DATA 557: HW Assignment 5

Hriday Baghar

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).

```
data <- read.csv("Sales_sample.csv")
str(data)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ BEDS           : int  4 4 4 3 6 4 3 5 5 3 ...
##  $ BATHS          : num  2.5 2 2.25 2 2.5 1.75 2.75 3.25 2.5 2 ...
##  $ LOT_SIZE       : int  22578 4000 5000 6400 7431 7200 5500 12345 4000 7000 ...
##  $ LAST_SALE_PRICE: int  678000 888000 682000 1600000 750000 682000 896000 425000 911000 425000 ...
##  $ SQFT           : int  2410 2660 2800 3790 2940 2240 3230 4550 3800 1820 ...
```

```
summary(data)
```

```
##       BEDS           BATHS         LOT_SIZE     LAST_SALE_PRICE
##  Min.   :1.000   Min.   :0.75   Min.   :  653   Min.   :  87050
##  1st Qu.:3.000   1st Qu.:1.75   1st Qu.: 4000   1st Qu.: 475000
##  Median :3.000   Median :2.00   Median : 5502   Median : 632134
##  Mean   :3.388   Mean   :2.12   Mean   : 6635   Mean   : 735809
##  3rd Qu.:4.000   3rd Qu.:2.75   3rd Qu.: 7634   3rd Qu.: 859250
##  Max.   :6.000   Max.   :6.00   Max.   :80791   Max.   :4325000
##       SQFT
##  Min.   : 510
##  1st Qu.:1640
##  Median :2185
##  Mean   :2285
##  3rd Qu.:2760
##  Max.   :8820
```

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 ~ .,  data = data)
summary(model.1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ ., data = 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
## BEDS        -60884.742  14461.536  -4.210 2.78e-05 ***
## BATHS       178177.446  17107.532  10.415  < 2e-16 ***
## LOT_SIZE         6.844      1.858   3.684 0.000242 ***
## SQFT           224.502     14.794  15.175  < 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
```

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

```
library(ggplot2)
```

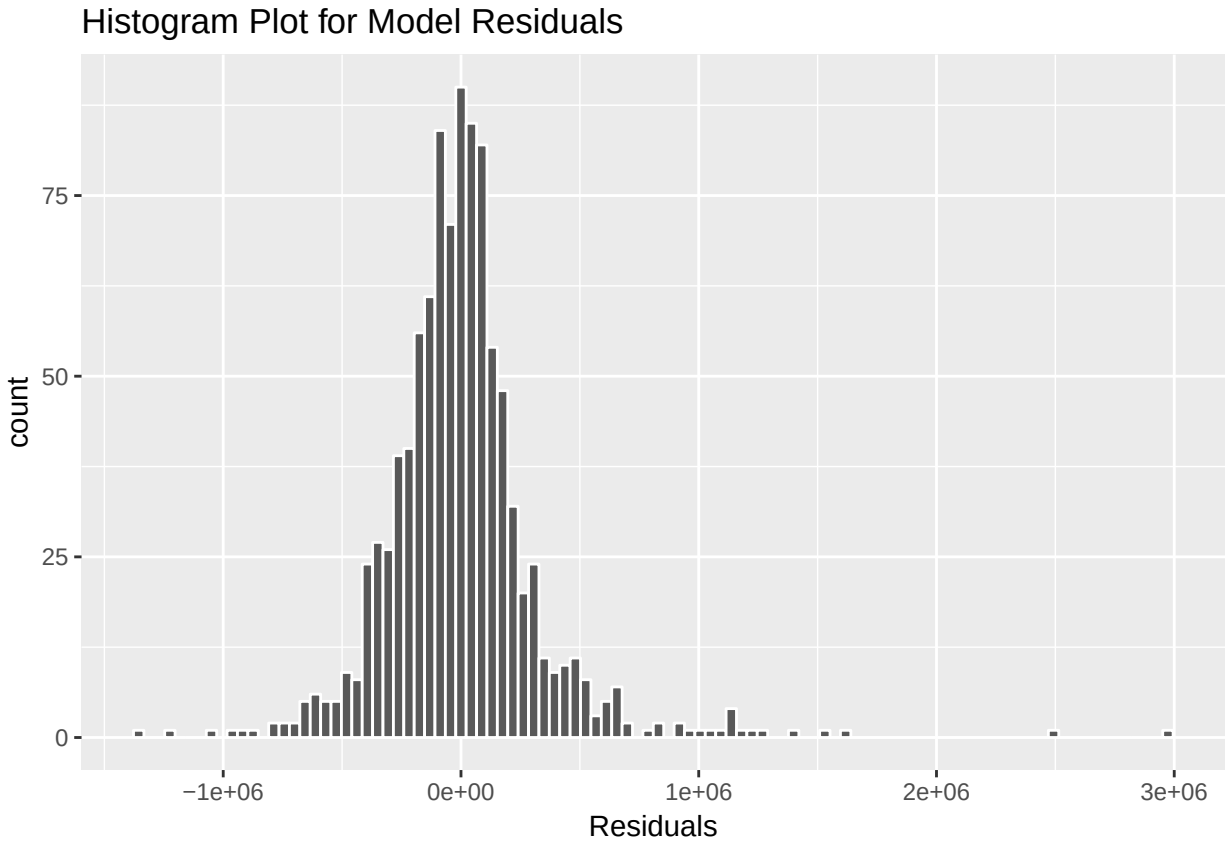
```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
#Using fortify will append the residuals and fitted values to a dataframe with original values
model.1.df <- fortify(model.1)
str(model.1.df)
```

```
## 'data.frame':    1000 obs. of  11 variables:
##  $ LAST_SALE_PRICE: int  678000 888000 682000 1600000 750000 682000 896000 425000 911000 425000 ...
##  $ BEDS           : int  4 4 4 3 6 4 3 5 5 3 ...
##  $ BATHS          : num  2.5 2 2.25 2 2.5 1.75 2.75 3.25 2.5 2 ...
##  $ LOT_SIZE       : int  22578 4000 5000 6400 7431 7200 5500 12345 4000 7000 ...
##  $ SQFT           : int  2410 2660 2800 3790 2940 2240 3230 4550 3800 1820 ...
##  $ .hat           : num  0.00986 0.00237 0.00186 0.00722 0.01161 ...
##  $ .sigma         : num  322141 322189 322189 321787 322218 ...
##  $ .cooksd        : num  9.86e-04 9.60e-05 7.46e-05 3.91e-03 5.07e-05 ...
##  $ .fitted        : num  903464 743351 826169 1074349 797013 ...
##  $ .resid         : num  -225464 144649 -144169 525651 -47013 ...
##  $ .stdresid      : num  -0.704 0.45 -0.448 1.638 -0.147 ...
##  - attr(*, "terms")=Classes 'terms', 'formula'  language LAST_SALE_PRICE ~ BEDS + BATHS + LOT_SIZE + SQFT
##   .. ..- attr(*, "variables")= language list(LAST_SALE_PRICE, BEDS, BATHS, LOT_SIZE, SQFT)
##   .. ..- attr(*, "factors")= int [1:5, 1:4] 0 1 0 0 0 0 0 1 0 0 ...
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:5] "LAST_SALE_PRICE" "BEDS" "BATHS" "LOT_SIZE" ...
##   .. .. .. ..$ : chr [1:4] "BEDS" "BATHS" "LOT_SIZE" "SQFT"
##   .. ..- attr(*, "term.labels")= chr [1:4] "BEDS" "BATHS" "LOT_SIZE" "SQFT"
##   .. ..- attr(*, "order")= int [1:4] 1 1 1 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(LAST_SALE_PRICE, BEDS, BATHS, LOT_SIZE, SQFT)
##   .. ..- attr(*, "dataClasses")= Named chr [1:5] "numeric" "numeric" "numeric" "numeric" ...
##   .. .. ..- attr(*, "names")= chr [1:5] "LAST_SALE_PRICE" "BEDS" "BATHS" "LOT_SIZE" ...
```

```
(hist.resid <- ggplot(data = model.1.df) +
  geom_histogram(aes(x = .resid), bins=100, color = "white") +
  ggtitle("Histogram Plot for Model Residuals") +
```

```
  labs(x = "Residuals")
  )
```



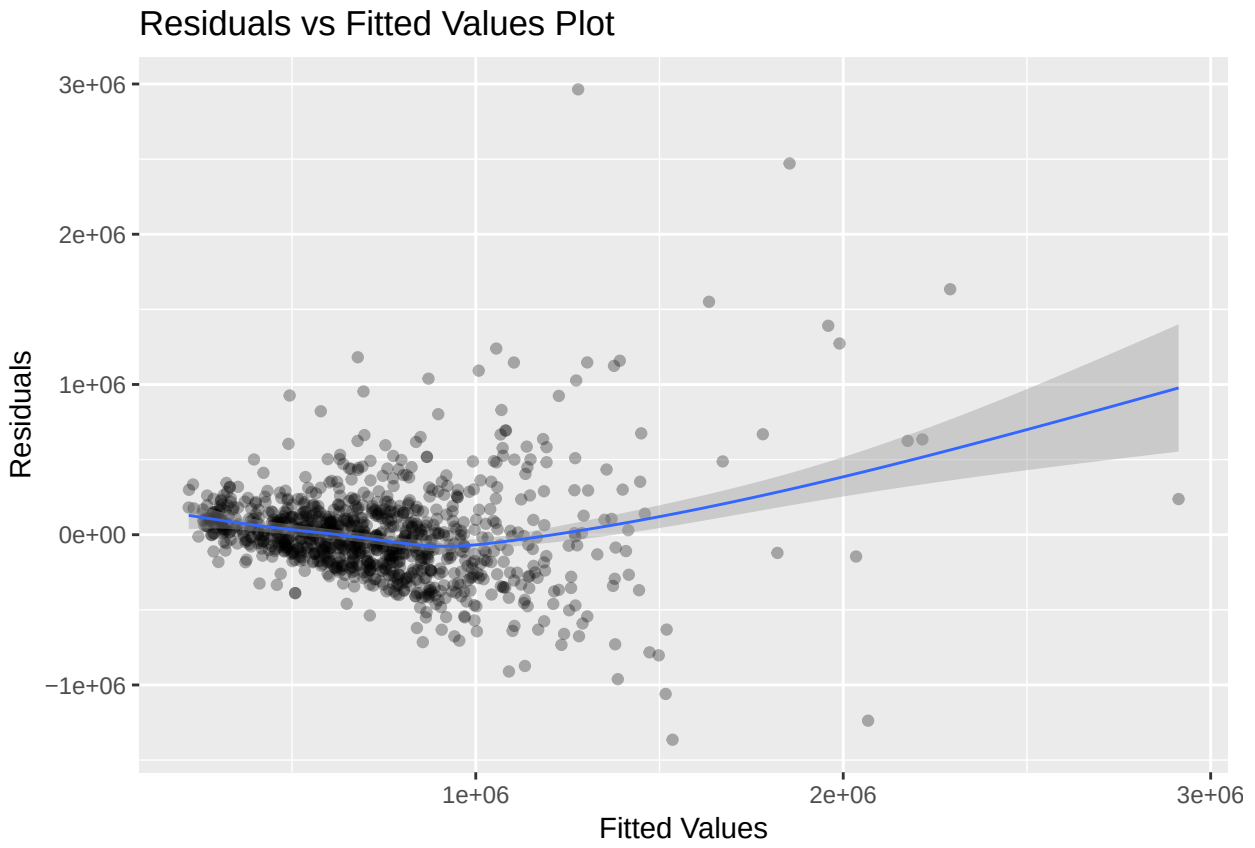
The distribution of the residuals is somewhat normal although it is slightly right skewed, possibly due to certain outliers.

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.

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

```
(scatter.resid <- ggplot(data = model.1.df, aes(y = .resid, x = .fitted)) +
   geom_point(alpha = 0.3) +
   geom_smooth(method = "auto", size = 0.5) +
   ggtitle("Residuals vs Fitted Values Plot") +
   labs(x = "Fitted Values", y = "Residuals")
   )
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

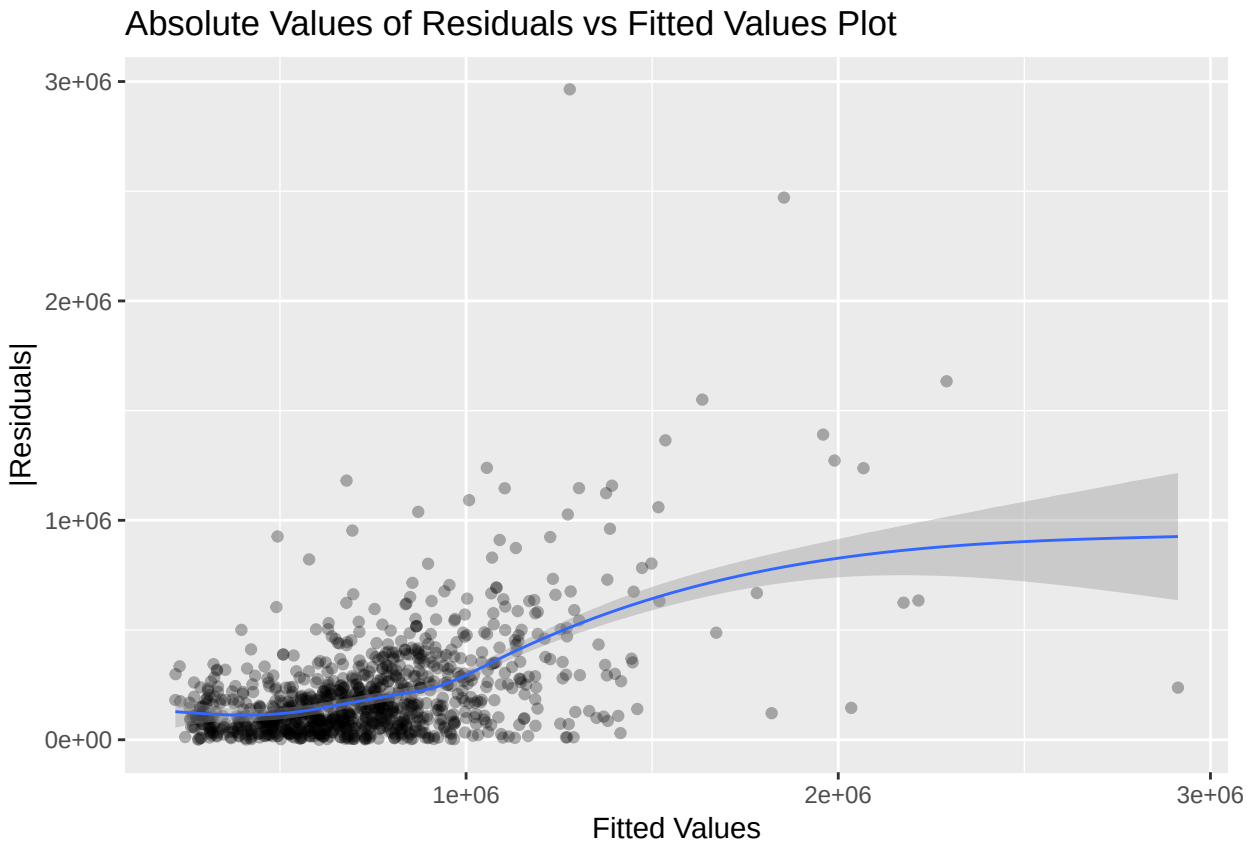


The linearity assumption appears to be violated for larger fitted values as we can see from the plot. For the bulk of the dataset, there appears so be some minor relationship between fitted values and residuals.

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

```
(scatter.abs.resid <- ggplot(data = model.1.df, aes(x = .fitted, y = abs(.resid)))+
   geom_point(alpha = 0.3) +
   geom_smooth(method = "auto", size = 0.5) +
   ggtitle("Absolute Values of Residuals vs Fitted Values Plot") +
   labs(x = "Fitted Values", y = "|Residuals|"))
```

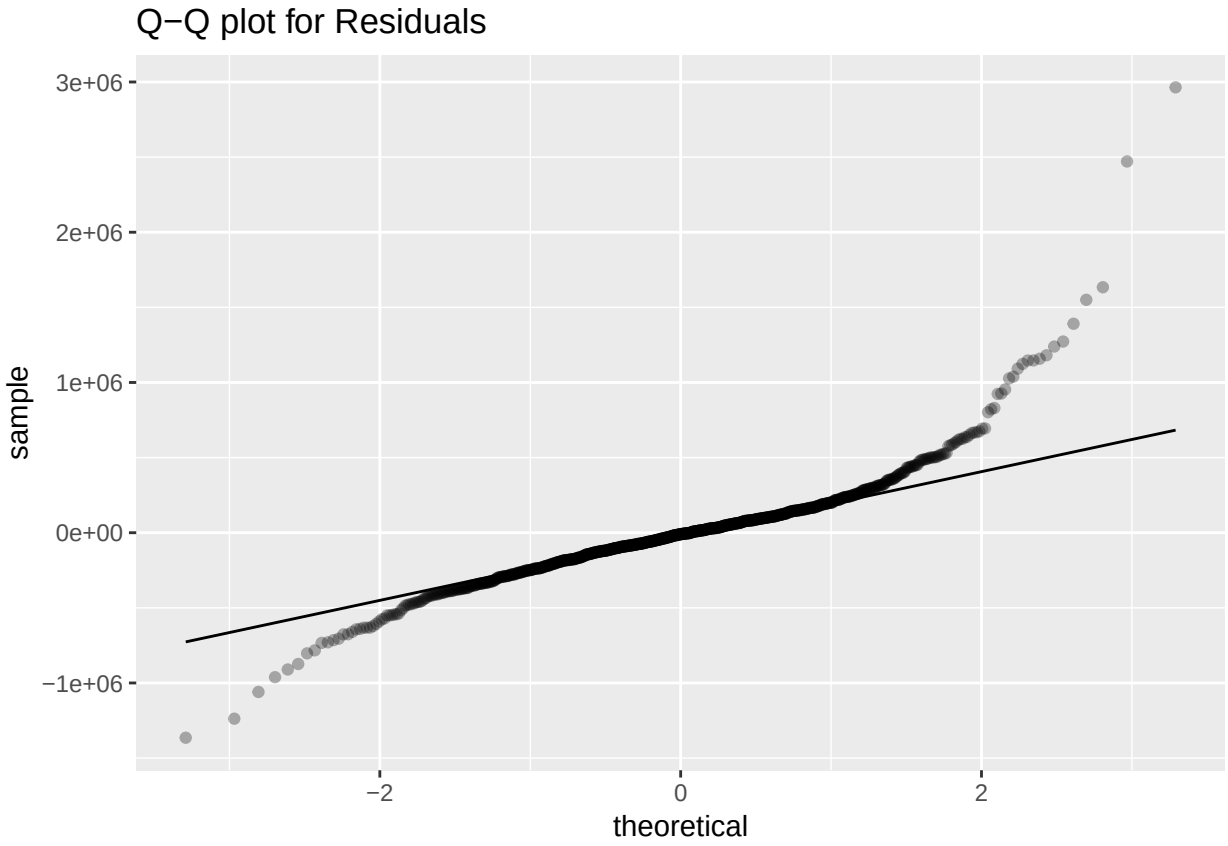
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The smoothing curve should be approximately horizontal for the homoscedasticity assumption to hold. We see that is not the case, hence the constant variance assumption does not hold. We also see that the spread of the points is higher for larger fitted values, implying variance is not constant.

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

```
(qq.resid <- ggplot(data = model.1.df, aes(sample = .resid)) +
  geom_qq(alpha = 0.3) +
  geom_qq_line() +
  ggtitle("Q-Q plot for Residuals"))
```



We see that the normality assumption is violated at both the tails of the distribution.

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

Overall, there appears to be a clear violation of the linearity, constant variance and normality assumption for the model.

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

Violation of constant variance and linearity assumptions makes statistical inferences based on this model unreliable. Since we have a large sample size we need not worry about the normality assumption.

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.

```
data$LOG_PRICE <- log10(data$LAST_SALE_PRICE)
summary(data)
```

```
##       BEDS           BATHS         LOT_SIZE     LAST_SALE_PRICE
##  Min.   :1.000   Min.   :0.75   Min.   :  653   Min.   :  87050
##  1st Qu.:3.000   1st Qu.:1.75   1st Qu.: 4000   1st Qu.: 475000
##  Median :3.000   Median :2.00   Median : 5502   Median : 632134
##  Mean   :3.388   Mean   :2.12   Mean   : 6635   Mean   : 735809
##  3rd Qu.:4.000   3rd Qu.:2.75   3rd Qu.: 7634   3rd Qu.: 859250
##  Max.   :6.000   Max.   :6.00   Max.   :80791   Max.   :4325000
##       SQFT        LOG_PRICE
##  Min.   : 510   Min.   :4.940
##  1st Qu.:1640   1st Qu.:5.677
##  Median :2185   Median :5.801
```

```
##  Mean   :2285   Mean   :5.813
##  3rd Qu.:2760   3rd Qu.:5.934
##  Max.   :8820   Max.   :6.636
```

```
model.2 <- lm(LOG_PRICE ~ . - LAST_SALE_PRICE, data = data)
summary(model.2)
```

```
##
## Call:
## lm(formula = LOG_PRICE ~ . - LAST_SALE_PRICE, data = 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 ***
## BEDS        -1.321e-02  7.012e-03  -1.884   0.0598 .
## BATHS        8.480e-02  8.295e-03  10.223   <2e-16 ***
## LOT_SIZE    -2.185e-06  9.007e-07  -2.426   0.0154 *
## SQFT         1.006e-04  7.173e-06  14.022   <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.

For a unit increase in sqft, holding all other variables at a constant value, the log10 of sale price increases by 1.006e-4 units. Additionally, we can interpret the coefficient in terms of percentage as follows:

```
cat("Percentage change:", round((10^model.2$coefficients["SQFT"]-1)*100,3),"%")
```

```
## Percentage change: 0.023 %
```

For a unit increase in sqft (holding all other variables constant!), the percentage increase in sale price is 0.023%

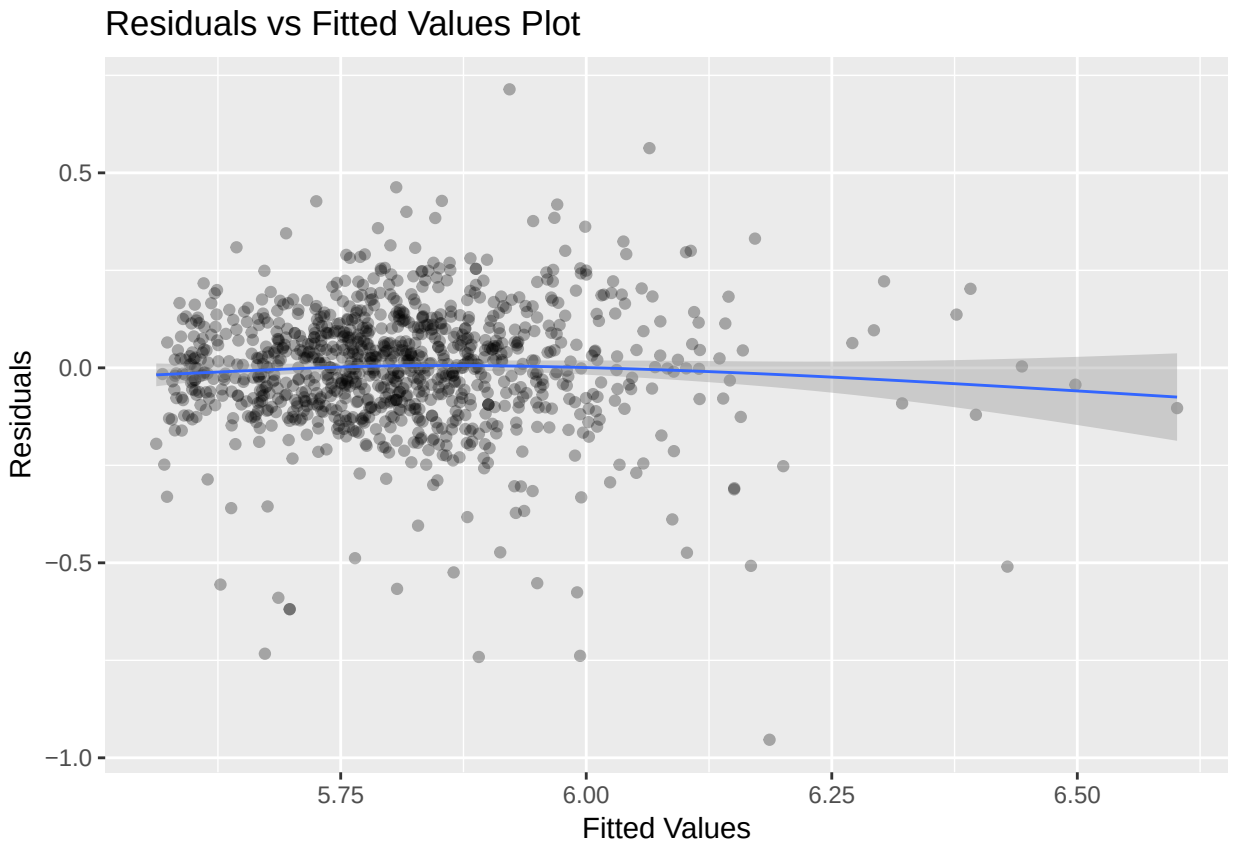
Answer the following questions using residual plots for Model 2. You do not need to display the plots in your submission.

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

```
model.2.df <- fortify(model.2)

(scatter.resid.2 <- ggplot(data = model.2.df, aes(y = .resid, x = .fitted)) +
   geom_point(alpha = 0.3) +
   geom_smooth(method = "auto", size = 0.5) +
   ggtitle("Residuals vs Fitted Values Plot") +
   labs(x = "Fitted Values", y = "Residuals")
   )
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

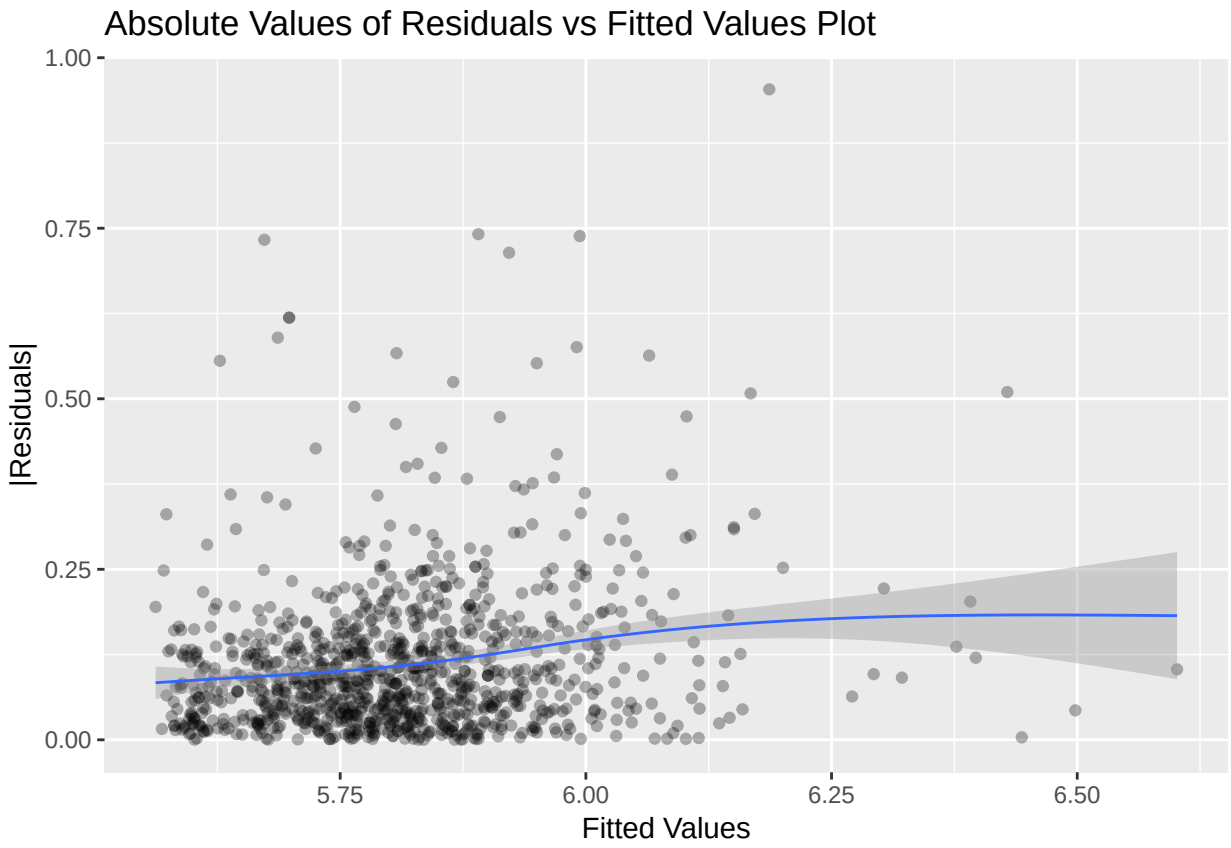


The linearity assumption appears to hold for this model. There is a very minor relationship between fitted values and residuals for outliers, as observed by the smoothing curve.

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

```
(scatter.abs.resid.2 <- ggplot(data = model.2.df, aes(x = .fitted, y = abs(.resid)))+
   geom_point(alpha = 0.3) +
   geom_smooth(method = "auto", size = 0.5) +
   ggtitle("Absolute Values of Residuals vs Fitted Values Plot") +
   labs(x = "Fitted Values", y = "|Residuals|"))
```

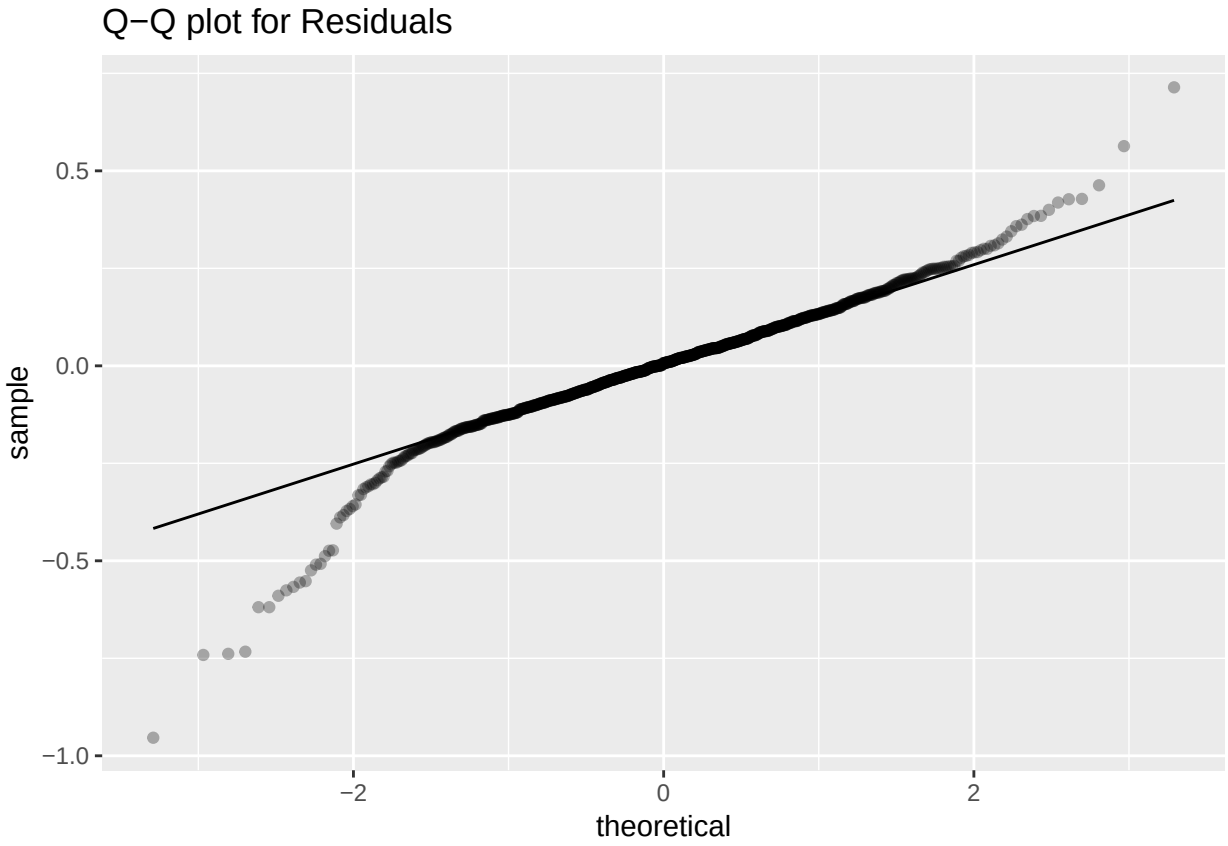
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The smoothing curve is nearly horizontal for the model, hence the constant variance assumption appears to hold.

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

```
(qq.resid.2 <- ggplot(data = model.2.df, aes(sample = .resid)) +
  geom_qq(alpha = 0.3) +
  geom_qq_line() +
  ggtitle("Q-Q plot for Residuals"))
```



The normality assumption is violated for this model at both the tails, although the number of points deviating considerably from the theoretical values appears to be lesser than model 1.

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

Linearity and constant variance assumptions hold based on the observed plots. The normality assumption does not hold, but we have a large sample size so we need not worry about this.

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

The major assumptions of independence, linearity and constant variance are met. If we have a sufficiently large sample size, statistical inferences based on model 2 should be valid.