ECON_203_HW5_Ancel_Charles

Charles Ancel

2024-03-02

```
library(readxl)
```

```
# Load the dataset
data <- read_excel("housing.xls")
```

(a) Compute the average price for colonial and non-colonial houses:

```
average_price_colonial <- mean(data$price[data$colonial == 1])
average_price_non_colonial <- mean(data$price[data$colonial == 0])

average_price_colonial
```

```
## [1] 302.9189
```

```
average_price_non_colonial
```

```
## [1] 272.3704
```

(b) Estimate a linear model relating house prices to the style of the house (colonial or non-colonial):

```
# Manual calculation
n <- length(data$price)
b1 <- (sum((data$price - mean(data$price)) * (data$colonial - mean(data$colonial)))) / (sum((data$colon
b0 <- mean(data$price) - b1 * mean(data$colonial)

# Using lm
linear_model <- lm(data$price ~ data$colonial)

b0
```

```
## [1] 272.3704
```

```
b1
```

```
## [1] 30.5485
```

```
n
```

```
## [1] 88
```

(c) Easier way to estimate a linear model in R:

```
linear_model <- lm(price ~ colonial, data = data)
coefficients(linear_model)
```

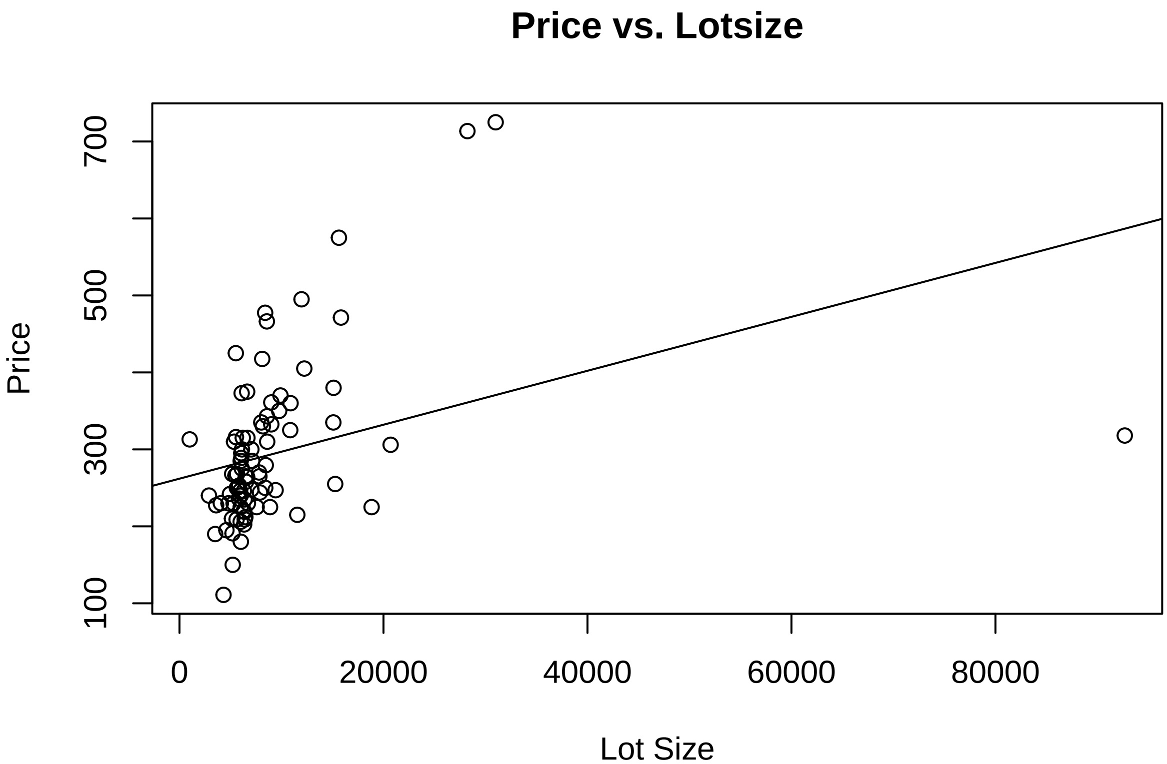
```
## (Intercept)    colonial
##    272.3704     30.5485
```

(d) Replace Colonial by sqrft and draw a scatter plot with the best-fitting line:

```r
linear_model_sqrft <- lm(price ~ sqrft, data = data)
plot(data$sqrft, data$price, main = "Price vs. sqrft", xlab = "Square Footage", ylab = "Price")
abline(linear_model_sqrft)
```

(e) Replace sqrft by lotsize and draw the corresponding scatter plot:

```r
linear_model_lotsize <- lm(price ~ lotsize, data = data)
plot(data$lotsize, data$price, main = "Price vs. Lotsize", xlab = "Lot Size", ylab = "Price")
abline(linear_model_lotsize)
```



Part (a): Average price for colonial houses

```r
average_price_colonial <- mean(data$price[data$colonial == 1])
cat("The average price for colonial houses is", round(average_price_colonial, 2), "\n")
```

```
## The average price for colonial houses is 302.92
```

Part (b): Average price for non-colonial houses

```r
average_price_non_colonial <- mean(data$price[data$colonial == 0])
cat("The average price for non-colonial houses is", round(average_price_non_colonial, 2), "\n")
```

```
## The average price for non-colonial houses is 272.37
```

Part (c): Estimated b0 and b1 for the model with the colonial variable

```r
linear_model_colonial <- lm(price ~ colonial, data = data)
coefficients_colonial <- coefficients(linear_model_colonial)
b0_colonial <- round(coefficients_colonial["(Intercept)"], 2)
b1_colonial <- round(coefficients_colonial["colonial"], 2)
cat("The estimated b0 and b1 for the colonial model are", b0_colonial, "and", b1_colonial, "\n")
```

```
## The estimated b0 and b1 for the colonial model are 272.37 and 30.55
```

Part (d): Is b0 the average price of non-colonial houses? Since b0_colonial is the intercept and represents the price when colonial is 0, it should be the average price of non-colonial houses

```r
average_non_colonial_check <- round(b0_colonial, 2) == round(average_price_non_colonial, 2)
cat("The statement that b0 is the average price of non-colonial houses is", average_non_colonial_check,
```

```
## The statement that b0 is the average price of non-colonial houses is TRUE
```

Part (e): Is b1 the average price of colonial houses? b1 represents the difference in price due to the house being colonial. It is not the average price of colonial houses.

```r
average_colonial_check <- round(b1_colonial, 2) == round(average_price_colonial, 2)
cat("The statement that b1 is the average price of colonial houses is", average_colonial_check, "\n")
```

```
## The statement that b1 is the average price of colonial houses is FALSE
```

Part (f): Model with sqrft

```r
linear_model_sqrft <- lm(price ~ sqrft, data = data)
coefficients_sqrft <- coefficients(linear_model_sqrft)
b0_sqrft <- round(coefficients_sqrft[1], 2)
b1_sqrft <- round(coefficients_sqrft[2], 2)
```

Part (g): Model with lotsize

```r
linear_model_lotsize <- lm(price ~ lotsize, data = data)
coefficients_lotsize <- coefficients(linear_model_lotsize)
b0_lotsize <- round(coefficients_lotsize[1], 2)
b1_lotsize <- round(coefficients_lotsize[2], 2)

# Output the coefficients for sqrft and lotsize models
cat("For sqrft model, b0 (intercept) is", b0_sqrft, "and b1 (slope) is", b1_sqrft, "\n")
```

```
## For sqrft model, b0 (intercept) is 11.2 and b1 (slope) is 0.14
```

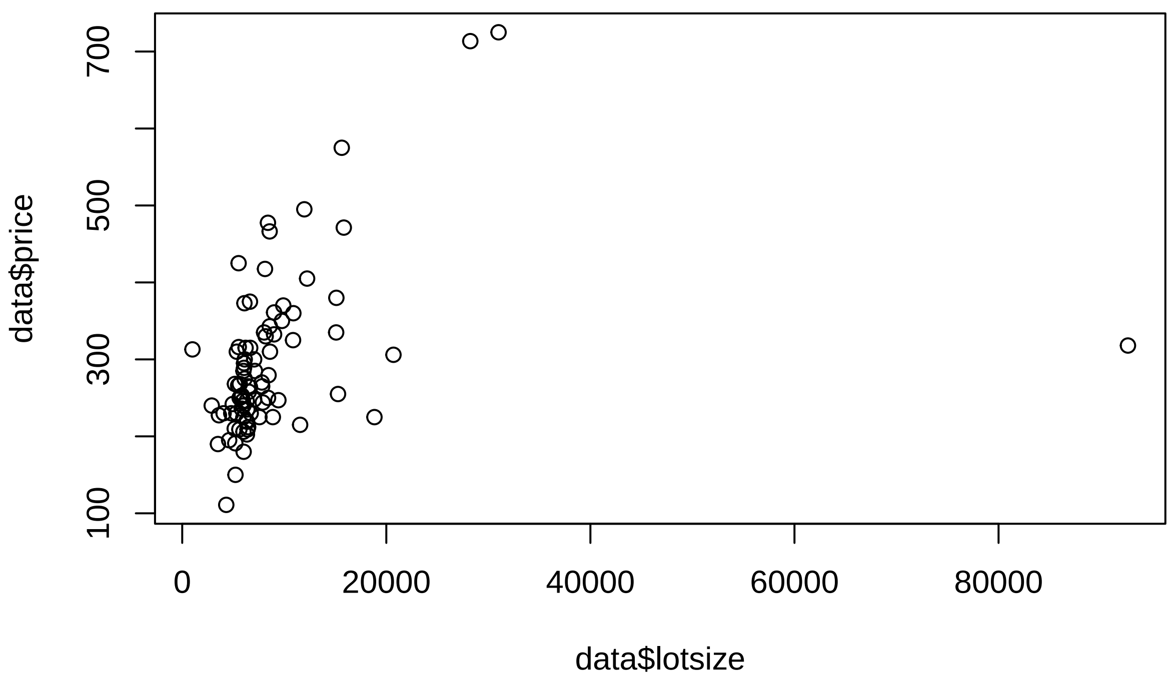
```r
cat("For lotsize model, b0 (intercept) is", b0_lotsize, "and b1 (slope) is", b1_lotsize, "\n")
```

```
## For lotsize model, b0 (intercept) is 261.94 and b1 (slope) is 0
```

```r
summary(data$lotsize)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    5733    6430    9020    8583   92681
```

```r
plot(data$lotsize, data$price)
```



```r
b1_lotsize <- coefficients_lotsize[2]  # Slope for lotsize model without rounding
cat("The unrounded slope for lotsize model is", b1_lotsize, "\n")
```

```
## The unrounded slope for lotsize model is 0.003504406
```

```r
# Check the structure of the data to ensure 'lotsize' is numeric
str(data)
```

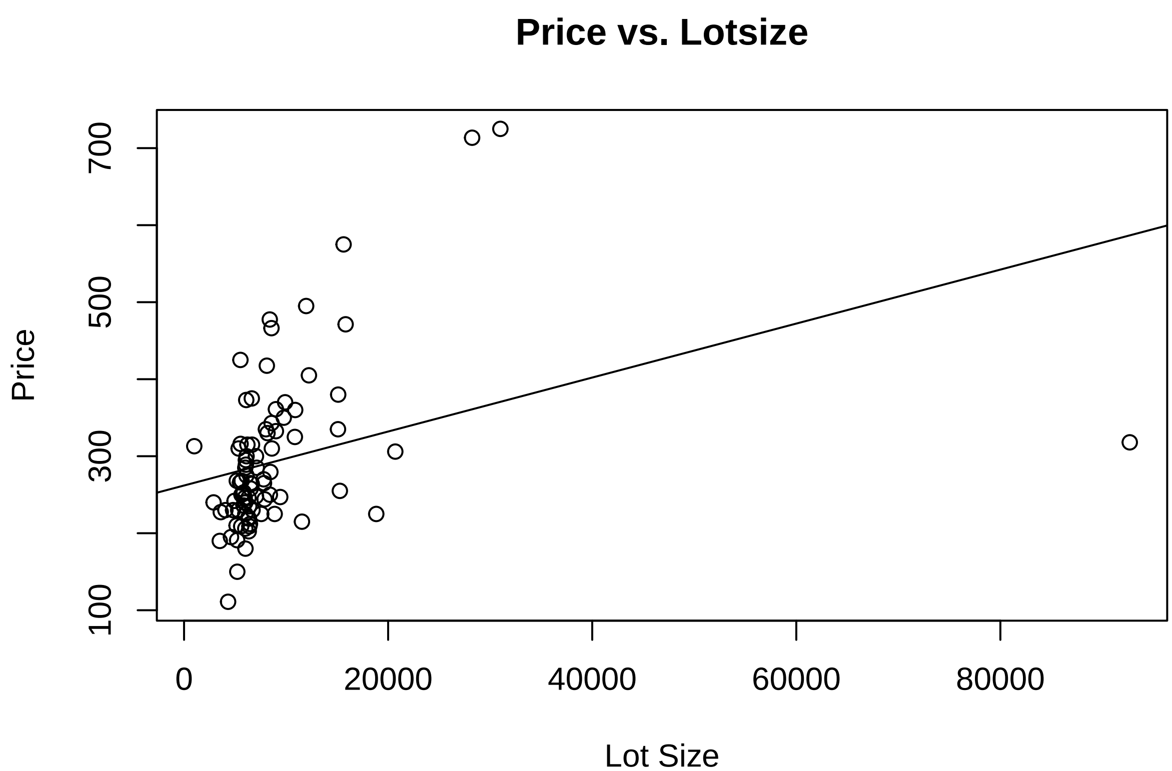
```
## tibble [88 x 6] (S3: tbl_df/tbl/data.frame)
##  $ price   : num [1:88] 300 370 191 195 373 ...
##  $ assess  : num [1:88] 349 352 218 232 319 ...
##  $ bdrms   : num [1:88] 4 3 3 3 4 5 3 3 3 3 ...
##  $ lotsize : num [1:88] 6126 9903 5200 4600 6095 ...
##  $ sqrft   : num [1:88] 2438 2076 1374 1448 2514 ...
##  $ colonial: num [1:88] 1 1 0 1 1 1 1 1 0 0 ...
```

```r
# Calculate coefficients without rounding
linear_model_lotsize <- lm(price ~ lotsize, data = data)
coefficients_lotsize <- coefficients(linear_model_lotsize)
b0_lotsize <- coefficients_lotsize[1]  # Intercept for lotsize model
b1_lotsize <- coefficients_lotsize[2]  # Slope for lotsize model

# Output the coefficients for the lotsize model without rounding
cat("For lotsize model, b0 (intercept) is", b0_lotsize, "and b1 (slope) is", b1_lotsize, "\n")
```

```
## For lotsize model, b0 (intercept) is 261.9368 and b1 (slope) is 0.003504406
```

```r
plot(data$lotsize, data$price, main = "Price vs. Lotsize", xlab = "Lot Size", ylab = "Price")
abline(a = b0_lotsize, b = b1_lotsize)  # Use the precise coefficients here
```



Part (h): Comparison of R-squared values

```r
r_squared_sqrft <- summary(linear_model_sqrft)$r.squared
r_squared_lotsize <- summary(linear_model_lotsize)$r.squared

# Output the R-squared values for sqrft and lotsize models
cat("R-squared for sqrft model is", r_squared_sqrft, "\n")
```

```
## R-squared for sqrft model is 0.6207967
```

```r
cat("R-squared for lotsize model is", r_squared_lotsize, "\n")
```

```
## R-squared for lotsize model is 0.1204954
```

```r
# Determine which model explains more variance
if (r_squared_sqrft > r_squared_lotsize) {
  cat("The variable sqrft explains more of the difference in Prices than the variable lotsize.\n")
} else {
  cat("The variable lotsize explains more of the difference in Prices than the variable sqrft.\n")
}
```

```
## The variable sqrft explains more of the difference in Prices than the variable lotsize.
```