DATA 557: Applied Statistics and Experimental Design
Homework 5

Abhishek Saini

February 2022

Question 1

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

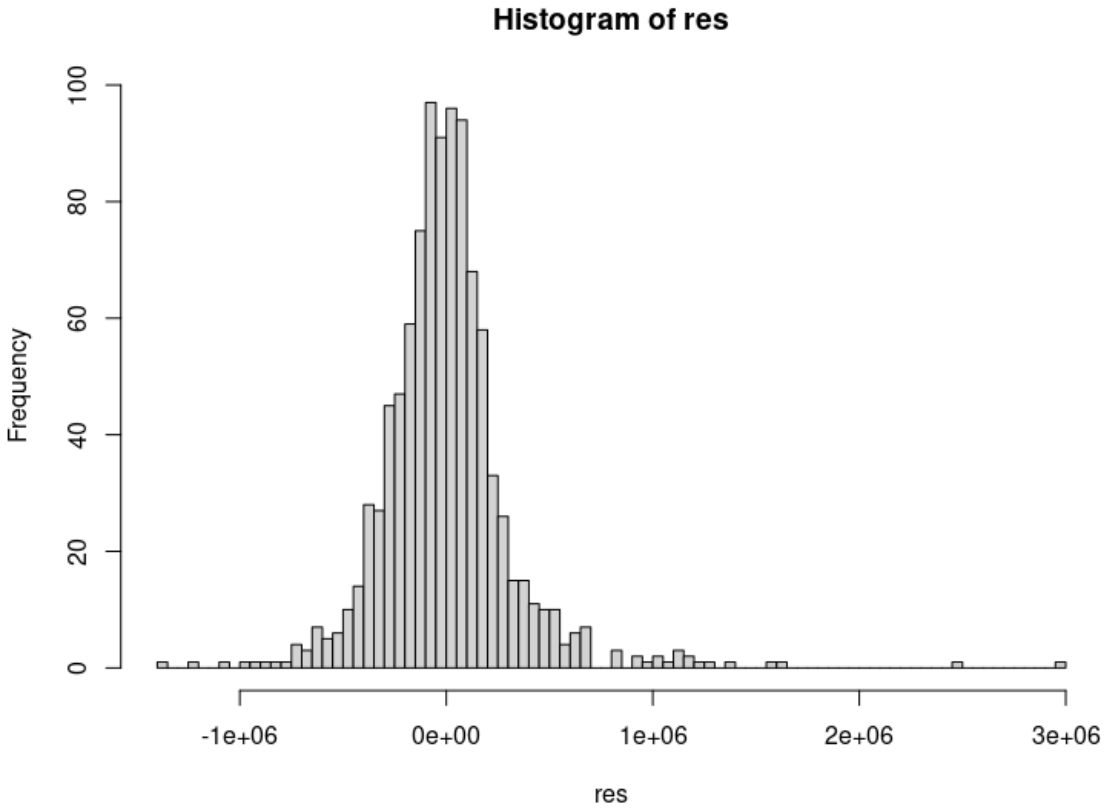
0.1 Fit a linear regression model (Model 1) with sale price as response variable and SQFT, LOT_SIZE, BEDS, and BATHS as predictor variables. Add the fitted values and the residuals from the models as new variables in your data set. Show the R code you used for this question.

Here is the R code for this question.

```
model1 = lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data = sales)
pred = model1$fitted.values
res = model1$residuals
```

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

The histogram of residuals is shown below. This looks like a normal distribution.



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.

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

Checking residual vs fitted values for the model, we see that the mean of residuals is close to zero for all range of fitted values, so the linearity assumption seems to hold.

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

In the scale location plot, we can see that the line isn't horizontal but shows a clear increasing trend which suggests that the constant variance assumption doesn't hold.

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

The QQ-plot shows a clear deviation from straight line at both extremes which suggests that the normality assumption doesn't hold. But we do have a large sample size of 1000.

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

Overall, one of the four assumptions doesn't hold - constant variance and the normality assumption which means we should be careful interpreting the results of this model.

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

The statistical inferences based on this model will not be valid since one of the four assumptions don't hold - constant variance. The inferences based on this model can either be too convervative or anti-conservative.

0.8 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.

The coefficient estimates is calculated as follows:

```
> model2 = lm(formula = LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data = df)
> summary(model2)$coefficients
                 Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  5.462318e+00 1.940579e-02 281.478856 0.000000e+00
SQFT         1.005839e-04 7.173043e-06  14.022481 6.739286e-41
LOT_SIZE    -2.185404e-06 9.007442e-07  -2.426221 1.543384e-02
BEDS        -1.321156e-02 7.011857e-03  -1.884174 5.983274e-02
BATHS        8.479850e-02 8.294801e-03  10.223090 2.137789e-23
```

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

In Model 2 the coefficient of SQFT is interpreted as the average difference in LOG_PRICE per unit difference in area(SQFT) for houses of the same LOT SIZE, BEDS, BATHS.

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

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

Checking residual vs fitted values for the model, we see that the mean of residuals is close to zero for all range of fitted values, so the linearity assumption seems to hold.

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

In the scale location plot, we can see that the line is relatively horizontal which suggests variance is constant for the range of fitted values. This suggests that the constant variance assumption holds.

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

The QQ-plot shows a relatively straight line compared to that for model 1 which suggests that the normality assumption holds. We also have a large sample size of 1000.

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

All the assumptions of linear regression model seems to hold - linearity, constant variance and normality.

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

Yes, statistical inferences based on this model would be valid. This is because all the assumptions required for linear regression hold and hence, the results of our model wouldn't be conservative or anti-conservative.