DATA 557: HW Assignment 4

Hriday Baghar

February 17, 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")
str(sales)
```

```
## 'data.frame':    4695 obs. of  5 variables:
##  $ LAST_SALE_PRICE: int  410000 229000 370000 436000 415000 505000 550000 385000 365000 399000 ...
##  $ SQFT           : int  950 446 1400 1610 2520 2048 2140 1270 1080 2280 ...
##  $ LOT_SIZE       : int  6697 446 6500 7200 14000 7200 13264 NA 8645 14850 ...
##  $ BEDS           : int  1 0 3 4 4 3 3 3 3 3 ...
##  $ BATHS          : num  1 1 1 2 2.75 1.75 1.5 1 1 1.75 ...
```

```
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
```

We see that variables have missing values. For the scope of this assignment, we will remove all records with any missing values.

```
sales <- na.omit(sales)
str(sales)
```

```
## 'data.frame':    4065 obs. of  5 variables:
##  $ LAST_SALE_PRICE: int  410000 229000 370000 436000 415000 505000 550000 365000 399000 400000 ...
##  $ SQFT           : int  950 446 1400 1610 2520 2048 2140 1080 2280 1940 ...
##  $ LOT_SIZE       : int  6697 446 6500 7200 14000 7200 13264 8645 14850 12061 ...
##  $ BEDS           : int  1 0 3 4 4 3 3 3 3 4 ...
```

```
##  $ BATHS          : num  1 1 1 2 2.75 1.75 1.5 1 1.75 1.75 ...
##  - attr(*, "na.action")= 'omit' Named int  8 26 44 49 52 58 77 88 95 103 ...
##   ..- attr(*, "names")= chr  "8" "26" "44" "49" ...
```

```r
summary(sales)
```

```
##  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
```

1. Calculate all pairwise correlations between all five variables.

```r
combs <- combn(names(sales), 2)
corr.df <- data.frame(variable1 = combs[1,], varaible2 = combs[2,], corr = rep(NA, length(combs[1,])),
                      stringsAsFactors = FALSE)

for(row in 1:nrow(corr.df)){
  x <- sales[,corr.df[row,1]]
  y <- sales[,corr.df[row,2]]
  corr.df[row,"corr"] <- cov(x,y)/(sd(x)*sd(y))
}
corr.df
```

```
##          variable1 varaible2      corr
## 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
```

```r
#Sanity check
cor(sales)
```

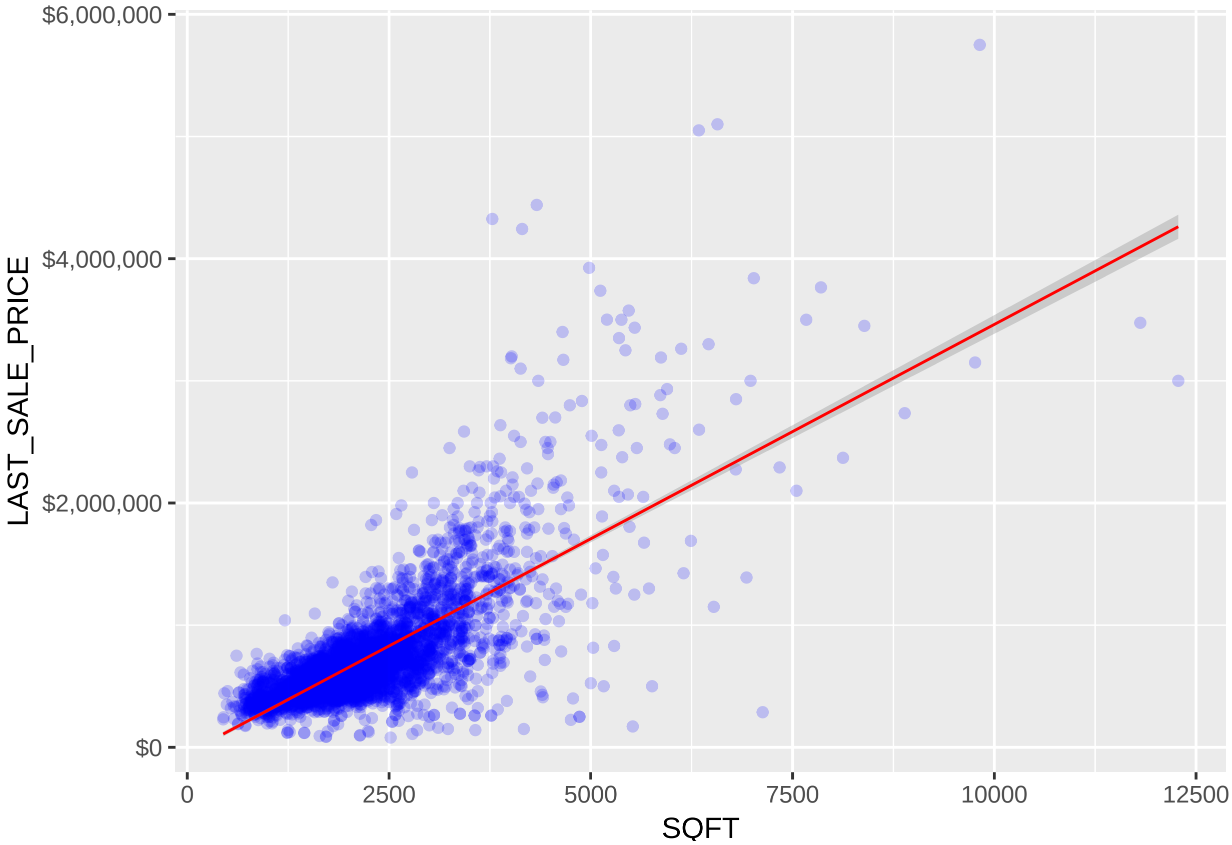
```
##                 LAST_SALE_PRICE      SQFT  LOT_SIZE      BEDS     BATHS
## LAST_SALE_PRICE       1.0000000 0.7408940 0.1349629 0.3785385 0.5980328
## SQFT                  0.7408940 1.0000000 0.2369659 0.6360399 0.7455693
## LOT_SIZE              0.1349629 0.2369659 1.0000000 0.1770005 0.1353978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

```
library(ggplot2)
```

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

```
scatter <- ggplot(data = sales, aes(y = LAST_SALE_PRICE, x = SQFT)) +
  scale_y_continuous(labels = scales::dollar_format()) +
  geom_point(alpha = 0.2, color = "blue") +
  geom_smooth(method = "lm", color = "red", size = 0.5)
scatter
```



There is a positive correlation between sale price and area. Most houses are priced under $2,000,000 and have area under 3750 sq. ft.

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)
summary(model1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales)
##
```

```
## 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
```

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

$$\hat{Price} = -47,566.52 + 350.91 \times Area$$

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

For an area of 0 sq. ft. the estimated mean sale price is -$47,566. This is due to extrapolation of data beyond the actual range.

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

For every unit increase in area, the sale price increases by $350.91.

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

```r
model2 <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales)
summary(model2)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales)
##
## 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 estimated coefficient for SQFT increased a little bit.

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

The average difference in sale price for houses at a given lot size is $355.7 per unit increase in area.

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

For the model with only area as a predictor variable $R^2 = 0.5489$. For the model with area and lot size as predictor variables $R^2 = 0.5507$

The R^2 value is a measure of how much variation in the response is explained by the model. A higher R^2 for model 2 corresponds to higher variation explained by the model on adding the lot size variable, i.e. a better fit.

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

```
(err.var.1 <- sum((model1$residuals)^2)/(nrow(sales)-length(model1$coefficients)))
```

```
## [1] 95895947932
```

```
(err.var.2 <- sum((model2$residuals)^2)/(nrow(sales)-length(model2$coefficients)))
```

```
## [1] 95548117507
```

```
err.var.1 < err.var.2
```

```
## [1] FALSE
```

```
#Sanity check

#Refer to Sum Sq value for residuals
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## SQFT         1 4.7414e+14 4.7414e+14  4944.4 < 2.2e-16 ***
## Residuals 4063 3.8963e+14 9.5896e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##             Df     Sum Sq    Mean Sq  F value    Pr(>F)
## SQFT         1 4.7414e+14 4.7414e+14 4962.350 < 2.2e-16 ***
## LOT_SIZE     1 1.5088e+12 1.5088e+12   15.791 7.197e-05 ***
## Residuals 4062 3.8812e+14 9.5548e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Error variance is lower in model 2 than 1 which reduces the standard error for $\hat{\beta}$. Model 2 would be a better fit than model 1 based on this.

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.)

```r
summary(model2)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales)
##
## 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
```

From the above summary, we see that the test statistic for SQFT is 69.397 and the $p - value\, is < 2 \times 10^{-16}$. In conclusion we reject the null hypothesis that the coefficient of SQFT is 0. There is an evidence of linear association between area and 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.

```r
reduced.model <- lm(LAST_SALE_PRICE ~ 1, data = sales)
anova(model2, reduced.model)
```

```
## Analysis of Variance Table
##
## Model 1: LAST_SALE_PRICE ~ SQFT + LOT_SIZE
## Model 2: LAST_SALE_PRICE ~ 1
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1   4062 3.8812e+14
## 2   4064 8.6377e+14 -2 -4.7565e+14 2489.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F - statistic = 2489.1$ this is same as the F-statistic in model 2's summary function. We reject the null hypothesis that the coefficient for area and lot size is 0.

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

The test statistic follows an F-distribution with 4062 degrees of freedom under the null hypothesis.

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

The p-value for this test is 2.2×10^{-16}