# hw2stats

Sherine

2023-11-06

Question 1 - 5 points 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.

```
# Load the data
realestate = read_excel("D:\\Downloads\\real+estate+valuation+data+set\\Real estate valuation data set.

# Fit a multiple linear regression model
model <- lm(`Y house price of unit area` ~
              `X4 number of convenience stores` + `X3 distance to the nearest MRT station`, realestate)

# Evaluate VIF

vif(model)
```

```
##        ‘X4 number of convenience stores‘
##                                 1.569931
## ‘X3 distance to the nearest MRT station‘
##                                 1.569931
```

Since the VIF values are typically small (Below 10) we can say that there are no concerns regarding collinearity of the two variables.

Question 2 - 20 points Print the summary of the model in R. In plain English, state the interpreta- tion of the coefficients associated with the predictors convenience_stores and distance.

```
summary(model)
```

```
##
## Call:
## lm(formula = ‘Y house price of unit area‘ ~ ‘X4 number of convenience stores‘ +
##     ‘X3 distance to the nearest MRT station‘, data = realestate)
##
## 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
## ‘X4 number of convenience stores‘         1.1975990  0.2025665   5.912 7.11e-09
## ‘X3 distance to the nearest MRT station‘ -0.0055780  0.0004728 -11.799  < 2e-16
##
## (Intercept)                                ***
## ‘X4 number of convenience stores‘          ***
## ‘X3 distance to the nearest MRT station‘ ***
## ---
## 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
```

The coefficient estimate associated with convenience_stores - `1.1975*1000=1197.59` represents the increase in the unit price by 1197.59 for a one-unit increase in the number of convenience stores nearby, holding all other variables constant. The coefficient estimate associated with distance - `-0.005578*1000=-5.578` represents the decrease in the unit price by 5.578 for a one-unit increase in the distance to the nearest MRT station, holding all other variables constant. Std. Error (0.2025665): The standard error of the coefficient for convenience stores is 0.2025665, indicating relatively low variability in the estimate. t value (5.912): The t value of 5.912 is also high, indicating that the coefficient for convenience stores is statistically significant. Each additional convenience store has a significant impact on the unit price. Std. Error (0.0004728): The standard error of the coefficient for distance is very low, suggesting high precision in the estimate. t value (-11.799): The t value of -11.799 is exceptionally high in absolute value, indicating that the coefficient for distance is highly statistically significant. Each additional meter of distance to the MRT station has a significant impact on the unit price.

Question 3 - 20 points In plain English, state the interpretation of the results of the F-test for this model.

```r
# Create a null model (intercept-only)
null_model <- lm(`Y house price of unit area` ~ 1, realestate)

# Perform the F-test to compare the full model to the null model

f_test_result <- anova(model,null_model , test = "F")

p_value <- f_test_result[2, "Pr(>F)"]

f_test_result[2,6]
```

```
## [1] 5.607584e-62
```

```r
p_value
```

```
## [1] 5.607584e-62
```

The F-test assesses whether the overall model (including both predictors) is a statistically significant improvement over a null model (no predictors). In plain English, if the F-test p-value is small (typically less than 0.05), it suggests that the model as a whole is a good fit for the data, meaning that at least one of the predictors significantly contributes to explaining the variance in the unit price.

The F-statistic in Model 1 is 202.7, and its p-value is very close to zero. This suggests that the full model, which includes both X4 number of convenience stores and X3 distance to the nearest MRT station, is a statistically significant improvement over the null model (Model 2). Therefore, at least one of the predictors significantly contributes to explaining the variance in Y house price of unit area.

Question 4 - 20 points In plain English, state the interpretation of the coefficient of determination R2 for this model (this can also be found using the summary function).

The coefficient of determination (R-squared) represents the proportion of the variance in the dependent variable (unit_price) that can be explained by the independent variables (convenience_stores and distance). In plain English, it tells you how well the model fits the data. A higher R-squared value (closer to 1) indicates that a larger proportion of the variance is explained by the model.

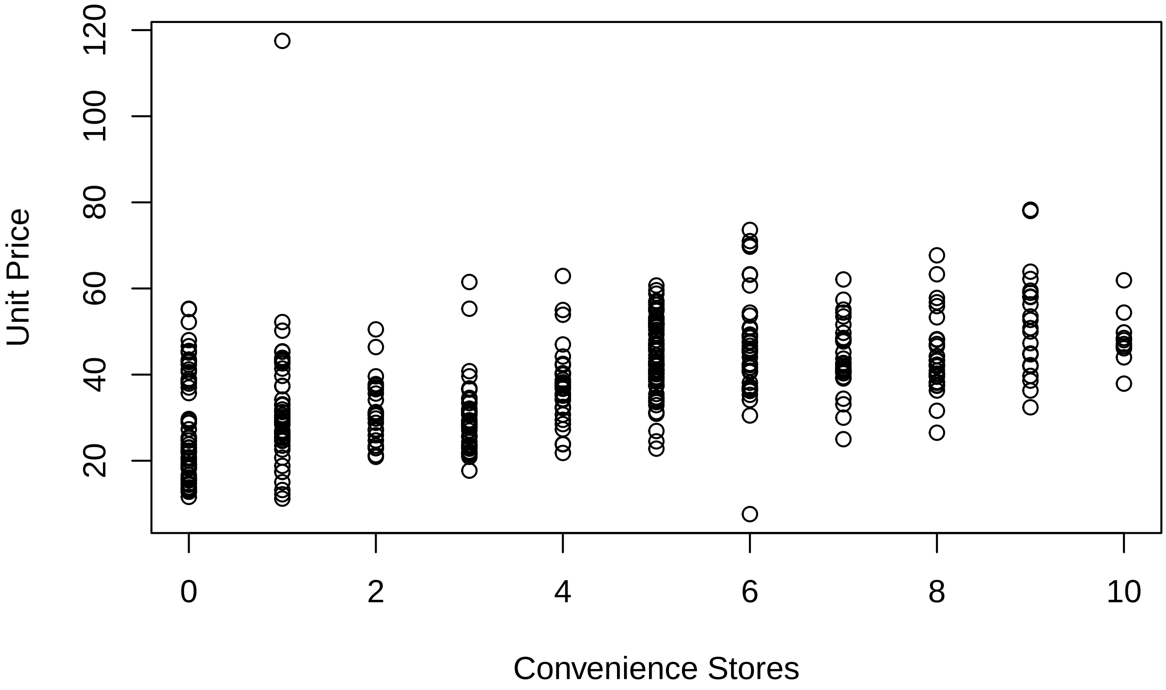
In this specific model, the R-squared value is approximately 0.4966, which means that about 49.66% of the total variance in house prices can be explained by the combination of the number of convenience stores and the distance to the nearest MRT station. The remaining 50.34% of the variance is unexplained and may be attributed to other factors not included in the model or random variation.

The Adjusted R-squared (Adjusted $R^2$) is also provided, and it is 0.4941 in this case. Adjusted $R^2$ takes into account the number of predictors in the model and can be a more conservative measure of goodness-of-fit. It penalizes models that include unnecessary predictors. In this case, the Adjusted R-squared is slightly lower than the R-squared, indicating that there might be some level of overfitting or the inclusion of less relevant predictors.

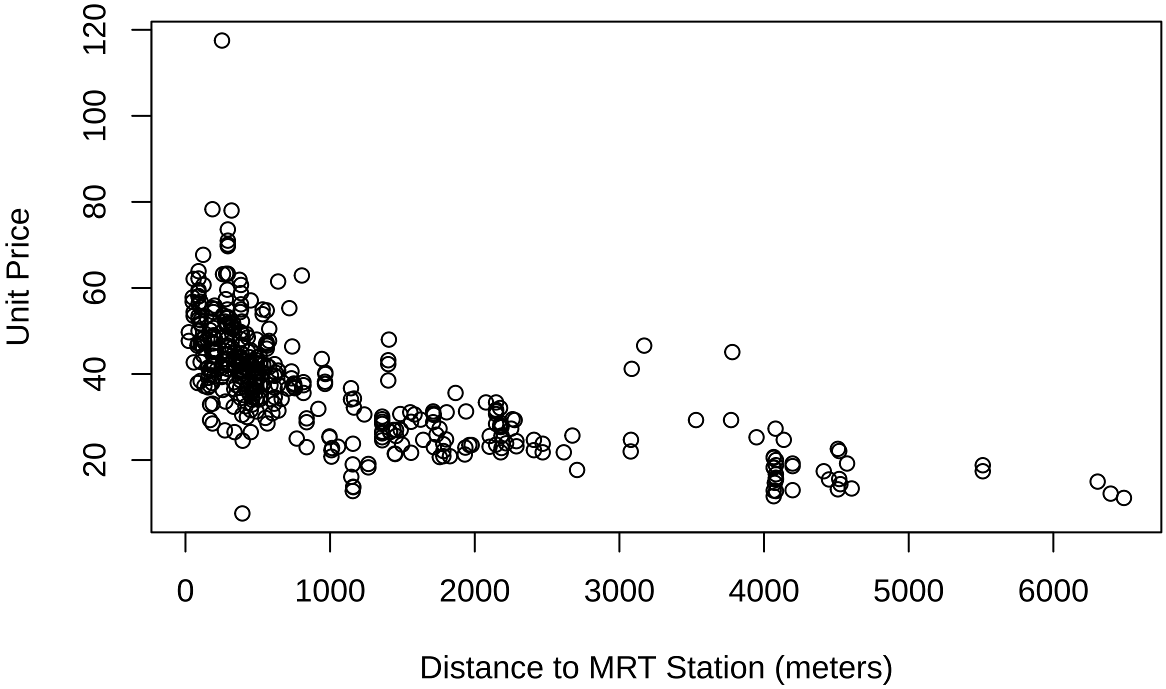
Overall, an R-squared of approximately 0.4966 suggests that the model, with the given predictors, explains a moderate amount of the variability in house prices. The remaining variability not explained by the model may be influenced by other factors not considered in this analysis.

Question 5: To create plots of unit_price vs. convenience_stores and unit_price vs. distance:

```
# Plot unit_price vs. convenience_stores
plot(realestate$`X4 number of convenience stores`,
     realestate$`Y house price of unit area`, xlab = "Convenience Stores",
     ylab = "Unit Price")
```



```r
# Plot unit_price vs. distance
plot(realestate$`X3 distance to the nearest MRT station`,
     realestate$`Y house price of unit area`, xlab = "Distance to MRT Station (meters)", ylab = "Unit P
```



Question 6 - 20 points Based on these plots, do you believe the multiple linear regression model that we just built is appropriate for these data? Explain.

Number of Convenient Stores vs. Unit Price (Seems Linear):

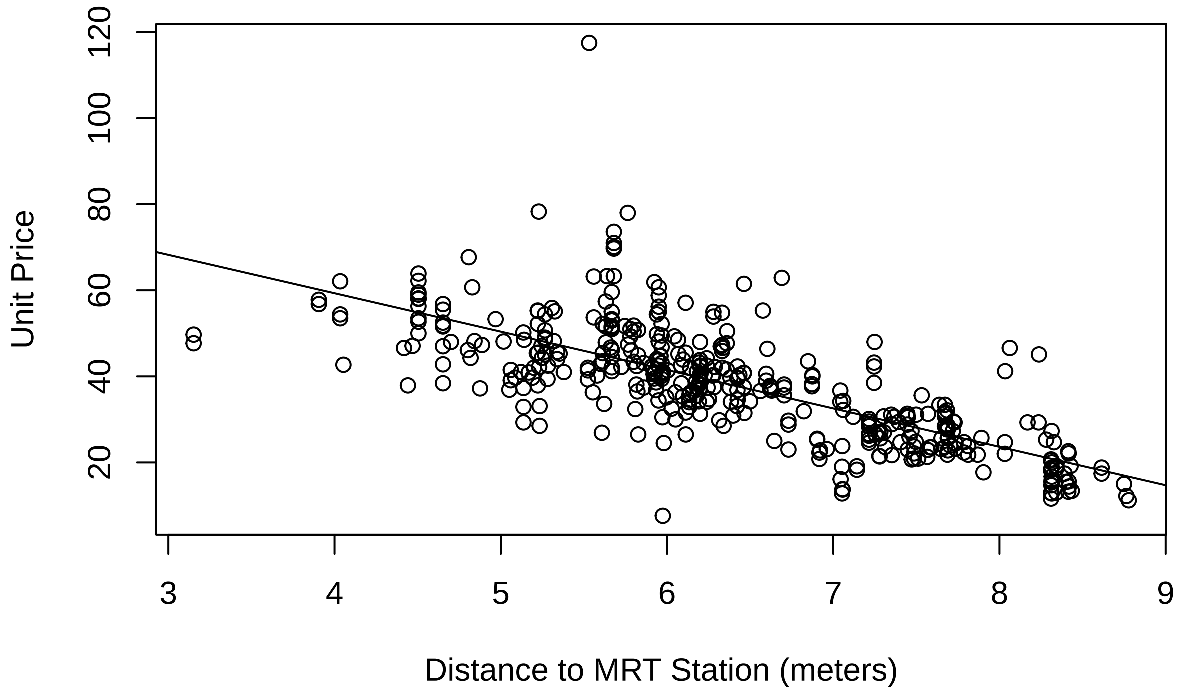
When the plot of unit price against the number of convenience stores appears somewhat linear, it suggests that there is a relatively consistent and predictable relationship between the number of convenience stores near a house and its unit price. An increase in the number of convenience stores is associated with an increase in unit price, and this relationship is somewhat consistent across the data points. The linear relationship indicates that the impact of adding one more convenience store on unit price is somewhat constant throughout the range of data, making it a good candidate for linear regression modeling. Distance to the Nearest MRT Station vs. Unit Price (Scattered):

When the plot of unit price against the distance to the nearest MRT station appears scattered or less linear, it suggests that there is more variability in the relationship. The unit price of houses does not seem to change in a consistent linear fashion as distance increases. There may be more complex or non-linear patterns at play. This scattered relationship indicates that a simple linear model may not fully capture the variation in unit price as distance changes. It may be worth exploring other modeling approaches, such as using transformations of the distance variable (e.g., taking the logarithm of distance) or considering additional predictors to better explain this relationship.

Question 7 - 10 points 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 log- arithm of distance. Does the relationship between these two variables look more linear?

```r
# Plot unit_price vs. log of distance
plot(log(realestate$`X3 distance to the nearest MRT station`),
     realestate$`Y house price of unit area`,
     xlab = "Distance to MRT Station (meters)", ylab = "Unit Price")
model1 <- lm(`Y house price of unit area` ~
               log(`X3 distance to the nearest MRT station`), realestate)
abline(model1,cil="blue")
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "cil" is
## not a graphical parameter
```



The relationship of distance with unit price looks inversely linear. Inversely Linear Relationship: An inversely linear relationship, also known as a negative exponential relationship, means that as the logarithm of the distance to the nearest MRT station increases, the unit price of houses decreases, and vice versa. This pattern indicates that the impact of increasing the distance on unit price is not linear; instead, it becomes more significant as the distance increases. In practical terms, it suggests that houses located further away from the MRT station may experience a more rapid decrease in unit price compared to houses located closer to the station.

Logarithmic Transformation: The use of the logarithm of distance is a common transformation in regression modeling when dealing with variables that have non-linear relationships. In this case, it helps capture the diminishing effect of distance on unit price. The logarithmic transformation can make the relationship more linear, and it's often used when there is a decreasing trend that becomes more gradual as the predictor variable (distance) increases.