**Rounok Joardar**

**FTEC 6334 HW2**

**Question 1.** Sheet Q1 contains data on house characteristics. Assume house prices follow the following equation. This equation shows the **actual**, not estimated, value of house. No one knows this equation, we will try to find something close to this using data we have.

Remodeled is a flag which is 1 if the house is remodeled, and 0 otherwise. Error has normal distribution with mean 5000 and standard deviation of 2000.

Moreover, price per SQFT of neighborhood 1 is $15 higher than price calculated above, and price per SQFT of neighborhood 2 is $5 less than price calculated above.

1. Simulate house prices using the above information (note: after one simulation of errors, copy the values, and use the copied ones; otherwise they would change each time you make any change)



**Note**: The error term e, is obtained by using =NORM.INV(RAND(),5000,2000)

1. Now we want to find a model for price of house, so we can use it to estimate price of other houses. Again, remember after this we have no idea about equation above. We just have some data and want to find a model for price of house.

Assume that no data is available on whether a house is remodeled. So we can not use this variable. Using the other variables, estimate a model for house prices. Write the equation for the linear model you are going to estimate, and find coefficients using Excel solver (don’t forget the intercept!)



 

**Equation of Model**: ModelPrice = b0 + (b1\* AreaN) + (b2\* Hood1) + (b3\* Hood2) + (b4\* AgeN)

**Note**: AreaN is the Area normalized by 1000, AgeN is normalized by 10, and PriceN by 500000. This is done to reduce the range of values and aid the Excel Solver in reaching the global minima.

**Note**: errorSq is calculated using normalized values, and the final Model column is denormalized. SSE is the Sum of Squared Errors, the sum of the errorSq column. Beta values obtained by minimizing SSE using Excel Solver.

1. Calculate average and standard deviation of errors of model in part B.



**Note**: Price from Part A, Model Price from Part B, Error is the difference between the two prices. AVERAGE() and STDEV() Excel functions used to calculate the AvgError and StdDevError.

1. Re-estimate model of part B, this time without intercept. Calculate average of errors, and compare it with average errors in part C. What does this comparison tell you?



**Note**: New data calculated using same model as previously, but with b0, the intercept, forced to 0.

**Note**: In this no intercept model, the average error is no longer 0. The intercept informs us the mean of the response (in this case, the Price) when all predictors, (Area, Neighborhood, and Age) are 0. Without the intercept, the model tries to fit itself through the origin point. But, as we can see from the equation for the Actual Price, when all predictors are set to 0, the Price is equal to the Error term, not to 0. This is the cause of the nonzero average error.

It is essential to include the intercept term when doing linear regression. The most important reason is that this term allows the average of the residual errors to be zero. Non-zero average of the residuals implies that a basic assumption of linear regression has been violated and the results may not be meaningful. Also, in this case we already know that the error term used to compute the actual prices had a non-zero mean (5000). The estimated intercept reflects this. Lastly, in mathematical terms, the intercept can be interpreted to be the mean value of the dependent variable when all predictor variables are zero. However, this interpretation is meaningless unless observations of the dependent variable are available around the region where all predictor variables are zero.

1. Estimate model of part B using the Golden Rule of Beta Hat (refer to lecture 2, slide 15), and calculate variance of errors.

From lecture 2, slide 15:

A picture containing bird, knife

Description automatically generated



**Note**: [Y] vector is the Y column, and the 5 X columns create the [X] matrix. These are used as inputs for python script HouseModel\_MatrixCalculations.py which performs the matrix calculations.

﻿

**Python Outputs:**

B\_hat:



Y\_hat:



**Note**: Y\_hat outputs match the model prices from the Excel model exactly.

﻿

Variance of Errors:

660509706.1192286

1. Estimate matrix of variance-covariance matrix for coefficients, using the following formula:

is variance of error terms from part E.

Variance-Covariance Matrix:



**Note**: All Values obtained from python script HouseModel\_MatrixCalculations.py

1. Diagonal of variance-covariance matrix, shows variance of each coefficient. Estimate t-statistic for each coefficient using the formula:

t-stats:



**Note**: All Values obtained from python script HouseModel\_MatrixCalculations.py

1. Identify which coefficients are statistically significant at 5% level of significance. Using only significant coefficients, estimate house price for a 3500 SQFT house in neighborhood 3, with 10 years age.

Given 20 samples and 4 factors we have a nu value of 16. Using an online calculator we can see that the 95% T-distribution boundary is at +/- 2.11991.

A screenshot of a cell phone

Description automatically generated

﻿If t-stat is within the 95% t-distribution then we reject that factor, so B0 and B3 can be rejected. From this a new model can be constructed.

**Equation of New Model**: NewModelPrice = (b1\* AreaN) + (b2\* Hood1) + (b4\* AgeN)

And from this equation, we can estimate the model of the 3500 house to be 534764.16, as shown below.

