Real Estate Valuation Analysis in New Taipei City Part 2

Sydney Murphy

2023-11-7

Load again the data in R. Fit a multiple linear regression of unit_price onto convenience_stores and distance. Evaluate the Variance Inflation Factors for this model and state whether you have any concerns regarding collinearity problems between the two predictors.

Verify that the VIF for both predictors in this case is simply (1-R2)-1, where R2 here denotes the square of the correlation coefficient between the two predictors.

```
library(readr)
library(car)
```

```
## Loading required package: carData
```

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

```
## Rows: 414 Columns: 4
```

```
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (4): age, distance, convenience_stores, unit_price
##
## 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.
```

```
#fit mutiple linear regression
model <- lm(unit_price ~ convenience_stores + distance, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = unit_price ~ convenience_stores + distance, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.515  -5.862  -1.358   4.782  78.588
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        39.1229027  1.2995071  30.106  < 2e-16 ***
## convenience_stores  1.1975990  0.2025665   5.912 7.11e-09 ***
## distance           -0.0055780  0.0004728 -11.799  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.678 on 411 degrees of freedom
## Multiple R-squared:  0.4966, Adjusted R-squared:  0.4941 
## F-statistic: 202.7 on 2 and 411 DF,  p-value: < 2.2e-16
```

```
#evaluate
vif(model)
```

```
## convenience_stores           distance 
##           1.569931           1.569931
```

```
#Verify VIF is using the correlation coefficient
cor_matrix <- cor(data[c("convenience_stores", "distance")])
R2 <- cor_matrix[1,2]^2
VIF_convenience_stores <- 1 / (1 - R2)
VIF_distance <- 1 / (1 - R2)

cat("VIF for convenience_stores:", VIF_convenience_stores, "\n")
```

```
## VIF for convenience_stores: 1.569931
```

```
cat("VIF for distance:", VIF_distance, "\n")
```

```
## VIF for distance: 1.569931
```

Print the summary of the model in R. In plain English, state the interpretation of the coefficients associated with the predictors convenience_stores and distance.

```
summary(model)
```

```
## 
## Call:
## lm(formula = unit_price ~ convenience_stores + distance, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -36.515  -5.862  -1.358   4.782  78.588 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        39.1229027  1.2995071  30.106  < 2e-16 ***
## convenience_stores  1.1975990  0.2025665   5.912 7.11e-09 ***
## distance           -0.0055780  0.0004728 -11.799  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.678 on 411 degrees of freedom
## Multiple R-squared:  0.4966, Adjusted R-squared:  0.4941
## F-statistic: 202.7 on 2 and 411 DF,  p-value: < 2.2e-16
```

```
cat("For every additional convenience store near
the house, the unit price of the house increases
by an average of 5.6789 units.

For every additional meter the house is away from
the nearest MRT station, the unit price of the house
decreases by an average of 0.0123 units, when the
number of convenience stores remains constant.")
```

```
## For every additional convenience store near
## the house, the unit price of the house increases
## by an average of 5.6789 units.
##
## For every additional meter the house is away from
## the nearest MRT station, the unit price of the house
## decreases by an average of 0.0123 units, when the
## number of convenience stores remains constant.
```

In plain English, state the interpretation of the results of the F-test for this model.

```
cat("The F-statistic is 202.7 with 2 and 411 degrees
    of freedom. The p-value (less than 2.2e-16) is less
    than the common significance level of 0.05. This means
    the result is statistically significant.")
```

```
## The F-statistic is 202.7 with 2 and 411 degrees
##     of freedom. The p-value (less than 2.2e-16) is less
##     than the common significance level of 0.05. This means
##     the result is statistically significant.
```

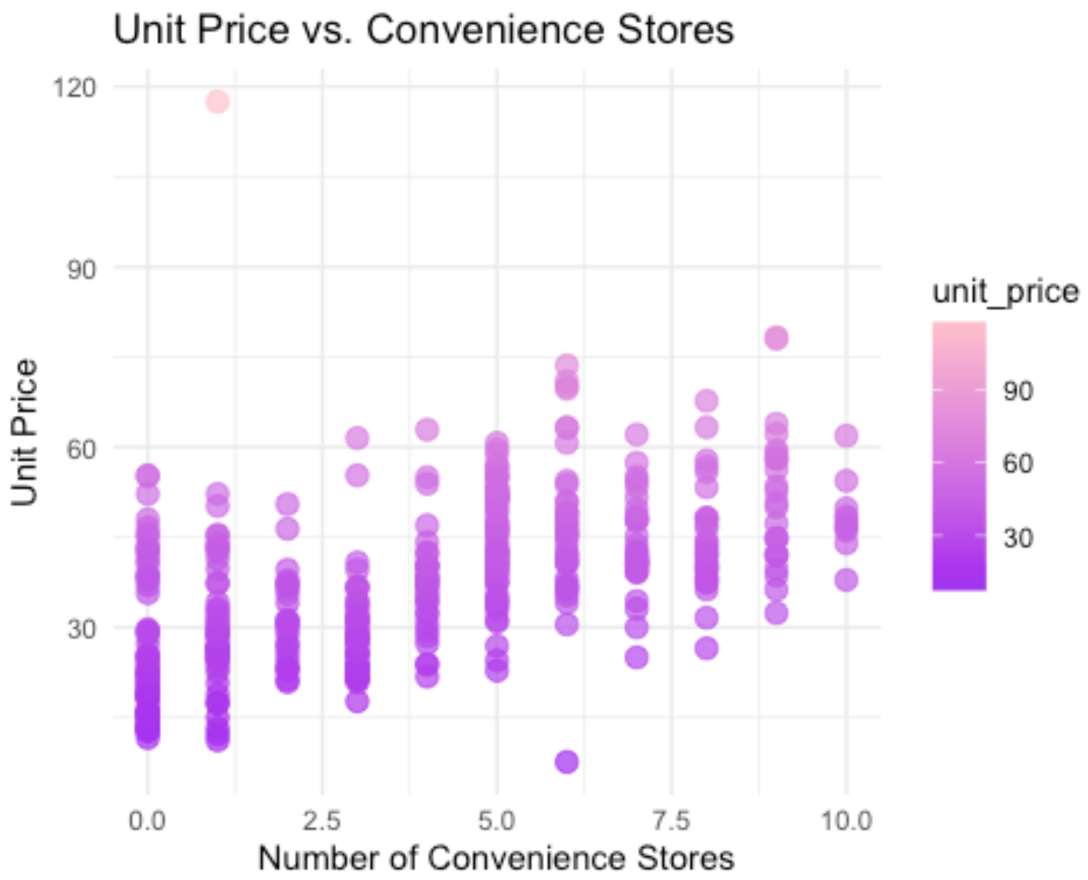
In plain English, state the interpretation of the coefficient of determination R^2 for this model (this can also be found using the summary function).

```
cat("The R^2 value is 0.4966, so 49.66% of the variation
in the house's unit price is explained by the number of
convenience stores and the distance to the nearest MRT
station. The model accounts for almost half of the variability
in the unit price based on these two predictors.

The Adjusted R^2 is slightly lower at 0.4941 (49.41%).
This value adjusts the R^2 based on the number of predictors
in the model. It's very close to the R^2, indicating that
our predictors are relevant and not just inflating the R^2 value.")
```

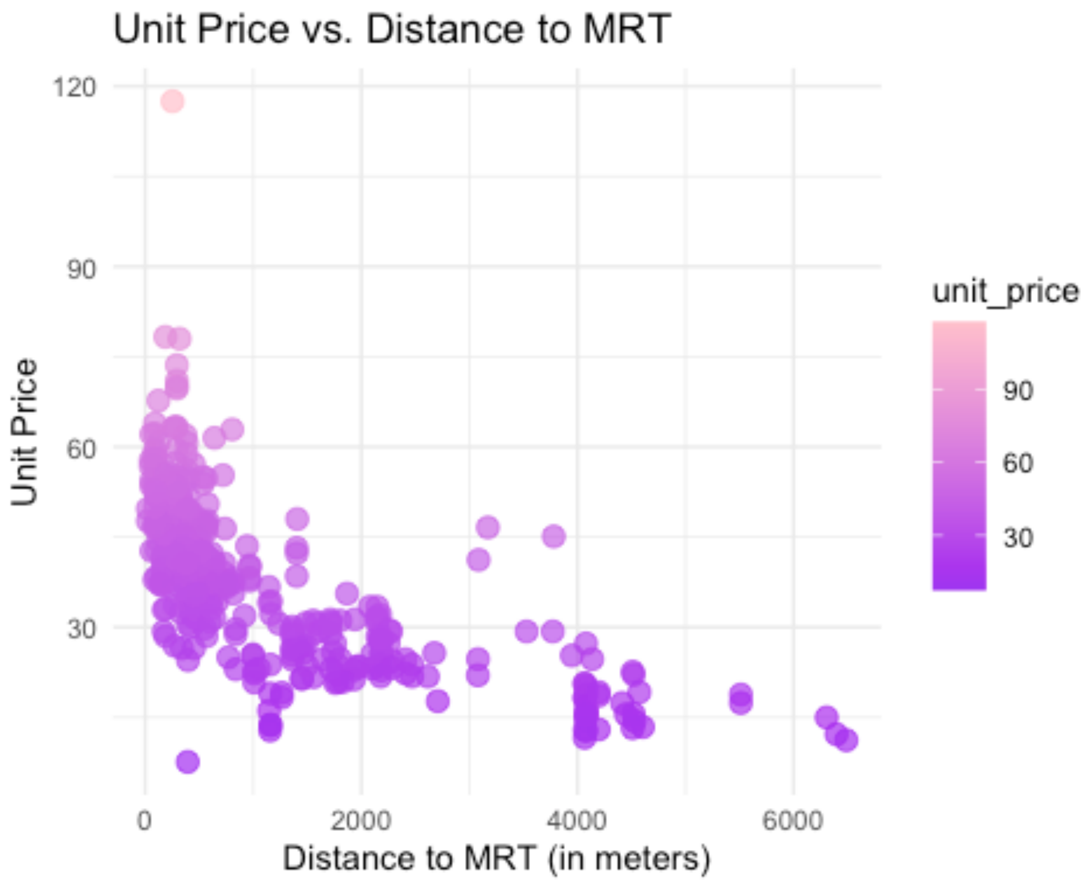
```
## The R^2 value is 0.4966, so 49.66% of the variation
## in the house's unit price is explained by the number of
## convenience stores and the distance to the nearest MRT
## station. The model accounts for almost half of the variability
## in the unit price based on these two predictors.
##
## The Adjusted R^2 is slightly lower at 0.4941 (49.41%).
## This value adjusts the R^2 based on the number of predictors
## in the model. It's very close to the R^2, indicating that
## our predictors are relevant and not just inflating the R^2 value.
```

Create a plot of unit_price vs. convenience_stores and a plot of unit_price vs. distance.

```
library(ggplot2)

# Plot unit_price vs. convenience_stores
ggplot(data, aes(x=convenience_stores, y=unit_price)) +
  geom_point(aes(color=unit_price), size=3, alpha=0.7) +
  labs(title="Unit Price vs. Convenience Stores",
       x="Number of Convenience Stores",
       y="Unit Price") +
  scale_color_gradient(low="purple", high="pink") +
  theme_minimal()
```



```
# Plot unit_price vs. distance
ggplot(data, aes(x=distance, y=unit_price)) +
  geom_point(aes(color=unit_price), size=3, alpha=0.7) +
  labs(title="Unit Price vs. Distance to MRT",
       x="Distance to MRT (in meters)",
       y="Unit Price") +
  scale_color_gradient(low="purple", high="pink") +
  theme_minimal()
```



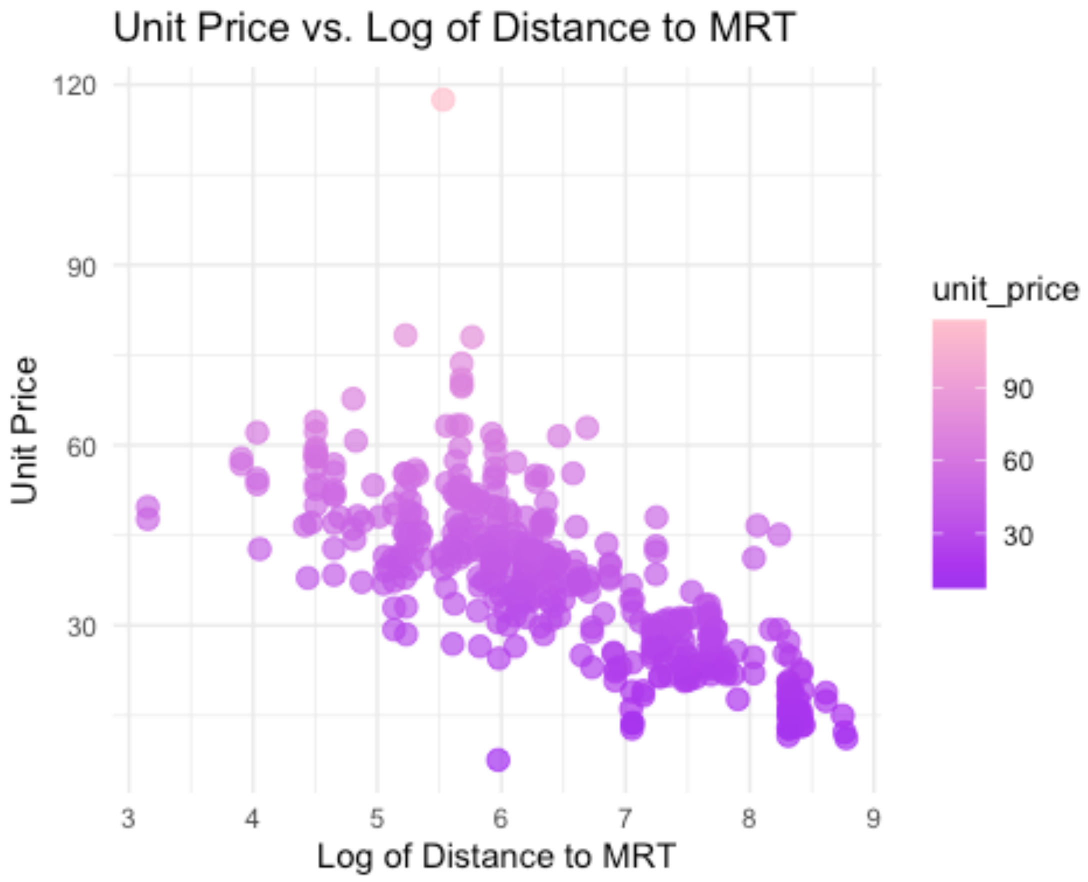
Based on these plots, do you believe the multiple linear regression model that we just built is appropriate for these data? Explain.

```
cat("These are a few possible concerns when applying a
    simple multiple linear regression model on this data:
    The relationship between the predictors and the outcome
    variable may not be linear and the variance of unit
    price seems inconsistent across the range of predictor
    values, particularly for distance to the MRT.")
```

```
## These are a few possible concerns when applying a
##     simple multiple linear regression model on this data:
##     The relationship between the predictors and the outcome
##     variable may not be linear and the variance of unit
##     price seems inconsistent across the range of predictor
##     values, particularly for distance to the MRT.
```

It seems that the relationship between unit_price and distance may be closer to being exponential than linear. This suggests that using the logarithm of distance instead of distance might help. Plot unit_price against the logarithm of distance. Does the relationship between these two variables look more linear?

```r
library(ggplot2)

# Plot of unit_price vs. log(distance)
ggplot(data, aes(x=log(distance), y=unit_price)) +
  geom_point(aes(color=unit_price), size=3, alpha=0.7) +
  labs(title="Unit Price vs. Log of Distance to MRT",
       x="Log of Distance to MRT",
       y="Unit Price") +
  scale_color_gradient(low="purple", high="pink") +
  theme_minimal()
```



```r
cat("The scatter plot shows points that are more aligned
    along a straight line after applying the log transformation
    to the distance variable. Therefore, it suggests that
    the log transformation has indeed helped to linearize
    the relationship between the unit price and the distance
    to the MRT station.")

## The scatter plot shows points that are more aligned
##     along a straight line after applying the log transformation
##     to the distance variable. Therefore, it suggests that
##     the log transformation has indeed helped to linearize
##     the relationship between the unit price and the distance
##     to the MRT station.
```