Homework Assignment 4

Shrusti Ghela

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:

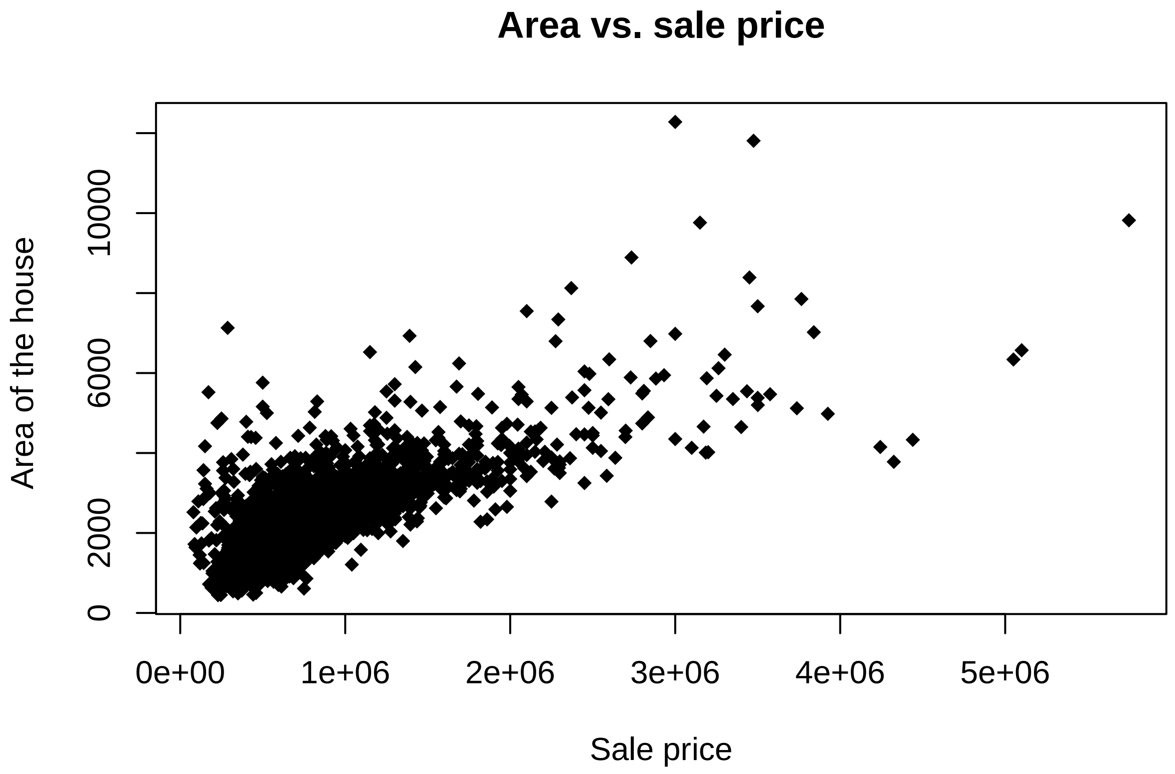
LAST_SALE_PRICE: the sale price of the home SQFT: area of the house (sq. ft.) LOT_SIZE: area of the lot (sq. ft.) BEDS: number of bedrooms BATHS: number of bathrooms

1. Calculate all pairwise correlations between all five variables.

```
##                 LAST_SALE_PRICE   SQFT LOT_SIZE   BEDS  BATHS
## LAST_SALE_PRICE          1.0000 0.7409   0.1350 0.3785 0.5980
## SQFT                     0.7409 1.0000   0.2370 0.6360 0.7456
## LOT_SIZE                 0.1350 0.2370   1.0000 0.1770 0.1354
## BEDS                     0.3785 0.6360   0.1770 1.0000 0.6163
## BATHS                    0.5980 0.7456   0.1354 0.6163 1.0000
```

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

```
plot(sales_clean$LAST_SALE_PRICE, sales_clean$SQFT, main="Area vs. sale price",
   xlab="Sale price", ylab="Area of the house ", pch=18)
```



According to the scatter plot, there seems to be a linear relationship between area and the sale price.

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_clean)
model1
```

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

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

LAST_SALE_PRICE = - 47566.5 + 350.9*SQFT

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

The interpretation of α is the mean of Y given $X = 0$, i.e., $E(Y|X = 0) = \alpha + \beta \times 0 = \alpha$. This is the point where the regression line crosses the y-axis. The mean of LAST_SALE_PRICE given SQFT=0 is - 47566.5

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. The average difference in the mean of LAST_SALE_PRICE per unit difference in SQFT is 350.9

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_clean)
model2
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_clean)
##
## Coefficients:
## (Intercept)         SQFT     LOT_SIZE
##  -32579.055      355.737       -3.965
```

Estimated coefficient of SQFT for model 1 = 350.9 Estimated coefficient of SQFT for model 1 = 355.737 There is a little change between the estimated coefficient for the SQFT variable.

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

The average difference in the mean of LAST_SALE_PRICE per unit difference in SQFT is 355.737

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

```r
summary(model1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales_clean)
##
## 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(model2)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales_clean)
##
## 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-squared values from the two models are different because model 2 has an additional variable(LOT_SIZE) which changes how much variation in the response is explained by the model.

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

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

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

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

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

The error variance is the variance of the errors. We estimate it using the sum of squares of residuals. Since we add one more variable in the model 2, there is a change in the error variance. Because we add one extra variable that explains something extra about the predictor, we see that the error variance is reduced. That means the the part that we can't explain is reduced.

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

If we have a model, we can explain part of the variance of the response from the variance of predictors. The part we can't explain is error variance. The variance of the errors for the model 2 is 95548117507.

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.)

Full-Model: $\mathrm{E}(LAST\_SALE\_PRICE) = \beta_0 + \beta_1 SQFT + \beta_2 LOT\_SIZE$

Null hypothesis: $H_0 : \beta_1 = 0$.

Reduced-Model: $\mathrm{E}(LAST\_SALE\_PRICE) = \beta_0 + \beta_2 LOT\_SIZE$

ANOVA table for the full model

```
anova(lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data=sales_clean), options(scipen=999))
```

```
## Warning in anova.lmlist(object, ...): models with response '"NULL"' removed
## because response differs from model 1
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##              Df          Sum Sq         Mean Sq  F value                 Pr(>F)
## SQFT         1 474143156081999 474143156081999 4962.350 < 0.0000000000000022
## LOT_SIZE     1   1508783132972   1508783132972   15.791            0.00007197
## Residuals 4062 388116453312974     95548117507
##
## SQFT      ***
## LOT_SIZE  ***
## Residuals
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

ANOVA table for the reduced model

```
anova(lm(LAST_SALE_PRICE ~ LOT_SIZE, data=sales_clean), options(scipen=999))
```

```
## Warning in anova.lmlist(object, ...): models with response ’"NULL"’ removed
## because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##              Df          Sum Sq        Mean Sq F value                 Pr(>F)
## LOT_SIZE     1  15733534826184 15733534826184  75.381 < 0.0000000000000022 ***
## Residuals 4063 848034857701759   208721353114
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

F-test for comparing full and reduced model

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

```
## [1] 4813.474
```

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

```
pf(4813.474, 1, 4062, lower.tail=FALSE)
```

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

We would reject the null hypothesis that the coefficient of the SQFT variable in Model 2 is equal to 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.

Full-Model: E$(LAST_SALE_PRICE) = \beta_0 + \beta_1 SQFT + \beta_2 LOT_SIZE$

Null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$.

Reduced-Model: E$(LAST_SALE_PRICE) = \beta_0$

ANOVA table for full-model

```
anova(lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data=sales_clean), options(scipen=999))
```

```
## Warning in anova.lmlist(object, ...): models with response '"NULL"' removed
## because response differs from model 1

## Analysis of Variance Table
##
## Response: LAST_SALE_PRICE
##              Df          Sum Sq         Mean Sq  F value                  Pr(>F)
## SQFT          1 474143156081999 474143156081999 4962.350 < 0.0000000000000022
## LOT_SIZE      1   1508783132972   1508783132972   15.791             0.00007197
## Residuals 4062 388116453312974     95548117507
##
## SQFT      ***
## LOT_SIZE  ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA table for reduced model

```
anova(lm(LAST_SALE_PRICE ~ 1, data=sales_clean), options(scipen=999))
```

```
## Warning in anova.lmlist(object, ...): models with response '"NULL"' removed
## because response differs from model 1

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

F-test for comparing full and reduced model

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

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

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

```
pf(2489.07, 2, 4062, lower.tail=FALSE)
```

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

We would reject the null hypothesis that the coefficient of both the SQFT and LOT_SIZE variables are equal to 0.

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

To test a null hypothesis we compare the sums of squares of residuals for the *full* model, which includes the coefficients being tested, with the *reduced* model, which has those coefficients set to 0.

If we define SSE_0 and SSE_1 as the sums of squares of residuals for the reduced and full models, respectively, the F-statistic is defined as:

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n - p_1)}$$

The F-statistic is referred to the $F_{p_1-p_0,n-p_1}$ distribution for calculation of the p-value.

So, the F-statistic is $F_{2,4062}$

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

```
pf(2489.07, 2, 4062, lower.tail=FALSE)
```

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