Project - Regression Analysis of Properties for Sale

Elijah Castro

December 07, 2022

Summary

Data containing a random sample of 83 properties for sale in a city is here analyzed using multiple linear regression techniques to investigate how the listed price of the property depends on the remaining variables provided.

Introduction

Data of the first few listed properties are shown below along with a description of each variable included in this data set.

Table 1: Data of First Five Properties For Sale

size	age	dc	dt	price
102.2	4.1	0.8	50.2	472
102.7	16	5.5	51.4	660.1
101.1	6.8	14.7	29.1	683
121.2	18	6.5	66.2	473.2
102.9	17	10.4	43.8	593

Variable Description:

1. *size*: size of the property (in square meters).
2. *age*: age of the property (in years).
3. *dc*: distance (in km) from the property to the city center.
4. *dt*: distance (in km) from the property to a toxic waste disposal site.
5. *price*: the listed price of the property, in thousands of dollars.

Our objective here is to produce a framework that accurately describes the relationship between the independent variables, size, age, distance to city center, and distance to toxic waste disposal site, and the response, price, for the properties in this city. We will achieve this by uncovering the regression model that best fits the data provided. Some of the methods employed to complete this objective include exploratory data analysis, model selection, diagnostics, and when relevant, applying transformations and identifying large leverages, outliers, and influential points.

Procedure

To begin, a few preliminaries are explored. This includes exploratory analysis of the response, price, against all other individual variables, fitting an initial regression model with all predictors, and exploring diagnostics. This is used as a starting point to see what can be improved.

Next, a stepwise procedure, specifically backward elimination using AIC criterion, is used in order to create the final model. For this analysis, backwards selection is preferred as forward approaches often allow for important variables to be missed due to other variables being entered into the model first.

After fitting the final model, diagnostics are again explored. When necessary, a transformation will be applied to improve the fit of the model. Lastly, we check for potential influential points in the final model. If any exist, findings will be reported with and without the influential observation present in the data set.

Note: To view the entire process of finding the final model, click on the link below:

In-Depth Approach to Find Best Model

Findings/Analysis

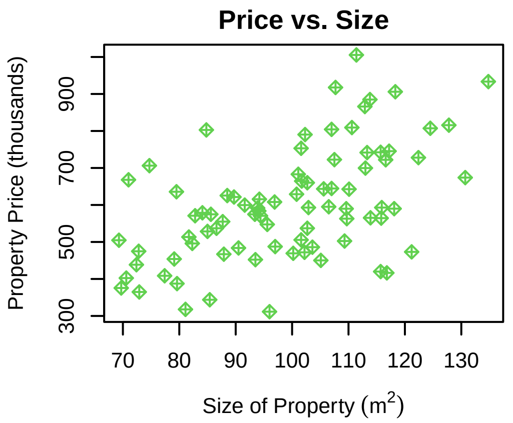
Here, findings and analysis from the final model are discussed. In the final model, the following changes were made from the initial model:

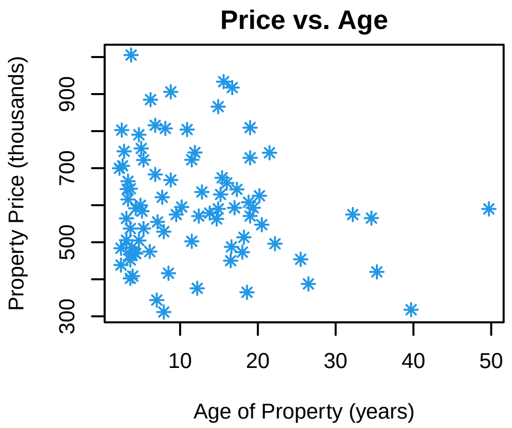
- Removed variable *dc*, distance from the property to the city center
- Added a higher order transformation of the *age* variable
- Identified an influential point, so findings/analyses will be reported with and without this point present

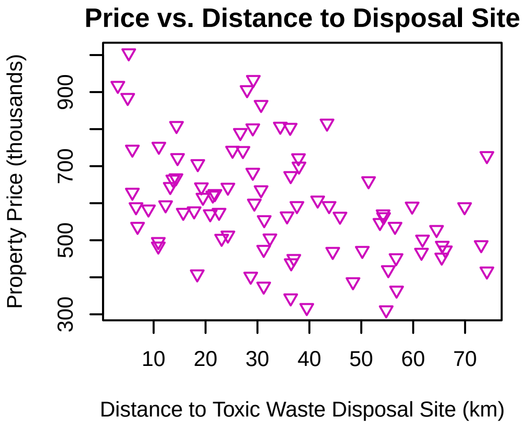
Influential Point Included

EDA

First, we display the exploratory data of the price of the property against each individual variable included in the final model:







For `size`, we notice there is a strong positive linear relationship with `price`. That is, as the size of the property increases, the price also increases, on average. For `age`, we find there is relatively mild negative relationship with `price`. That is, as the age of the property increases, the price decreases somewhat, on average. Lastly, for `dt`, we notice there is a relatively strong negative linear relationship with `price`. That is, as the distance from the property to the toxic waste disposal site increases, the price of the property decreases, on average.

Final Model

Next, we display the summary statistics of this final model:

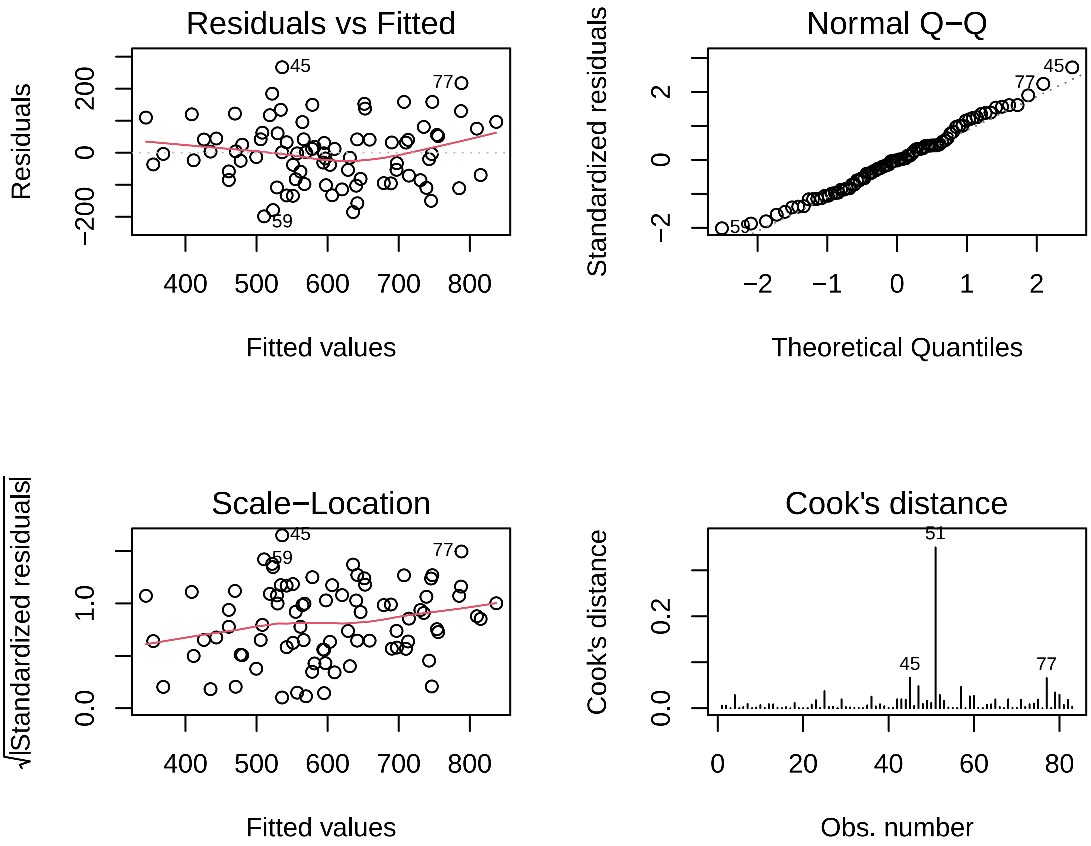
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	161.5	72.15	2.238	0.02805
size	5.683	0.706	8.049	7.56e-12
age	4.739	3.37	1.406	0.1636
I(age^2)	-0.1876	0.08229	-2.28	0.02534
dt	-4.053	0.5937	-6.827	1.678e-09

Table 3: Fitting linear model: price ~ size + age + I(ageˆ2) + dt

Observations	Residual Std. Error	R^2	Adjusted R^2
83	100.2	0.5804	0.5589

We notice including age^2 is significant at a 5% level, but similar to the initial model, the original variable age is still not significant. From the R^2 value, we also note that close to 58% of the variation in the response can be explained by our independent variables in this final model.

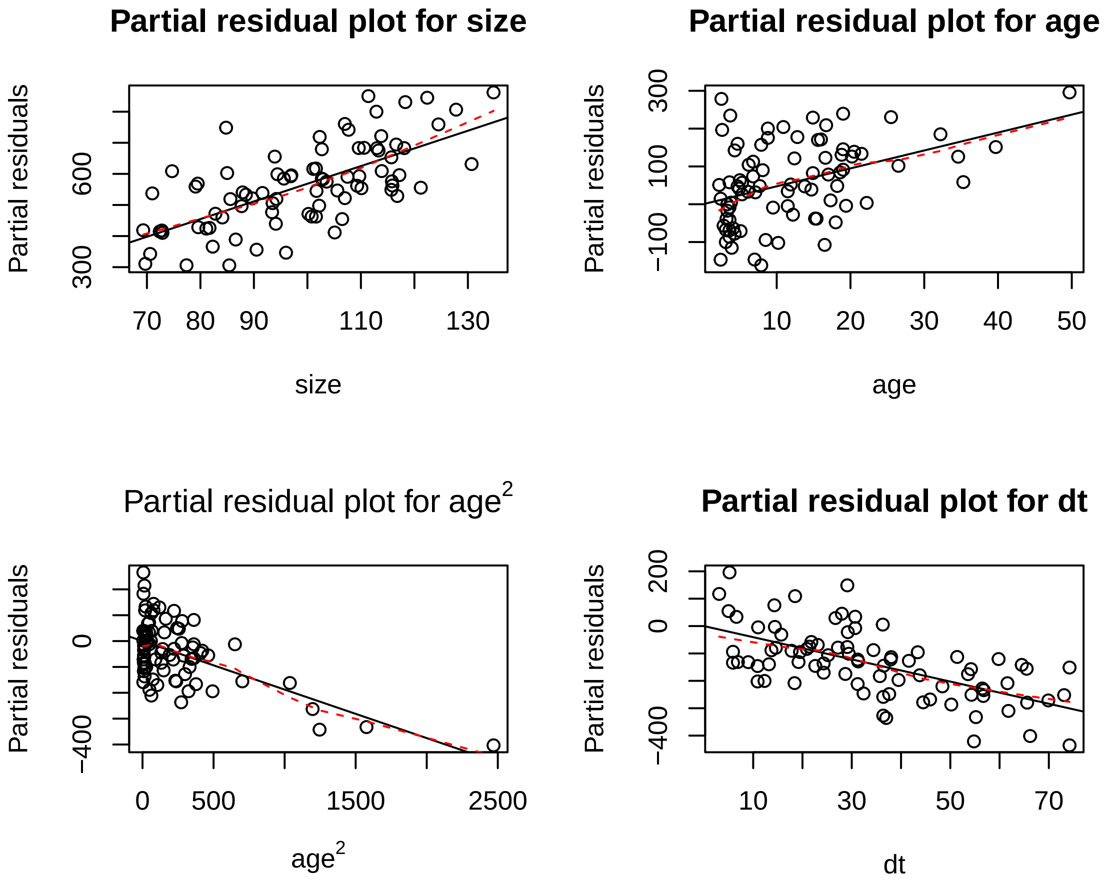
Diagnostics



From the diagnostic plots, we determine that the constant variance and normality assumptions are met since the residuals vs. fitted values and scale-location plots show random scatter (homoscedasticity) and the points in the Q-Q plot follow a straight line, respectively. Looking at the plot of Cook's distance, we will demonstrate our proof in the next section that observation 51 is an influential point, as its rather large distance value peaked our interest.

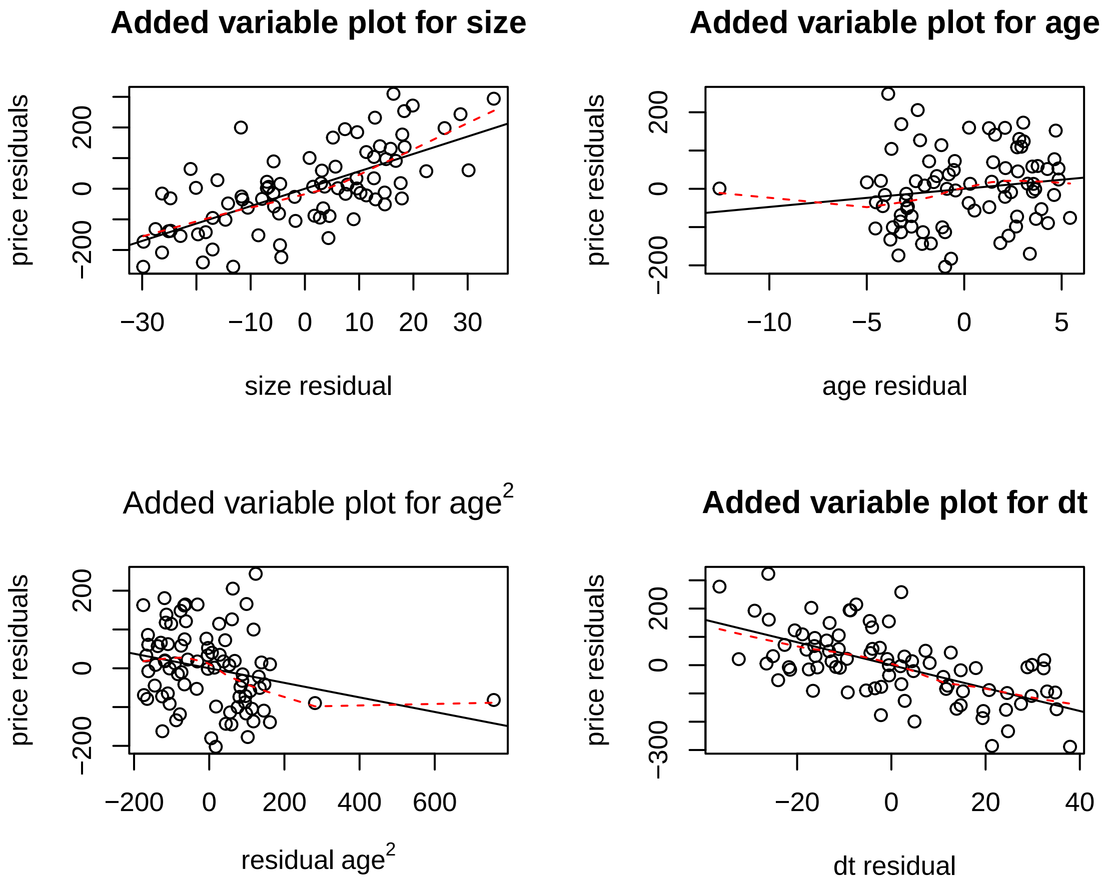
We also confirm the linearity condition by constructing added variable and partial residual plots. Added variable plots inspect any correlation between our predictors and response and partial residual plots help determine if our model is correct.

Partial Residual Plots



For the partial residual plots, there should be a straight line if the model is correct. A nonlinear pattern suggests we may need a higher order term or a transformation. For the partial residual plots of *size*, and *dt*, there is almost a perfect straight line. We also note by adding a higher order term for `age`, we fix the nonlinear pattern from the initial model. Therefore, we can now determine this model is appropriate.

Added Variable Plots

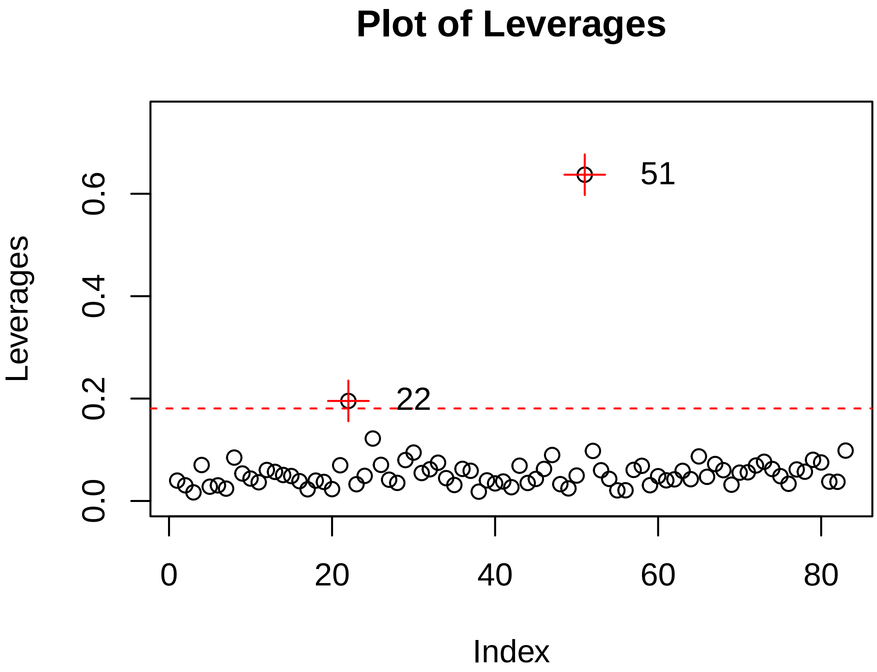


An added-variable plot is a effective way to show the correlation between our independent variables and response conditional on other independent variables. A strong linear relationship in the added variable plot indicates the increased importance of the contribution of the regressor to the model already containing the other predictors.

For the predictors *size* and *dt*, there is strong evidence of a strong linear relationship in the added variable plot. Thus, these variables add strong contribution to the model containing all the other predictors. However, we notice *age* and the higher order term of *age* still do not necessarily meet the linearity assumption. It is possible we can attribute this to the observation residual that is very far from the other residuals in the plot. It will be shown that this issue is from the influential point we found in the model, and when we report the results with the influential observation removed, the changes are drastic.

Influential Point

As mentioned previously, we will now demonstrate our proof observation 51 is an influential point.



Looking at the plot of leverages, we notice observations 51 and 22 are large leverage points since they are above the red threshold line, which indicates three times the mean leverage value.

We also find there are no outliers present. Therefore, we only examine if observations 22 and 51, which are large leverage points, are influential points. The rule of thumb to determine if an observation is influential is if the observation has a distance (using Cook's distance) greater than 4 divided by the total number of observations.

Using this criteria, we find 3 observations that meet the condition of being an influential point. However, we only check for influential points from large leverage points and outliers. Therefore, observation 51, a large leverage point, is the only influential point here.

Influential Point NOT Included

Now, since an influential point was identified, the next step is to remove this observation from the data set, and compare the results.

Examining Influential Observation

When examining the data from this observation, we note that the age of this property is almost 50 years old.

Table 4: Data of Influential Point

size	age	dc	dt	price
109.6	49.7	9.1	6.6	589.9

From the previous scatterplot of *price vs. age*, this observation is the isolated point and has the largest value in the *age* variable. This is evidence for why this observation had such large leverage, and most likely, why it contributed to problems in some of the diagnostics.

Final Model

Below, we display the summary statistics of this final model:

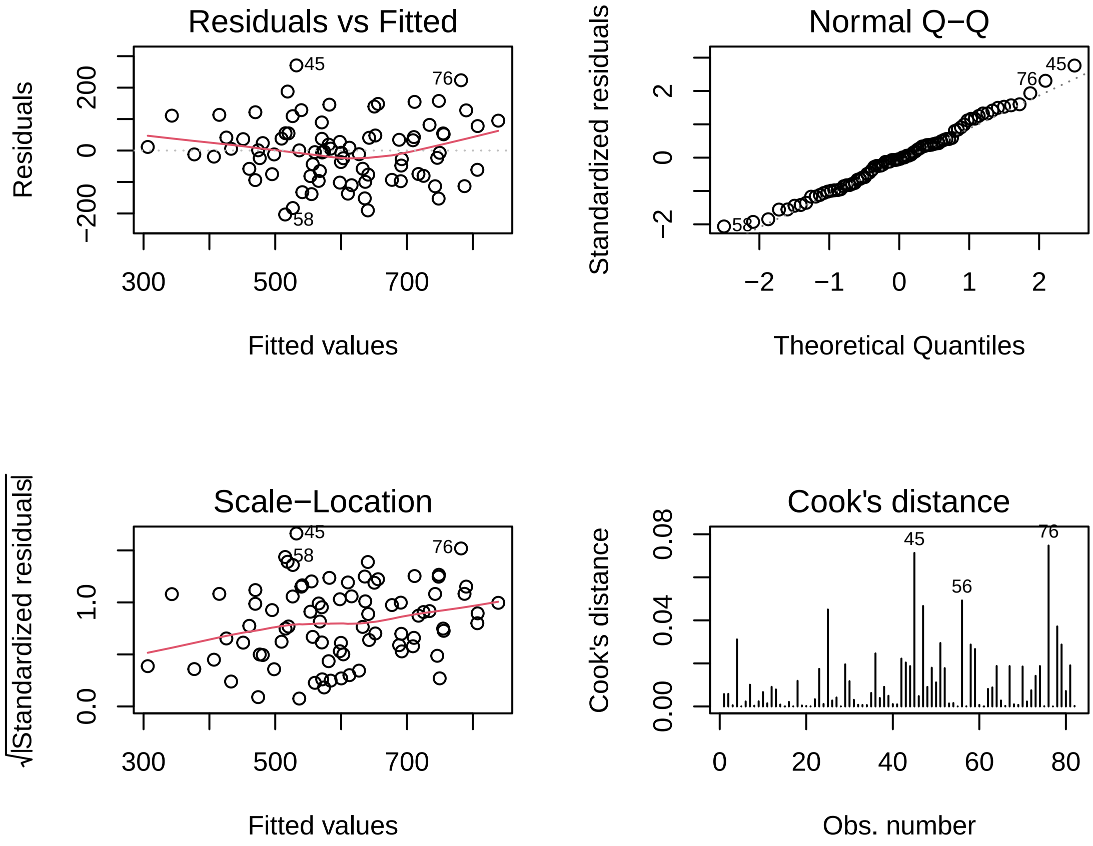
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	159.9	72.17	2.215	0.0297
size	5.569	0.7152	7.786	2.619e-11
age	7.096	4.115	1.724	0.08866
I(age^2)	-0.2724	0.1183	-2.303	0.02396
dt	-3.988	0.5973	-6.677	3.377e-09

Table 6: Fitting linear model: price ~ size + age + I(age^2) + dt

Observations	Residual Std. Error	R^2	Adjusted R^2
82	100.2	0.5858	0.5642

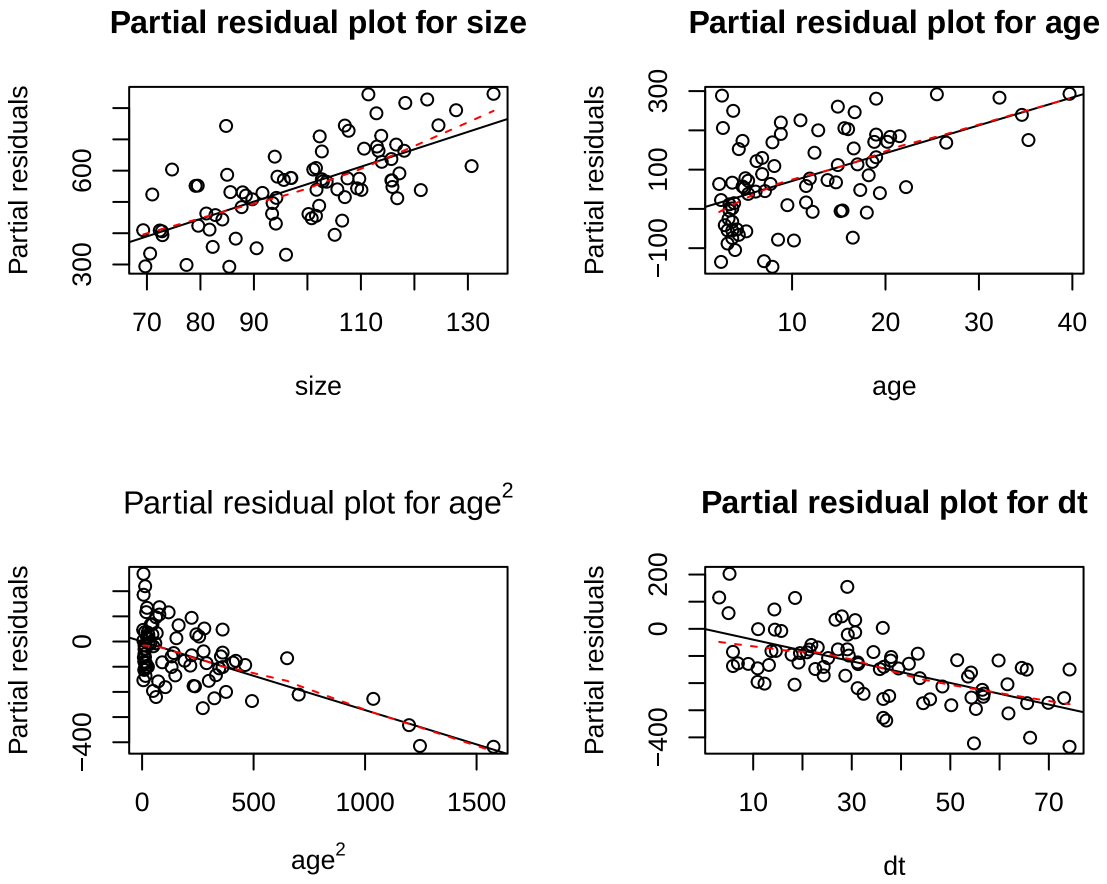
We notice including age^2 is still significant at a 5% level, and the original `age` variable is still not significant at the 5% level, although the p-value did decrease. The value of R^2 did slightly increase with the removal of this influential point.

Diagnostics



Looking at the diagnostic plots above, constant variance and normality assumptions are still met. Looking at the plot of Cook's distance, there are a few observations that stand out for investigation. However, the range of values displayed in the y-axis is much smaller, so it is less likely that these points have as much influence as before.

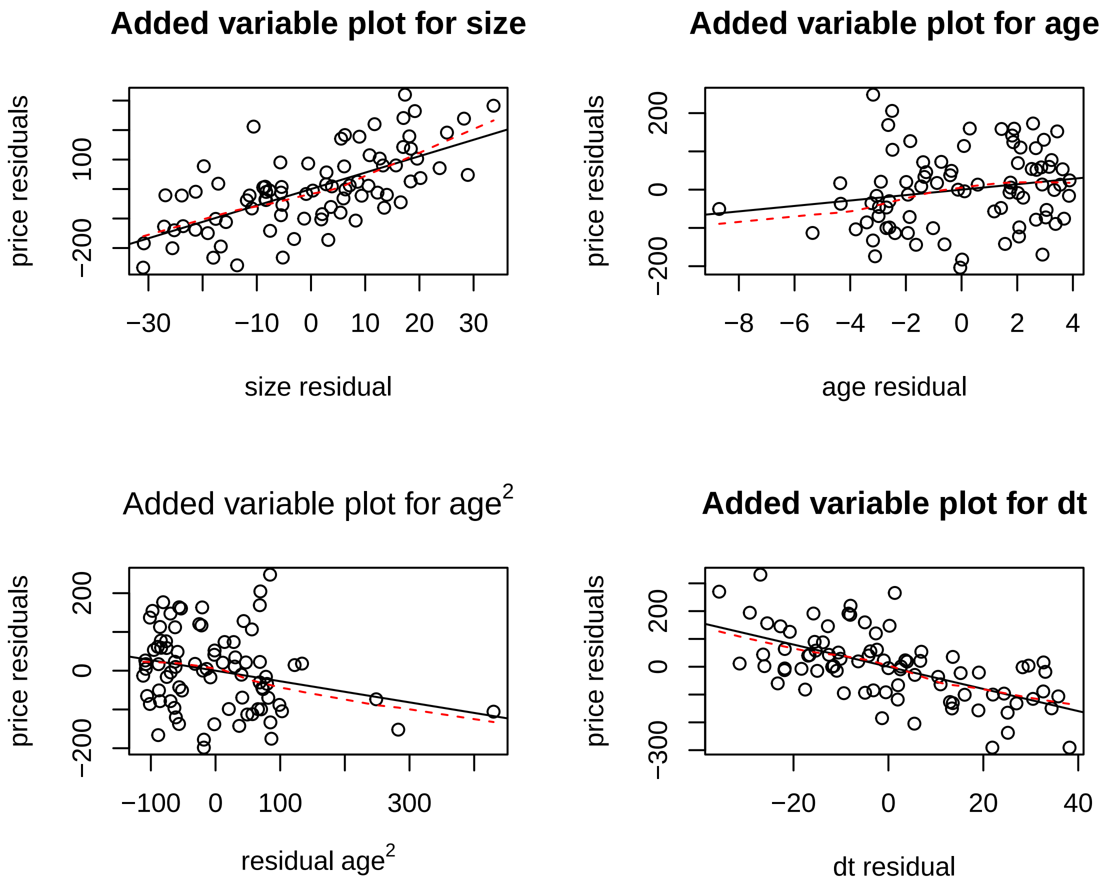
Partial Residual Plots



Again, we note the model is appropriate as there are almost perfect straight lines in each partial residual plot after removing the influential observation.

Added Variable Plots

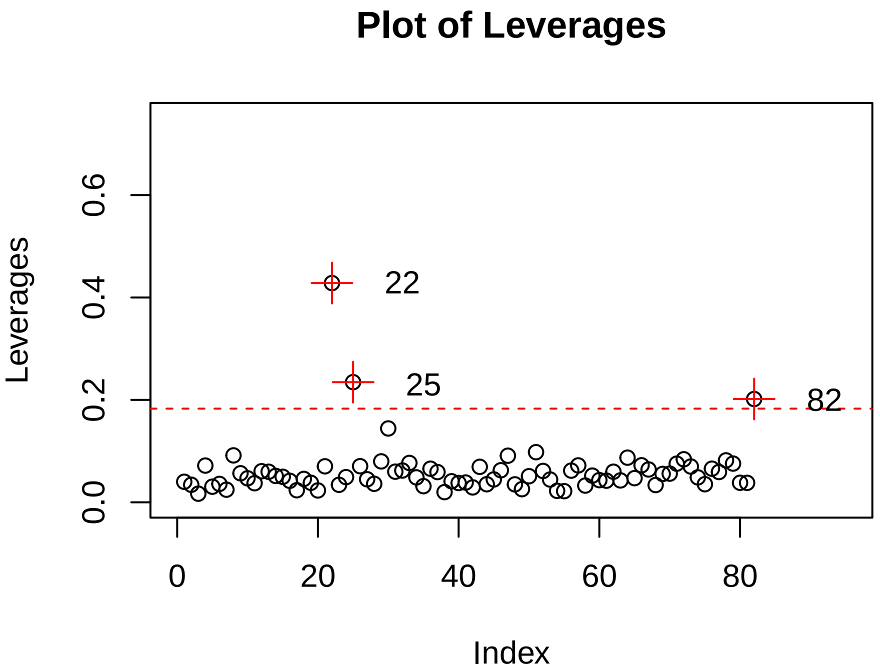
Lastly, we inspect if removing the influential observation fixes the linearity assumption in the *age* and higher order term *age* variables.



In fact, this influential point did have a large effect on the linearity. Now, all added variable plots meet the linearity assumption quite well.

Potential New Influential Points?

After removing the previous influential observation, we still need to ensure there are not any new observations that can be classified as influential points.



Looking at the plot of leverages, we see observations 22, 25, and 82 are large leverage points since they are above the red threshold line, which again, indicates three times the mean leverage value.

We again find there are no outliers present. Therefore, we only examine if observations 22, 25, and 82, large leverage points, are also influential points. We use the same criteria as before to determine if an observation is influential: the observation has a distance (using Cook's distance) greater than 4 divided by the total number of observations.

We find 3 observations meeting the condition of being an influential point. However, none of the observations 22, 25, and 82, are among the values considered. Hence, none of these observations meet the requirements of being an influential point. Therefore, removing the previous influential observation results in no new influential points in this model.

Conclusions

While our findings and analysis include a strong model that represents the given data well, there are limitations because of the size of the data. Since we were provided with only a random sample of 83 observations, it may be difficult to accurately replicate the relationship between property prices against the same independent variables on new data using the same model we found. Similarly, this then means predictions may not be the most precise on new data since the provided data set was so small. Therefore, if we were to extend this research further, we would have to proceed with our best model with caution.

Additionally, to produce an even more powerful and accurate model, it would have been ideal to have a much higher amount of observations, so that we could split the data into training and test data. This would have prevented the potential of overfitting, which unfortunately, is less feasible for our model given the limitations presented.