Robert Stephenson

Dr. Tan

MSC 450-01

5 October 2023

Midterm

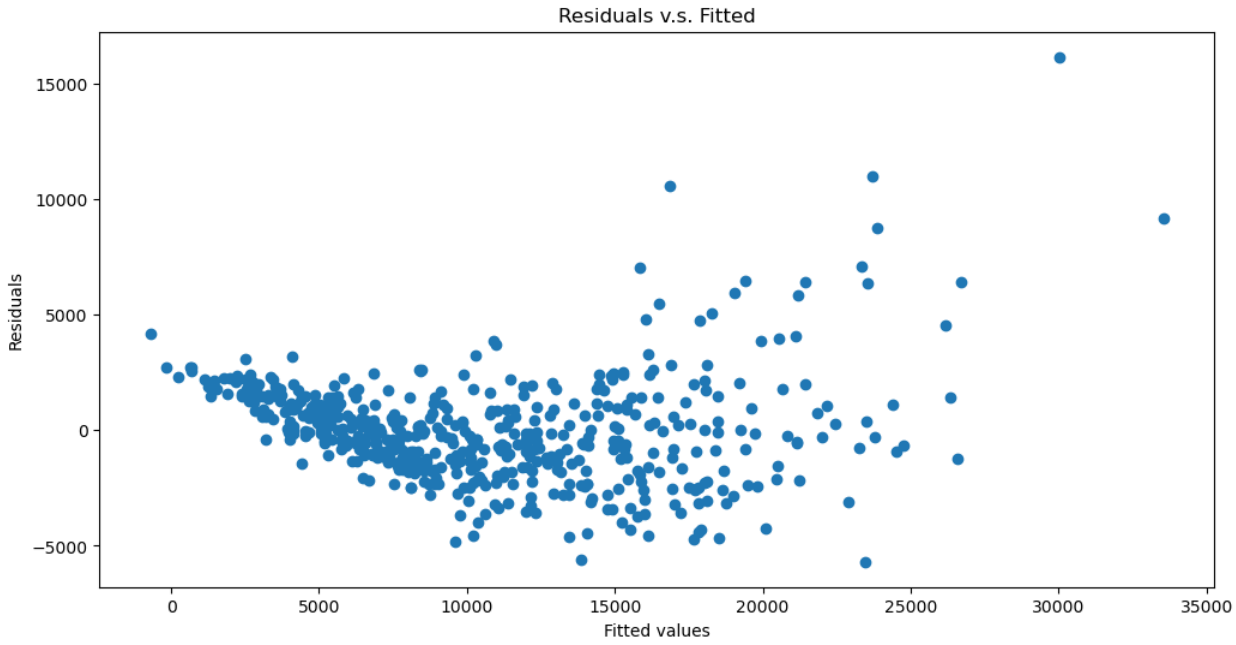
First we will import the data and run a basic linear regression on the raw data without manipulating any data.

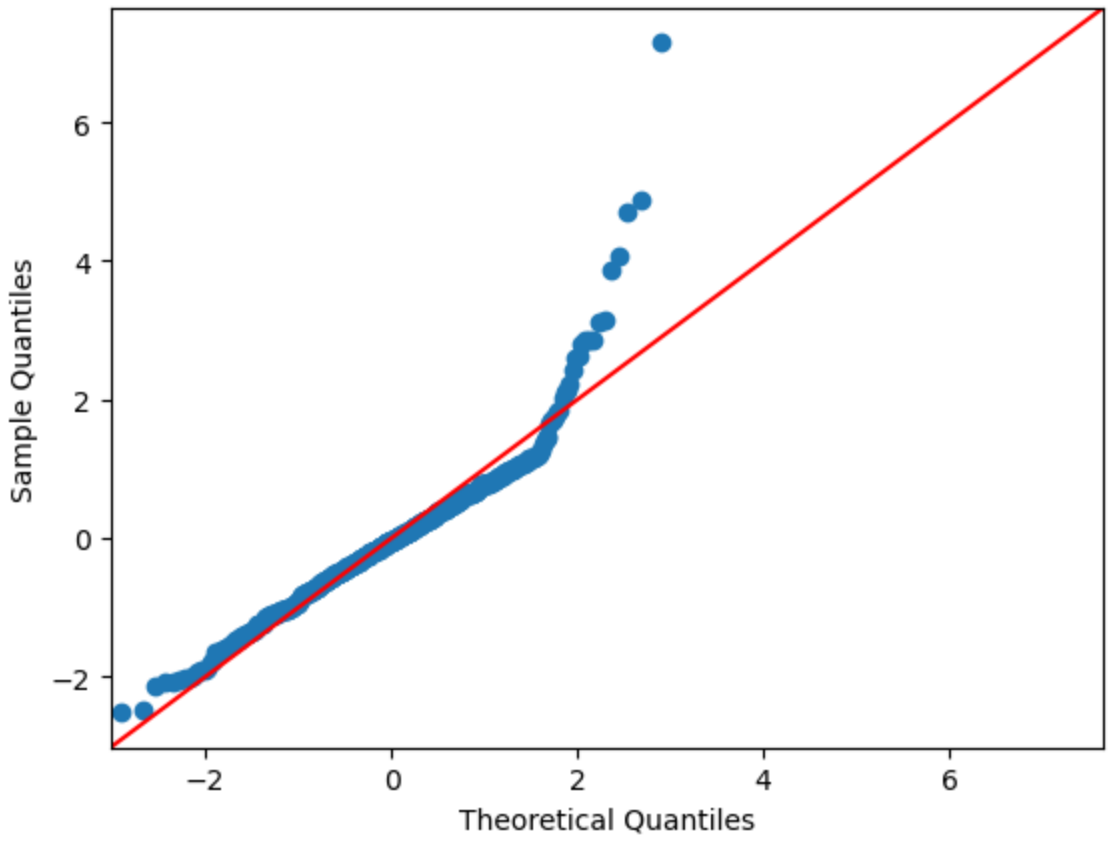
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.865
Method:                 Least Squares   F-statistic:                     346.4
Date:                Thu, 05 Oct 2023   Prob (F-statistic):          7.67e-225
Time:                        03:07:12   Log-Likelihood:                -4935.6
No. Observations:                 540   AIC:                             9893.
Df Residuals:                     529   BIC:                             9940.
Df Model:                          10
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          -1.99e+04    631.814    -31.500      0.000   -2.11e+04   -1.87e+04
airconditioning -627.6635    232.231     -2.703      0.007   -1083.871    -171.456
hotwaterheating -358.8964    483.557     -0.742      0.458   -1308.824     591.031
basement        1331.7594    232.131      5.737      0.000     875.747    1787.772
guestroom       -188.2469    284.738     -0.661      0.509    -747.603     371.109
mainroad        -551.6684    301.545     -1.829      0.068   -1144.041      40.704
area               6.5789      0.158     41.578      0.000       6.268       6.890
bedrooms         308.4766    154.663      1.995      0.047       4.647     612.306
bathrooms       2400.1873    221.901     10.816      0.000    1964.273    2836.102
stories         1702.2485    137.413     12.388      0.000    1432.306    1972.191
parking          775.0578    125.123      6.194      0.000     529.258    1020.857
==============================================================================
Omnibus:                      197.678   Durbin-Watson:                   1.748
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1321.689
Skew:                           1.444   Prob(JB):                     9.97e-288
Kurtosis:                      10.099   Cond. No.                     2.35e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.35e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

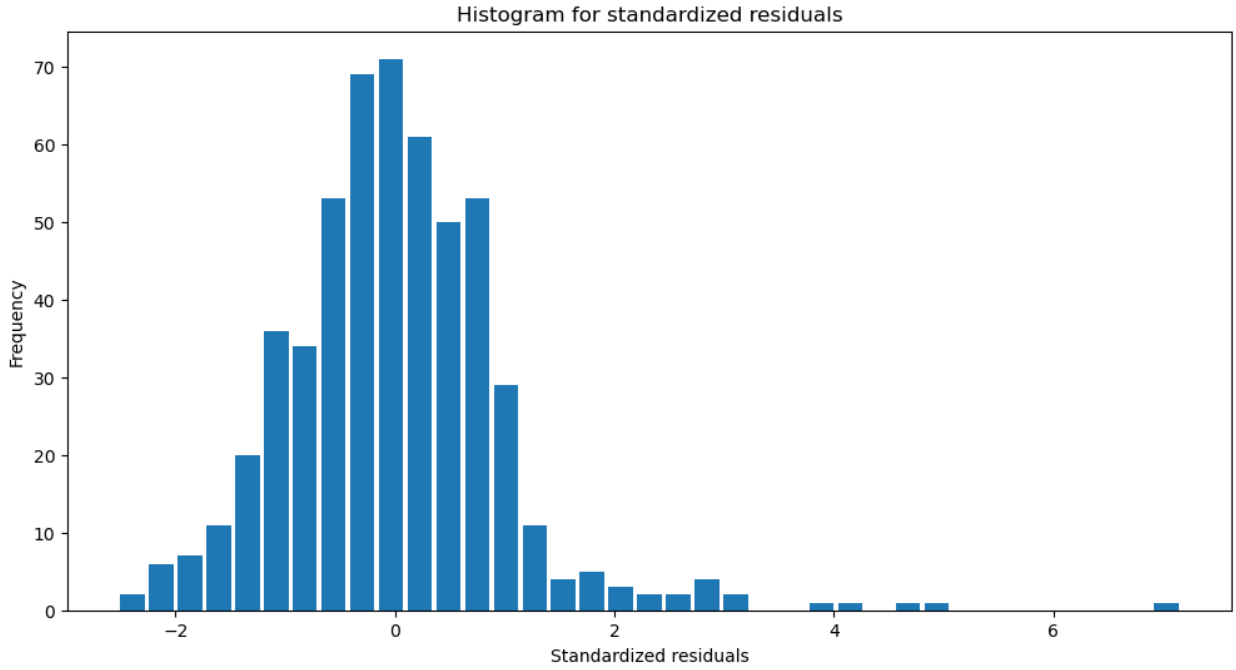
The fitted model is price = -1.99e^4 - airconditioning 627.6635 - hotwaterheating 358.8964 + basement 1331.7594 - guestroom 188.2469 - mainroad 551.6684 + area 6.5789 + bedrooms 308.4766 + bathrooms 2400.1873 + stories 1702.2485 + parking 775.0578.

Our adjusted R^2 is 0.865.

Here are the diagnostic graphs for this raw model.

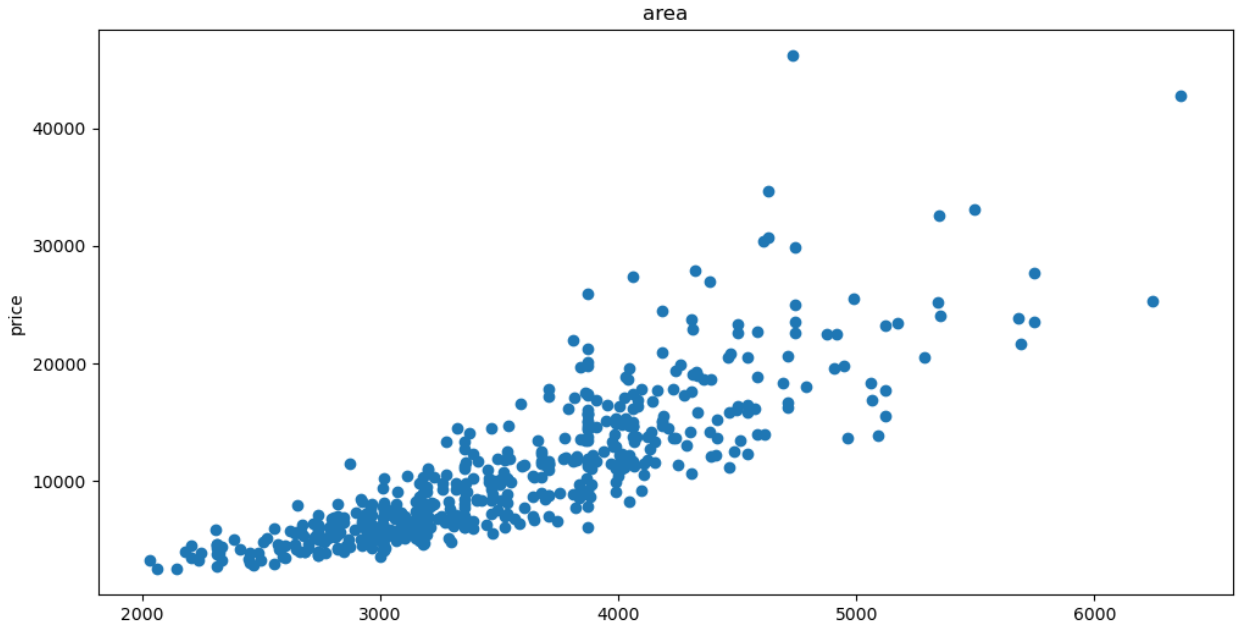



Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−2
0
2
4
6


Histogram for standardized residuals
Frequency
Standardized residuals
70
60
50
40
30
20
10
0
−2
0
2
4
6

We can see a possible violation of the zero mean assumption as the residual plot could be seen to have a very slight u-shaped curve. There is also a violation of the constant variance assumption as the data starts tight and the scatters towards the end. Our QQ plot also shows us a heavy positive tail which indicates a violation of the normality assumption. Our data also seems to be very left skewed.

We then see that area is our only variable that can be easily identified as to whether or not it has a linear relationship with the response variable as it is the only non-categorical or continuous variable. Our graph shows us that it appears to be linear.



We will then try a log-transformation of the response variable. Here is the model we then construct with the log-transformed response variable.

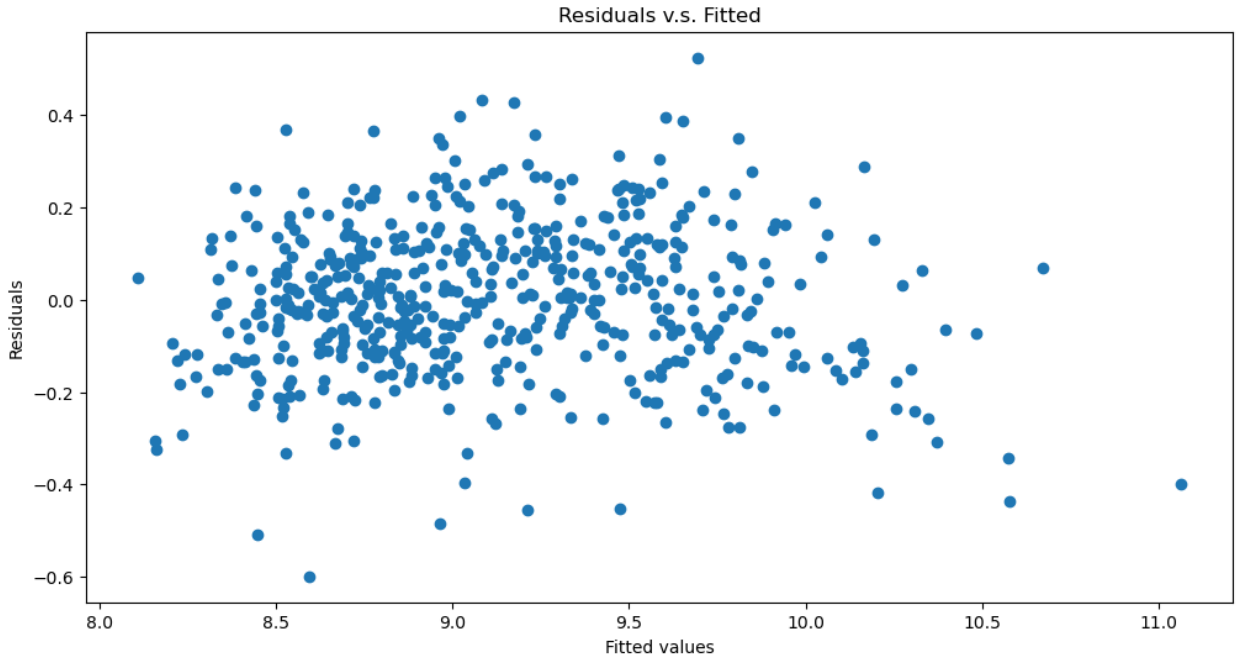
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Log_price   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.909
Method:                 Least Squares   F-statistic:                     537.2
Date:                Thu, 05 Oct 2023   Prob (F-statistic):          1.26e-269
Time:                        03:07:12   Log-Likelihood:                 218.62
No. Observations:                 540   AIC:                            -415.2
Df Residuals:                     529   BIC:                            -368.0
Df Model:                          10
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              6.3858      0.045    141.216      0.000       6.297       6.475
airconditioning   -0.0123      0.017     -0.738      0.461      -0.045       0.020
hotwaterheating    0.0243      0.035      0.702      0.483      -0.044       0.092
basement           0.1497      0.017      9.013      0.000       0.117       0.182
guestroom          0.0250      0.020      1.229      0.220      -0.015       0.065
mainroad           0.0160      0.022      0.741      0.459      -0.026       0.058
area               0.0006   1.13e-05     51.754      0.000       0.001       0.001
bedrooms           0.0373      0.011      3.372      0.001       0.016       0.059
bathrooms          0.1541      0.016      9.704      0.000       0.123       0.185
stories            0.1405      0.010     14.283      0.000       0.121       0.160
parking            0.0600      0.009      6.695      0.000       0.042       0.078
==============================================================================
Omnibus:                        4.442   Durbin-Watson:                   1.955
Prob(Omnibus):                  0.108   Jarque-Bera (JB):                4.840
Skew:                          -0.118   Prob(JB):                       0.0889
Kurtosis:                       3.400   Cond. No.                     2.35e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.35e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

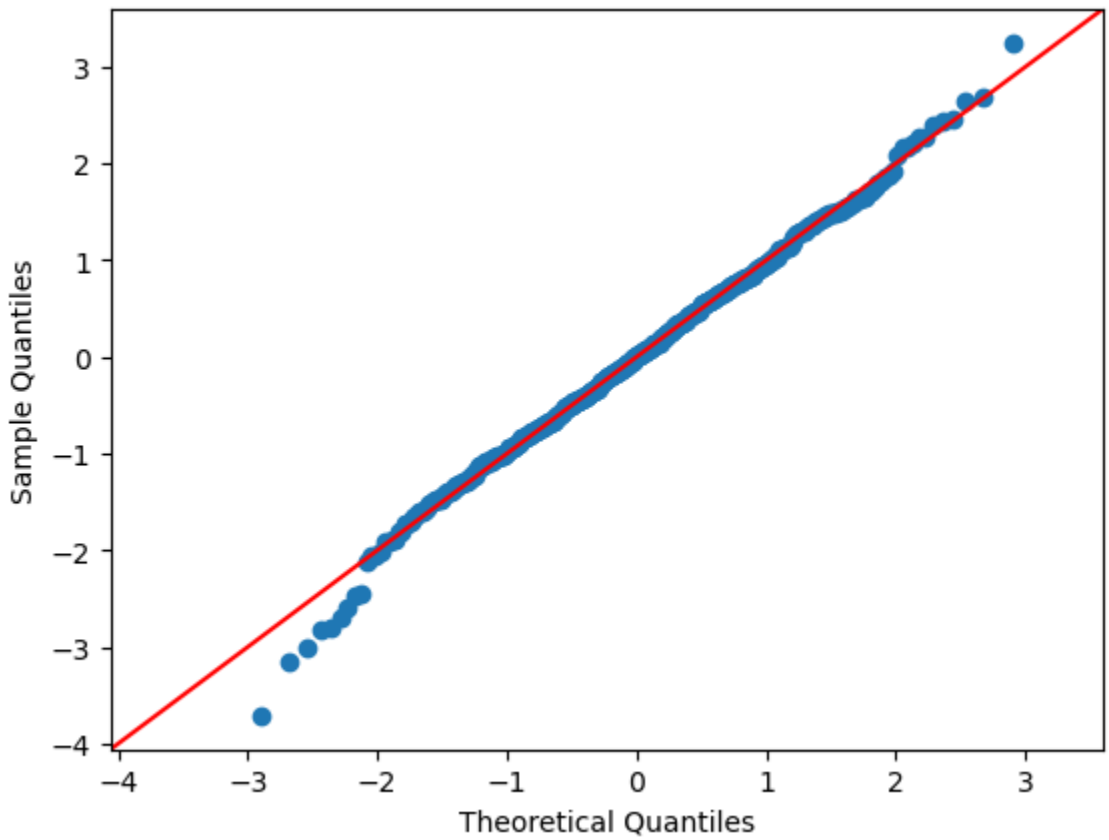
We see an improvement of R^2 in our new model and as it will be shown later, we also see improvements in our assumption violations.

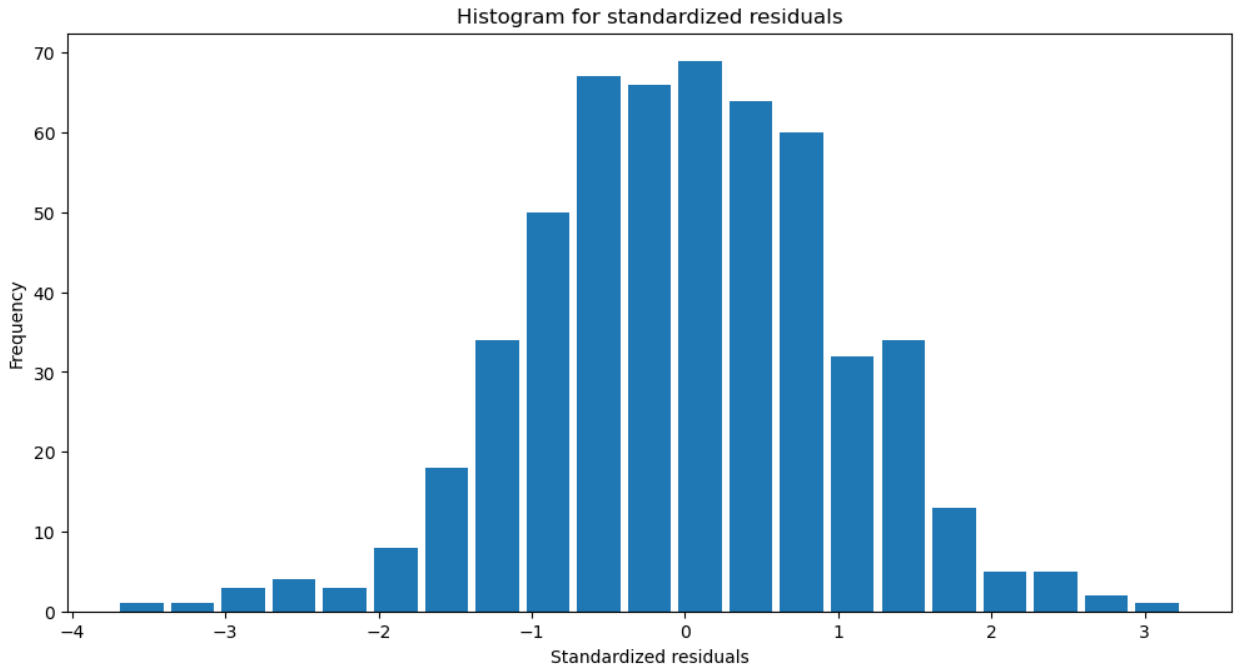
The new fitted model is Log_price = 6.3858 - airconditioning 0.0123 + hotwaterheating 0.0243 + basement 0.1497 + guestroom 0.0250 + mainroad 0.0160 + area 0.0006 + bedrooms 0.0373 + bathrooms 0.1541 + stories 0.1405 + parking 0.0600.

Our adjusted R^2 is 0.909

Here are our new diagnostic plots.







We can see from our residual plot that our zero mean and constant variance violations now seem to be rectified as the curvature shape is gone and the variance seems to be much more consistent.

We can also see that the positive tail that was in our previous QQ plot is now gone and our data is now only slightly left skewed. Which would indicate that we may have repaired our possible violation of the normality assumption.

We will now try to remove outliers with the following method.

```
data[abs(residual_norm) > 3]
```

	Log_price	airconditioning	hotwaterheating	basement	guestroom	mainroad	area	bedrooms	bathrooms	stories	parking
157	10.219759	0	0	1	1	1	4062.019202	4	2	2	0
406	7.936753	0	0	0	0	1	2315.707235	3	1	3	0
407	7.996794	0	0	1	0	1	2554.407955	3	1	2	0
469	8.478565	0	0	0	0	0	3298.484500	4	1	2	1

```
data = data[abs(residual_norm) <= 3]
```

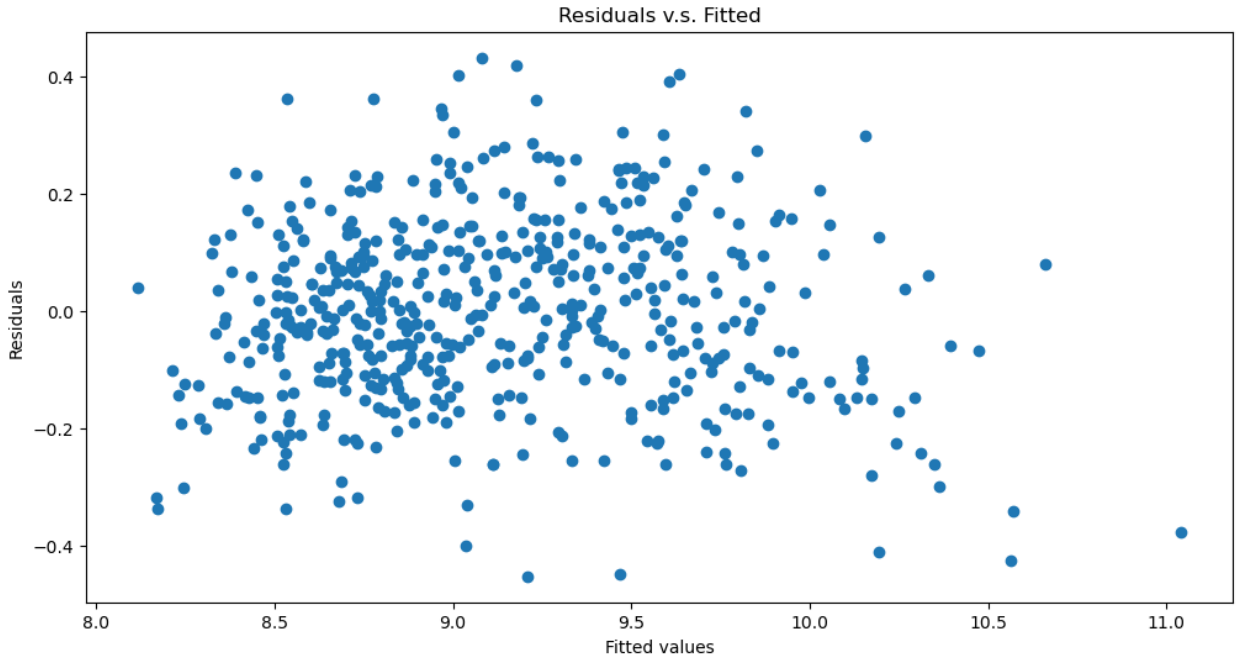
Now that we have removed the outliers we will do another linear regression model.

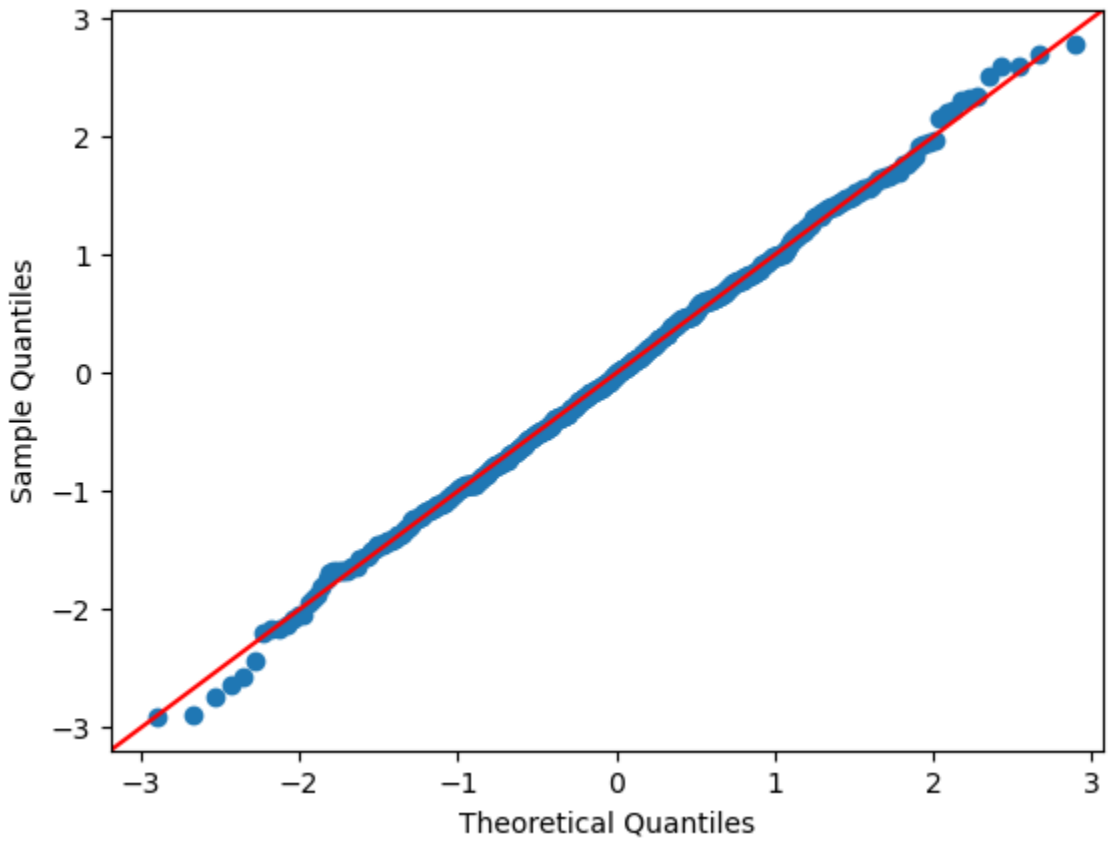
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Log_price   R-squared:                       0.915
Model:                            OLS   Adj. R-squared:                  0.914
Method:                 Least Squares   F-statistic:                     568.1
Date:                Thu, 05 Oct 2023   Prob (F-statistic):          3.89e-274
Time:                        03:07:13   Log-Likelihood:                237.92
No. Observations:                 536   AIC:                            -453.8
Df Residuals:                     525   BIC:                            -406.7
Df Model:                          10
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              6.4072      0.044    146.811      0.000       6.321       6.493
airconditioning   -0.0142      0.016     -0.887      0.376      -0.046       0.017
hotwaterheating    0.0218      0.033      0.656      0.512      -0.044       0.087
basement           0.1519      0.016      9.472      0.000       0.120       0.183
guestroom          0.0171      0.020      0.871      0.384      -0.022       0.056
mainroad           0.0132      0.021      0.632      0.528      -0.028       0.054
area               0.0006   1.09e-05     53.164      0.000       0.001       0.001
bedrooms           0.0380      0.011      3.560      0.000       0.017       0.059
bathrooms          0.1470      0.015      9.602      0.000       0.117       0.177
stories            0.1445      0.009     15.235      0.000       0.126       0.163
parking            0.0617      0.009      7.145      0.000       0.045       0.079
==============================================================================
Omnibus:                        0.073   Durbin-Watson:                   2.024
Prob(Omnibus):                  0.964   Jarque-Bera (JB):                0.158
Skew:                           0.004   Prob(JB):                        0.924
Kurtosis:                       2.916   Cond. No.                     2.35e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.35e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

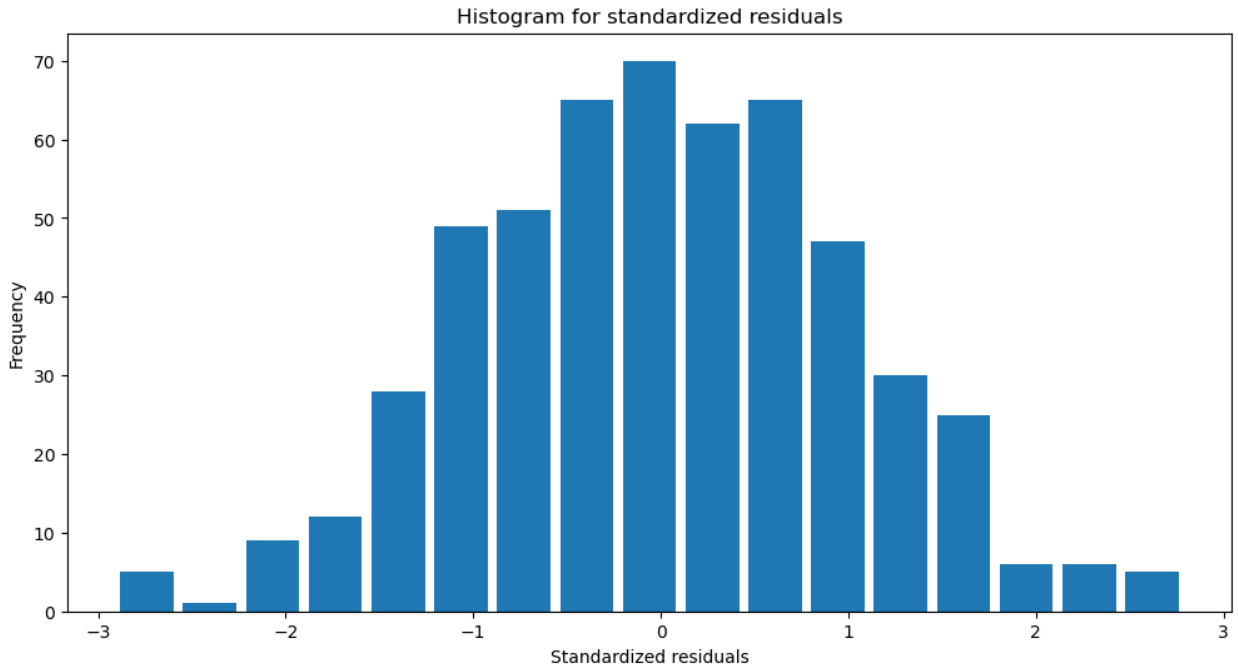
Our new fitted model is Log_price = 6.4072 - airconditioning 0.0142 + hotwaterheating 0.0218 + basement 0.1519 + guestroom 0.0171 + mainroad 0.0132 + area 0.0006 + bedrooms 0.0380 + bathrooms 0.1470 + stories 0.1445 + parking 0.0617.

Our new adjusted R^2 is 0.914

Here are our new diagnostics plots.







We can see that while our charts stay the same we do see that the data is now almost not skewed at all and seems to have a near-perfect bell shaped curve. Which is likely because of our removal of outliers.

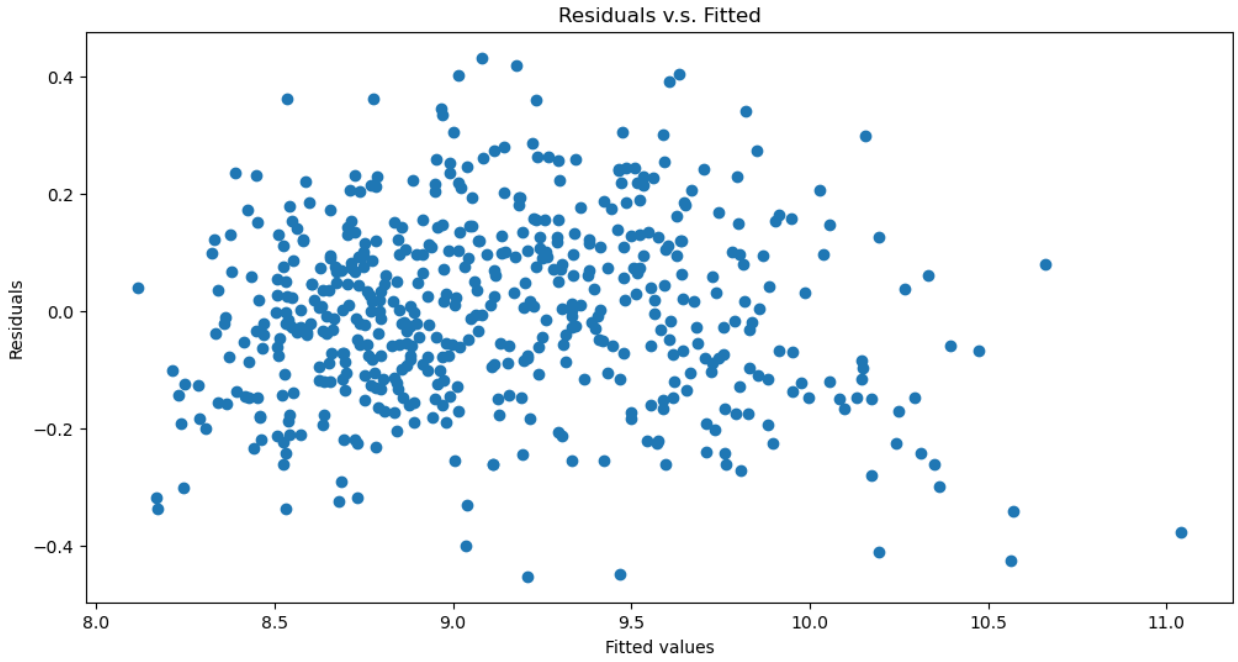
We then remove unnecessary predictors by doing a subset selection. We determined the model with 6 predictors to be the best as it had the highest R^2 value as well as the lowest AIC and BIC. We will then create another linear regression model on this new subset.

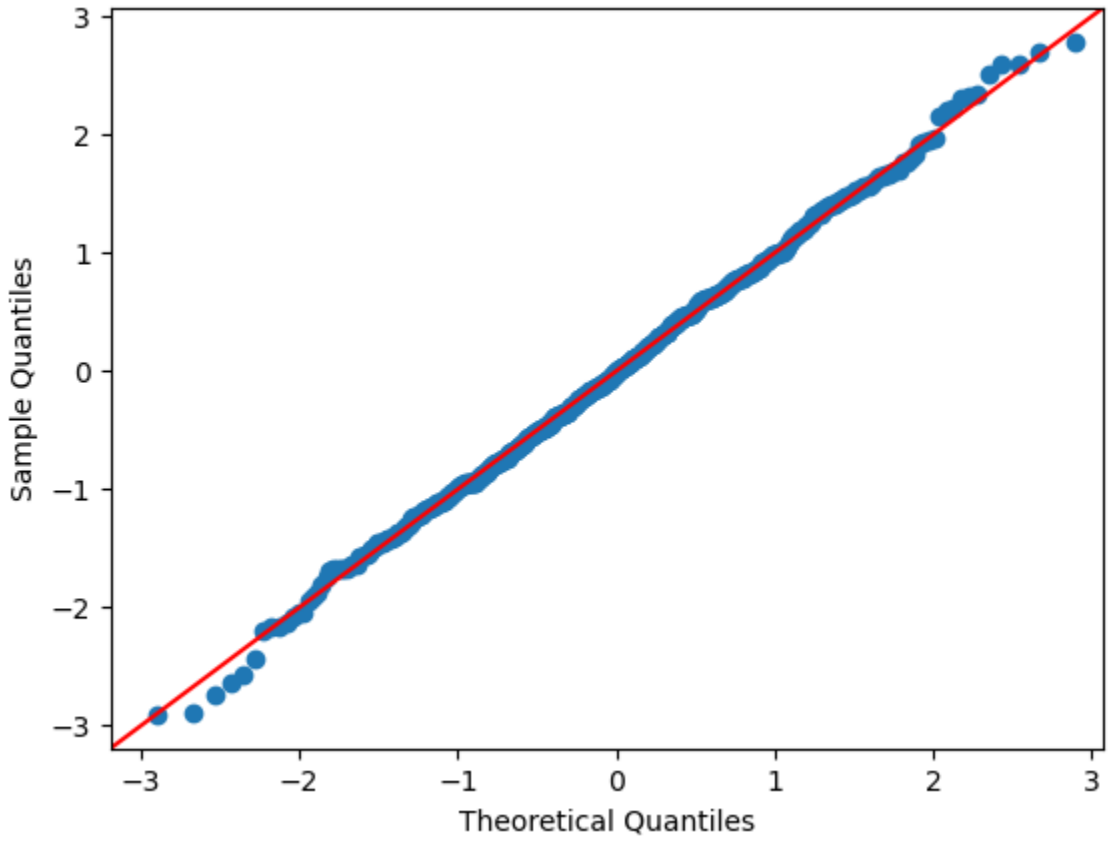
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Log_price   R-squared:                       0.915
Model:                            OLS   Adj. R-squared:                  0.914
Method:                 Least Squares   F-statistic:                     949.0
Date:                Thu, 05 Oct 2023   Prob (F-statistic):          2.03e-279
Time:                        03:07:16   Log-Likelihood:                236.64
No. Observations:                 536   AIC:                            -459.3
Df Residuals:                     529   BIC:                            -429.3
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.4154      0.042    153.296      0.000       6.333       6.498
basement       0.1564      0.015     10.568      0.000       0.127       0.185
area           0.0006   1.03e-05     56.578      0.000       0.001       0.001
bedrooms       0.0373      0.011      3.519      0.000       0.016       0.058
bathrooms      0.1472      0.015      9.667      0.000       0.117       0.177
stories        0.1439      0.009     15.964      0.000       0.126       0.162
parking        0.0618      0.009      7.261      0.000       0.045       0.078
==============================================================================
Omnibus:                        0.134   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.935   Jarque-Bera (JB):                0.234
Skew:                          -0.005   Prob(JB):                        0.889
Kurtosis:                       2.898   Cond. No.                     2.24e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.24e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

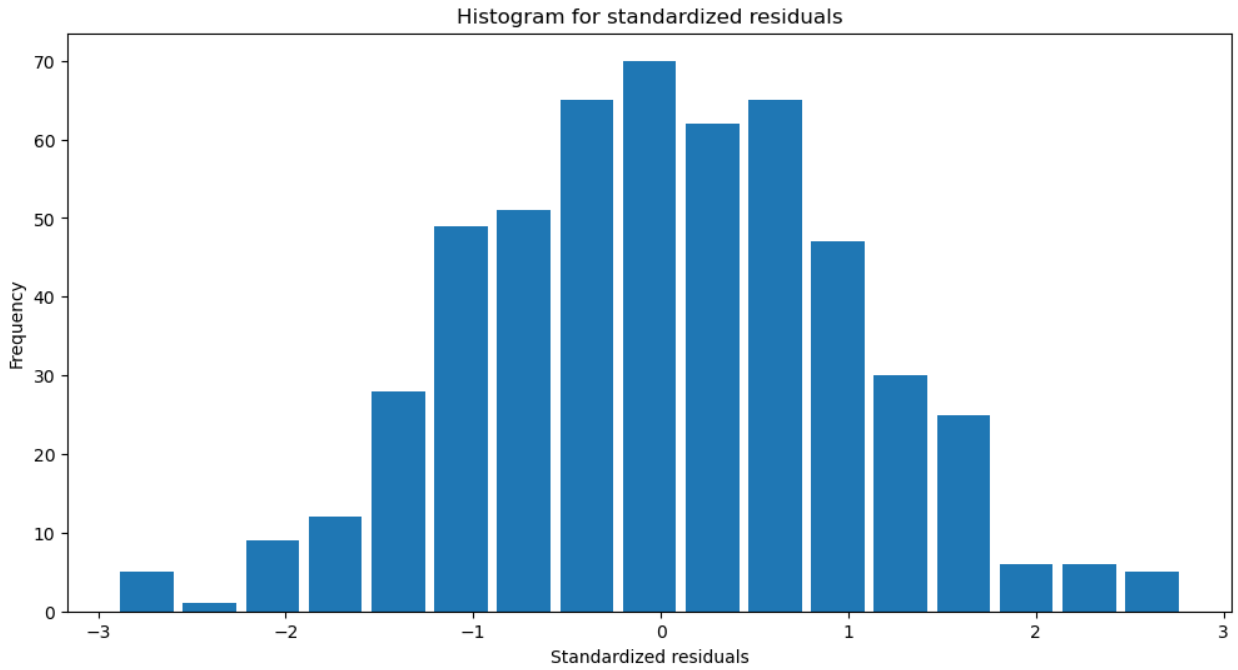
Our new fitted model is Log_price = 6.4154 + basement 0.1564 + area 0.0006 + bedrooms 0.0373 + bathrooms 0.1472 + stories 0.1439 + parking 0.0618.

Our new adjusted R^2 is 09.14.

Interestingly our diagnostics charts seem to remain the exact same from our last model.


Residuals v.s. Fitted
Residuals
Fitted values
0.4
0.2
0.0
−0.2
−0.4
8.0
8.5
9.0
9.5
10.0
10.5
11.0


Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3



I believe this model has adequately addressed violations of the error term assumptions, removed unnecessary predictors, and has dealt with outliers. This will be the final model that I will report.

1. In the final model: guestroom, hot water heater, and air conditioning do not significantly influence housing price as they are not in the final model. Basement however, does influence the housing price as its p-value of 0.000 is less than a 0.05 significance level. Basement's relation to the independent variable is that for every 1 unit increase in basement there is a 0.1564 unit increase in Log_price.
2. We make the predictions on the housing_test.csv and our MSE is 5999659.242668923.