Section 2

2024-04-26

Regression Analysis

The data set was split into a training and test set, so that we could perform valid inferences about the the relationship of certain predictors with the sales price of homes. We also ensured that the neighborhoods in the data set were appropriately identified as categorical predictors. We then began the process of variable selection by using the best subset method using both AIC and adjusted R2 for variable selection. The choice of the following selection procedure was due to the fact that its an exhaustive method that can check all possible models. The two models chosen from the search algorithm were then assessed using the rmseloocv, which gives us an estimate of the test rmse, to determine the best fitting model using this method.

Table 1: RMSEloocv for Quality Criterion

Criterion	RMSEloocv
AIC	12417.59
Adjusted R^2	12465.33
Hypothesized Model	19399.57

The chosen model was the AIC model given its smaller RMSEloocv.

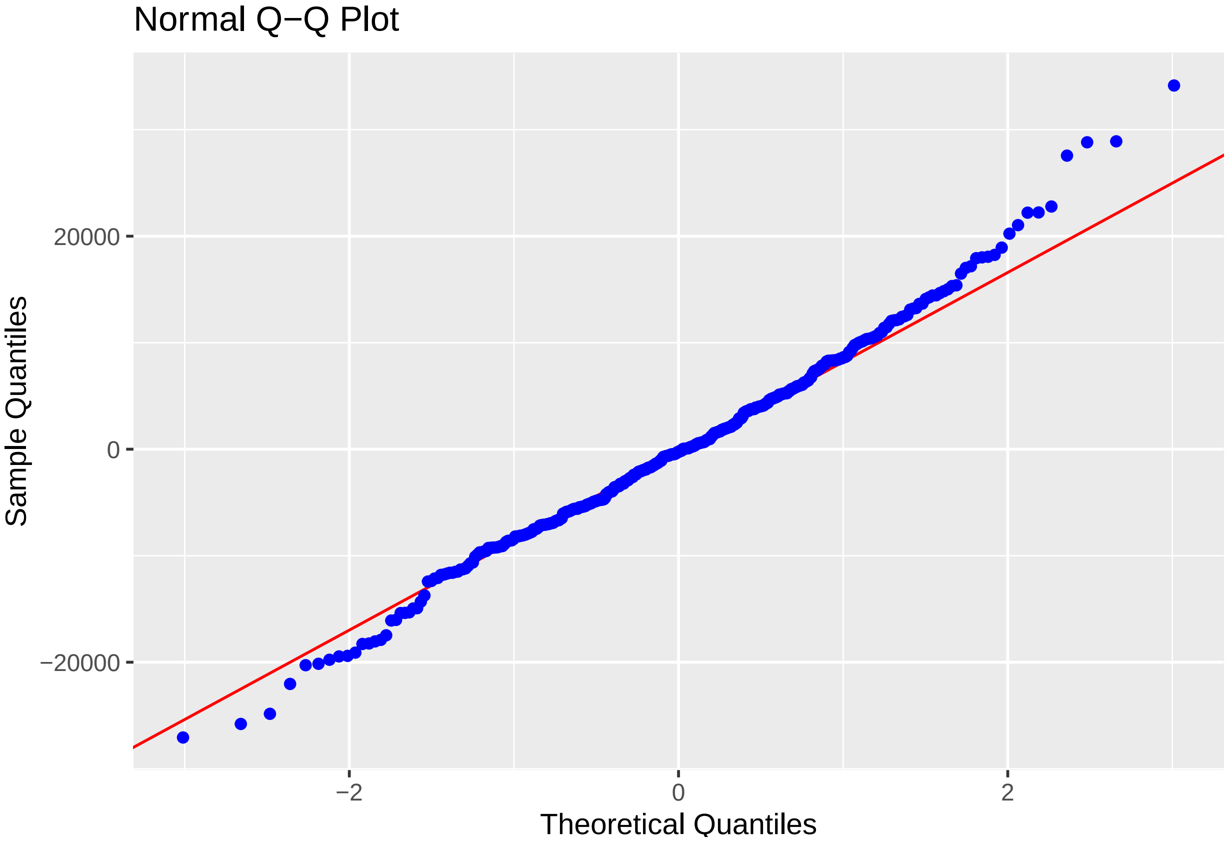
Once chosen, the model was fit to the test data for further analysis. We first checked whether their were issues with collinearity since this would decrease the power of our hypothesis test. The condition number was 910, which seems to be due to the relationship between the categorical predictors. The VIFs, as shown in the table below, were all below 5, except for NeighborhoodOldTown, which again likely has to do with the relationship between the categorical predictors. We should be cautious about the large condition number, but since the VIFs are relatively low, we're not losing to much power relative to the uncorrelated case.

Table 2: VIF Values for Regression Variables

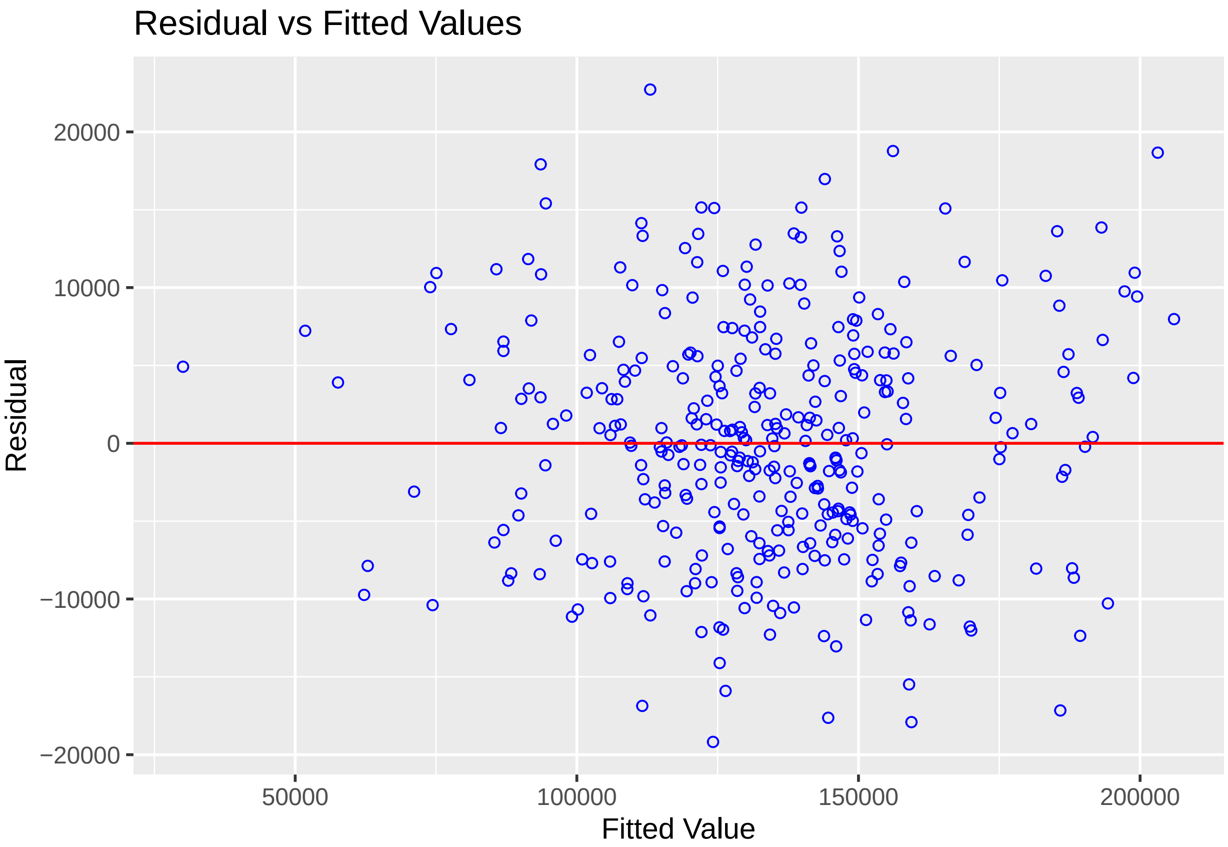
Variable	VIF
LotArea	1.300574
NeighborhoodEdwards	2.649184
NeighborhoodNAmes	3.966050
NeighborhoodOldTown	5.012429
NeighborhoodSawyer	2.275655
OverallQual	1.993924
OverallCond	1.701585
YearBuilt	3.572031
YearRemodAdd	1.845922
BsmtFinSF1	1.913897
TotalBsmtSF	1.882527
GrLivArea	3.111682
BsmtFullBath	1.679272
BsmtHalfBath	1.112376
FullBath	1.772713

Variable	VIF
BedroomAbvGr	1.996536
KitchenAbvGr	1.570560
TotRmsAbvGrd	2.874547
Fireplaces	1.344416
GarageCars	1.335036
WoodDeckSF	1.170969
OpenPorchSF	1.164659
ScreenPorch	1.146769

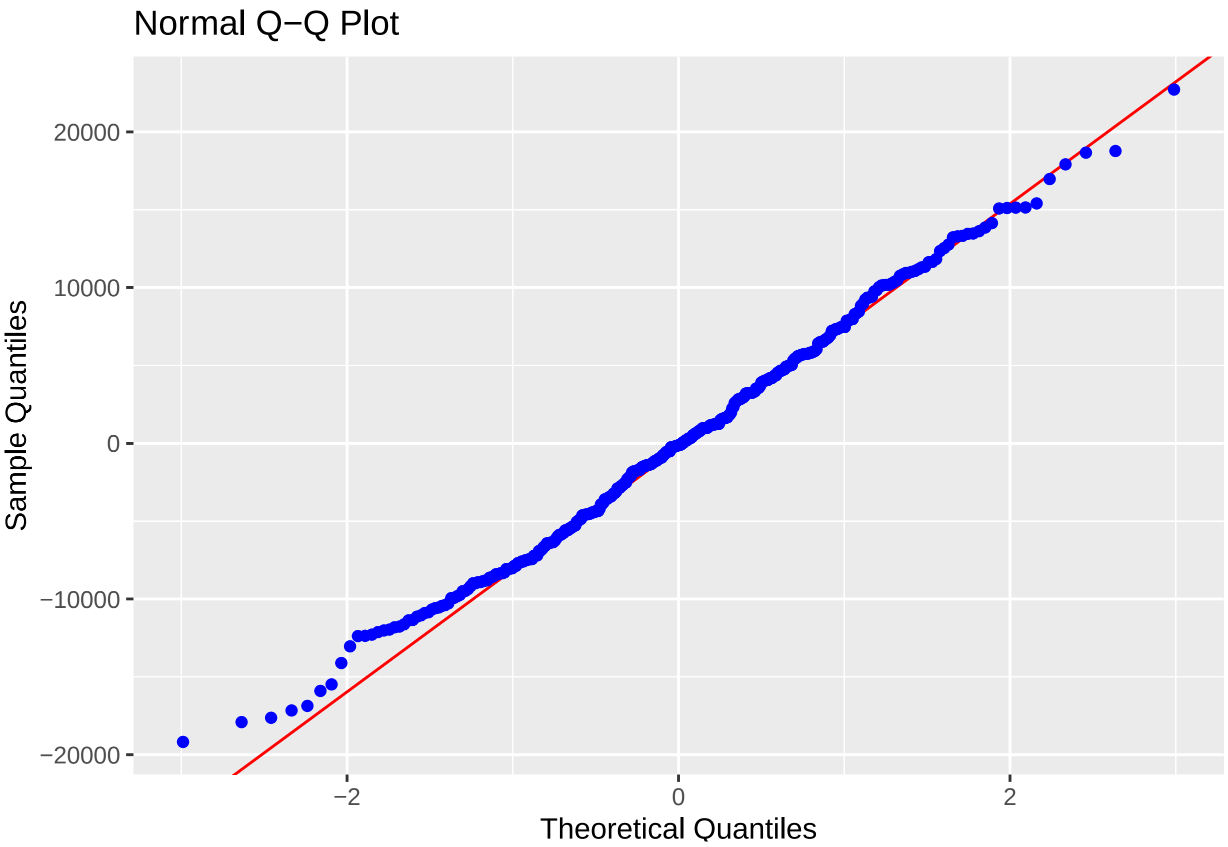
The LINE assumptions were checked using the following plots and tests to determine if there were any violations. Our first analysis before checking for any highly influential observations showed that only the normality assumption was violated. The Shapiro-wilks test had a p-value of .03, leading us to reject the null hypothesis at the 5% significance level and conclude that the errors did not follow a normal distribution. The Q-Q plot also showed moderately discrepancies from the line at the tails as seen in the plot below.



The highly influential observations were then removed from the model and the model assumptions were checked again. We concluded that the linearity assumption wasn't violated by checking the fitted versus residual plot, which showed that the residuals were centered around zero. The spread of the residuals remains constant, which suggest that the constant variance assumption holds.



The Breush-Pagan test was used to validate what was suggested by the fitted versus residual plot. The reported p-value from the BP-test was .465, so we fail to reject the null hypothesis at any reasonable significance level and conclude that the errors are homoscedastic. The Q-Q plot was then checked after removing the highly influential points giving us the following plot. Observe that the tails closely follow the 45 degree line, suggesting normality is not violated. The p-value from the shapiro-wilk test was .35, so we again fail to reject the null hypothesis at any reasonable significance level and conclude that the errors follow a normal distribution.



The final model after checking for collinearity, removing highly influential observations, and verifying the LINE assumptions:

$$
\begin{aligned}
\text{SalePrice} = & -1,140,328 + 1.24\text{LotArea} - 6,632.58\text{NeighborhoodEdward} \\
+76.93\text{NeighborhoodOldTown} + 3,675.31\text{NeighborhoodSawyer} & \\
& + 6,755.23 \times \text{OverallQual} \\
& + 4,651.67 \times \text{OverallCond} \\
& + 452.35 \times \text{YearBuilt} \\
& + 121.07 \times \text{YearRemodAdd} \\
& + 5.04 \times \text{BsmtFinSF1} \\
& + 12.90 \times \text{TotalBsmtSF} \\
& + 36.51 \times \text{GrLivArea} \\
& + 4,320.65 \times \text{BsmtFullBath} \\
& + 2,167.45 \times \text{BsmtHalfBath} \\
& + 5,622.58 \times \text{FullBath} \\
& - 1,248.01 \times \text{BedroomAbvGr} \\
& - 7,163.52 \times \text{KitchenAbvGr} \\
& + 1,867.20 \times \text{TotRmsAbvGrd} \\
& + 4,221.25 \times \text{Fireplaces} \\
& + 6,027.75 \times \text{GarageCars} \\
& + 13.93 \times \text{WoodDeckSF} \\
& - 0.88 \times \text{OpenPorchSF} \\
& + 1.67 \times \text{ScreenPorch}
\end{aligned}
$$

#Code Appendix