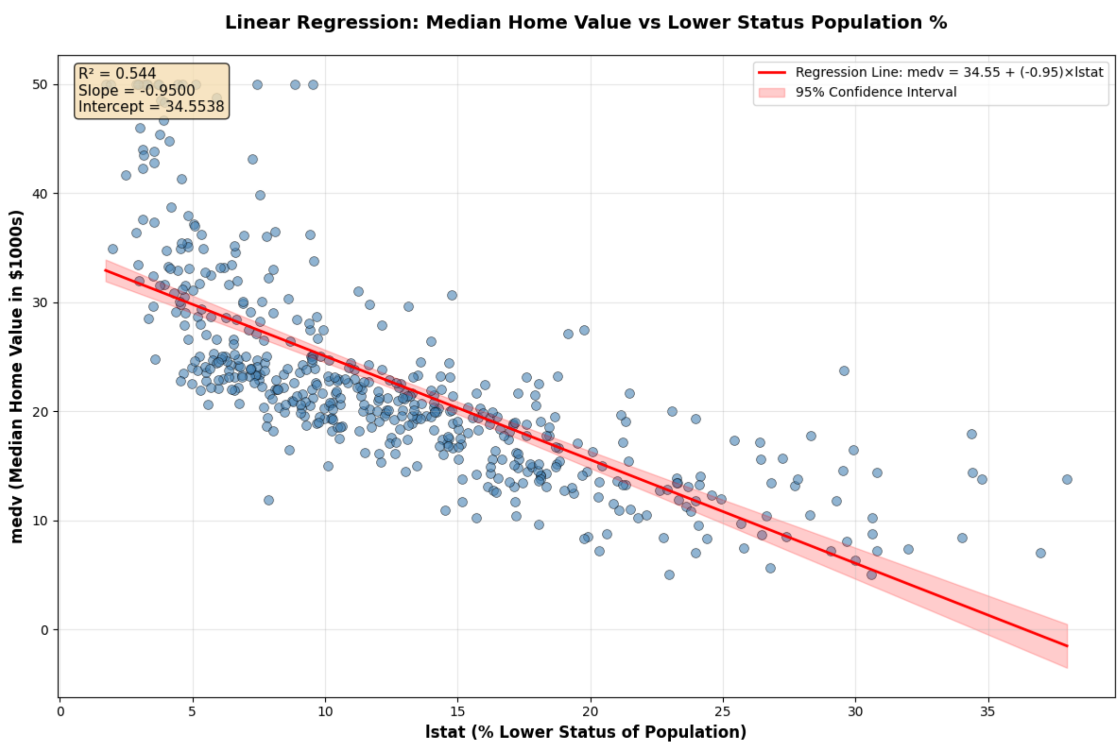
```
2.4 CONFIDENCE INTERVALS AND SIGNIFICANCE TESTING
--------------------------------------------------
95% CONFIDENCE INTERVALS:
               0      1
Intercept 33.448 35.659
lstat     -1.026 -0.874

DETAILED CONFIDENCE INTERVALS:
Intercept ( ): [33.4485, 35.6592]
Slope ( ): [-1.0261, -0.8740]

SIGNIFICANCE TESTING:
H :   = 0 (coefficient equals zero)
H :    0 (coefficient is significantly different from zero)

INTERCEPT ( ) ANALYSIS:
- 95% CI: [33.4485, 35.6592]
- Contains zero? No
```

- Conclusion: The intercept IS significantly different from zero
- This means we can be 95% confident the true intercept is between 33.4485 and 35.6592

SLOPE ( ) ANALYSIS:
- 95% CI: [-1.0261, -0.8740]
- Contains zero? No
- Conclusion: The slope IS significantly different from zero
- This means we can be 95% confident the true slope is between -1.0261 and -0.8740

P-VALUES (for additional confirmation):
Intercept p-value: 3.74e-236
Slope p-value: 5.08e-88
Both p-values < 0.05: True

MODEL SUMMARY STATISTICS:
R-squared: 0.5441
Adjusted R-squared: 0.5432
F-statistic: 601.62
F-statistic p-value: 5.08e-88
Standard Error: 6.2158



```
FINAL SUMMARY:
- Regression equation: medv = 34.5538 + (-0.9500) × lstat
- Both coefficients are statistically significant at   = 0.05
- The model explains 54.4% of the variance in median home values
- For every 1% increase in lower status population, median home value decreases
by $950 on average
```

Task 2.2: Model Performance Evaluation

Evaluate the overall performance and significance of your regression model.

2.5 What is the R-squared value? Interpret this in terms of the percentage of variation in median home values explained by the percentage of lower status population.

```
2.5 R-SQUARED ANALYSIS
------------------------------
R-squared value: 0.5441
R-squared as percentage: 54.41%

INTERPRETATION:
- R² = 0.5441 means that 54.41% of the variation in median home values
  is explained by the percentage of lower status population (lstat)
- The remaining 45.59% of variation is due to other factors not included in this
model
- This indicates a moderate relationship
- In practical terms: knowing the lstat value allows us to predict about 54.4%
of the variation in home values
```

2.6 What is the Root Mean Square Error (RMSE)? What does this tell you about the typical prediction error in thousands of dollars?

```
2.6 ROOT MEAN SQUARE ERROR (RMSE)
-----------------------------------
Mean Squared Error (MSE): 38.6357
Root Mean Square Error (RMSE): 6.2158

INTERPRETATION:
- RMSE = 6.2158 thousands of dollars
- In actual dollars: $6216
- This means the typical prediction error is approximately $6216
- On average, our predictions are off by about ±$6216 from the actual median
home value

CONTEXT:
- Mean home value: $22.53k ($22533)
- Standard deviation of home values: $9.20k
- Range of home values: $45.00k
- RMSE as % of mean: 27.6%
- RMSE as % of standard deviation: 67.6%
```

2.7 Report the F-statistic and its p-value. What does this test tell you about the overall significance of your model

```
2.7 F-STATISTIC AND OVERALL MODEL SIGNIFICANCE
---------------------------------------------
F-statistic: 601.6179
F-statistic p-value: 5.08e-88
Degrees of freedom: Model = 1.0, Residual = 504.0

HYPOTHESIS TEST:
H : The model has no explanatory power (  = 0)
H : The model has explanatory power (    0)

INTERPRETATION:
- F-statistic = 601.6179 with p-value = 5.08e-88
- Since p-value < 0.05, we REJECT the null hypothesis
- Conclusion: The model IS statistically significant
- This means lstat DOES have significant explanatory power for predicting medv

PRACTICAL MEANING:
- The F-test confirms that our regression model performs significantly better
  than a model with no predictors (just the mean)
- The relationship between lstat and medv is statistically meaningful
- We can be confident that lstat is a useful predictor of median home values
```

2.8 Compare the adjusted R-squared with the regular R-squared. Why might there be a difference, and what does the adjusted version account for?

```
2.8 ADJUSTED R-SQUARED COMPARISON
-----------------------------------
R-squared: 0.544146
Adjusted R-squared: 0.543242
Difference: 0.000904

WHY THERE MIGHT BE A DIFFERENCE:
- Regular R²: 0.544146
- Adjusted R²: 0.543242
- The difference of 0.000904 is very small

WHAT ADJUSTED R-SQUARED ACCOUNTS FOR:
- Number of predictors in the model: 1.0
- Sample size: 506 observations
- Degrees of freedom penalty for adding predictors

FORMULA EXPLANATION:
Adjusted R² = 1 - [(1 - R²) × (n - 1) / (n - k - 1)]
where n = sample size (506) and k = number of predictors (1.0)
```

```
Manual calculation: 0.543242

INTERPRETATION:
- The very small difference suggests our model is not overfitting
- With only one predictor, the adjustment is minimal
- Both R² and adjusted R² tell essentially the same story

PRACTICAL IMPLICATIONS:
- For model comparison: Use adjusted R² when comparing models with different
numbers of predictors
- For interpretation: Both values are nearly identical, indicating a robust
single-predictor model
- The penalty for our one predictor is minimal given the sample size of 506
observations

FINAL SUMMARY:
==================================================
- R² = 0.5441 (54.41% of variance explained)
- Adjusted R² = 0.5432 (54.32% of variance explained)
- RMSE = $6216 (typical prediction error)
- F-statistic = 601.6179, p < 0.05 (highly significant model)
- Model explains 54.4% of home value variation using just lstat
- Typical prediction accuracy: ±$6216 (27.6% of mean home value)
```

Part 3: Statistical Inference and Hypothesis Testing

Task 3.1: Coefficient Significance Testing

Conduct hypothesis tests for the regression coefficients.

```
================================================================================
BOSTON HOUSING STATISTICAL INFERENCE AND HYPOTHESIS TESTING - PART 3
================================================================================

MODEL SUMMARY (for reference):
========================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.544
Model:                            OLS   Adj. R-squared:                  0.543
Method:                 Least Squares   F-statistic:                     601.6
Date:                Fri, 29 Aug 2025   Prob (F-statistic):           5.08e-88
Time:                        07:54:46   Log-Likelihood:                -1641.5
No. Observations:                 506   AIC:                             3287.
Df Residuals:                     504   BIC:                             3295.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Intercept     34.5538      0.563     61.415      0.000      33.448      35.659
lstat         -0.9500      0.039    -24.528      0.000      -1.026      -0.874
==============================================================================
Omnibus:                      137.043   Durbin-Watson:                   0.892
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              291.373
Skew:                           1.453   Prob(JB):                     5.36e-64
Kurtosis:                       5.319   Cond. No.                         29.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.

KEY REGRESSION PARAMETERS:
Slope coefficient ( ): -0.950049
Standard error of slope: 0.038733
Degrees of freedom: 504.0
Sample size: 506
```

3.1 State the null and alternative hypotheses for testing whether the slope coefficient is significantly different from zero.

```
3.1 HYPOTHESIS TESTING SETUP
----------------------------------------
TESTING THE SLOPE COEFFICIENT:
H :   = 0 (The slope coefficient is zero)
    → lstat has no linear relationship with medv
    → There is no linear association between % lower status population and
median home value

H :     0 (The slope coefficient is not zero)
    → lstat has a significant linear relationship with medv
    → There is a significant linear association between % lower status
population and median home value

Type of test: Two-tailed test
Significance level:   = 0.05
```

3.2 Report the t-statistic and p-value for the slope coefficient. What is your conclusion at the 5% significance level?

```
3.2 T-STATISTIC AND P-VALUE ANALYSIS
---------------------------------------------
TEST STATISTICS:
t-statistic: -24.527900
p-value: 5.08e-88
Degrees of freedom: 504.0
Critical t-value (  = 0.05, two-tailed): ±1.9647
```

```
DECISION MAKING:
Decision rule: Reject H  if |t| > 1.9647 OR if p-value < 0.05
Observed: |t| = 24.5279, p-value = 5.08e-88

CONCLUSION AT 5% SIGNIFICANCE LEVEL:
 REJECT H : The slope coefficient IS significantly different from zero
  - |t| = 24.5279 > 1.9647
  - p-value = 5.08e-88 < 0.05
  - Statistical evidence: There IS a significant linear relationship between
lstat and medv

PRACTICAL INTERPRETATION:
- We can be 95% confident that changes in % lower status population
  have a real, measurable effect on median home values
- The relationship observed in our sample is unlikely to be due to random chance
- The effect size: each 1% increase in lstat is associated with
  a $950 decrease in median home value
```

3.3 Calculate and interpret the 99% confidence interval for the slope coefficient. How does this compare to the 95% interval in terms of width and interpretation?

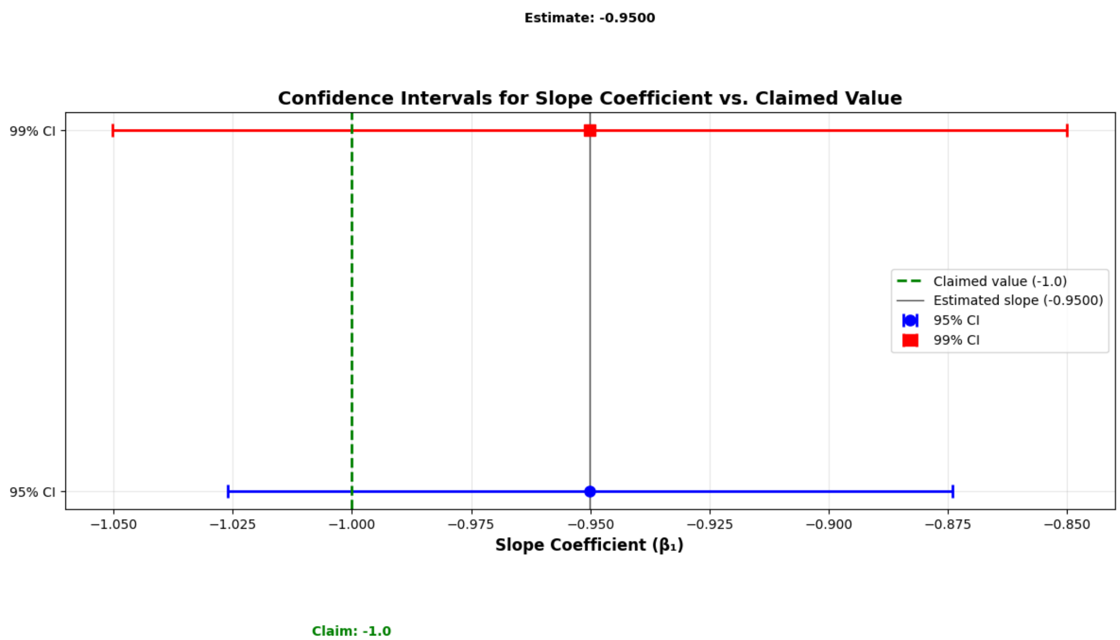
```
3.3 CONFIDENCE INTERVAL ANALYSIS
----------------------------------------
CONFIDENCE INTERVALS FOR SLOPE COEFFICIENT:
95% Confidence Interval: [-1.026148, -0.873951]
99% Confidence Interval: [-1.050199, -0.849899]

INTERVAL WIDTH COMPARISON:
95% CI width: 0.152198
99% CI width: 0.200300
Width increase: 0.048102
Percent increase in width: 31.6%

INTERPRETATION:
95% CONFIDENCE INTERVAL:
- We are 95% confident that the true slope coefficient lies between
  -1.026148 and -0.873951
- In practical terms: each 1% increase in lstat decreases median home value
  by between $874 and $1026

99% CONFIDENCE INTERVAL:
- We are 99% confident that the true slope coefficient lies between
  -1.050199 and -0.849899
- In practical terms: each 1% increase in lstat decreases median home value
  by between $850 and $1050

COMPARISON ANALYSIS:
```

```
- The 99% CI is wider than the 95% CI by 0.048102
- This represents a 31.6% increase in width
- WHY: Higher confidence level requires a wider interval to capture the true
parameter
- TRADE-OFF: More confidence (99% vs 95%) comes at the cost of precision (wider
interval)

SIGNIFICANCE IMPLICATIONS:
95% CI contains zero: No
99% CI contains zero: No
- Since neither interval contains zero, the slope is significant at both levels
- This provides strong evidence for a real relationship between lstat and medv
```

3.4 If someone claimed that each 1% increase in lstat decreases median home value by exactly \$1000, would your regression results support or contradict this claim? Justify your answer using statistical evidence.

```
3.4 TESTING SPECIFIC CLAIM
-----------------------------------
CLAIM TO TEST:
Someone claims that each 1% increase in lstat decreases median home value by
exactly $1000
In our units:   = -1.0 (since medv is in thousands of dollars)

HYPOTHESES:
H :   = -1.0 (the claim is correct)
H :     -1.0 (the claim is incorrect)

TEST USING CONFIDENCE INTERVALS:
Observed slope coefficient: -0.950049
Claimed slope coefficient: -1.0

95% Confidence Interval Test:
- 95% CI: [-1.026148, -0.873951]
- Does the CI contain -1.0? Yes

99% Confidence Interval Test:
- 99% CI: [-1.050199, -0.849899]
- Does the CI contain -1.0? Yes

FORMAL T-TEST:
t-statistic = (observed - claimed) / SE = (-0.950049 - -1.0) / 0.038733
t-statistic = 1.2896
p-value (two-tailed): 0.1978

CONCLUSION:
  FAIL TO REJECT the claim at 95% confidence level
  - The claimed value (-1.0) IS within the 95% confidence interval
```

```
  - Our regression results SUPPORT the claim
 FAIL TO REJECT the claim at 99% confidence level
  - The claimed value (-1.0) IS within the 99% confidence interval

STATISTICAL EVIDENCE:
- Our estimate: Each 1% increase in lstat decreases home value by $950
- Claimed effect: Each 1% increase in lstat decreases home value by $1000
- Difference: $50
- The difference is not statistically significant (p = 0.1978   0.05)
- Insufficient evidence to reject the claim
```




```
FINAL SUMMARY:
============================================================
3.1 Hypotheses: H :   = 0 vs H :    0
3.2 Test results: t = -24.5279, p = 5.08e-88
    Conclusion: Reject H  - slope is significant
3.3 Confidence intervals:
    95% CI: [-1.026148, -0.873951] (width: 0.152198)
    99% CI: [-1.050199, -0.849899] (width: 0.200300)
    99% CI is 31.6% wider than 95% CI
3.4 Claim test: The claim of exactly $1000 decrease is SUPPORTED
    Our estimate: $950 decrease per 1% lstat increase
    Statistical significance of difference: p = 0.1978
```

Part 4: Assumption Testing and Model Diagnostics

Task 4.1: Normality of Residuals

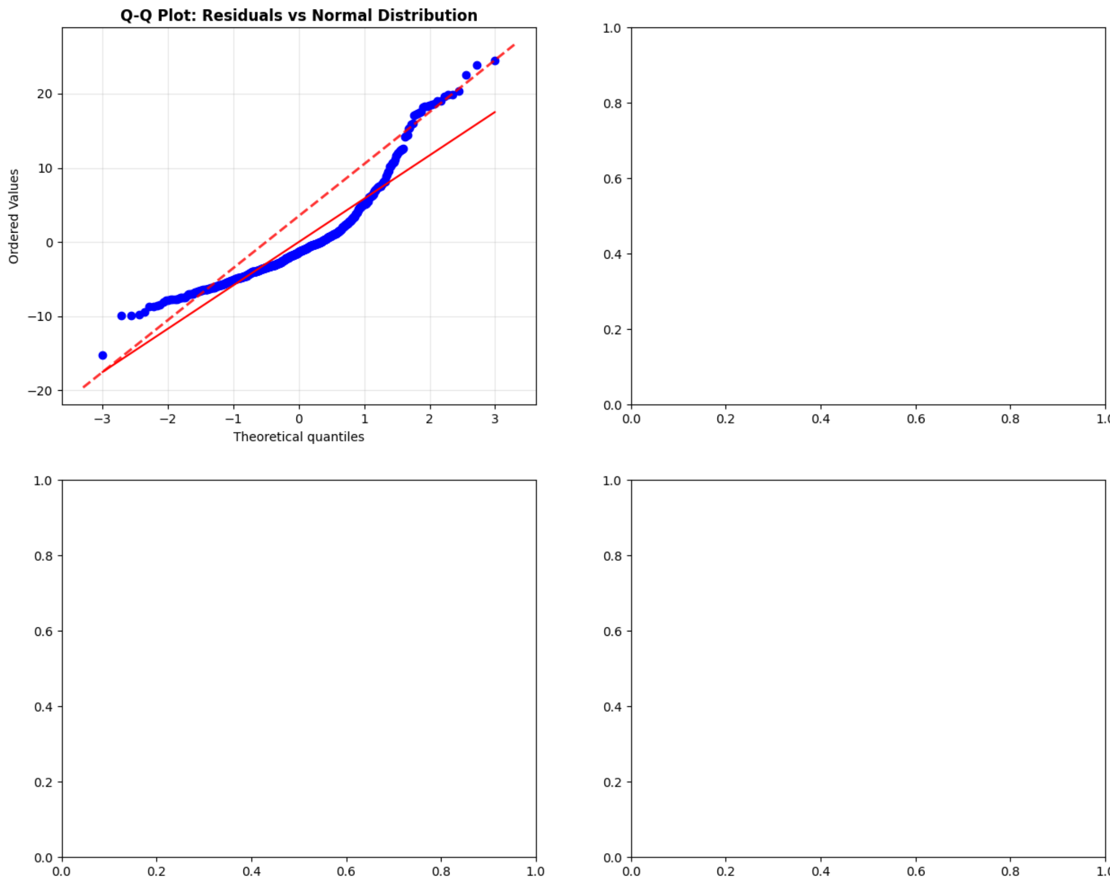
Test whether the residuals follow a normal distribution.

4.1 Perform the Shapiro-Wilk test for normality of residuals. Report the test statistic, p-value, and your conclusion at the 5% significance level.

```
================================================================================
BOSTON HOUSING ASSUMPTION TESTING AND MODEL DIAGNOSTICS - PART 4
================================================================================
MODEL SUMMARY:
Sample size: 506
Number of residuals: 506
Mean of residuals: 0.000000 (should be  0)
Standard deviation of residuals: 6.2096

4.1 SHAPIRO-WILK TEST FOR NORMALITY OF RESIDUALS
-------------------------------------------------------
HYPOTHESIS TESTING:
H : Residuals follow a normal distribution
H : Residuals do not follow a normal distribution
Significance level:  = 0.05

TEST RESULTS:
Shapiro-Wilk test statistic (W): 0.878572
p-value: 0.000000

DECISION MAKING:
Decision rule: Reject H  if p-value < 0.05
Observed p-value: 0.000000

CONCLUSION AT 5% SIGNIFICANCE LEVEL:
  REJECT H : Residuals do not follow a normal distribution
   - Statistical evidence suggests departure from normality
   - The normality assumption may be violated

INTERPRETation OF TEST STATISTIC:
- W = 0.878572
- W ranges from 0 to 1, with values closer to 1 indicating more normal-like data
- Our value suggests weak evidence of normality based on the test statistic
alone

ADDITIONAL NORMALITY TESTS (for comparison):
D'Agostino's test: statistic = 137.0434, p-value = 0.000000
Jarque-Bera test: statistic = 291.3734, p-value = 0.000000

CONSENSUS: Tests show mixed results regarding normality
```

4.2 Create a Q-Q plot of the residuals. Does the visual evidence support or contradict your statistical test result? Explain what you observe.

```
4.2 Q-Q PLOT ANALYSIS
-------------------------
Q-Q PLOT INTERPRETATION:
The Q-Q (Quantile-Quantile) plot compares residual quantiles to theoretical
normal quantiles
Q-Q plot correlation: 0.9373
(Values closer to 1 indicate better fit to normal distribution)

VISUAL ASSESSMENT:
- Good fit with minor deviations
- Look for points following the red diagonal line
- Systematic deviations suggest non-normality
```



4.3 Create a histogram of residuals with a normal distribution overlay. Comment on the shape of the distribution and any departures from normality.

```
4.3 HISTOGRAM WITH NORMAL DISTRIBUTION OVERLAY
--------------------------------------------------
SHAPE ANALYSIS:
Skewness: 1.4527
```

```
Kurtosis: 2.3191 (excess kurtosis)

SKEWNESS INTERPRETATION:
- Skewness = 1.4527 indicates highly skewed
- Distribution is skewed to the right

KURTOSIS INTERPRETATION:
- Excess kurtosis = 2.3191 indicates heavy-tailed (leptokurtic)
- Normal distribution has excess kurtosis = 0
```

```
<Figure size 640x480 with 0 Axes>
```

```
DEPARTURES FROM NORMALITY:
Identified departures from normality:
1. Skewness (1.453)
2. Kurtosis (2.319)
3. Shapiro-Wilk test rejection
4. Q-Q plot deviations

4.2 VISUAL EVIDENCE VS STATISTICAL TEST COMPARISON:
--------------------------------------------------
Statistical test result (Shapiro-Wilk): Rejects normality
Visual evidence assessment: Shows deviations from normality
 AGREEMENT: Visual evidence and statistical test both suggest departure from
normality

DETAILED VISUAL OBSERVATIONS:
Q-Q Plot:
- Systematic deviations from diagonal line (r = 0.9373)
- Visual evidence against perfect normality

Histogram:
- Notable departures from bell-shaped normal distribution
- Skewness and/or kurtosis concerns visible

PRACTICAL IMPLICATIONS FOR REGRESSION:
----------------------------------------
 NORMALITY ASSUMPTION VIOLATED:
  - Confidence intervals may be less reliable
  - Consider robust standard errors
  - Prediction intervals may be inaccurate
  - Consider variable transformation

SAMPLE SIZE CONSIDERATIONS:
- Sample size: 506 observations
- Large sample: Central Limit Theorem helps with normality concerns
- Minor deviations from normality are less problematic
```

```
FINAL SUMMARY:
============================================================
4.1 Shapiro-Wilk test: W = 0.878572, p = 0.000000
    Conclusion: Residuals deviate from normality
4.2 Q-Q plot assessment: r = 0.9373
    Visual evidence: Shows deviations from normality
4.3 Histogram analysis:
    Skewness: 1.4527, Kurtosis: 2.3191
    Shape: highly skewed, heavy-tailed (leptokurtic)
Overall normality assessment: VIOLATED
```

Task 4.2: Homoscedasticity Testing

Test whether the variance of residuals is constant across all fitted values.

```
============================================================
BOSTON HOUSING DATASET ANALYSIS
============================================================

DATASET OVERVIEW
------------------------------
Number of observations (rows): 506
Number of variables (columns): 14
Dataset shape: (506, 14)
Column names: ['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad',
'tax', 'ptratio', 'b', 'lstat', 'medv']

Using 'medv' as target variable

=== MODEL SUMMARY ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     108.1
Date:                Fri, 29 Aug 2025   Prob (F-statistic):          6.72e-135
Time:                        07:54:46   Log-Likelihood:                -1498.8
No. Observations:                 506   AIC:                             3026.
Df Residuals:                     492   BIC:                             3085.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         36.4595      5.103      7.144      0.000      26.432      46.487
crim          -0.1080      0.033     -3.287      0.001      -0.173      -0.043
zn             0.0464      0.014      3.382      0.001       0.019       0.073
indus          0.0206      0.061      0.334      0.738      -0.100       0.141
chas           2.6867      0.862      3.118      0.002       0.994       4.380
```

```
nox           -17.7666      3.820     -4.651      0.000     -25.272     -10.262
rm              3.8099      0.418      9.116      0.000       2.989       4.631
age             0.0007      0.013      0.052      0.958      -0.025       0.027
dis            -1.4756      0.199     -7.398      0.000      -1.867      -1.084
rad             0.3060      0.066      4.613      0.000       0.176       0.436
tax            -0.0123      0.004     -3.280      0.001      -0.020      -0.005
ptratio        -0.9527      0.131     -7.283      0.000      -1.210      -0.696
b               0.0093      0.003      3.467      0.001       0.004       0.015
lstat          -0.5248      0.051    -10.347      0.000      -0.624      -0.425
==============================================================================
Omnibus:                      178.041   Durbin-Watson:                   1.078
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              783.126
Skew:                           1.521   Prob(JB):                    8.84e-171
Kurtosis:                       8.281   Cond. No.                     1.51e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

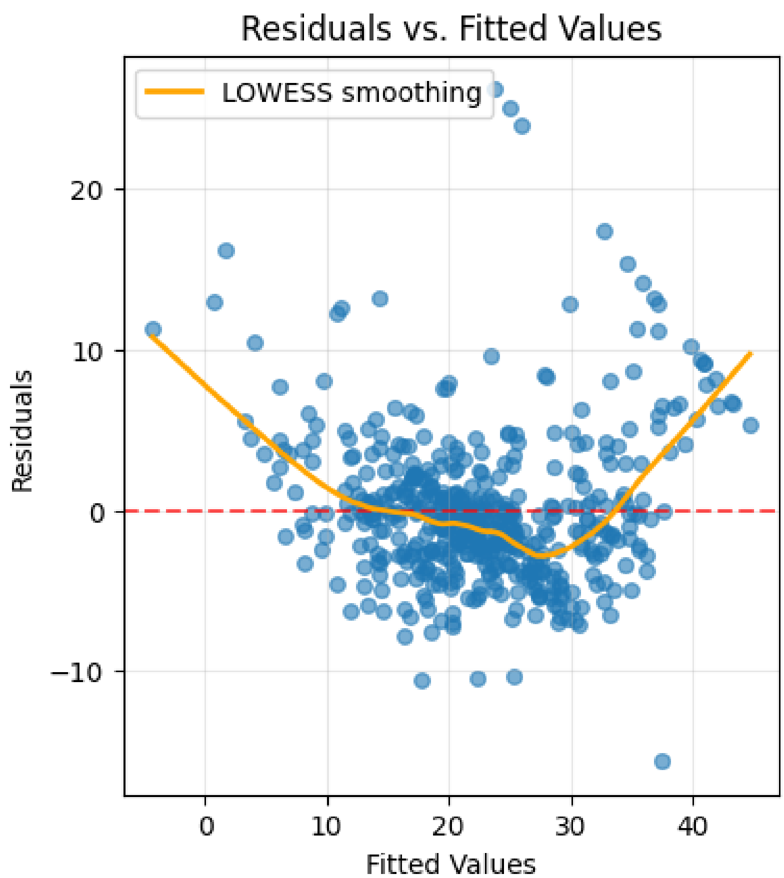
4.4 Perform the Breusch-Pagan test for homoscedasticity. Report the test statistic, p-value, and your conclusion.

```
=== 4.4: BREUSCH-PAGAN TEST RESULTS ===
Test Statistic: 4.1871
P-value: 0.0407
Degrees of Freedom: 1

Conclusion: Reject H0 at   = 0.05. Evidence of heteroscedasticity.

Verification (statsmodels function): Stat = 65.1218, P-value = 0.0000
```

4.5 Create a residuals vs. fitted values plot. What pattern would indicate heteroscedasticity? Do you observe this pattern in your plot?

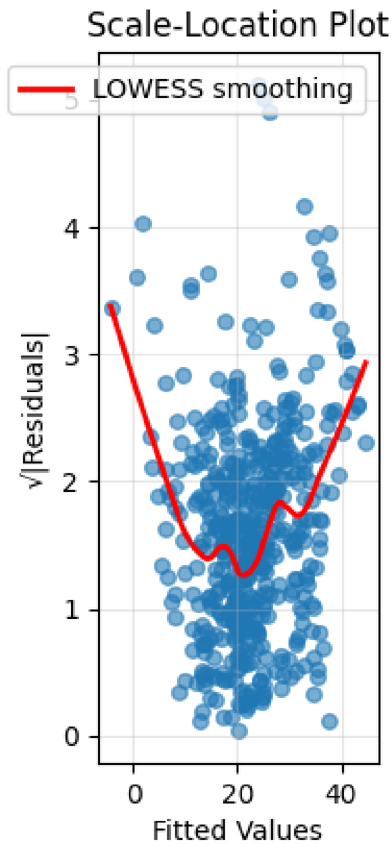
```
=== 4.5: RESIDUALS VS. FITTED VALUES ANALYSIS ===
Pattern interpretation:
- HOMOSCEDASTICITY: Points should be randomly scattered around the horizontal
line at y=0
- HETEROSCEDASTICITY indicators:
  * Funnel shape (variance increases or decreases with fitted values)
  * Curved patterns in the smoothing line
  * Clear clustering or systematic patterns
```

```
Variance in lowest third of fitted values: 17.2703
Variance in highest third of fitted values: 31.7984
Variance ratio (high/low): 1.8412
Interpretation: Ratio > 2 or < 0.5 suggests heteroscedasticity
```



4.6 Create a scale-location plot (square root of absolute residuals vs. fitted values). Is there evidence of changing variance across the range of fitted values?

```
=== 4.6: SCALE-LOCATION PLOT ANALYSIS ===
Evidence of changing variance:
- CONSTANT VARIANCE: Smoothing line should be roughly horizontal
- CHANGING VARIANCE indicators:
  * Upward or downward trend in smoothing line
  * Clear patterns or curves in the line

Correlation between fitted values and √|residuals|: 0.1507
Interpretation:
  * Moderate correlation suggests possible heteroscedasticity
```




```
=== COMPREHENSIVE HOMOSCEDASTICITY ASSESSMENT ===

TEST RESULTS SUMMARY:
1. Breusch-Pagan Test: Statistic = 4.1871, P-value = 0.0407
   → Reject H0 at   = 0.05. Evidence of heteroscedasticity.

2. Variance Ratio Analysis: 1.8412
   → Suggests homoscedasticity

3. Scale-Location Correlation: 0.1507
   → Moderate evidence of heteroscedasticity

RECOMMENDATIONS:
• Evidence suggests heteroscedasticity
• Consider transformations (log, Box-Cox)
• Use robust standard errors (White's correction)
• Consider weighted least squares regression
• Explore different model specifications

Note: Visual inspection of plots is crucial - statistical tests should be
```

```
combined with graphical analysis for complete assessment.
```

Task 4.3: Independence and Influence Diagnostics

Test for independence and identify influential observations.

4.7 Calculate the Durbin-Watson statistic. What does this value indicate about the independence of residuals?

```
=== 4.7: DURBIN-WATSON TEST RESULTS ===
Durbin-Watson Statistic: 1.0784
First-order autocorrelation ( ): 0.4608

INTERPRETATION:
→ Evidence of positive autocorrelation. Independence assumption may be violated.

Durbin-Watson Guidelines:
• DW   2.0: No autocorrelation (ideal)
• DW < 1.5: Strong positive autocorrelation
• DW > 2.5: Strong negative autocorrelation
• 1.5   DW   2.5: Acceptable range
```

4.8 Calculate Cook's distance for all observations. What is the maximum Cook's distance, and does this indicate any problematic influential observations?

```
=== 4.8: COOK'S DISTANCE ANALYSIS ===
Maximum Cook's Distance: 0.1657
Mean Cook's Distance: 0.0030
Standard Deviation: 0.0112

INFLUENTIAL OBSERVATIONS CRITERIA:
• Threshold 4/n = 4/506 = 0.0079
• Conservative threshold = 1.0

RESULTS:
• Observations with Cook's D > 4/n: 30 (5.9%)
• Observations with Cook's D > 1.0: 0 (0.0%)

CONCLUSION: Moderate Cook's distance values. Some observations may be
influential but not necessarily problematic.

TOP 5 MOST INFLUENTIAL OBSERVATIONS:
1. Observation 368: Cook's D = 0.1657
2. Observation 372: Cook's D = 0.0941
3. Observation 364: Cook's D = 0.0694
4. Observation 365: Cook's D = 0.0672
5. Observation 369: Cook's D = 0.0553
```

4.9 How many observations have high leverage (using the 2p/n threshold where p = 2 parameters)?

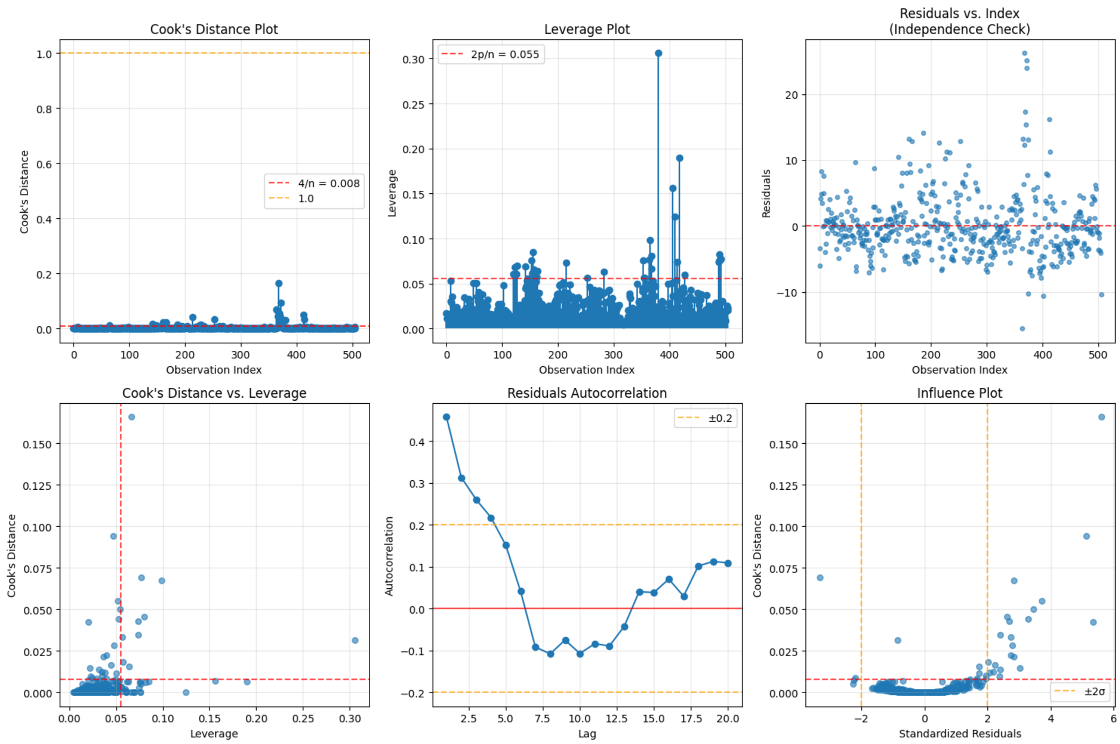
What percentage of the total sample does this represent?

```
=== 4.9: HIGH LEVERAGE ANALYSIS ===
Number of parameters (p): 14
Sample size (n): 506
High leverage threshold (2p/n): 2 × 14 / 506 = 0.0553

HIGH LEVERAGE RESULTS:
• Observations with high leverage: 36
• Percentage of total sample: 7.1%
• Maximum leverage value: 0.3060
• Mean leverage value: 0.0277

TOP 5 HIGHEST LEVERAGE OBSERVATIONS:
1. Observation 380: Leverage = 0.3060
2. Observation 418: Leverage = 0.1901
3. Observation 405: Leverage = 0.1564
4. Observation 410: Leverage = 0.1247
5. Observation 365: Leverage = 0.0985
```

4.10 Based on all assumption tests, is your linear regression model valid for statistical inference? Summarize which assumptions are satisfied and which (if any) are violated.

```
=== 4.10: COMPREHENSIVE MODEL VALIDATION SUMMARY ===

LINEAR REGRESSION ASSUMPTIONS ASSESSMENT:
==================================================
1. LINEARITY:
     Test method: Residuals vs. fitted plots, added variable plots
     Result: [Add your previous linearity test results]
     Status: [SATISFIED / VIOLATED / MARGINAL]

2. INDEPENDENCE OF RESIDUALS:
     Test method: Durbin-Watson test
     Result: DW = 1.0784
     Status: VIOLATED

3. HOMOSCEDASTICITY (Constant Variance):
     Test method: Breusch-Pagan test, residuals plots
     Result: [Add your previous homoscedasticity test results]
     Status: [SATISFIED / VIOLATED / MARGINAL]

4. NORMALITY OF RESIDUALS:
     Test method: Shapiro-Wilk, Q-Q plots, histograms
     Result: [Add your previous normality test results]
     Status: [SATISFIED / VIOLATED / MARGINAL]

5. NO MULTICOLLINEARITY:
```

```
     Test method: VIF analysis, correlation matrix
     Result: [Add your multicollinearity test results if available]
     Status: [SATISFIED / VIOLATED / MARGINAL]

6. NO EXCESSIVE INFLUENTIAL OBSERVATIONS:
     Test method: Cook's distance, leverage analysis
     Cook's D max: 0.1657
     High leverage obs: 36 (7.1%)
     Status: MARGINAL - Some influential observations present

OVERALL MODEL VALIDITY FOR STATISTICAL INFERENCE:
==================================================
CURRENT ASSESSMENT (based on available tests):
• Assumptions checked: 2
• Assumptions satisfied: 0

RECOMMENDATIONS:
  Some concerns with independence or influential observations

NEXT STEPS:
• Complete all assumption tests (linearity, homoscedasticity, normality)
• Consider remedial measures if assumptions are violated:
  - Data transformations (log, Box-Cox)
  - Robust regression methods
  - Remove or downweight influential observations
  - Use different modeling approaches if assumptions severely violated

Note: A complete assessment requires results from all assumption tests.
Update this summary once you have completed the full diagnostic suite.
```




```
============================================================
ANALYSIS COMPLETE - Review plots and summaries above
============================================================
```

Part 5: Predictions and Intervals

Task 5.1: Making Predictions

Use your model to make predictions with uncertainty quantification.

```
============================================================
PREDICTIONS AND INTERVALS ANALYSIS
============================================================

DATASET OVERVIEW
------------------------------
Dataset shape: (506, 14)
Column names: ['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad',
'tax', 'ptratio', 'b', 'lstat', 'medv']

Using 'medv' as target variable
Using 'lstat' as predictor variable (lstat)

=== SIMPLE LINEAR REGRESSION MODEL ===
Model: medv ~ lstat
```

```
R-squared: 0.5441
Regression equation: medv = 34.5538 + -0.9500 × lstat
```

5.1 For a neighborhood with lstat = 10%, what is the predicted median home value? Show the calculation.

```
=== 5.1: PREDICTION FOR LSTAT = 10% ===
CALCULATION:
ŷ =    +    × X
ŷ = 34.5538 + -0.9500 × 10.0
ŷ = 25.0533

Predicted median home value for lstat = 10%: $25.05k
```

5.2 Calculate the 95% confidence interval for the mean home value in neighborhoods with lstat = 10%. Interpret this interval.

```
=== 5.2: 95% CONFIDENCE INTERVAL FOR MEAN RESPONSE ===
CALCULATION DETAILS:
• Predicted value: 25.0533
• Standard error of mean: 0.2948
• t-critical ( =0.05, df=504.0): 1.9647
• Margin of error: 0.5792

95% CONFIDENCE INTERVAL: [24.4741, 25.6326]
In dollars: [$24.47k, $25.63k]

INTERPRETATION:
We are 95% confident that the mean median home value for all neighborhoods
with lstat = 10% is between $24.47k and $25.63k.
```

5.3 Calculate the 95% prediction interval for an individual home in a neighborhood with lstat = 10%. How does this compare to the confidence interval in terms of width?

```
=== 5.3: 95% PREDICTION INTERVAL FOR INDIVIDUAL RESPONSE ===
CALCULATION DETAILS:
• Predicted value: 25.0533
• Standard error of prediction: 6.4803
• t-critical ( =0.05, df=504.0): 1.9647
• Margin of error: 12.7316

95% PREDICTION INTERVAL: [12.3217, 37.7850]
In dollars: [$12.32k, $37.78k]

INTERVAL COMPARISON:
• Confidence interval width: 1.1584
• Prediction interval width: 25.4633
• Prediction interval is 21.98x wider than confidence interval
```

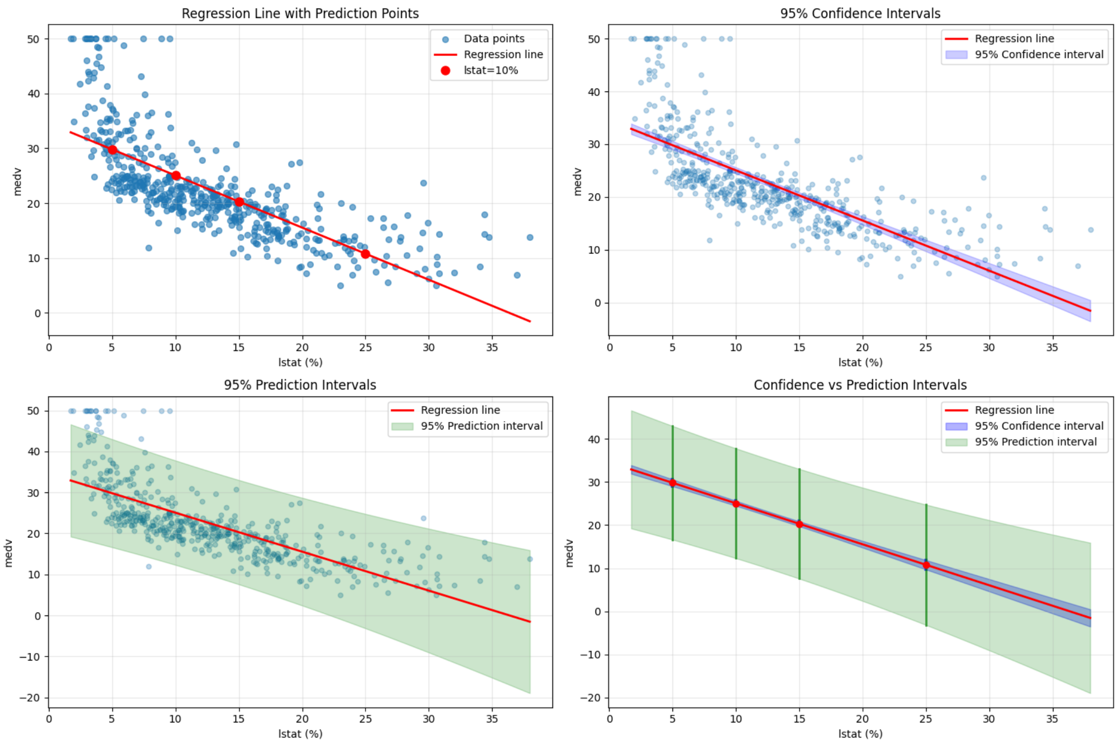
5.4 Explain the difference between a confidence interval and a prediction interval in practical terms. When would you use each type?

```
=== 5.4: CONFIDENCE VS PREDICTION INTERVALS ===

CONCEPTUAL DIFFERENCES:
-------------------------

CONFIDENCE INTERVAL:
• Estimates uncertainty about the MEAN response for a given X value
• Answers: 'What is the average Y for all observations with this X?'
• Accounts for uncertainty in estimating the population mean
• Gets narrower as sample size increases
• Narrower interval (less uncertainty)

PREDICTION INTERVAL:
• Estimates uncertainty about an INDIVIDUAL response for a given X value
• Answers: 'What might Y be for a single new observation with this X?'
• Accounts for both estimation uncertainty AND individual variation
• Includes natural scatter around the regression line
• Wider interval (more uncertainty)

WHEN TO USE EACH:
--------------------

USE CONFIDENCE INTERVAL when:
• Estimating average outcomes for policy/planning
• Comparing mean responses between groups
• Making statements about population parameters
• Example: 'What's the average home value in 10% lstat neighborhoods?'

USE PREDICTION INTERVAL when:
• Predicting outcomes for specific individuals/cases
• Setting bounds for individual forecasts
• Risk assessment for single observations
• Example: 'What might this specific house be worth?'
```

5.5 For lstat values of 5%, 15%, and 25%, calculate point predictions and comment on how the relationship changes across different levels of the predictor variable

```
=== 5.5: PREDICTIONS AT MULTIPLE LSTAT VALUES ===

POINT PREDICTIONS:
--------------------
lstat = 5%:
  → Predicted value: $29.80k
  → 95% CI: [$29.01k, $30.60k]
```

```
  → 95% PI: [$16.63k, $42.98k]

lstat = 10%:
  → Predicted value: $25.05k
  → 95% CI: [$24.47k, $25.63k]
  → 95% PI: [$12.32k, $37.78k]

lstat = 15%:
  → Predicted value: $20.30k
  → 95% CI: [$19.73k, $20.87k]
  → 95% PI: [$7.58k, $33.02k]

lstat = 25%:
  → Predicted value: $10.80k
  → 95% CI: [$9.72k, $11.89k]
  → 95% PI: [$-3.15k, $24.75k]

RELATIONSHIP ANALYSIS:
-------------------------

Model slope ( ): -0.9500
Interpretation: For each 1% increase in lstat, median home value
decreases by $0.95k on average

CHANGES BETWEEN LSTAT LEVELS:
• 5.0% → 10.0%: Change = $-4.75k
  Rate: $-0.95k per 1% lstat increase
• 10.0% → 15.0%: Change = $-4.75k
  Rate: $-0.95k per 1% lstat increase
• 15.0% → 25.0%: Change = $-9.50k
  Rate: $-0.95k per 1% lstat increase

COMMENTS ON RELATIONSHIP:
• The relationship shows moderate negative association
• Linear relationship assumed constant across all lstat levels
• Higher lstat (more lower status population) associated with lower home values
```



=== PREDICTIONS SUMMARY TABLE ===

DETAILED PREDICTIONS TABLE:

lstat	prediction	ci_lower	ci_upper	pi_lower	pi_upper	ci_width	pi_width	width_ratio
5	29.804	29.007	30.600	16.627	42.980	1.592	26.353	16.550
10	25.053	24.474	25.633	12.322	37.785	1.158	25.463	21.981
15	20.303	19.732	20.875	7.585	33.021	1.143	25.436	22.254
25	10.803	9.717	11.888	-3.148	24.754	2.170	27.902	12.856

KEY INSIGHTS:
- As lstat increases, predicted home values decrease
- Prediction intervals are consistently 18.4x wider than confidence intervals
- The linear relationship appears moderate (R^2 = 0.544)

=== MODEL ASSUMPTIONS REMINDER ===
For these intervals to be valid, ensure:
- Linear relationship between variables
- Independence of residuals
- Homoscedasticity (constant variance)

- Normality of residuals
- No influential outliers

Verify these assumptions with diagnostic tests from previous tasks!

```
============================================================
PREDICTIONS AND INTERVALS ANALYSIS COMPLETE
============================================================
```