Statistical Foundations of Business Analytics
46-883

Carnegie Mellon University
Tepper School of Business

Homework 1

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).

Question 1 - 5 points

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

Question 2 - 15 points

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

Question 3 - 5 points

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

Question 4 - 5 points

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

Question 5 - 5 points

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.
[optional] Verify that the VIF for both predictors in this case is simply $(1-R^2)^{-1}$*, where* R^2 *here denotes the square of the correlation coefficient between the two predictors.*

Question 6 - 15 points

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`.

Question 7 - 15 points

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

Question 8 - 15 points

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).

Question 9 - 5 points

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

Question 10 - 15 points

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