Data 557 HW5

Anuhya B S

2/23/2022

Data: "Sales_sample.csv".

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

```
data_sales = read.csv('Sales_sample.csv')
```

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.

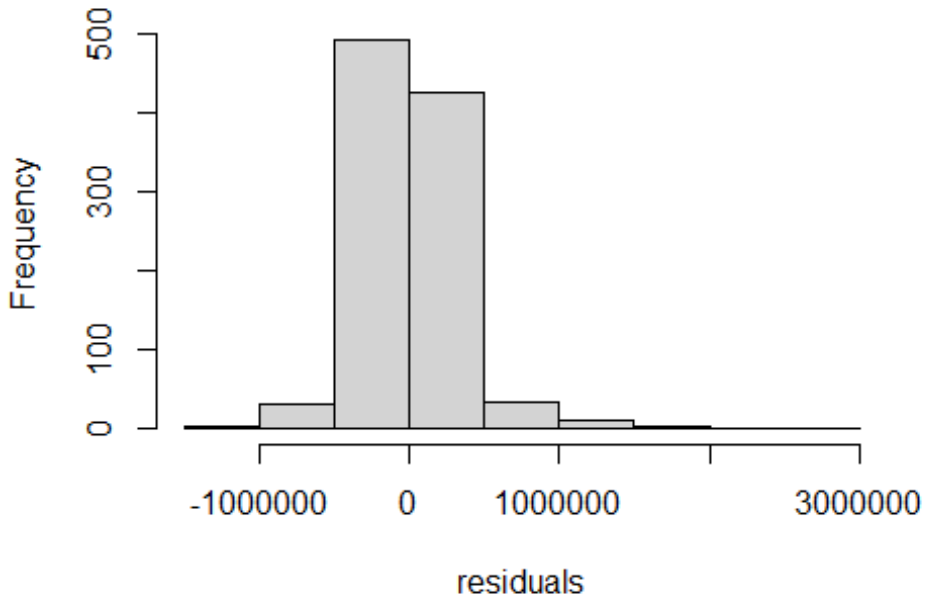
```
summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data =
data_sales))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,
##     data = data_sales)
##
## 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
```

```
model_1 = (lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,
data =  data_sales))
residuals = model_1$residuals
fit = model_1$fitted.values
```

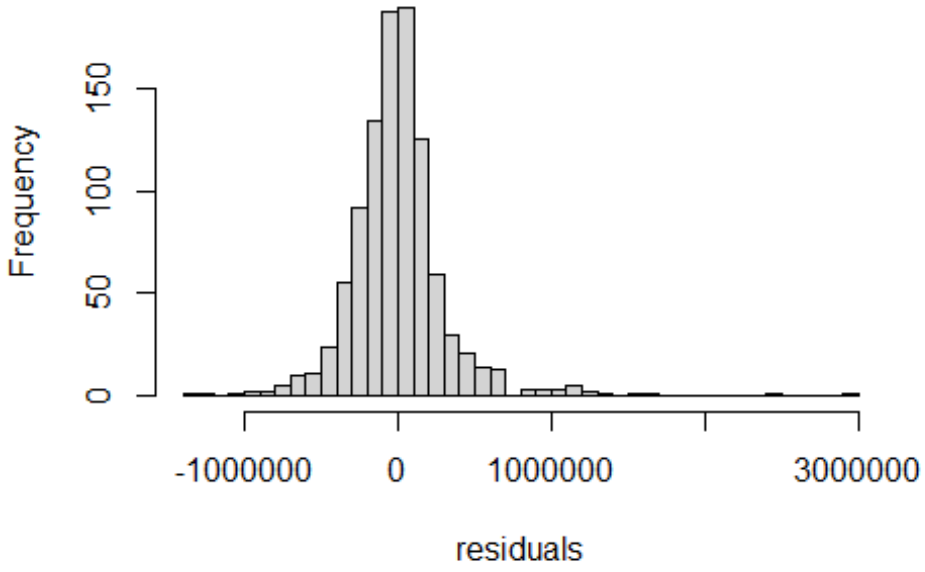
1.2. Create a histogram of the residuals. Based on this graph does the normality assumption hold?

```
options(scipen = 999)
hist(residuals,main="")
```



The graph looks like it follows normal distribution however the data is too peaked in the middle. If we add more bins to the histogram, it seems more normally distributed and having longer tails.

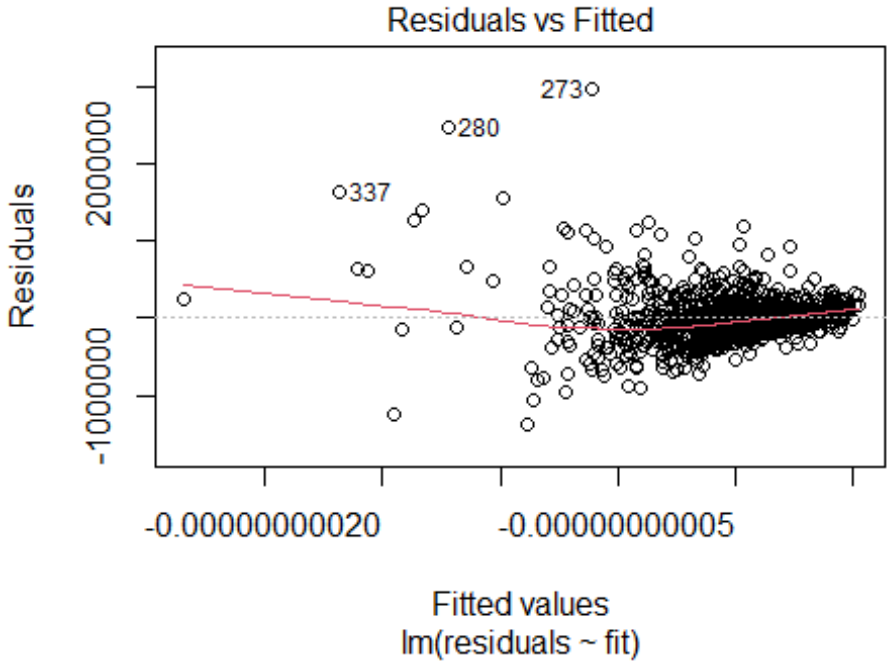
```
hist(residuals,breaks = 50,main="")
```

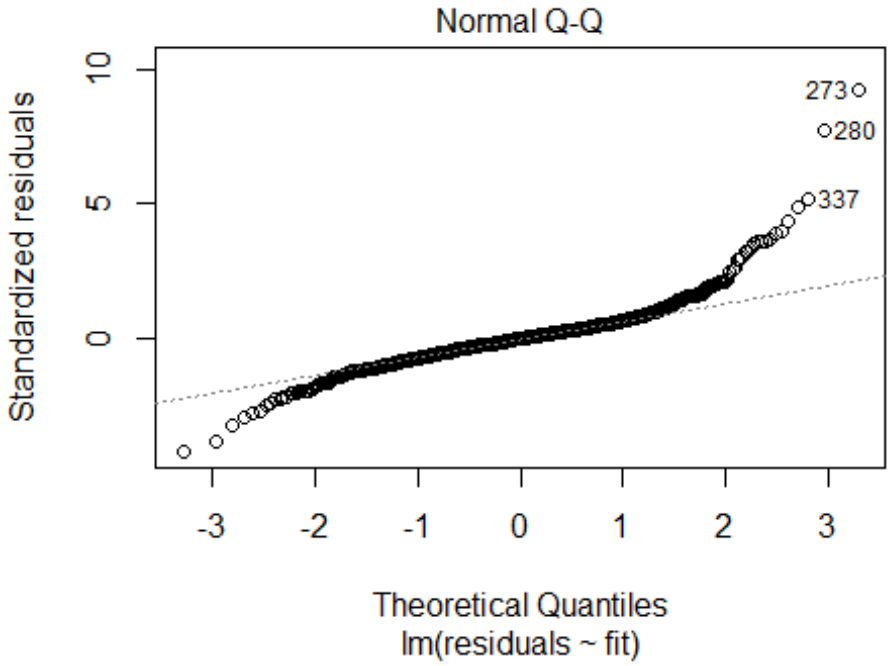


Based on these graphs, I would assume that the normality holds.

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(lm(residuals ~ fit))
```


Residuals vs Fitted
Residuals
2000000
-1000000
273
280
337
-0.00000000020
-0.00000000005
Fitted values
lm(residuals ~ fit)


Normal Q-Q
Standardized residuals
10
5
0
273
280
337
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(residuals ~ fit)

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

We can assess the linearity of a regression model using the regression vs fitted values plot. If linearity holds, the mean of all the residuals should be 0 which means that there is no relationship that can be observed in the plot.

The pattern in the plot shows that all the residual values are too clustered around the right edge of the plot and bounce randomly around the 0 line. If we avoid the outliers, we can say that there is some linearity.

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

The plot of residuals against the fitted values shows some evidence of non constant variance - the residuals are more spread out for lower fitted values while they are closely spread for positive fitted values. There is a slight evidence of funnel shape in the plot, which indicates that there is heteroskedasticity.

Thus the constant variance assumption is not satisfied.

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

We use the residuals to check the normality by applying histograms and q-q plot to residuals. The histogram of the residuals has fairly longer tail and a lot of data peaked in the middle. In the Q-Q plot the residuals are plotted on the y-axis against the fitted values of the x-axis. Both ends ((-3,-1.7) and (1.5,3)) of the Q-Q plot deviate from the straight line and it's center follows the straight line. This plot kind of makes sense intuitively when the distribution presented on the histogram is taken into consideration however, the residuals on the Q-Q plot do deviate a lot from the straight line and hence, I would not consider this a perfect fit for the Normal Distribution.

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

The overall assumptions for linear model to justify statistical inference for the regression coefficients are: 1. Independence 2. Linearity 3. constant variance 4. Normality

Independence - based in random sampling, we can assume independence

Linearity - satisfied

Constant variance - not satisfied

Normality - not satisfied

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

Since assumptions 1-2 are satisfied but 3 is not satisfied, we can not justify the statistic inference for regression coefficient. It can lead to the confidence intervals and hypothesis

test being invalid. The problems can be described as inferences being either too conservative or too anti-conservative. The non-normality can be handled using a large sample size. However, for non-constant variance solutions such as robust standard errors, GLMs and Transformations can be used to help.

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.

```
log_price = log10(data_sales$LAST_SALE_PRICE)

summary(lm(formula = log_price ~ SQFT + LOT_SIZE + BEDS + BATHS, data =
data_sales))

##
## Call:
## lm(formula = log_price ~ SQFT + LOT_SIZE + BEDS + BATHS, data =
data_sales)
##
## 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)  5.4623182142  0.0194057852 281.479 <0.0000000000000002 ***
## SQFT         0.0001005839  0.0000071730  14.022 <0.0000000000000002 ***
## LOT_SIZE    -0.0000021854  0.0000009007  -2.426              0.0154 *
## BEDS        -0.0132115612  0.0070118572  -1.884              0.0598 .
## BATHS        0.0847985006  0.0082948013  10.223 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

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.

$\hat{\beta} = 0.0001005839$ is the estimated average difference in log transformed sale price(in dollars) per unit difference in area keeping other predictors constant.

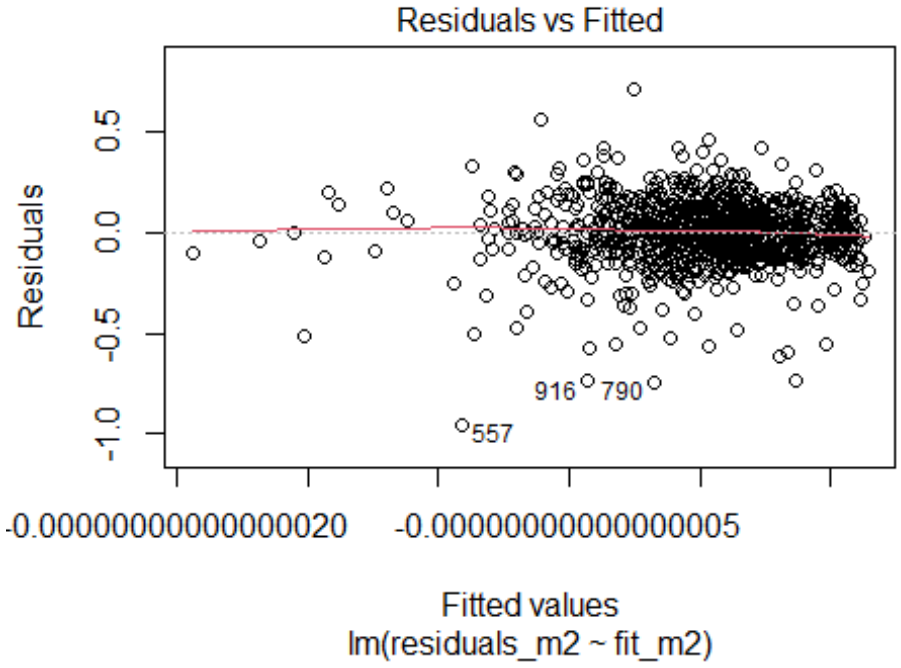
Note: Transforming the log value ratio of the sale price 1.000232.

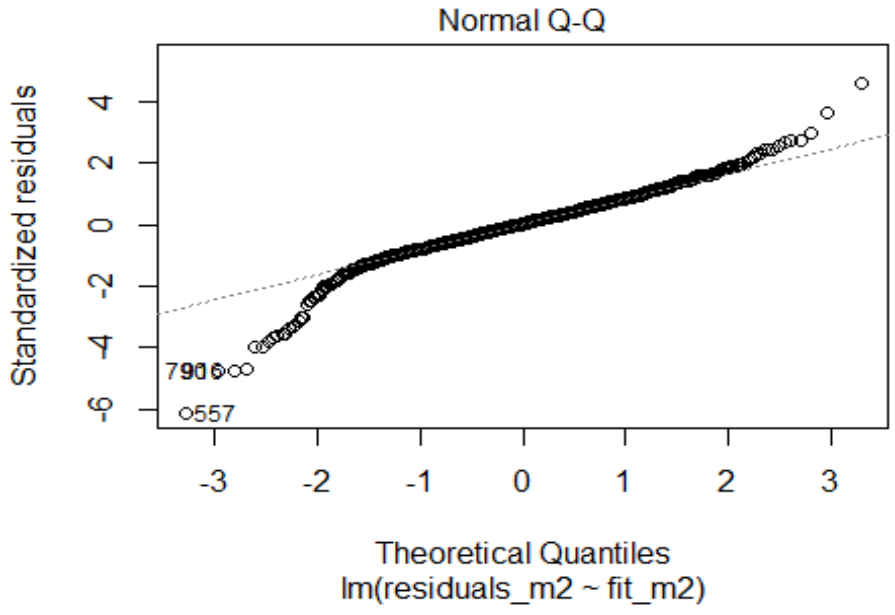
```
10^0.0001005839
```

```
## [1] 1.000232
```

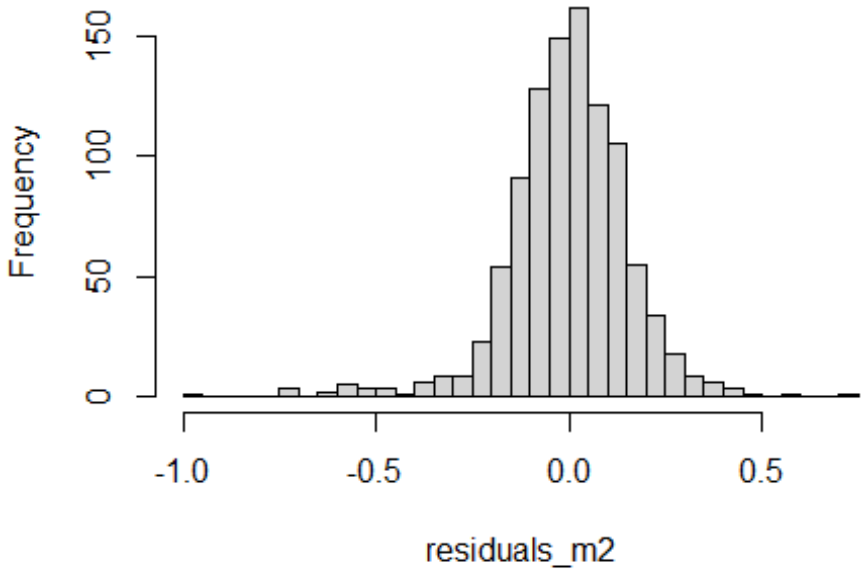
Answer the following questions using residual plots for Model 2. You do not need to display the plots in your submission.

```
model_2 = (lm(formula = log_price ~ SQFT + LOT_SIZE + BEDS + BATHS, data =
data_sales))
residuals_m2 = model_2$residuals
fit_m2 = model_2$fitted.values
plot(lm(residuals_m2 ~ fit_m2))
```





```
hist(residuals_m2, breaks= 50, main="")
```



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

We can assess the linearity of a regression model using the regression vs fitted values plot. If linearity holds, the mean of all the residuals should be 0 which means that there is no relationship that can be observed in the plot.

The pattern in the plot shows that the residuals bounce randomly around the 0 line (mostly towards the right edge of the plot) which indicates that the assumption of linearity is reasonable.

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

The residuals roughly form a horizontal band around the 0 line (if otliers are ignored). There is no evidence of funnel shaped residuals on the plot which is an indication of homoscedasticity.

Thus the constant variance assumption is satisfied.

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

We use the residuals to check the normality by applying histograms and q-q plot to residuals. The histogram of the residuals is fairly close to normal and have long tails. In the

q-q plot the residuals are plotted on the y-axis against the fitted values of the x-axis. Both ends ((-3,-2) and (2,3)) of the Q-Q plot deviate from the straight line and it's center follows the straight line. This plot kind of makes sense intuitively when the distribution presented on the histogram is taken into consideration. The deviation of the residuals at the tails can be expected in the Q-Q plot and hence, I would consider this a perfect fit for the Normal Distribution 9if the outliers are not considered).

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

The overall assumptions for linear model to justify statistical inference for the regression coefficients are: 1. Independence 2. Linearity 3. constant variance 4. Normality

Independence - due to random sampling, we are assuming independence

Linearity - satisfied

Constant variance - satisfied

Normality - satisfied

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

All assumptions 1-4 are met then we can justify the statistic inference for regression coefficients which includes confidence intervals as well as hypothesis tests for the coefficients. Since the regression assumptions have been met, we would not have to worry about the inferences being too conservate or anti-conservative. The statistical inferences based on Model-2 would be valid.