Solution

ST3131 Midterm Sem 1-AY2526

Q1

1. (2 points) Rename the first column of data frame **house** from **HP.in.thousands** to **price**.

```r
house= read.csv("~/Documents/Data/house-price-OR.csv")
head(house)
```

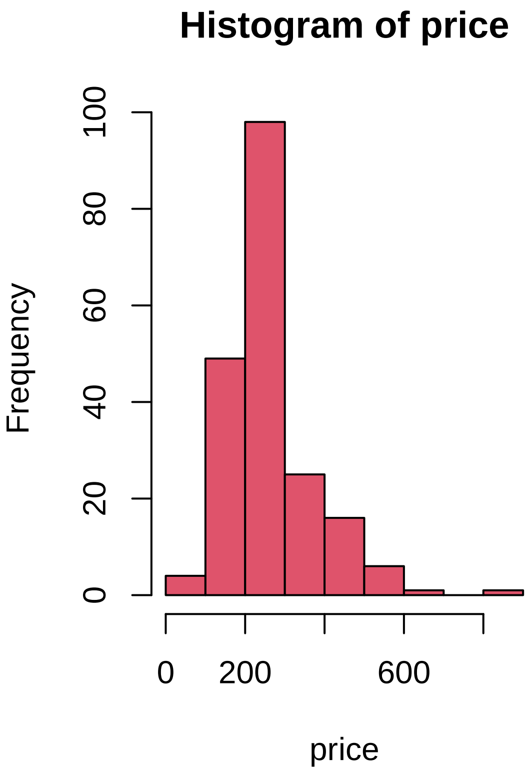
```
##   HP.in.thousands House.Size Acres Bedrooms Garage Age.Category
## 1           232.5       1679  0.23        3      1            M
## 2           470.0       4494  0.52        5      1            M
## 3           150.0       2542  0.11        4      1            N
## 4           167.5       1094  0.18        2      0            O
## 5           210.0       1838  0.19        4      1            M
## 6           522.0       4156  0.22        3      1            N
```

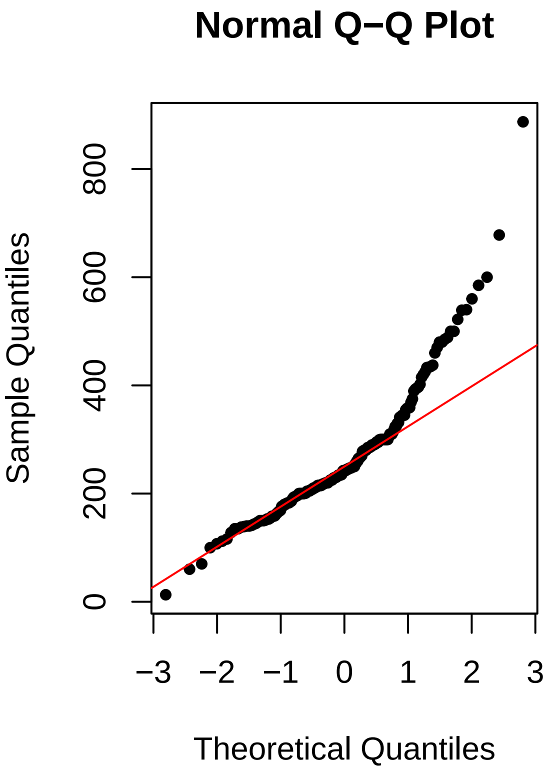
```r
names(house)[1] = "price"
house$Garage = as.factor(house$Garage)
attach(house)
```

2. (2 points) Form a QQ plot and a histogram for **price** and comment if it is suitable to be the response for a linear model. Explain.

```r
par(mfrow = c(1,2))
hist(price, col = 2)

qqnorm(price, pch = 20)
qqline(price, col = "red")
```





Comments: The histogram of **price** shows that it's not symmetric (right skewed). The QQ plot further confirms that the right tail is longer than normal. Hence, variable **price** is not normal, even not symmetric.

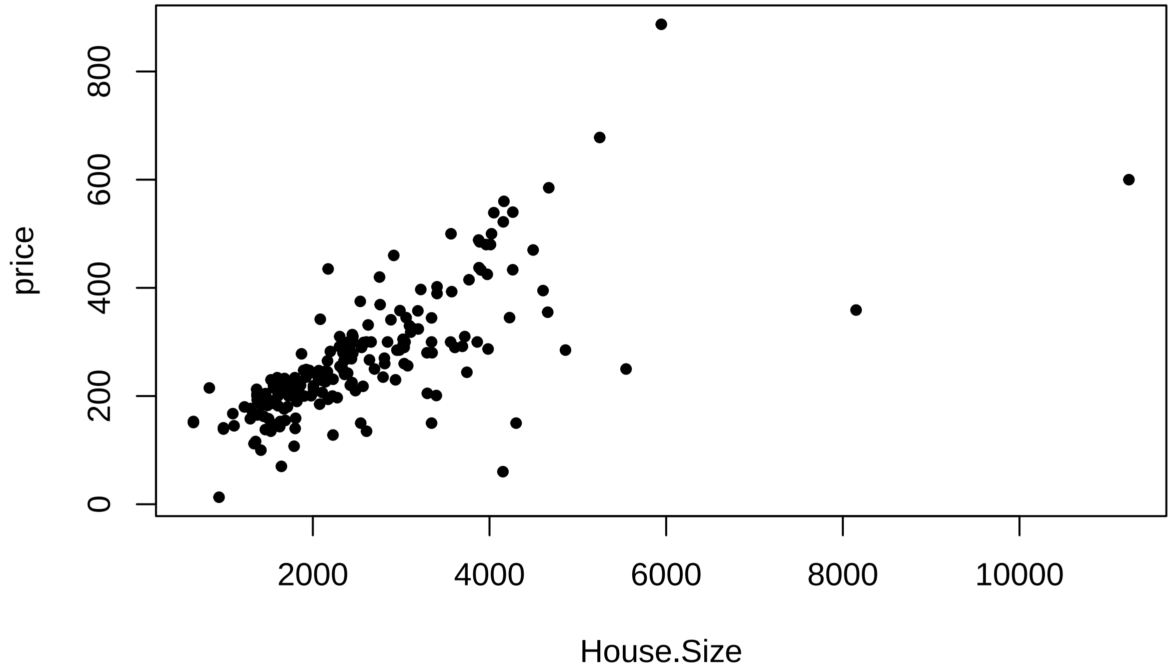
With that, it's not suitable to be the response of a linear model. This is because, very almost sure that the normality assumption will be violated (when we check the residual plots after fitting the model with **price** being the response).

3. (4 points) Plot a scatter plot between **price** and **House.Size**. Derive the correlation coefficient between them. Give your comments.

```
cor(price, House.Size) # 0.711
```

```
## [1] 0.7105131
```

```
plot(price~House.Size, pch = 20)
```



Comments: The scatter plot shows a positive and quite linear relationship between the 2 variables. The association is quite strong with correlation coefficient 0.711.

One **might** observe that the variability of price is not very stable when the house size change from small to large, a **possible** outward funnel shape, but not obviously shown.

4. (2 points) Fit a multiple linear model for **price** that depends on all the input variables in the data given, name it as **M1**. Write down the fitted model.

```r
M1 = lm(price ~ House.Size + Acres + Bedrooms +  Garage + Age.Category, data = house)
summary(M1)
```

```
##
## Call:
## lm(formula = price ~ House.Size + Acres + Bedrooms + Garage +
##     Age.Category, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -301.27  -42.27   -2.45   33.75  346.64
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.104668  22.397599   1.478   0.1410
## House.Size      0.058515   0.005383  10.871   <2e-16 ***
## Acres           5.590631   5.786251   0.966   0.3352
## Bedrooms       14.111131   5.713079   2.470   0.0144 *
## Garage1        30.728255  14.334280   2.144   0.0333 *
## Age.CategoryN  31.842047  13.346063   2.386   0.0180 *
## Age.CategoryO  -4.895751  16.285570  -0.301   0.7640
## Age.CategoryVO 56.221203  24.700374   2.276   0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.97 on 192 degrees of freedom
## Multiple R-squared:  0.5627, Adjusted R-squared:  0.5467
## F-statistic: 35.29 on 7 and 192 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\begin{aligned}\widehat{\text{price}} =& 33.105 + 0.0585 * \text{House.Size} + 5.591 * \text{Acres} + 14.111 * \text{Bedrooms} + 30.728 * I(\text{Garage} = 1) \\ & + 31.842 * I(\text{Age.Cate} = \text{N}) - 4.896 * I(\text{Age.Cate} = \text{O}) + 56.221 * I(\text{Age.Cate} = \text{VO}).\end{aligned}$$

5. (4 points) Test the significance of variable **Age.category** in model **M1** by reporting the hypotheses, the test statistic value, the null distribution and the p-value.

ANS: The coefficients of variable **Age.Category** are denoted as $\beta_5, \beta_6, \beta_7$.

Hypotheses: $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$, vs H_1: at least one of them is non-zero.

```
anova(M1)
```

```
## Analysis of Variance Table
##
## Response: price
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## House.Size     1 1347323 1347323 221.6410 < 2.2e-16 ***
## Acres          1    2515    2515   0.4137  0.520840
## Bedrooms       1   49872   49872   8.2042  0.004644 **
## Garage         1   33755   33755   5.5529  0.019458 *
## Age.Category   3   68265   22755   3.7433  0.012043 *
## Residuals    192 1167139    6079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test statistic is F, derived from Anova table.

$$F = \frac{SS_R(\text{Age.Category} \mid \text{all other variables})/3}{SS_{Res}/192} = \frac{68265/3}{1167139/192} = 3.74 \sim F_{3,192}.$$

Null distribution (the distribution of the test statistic under the null hypothesis) is $F_{3,192}$.

p-value is 0.012043.

Conclude: Variable Age.Category is significant in model M1 at $\alpha = 0.05$.

6. (2 points) Report the coefficient of variable **Garage** in model **M1** and interpret it.

ANS: coefficient of Garage in Model 1 is 30.728. That means, fixing other variables, a house with garage is predicted to have mean price higher than the one without garage by 30.728 thousand dollars.

7. (4 points) Let **SR1** denotes the standardized residuals of model **M1**. Give your comments on the adequacy of model **M1** based on the residual plots.

ANS: The residual plots are shown next page where it must have: a QQ plot of SR1; a scatter plot between SR1 and fitted value; and at least a scatter plot of SR1 with any of the 3 quantitative regressors (student could choose any scatter plot between SR1 with a quantitative regressor to plot and to give comments). In real practice, we should have all 3 scatter plots of SR1 and 3 quantitative regressors.

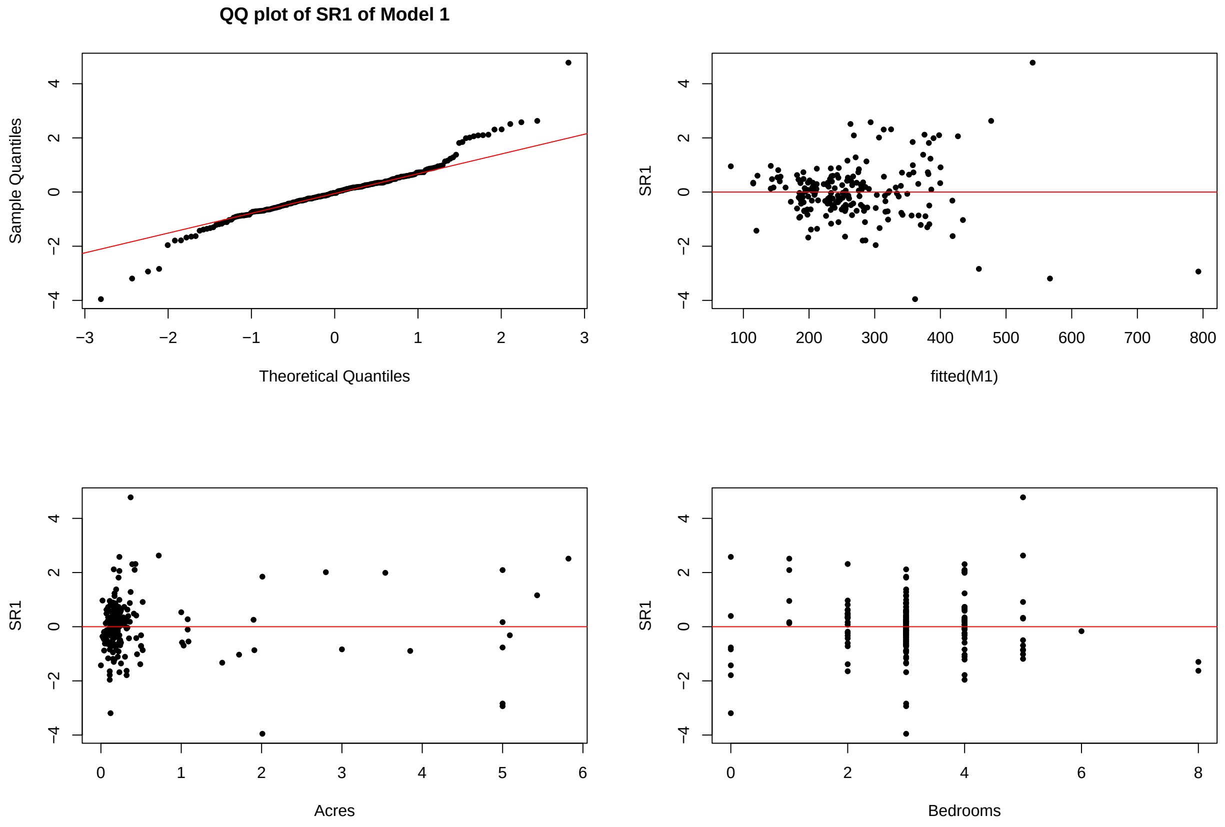
Comments: Model seems violating the normality since the QQ plot of SR1 has 2 tails might not be normal (heavy tails).

Besides, the scatter plot of SR1 vs fitted values shows some outliers (both large and small outliers) which could be the reason to cause a possible funnel shape, indicating non-constant variance. Other than that, the plot doesnt show non-linearity pattern.

The scatter plot of SR1 vs House.Size has similar pattern as SR1 vs fitted.

The scatter plot of SR1 vs Acres only shows that most of data points have small acreas, form 0 to 0.5. Other than that, the plot doesn't show any trend nor pattern, just shows some outliers.

The scatter plot of SR1 vs Bedrooms is good, except some outliers.



8. (2 points) Does model **M1** have any outliers? If yes, report the number of outliers.

ANS: Model M1 does have some outliers defined by SR1. Either way below is accepted but the latter is more commonly used and preferred.

Method 1: we could use the boxplot of SR1 to get the outliers.

Method 2: we use the rule of thumb to define outliers of a model as any points that has $|SR1| \geq 3$.

Method 1 results 16 outliers.

```
b = boxplot(SR1)$out; length(b)
```

```
## [1] 16
```

Method 2 (less strict, more flexible) results 3 outliers.

```
length(which(SR1>=3 | SR1<=(-3)) )
```

```
## [1] 3
```

9. (2 points) Does model **M1** have any influential points? If yes, report the index of influential points in the data set given.

ANS: Using the common rule that “any point that has Cook’s distance greater than 1’’ is an influential point, then Model 1 doesn’t have any influential point.

```
which(cooks.distance(M1)>1)
```

```
## named integer(0)
```

10. (10 points) Given the three houses, A, B and C, listed below. Before using model **M1** to predict the selling price for each of them, we would need to check if any of them is extrapolation.

A: House.Size = 2551, Acres = 0.4, Bedrooms = 3, Garage = 1, Age.Category = O;

B: House.Size = 600, Acres = 0.05, Bedrooms = 1, Garage = 0, Age.Category = VO;

C: House.Size = 3100, Acres = 0.3, Bedrooms = 3, Garage = 1, Age.Category = M;

(a) Using only the quantitative input variables, form Hat matrix for model **M1** in R, denoted as **H**. Report the value of the largest diagonal h_{ii} of **H**, denote it as **h.max**. **ANS:**

```
X = cbind(1, House.Size, Acres, Bedrooms)

H = X%*%solve(t(X)%*%X)%*%t(X)

h.max = max(diag(H)); h.max #0.278
```

```
## [1] 0.2778754
```

(b) With the three new points given above, form matrix **x**, then find and report the value h_{00} for each point. Any of the three houses given is extrapolation? *Hint: You could refer to Slide 50 – 54 of Topic 2 - Description.*

ANS: None of the three shourse given is extrapolation because their h_{00} are smaller than **h.max**.

```
new = data.frame(House.Size = c(2551, 600, 3100), Acres = c(0.4, 0.05, 0.3), Bedrooms = c(3,1,3), Garag

new.X = cbind(1, new$House.Size, new$Acres, new$Bedrooms)

new.H =  new.X%*%solve(t(X)%*%X)%*%t(new.X) ; new.H
```

```
##             [,1]        [,2]        [,3]
## [1,] 0.005153124 0.005808262 0.005448371
## [2,] 0.005808262 0.029688286 0.003671066
## [3,] 0.005448371 0.003671066 0.007301740
```

```
diag(new.H) # all smaller than h.max # 0.005153124 0.029688286 0.007301740
```

```
## [1] 0.005153124 0.029688286 0.007301740
```

(c) For house A, report the predicted selling price and a 99% confidence interval for its selling price.

ANS: The selling price for house A is predicted as 252.78k with 99% CI from 215.941k to 289.618k.

```r
predict(M1, newdata = new[1,], interval = "confidence", level = 0.99)
```

```
##        fit      lwr      upr
## 1 252.7796 215.9411 289.6182
```

11. (2 points) Remove the variable that is most insignificant in model **M1**, and re-fit a new model without it, called **M2**. Report the adjusted R^2 of **M2** to 4 decimal places.

ANS: The most insignificant variable in M1 is Acres with highest p-value in summary(M1). Adjusted R^2 of model M2 is 0.547.

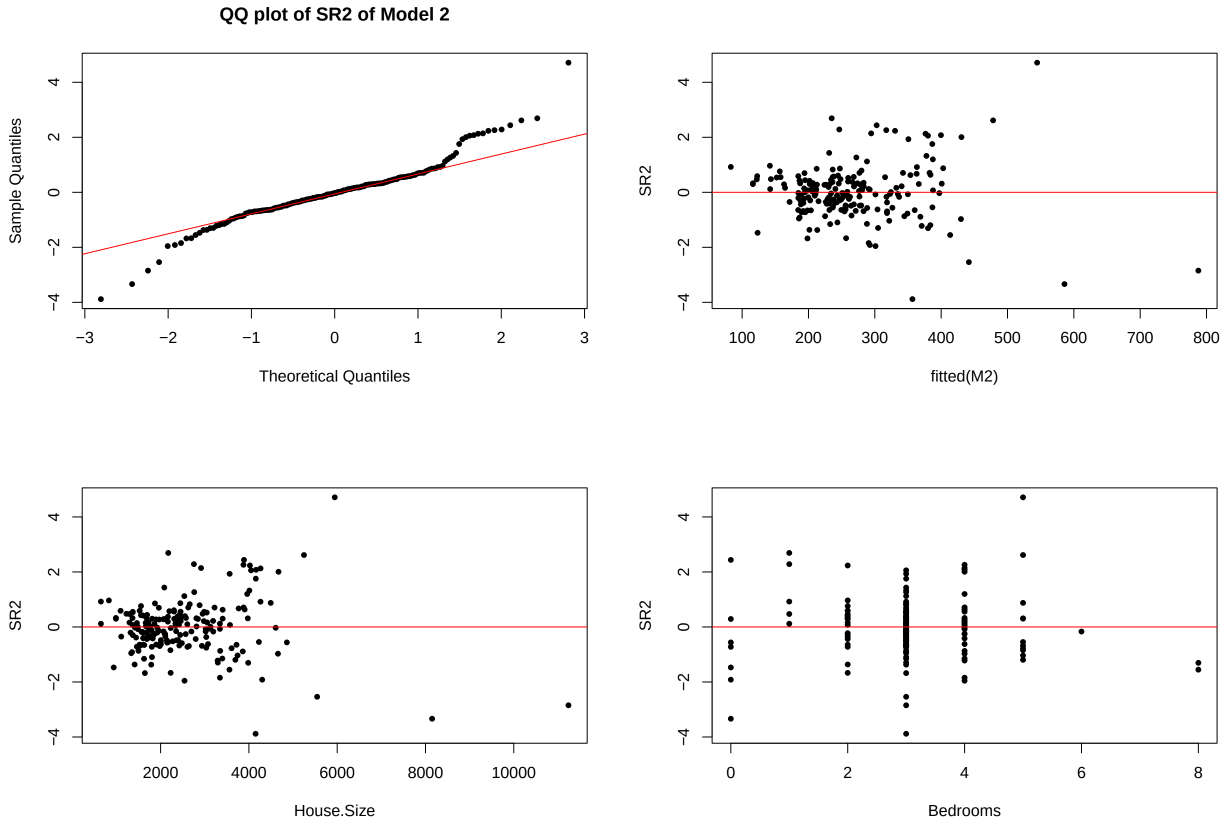
```r
M2 = lm(price ~ House.Size +  Bedrooms + Age.Category + Garage , data = house)
summary(M2)
```

```
##
## Call:
## lm(formula = price ~ House.Size + Bedrooms + Age.Category + Garage,
##     data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -296.50  -41.58   -2.11   33.11  342.51
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.883522  22.049680   1.673   0.0960 .
## House.Size      0.060819   0.004825  12.605   <2e-16 ***
## Bedrooms       12.635980   5.504345   2.296   0.0228 *
## Age.CategoryN  29.384232  13.099133   2.243   0.0260 *
## Age.CategoryO  -5.974183  16.244477  -0.368   0.7134
## Age.CategoryVO 53.356050  24.517499   2.176   0.0308 *
## Garage1        29.699655  14.292230   2.078   0.0390 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.95 on 193 degrees of freedom
## Multiple R-squared:  0.5606, Adjusted R-squared:  0.5469
## F-statistic: 41.03 on 6 and 193 DF,  p-value: < 2.2e-16
```

12. (2 points) Compare the adequacy and the goodness of fit of model **M1** and model **M2**. Which model would you prefer among these two? Explain.

ANS: Model M1 has adjusted $R^2 = 0.547$, same as that of model M2. Both model of course have significant F-test. Hence, in term of goodness-of-fit, they are similar.

In term of the adequacy, deriving the residual plots for M2 (shown next page) will help to check its adequacy.

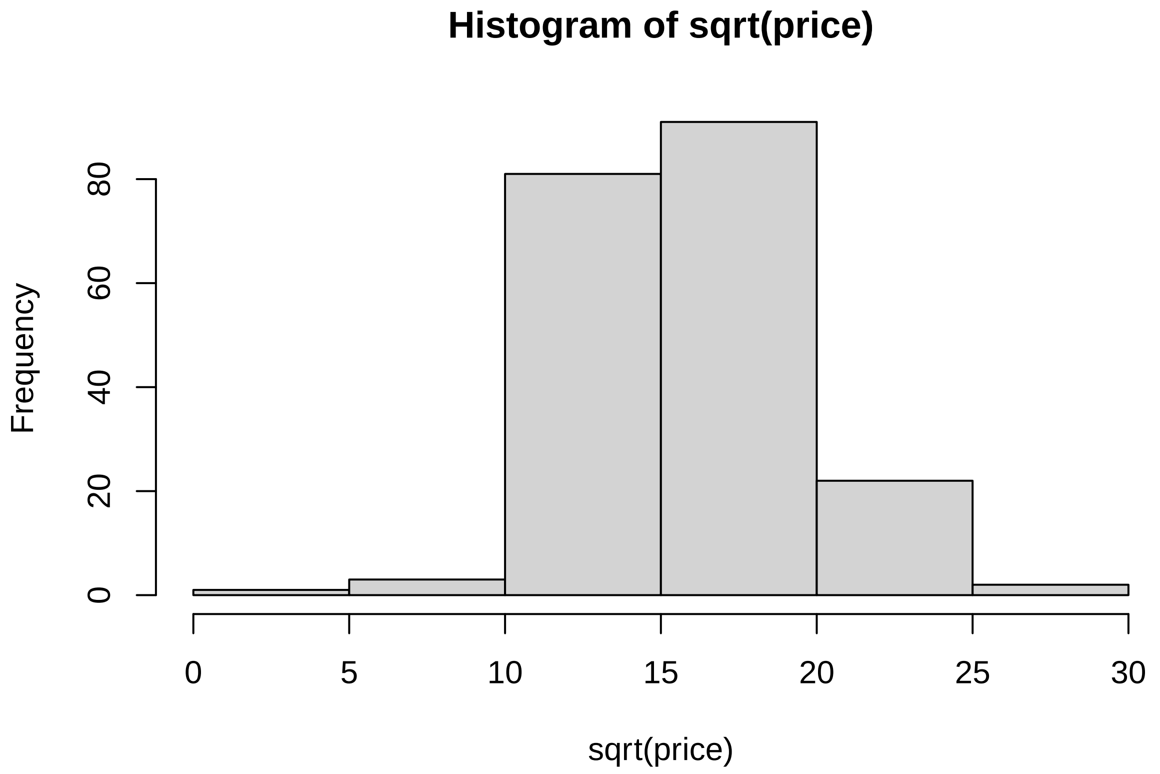


For Model 2, the model adequacy doesn't change compared to that of Model 1.

Hence, between M1 and M2, then M2 could be preferred since both the adequacy and the goodness-of-fit is similar but M2 is simpler.

13. (2 points) One suggests that we should use square root of the house's price as the response instead of the original price. Explain for that suggestion.

ANS: Consider the square root of the house's price. A histogram of it is plotted below, which is more symmetric than the histogram of price itself. Hence, $\sqrt{\text{price}}$ could be more suitable to be the response of a linear model, that might help to correct the non-normality that model M1 and model M2 have.



Q2

A data set is collected with n observations, $(x_1, y_1), ..., (x_n, y_n)$. A simple linear model $y = \beta_0 + \beta_1 x + \epsilon$ is fitted using the data set above by OLS method. The equation of the fitted model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators of β_0 and β_1 by OLS method, which yields a coefficient of determination R^2.

Recall some definitions:

For a sample $(x_1, ..., x_n)$, sample variance is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

For two samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, the sample covariance is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

For two samples $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$, the correlation coefficient is

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}}.$$

1. (3 points) Prove that $\operatorname{Cov}(y, \hat{y}) = \operatorname{Var}(\hat{y})$.

   ANS: The fitted model has the mean of residuals be zero, meaning $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$. That means, the average $\bar{y} = \bar{\hat{y}}$.

$$Var(\hat{y}) = \frac{1}{n-1}\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{n-1}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$Cov(y, \hat{y}) = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})$$

$$\begin{aligned}
Cov(y,\hat{y}) - Var(\hat{y}) &= \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) - \frac{1}{n-1}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^n \big(\hat{y}_i - \bar{y}\big)\big(y_i - \bar{y} - \hat{y}_i + \bar{y}\big) \\
&= \frac{1}{n-1}\sum_{i=1}^n \big(\hat{y}_i - \bar{y}\big)(y_i - \hat{y}_i) \\
&= \frac{1}{n-1}\sum_{i=1}^n \big(\hat{y}_i - \bar{y}\big)e_i \\
&= \frac{1}{n-1}\sum_{i=1}^n \hat{y}_i e_i + \frac{1}{n-1}\sum_{i=1}^n \bar{y} e_i \\
&= 0 + \frac{1}{n-1}\bar{y}\sum_{i=1}^n e_i \\
&= 0 + 0 = 0
\end{aligned}$$

   Note: $\sum_{i=1}^n \hat{y}_i e_i = 0$ was proved in Tutorial 1.

2. (3 points) Prove that $R^2 = [\operatorname{Cor}(y, \hat{y})]^2$.

   ANS:

$$\operatorname{Cor}(y,\hat{y})]^2 = \frac{\operatorname{Cov}(y,\hat{y})^2}{Var(y)Var(\hat{y})} = \frac{Var(\hat{y})^2}{Var(y)Var(\hat{y})} = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2)}{\sum_{i=1}^n (y_i - \bar{y})^2)} = \frac{SS_R}{SS_T} = R^2.$$

3. (4 points) Given $n = 25, R^2 = 0.9$. Test the significance of the model at $\alpha = 0.05$.

   ANS: Model is of the form $y = \beta_0 + \beta_1 x + \epsilon$, which has $k = 1$ regressor and $p = 2$ coefficients including intercept.

   The test has H_0 : all coefficients in model are zero, except β_0 vs H_1 : at least one of them is non-zero, or equivalently, we have

   $H_0 : \beta_1 = 0, \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$

   Test statistic

$$F = \frac{SS_R/k}{SS_{res}/(n-p)} = \frac{(n-p)}{k}\frac{SS_R}{SS_{res}} = \frac{(n-p)}{k}\frac{SS_R/SS_T}{SS_{res}/SS_T} = \frac{(n-p)}{k}\frac{R^2}{1-R^2} \sim F_{k,n-p}.$$

   Substitute $k = 1,\ p = 2,\ n = 25$ and $R^2 = 0.9$, we have $F = 207 \sim F_{1,23}$. The critical value is $F_{1,23}(\alpha) = F_{1,23}(0.05) = 4.28$.

   Since $F = 207 > F_{1,23}(0.05)$, we reject H_0 and conclude that the given model is significant at $\alpha = 0.05$.