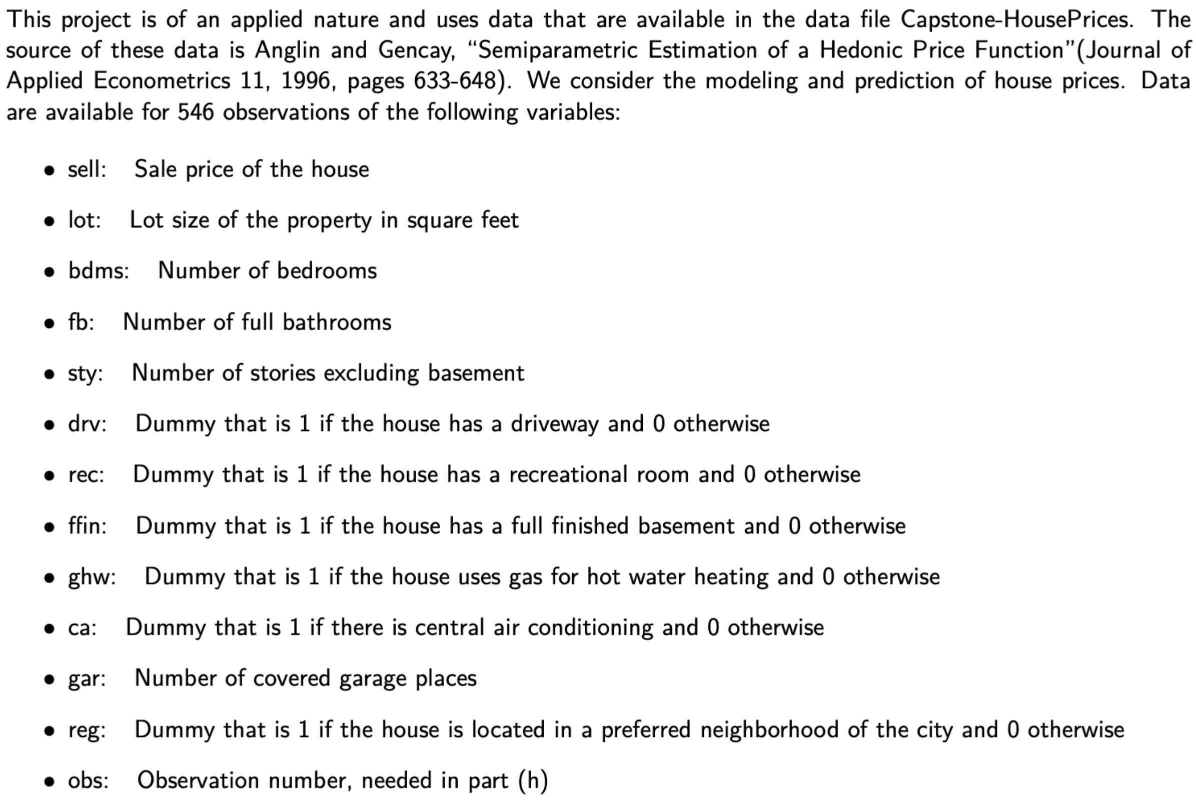
MIMANI ANIKET

CASE PROJECT

04/07/2025

Name of course: Econometrics – Methods and Applications
Offered by: Erasmus University Rotterdam

Background



This project is of an applied nature and uses data that are available in the data file Capstone-HousePrices. The source of these data is Anglin and Gencay, "Semiparametric Estimation of a Hedonic Price Function"(Journal of Applied Econometrics 11, 1996, pages 633-648). We consider the modeling and prediction of house prices. Data are available for 546 observations of the following variables:

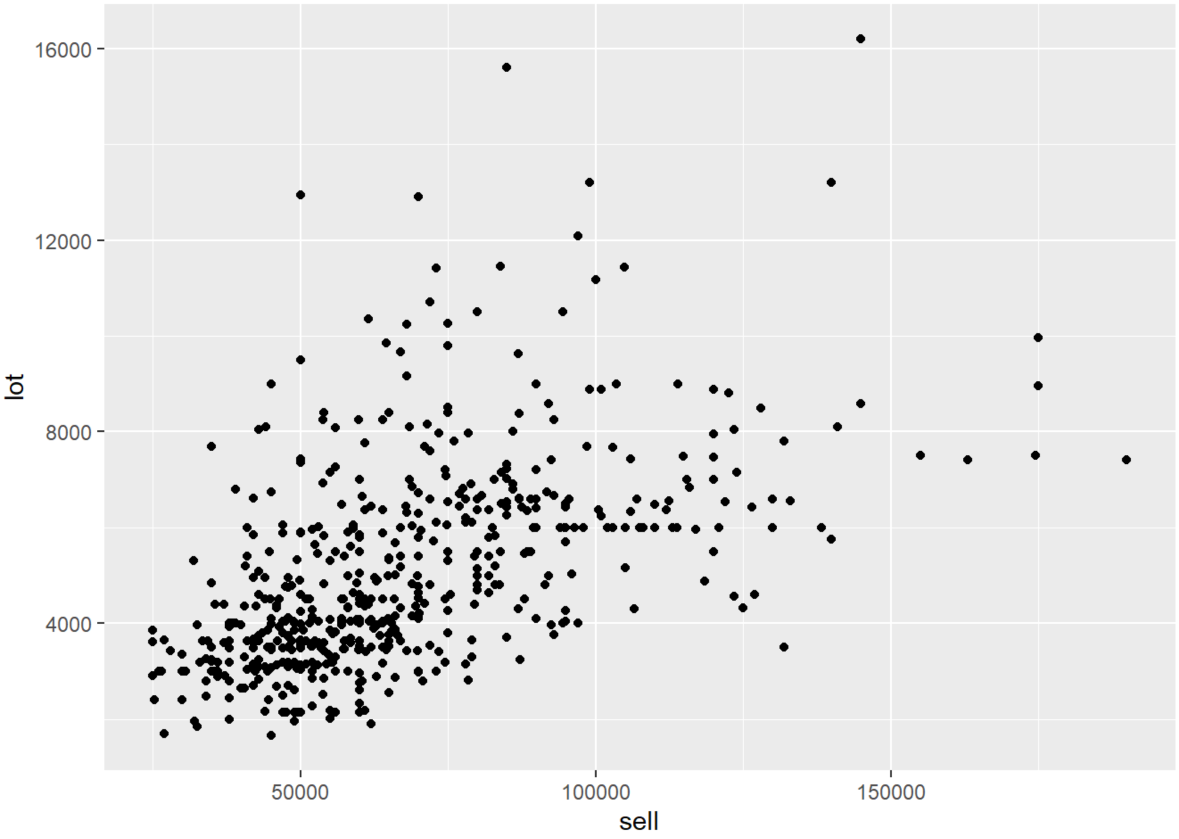
- sell: Sale price of the house
- lot: Lot size of the property in square feet
- bdms: Number of bedrooms
- fb: Number of full bathrooms
- sty: Number of stories excluding basement
- drv: Dummy that is 1 if the house has a driveway and 0 otherwise
- rec: Dummy that is 1 if the house has a recreational room and 0 otherwise
- ffin: Dummy that is 1 if the house has a full finished basement and 0 otherwise
- ghw: Dummy that is 1 if the house uses gas for hot water heating and 0 otherwise
- ca: Dummy that is 1 if there is central air conditioning and 0 otherwise
- gar: Number of covered garage places
- reg: Dummy that is 1 if the house is located in a preferred neighborhood of the city and 0 otherwise
- obs: Observation number, needed in part (h)

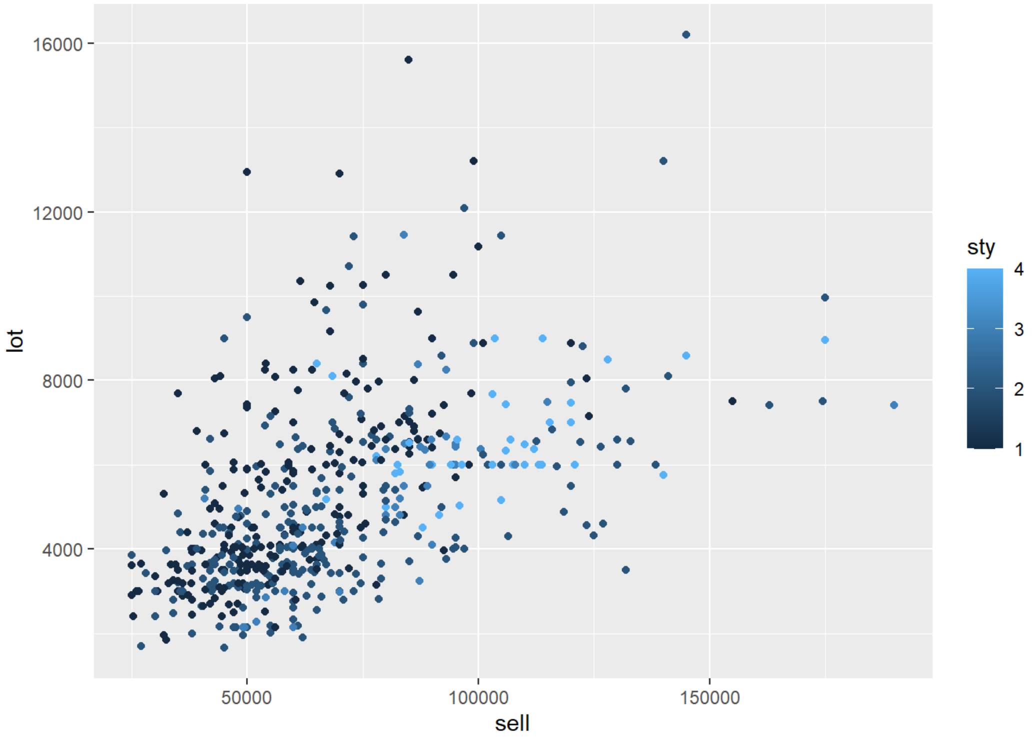
QUESTIONS AND ANSWERS:

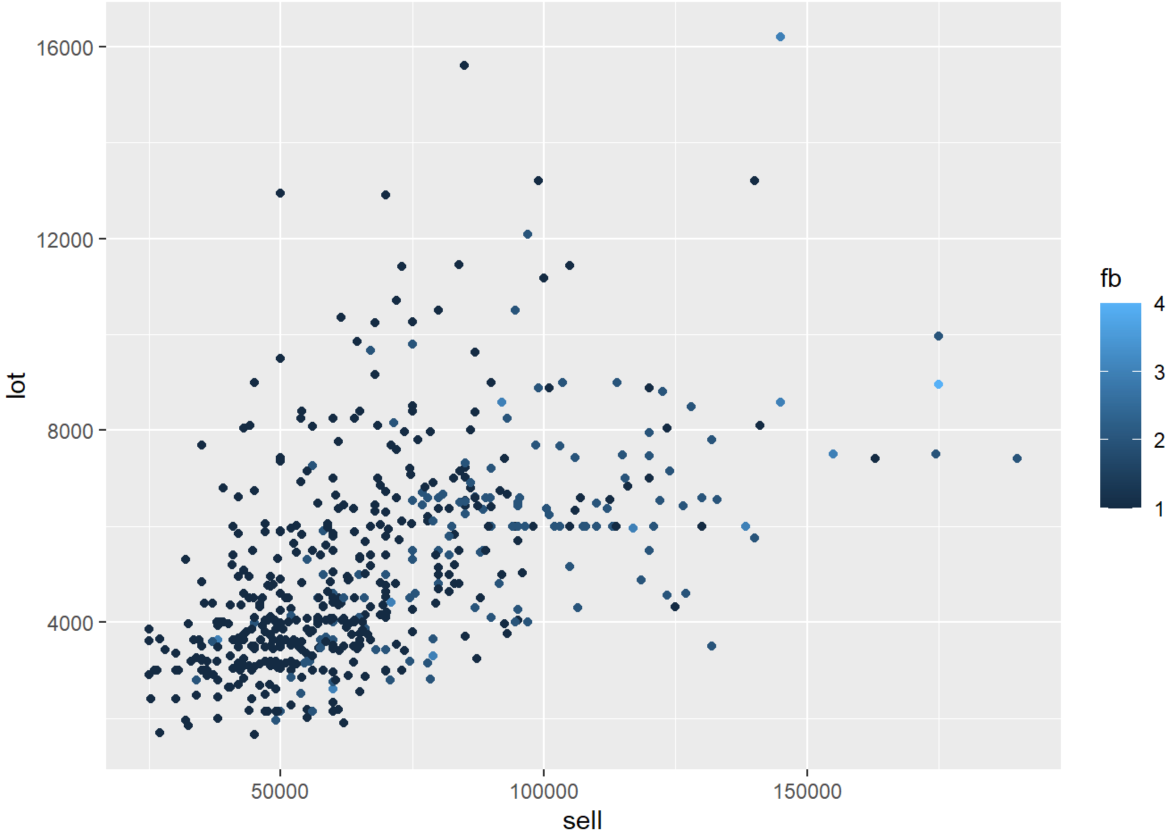
a) Consider a linear model where the sale price of a house is the dependent variable and the explanatory variables are the other variables given above. Perform a test for linearity. What do you conclude based on the test result?

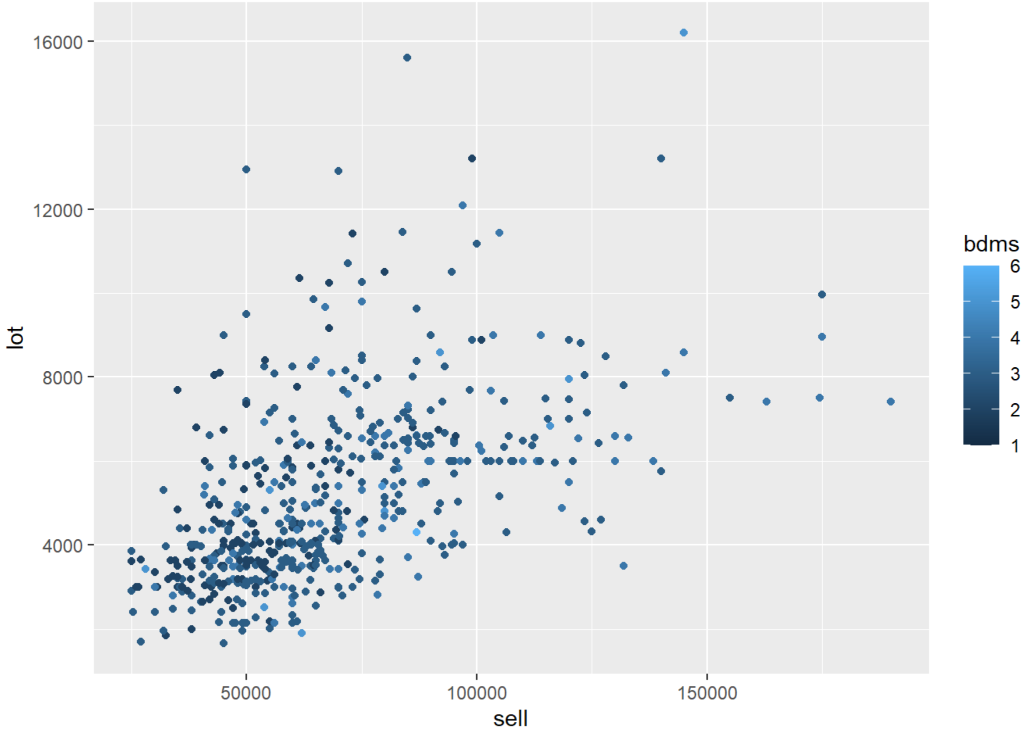
Answer)

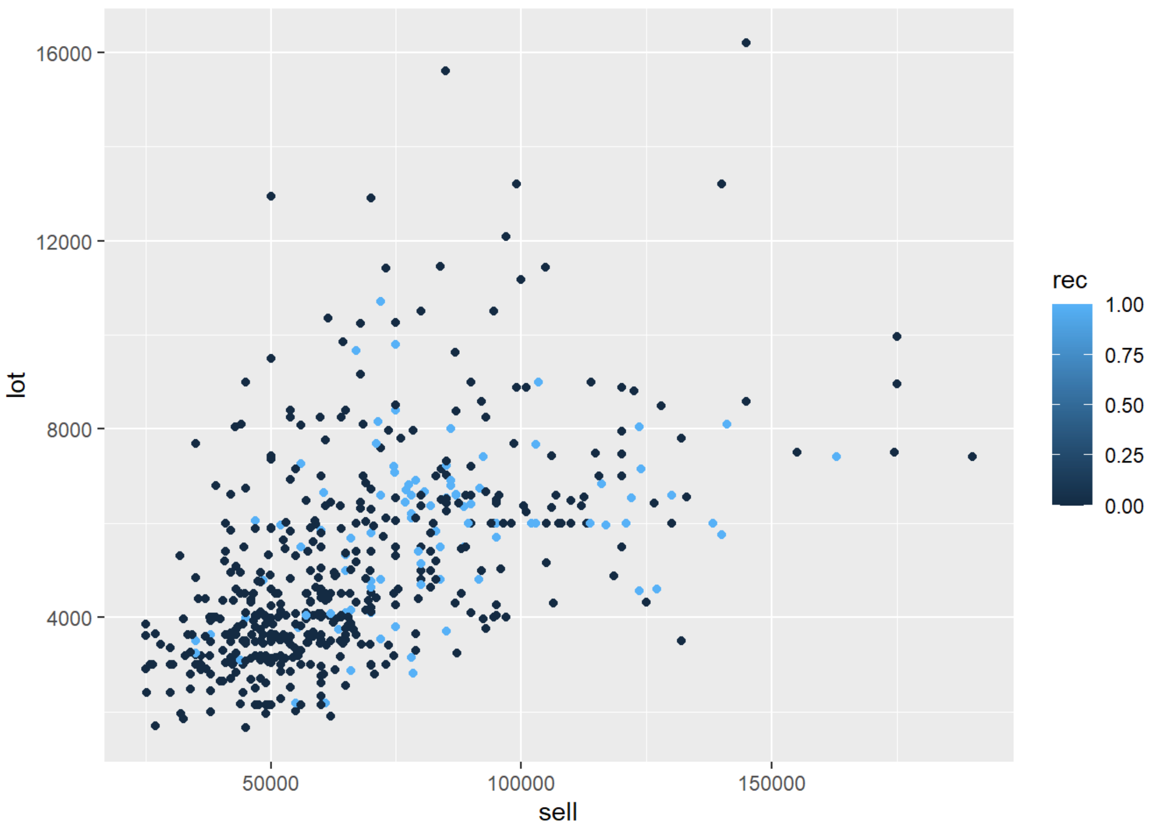
After loading the data, we perform Exploratory Data Analysis to get these plots:

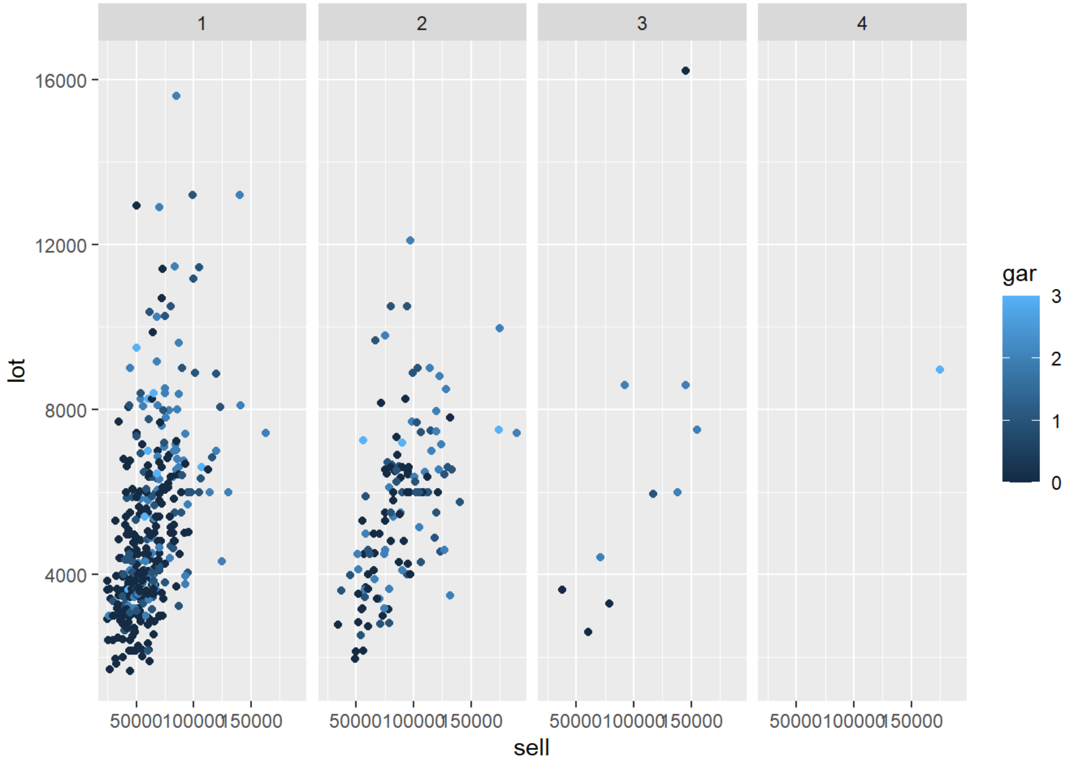





16000
12000
8000
4000
lot
50000
100000
150000
sell
fb
4
3
2
1


16000
12000
8000
4000
lot
50000
100000
150000
sell
bdms
6
5
4
3
2
1


16000
12000
8000
4000
lot
50000
100000
150000
sell
rec
1.00
0.75
0.50
0.25
0.00


1
2
3
4
16000
12000
8000
4000
lot
50000100000150000
50000100000150000
50000100000150000
50000100000150000
sell
gar
3
2
1
0

Then we need to fit the first model.

Model characteristics: R2 = 0.6731 , F-statistic: 99.97 on 11 and 534 DF

Model Linearity testing:

Ramsey's RESET (linearity) testing

```
##
##  RESET test
##
## data:  fit
## RESET = 26.986, df1 = 1, df2 = 533, p-value = 2.922e-07
```

With a statistic of 26.986 and a p-value of ~0.000, the Ramsey's RESET test suggests that the linear model is NOT correctly specified. So we reject Ho.

Jarque-Bera

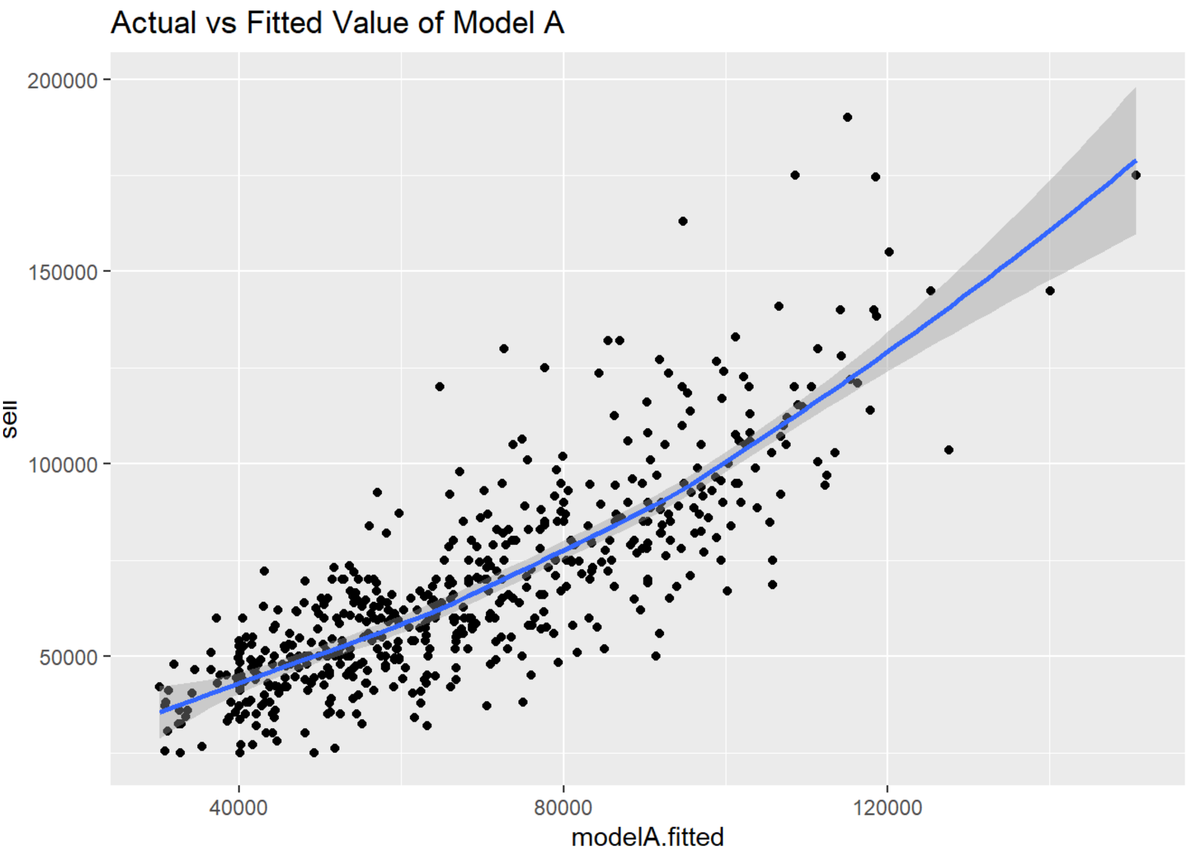
```
##
##  Jarque Bera Test
##
## data:  fit$residuals
## X-squared = 247.62, df = 2, p-value < 2.2e-16
##
##
##  Skewness
##
## data:  fit$residuals
## statistic = 0.85278, p-value = 4.119e-16
##
##
##  Kurtosis
##
## data:  fit$residuals
## statistic = 5.8241, p-value < 2.2e-16
```

With a statistic of ~247.62 and a p-value of ~0, the Jarque-Bera test suggests that the linear model residuals are *NOT* normally distributed, therefore the linear model is

NOT correctly specified. Also is not distributed as normal distribution, and its distribution is leptokurtic.

Both Ramsey's RESET and Jarque-Bera tests suggest that the considered linear model is NOT correctly specified.

Real to fitted-values diagram



b) Now consider a linear model where the log of the sale price of the house is the dependent variable and the explanatory variables are as before. Perform again the test for linearity. What do you conclude now?

Answer)

Second model estimation – Logarithmic

Model characteristics: R2 = 0.6766 , F-statistic: 101.56 on 11 and 534 DF

Model linearity testing:

Ramsey's RESET

```
## RESET = 0.27031, df1 = 1, df2 = 533, p-value = 0.6033
```

With a statistic of ~0.27 and a p-value of ~0.6033, the Ramsey's RESET test suggests that the second linear model might be correctly specified. We accept Ho, at the 5% level of significance).

Jarque-Bera

```
##
##  Jarque Bera Test
##
## data:  fit2$residuals
## X-squared = 8.4432, df = 2, p-value = 0.01467
##
##
##  Skewness
##
## data:  fit2$residuals
## statistic = 0.19898, p-value = 0.05768
##
##
##  Kurtosis
##
## data:  fit2$residuals
## statistic = 3.4613, p-value = 0.0278
```

Also with this model, With a statistic of ~8.443 and a p-value of ~0.0147, the Jarque-Bera test suggests that the linear model residuals are still NOT normally distributed.

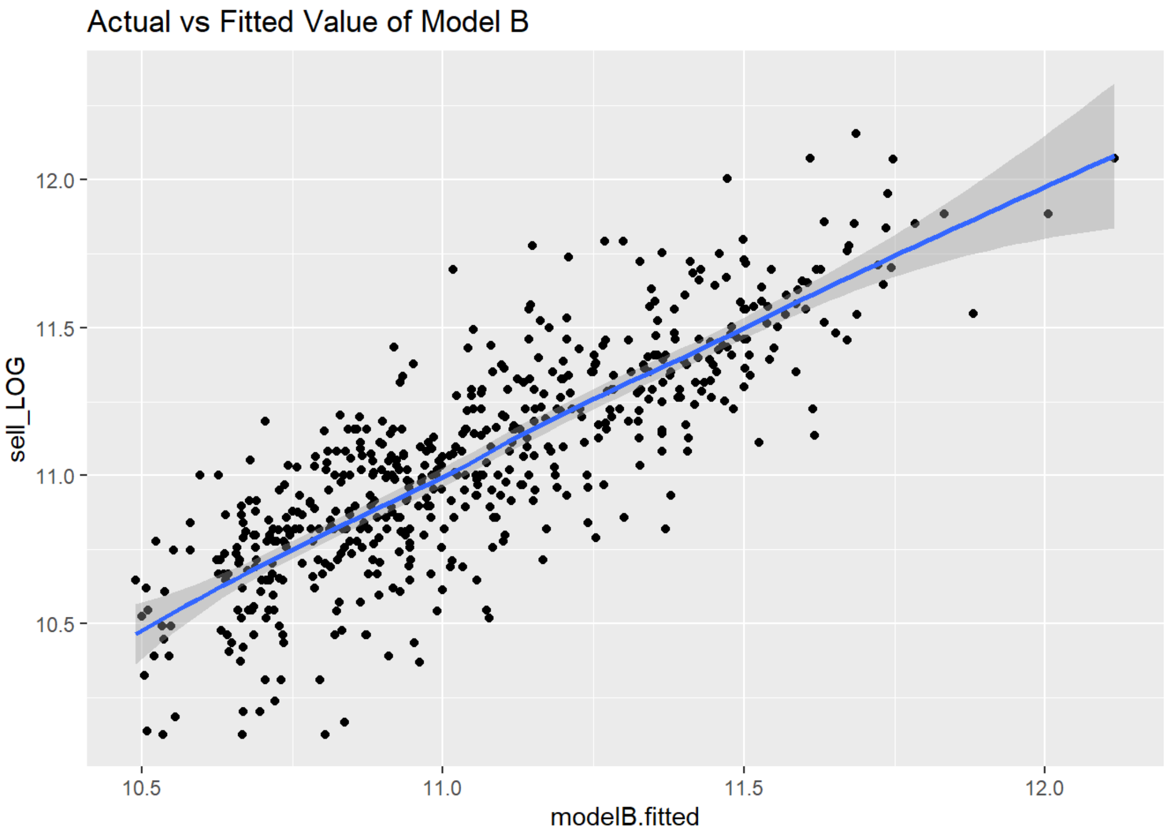
Therefore the linear model is still NOT correctly specified, although the second model's JB statistic is significantly decreased (and therefore the model significantly improved).

Both Ramsey's RESET and Jarque-Bera tests suggest that the second model is significantly improved than the model considered first, but has a problem:

The Ramsey's RESET test suggests that the second linear model might be correctly specified.

while the Jarque-Bera test suggests that it is still NOT correctly specified (although significantly improved).

Real to fitted-values diagram



c) Continue with the linear model from question (b). Estimate a model that includes both the lot size variable and its logarithm, as well as all other explanatory variables without transformation. What is your conclusion, should we include lot size itself or its logarithm?

Answer)

Model characteristics: R2 = 0.687 , F-statistic: 97.51 on 12 and 533 DF

Model linearity testing:

Ramsey's RESET

```
## RESET = 0.06769, df1 = 1, df2 = 532, p-value = 0.7948
```

With a statistic of ~0.068 and a p-value of ~0.7948, the Ramsey's RESET test suggests that the third linear model might be correctly specified We accept Ho, at the 5% level of significance).

It also suggests that this is the best model constructed so far, as it has the lowest statistic and the highest p-value scored by all Ramsey's RESET tests ran so far.

Jarque-Bera

```
##  Jarque Bera Test
##
## data:  fit3$residuals
## X-squared = 9.3643, df = 2, p-value = 0.009259
##
##
##  Skewness
##
## data:  fit3$residuals
## statistic = 0.18025, p-value = 0.08553
##
##
##  Kurtosis
##
## data:  fit3$residuals
## statistic = 3.5307, p-value = 0.01136
```

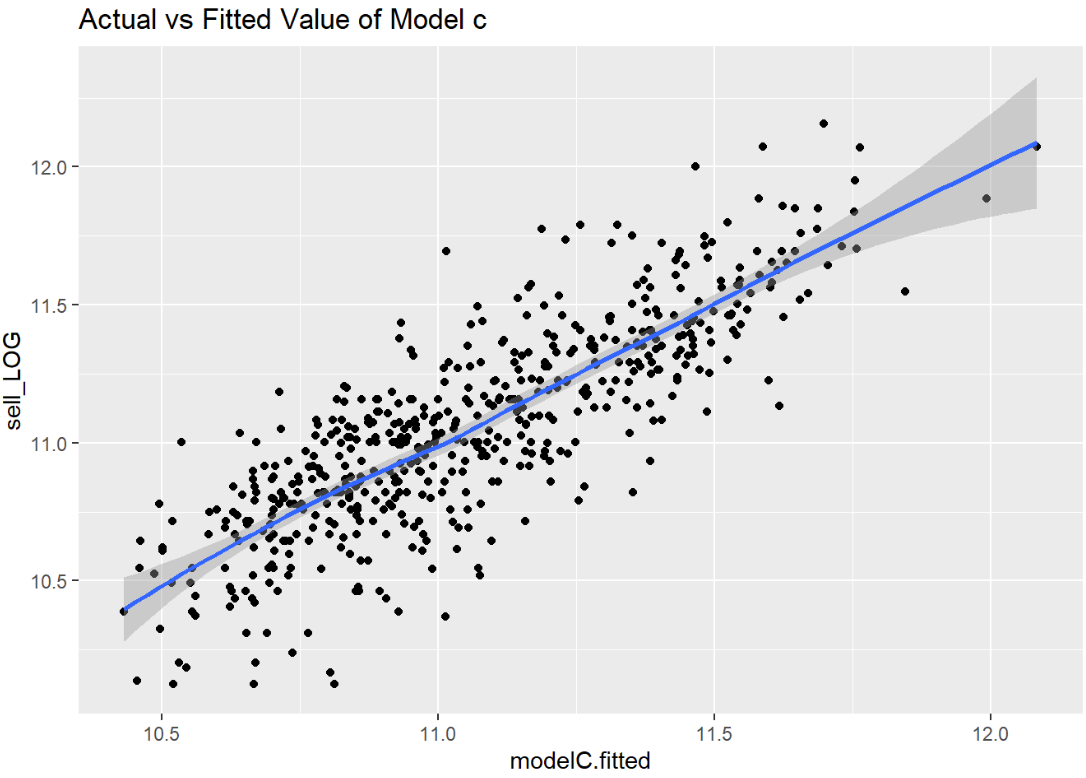
With a statistic of ~9.364 and a p-value of ~0.0093, the Jarque-Bera test suggests that the linear model residuals are still NOT normally distributed; therefore the linear model is still NOT correctly specified.

No further model improvement is indicated by the Jarque-Bera residuals normality test; in fact the second model's residuals were slightly more normal than the third's.

Both Ramsey's RESET and Jarque-Bera tests suggest that the third model is significantly improved than the model considered first, while the Ramsey's RESET test suggests that it is even more improved than the model considered second.But the problems are:

- The Ramsey's RESET test suggests that the third linear model might be correctly specified.
- while the Jarque-Bera test suggests that it is still NOT correctly specified.

Real to fitted-values diagram



It is concluded that it would be better to include the lot size logarithm in the model, rather than the lot size variable itself, due to the following reasons:

- The three models testing performed so far, see Table 4 (above): “Models linearity test results’ comparison chart”. The Ramsey’s RESET tests showed that the lot size logarithm variable significantly improves the model linearity, while the Jarque-Bera tests showed that it produces a satisfactory (so far) level of residuals normality.
- The (much better) lot size logarithm variable coefficient p-value (0), compared to the lot size variable itself coefficient p-value (0.359), when used together. See Table 5 (above): “Third model lot related variables’ coefficients’ comparison chart”.
- The fact that lot variable ended with a ~zero (0) coefficient anyway at the (improved) second and third models.

d) Consider now a model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables as before. We now consider interaction effects of the log lot size with the other variables. Construct these interaction variables. How many are individually significant?

Answer)

Model characteristics: R2 = 0.6951 , F-statistic: 56.89 on 21 and 524 DF, without "Lot" variable. Notice also the ten (10) interaction variables introduction, between the log lot size and each one of all other variables.

Model linearity testing:

Ramsey's RESET

```
## RESET = 0.011571, df1 = 1, df2 = 523, p-value = 0.9144
```

With a statistic of ~0.012 and a p-value of ~0.9144, the Ramsey's RESET test suggests that the fourth model might be correctly specified, We accept Ho at the 5% level of significance.

It also suggests that this is the best model constructed so far, as it has an even lower statistic and the highest p-value scored by all Ramsey's RESET tests ran so far.

Jarque-Bera

```
##  Jarque Bera Test
##
## data:  fit4$residuals
## X-squared = 8.2029, df = 2, p-value = 0.01655
##
##
##  Skewness
##
## data:  fit4$residuals
## statistic = 0.17281, p-value = 0.09924
##
##
##  Kurtosis
##
```

```
## data:  fit4$residuals

## statistic = 3.491, p-value = 0.01918
```

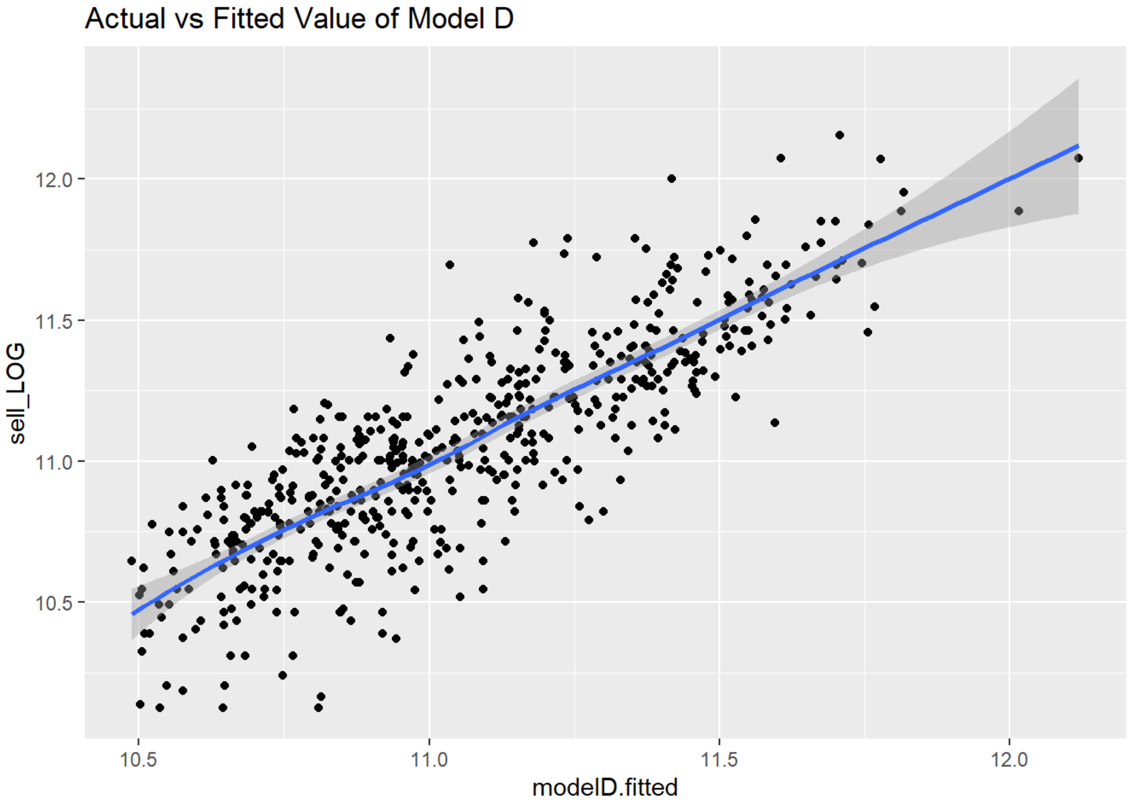
With a statistic of ~8.203 and a p-value of ~0.0165, the Jarque-Bera test suggests that the model residuals are still NOT normally distributed; therefore the model is still NOT correctly specified.

This Jarque-Bera test result, however, is the best scored so far. It seems that the interaction variables introduction slightly improves the (previous best) second model residuals normality.

Both Ramsey’s RESET and Jarque-Bera tests suggest that the fourth model is significantly improved than the models previously considered.But continues with the problems:

- The Ramsey’s RESET test suggests that the fourth model might be correctly specified
- While the Jarque-Bera test suggests that it is still NOT correctly specified.

Real to fitted-values diagram



Using the 5% significance level, only two (2) of the ten (10) interaction variables used are individually significant:

- LOG(lot)·drv
- LOG(lot)·rec

e) Perform an F-test for the joint significance of the interaction effects from question (d).

Answer)

The corresponding restricted model (fifth model) will use only the two (2) significant interaction variables, as identified at the previous section (d):

- LOG(lot)·drv
- LOG(lot)·rec

It will also use the rest of the fourth model variables (as is)

Fifth model estimation

Model characteristics: R2 = 0.6916 , F-statistic: 91.79 on 13 and 532 DF

Model linearity testing:

Ramsey’s RESET

```
## RESET = 0.045677, df1 = 1, df2 = 531, p-value = 0.8308
```

With a statistic of ~0.046 and a p-value of ~0.8308, the Ramsey’s RESET test suggests that the fifth model might be correctly specified (H0 of correct/linear specification NOT rejected, at the 5% level of significance).

It also suggests that this is not the best model constructed so far, as the fourth model had scored an even lower statistic and a higher p-value than this.

Jarque-Bera

```
##  Jarque Bera Test
##
## data:  fit5$residuals
## X-squared = 9.2372, df = 2, p-value = 0.009866
##
##
##  Skewness
##
## data:  fit5$residuals
## statistic = 0.18929, p-value = 0.07096
```

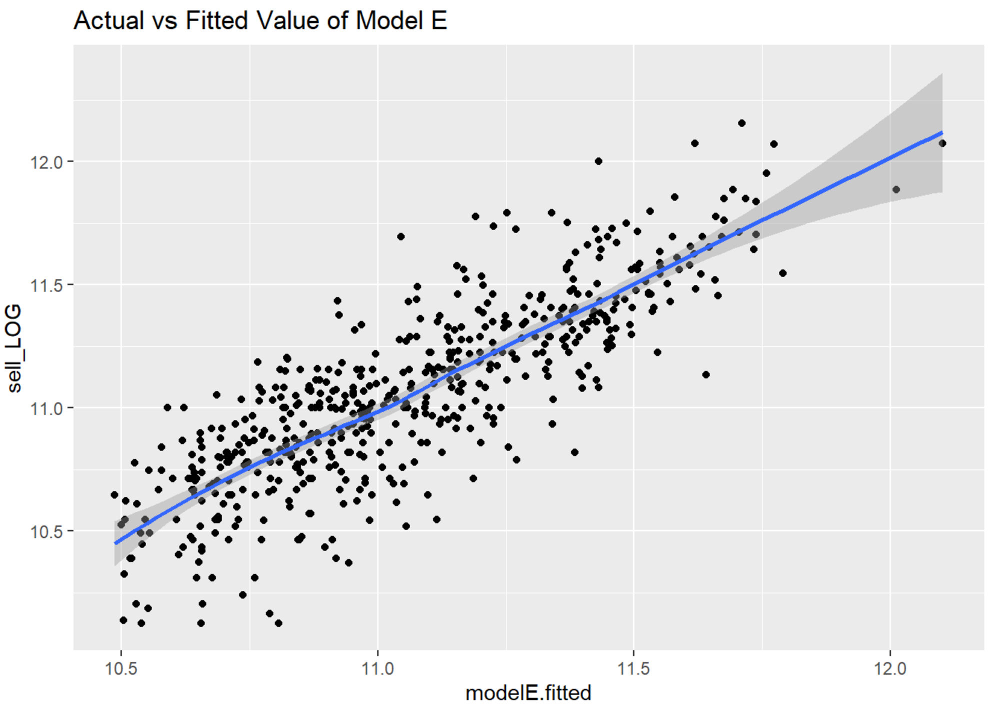
```
##
##
##   Kurtosis
##
## data:  fit5$residuals
## statistic = 3.5125, p-value = 0.0145
```

With a statistic of ~9.237 and a p-value of ~0.0099, the Jarque-Bera test suggests that the model residuals are still NOT normally distributed; therefore the model is still NOT correctly specified.

This Jarque-Bera test result is not the best scored so far either, as the fourth model related test had indicated an even better residuals normality.

Both Ramsey’s RESET and Jarque-Bera tests suggest that the fourth model was better than this fifth one (restricted model); which, however, seems to be the second best (according to Ramsey’s RESET testing) or the third best (according to Jarque-Bera testing) model constructed so far, but with the same problems before showed.

Real to fitted-values diagram



Using SSR

An interaction effects joint significance F-test can be performed either:

- using the sum of square residuals (SSR) of the restricted and the unrestricted models (p.valueforF~F(10,524)=0.1833), or
- using the R2 of both the restricted and the unrestricted models. (p.valueforF~F(10,524)=0.1948)
- Both SSR and R2 methods produced an F-test statistic of 0.6–0.61, with a p-value of 0.18–0.19.

The above results conclude that interactions are jointly significant at the 5% significance level.

f) Now perform model specification on the interaction variables using the general-to-specific approach. (Only eliminate the interaction effects.)

General-to-specific model:

- Variables elimination after regression, one at a time, produced the following results:
- After regression round #1: LOG(lot)·reg interaction variable was chosen to be removed.
- After regression round #2: LOG(lot)·bdms interaction variable was chosen to be removed.
- After regression round #3: LOG(lot)·ffin interaction variable was chosen to be removed.
- After regression round #4: LOG(lot)·ghw interaction variable was chosen to be removed.
- After regression round #5: LOG(lot)·ca interaction variable was chosen to be removed.
- After regression round #6: LOG(lot)·gar interaction variable was chosen to be removed.
- After regression round #7: LOG(lot)·fb interaction variable was chosen to be removed.
- After regression round #8: LOG(lot)·sty interaction variable was chosen to be removed.
- After regression round #9: LOG(lot)·drv interaction variable was chosen to be removed.
- After regression round #10: all remaining variables were found to be significant; variables removal stops here. Conclusively, the only interaction variable found to be significant is LOG(lot)·rec.

```
## Residual standard error: 0.2096 on 533 degrees of freedom
## Multiple R-squared:  0.6894, Adjusted R-squared:  0.6824
## F-statistic: 98.59 on 12 and 533 DF,  p-value: < 2.2e-16
```

Model linearity testing:

Ramsey's RESET

```
## RESET = 0.43102, df1 = 1, df2 = 532, p-value = 0.5118
```

With a statistic of ~0.431 and a p-value of ~0.5118, the Ramsey's RESET test suggests that the sixth model might be correctly specified (H0 of correct/linear specification NOT rejected, at the 5% level of significance).

It also suggests that this is far from a linear model, as all the previous models (except from the first) had scored a lower statistic and a higher p-value than this.

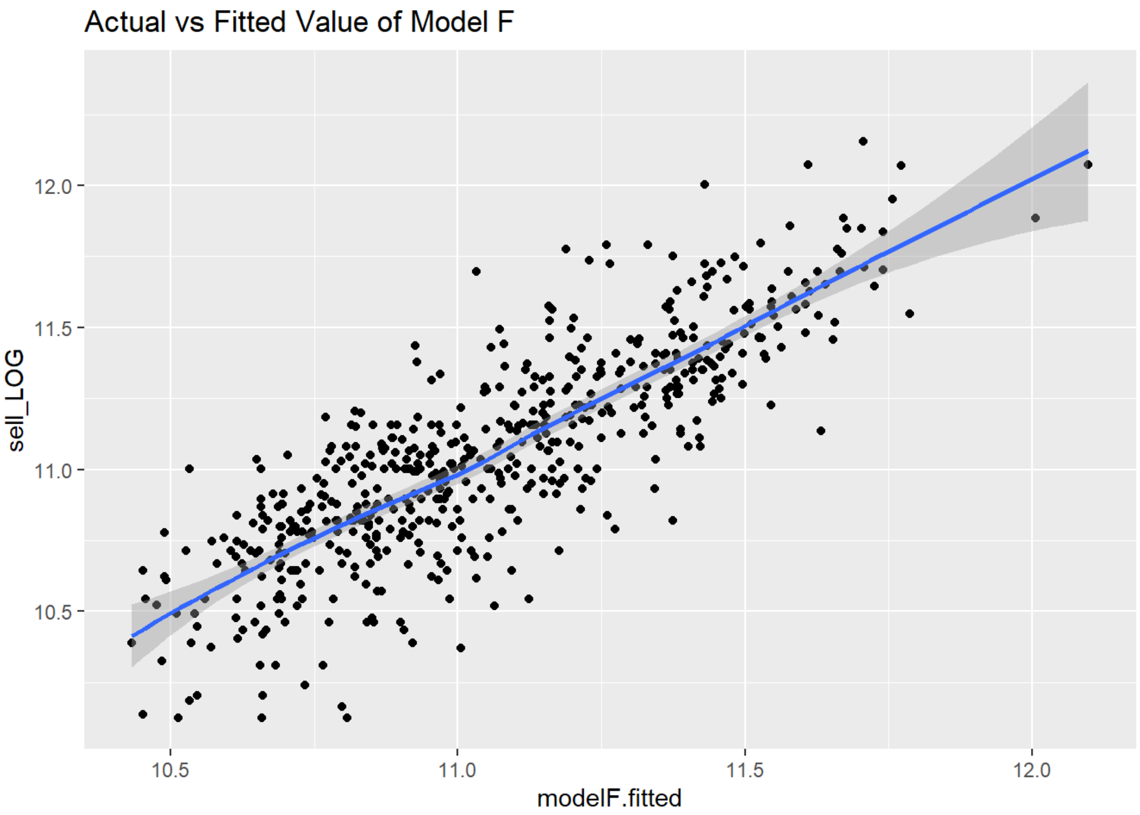
Jarque-Bera

```
##  Jarque Bera Test
##
## data:  fit6$residuals
## X-squared = 10.348, df = 2, p-value = 0.005661
##
##
##  Skewness
##
## data:  fit6$residuals
## statistic = 0.18972, p-value = 0.07032
##
##
##  Kurtosis
##
## data:  fit6$residuals
## statistic = 3.5576, p-value = 0.007826
```

With a statistic of ~10.348 and a p-value of ~0.0057, the Jarque-Bera test suggests that the model residuals are still NOT normally distributed; therefore the model is still NOT correctly specified.

This Jarque-Bera test result is not the best scored so far. All the previous models (except from the first) related test had indicated an even better residuals normality.

Real to fitted-values diagram



Both Ramsey's RESET and Jarque-Bera tests suggest that the sixth model is less linear and with less residuals normality than the models tested before (except from the first model).

The Ramsey's RESET test suggests that the sixth model might be correctly specified, while the Jarque-Bera test suggests that it is still NOT correctly specified.

g) One may argue that some of the explanatory variables are endogenous and that there may be omitted variables. For example, the 'condition' of the house in terms of how it is maintained is not a variable (and difficult to measure) but will affect the house price. It will also affect, or be reflected in, some of the other variables, such as whether the house has an air conditioning (which is mostly in newer houses). If the condition of the house is missing,

will the effect of air conditioning on the (log of the) sale price be over- or underestimated? (For this question no computer calculations are required.)

Answer)

We know that only newer houses are typically equipped with air conditioning (captured by the ca variable), and that newer homes—due to their better condition—tend to sell for higher prices. However, because we do not have a separate variable representing the **condition** or **age** of the house in our model, their positive effects on the sale price are indirectly captured by the ca variable.

As a result, the estimated impact of air conditioning on the logarithm of the sale price (LOG(sell)) is **overstated**. This is because ca is absorbing not only the effect of air conditioning itself but also the unobserved positive effects of newer age and better condition. Therefore, in the absence of a condition or age variable, the ca coefficient in the model reflects more than just the value of air conditioning—it partially reflects omitted variables, leading to **omitted variable bias** and an **overestimation** of the air conditioning effect.

h) Finally we analyze the predictive ability of the model. Consider again the model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables in their original form (and no interaction effects). Estimate the parameters of the model using the first 400 observations. Make predictions on the log of the price and calculate the MAE for the other 146 observations. How good is the predictive power of the model (relative to the variability in the log of the price)?

Answer)

We separate the data sample into two groups.

Dataset 1:

```
##       obs              sell             lot             bdms      
##  Min.   :  1.0   Min.   : 25000   Min.   : 1650   Min.   :1.00  
##  1st Qu.:100.8   1st Qu.: 46150   1st Qu.: 3495   1st Qu.:2.00  
##  Median :200.5   Median : 59250   Median : 4180   Median :3.00  
##  Mean   :200.5   Mean   : 64977   Mean   : 4905   Mean   :2.95  
##  3rd Qu.:300.2   3rd Qu.: 78000   3rd Qu.: 6000   3rd Qu.:3.00  
##  Max.   :400.0   Max.   :190000   Max.   :16200   Max.   :6.00  
##        fb             sty             drv              rec        
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :1.000   Median :2.000   Median :1.0000   Median :0.0000  
##  Mean   :1.278   Mean   :1.718   Mean   :0.8125   Mean   :0.1625  
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.0000  
##       ffin              ghw              ca              gar        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.00   Median :0.000   Median :0.0000  
##  Mean   :0.3475   Mean   :0.05   Mean   :0.285   Mean   :0.6925  
##  3rd Qu.:1.0000   3rd Qu.:0.00   3rd Qu.:1.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.00   Max.   :1.000   Max.   :3.0000  
##       reg             sell_LOG         lot_LOG     
##  Min.   :0.000   Min.   :10.13   Min.   :7.409  
##  1st Qu.:0.000   1st Qu.:10.74   1st Qu.:8.159  
##  Median :0.000   Median :10.99   Median :8.338  
##  Mean   :0.105   Mean   :11.01   Mean   :8.420  
##  3rd Qu.:0.000   3rd Qu.:11.26   3rd Qu.:8.700  
##  Max.   :1.000   Max.   :12.15   Max.   :9.693  
```

Dataset 2:

```
##       obs              sell             lot              bdms      
##  Min.   :401.0   Min.   : 31900   Min.   : 1950   Min.   :2.000  
##  1st Qu.:437.2   1st Qu.: 60000   1st Qu.: 4678   1st Qu.:3.000  
##  Median :473.5   Median : 72750   Median : 6000   Median :3.000  
##  Mean   :473.5   Mean   : 76737   Mean   : 5821   Mean   :3.007  
##  3rd Qu.:509.8   3rd Qu.: 91125   3rd Qu.: 6652   3rd Qu.:3.000  
##  Max.   :546.0   Max.   :174500   Max.   :12944   Max.   :5.000  
##        fb             sty             drv              rec        
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :1.000   Median :2.000   Median :1.0000   Median :0.0000  
##  Mean   :1.308   Mean   :2.055   Mean   :0.9863   Mean   :0.2192
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :2.000   Max.   :4.000   Max.   :1.0000   Max.   :1.0000
##       ffin             ghw                 ca              gar
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.3562   Mean   :0.03425   Mean   :0.4041   Mean   :0.6918
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :3.0000
##       reg           sell_LOG        lot_LOG
##  Min.   :0.000   Min.   :10.37   Min.   :7.576
##  1st Qu.:0.000   1st Qu.:11.00   1st Qu.:8.451
##  Median :1.000   Median :11.19   Median :8.700
##  Mean   :0.589   Mean   :11.21   Mean   :8.595
##  3rd Qu.:1.000   3rd Qu.:11.42   3rd Qu.:8.803
##  Max.   :1.000   Max.   :12.07   Max.   :9.468
```

Seventh model estimation

Model characteristics: R2 = 0.6705 , F-statistic: 71.77 on 11 and 388 DF

Model linearity testing:

Ramsey’s RESET

```
## RESET = 0.03955, df1 = 1, df2 = 387, p-value = 0.8425
```

With a statistic of ~0.04 and a p-value of ~0.8425, the Ramsey’s RESET test suggests that the seventh model might be correctly specified (H0 of correct/linear specification NOT rejected, at the 5% level of significance).

It also suggests that this is not the best model constructed so far, as the fourth model had scored an even lower statistic and a higher p-value than this; this seventh model scored the second best Ramsey’s RESET test score.

Jarque-Bera

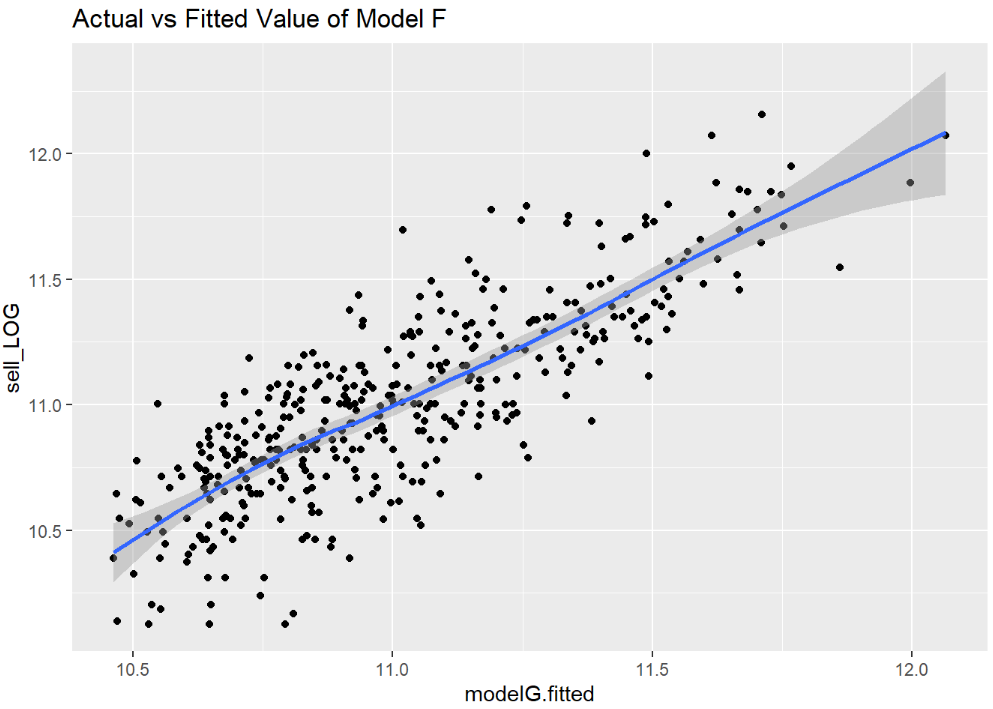
```
##  Jarque Bera Test
##
## data:  fit7$residuals
## X-squared = 0.69757, df = 2, p-value = 0.7055
```

```
##
##
##   Skewness
##
## data:  fit7$residuals
## statistic = 0.08868, p-value = 0.469
##
##
##   Kurtosis
##
## data:  fit7$residuals
## statistic = 3.102, p-value = 0.6772
```

With a statistic of ~0.698 and a p-value of ~0.7055, the Jarque-Bera test suggests that the model residuals are normally distributed; therefore the model is considered correctly specified.

This Jarque-Bera test result is the best scored so far, and indicates a sufficient residuals normality.

Real to fitted-values diagram



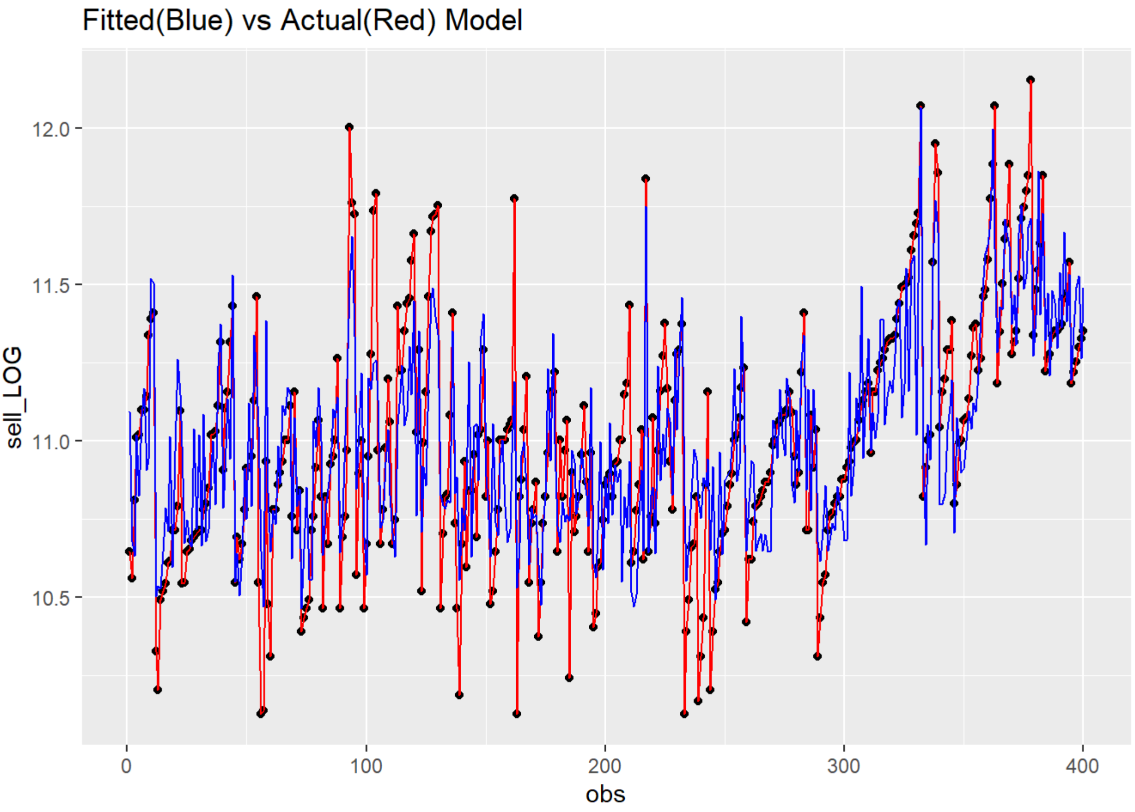
Both Ramsey's RESET and Jarque-Bera tests suggest that the seventh model is sufficiently linear and with good residuals normality.

Both Ramsey's RESET and Jarque-Bera tests suggest that the seventh model might be correctly specified.

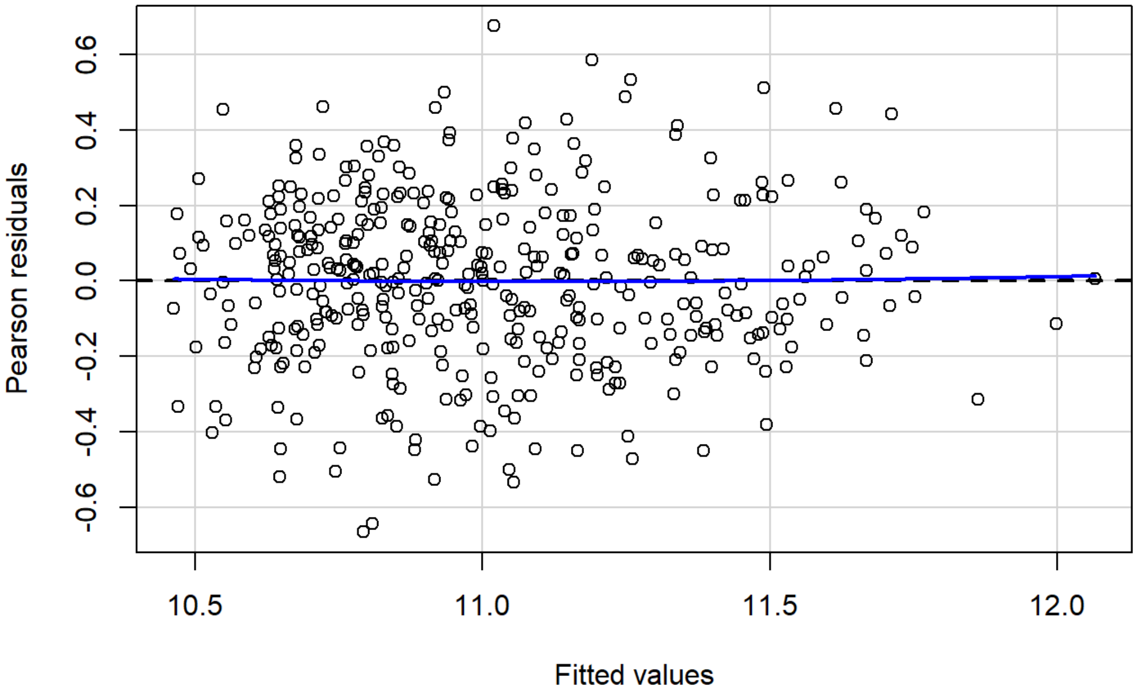
This is also intuitively demonstrated by the seventh model real to fitted-values diagram shown at the next page (looks about the same or more like a linear relationship).

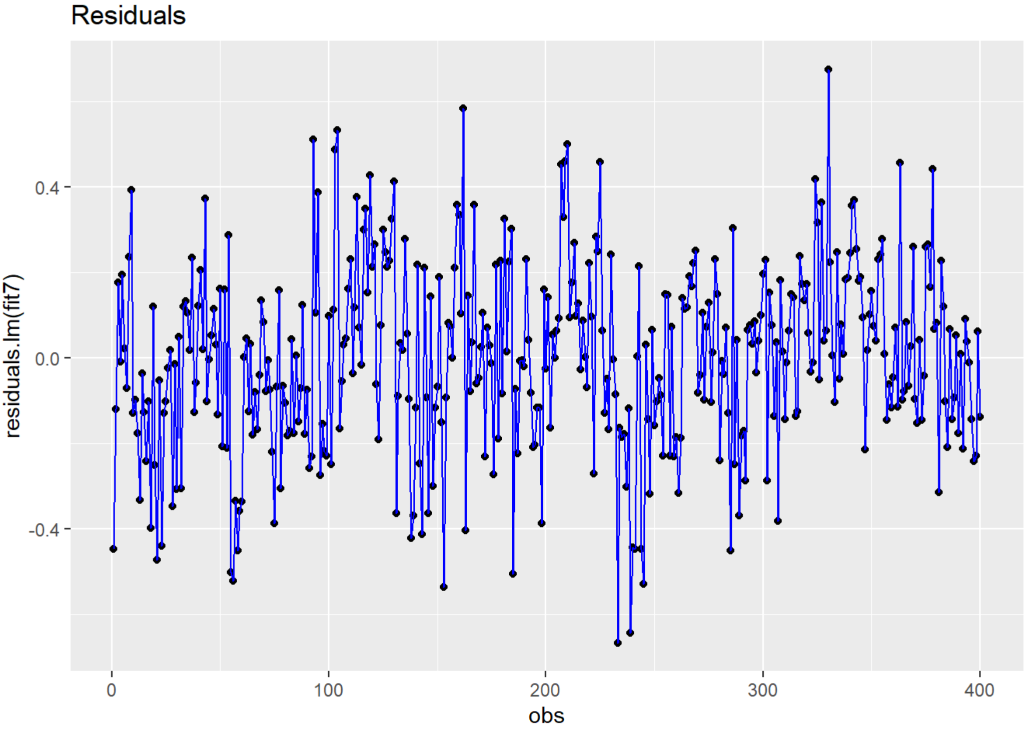
Model predictive ability

The seventh model, as estimated using the first data group, produced the following LOG(lot)^ values on the second data group:



Residual Plots:





Calculated MAE value of 0.128 is much less than the dependent variable standard deviation half, which leads to the conclusion that the model has some/significant predictive ability.