PSTAT 220A Project 2 Scratch Work

Elijah Castro

December 06, 2022

In a good data analysis, one usually needs to go through several exploratory "iterations" before reaching at the final results. In the process, tentative models should be evaluated and reevaluated by both statistical analytical tools and by common sense. In the presentation of the final results, however, one should avoid tedious reporting, but instead focus on the important findings. When appropriate, do use plots in your exploration, and do include good ones in your presentation.

The data set *property.txt* in the same folder contains a random sample of 83 properties for sale in a city. It contains 5 variables:

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.

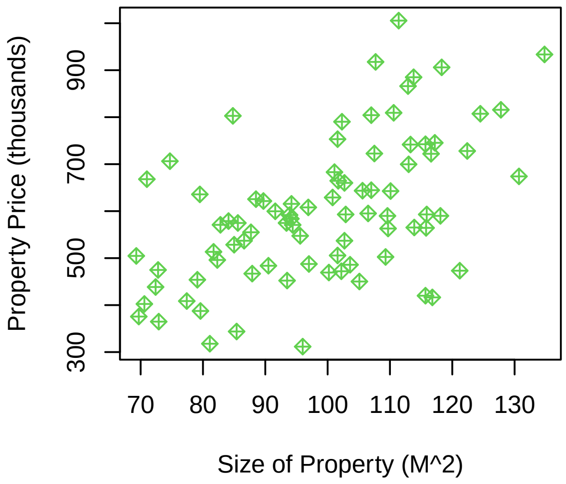
Investigate how listed price depends on other variables.

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
94.4	12.4	15.1	54.2	570.6

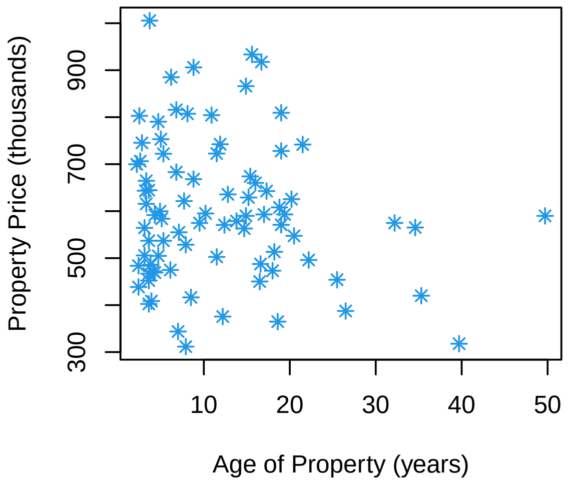
Scratch Work

First, we will plot the response (`price`) vs. each independent variable

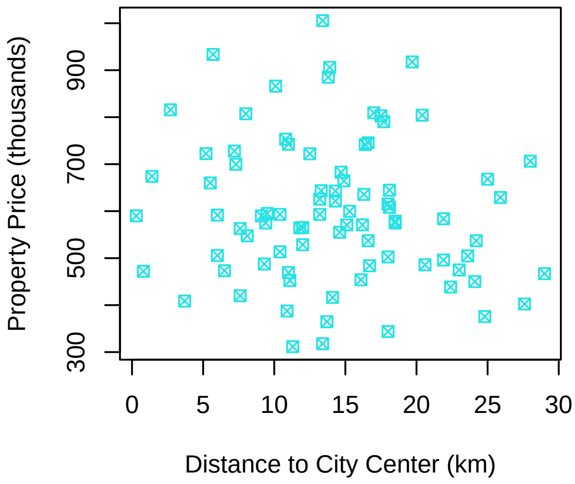
Price vs. Size



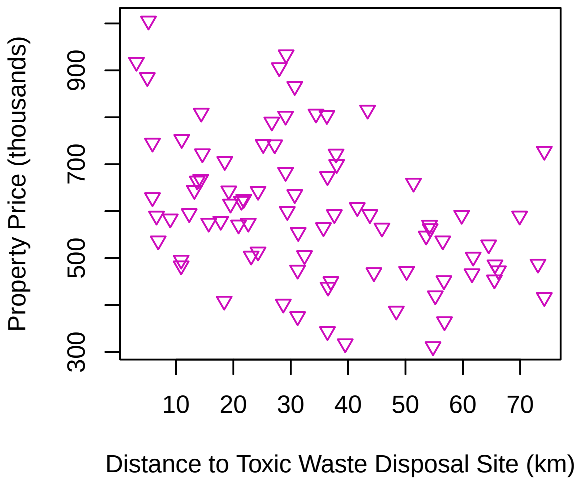
Price vs. Age



Price vs. Distance to City Center



Price vs. Distance to Disposal Site



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 notice there is relatively mild negative relationship with `price`. That is, as the age of the property increases, the price decreases somewhat, on average. For `dc`, there is almost no positive or negative linear relationship with `price`. This will require further investigation later. 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.

Now, we will fit the model on all independent variables and display summary statistics.

```
fit_property <- lm(price ~ ., property)
pander(summary(fit_property))
```

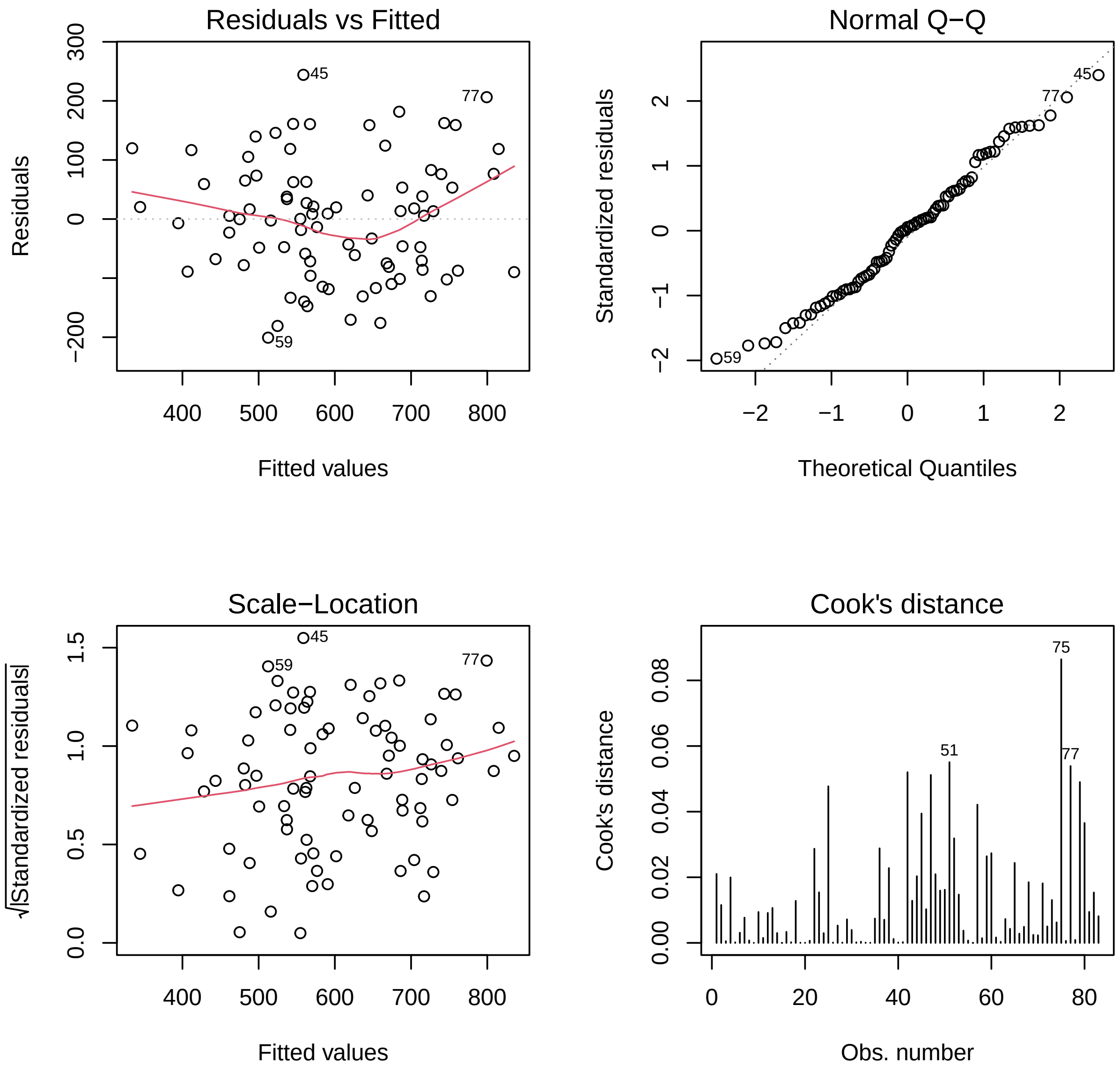
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	158.9	111.9	1.42	0.1597
size	5.918	0.8531	6.937	1.039e-09
age	-2.38	1.246	-1.91	0.05977
dc	0.6866	2.207	0.3112	0.7565
dt	-3.714	0.6255	-5.938	7.581e-08

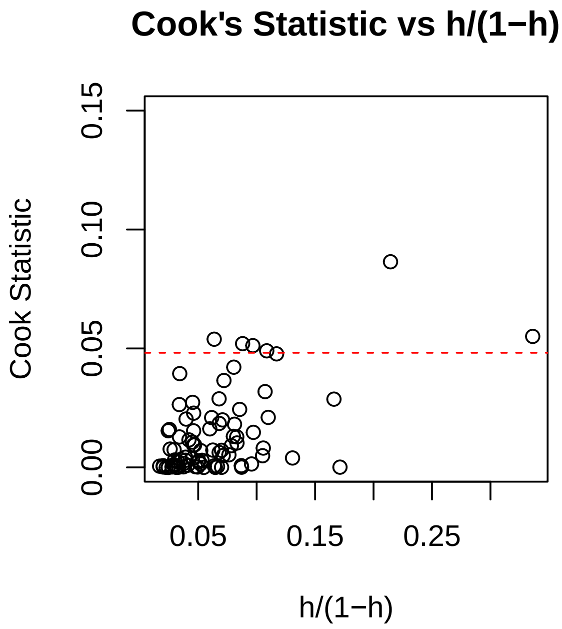
Table 3: Fitting linear model: price ~ .

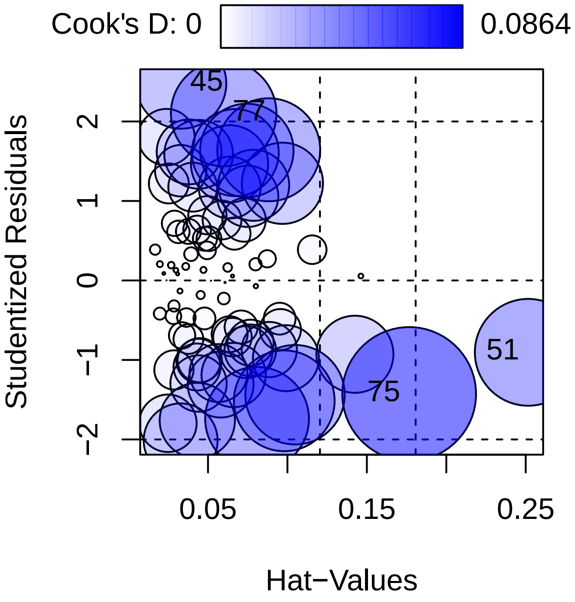
Observations	Residual Std. Error	R^2	Adjusted R^2
83	103.5	0.553	0.5301

We notice at a 5% significance level, both `age` and `dc` variables are not significant. This will require further investigation to determine if these predictors should still be included in the model.

Let’s also check if the model fits the data well by running diagnostic plots.





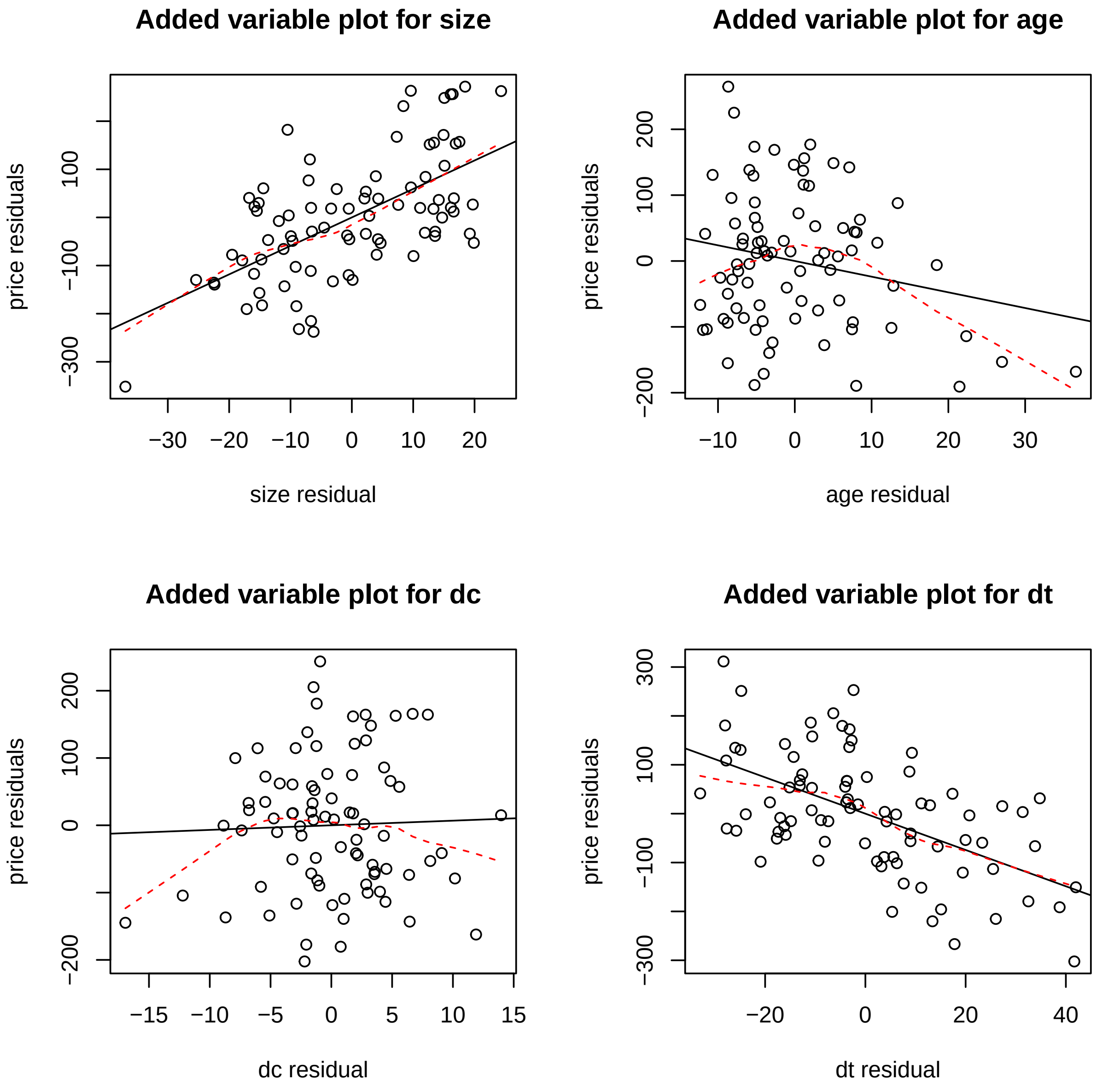


```
##    StudRes   Hat CookD
## 45     2.5 0.033 0.039
## 51    -0.9 0.252 0.055
## 75    -1.4 0.177 0.086
## 77     2.1 0.060 0.054
```

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. We will investigate for influential points and outliers and how they impact our model once we determine our best model.

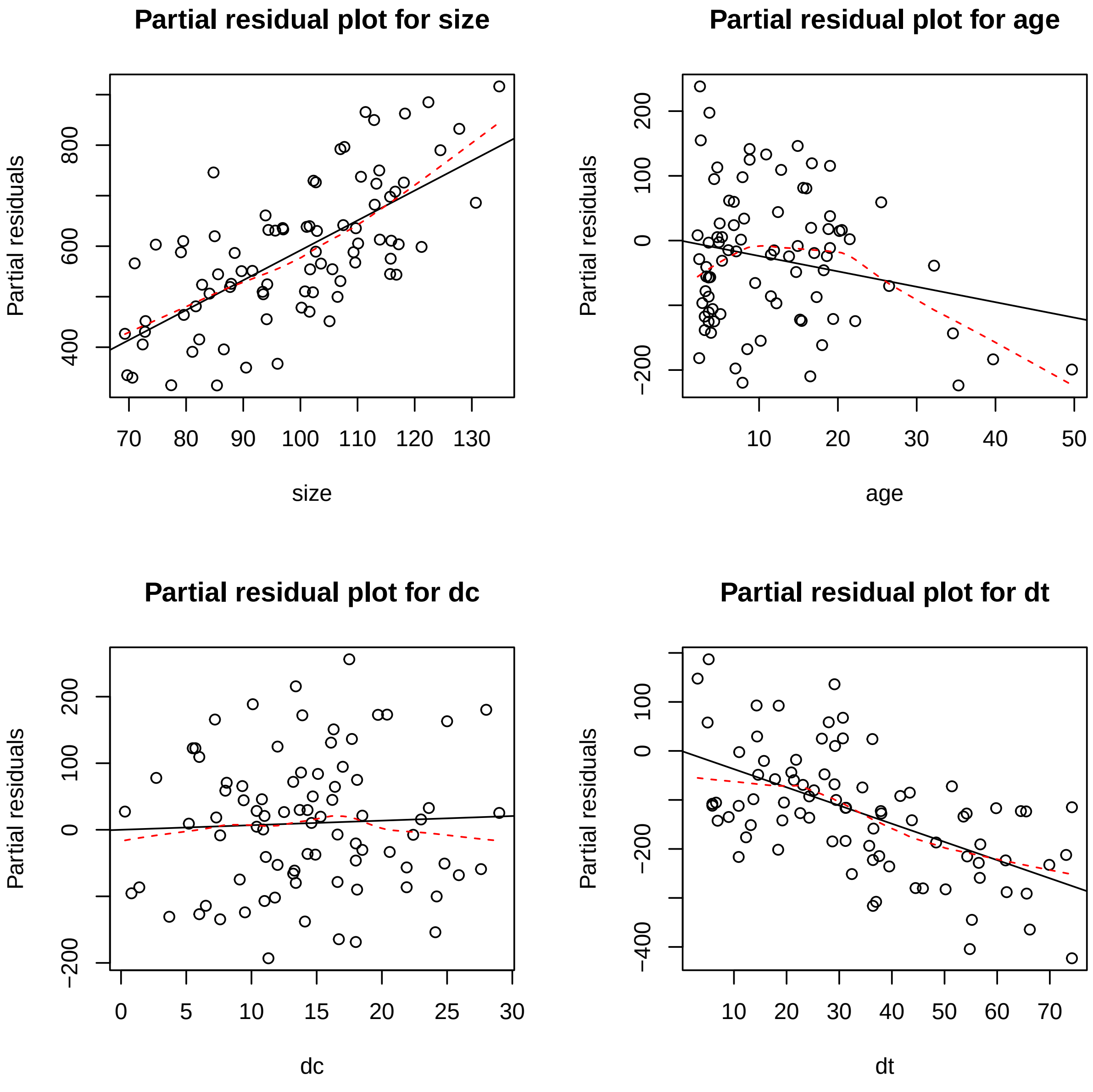
Lastly, we will check the linearity condition by constructing plots on each predictor. We want to inspect any correlation between our predictors and response and to see if our model is correct. That is, we will construct added variable and partial residual plots respectively.

Added Variable Plots



An added-variable plot is a effective way to show the correlation between our independent variables (`size`,`age`, `dc`, and `dt`) and `price` 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. Here we can see 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. This is similar for the variable `age`, though not as much of a strong linear relationship. Lastly, for the `dc` variable, we notice that there is almost no linear relationship in the added variable plot. Perhaps this means that a property's distance to the city center (`dc`) adds very minimal contribution to the model. This is an indication that `dc` may not be needed in the investigation on how a property's listed price depends on other variables.

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. So, for the partial residual plots of `size`, `dc`, and `dt`, there is almost a perfect straight line. Therefore, we will say the model is correct. However, for the variable `age`, there is not a straight line in the partial residual plot. Therefore, a higher order term or other transformation is most likely necessary to remedy this problem.

Model Selection

From our previous summary statistics as well as the added variable plot of `dc`, we have evidence to believe that this predictor may not be necessary to include in the final model. Therefore, we will perform model selection using AIC criterion to confirm this belief.

```r
# forward
step(lm(price~1, data=property),
scope=list(upper=formula(fit_property)),
direction="forward")
```

```
## Start:  AIC=834
## price ~ 1
##
##         Df Sum of Sq     RSS AIC
## + size  1    542104 1325483 807
## + dt    1    333814 1533773 819
## <none>              1867587 834
## + age   1     39145 1828443 834
## + dc    1     27995 1839593 835
##
## Step:  AIC=807
## price ~ size
##
##         Df Sum of Sq     RSS AIC
## + dt    1    446697  878787 775
## + dc    1     70060 1255424 805
## + age   1     66081 1259403 805
## <none>              1325483 807
##
## Step:  AIC=775
## price ~ size + dt
##
##         Df Sum of Sq    RSS AIC
## + age   1     42986 835801 773
## <none>              878787 775
## + dc    1      4966 873821 777
##
## Step:  AIC=773
## price ~ size + dt + age
##
##         Df Sum of Sq    RSS AIC
## <none>              835801 773
## + dc    1      1036 834765 775

##
## Call:
## lm(formula = price ~ size + dt + age, data = property)
##
## Coefficients:
```

```
## (Intercept)         size           dt          age
##      185.11         5.78        -3.77        -2.45
```

```
# backward
step(fit_property, direction="backward")
```

```
## Start:  AIC=775
## price ~ size + age + dc + dt
##
##        Df Sum of Sq     RSS AIC
## - dc    1      1036  835801 773
## <none>               834765 775
## - age   1     39056  873821 777
## - dt    1    377327 1212091 804
## - size  1    514983 1349748 813
##
## Step:  AIC=773
## price ~ size + age + dt
##
##        Df Sum of Sq     RSS AIC
## <none>               835801 773
## - age   1     42986  878787 775
## - dt    1    423602 1259403 805
## - size  1    675485 1511286 820
```

```
##
## Call:
## lm(formula = price ~ size + age + dt, data = property)
##
## Coefficients:
## (Intercept)         size          age           dt
##      185.11         5.78        -2.45        -3.77
```

So, we will remove the variable dc (distance (in km) from the property to the city center) from the model, re-fit the model and run diagnostics again. We also attempt to remedy the non-linearity problem in the age variable that we noticed from the partial residual plot by adding higher order terms.

Below are the summary statistics. We notice including age^2 is significant at a 5% level, but the original variable age is still not significant.

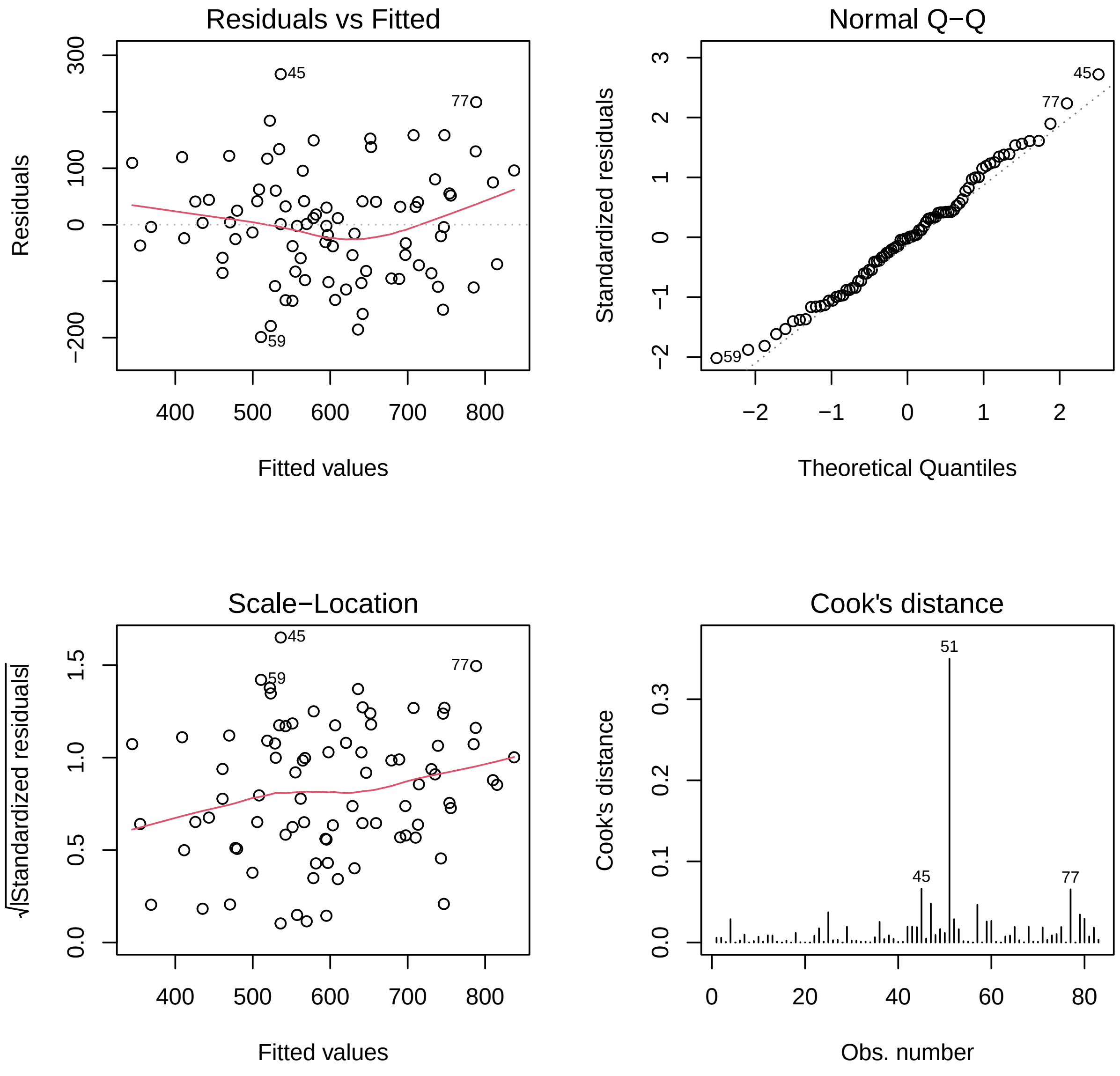
```
fit_property2 <- update(fit_property, price ~ .-dc)
fit_property2 <- update(fit_property2, price ~ size + age + I(age^2) + dt)

pander(summary(fit_property2))
```

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

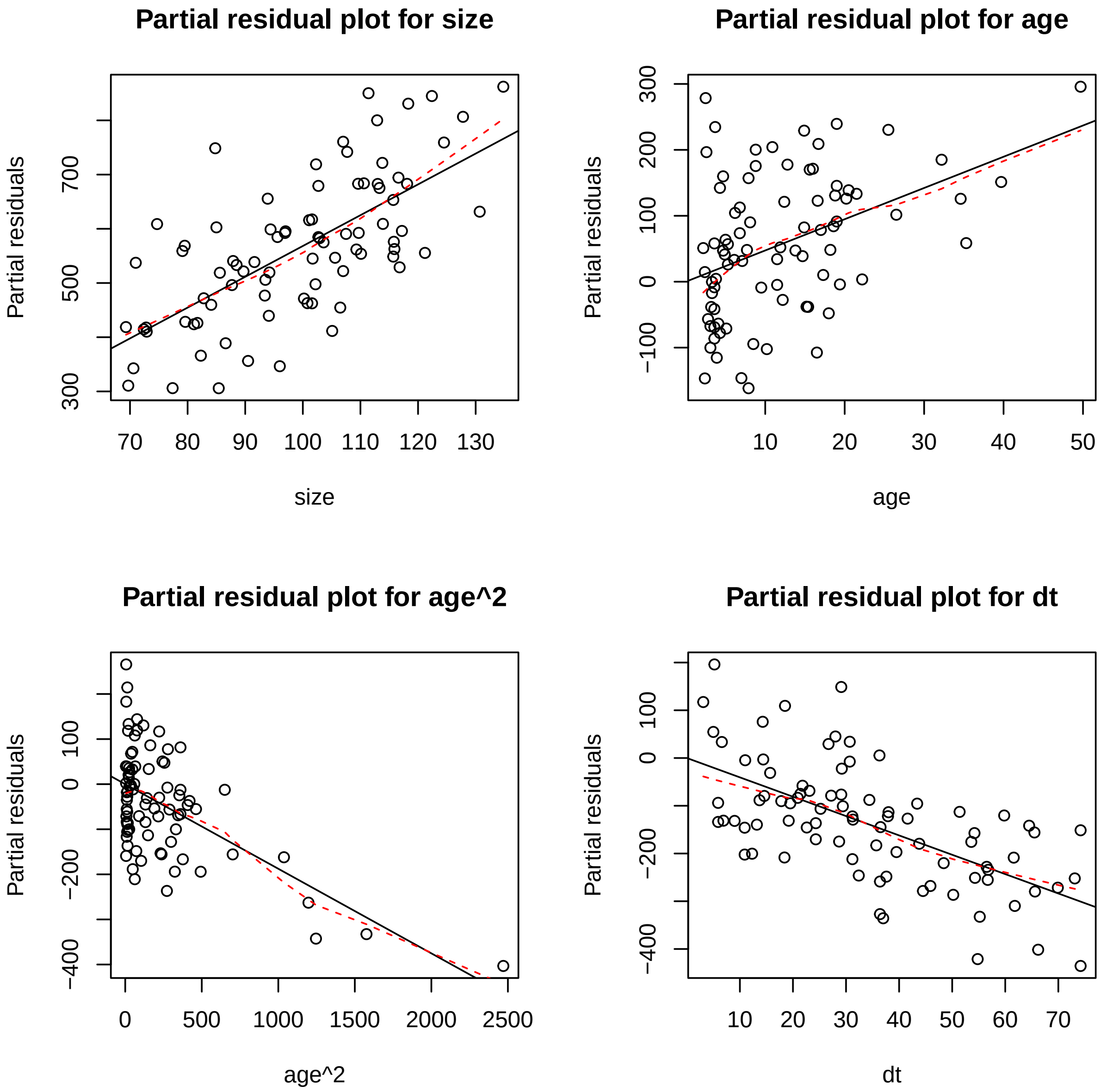
Observations	Residual Std. Error	R^2	Adjusted R^2
83	100.2	0.5804	0.5589

Looking at the diagnostic plots below, constant variance and normality assumptions are still met. Looking at the plot of Cook’s distance, we will investigate observation 51 to see if it is an outlier or influential point since it has a rather large distance value.



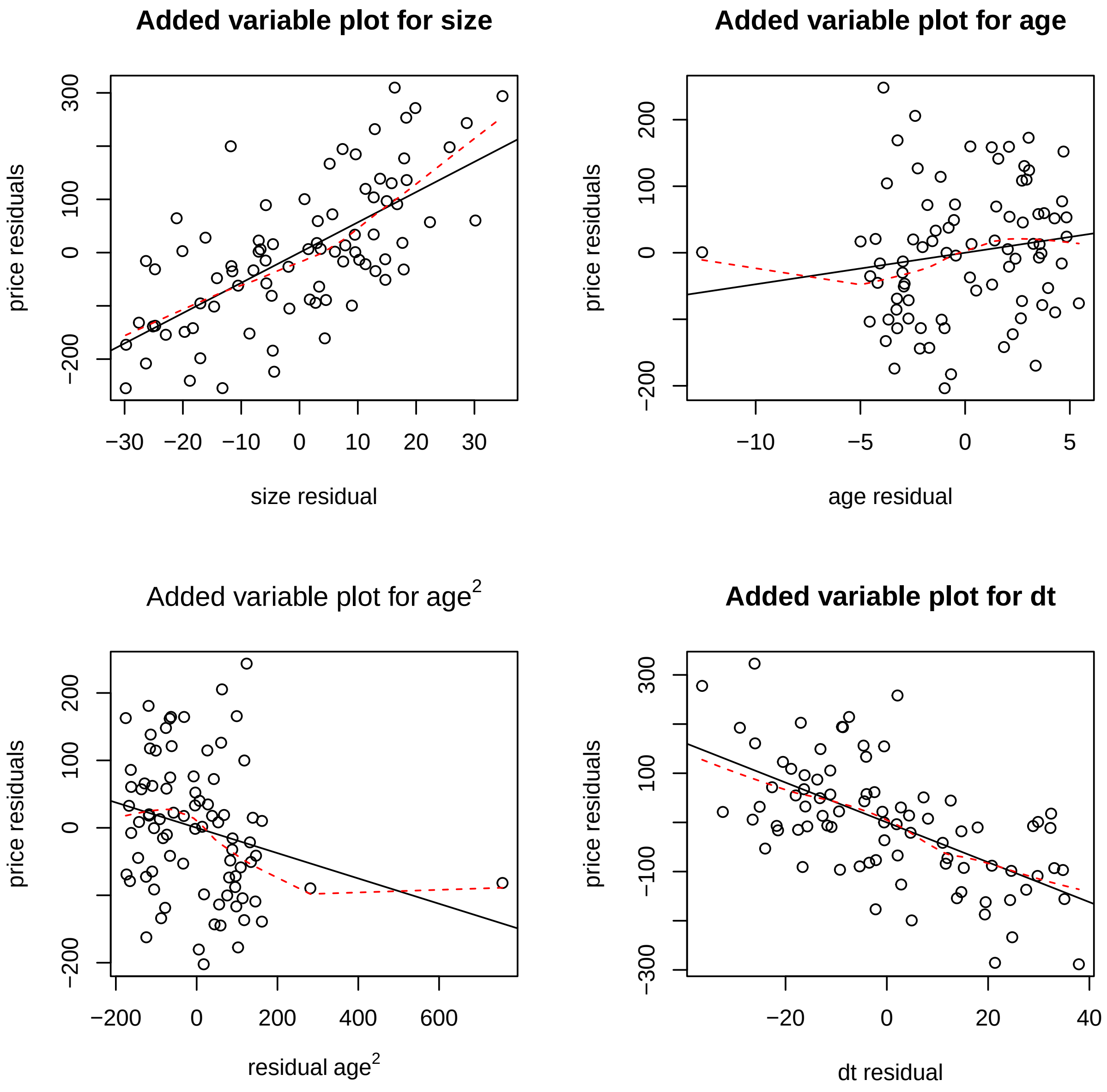
Partial Residual Plots

We also note by adding a higher order term for `age`, we also met the linearity assumption in the partial residual plots.



Added Variable Plots

However, when examining the added variable plots, 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. Perhaps if this point is an influential point and we remove it, it would fix this issue. We will investigate further.

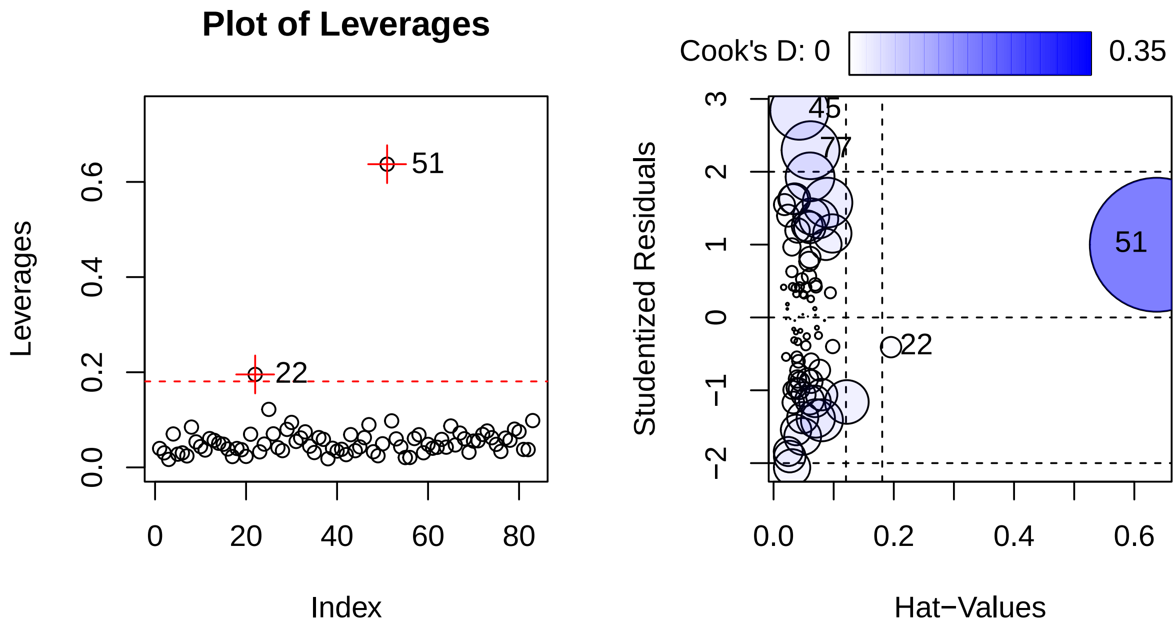

Added variable plot for size

Added variable plot for age

Added variable plot for age^2

Added variable plot for dt

We will now investigate observation 51 to determine its influence on this model.



```
##    StudRes   Hat  CookD
## 22   -0.41 0.195 0.0082
## 45    2.84 0.043 0.0665
## 51    1.00 0.637 0.3499
## 77    2.30 0.062 0.0656
```

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

```
outlierTest(fit_property2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 45      2.8             0.0058         0.48
```

Using the provided outlier test, we find there are no outliers present. Therefore, we will 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. Let's check this:

```
cd.prop2 <- cooks.distance(fit_property2) # Cook's statistic

n <- nrow(property)
which(cd.prop2 > 4/n)
```

```
## 45 51 77
## 45 51 77
```

Here we see 3 observations meet the condition of being an influential point. However, we have to remember that we check for influential points from large leverage points and outliers. Therefore, observation 51, a large leverage point, is an influential point here.

Regression Model Without Influential Point

Now, since we have an influential point, our next step is to fit the best regression model without this observation present in the data, and report these results.

```r
# remove observation 51
property_new <- property %>% filter(!row_number() %in% 51)
```

Below are the summary statistics. 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. Adjusted R^2 did slightly increase, however.

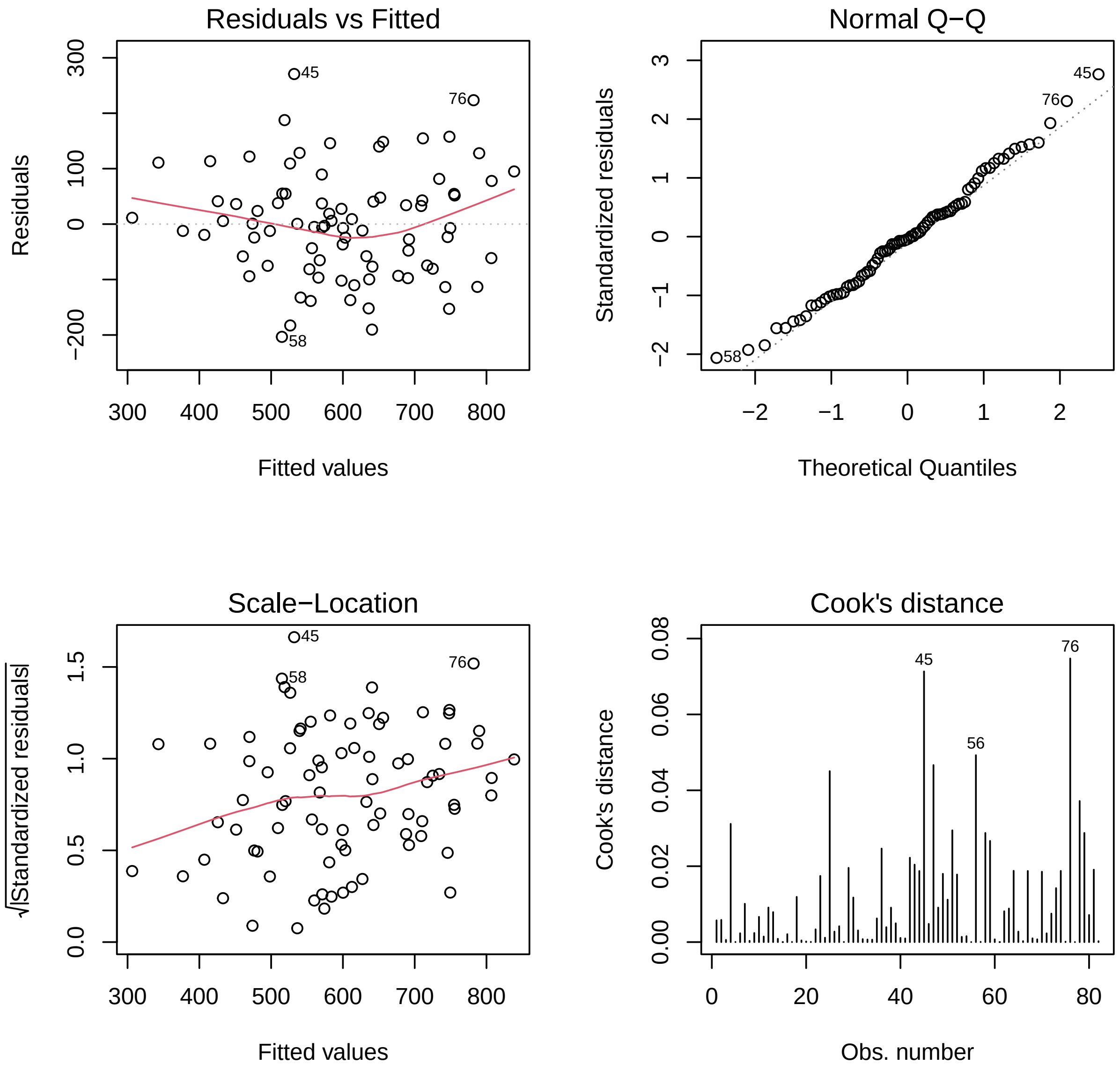
```r
fit_property_new <- lm(price ~ size + age + I(age^2) + dt, data = property_new)

pander(summary(fit_property_new))
```

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

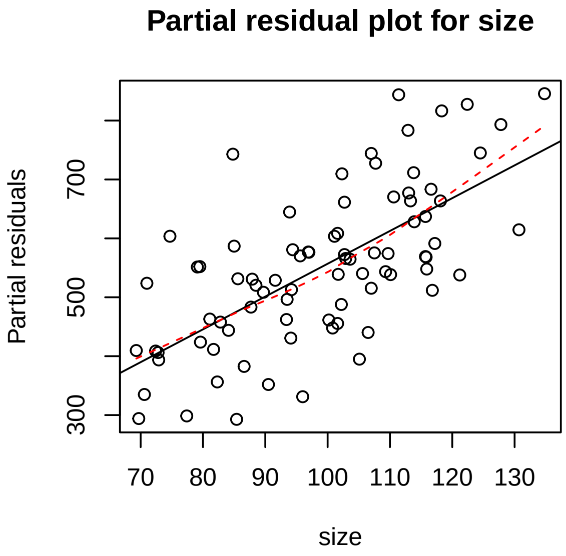
Observations	Residual Std. Error	R^2	Adjusted R^2
82	100.2	0.5858	0.5642

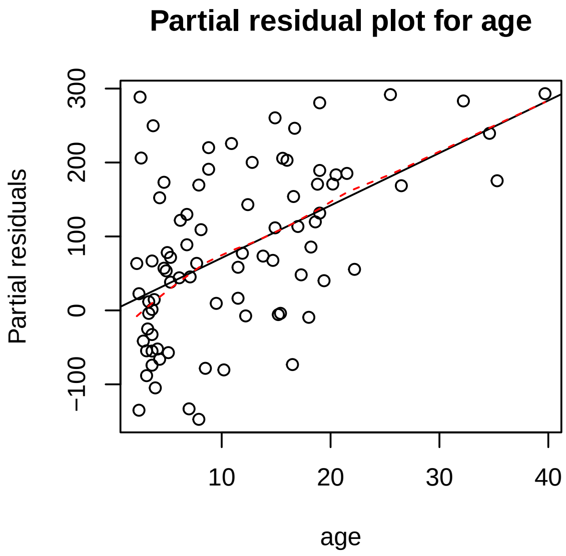
Looking at the diagnostic plots below, constant variance and normality assumptions are still met. Looking at the plot of Cook's distance, there are a few observations that stand out for investigation. However, the y-axis in the plot of Cook's distance is much smaller, so it is less likely that these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

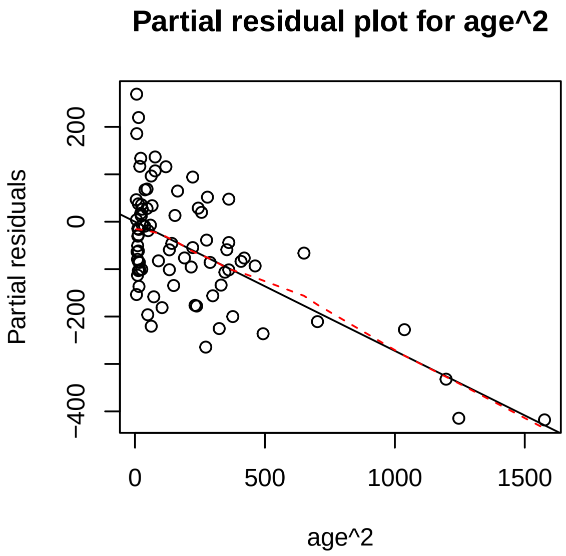


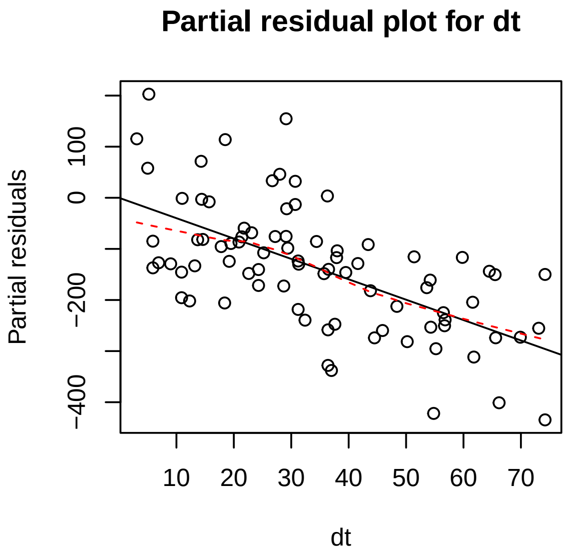
Partial Residual Plots

We also note the linearity assumption is still met in the partial residual plots after removing the influential observation. The plots even slighlty improve with even straighter lines.


Partial residual plot for size

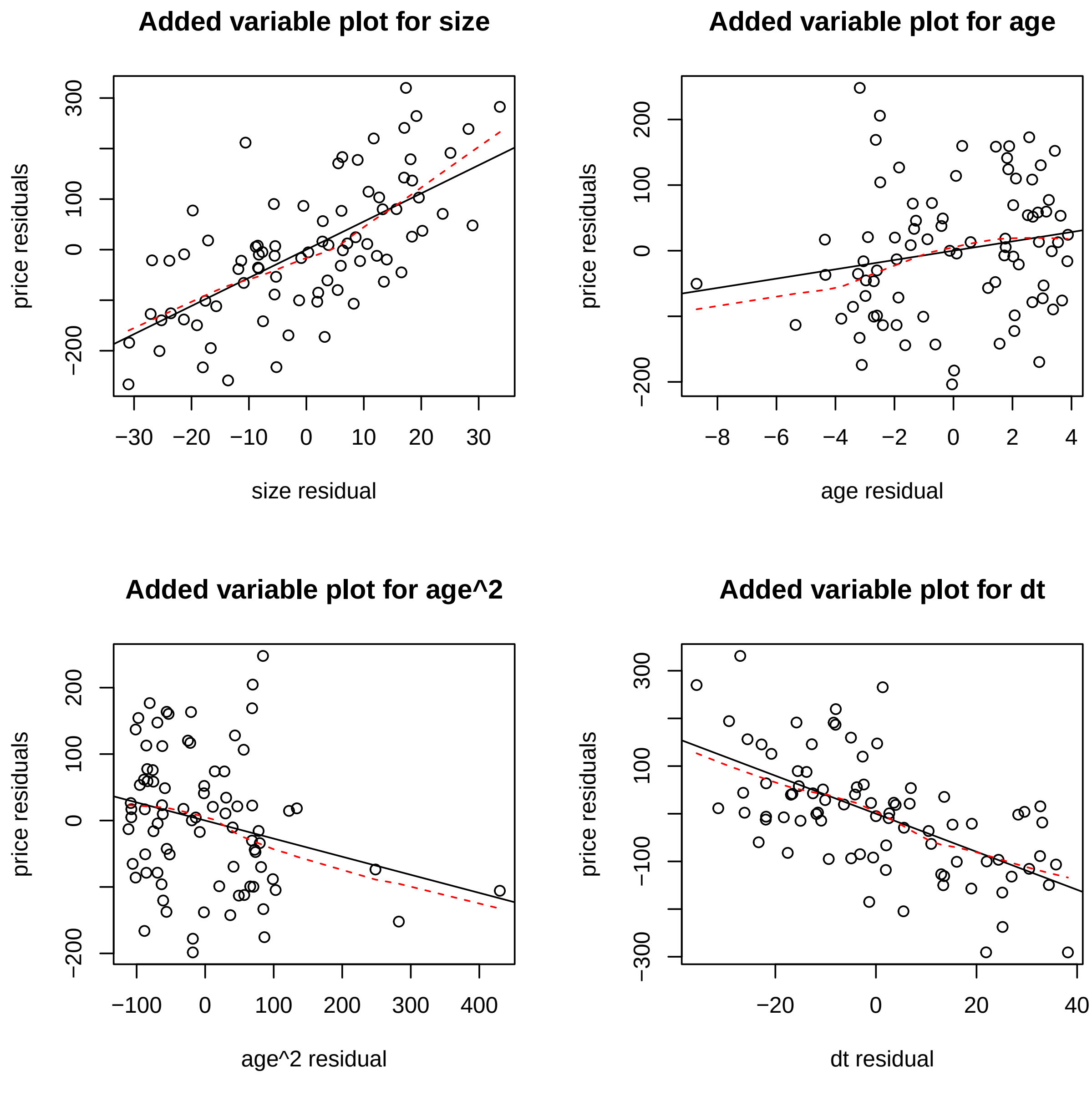

Partial residual plot for age


Partial residual plot for age^2

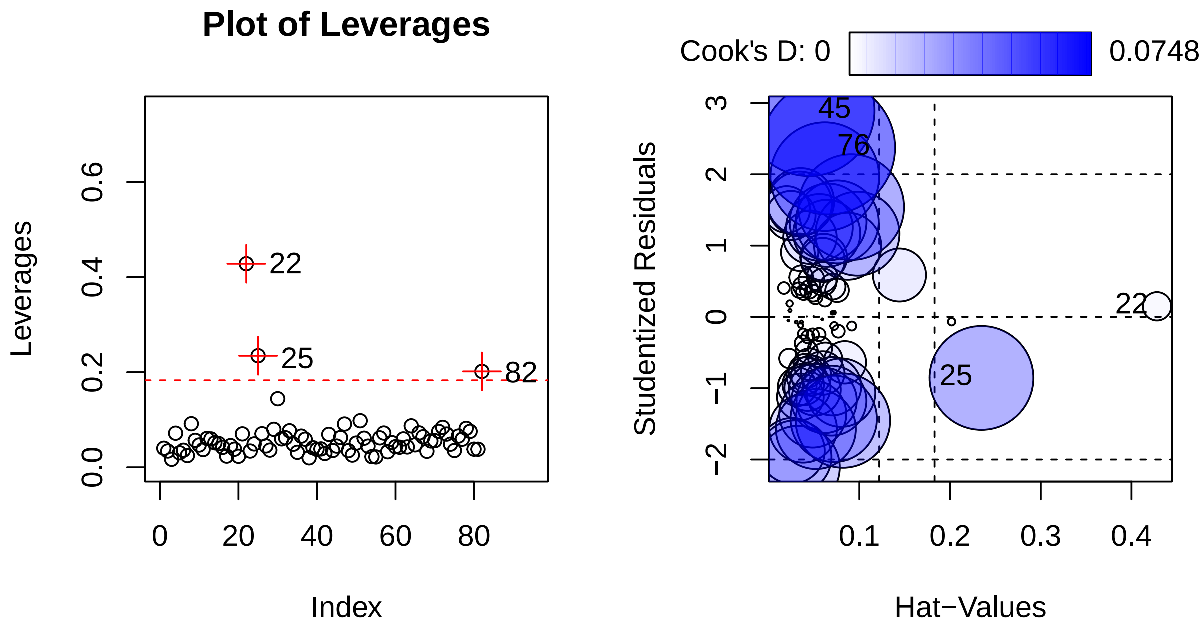

Partial residual plot for dt

Added Variable Plots

We now inspect if removing the influential observation fixes the linearity assumption in the `age` and higher order term `age` variables. We notice that this influential point actually did have a huge effect on the linearity. Now, all added variable plots meet the linearty assumption quite well.



After removing the previous influential observation (51), we will now investigate if any new observations have large leverage or are outliers and then see if they will be influential points.



```
##    StudRes   Hat  CookD
## 22    0.15 0.428 0.0034
## 25   -0.86 0.235 0.0451
## 45    2.89 0.045 0.0713
## 76    2.37 0.066 0.0748
```

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 indicates three times the mean leverage value.

```
outlierTest(fit_property2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 45      2.8           0.0058         0.48
```

Using the provided outlier test, we find there are still no outliers present. Therefore, we will only examine if observations 22, 25, and 82, 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. Let's check this:

```
cd.prop_new <- cooks.distance(fit_property_new) # Cook's statistic

n_new <- nrow(property_new)
which(cd.prop_new > 4/n_new)
```

```
## 45 56 76
## 45 56 76
```

Here, we see 3 observations meet the condition of being an influential point. However, we have to remember that we check for influential points from large leverage points and outliers. Since observations 22, 25, and 82 were the only large leverage points, and none of these observations were shown above to meet the requirements of being an influential point, we in fact, have no influential points in this model. We can see removing the previous influential observation results in no new influential points in this model.

Checklist To Find a Good Model

- Scatterplots of each Independent Variable
- Non-Constant Variance Assumption (Residuals vs Fitted, Scale Location)
- Normality Assumption (Q-Q Plot)
- Correlated Errors (2 plots, see HW4, Q3b)
  - Autocorrelation test (Durbin-Watson)
- Large Leverage Points (leverage value $>$ 3 times mean leverage, plot leverages)
- Outliers (outlier test, influence plot)
- Influential Points (Cook's Distance)
- At a Coefficient Level
  - Plotting Influence on each coefficient
  - Added Variable Plot
  - Partial Residual Plot
- Check if Transformation is Necessary
  - Log (only on response or both)
  - Sqrt (only on response or both)
  - Box-Cox (only on response)
- Check if model improves adding higher order terms
- Check if model improves adding interaction terms
- Model Selection?
  - AIC (forward, backward, both)
  - p-values
  - etc…