Assignment 5

Amrit Bhat

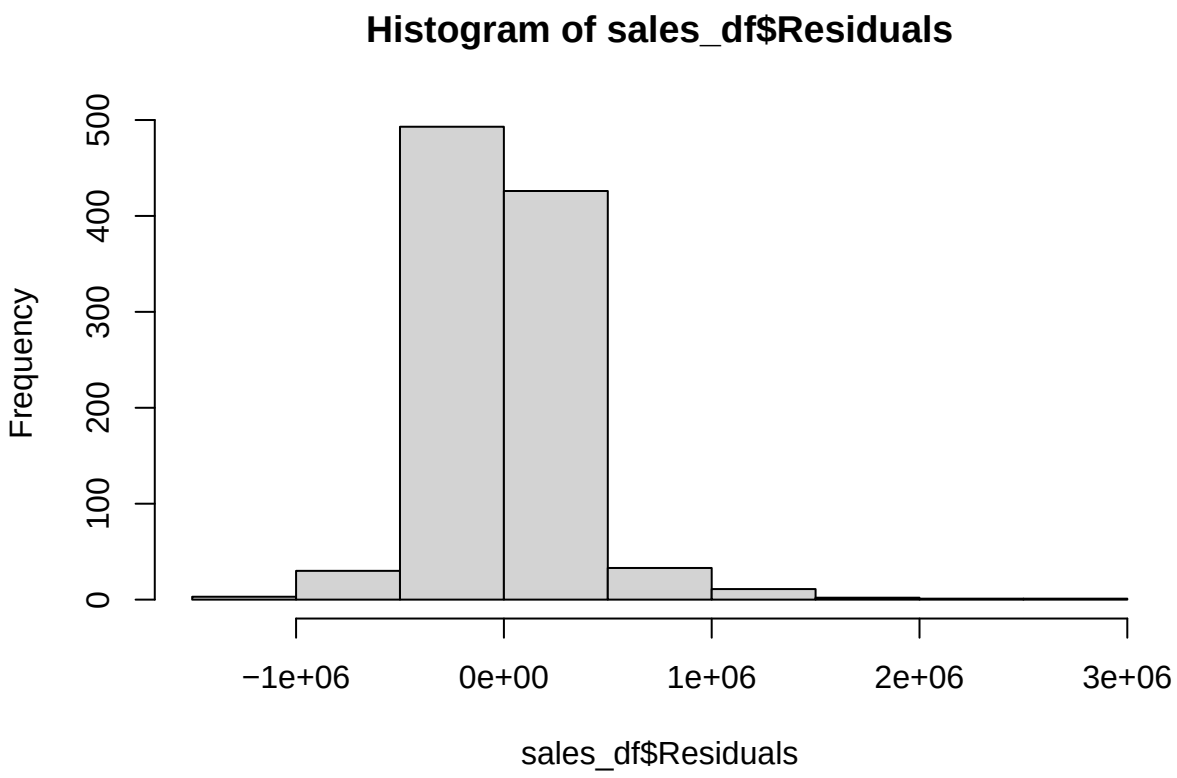
2/24/2022

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_df)
sales_df$Predicted <- model_1$fitted.values
sales_df$Residuals <- model_1$residuals
```

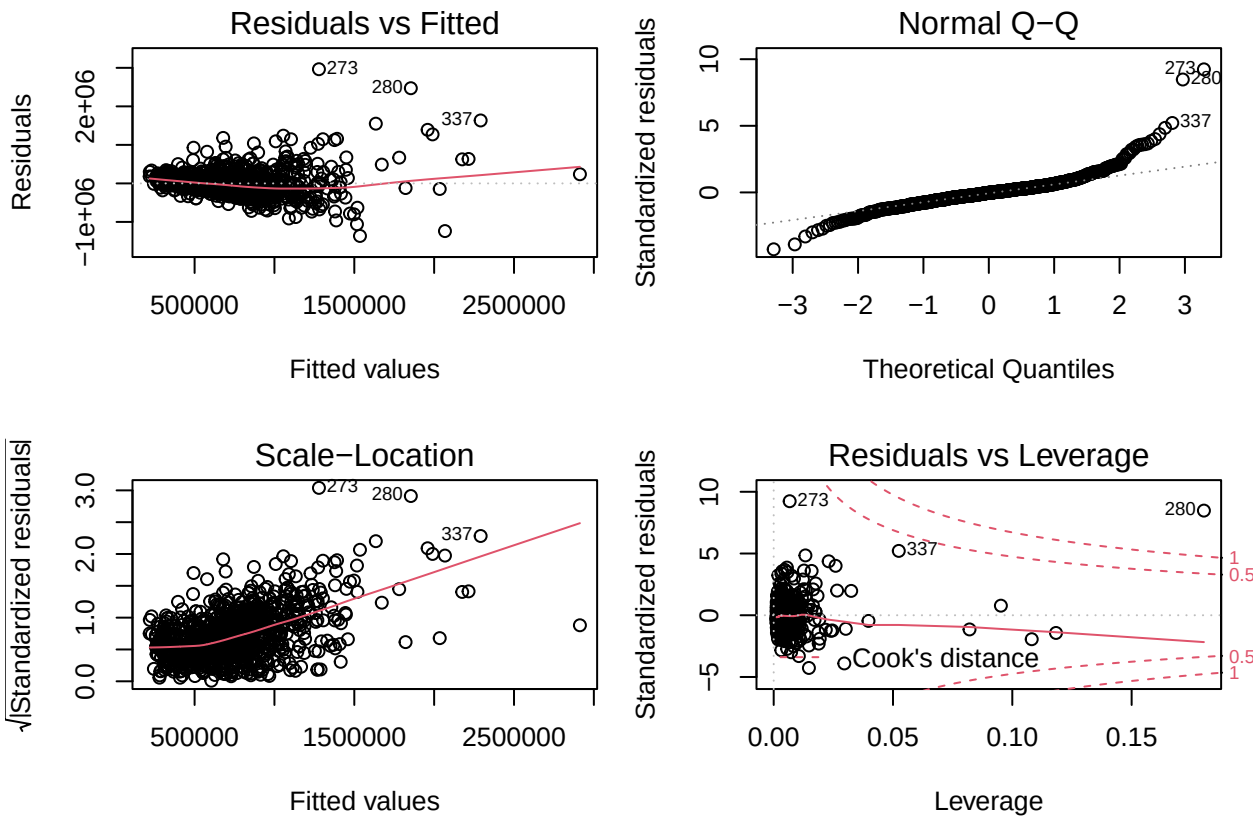
1.2. Create a histogram of the residuals. Based on this graph does the normality assumption hold?

```
hist(sales_df$Residuals)
```



From the above graph, the normality assumption seems to be met as it approximated symmetrical around 0.

```
par(mfrow=c(2,2),mar=c(5,4,2,1))
plot(model_1)
```



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

Linearity holds true as the expected value of the residual is approximately 0 over all values. The graph goes below 0 for approximately as much points compared to when it goes above. So the expected value over all fitted values is close to 0, hence no relationship.

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

The plots shows some evidence of non constant variance as the residuals are more spread out for the larger house price

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

The Q-Q plot of the residual suggests an approximate fit to a normal distribution, with the exception of a few large outliers.

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

The regression model approximately holds the linearity and normality assumption but doesn't hold variance assumption. We cannot make conclusions about independence assumption as we are not sure about the

experimental design of the generated sample.

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

As the constant variance assumption is not met and we do not have any idea about the experimental design, we cannot be sure about the validity of the statistical inferences.

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_df$LOG_PRICE <- log10(sales_df$LAST_SALE_PRICE)
model_2 <- lm(LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,data = sales_df)
summary(model_2)
```

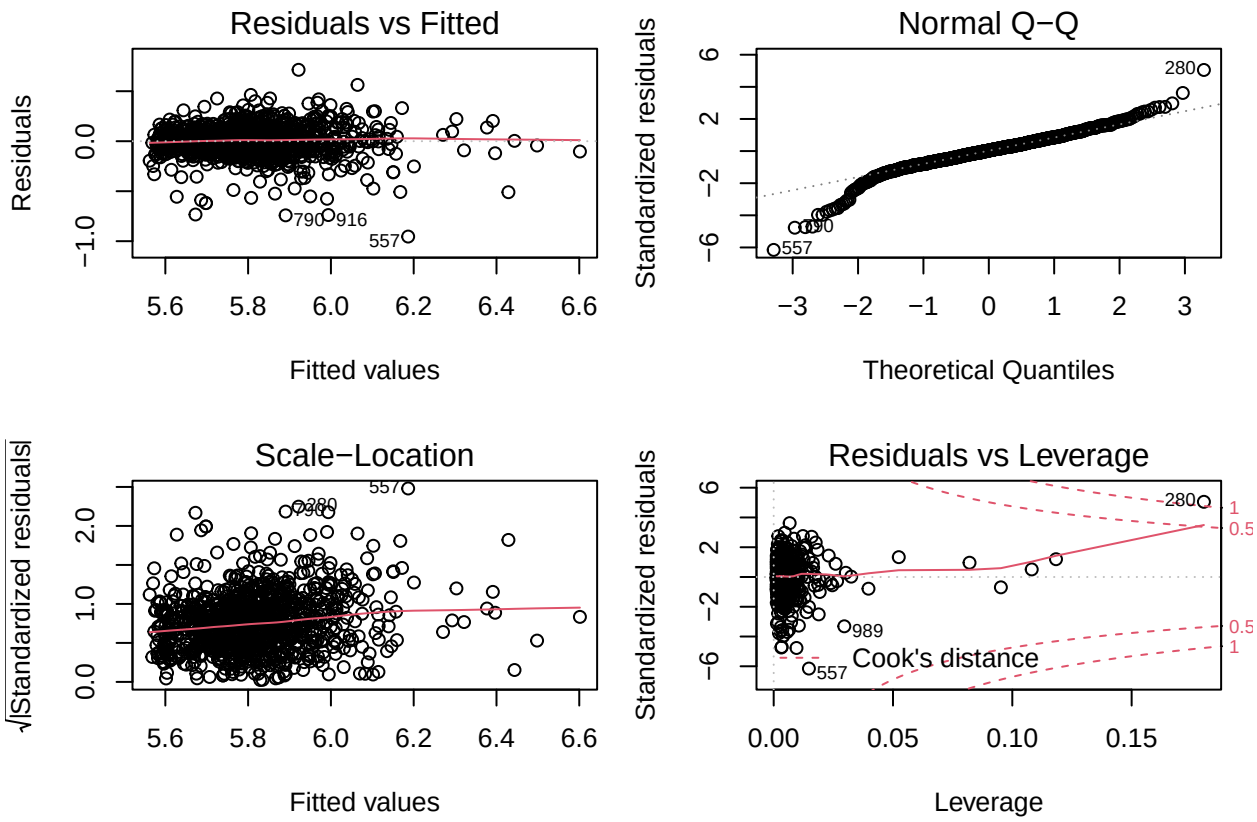
```
Call:
lm(formula = LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data = sales_df)

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 estimated coefficient of the variable SQFT is interpreted as the average rise in the price of the house(0.000106$) for every sqft increase in area of the house.

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

```
par(mfrow=c(2,2),mar=c(5,4,2,1))
plot(model_2)
```



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

Linearity holds true as the expected value of the residual is very close to 0 over all values.

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

The plots shows some evidence of non constant variance as the residuals are more spread out for the larger log of house prices.

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

The Q-Q plot of the residual suggests an approximate fit to a normal distribution, with the exception of a few large outliers.

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

Similar to Model 1, regression Model 2 approximately holds the linearity and normality assumption but doesn't hold variance assumptions. We cannot make conclusions about independence assumption as we are again not sure about the experimental design of the generated sample.

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

Similar to Model 1, the constant variance assumption is not met and we do not have any idea about the experimental design, we cannot be sure about the validity of the statistical inferences.