Data 557 HW4

Anuhya B S

2/13/2022

Data: 'Sales.csv'

The data consist of sales prices for a sample of homes from a US city and some features of the houses.

Variables:

LAST_SALE_PRICE: the sale price of the home SQFT: area of the house (sq. ft.) LOT_SIZE: area of the lot (sq. ft.) BEDS: number of bedrooms BATHS: number of bathrooms

```
sales = read.csv('Sales.csv')
colnames(sales)
```

```
## [1] "LAST_SALE_PRICE" "SQFT"            "LOT_SIZE"        "BEDS"
## [5] "BATHS"
```

```
summary(sales)
```

```
##  LAST_SALE_PRICE       SQFT          LOT_SIZE           BEDS
##  Min.   :  20100   Min.   :  400   Min.   :   446   Min.   : 0.000
##  1st Qu.: 462000   1st Qu.: 1550   1st Qu.:  4000   1st Qu.: 3.000
##  Median : 622050   Median : 2040   Median :  5500   Median : 3.000
##  Mean   : 728308   Mean   : 2189   Mean   :  6572   Mean   : 3.358
##  3rd Qu.: 830000   3rd Qu.: 2660   3rd Qu.:  7610   3rd Qu.: 4.000
##  Max.   :5750000   Max.   :12280   Max.   :120542   Max.   :11.000
##  NA's   :97        NA's   :24      NA's   :506      NA's   :8
##      BATHS
##  Min.   :0.500
##  1st Qu.:1.500
##  Median :2.000
##  Mean   :2.051
##  3rd Qu.:2.500
##  Max.   :7.750
##  NA's   :22
```

```
sales_new = na.omit(sales)
summary(sales_new)
```

```
##  LAST_SALE_PRICE       SQFT          LOT_SIZE          BEDS
##  Min.   :  79950   Min.   :  446   Min.   :  446   Min.   : 0.000
##  1st Qu.: 476950   1st Qu.: 1620   1st Qu.: 4000   1st Qu.: 3.000
##  Median : 631268   Median : 2110   Median : 5500   Median : 3.000
##  Mean   : 742552   Mean   : 2252   Mean   : 6522   Mean   : 3.408
```

```
##   3rd Qu.: 849950   3rd Qu.: 2710    3rd Qu.: 7609    3rd Qu.: 4.000
##   Max.   :5750000   Max.   :12280   Max.   :94089   Max.   :11.000
##       BATHS
##   Min.   :0.500
##   1st Qu.:1.500
##   Median :2.000
##   Mean   :2.122
##   3rd Qu.:2.500
##   Max.   :7.750
```

```r
nrow(sales_new)
```

```
## [1] 4065
```

1. Calculate all pairwise correlations between all five variables.

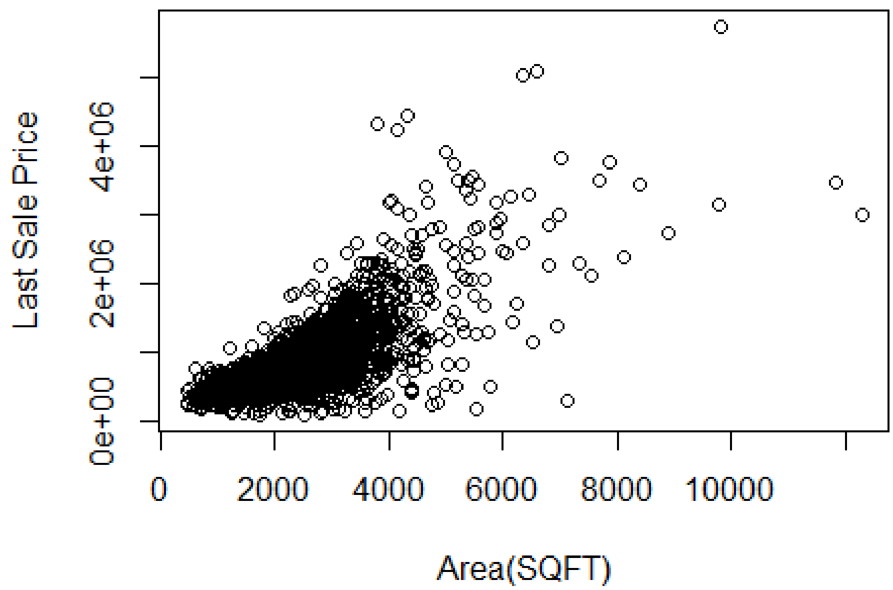
```r
cor(sales_new)
```

```
##                 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
```

The correlations between the five variables are as follows:
1. LAST_SALE_PRICE, SQFT = **0.7408940**
2. LAST_SALE_PRICE, LOT_SIZE = **0.1349629**
3. LAST_SALE_PRICE, BEDS = **0.3785385**
4. LAST_SALE_PRICE, BATHS = **0.5980328**
5. SQFT, LOT_SIZE = **0.2369659**
6. SQFT, BEDS = **0.6360399**
7. SQFT, BATHS = **0.7455693**
8. LOT_SIZE, BEDS = **0.1770005**
9. LOT_SIZE, BATHS = **0.1353978**
10. BEDS, BATHS = **0.6163141**

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

```r
plot(sales_new$LAST_SALE_PRICE ~ sales_new$SQFT, data=sales_new,xlab = "Area(
SQFT)", ylab = "Last Sale Price")
```



From the above displayed scatterplot, it can be inferred that there is a strong linear correlation between the two variables Sale Price and Area (SQFT)

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.

```
lm(LAST_SALE_PRICE ~ SQFT, data=sales_new)

##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_new)
##
## Coefficients:
## (Intercept)         SQFT
##    -47566.5        350.9
```

The estimated value of the intercerpt is **-47566.5**. The estimated coefficient for the area variable is **350.9**.

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

α is the *intercept* = -47566.5

β is the *regression coefficient* for *Area* = 350.9

The equation of the fitted line is

$$\text{sale price} = -47566.5 + 350.9 \times \text{area}$$

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

The interpretation of α is the mean of Y given $X = 0$, i.e., $\mathrm{E}(Y|X = 0) = \alpha + \beta \times 0 = \alpha$. This is the point where the regression line crosses the y-axis.

For a given data set, the fitted regression model is written as $\mathrm{E}(Y) = \hat{\alpha} + \hat{\beta}X$, where $\hat{\alpha}$ is the point where the fitted regression line crosses the y-axis and $\hat{\beta}$ is the slope of the fitted regression line.

$\hat{\alpha} = -47566.5$ is the estimated mean sale price if the area is set to 0.

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

The interpretation of β is the average *difference* in the mean of Y per unit *difference* in X.

Sometimes this is expressed as the average difference in Y corresponding to a 1-unit difference in X, i.e.,

$$\mathrm{E}(Y|X = x + 1) - \mathrm{E}(Y|X = x) = \alpha + \beta(x + 1) - (\alpha + \beta x) = \beta.$$

For a given data set, the fitted regression model is written as $\mathrm{E}(Y) = \hat{\alpha} + \hat{\beta}X$, where $\hat{\alpha}$ is the point where the fitted regression line crosses the y-axis and $\hat{\beta}$ is the slope of the fitted regression line.

$\hat{\beta} = 350.9$ is the estimated average difference in sale price per unit difference in area.

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

```r
summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_new))$coef
```

```
##                Estimate   Std. Error   t value    Pr(>|t|)
## (Intercept) -47566.522 12241.465236 -3.885689 0.000103666
## SQFT           350.909     4.990453 70.316074 0.000000000
```

```r
summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new))$co
ef
```

```
##                   Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -32579.055135 1.278808e+04 -2.547612 1.088285e-02
## SQFT           355.737262 5.127433e+00 69.379206 0.000000e+00
## LOT_SIZE        -3.965089 9.978163e-01 -3.973766 7.197273e-05
```

The estimate of the coefficient of SQFT variable is different in the two models: The estimated value in the second model is higher.

First model: The coefficient of `SQFT' is > 0 and statistically significant

Second model: The coefficient of `SQFT' is > 0 and statistically significant

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

In the first model the coefficient of `SQFT` is the average difference in sales price comparing different area sizes (in sqft).

In the second model the coefficient of `SQFT` is interpreted as the average difference in sales price comparing different area sizes(in sqft) **having the same lot size(in sqft)**.

Due to the addition of the lot size, there is a certain amount if change in the coefficient of the Area variable however, this addition does not have a significant impact on the estimated coefficient of area i.e. the Lot size variable does not have a confounding effect.

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

```r
summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_new))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2166915  -147629    -9306   124458  3046130
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -47566.52   12241.47  -3.886 0.000104 ***
## SQFT           350.91       4.99  70.316  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 309700 on 4063 degrees of freedom
## Multiple R-squared:  0.5489, Adjusted R-squared:  0.5488
## F-statistic:  4944 on 1 and 4063 DF,  p-value: < 2.2e-16
```

```r
summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new))
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2162244  -146163   -11297   119938  3333236
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.258e+04  1.279e+04  -2.548   0.0109 *
## SQFT         3.557e+02  5.127e+00  69.379  < 2e-16 ***
## LOT_SIZE    -3.965e+00  9.978e-01  -3.974  7.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 309100 on 4062 degrees of freedom
## Multiple R-squared:  0.5507, Adjusted R-squared:  0.5504
## F-statistic:  2489 on 2 and 4062 DF,  p-value: < 2.2e-16
```

The R^2 value from the first model:$R^2 = 0.5489$.
The R^2 value from the second model:$R^2 = 0.5507$.

For simple linear regression models, the R-squared is just the square of the Pearson correlation coefficient. For models with more than 1 predictor R-squared has an interpretation in terms of correlation between observed and fitted values and also as a percentage of variance explained by the model. The R squared values are different for the two models as one is a simple linear regression model with one variable and the other is a model having two predictors.

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

The error variance is the variance of the errors ϵ_i, is calculated using the sum of squares of residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p},$$

```
(summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_new))$sigma)**2

## [1] 95895947932

(summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new))$s
igma)**2

## [1] 95548117507
```

The estimated error variance of Model 1 is **95895947932**. The estimated error variance of Model 2 is **95548117507**.

The estimated variance basically tells you about the variance of the standard errors. The estimated variance of the first model tells us about the variance of the standard errors when we take only one predictor into consideration. The estimated variance of the second model tells us about the variance of the standard error when we take take 2 predictors (SQFT and LOT_SIZE) into consideration which is the reason why that the standard error variance differ for the two models.

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

Estimated variance essentially tells us about the variance of the residuals. In the case of model two, there are multiple predictors. In such a case, the standard errors do not depend on just the sums of squares of the standard error but also on the sums of cross-products of the different predictor variables.

The standard error of the regression coefficient can change when a variable is added to the modeled and whether or not it changes depends on the the sum of the squares of cross -

products of predictors as well as whether the estimated of error variance changes. In model two, we can see that the estimated error variance has changes, which indicates that the standard error of the regression model has also changed.

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.)

The full model is

$$\text{sale price} = \beta_0 + \beta_1 \times \text{area} + \beta_2 \times \text{lot\_size}$$

Testing that the coefficient of the SQFT variable is 0 in the model, the null hypothesis is

$$H_0 : \beta_1 = 0$$

The reduced model is

$$\text{sale price} = \beta_0 + \beta_2 \times \text{lot\_size}$$

The F-test for full model is

```
options(scipen = 999)
anova(lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new))#["Residuals","
Sum Sq"]
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df          Sum Sq         Mean Sq  F value                Pr(
>F)
## SQFT         1 474143156081999 474143156081999 4962.350 < 0.00000000000000
022
## LOT_SIZE     1   1508783132972   1508783132972   15.791            0.00007
197
## Residuals 4062 388116453312974     95548117507
##
## SQFT      ***
## LOT_SIZE  ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test for reduced models

```
anova(lm(LAST_SALE_PRICE ~ LOT_SIZE, data = sales_new))
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df          Sum Sq        Mean Sq F value                Pr(>F
)
## LOT_SIZE     1  15733534826184 15733534826184  75.381 < 0.0000000000000002
```

```
2 ***
## Residuals 4063 848034857701759   208721353114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is defined as:

$$F = \frac{(SSE_0 - SSE_1)/(df_1 - df_0)}{SSE_1/(df_1)}$$

The F-test for comparing full and reduced models

```
((848034857701759-388116453312974)/(4063-4062))/(388116453312974/4062)

## [1] 4813.474
```

The p-value obtained for the tail probability for the value **4813.474 in the F-distribution with 1 numerator df and 4062 denominator df** is:

```
1-pf(4813.474,1,4062)

## [1] 0
```

We **reject the null hypothesis** as the p value is less than the level of significance which mean that the SQFT variable is statistically significant and there is evidence for association between the SQFT and Last Sale Price.

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 full model is

$$\text{sale price} = \beta_0 + \beta_1 \times \text{area} + \beta_2 \times \text{lot\_size}$$

Testing that the coefficient of the SQFT and LOT_SIZE variable is 0 in the model, the null hypothesis is

$$H_0: \beta_1 = \beta_2 = 0$$

The reduced model is

$$\text{sale price} = \beta_0$$

The F-test for full model is

```
anova(lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_new))

## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df          Sum Sq         Mean Sq  F value                Pr(
>F)
## SQFT         1 474143156081999 474143156081999 4962.350 < 0.00000000000000
```

```
022
## LOT_SIZE     1   1508783132972   1508783132972   15.791            0.00007
197
## Residuals 4062 388116453312974     95548117507
##
## SQFT      ***
## LOT_SIZE  ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test for reduced model is

```
anova(lm(LAST_SALE_PRICE ~ 1, data = sales_new))
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df          Sum Sq      Mean Sq F value Pr(>F)
## Residuals 4064 863768392527944 212541435169
```

The F-test for comparing full and reduced models

```
((863768392527944-388116453312974)/(4064-4062))/(388116453312974/4062)
```

```
## [1] 2489.07
```

The F-statistic is **2489.07 with 2 numerator df and 4062 denominator df**.

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

The F-statistic is referred to the $F_{p_1-p_0,n-p_1}$ distribution for calculation of the p-value : $F_{2,4062}$. This means that assuming that the model assumptions are met, we need to find the p value for the tail probability for the value 2489.07 in the F-distribution with 2 numerator df and 4062 denominator df.

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

The p-value obtained for the tail probability for the value 2489.07 in the F-distribution with 2 numerator df and 4062 denominator df is:

```
1-pf(2489.07,2,4062)
```

```
## [1] 0
```

The p value is 0.

We **reject the null hypothesis** as the p value is less than the level of significance which mean that the SQFT and LOT_SIZE variables are statistically significant and there is evidence for association between the SQFT, LOT_SIZE and Last Sale Price.