In-Class Assignment 1: Simple Linear Regression (SLR) in Practice

Boston Housing: Predicting `medv` with one predictor

Team Members: *Ezequiel Andrade* *Adrian Muro* *Krrithik Ezhilarasan*

California State University, Bakersfield

1/30/2026

Problem + Data + Variables

- **Goal:** Predict median home value using one predictor at a time (SLR).
- **Dataset:** Boston Housing (R: `MASS::Boston`).
- **Response:** $Y = \texttt{medv}$ (median home value, in \$1000s).

Predictors (one per model):

- Model 1: $X_1 = \texttt{rm}$ (avg rooms per dwelling)
- Model 2: $X_2 = \texttt{crim}$ (per-capita crime rate)
- Model 3: $X_3 = \texttt{lstat}$ (% lower status)

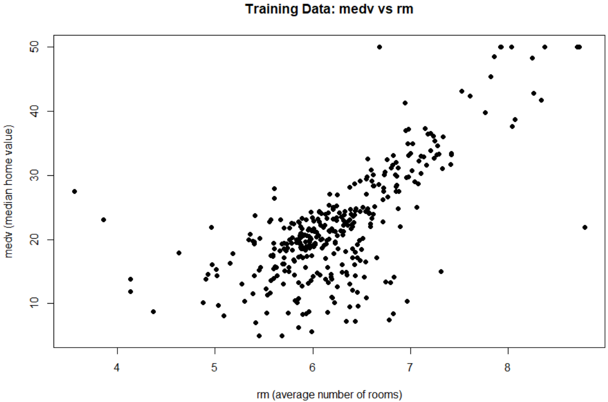
Evaluation Plan (Fair Comparison)

- Split data into **Training (70%)** and **Test (30%)**.
- Record the random seed so results are reproducible.
- **No data leakage:** Outlier rules / transformations decided using training only.
- Compare final models using **Test MSE** (smaller is better).
- Also report **Training** R^2 (for interpretation, not ranking).

$$\text{MSE}_{test} = \frac{1}{n_{test}} \sum_{i \in test} (y_i - \hat{y}_i)^2$$

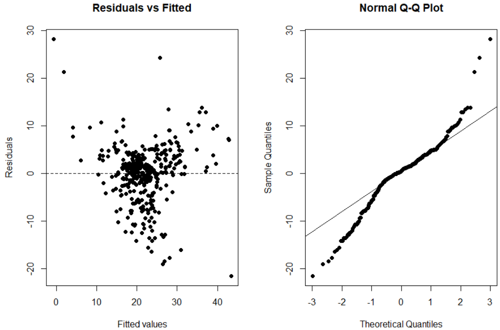
Model 1: medv ∼ rm

Key scatterplot (train):



Final fitted equation: -30.886 + 8.459 * rm

One diagnostic plot:

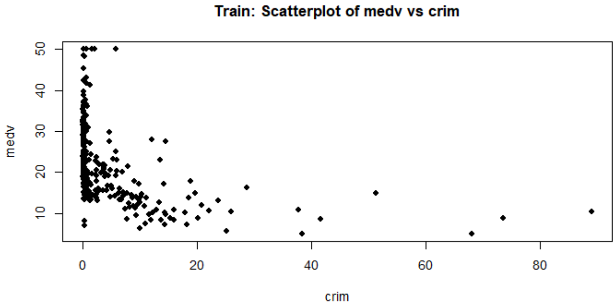


Results (final model):

- Training R^2: *0.491*
- Test MSE: *56.84917*

Model 2: `medv` ~ `crim` (Outliers Present)

Scatterplot (train):



Fix (training only):

- 1.5×IQR rule on `crim`
- Removed: **52** training points
- Upper bound: **8.3674**

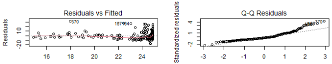
Final equation (cleaned train):

$$\hat{y} = 24.9100 - 1.2264(\texttt{crim})$$

Results (final model):

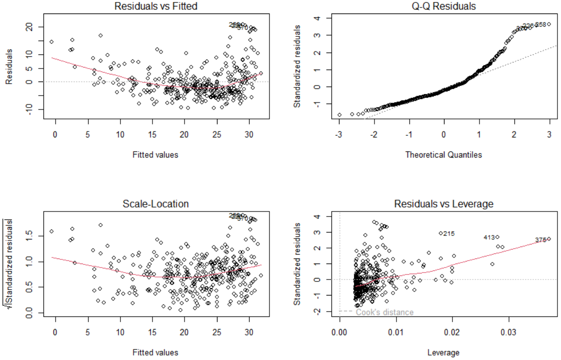
- Training R^2: **0.0793**
- Test MSE: **112.2664**

One diagnostic plot (cleaned model):

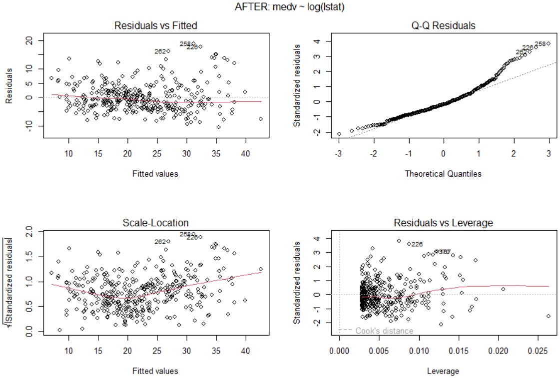


Model 3: `medv` ∼ `lstat` (Transformation Needed)

Untransformed diagnostics show problems:



Improved diagnostics (after transform):



Fix (training only): Log transformation

Final equation:

*50.70 - 11.99 * log(lstat)*

Final Comparison + Recommendation

Comparison table (final models):

Model	Predictor	Training R^2	Test MSE	Notes (issues / fix)
3	lstat	*0.675*	*37.74*	*Nonlinearity/heterosced; used log(lstat) to improve diagnostics*
1	rm	*0.491*	*56.85*	*Clean SLR; assumptions mostly OK*
2	crim	0.0793	112.2664	Outliers in `crim`; removed 52 train points (1.5×IQR)

Ranking (by Test MSE): *(fill in best → worst)*

Recommendation (1–2 sentences):

- Choose the model with the **lowest Test MSE**.
- Mention interpretability (simple story) + performance (test MSE).

Limitations + Next Steps (Optional)

- Only one predictor at a time (SLR can miss important factors).
- Outlier handling / transformations can change the fitted story.
- Next step: multiple regression using more predictors.
- Next step: cross-validation to check stability.

Backup: Definitions (if asked)

- **Outlier:** unusual response relative to the fitted model.
- **Influential point:** a point that noticeably changes the fit if removed.
- **Training R^2:** fraction of training variability explained by the line.
- **Test MSE:** average squared prediction error on unseen data.