DATA 557: Applied Statistics and Experimental Design
Homework 4

Abhishek Saini

February 2022

Question 1

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

0.1 Calculate all pairwise correlations between all five variables.

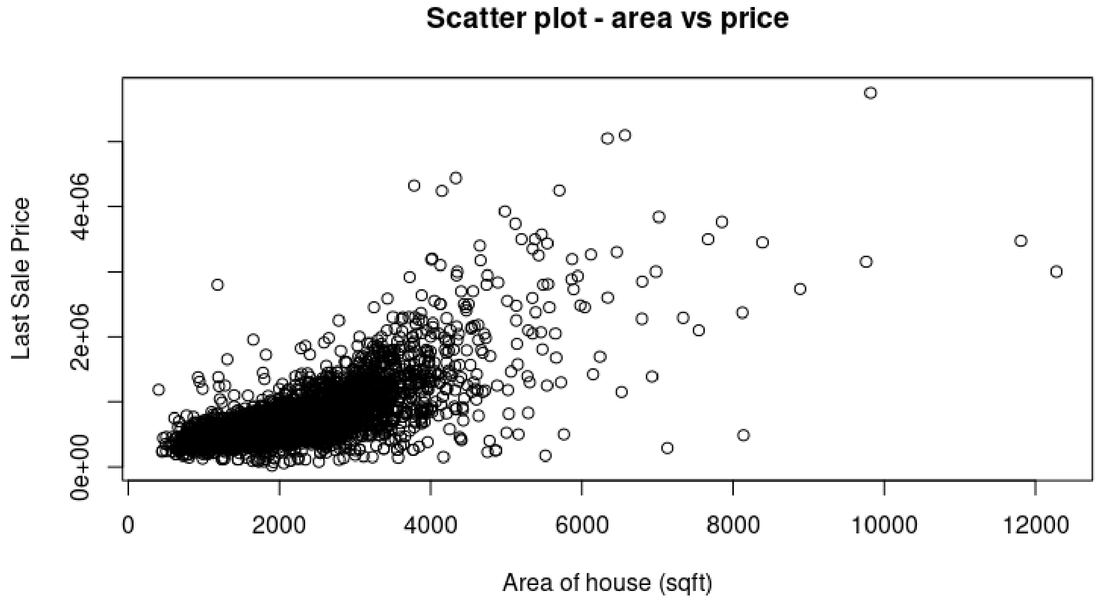
The pairwise correlations are shown below.

```
> sales = read.csv("sales.csv")
> cor(sales, 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
```

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

Sale price and area of the house are positively correlated with coefficient of correlation = 0.7408940.

```
plot(sales$SQFT, sales$LAST_SALE_PRICE, main="Scatterplot Example",
     xlab="Area of house (sqft)", ylab="Last Sale Price", pch=1)
```



0.3 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.

The estimated intercept is -49597.0880 and the estimated coefficient of the area variable is 351.7581.

```
> sales = sales[! (is.na(sales$SQFT) | is.na(sales$LAST_SALE_PRICE) | is.na(sales$LOT_SIZE)),]
> summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales))$coef
               Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -49597.0880 12234.793674 -4.053774 5.133843e-05
SQFT           351.7581     4.989883 70.494256 0.000000e+00
```

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

$$
\begin{aligned}
E[LAST\_SALE\_PRICE\,|\,SQFT] &= \hat{\alpha} + \hat{\beta} \times SQFT \\
&= -49597.0880 + 351.7581 \times SQFT
\end{aligned}
$$

0.5 State the interpretation in words of the estimated intercept.

$\hat{\alpha}$ = -49597.0880 is the estimated mean last sale price if the area in sqft is set to 0.

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

$\hat{\beta} = 351.7581$ is the estimated average difference in last sale price per unit difference in area (sqft).

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

The estimated coefficient for the SQFT variable are both positive and not very different in both the models.

```
> summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales))$coef
                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -34388.445095 1.278692e+04 -2.689344 7.188411e-03
SQFT           356.615404 5.125678e+00 69.574283 0.000000e+00
LOT_SIZE        -4.008392 9.990188e-01 -4.012329 6.120175e-05
```

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

In the second model the coefficient of SQFT is interpreted as the average difference in LAST_SALE_PRICE per unit difference in area(SQFT) for houses of the same LOT_SIZE.

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

The R-squared for the second model increases to 0.5511594 compared to that of 0.5493862 for the first model. The increase is due to the addition of another variable which helps explain more of the variance in the predictor variable.

```
> summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales))$r.squared
[1] 0.5493862
> summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales))$r.squared
[1] 0.5511594
```

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

The estimated error variance for the first model is 96282502457.
The estimated error variance for the second model is 95927158155.
The second model more fully captures the effects of the predictors and so gives a better estimate of pure error variance, compared with the first model which has one lesser predictor variable.

```
> model1 = summary(lm(formula = LAST_SALE_PRICE ~ SQFT, data = sales))
> model2 = summary(lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE, data = sales))
> sum(model1$residuals*model1$residuals)/(length(model1$residuals) - 2)
[1] 96282502457
> sum(model2$residuals*model2$residuals)/(length(model2$residuals) - 3)
[1] 95927158155
```

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

The estimated error variance can be interpreted as the variance of the errors ϵ_i, which we estimate using the sum of squares of residuals.

0.12 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.)

Since the p-value - 0.000000e+00 is very small, we reject the null hypothesis that the mean last sale price is same for any area of the house.

```
> model2$coefficients
                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -34388.445095 1.278692e+04 -2.689344 7.188411e-03
SQFT           356.615404 5.125678e+00 69.574283 0.000000e+00
LOT_SIZE        -4.008392 9.990188e-01 -4.012329 6.120175e-05
```

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.

Using Composite Hypothesis test, we get the F-statistic = 2501.974. The null hypothesis of our composite test is that both SQFT and LOT_SIZE have no linear association with LAST_SALE_PRICE.

```
> # full model
> SSE1 = sum(model2$residuals*model2$residuals)
> p1 = 3
>
> # reduced model
> last_sale_price = sales[!is.na(sales$LAST_SALE_PRICE),]$LAST_SALE_PRICE
> mean0 = mean(last_sale_price)
> res0 = last_sale_price - mean0
> SSE0 = sum(res0*res0)
> p0 = 1
>
```

```
> F_statistic = (SSE0 - SSE1)/(p1-p0)/SSE1*model2$df[2] ;F_statistic
[1] 2501.974
```

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

The distribution of the test statistic under the null hypothesis is F-distribution with 2 numerator df and 4075 denominator df.

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

The p-value for the test is 0. Based on the extremely low p-value, we can reject the null hypothesis.

```
> pf(F_statistic, p1-p0, model2$df[2], lower.tail = FALSE)
[1] 0
```