Homework Assignment 5

Shrusti Ghela

February 24, 2022

Data: "Sales_sample.csv".

The data are a random sample of size 1000 from the "Sales" data (after removing observations with missing values).

Variables:

LAST_SALE_PRICE: the sale price of the home SQFT: area of the house (sq. ft.) LOT_SIZE: area of the lot (sq. ft.) BEDS: number of bedrooms BATHS: number of bathrooms

1.1. Fit a linear regression model (Model 1) with sale price as response variable and SQFT, LOT_SIZE, BEDS, and BATHS as predictor variables. Add the fitted values and the residuals from the models as new variables in your data set. Show the R code you used for this question.

```
model_1 <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data=sales_data)

summary(model_1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,
##     data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1364578  -166436    -9884   122468  2964364
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5982.604  40023.271   0.149 0.881207
## SQFT           224.502     14.794  15.175  < 2e-16 ***
## LOT_SIZE         6.844      1.858   3.684 0.000242 ***
## BEDS        -60884.742  14461.536  -4.210 2.78e-05 ***
## BATHS       178177.446  17107.532  10.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322100 on 995 degrees of freedom
## Multiple R-squared:  0.4691, Adjusted R-squared:  0.467
## F-statistic: 219.8 on 4 and 995 DF,  p-value: < 2.2e-16
```

```r
names(model_1)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```r
model_1$coefficients
```

```
##   (Intercept)          SQFT      LOT_SIZE          BEDS         BATHS
##   5982.604259    224.502066      6.844143 -60884.742104 178177.446061
```

```r
sales_data$fitted_val_m1 <- model_1$fitted.values
head(sales_data)
```

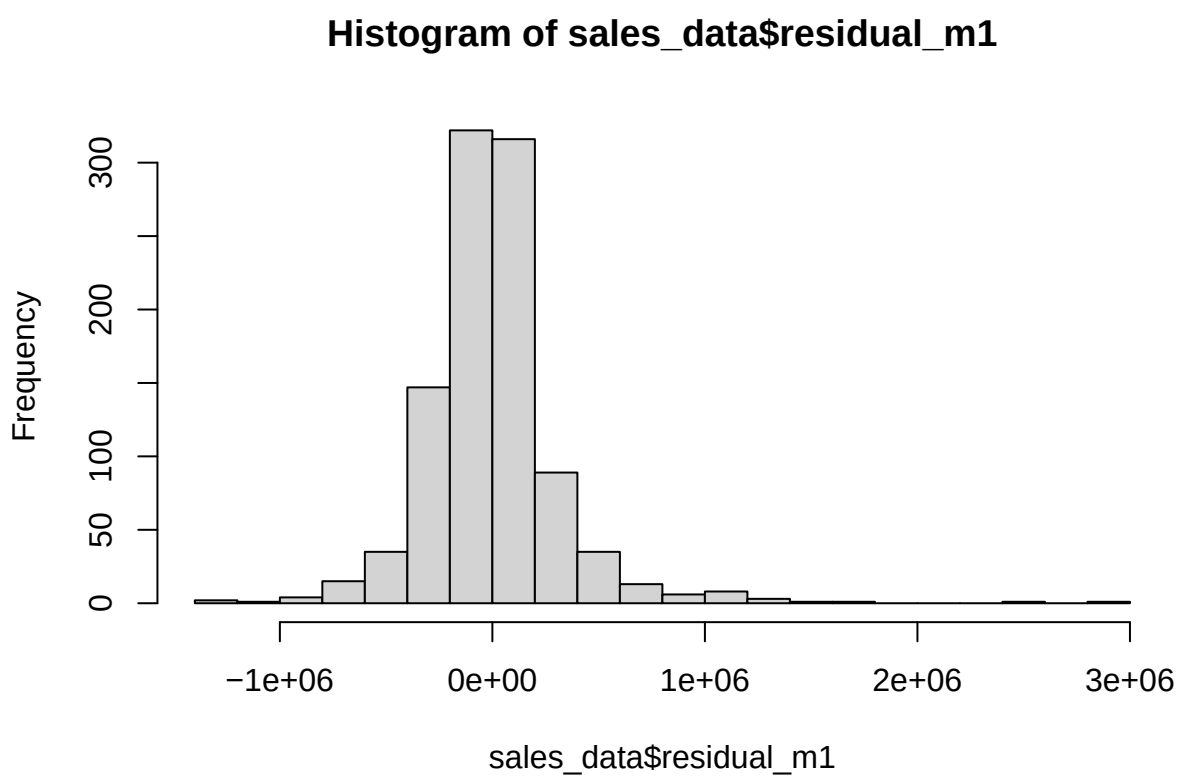
```
##   BEDS BATHS LOT_SIZE LAST_SALE_PRICE SQFT fitted_val_m1
## 1    4  2.50    22578          678000 2410      903464.3
## 2    4  2.00     4000          888000 2660      743350.6
## 3    4  2.25     5000          682000 2800      826169.4
## 4    3  2.00     6400         1600000 3790     1074348.6
## 5    6  2.50     7431          750000 2940      797012.7
## 6    4  1.75     7200          682000 2240      626416.6
```

```r
sales_data$residual_m1 <- sales_data$LAST_SALE_PRICE - sales_data$fitted_val_m1
head(sales_data)
```

```
##   BEDS BATHS LOT_SIZE LAST_SALE_PRICE SQFT fitted_val_m1 residual_m1
## 1    4  2.50    22578          678000 2410      903464.3  -225464.30
## 2    4  2.00     4000          888000 2660      743350.6   144649.40
## 3    4  2.25     5000          682000 2800      826169.4  -144169.39
## 4    3  2.00     6400         1600000 3790     1074348.6   525651.38
## 5    6  2.50     7431          750000 2940      797012.7   -47012.67
## 6    4  1.75     7200          682000 2240      626416.6    55583.37
```

1.2. Create a histogram of the residuals. Based on this graph does the normality assumption hold?

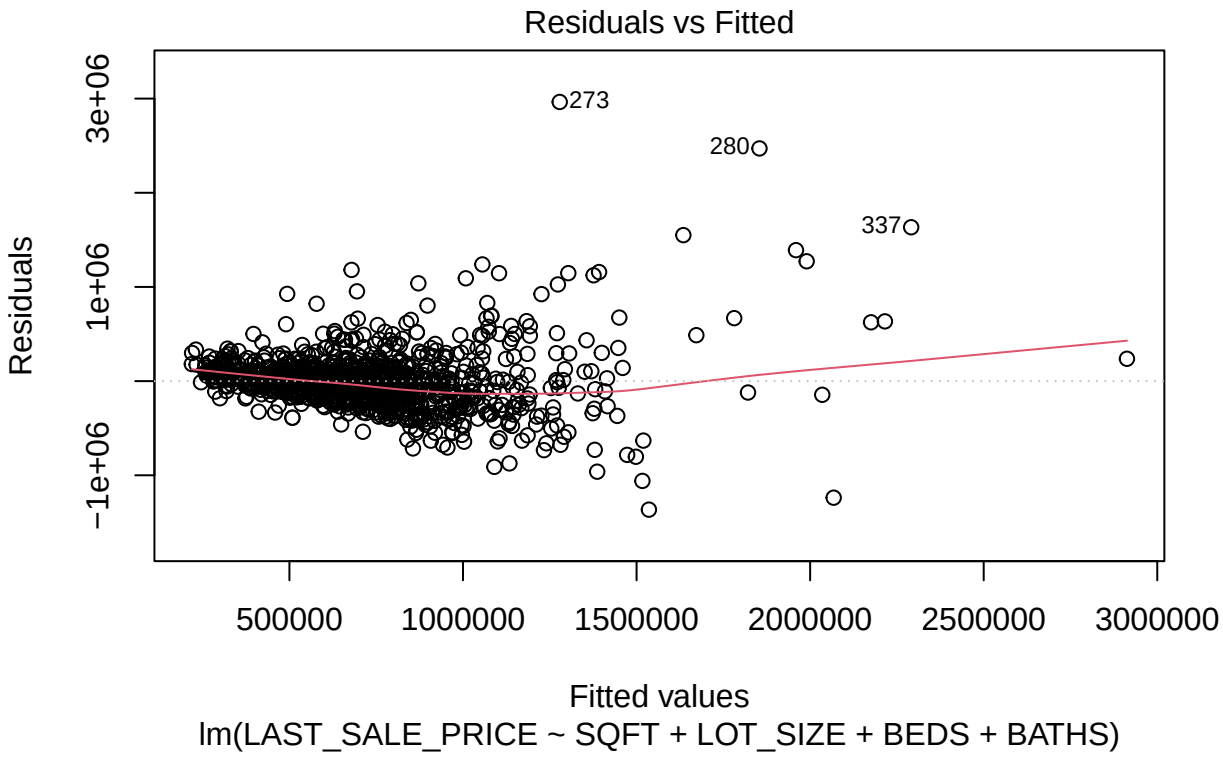
```r
hist(sales_data$residual_m1, breaks=25)
```

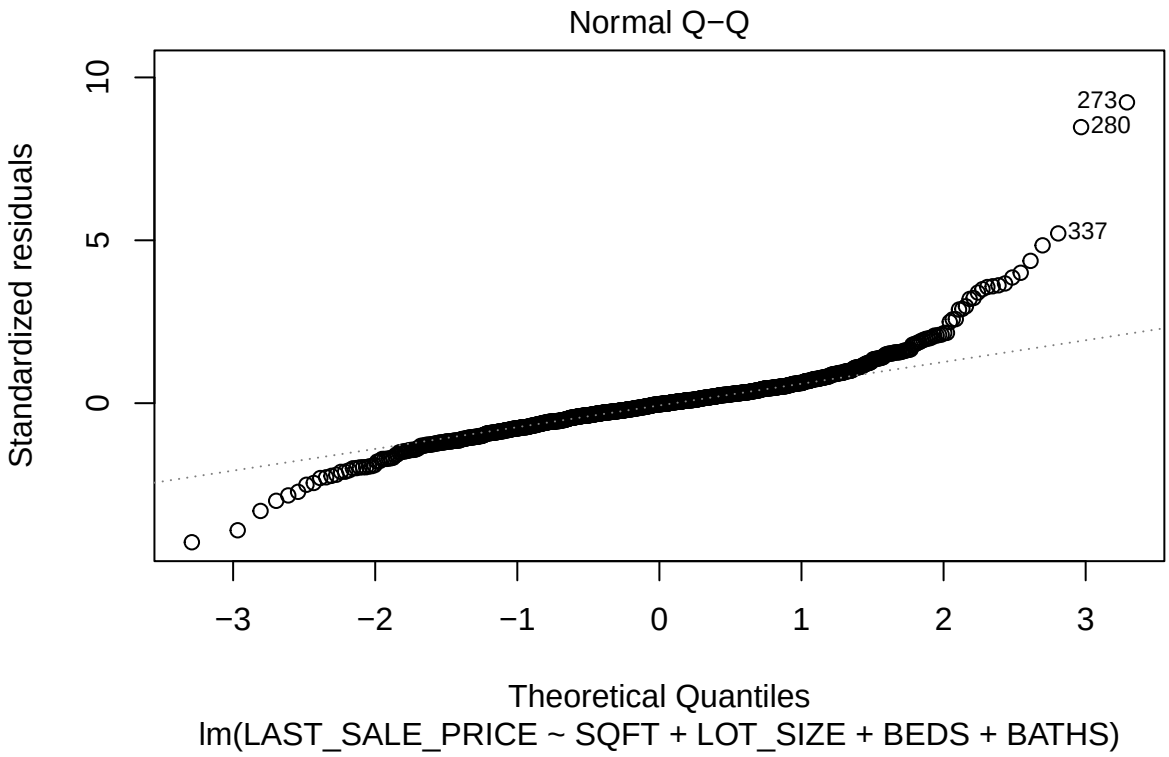


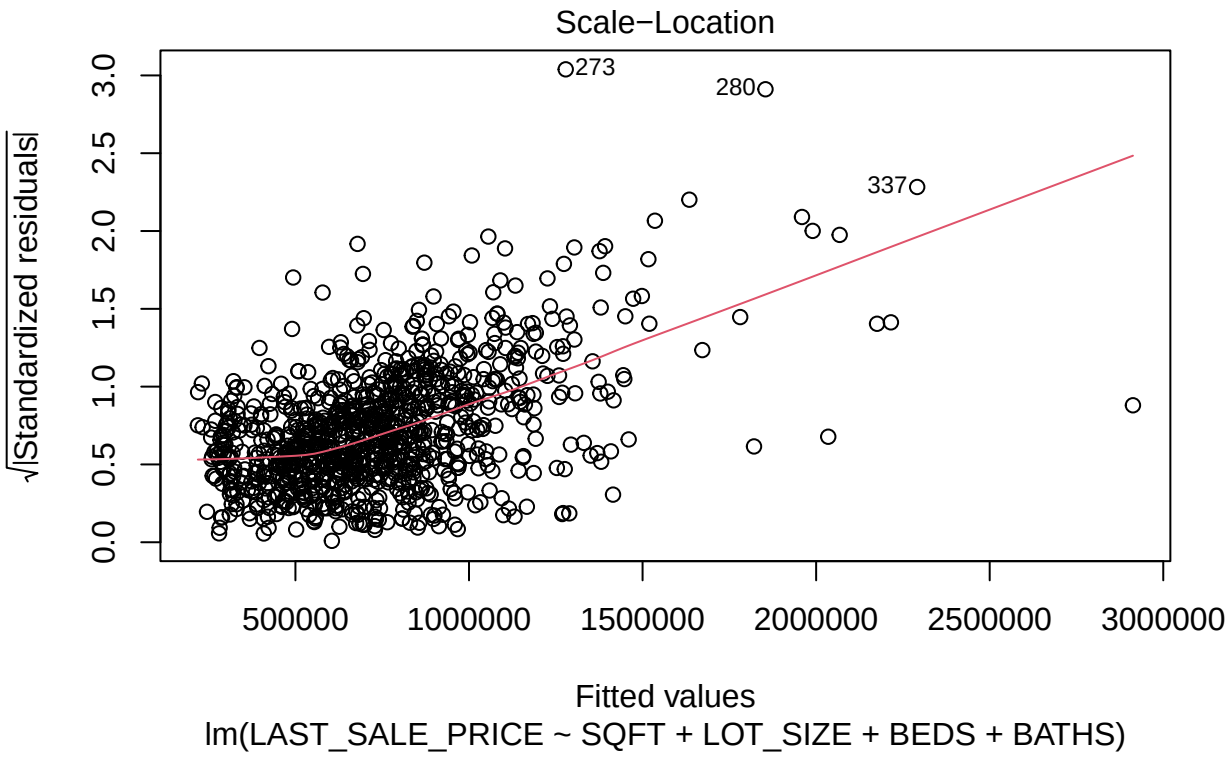
The histogram of the residuals from the Sales data looks fairly close to normal.(with exception of a few outliers)

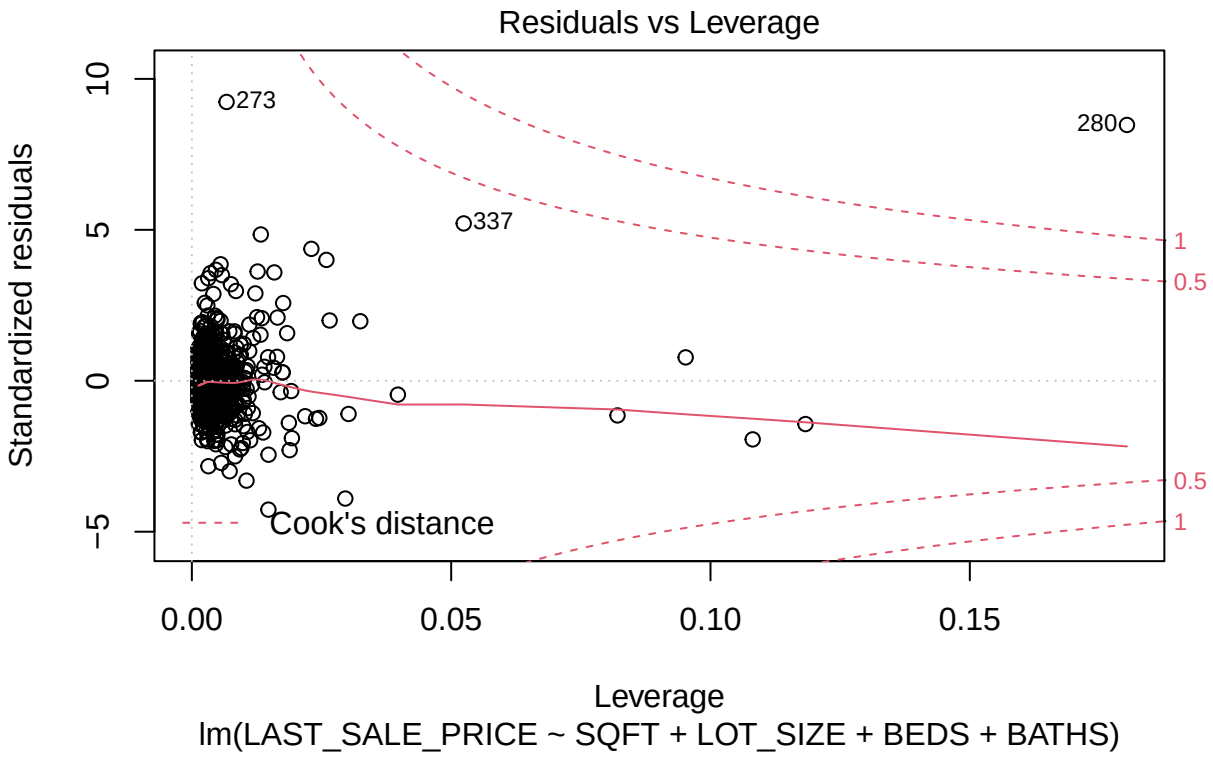
Answer the following questions using residual plots for the model. You may make the plots using the residuals and fitted variables added to your data set or you may use the 'plot' function. You do not need to display the plots in your submission.

```
plot(model_1)
```


Residuals vs Fitted
Residuals
3e+06
1e+06
−1e+06
273
280
337
500000
1000000
1500000
2000000
2500000
3000000
Fitted values
lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS)


Normal Q−Q
Standardized residuals
10
5
0
273
280
337
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS)


Scale−Location
273
280
337
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
500000
1000000
1500000
2000000
2500000
3000000
Fitted values
lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS)



1.3. Assess the linearity assumption of the regression model. Explain by describing a pattern in one or more residual plots.

The assumption of linearity means that the relationships between mean response and each predictor variable are linear. In terms of the model, this is stated as $\mathrm{E}(\epsilon_i) = 0$ for all i.

Linearity is necessary for linear regression, but should not be interpreted too strictly because it rarely if ever is exactly true.

To check for the linearity assumption, we observe the Residuals vs Fitted plot. From that plot, we can see that it is clustered towards lower sales price. There doesn't necessarily seem to be a pattern here, but also it doesn't look as random as its supposed to. So, I think the assumption of linearity is not met.

1.4. Assess the constant variance assumption of the regression model. Explain by describing a pattern in one or more residual plots.

The constant variance assumption is that the errors ϵ_i all have the same variance, i.e., $\mathrm{var}(\epsilon_i) = \sigma^2$ for some (usually unknown) σ^2. It is also known as *homoscedasticity*.

Non-constant variance can have a large effect on the performance of confidence intervals and hypothesis tests for regression coefficients. The effect can be to make inferences either overly conservative or anti-conservative.

Non-constant variance (like non-independence) is a problem *no matter how large the sample size*.

The plot of residuals against fitted values shows some evidence of non-constant variance - the residuals are more spread out for the higher sale price as compared to the lower sale price.

1.5. Assess the normality assumption of the linear regression model. Explain by describing a pattern in one or more residual plots.

The normality assumption is that the errors ϵ_i are normally distributed.

We use the residuals to check normality by applying histograms and q-q plots to the residuals.

In the q-q plot, We can see that the points lie mostly along the straight diagonal line with some deviations along each of the tails. Based on this plot, we could safely assume that this set of data is normally distributed.

Also, normality of the error distribution is only necessary if the sample sizes are not sufficiently large. Since we have sufficiently large sample size of data, normality holds irrespective.

1.6. Give an overall assessment of how well the assumptions hold for the regression model.

Overall, the linearity and constant variance assumption doesn't hold. But the normality assumption holds.

1.7. Would statistical inferences based on this model be valid? Explain.

The assumptions are needed to justify **statistical inference** for the regression coefficients. This includes confidence intervals as well as hypothesis tests for the coefficients.

It is important to note that the assumptions are not needed for fitting the model and using the results for purposes other than statistical inference. Thus, we can use linear regression models in a descriptive or exploratory fashion without worrying about assumptions, as long as we don't make inferential statements about the model or the parameters.

So, since all our assumptions are not met, the statistical inferences based on model 1 would not be valid.

1.8. Create a new variable (I will call it LOG_PRICE) which is calculated as the log-transformation of the sale price variable. Use base-10 logarithms. Fit a linear regression model (Model 2) with LOG_PRICE as response variable and SQFT, LOT_SIZE, BEDS, and BATHS as predictor variables. Report the table of coefficient estimates with standard errors and p-values.

```
sales_data$LOG_PRICE <- log10(sales_data$LAST_SALE_PRICE)
head(sales_data)
```

```
##   BEDS BATHS LOT_SIZE LAST_SALE_PRICE SQFT fitted_val_m1 residual_m1 LOG_PRICE
## 1    4  2.50    22578          678000 2410      903464.3  -225464.30  5.831230
## 2    4  2.00     4000          888000 2660      743350.6   144649.40  5.948413
## 3    4  2.25     5000          682000 2800      826169.4  -144169.39  5.833784
## 4    3  2.00     6400         1600000 3790     1074348.6   525651.38  6.204120
## 5    6  2.50     7431          750000 2940      797012.7   -47012.67  5.875061
## 6    4  1.75     7200          682000 2240      626416.6    55583.37  5.833784
```

```
model_2 <- lm(LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data=sales_data)
summary(model_2)
```

```
##
## Call:
## lm(formula = LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95365 -0.08261  0.00690  0.08986  0.71410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.462e+00  1.941e-02 281.479   <2e-16 ***
## SQFT         1.006e-04  7.173e-06  14.022   <2e-16 ***
## LOT_SIZE    -2.185e-06  9.007e-07  -2.426   0.0154 *
```

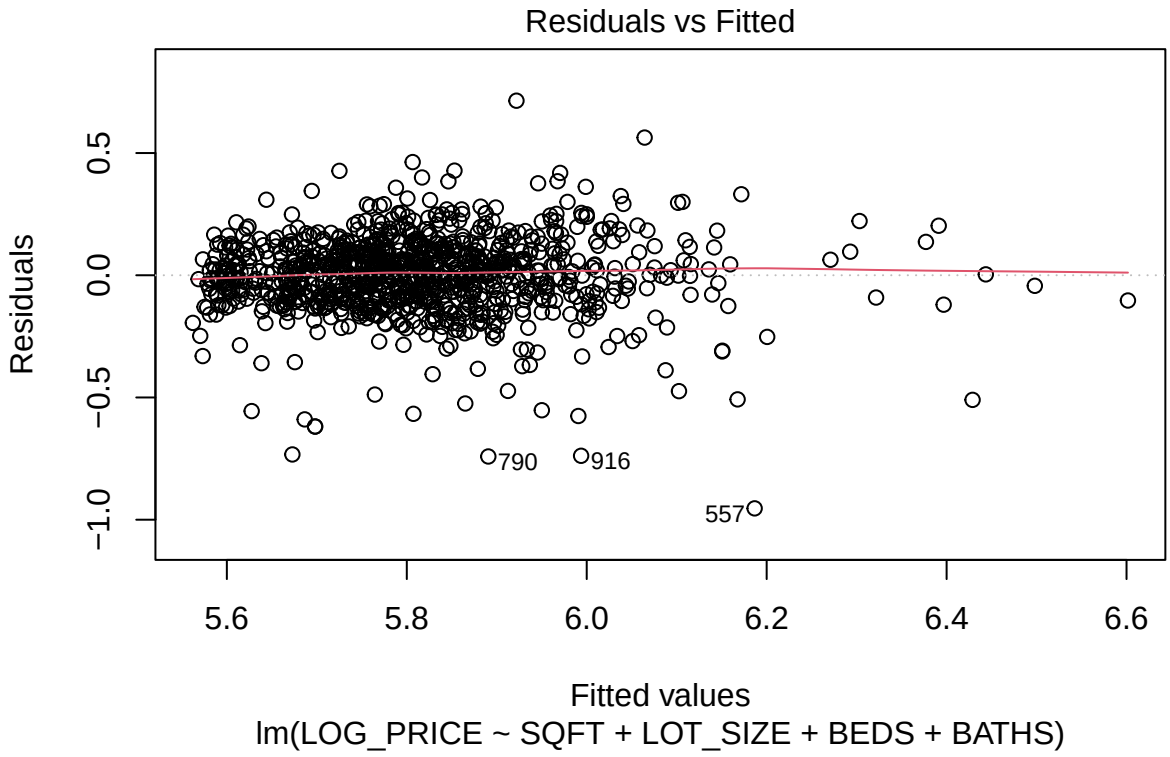
```
## BEDS        -1.321e-02  7.012e-03  -1.884   0.0598 .
## BATHS        8.480e-02  8.295e-03  10.223   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1562 on 995 degrees of freedom
## Multiple R-squared:  0.4446, Adjusted R-squared:  0.4424
## F-statistic: 199.1 on 4 and 995 DF,  p-value: < 2.2e-16
```

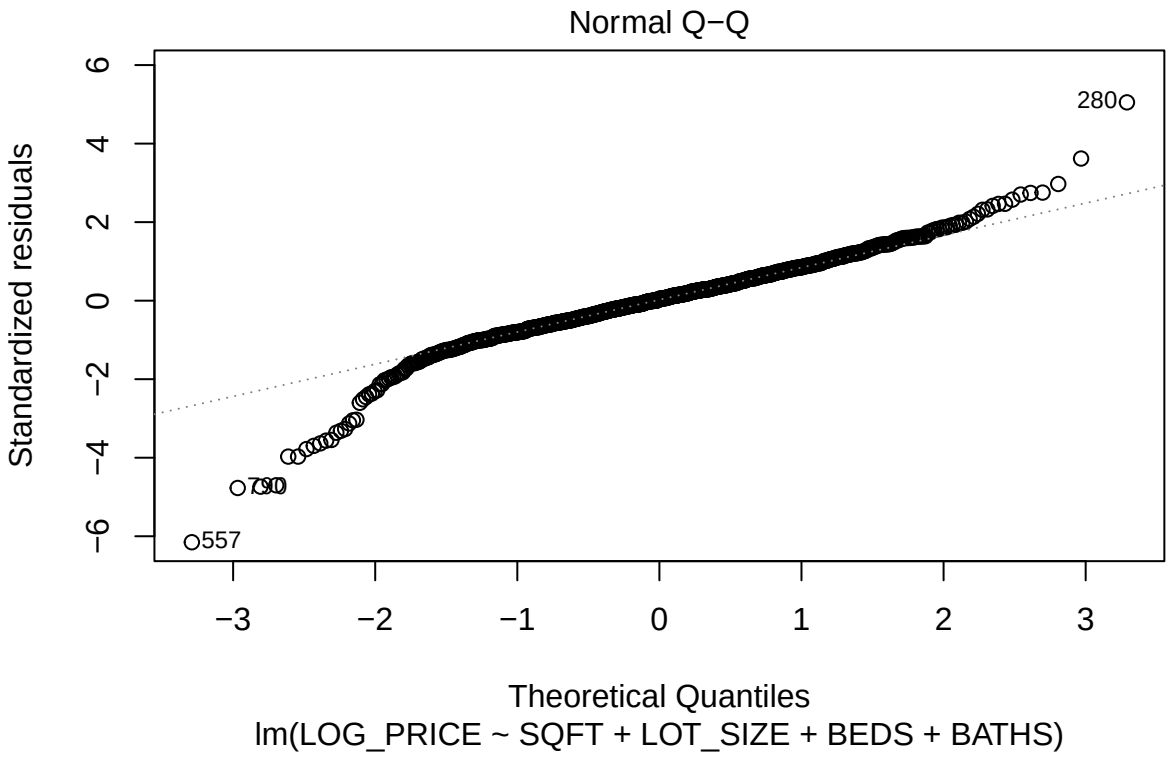
1.9. Give an interpretation of the estimated coefficient of the variable SQFT in Model 2.

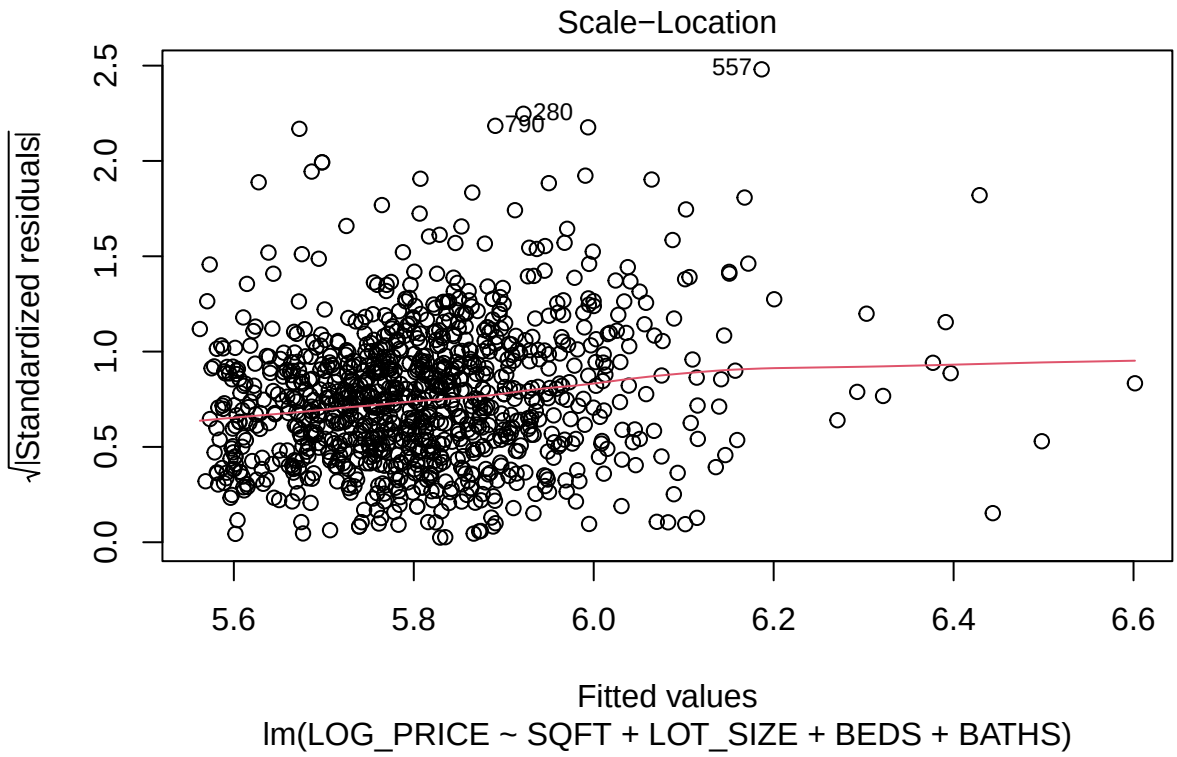
The interpretation of β is the average *difference* in the mean of Y per unit *difference* in X. The average difference in the mean of log of LAST_SALE_PRICE(LOG_PRICE) per unit difference in SQFT is 0.0001005839 considering the other varibles remain constant.

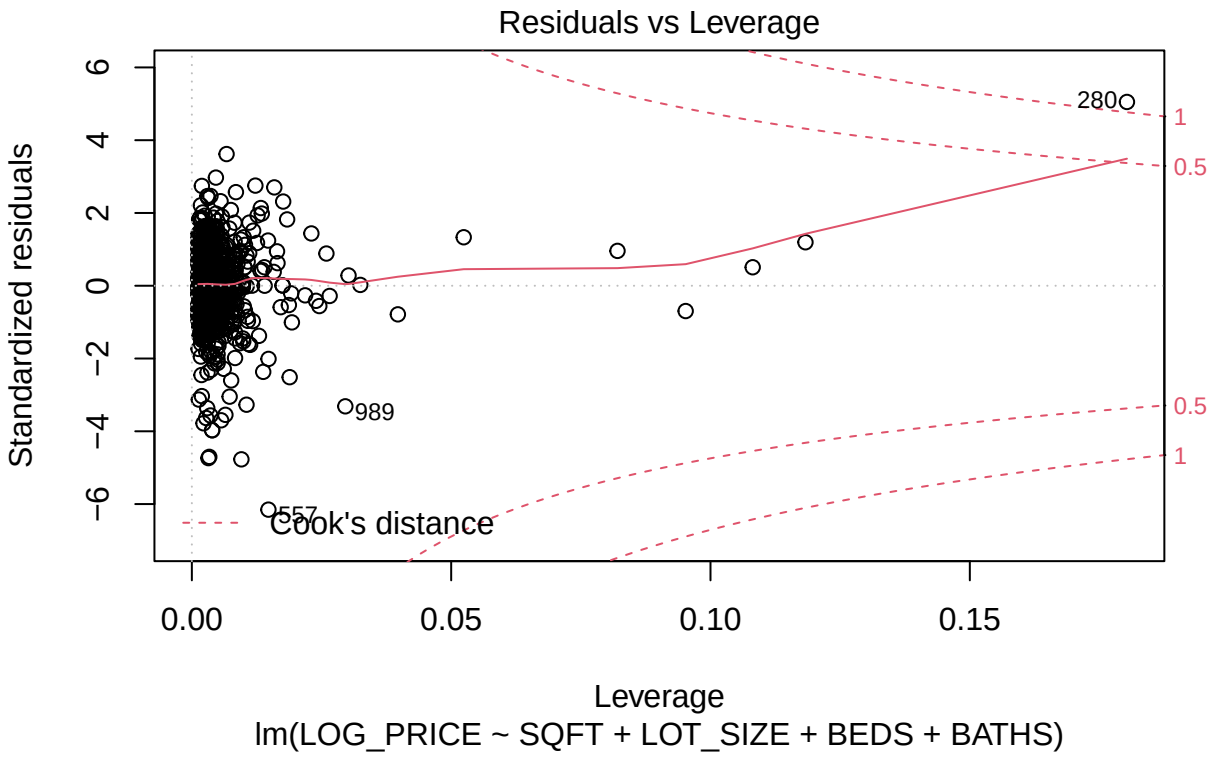
Answer the following questions using residual plots for Model 2. You do not need to display the plots in your submission.

```
plot(model_2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS)
280
557
790
90
−6
−4
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
557
280
790
5.6
5.8
6.0
6.2
6.4
6.6
Fitted values
lm(LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS)



1.10. Assess the linearity assumption of Model 2. Explain by describing a pattern in one or more residual plots.

To check for the linearity assumption, we observe the Residuals vs Fitted plot. From that plot, we can see that it is still a little bit clustered towards lower sales price. There doesn't necessarily seem to be a pattern here, but it looks fairly random as its supposed to. So, I think the assumption of linearity is met.

As we discussed earlier, Linearity is necessary for linear regression, but should not be interpreted too strictly because it rarely if ever is exactly true. So, I didn't interpret the clustering in the graph very strictly.

1.11. Assess the constant variance assumption of Model 2. Explain by describing a pattern in one or more residual plots.

Using the residual vs fitted plot, we can see that most of the data lie evenly on either side of 0 residual (excluding a few outliers). Unlike the graph for the model 1, we cannot see the spread of residual grow larger with the increase in the fitted value. This graph suggests that the assumption of constant-variance is met.

1.12. Assess the normality assumption of Model 2. Explain by describing a pattern in one or more residual plots.

The normality assumption is that the errors ϵ_i are normally distributed.

We use the residuals to check normality by applying histograms and q-q plots to the residuals.

In the q-q plot, We can see that the points lie mostly along the straight diagonal line with some deviations along each of the tails. Based on this plot, we could safely assume that this set of data is normally distributed.

1.13. Give an overall assessment of how well the assumptions hold for Model 2.

Considering model 2, the linearity, constant variance and normality assumptions hold.

1.14. Would statistical inferences based on Model 2 be valid? Explain.

The assumptions are needed to justify **statistical inference** for the regression coefficients. This includes confidence intervals as well as hypothesis tests for the coefficients.

It is important to note that the assumptions are not needed for fitting the model and using the results for purposes other than statistical inference. Thus, we can use linear regression models in a descriptive or exploratory fashion without worrying about assumptions, as long as we don't make inferential statements about the model or the parameters.

So, since all our assumptions hold, the statistical inferences based on model 2 would be valid.