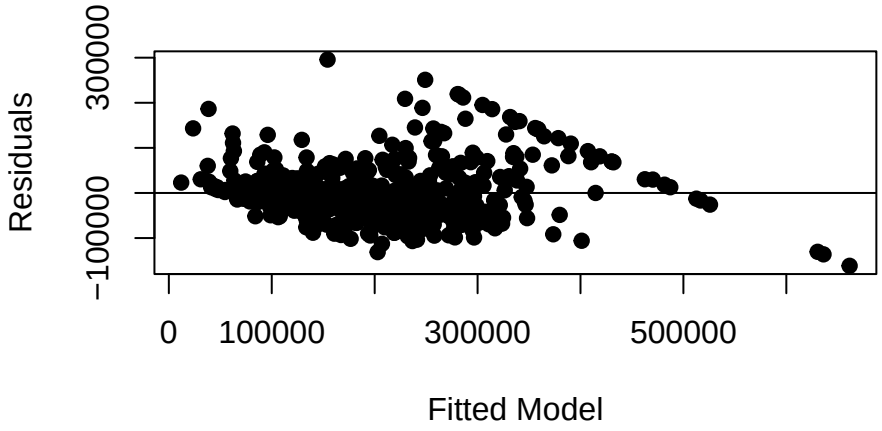
California Housing Price in 1990: Innovation

Phuc Lu Meghna Chandrasekar Sophia Li Youngju Kwon

Issues with our Data

When we use the best model from our housing data with 11 predictors and plot the residuals versus the fitted values, we see heteroscedasticity in the residual distribution, especially towards the right side of the plot.



Heteroscedasticity in the the residuals can lead to several issues in a linear model, including: Inefficient OLS estimators that do not have the smallest possible variance, incorrect standard errors of the estimated coefficients, and invalid inferences using test statistics.

Weighted Least Squares

Weighted Least Squares is appropriate to use when homoscedasticity in OLS linear regression is violated, which we see in our data.

Visually, it is clear that our residual errors are not constant. The Breusch-Pagan Test for heteroscedacity tests for it statistically.

H_0 : The variance of the residuals in our linear model is constant, indicating homoscedacity.

H_a : The variance of the residuals in our linear model is not constant, indicating heteroscedacity.

The test statistic for the Breusch-Pagan test is 41.338 with 11 degrees of freedom. The p-value for the Breusch-Pagan test for heteroskedacity is very low, at 0.00002107. Thus, we can reject the null hypothesis that the variance of the residuals is constant.

Theory for Weighted Least Squares

The method of ordinary least squares assumes that there is constant variance in the errors (homoscedasticity), and weighted least squares can be used when this assumption is violated (heteroscedasticity).

Given the linear model

$$Y = X\beta + \epsilon^*$$

The error term, ϵ^*, is assumed to have a standard normal distribution (centereda at 0) and has the nonconstant variance-covariance matrix:

$$W = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{bmatrix}$$

However, when this assumption is violated we can add weights to the variance of the residuals. The weight is defined by the reciprocal of each variance, $\hat{\sigma_i}^2$

$$w_i = \frac{1}{\hat{\sigma_i}^2}$$

Let matrix W be a diagonal matrix containing these weights.

$$W = \begin{bmatrix} w_1 & 0 & ... & 0 \\ 0 & w_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & w_n \end{bmatrix}$$

The weighted least squares estimate is

$$\hat{\beta}_{WLS} = \text{argmin} \sum_i^n \epsilon_i^* 2$$

$$= (X^T W X)^{-1} X^T W Y$$

Each weight is inversely proportional to the error variance, so an observation with a large error variance will have a smaller weight, and vice versa. Weighted least squares is useful to bring proportionality when some observations and correspodingly their variances are much larger than others.

Implementing Weighted Least Squares

Since heteroscedasticity is present, we can use Weighted Least Squares to improve our model.

Weights can be calculated based off either the inverse of the residuals or the fitted values.

Comparing Models

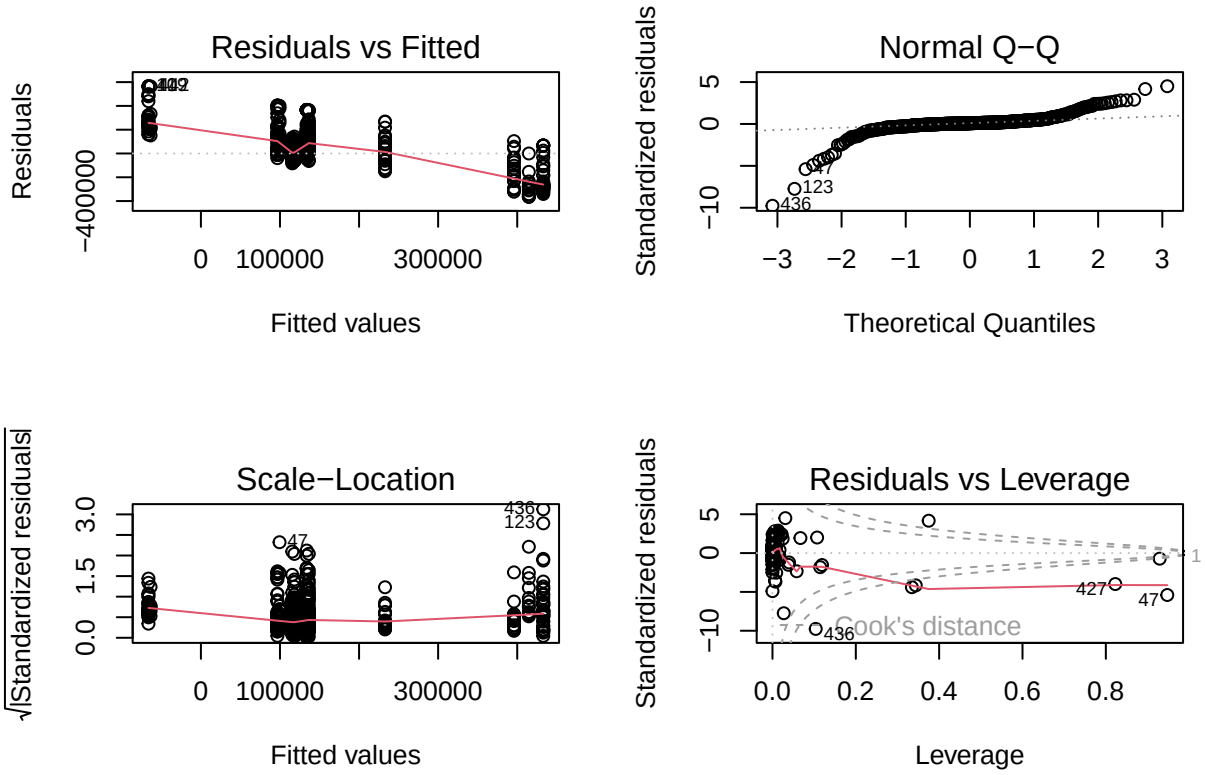
When we perform weighted least squares by weighted the residuals, the R^2 increases from 0.696451 to 0.9726604.

When we perform weighted least squares by weighted the fitted values, the R^2 decreases from 0.696451 to 0.5754741.

This suggests that the weighted least squares by residuals significantly increases the R^2 value, indicating that our model's explanatory power of the variance improved.

However, the AIC (Aikake Information criterion) in the weighted least squared model by residuals is 13584.62 vs 12008.51 in the OLS model, and the BIC (Bayesian Information Criterion) is 13613.84 in the WLS versus 12062.76 in the OLS model, were lower values indicate a better model.

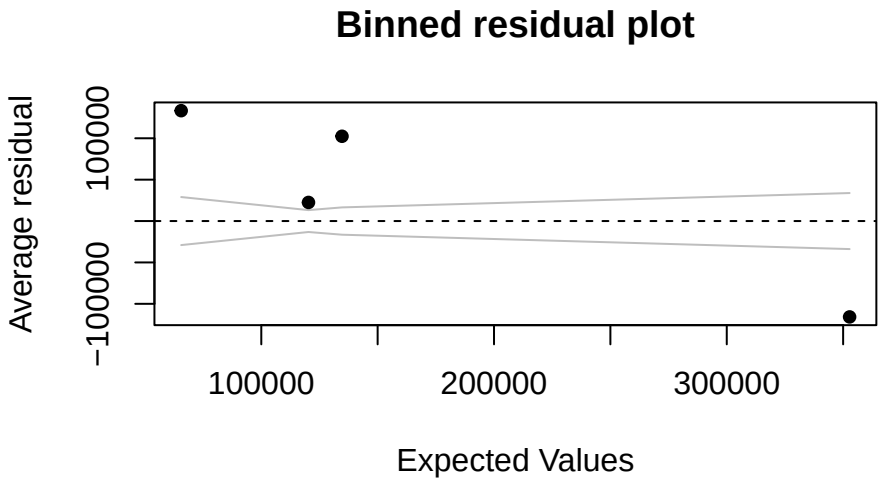
```
## Warning: not plotting observations with leverage one:
##   14
```



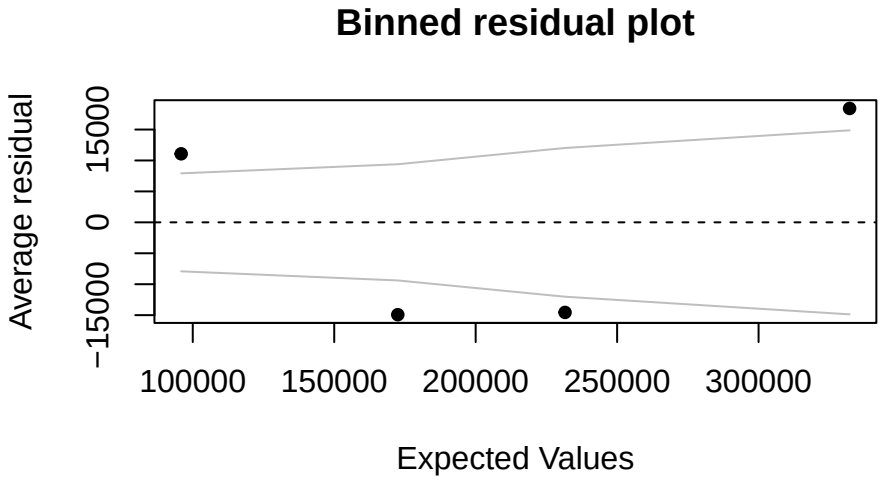
The residuals vs fitted plot in the WLS model shows that the data is fitted into clusters. This is likely because the predictors are organized into categorical variables. All clusters have both positive and negative values on the residuals axis.

Binned residual plot

A binned plot groups the data into bins and plots the average residuals within each bin. This may show a more representative view of the nature of the residuals variance. The number of bins chosen is 4.



In the weighted least squared model by residuals, all of the binned residuals fall outside the 95% CI bounds, indicating that the residuals are not evenly distributed.



In the unaltered model, the residuals also fall out of the 95% CI interval.

```
##
##  studentized Breusch-Pagan test
##
## data:  wls_model_residuals
## BP = 237.68, df = 5, p-value < 0.00000000000000022
```

The bp test fails for the new model as well. Overall, the cross validation failed in WLS method improving our linear model, and further methods should be implemented to fix the heteroskedasticity in variances.