IN-CLASS ASSIGNMENT 1

Paul Gascon ,Christian Rodriguez, Juan Rodriguez

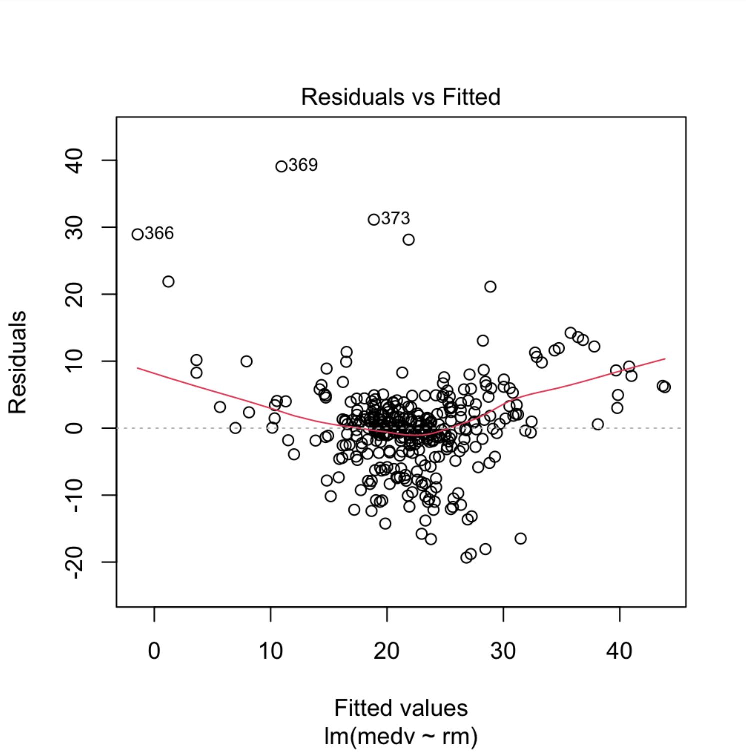
Problem Statement, dataset, and variables

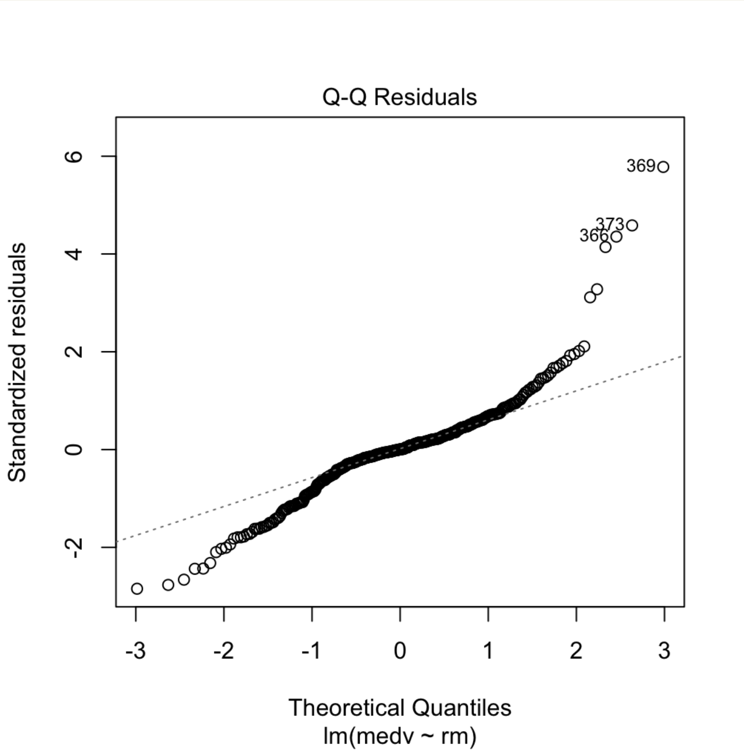
- In the dataset, *Boston*, we are provided 506 observations with the predictors:
  - medv, rm, crim, lstat
- Using seed 123, we split the data into a 70% training set and 30% test set.

- Test data: 152 observations
- Train data: 354 observations

Model 1: X_1: Well-Behaved Relationship

- Plot shows a strong positive correlation
- There is a mild visible curvature
- There are a few obvious outliers
- $\widehat{medv} = -32.677 + 8.773\, rm$





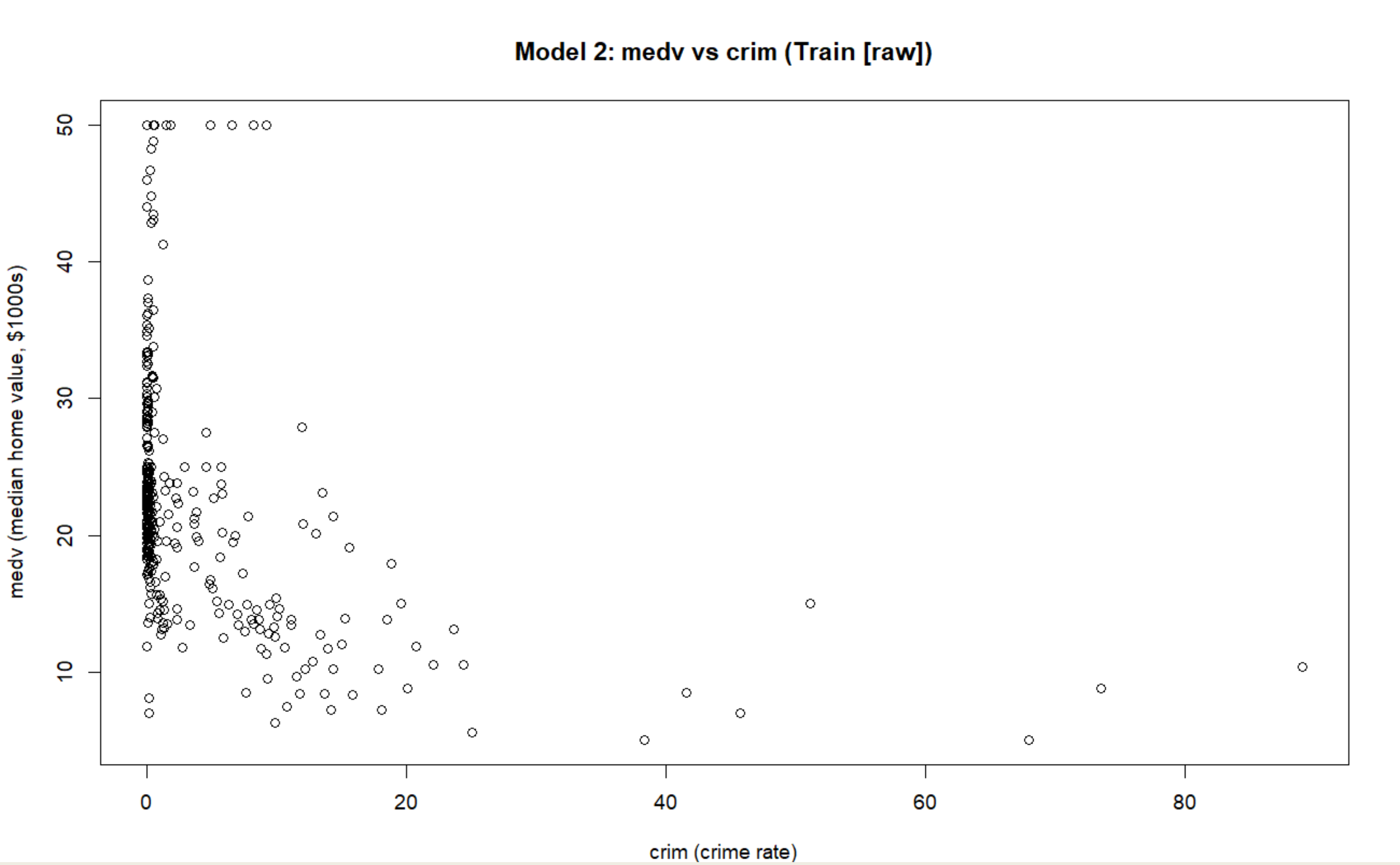
MODEL 1: RESIDUAL VS FITTED VALUES & NORMAL Q-Q PLOT OF RESIDUALS

- Residuals are centered around zero with only mild curvature ⇒ linearity is reasonably satisfied.
- The residual spread is fairly consistent, with a slight increase at higher fitted values ⇒ no serious heteroscedasticity.
- Residuals follow the reference line closely with some upper-tail deviation ⇒ residuals are approximately normal.

Model 1: ANOVA Interpretation

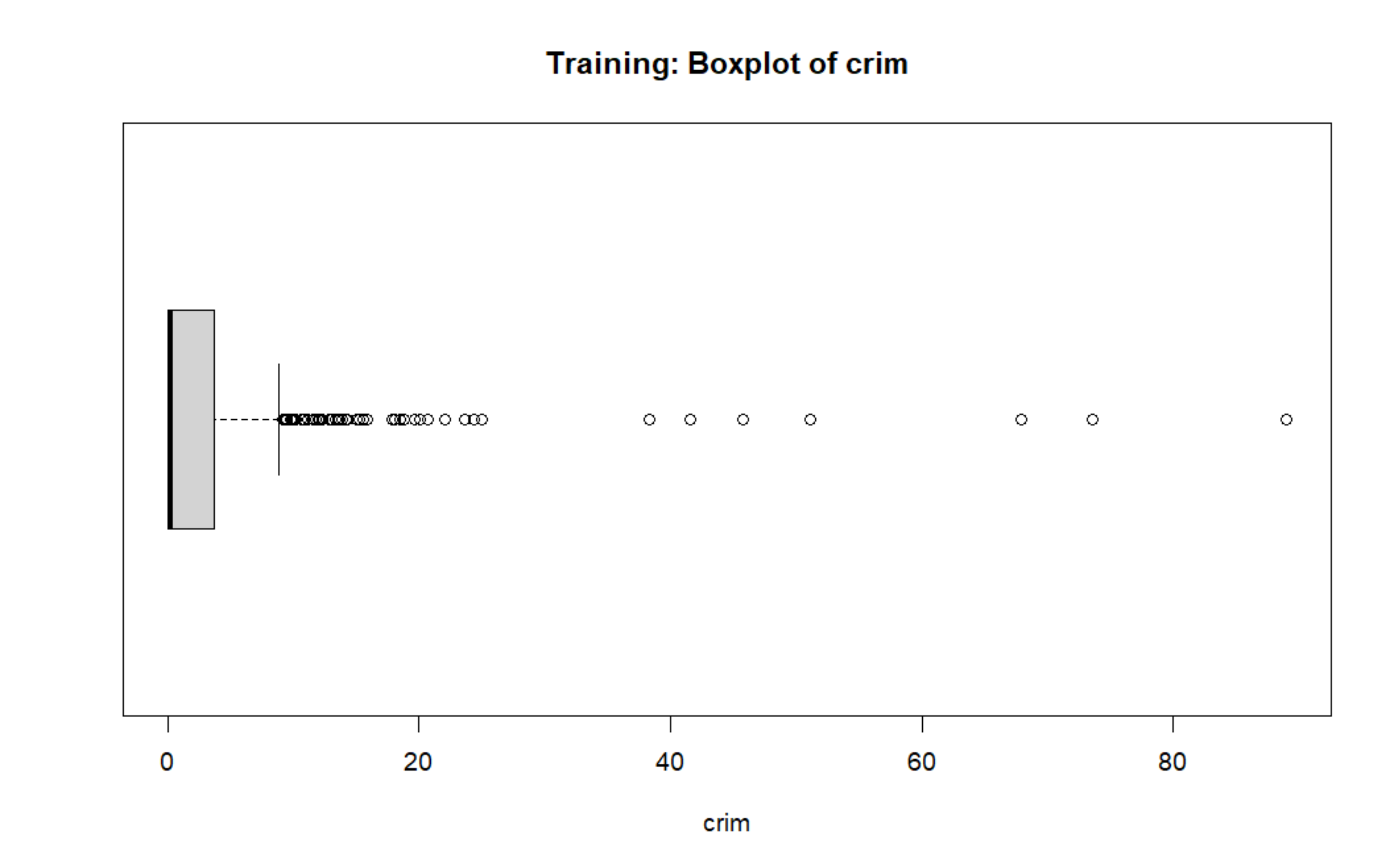
Source	Df	Sum Sq	Mean Sq	F value	P-value
rm	1	12893	12892.8	278.59	2.2 e-16
Residuals	352	16290	46.3		

- The F-test tests whether the regression using rm explains significantly more variation in medv than a model with no predictor
- Since α = 0.05 > p-value = 2.2 e-16, the regression is statistically significant
- About 44.18% of the variability in median home value ($R^2 = 0.4418$)
- Prediction: 38.1652 (predicted median home value in $1000)
- Test MSE: 38.17: the average squared difference between the predicted and actual median home is about 38 ($1000)^2



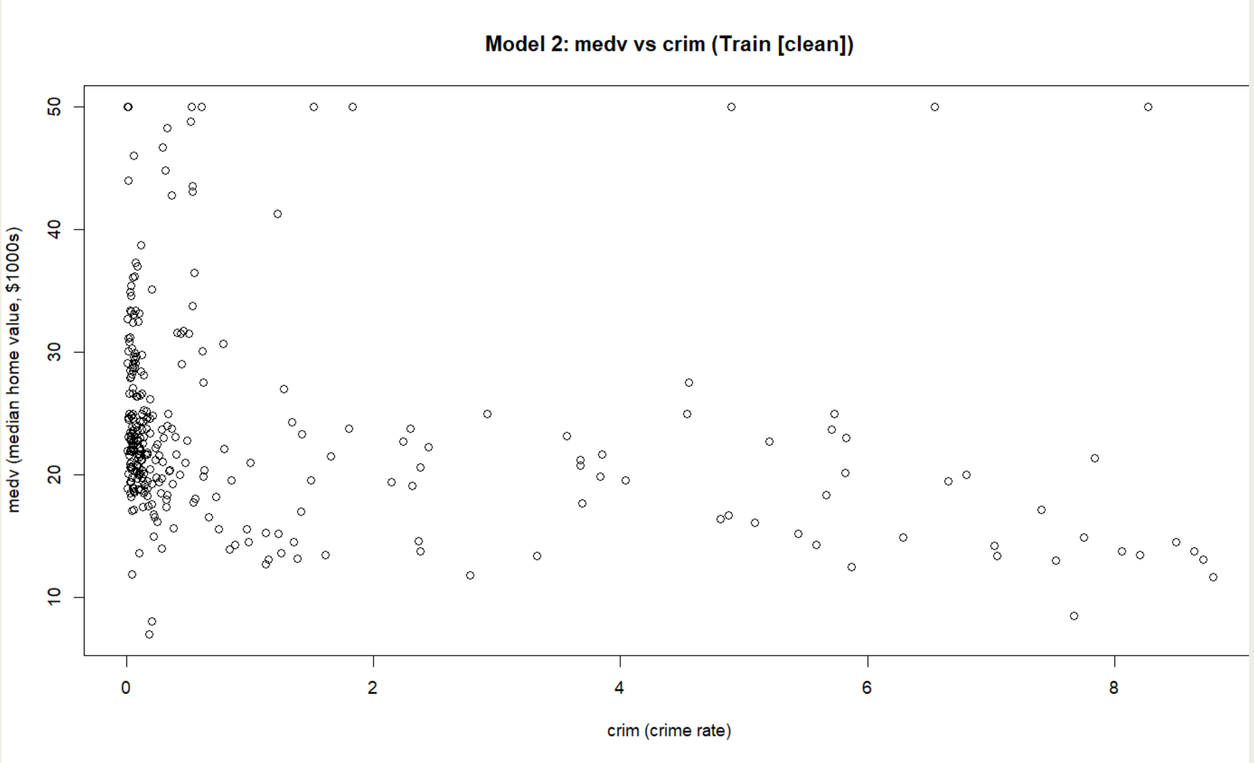
Model 2: medv vs crim

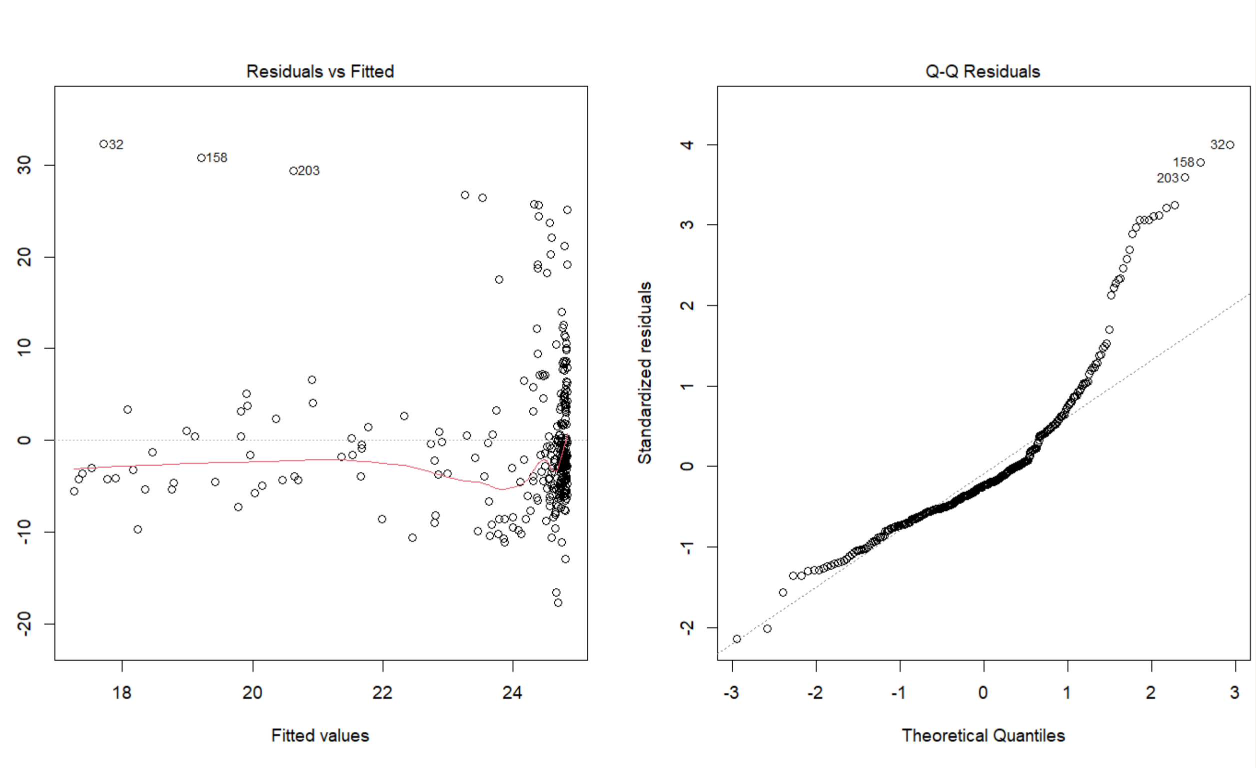
- Median home value generally decreases as crime rate increases.
- Most neighborhoods have very low crime rates and are tightly clustered high valued median home value.
- A few neighborhoods have extremely high crime rates and stand far apart from the rest of the data.
- These extreme points are likely influencing the regression line strongly.



Detect outliers / influential points

- The boxplot shows several extreme crime rate values beyond the upper whisker, indicating the presence of outliers in the training data.
- These outliers were removed using the standard 1.5×IQR rule before refitting the model.





Model 2: refit the model

- •After removing extreme crime-rate outliers, the relationship between crime rate and median home value appears more stable and interpretable.
- •The residuals vs fitted plot shows no strong systematic pattern, suggesting the linearity assumption is reasonably satisfied.
- •The Q–Q plot indicates that residuals are approximately normally distributed, with minor deviations in the tails.
- •Overall, the cleaned model meets the key regression assumptions and supports valid inference.

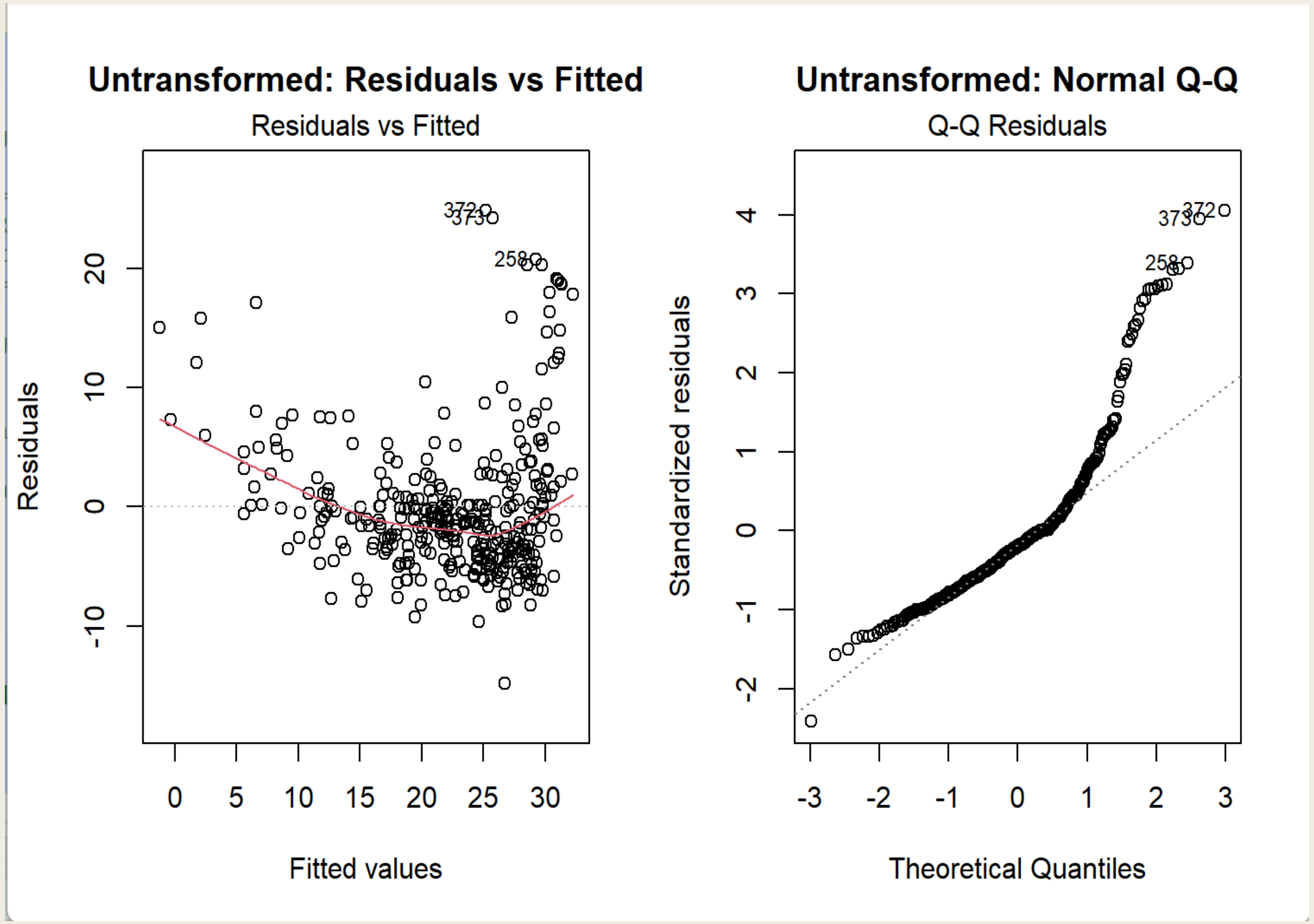
Model 2: ANOVA Interpretation

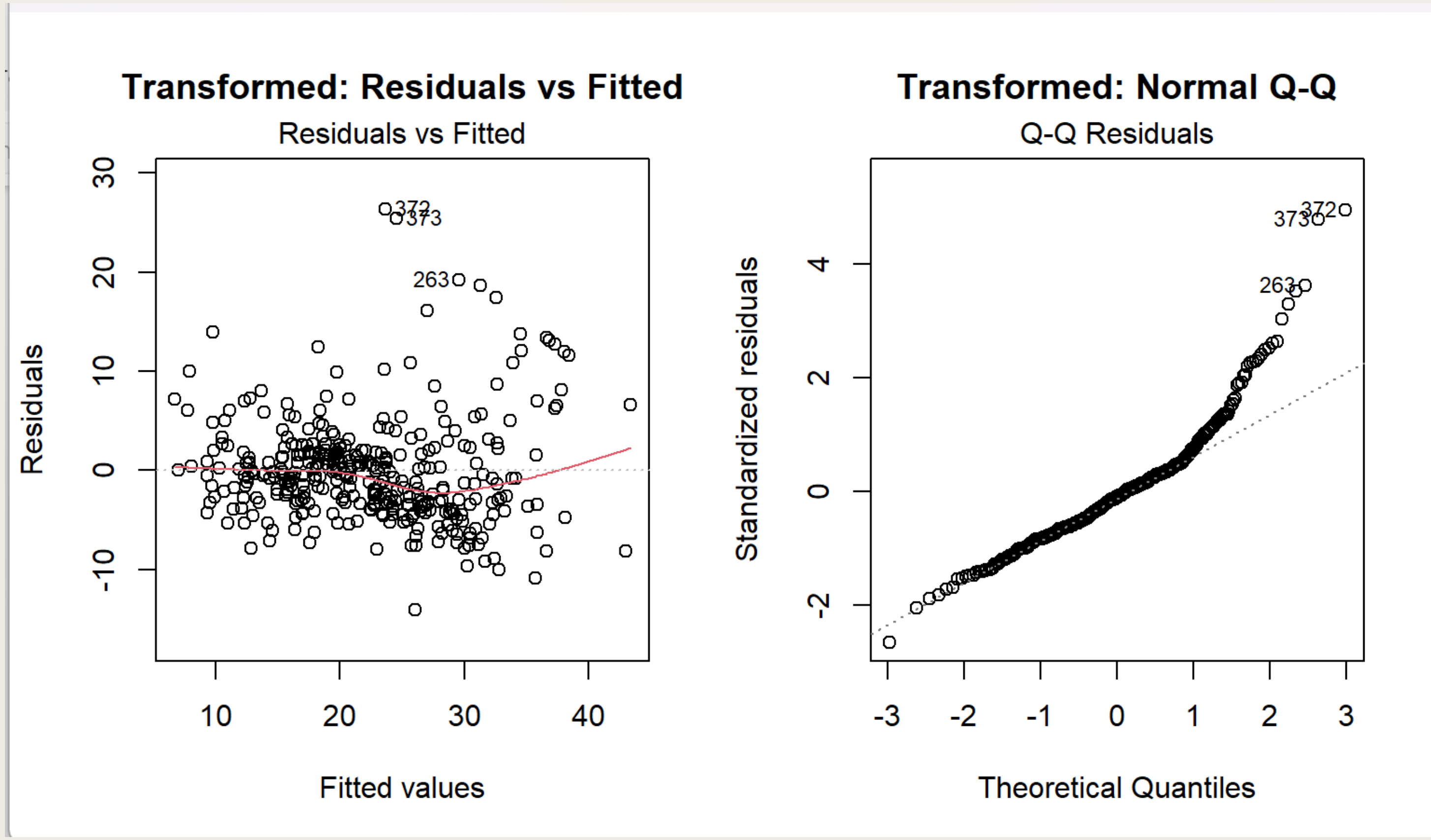
Source	Df	Sum Sq	Mean Sq	F value	P-value
rm	1	987.7	987.74	14.513	0.0001688
Residuals	301	20486.0	68.06		

- Since **p-value = 0.0001688 < 0.05**, Model 2 is statistically significant.
- Crime rate is a significant predictor of median home value after outlier removal.
- The negative slope indicates that neighborhoods with higher crime rates tend to have lower median home values.
- The test MSE summarizes the average squared prediction error on unseen data and reflects Model 2’s predictive performance.

Model 3: X_1: SLR with violated assumption

- Residuals vs fitted shows **curvature** ⇒ nonlinearity
- Spread changes across fitted values ⇒ heteroscedasticity
- Q–Q shows strong tail deviation ⇒ non-normal residuals
- **SLR assumptions violated → apply transformation**





Model 3 Transformed **medv~log(lstat)**

- Reduced curvature in residuals ⇒ improved linearity
- More even residual spread ⇒ improved variance stability
- Q–Q plot closer to line in center (heavy upper tail remains)
- Log transformation improves, but does not fully eliminate, violations

Model 3 Results (Anova Table)

- Final model: medv = $\beta_0 + \beta_1 \cdot \log(\text{lstat})$
- Training $R^2 = 0.659$ (approximately 66% of variability explained by this model)
- ANOVA F-test: F = 680.9, p = 2.73×10^{-84}
- Since p<.001, there exists a sig. effect
- The relationship between log(lstat) and medv is highly statistically significant
- **MSE = 29.06**, measuring average squared prediction error on unseen data

Source	Df	Sum Sq	Mean Sq	F value	P-value
Log(lstat)	1	19237.8	19237.8	680.9	2.2 e-16
Residuals	342	9945.3	28.3		

Comparison Table

Model	Predictor	Test_MSE
<chr>	<chr>	<dbl>
Model 3	log(lstat)	29.05586
Model 1	rm	38.16522
Model 2	crim (cleaned train)	74.71691

- **Model 3 (log(lstat))** has the **lowest Test MSE (29.06) →** best predictive performance
- **Model 1 (rm)** shows **moderate accuracy** (Test MSE = 38.17)
- **Model 2 (crim)** performs **poorly** (Test MSE = 74.72), even after cleaning