MMA 860 Final Practice Questions

1. Many university and college students work during their summer vacations. One particular school was very interested in understanding student employment prospects. To develop insights, a confidential survey was conducted for second-year students and that data was linked to records on their academic performance and other demographic information for those who worked in the summer.

You have been hired by the school’s principal to analyze the data. In general he would like you to analyze student earnings in terms of their grade performance and demographic characteristics. He has provided you with a list of questions below that he would like to have answered by your analysis. For each question, he would like you to **do whatever analysis is required to answer the questions and provide an explanation in language he is likely to understand.** He encourages you to read all of the parts of the question first as he believes this questions you develop the model(s) necessary to answer them.

* 1. Develop a model of student earnings based on the data on tab ‘Students’. Explain why you chose the model you did and what the results tell you.
  2. Is heteroskedasticity an issue with this model? Regardless of your findings, assume that heteroskedasticity is not an issue for the balance of your analysis.
  3. The principal claims that summer earnings increase as grades in calculus increase. If so, he would like to offer free tutoring support to students to increase their calculus grades so that they will earn more. Does the evidence support the existence of such a relationship? If such a relationship existed, would it justify his strategy?
  4. Is there evidence of any difference in performance between male and female students in terms of the relationship between at least some of their grades and their summer earnings?

Q1a

We should start with an examination of the data to look for outliers, missing values and other ‘obvious’ problems. This can be done with summary statistics in R / Excel using descriptive statistics and correlation analysis. To do so, I needed to recode Male / Female as a dummy variable for Male. I did this with an if-else statement.

Then we can start modelling – which will allow us to test for other problems such as multi-dimensional outlier / influential observations, the degree of problem associated with collinearity.

From a modelling perspective we need a single model that will address a number of things so we should think from the beginning about what might be required:

* Earnings increase with calculus grades -> this requirement implies that we must have calculus in the model regardless of whether or not it is significant so that we can assess its significance. Any of the other variables might be included or not with the exception that we need to test for some kind of interaction between grades and gender.
* We need some way to capture a potential interaction between grades and gender in determining income. Here the client is not very specific about what is required so we could reflect this in any number of ways. One approach (which I will use) is to determine if any of the four grades appear to be significant in explaining earnings and then check to see if that those interact with gender to yield a different result. It is worth pondering what other approaches you might use here. I created interactions between each of the grades and male in Excel.
* We will need to check for heteroscedasticity. As we know, there are several ways to do this – we might start with a theory for the possible cause – in this case, it seems hours worked would be a good source to consider. (It is worth pondering why theory might suggest hours worked as a source of heteroscedasticity in this case.) We may also use graphical analysis to see if we can spot patterns before we start testing.

I started building my model by including the following variables.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

From these results, I will eliminate Age, English, OB and Male and retain the others. A screenshot of a computer screen

Description automatically generated

The remaining variables all seem to be significant so I need to address the final issue which is to include the gender impact. Here we could include Male\_Calc and Male\_Acc to capture the gender interaction.

A screenshot of a computer

Description automatically generated

We now have the model necessary to answer our client’s questions:

Q1b – Heteroskdasticity does not appear to be an issue either as evidenced by graphical results nor by the Breusch-Pagan test.

Q1c – Higher calculus grades are associated with higher earnings – for each 1% increase is related to an increase in predicted earnings of about $33.65. By itself this is insufficient to justify hiring tutors for students in the anticipation of increasing their summer earnings for two reasons. The first is that correlation does not imply causation, perhaps being smart / hard working leads to success in both calculus and summer employment. Even if one were to accept that a causal relationship existed, one would have to determine that the earning would justify the cost of the tutors – it may cost $100 in tutoring to raise a calculus grade by 1% which would not be justified by summer earnings.

Q1d – There does not appear to be any difference in the relationship between earnings and grades that depend on gender.

1. A Kingston real estate company believes that the most important factor in selling a house is setting the initial price correctly. This company has hired you as a consultant to develop a model to predict housing prices so they can set initial prices accordingly. A collection of data has been provided to you on the tab ‘housing’. The client insists that the model should include all the explanatory variables they have provided: Price = B0 + B1 N\_Bedrooms + B2 N\_Bathrooms + B3 House\_Size + B4 Age + B5 Renovated\_Kitchen + B6 Finished\_Basement + B7 Close\_to\_Campus.
   1. In principle, how would you look for outliers in the data? If you found any how would you recommend that the client deal with them? Note: you do not need to actually do anything in R for this question – there are no outliers in the data.
   2. If there had been missing data elements, how would you recommend the client deal with that in their data? Note: you do not need to actually do anything in Python for this question – there are no missing values in the data.
   3. Aside from the size variables House\_Size and N\_Bathrooms, do all the variables appear to belong? If not, which would you remove if the client did not insist on including them? What is the consequence of including irrelevant variables?
   4. The client believes that the variables House\_Size and N\_Bathrooms should both be important in determining the selling price of homes.
      1. Do these variables appear to belong in the model according to the individual t-tests?
      2. Perform a test to see if any ‘questionable’ variables jointly belong in the model. What is the most likely explanation for the results?
      3. If you ran the model without House\_Size, what happens to the apparent significance of N\_Bedrooms? Would this make for a better model?
   5. Suppose I want to sell my house. Does your first model suggest that I should add a bathroom at a cost of $3,000 and that doing so would be justified by the resulting increase in selling price? Be sure to justify your answer.

2a – See notes / previous discussions on this from Session 3.

2b – See notes / previous discussions on this – Session ¾.

2c – Here we need to run the model that the client requested. The results are:

A screenshot of a computer

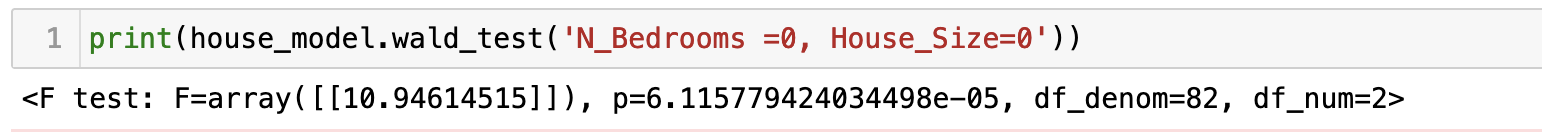
Description automatically generated

Several variables do not appear to be significant. N\_Bedrooms, Age, Renovated\_Kitchen, Finished\_Basement all appear to be insignificant. See text for discussion of the consequences of not including variables.

2d

i – House size does not appear to be significant, though N\_Bathrooms does.

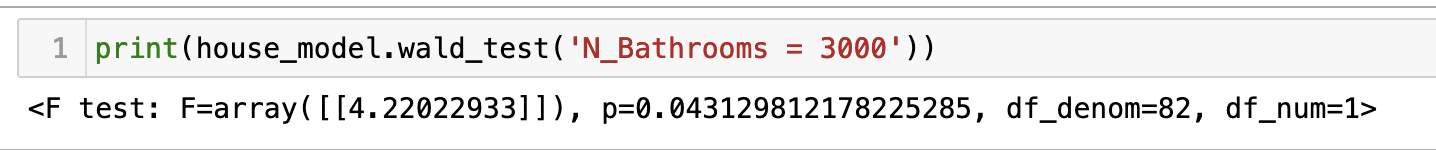
Ii – N\_Bedrooms and House\_Size are both suspect variables, we should run a joint test for this by using the following code.

The results are:

So they are not simultaneously equal to 0.

Iii – Removing House Size makes N\_Bedrooms Statistically significant with a p-value less than 0.0001.

2e – An additional bathroom is associated with an increase in price of about 29000, but there is a range around this so we should test the null hypothesis H0: B\_Bathroom <= 3000 vs H1: B\_Bathroom > 3000. We can test this with the code:



Which, for a one-way alternative, should be interpreted as a p-value of ~ 0.0215. This is not a particularly profound rejection, but it suggests that an additional bathroom is associated with a higher price. Subject to the standard observations of correlation vs. causation, this is likely to be a good idea.

1. The tab ‘Sales’ contains 12 months of sales data for a sample of stores from across Canada, along with their location, prices, and advertising budget.
   1. Consider the simple linear regression model: (i.e. ignoring the location and time data). Comment on the significance of Ad\_Budget and Price in this model? Based on this model, can you say with confidence that an increase in price of one unit is associated with a reduction in sales of more than 400 units?
   2. A new corporate strategy came into effect in month seven. Is there any evidence that the relationship between Sales and Ad\_Budget changes at that point going forward? Run an appropriate test and explain your results.

A)

- Consider significance of overall regression – not very significant, but one could let this go.

- Consider parameter estimates:

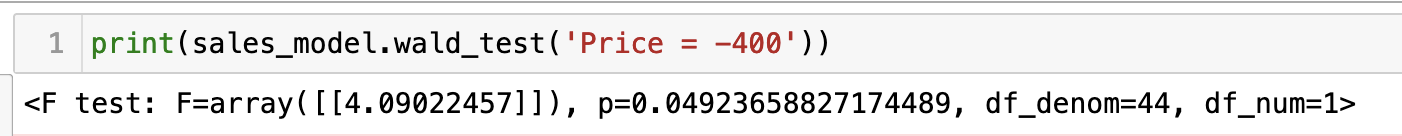
- Think here about the statistical significance, the sign and the practical significance.

- Might also consider the confidence interval around the estimates or run a hypothesis test

- Recognize correlation vs. causation

A screenshot of a computer

Description automatically generated



B)

-The best way is with the model:

Sales = B0 + B1 Ad\_B + B2\_Price + B3\_Dummy\_After\_Month\_6 \* Ad\_Budget

To allow a change in slope.

Looks like no difference to me. A screenshot of a computer

Description automatically generated

1. It is widely believed that the transit system’s revenue has been trending up over the course of the year, subject to the fact that sales are normally lower in the months of July and August, however, there are those that believe that somewhere around the beginning of May, the rate of growth increased significantly. Like most businesses, advertising budget and the cost of alternatives (i.e. gasoline for car commutes) should also play a role.

Using the data on the ‘TransitData’ tab, consider each of the issues below in isolation and construct a fairly complete test of each them alone (i.e. do not attempt to build a single model to test all of these things simultaneously.)

* + That there is an upwards trend in transit revenue as a function of time and that this trend has increased since early May.
  + Do advertising budget and gasoline each have the anticipated impact on transit revenue?
  + Is there evidence that the impact of gasoline price and advertising budget are somehow different in the summer, controlling for the 'summer effect' with dummy variables?

1. Will need a variable that increments 1 each observation starting in May. You can create it with this code:A screenshot of a computer

   Description automatically generated

It is not significant, suggesting the effect does not exist

1. No. Ad\_Budget does, but gas price does not.
2. We’ll need interaction terms with July/August and Gas\_Prices and Ad\_Budget:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedSummerAd is significant, but SummerGas is not. Together, they are significant (as we’d expect).

1. Consider the data on the CANvUS tab, which has data on Canadian and American respondents. Use this data to answer the following questions.

Build a model to explain Y using the X1..X4 variables and the US dummy.

1. Does being a US respondent appear to explain the results?
2. If being a US respondent does not explain the results, does that mean that Canadians and Americans are the same?
3. Run a model to explain Y using X1..X4 then perform a test to determine if Canadians and Americans are in fact the same based or different. Explain the results.
4. Without checking the data, what differences, in principle, could still exist between Canadians and Americans?

A screenshot of a computer

Description automatically generated

* 1. No.
  2. No – just that whatever effect we’re measuring does not exist within this dataset
  3. Run a chow test:A screenshot of a computer

     Description automatically generated

A screenshot of a computer

Description automatically generated

Appears to be no structural break

* 1. Any number of things – price sensitivity, population density, incomes, etc.