Real Estate Valuation Analysis in New Taipei City

Sydney Murphy

2023-10-27

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

```
library(readr)

# Load the data from the CSV file
data <-
read_csv("https://raw.githubusercontent.com/sydneymcolumbia/CMU/main/Real%20estate%20valuation%20data%20set.csv")

## Rows: 414 Columns: 8
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (8): No, X1 transaction date, X2 house age, X3 distance to the nearest
M...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

str(data)

## spc_tbl_ [414 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ No                                    : num [1:414] 1 2 3 4 5 6 7 8 9
10 ...
##  $ X1 transaction date                   : num [1:414] 2013 2013 2014 2014
2013 ...
##  $ X2 house age                          : num [1:414] 32 19.5 13.3 13.3 5
7.1 34.5 20.3 31.7 17.9 ...
##  $ X3 distance to the nearest MRT station: num [1:414] 84.9 306.6 562 562
390.6 ...
##  $ X4 number of convenience stores       : num [1:414] 10 9 5 5 5 3 7 6 1
3 ...
##  $ X5 latitude                           : num [1:414] 25 25 25 25 25 ...
##  $ X6 longitude                          : num [1:414] 122 122 122 122 122
...
##  $ Y house price of unit area            : num [1:414] 37.9 42.2 47.3 54.8
43.1 32.1 40.3 46.7 18.8 22.1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   No = col_double(),
```

```
##   ..   `X1 transaction date` = col_double(),
##   ..   `X2 house age` = col_double(),
##   ..   `X3 distance to the nearest MRT station` = col_double(),
##   ..   `X4 number of convenience stores` = col_double(),
##   ..   `X5 latitude` = col_double(),
##   ..   `X6 longitude` = col_double(),
##   ..   `Y house price of unit area` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Rename columns to ensure no spaces or special characters
colnames(data) <- c("X1", "X2", "age", "distance", "convenience_stores",
"latitude", "longitude", "unit_price")

# View the first few rows of the data
head(data)
```

```
## # A tibble: 6 × 8
##      X1    X2   age distance convenience_stores latitude longitude
unit_price
##   <dbl> <dbl> <dbl>    <dbl>              <dbl>    <dbl>     <dbl>
<dbl>
## 1     1 2013.  32       84.9                10     25.0      122.
37.9
## 2     2 2013.  19.5    307.                  9     25.0      122.
42.2
## 3     3 2014.  13.3    562.                  5     25.0      122.
47.3
## 4     4 2014.  13.3    562.                  5     25.0      122.
54.8
## 5     5 2013.   5      391.                  5     25.0      122.
43.1
## 6     6 2013.   7.1   2175.                  3     25.0      122.
32.1
```

```
# Fit a simple linear regression model
model <- lm(unit_price ~ convenience_stores, data = data)

# Print the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = unit_price ~ convenience_stores, data = data)
##
## 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 ***
## 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
```

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

```
# Print the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = unit_price ~ convenience_stores, data = data)
##
## 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 ***
## 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
```

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

```
# Print association description
cat("The more convenience stores near a house, the higher its price tends to
be. Specifically,
    each extra store nearby can raise the house's price by about 2.6377 (in
10,000 New Taiwan Dollars/Ping).
    The data strongly supports this finding. There is a a strong and
statistically significant association
    between the number of convenience stores and the unit price of houses.
House prices are clearly influenced
    by the number of nearby convenience stores. ")
```

```
## The more convenience stores near a house, the higher its price tends to
be. Specifically,
##     each extra store nearby can raise the house's price by about 2.6377
```

```
(in 10,000 New Taiwan Dollars/Ping).
##     The data strongly supports this finding. There is a a strong and
statistically significant association
##     between the number of convenience stores and the unit price of houses.
House prices are clearly influenced
##     by the number of nearby convenience stores.
```

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

```
# Create a 99% confidence interval for the coefficient associated with the
predictor convenience_stores.
# Compute the 99% confidence interval
conf_interval <- confint(model, "convenience_stores", level = 0.99)

print(conf_interval)
```

```
##                       0.5 %   99.5 %
## convenience_stores 2.154175 3.121132
```

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

```
# Print R squared
cat("An R^2 value of 0.326 says that 32.6% of the variation in house prices
is explained
    by the number of nearby convenience stores.")
```

```
## An R^2 value of 0.326 says that 32.6% of the variation in house prices is
explained
##     by the number of nearby convenience stores.
```

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

```
# Load necessary libraries
library(ggplot2)

# Create a scatterplot of unit_price vs. convenience_stores
ggplot(data, aes(x=convenience_stores, y=unit_price)) +
  geom_point(aes(color=convenience_stores), size=2, alpha=0.6) +  # Add
scatter points with color gradient and adjust size and transparency
  geom_smooth(method="lm", se=FALSE, color="black", linewidth=1) +  # Add
thicker regression line in red
  labs(title="Unit Price and Number of Convenience Stores",
       x="Number of Convenience Stores",
       y="Unit Price (in 10,000 New Taiwan Dollars/Ping)",
       caption="Source: Real Estate Valuation Data Set") +
  theme_light() +  # Use a light theme
  theme(legend.position="none",  # Remove legend
        panel.grid.major = element_line(linetype = "dashed", color =
```

```
"grey80"),
        panel.grid.minor = element_blank()) +
  scale_color_gradient(low="blue", high="pink")  # Gradient color for points
based on the number of stores

## `geom_smooth()` using formula = 'y ~ x'
```