**MGMT 2050 - Assignment 1B:**

Refer to the posted data file **Assignment 1B- Data file.xls**; it’s the data you have been collecting with your friends in class.

The file contains the following columns:

1. Column A: How much the person spent on their meal today ($)
2. Column B: Age
3. Column C: How many people the person dined with (=0 if the person dined alone)
4. Column D: Gender: Male=1 , Female=2
5. Column E: Location : York lanes, Tel, Seneca

You have been analyzing a similar file in class (specific to your location) using columns A to D.   
Now, we have added column E to represent the locations and consolidated the data from all files to one.

Your job is to translate columns D and E to dummy variables, run a multiple regression, and analyze the results.

Hint: You will need to adjust the both the Gender column (1 = Male,2= Female) and the Location column (York lanes, Tel, Seneca ) to dummy (0,1) variables. For the Location, create 2 columns for the 2 dummy variables (Location 1 and Location 2) and use the following 0,1 assignments:

|  |  |  |
| --- | --- | --- |
| **Original Location value** | **Location 1** | **Location 2** |
| York | 1 | 0 |
| TEL | 0 | 1 |
| Seneca | 0 | 0 |

1. Interpret the slopes

To begin, we must define all the variables that affect the slope:

B0 = Y-Intercept

B1 = Age

B2 = Number of people dining

B3 = Gender; This variable is a dummy variable where 0 = male and 1 = female

B4 & B5 = Variables that refer to the location of where people were surveyed; This is a dummy variable as well (B4,B5 = 0 if Seneca; B4 = 0, B5 = 1 if TEL; B4 = 1, B5 = 0 if York Lanes)

**Coefficient Function**

ŷ = 7.349022573 – 0.0365155B1 – 0.037830709B2 – 0.493212308B3 + 3.727923151B4 + 1.12452705B5

*B0 = 7.349022573 = Y-Intercept*

This is the y-intercept. Simply stated, the y intercept is $7.35. This is meaningless though, however, because age cannot be 0, and therefore the age variable will always be greater than 0

*B1 = -0.0365155 = Age*

This is the variable indicating age. An increase in the variable of age would have a resulting decrease in the ŷ by $0.036. This is assuming all other variables stay constant.

*B2 = -0.037830709 = number of people dining*

This is the variable indicating the number of people dining. An increase in the variable of number of people dining would have a resulting decrease in the ŷ by $0.038. This is assuming all other variables stay constant.

*B3 = -0.493212308 = Gender*

This is the variable indicating gender. This is a binary variable. The base case is male. Therefore, if the gender became female, there would have a resulting decrease in the ŷ by $0.493. This is assuming all other variables stay constant.

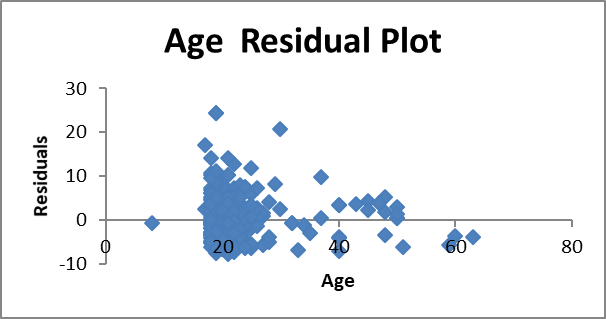
*B4 = 3.727923151 = Location 1 = York Lanes*

This is the variable indicating the location. This is a binary variable where the base location is Seneca. If the location happened to be York Lanes, there would be a resulting increase in the ŷ by $3.728. This is assuming all other variables stay constant.

*B5 = 1.124252705 = Location 2 = TEL Building*

This is the variable indicating the location. This is a binary variable where the base location is also Seneca. If the location happened to be the TEL building, there would be a resulting increase in the ŷ by $1.124. This is assuming all other variables stay constant.

1. Assess the model quality using all the types of measures we discussed for overall model quality as well as individual variables
2. Check assumptions

Assumption 1: Linearity  
Upon examination of the two graphs, it is visible that linearity is not present within the two graphs, the age residual plot and the number of people dined graph. There is no pattern present in the two graphs, and therefore, the linearity assumption is not violated when examining the non-dummy variables; the linearity assumption holds for both variables

Assumption 2: Normality  
  
After analyzing the normal probability plot, one can determine that the error is normally distributed for the majority of the graph. The graph resembles a straight line, a positive linear relation, with a slight increase at the very end. These indications suggest that the model is normal.

Assumption 3: Homoscedasticity  
  
When analyzing the chart, we must determine whether there is a significant dispersion between the data points. Upon seeing the dispersion of the data points, we can see a slight dispersion between the data points; between around 6-8 residuals and 10 residuals. however, despite the slight spread, the data points are all distributed in the same region, and therefore we can assume that homoscedasticity is not violated.

Assumption 4: Outliers  
  
There are some slight outliers when examining the data points. There are three data points grouped around the age of 60 and one around the age of 10. These outliers, however, do not play a huge role on the overall pattern of the model because the model has a very large sample size. The outliers will not have a large impact on the model.

Assumption 5: Independence of errors  
This final assumption does not play a role in this particular example as there isn’t cross sectional data in this model that involves cyclical and time factors.

1. Comment on whether you are satisfied with this model for prediction or not. If not, state what model you would try next (which independent variables). DO NOT run the additional model.

This model is a good model as all the assumptions fit the model and have not been violated. However, despite not violating the assumptions, we are not satisfied as the standard error of 4.70 compared to the mean of 7.72 is not a good sign of a strong model. To make this model stronger, I would look to try removing variables to reduce the standard error relative to the mean. By testing the lack of different variables, we would be able to make the model more or less accurate, and we will be able to determine which mix of variables will be the most accurate. I would try losing age as, if looking at the age residual plot, age does not play too much of a significant role on the amount of money spent.

Use your model (with all the variables – whether you are satisfied with it or not) to provide a prediction of the amount a person with the following characteristics will spend on a meal:

The person’s age is 19, it’s a male, he will dine alone and he is planning to eat at the TEL building cafeteria.

[NOTE: to keep the assignment manageable, we are not asking you to run additional models until you find one that you are happy with. Normally you would go through the iterative model building process to find a model you are satisfied with and only then would you make predictions using it.]

\*MAC users – please note that some users gets the wrong residual plots, we are not sure but it might be related to the version of the Excel or add-in in your MAC. Therefore it is recommended to use PC for the assignment, you can use the lab’s PCs for creating the regression files.