Assignment 4

Amrit Bhat

2/18/2022

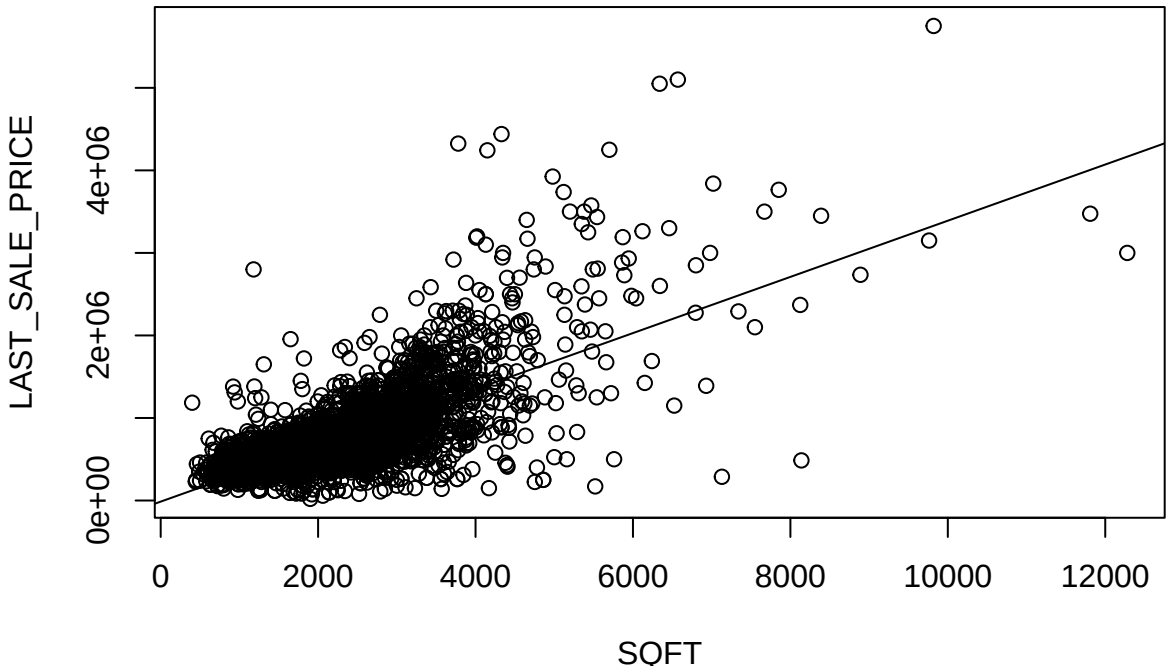
1. Calculate all pairwise correlations between all five variables.

```
cor(sales_df,use="complete.obs")
```

```
                 LAST_SALE_PRICE      SQFT  LOT_SIZE      BEDS     BATHS
  LAST_SALE_PRICE       1.0000000 0.7408940 0.1349629 0.3785385 0.5980328
  SQFT                  0.7408940 1.0000000 0.2369659 0.6360399 0.7455693
  LOT_SIZE              0.1349629 0.2369659 1.0000000 0.1770005 0.1353978
  BEDS                  0.3785385 0.6360399 0.1770005 1.0000000 0.6163141
  BATHS                 0.5980328 0.7455693 0.1353978 0.6163141 1.0000000
```

2. Make a scatterplot of the sale price versus the area of the house. Describe the association between these two variables.

```
reg1 <- lm(LAST_SALE_PRICE~SQFT,data=sales_df)
with(sales_df,plot(SQFT,LAST_SALE_PRICE))
abline(reg1)
```



We can visualize that the area of the house has a strong positive linear association with the sale price of a house. This can be confirmed by the positive correlation calculated in the previous question (0.7408940)

3. Fit a simple linear regression model (Model 1) with sale price as response variable and area of the house (SQFT) as predictor variable. State the estimated value of the intercept and the estimated coefficient for the area variable.

```
model1 <- lm(LAST_SALE_PRICE~SQFT,data=sales_df)
print(model1)
```

```

  Call:
  lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_df)

  Coefficients:
  (Intercept)         SQFT
     -13574.8        340.4
```

Estimated intercept: -13574.8 Estimated coefficient: 340.4

4. Write the equation that describes the relationship between the mean sale price and SQFT.

Linear regression equation between mean sale price vs SQFT: LAST_SALE_PRICE = -13574.8 + 340.4 * SQFT

5. State the interpretation in words of the estimated intercept.

The straight-forward interpretation of the estimated intercept is that if you don't buy a house you get 13574.8$, unlike the real world!

The mathematical interpretation is that when SQFT is 0, the LAST_SALE_PRICE is equal to the intercept (-13574.8). This is the point where the regression line crosses the y-axis.

6. State the interpretation in words of the estimated coefficient for the area variable.

The estimated coefficient in the above equation is interpreted as the average rise in the price of the house ($340.4), for every sqft increase in area of the house. This is also the slope of the regression line.

7. Add the LOT_SIZE variable to the linear regression model (Model 2). How did the estimated coefficient for the SQFT variable change?

```
model2 <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE,data=sales_df)
print(model2)
```

```

  Call:
  lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_df)

  Coefficients:
  (Intercept)         SQFT     LOT_SIZE
   -34388.445      356.615       -4.008
```

The estimated coefficient of the SQFT variable increased from 340.4 to 356.615. This also means there is a correlation between the area of the house and the area of the lot.

8. State the interpretation of the coefficient of SQFT in Model 2.

The average increase in the price/LAST_SALE_PRICE ($356.615) per unit increase in the area of the house (SQFT), given that the LOT_SIZE remains constant.

9. Report the R-squared values from the two models. Explain why they are different.

```
summary(model1)$r.squared
```

```
[1] 0.5243401
```

```
summary(model2)$r.squared
```

```
[1] 0.5511594
```

Model1 R^2: 0.5243401

Model2 R^2: 0.5511594

Upon adding the LOT_SIZE variable, the models ability to explain the variation in the price of the house (LAST_SALE_PRICE) increases.

10. Report the estimates of the error variances from the two models. Explain why they are different.

```
summary(model1)$sigma**2
```

```
[1] 99830208725
```

```
summary(model2)$sigma**2
```

```
[1] 95927158155
```

These values are different as adding more variables to the linear regression model will modify the variation of the response variable being explained by the predictor variable.

11. State the interpretation of the estimated error variance for Model 2.

The variance of Model 2 is lower compared to Model 1 as adding more variables to the regression model will only decrease the error variance.

12. Test the null hypothesis that the coefficient of the SQFT variable in Model 2 is equal to 0. (Assume that the assumptions required for the test are met.)

$H0 : \beta_1 = 0$ $H1 : \beta_1 ! = 0$

```
sqft_beta <- summary(model2)$coefficients[2]
sqft_beta_se <- summary(model2)$coefficients[5]
n=nrow(sales_df)
z = sqft_beta / sqft_beta_se
p = 2*(1-pt(abs(z),df=n-3))
data.frame(z,p)
```

```
          z p
1 69.57428 0
```

Assuming that $\alpha = 0.05$, in above result as the p-value is < 0.05, we reject the null hypothesis that the coefficient of SQFT in model 2 is 0

13. Test the null hypothesis that the coefficients of both the SQFT and LOT_SIZE variables are equal to 0. Report the test statistic.

The above question can be answered using the concept of composite hypothesis. We do the F-Test for testing the composite hypothesis that both the SQFT and LOT_SIZE variables are equal to 0.

Model2: $E(Y) = \beta_0 + \beta_1 SQFT + \beta_2 * LotSize + beta_3 * SQFT * LotSize$ Null hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Testing this hypothesis below:

```
## Handling missing data by omitting.

sales_sub_df <- na.omit(sales_df[c('LAST_SALE_PRICE','SQFT','LOT_SIZE')])
full=lm(LAST_SALE_PRICE ~ SQFT*LOT_SIZE, data=sales_sub_df)
reduced=lm(LAST_SALE_PRICE ~ 1, data=sales_sub_df)
anova(reduced,full)
```

```
  Analysis of Variance Table

  Model 1: LAST_SALE_PRICE ~ 1
  Model 2: LAST_SALE_PRICE ~ SQFT * LOT_SIZE
    Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
  1   4077 8.7092e+14
  2   4074 3.8772e+14  3 4.8319e+14 1692.4 < 2.2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("The test statistic is: ")
```

```
  [1] "The test statistic is: "
```

```
anova(full,reduced)$F[2]
```

```
  [1] 1692.388
```

```
summary(lm(len ~ supp*dose, data=ToothGrowth))$fstatistic
```

```
     value    numdf    dendf
  50.35522  3.00000 56.00000
```

Since p<0.05, we reject the null hypothesis.

14. What is the distribution of the test statistic under the null hypothesis (assuming model assumptions are met)?

The test statistic under the null hypothesis follows the F distribution with 3 numerator degrees of freedom and 56 denominator degrees of freedom

15. Report the p-value for the test in Q13.

The p-value from results in Q13 is 2.2e-16, which is < 0.05, hence, we reject the null hypothesis that the coefficients of both the SQFT and LOT_SIZE variables are equal to 0.