hw1stats

Sherine

2023-10-30

The data in real-estate-valuation-data-set.csv is a subset of the dataset hosted at https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+ data+set that contains information about the unit price of houses in New Taipei City, Taiwan. The subset of the data that we will use contains the following columns: • age: the age of the house in years • distance: the distance to the nearest Mass Rapid Transit (MRT) station from the house (in meters) • convenience_stores: the number of convenience stores near the house • unit_price: the unit price of the house, measured in 10,000 New Taiwan Dollars/Ping (where 1 Ping = 3.3 squared meters)

Q1. Load the data in R and fit a simple linear regression of unit_price onto convenience_stores.

```
realestate = read_excel("D:\\Downloads\\real+estate+valuation+data+set\\Real estate valuation data set.
data_filtered <- select(realestate, `X2 house age`, `X3 distance to the nearest MRT station`,
                        `X4 number of convenience stores`, `Y house price of unit area`)
#Filtered Dataset
data_filtered
```

```
## # A tibble: 414 x 4
##    `X2 house age` X3 distance to the nearest MRT statio~1 X4 number of conveni~2
##             <dbl>                                   <dbl>                  <dbl>
##  1           32                                     84.9                      10
##  2           19.5                                  307.                        9
##  3           13.3                                  562.                        5
##  4           13.3                                  562.                        5
##  5            5                                    391.                        5
##  6            7.1                                 2175.                        3
##  7           34.5                                  623.                        7
##  8           20.3                                  288.                        6
##  9           31.7                                 5512.                        1
## 10           17.9                                 1783.                        3
## # i 404 more rows
## # i abbreviated names: 1: `X3 distance to the nearest MRT station`,
## #   2: `X4 number of convenience stores`
## # i 1 more variable: `Y house price of unit area` <dbl>
```

```
# Fit a simple linear regression model
model <- lm(`Y house price of unit area` ~ `X4 number of convenience stores`, data = data_filtered)
```

Q2. Print the summary of the model in R. In plain English, state the interpretation of the coefficient estimate associated with the predictor convenience_stores.

```r
summary(model)
```

```
##
## Call:
## lm(formula = `Y house price of unit area` ~ `X4 number of convenience stores`,
##     data = data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.407  -7.341  -1.788   5.984  87.681
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      27.1811     0.9419   28.86   <2e-16 ***
## `X4 number of convenience stores`   2.6377     0.1868   14.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 412 degrees of freedom
## Multiple R-squared:  0.326,  Adjusted R-squared:  0.3244
## F-statistic: 199.3 on 1 and 412 DF,  p-value: < 2.2e-16
```

Ans 2. Since the coefficient estimate is positive, it means that as the number of convenience stores near the house increases, the unit price of the house is expected to increase by that coefficient, all else being equal.Coefficient -> 2.637 for convenience stores represents that the change in the unit_price is 2.637 for every one-unit change in the number of convenience stores.

Q3.Does the model indicate a statistically significant association between convenience_stores and unit_price? Explain.

Ans 3. Since the p value is less than 0.05 we can say that the model indicates a statistically significant association between convenience_stores and unit_price at 95% confidence level.

Q4.Create a 99% confidence interval for the coefficient associated with the predictor convenience_stores.

```r
# Create a 99% confidence interval for the coefficient of convenience_stores
conf_interval <- confint(model, level = 0.99)

# Print the confidence interval
conf_interval
```

```
##                                     0.5 %    99.5 %
## (Intercept)                      24.743591 29.618618
## `X4 number of convenience stores`  2.154175  3.121132
```

Q5. In plain English, state the interpretation of the coefficient of determination R2 for this model (this can also be found using the summary function)?

Ans 5. R squared, or the coefficient of determination, is a measure of how well the independent variable(s) (in this case, convenience_stores) explain the variation in the dependent variable (unit_price). The R squared value ranges from 0 to 1, where: If R^2 is 0, it means that the independent variable(s) do not explain any of the variation in the dependent variable. If R^2 is 1, it means that the independent variable(s) perfectly explain all the variation in the dependent variable.

So an R squared value of 0.326 implies that the number of convenience stores explains very little variation in house unit prices i.e 0.326 of the variance in house unit prices is explained by number of convenience stores, which shows that the linear model is weak.

Q6.Create a scatterplot of unit_price vs. convenience_stores that includes the regression line of the model.

```
plot(`Y house price of unit area` ~ `X4 number of convenience stores`, data = data_filtered)

# Fit a simple linear regression.
lm_fit <- lm(`Y house price of unit area` ~ `X4 number of convenience stores`, data = data_filtered)
abline(lm_fit, col = "blue")
```