**COURSE\_NUM – ANOVA / Simple linear regression – Workshop activity**

**Task:** The goal of this workshop is to introduce you to running linear regression in Excel. This will prepare you for running regressions as part of the Module 3 and 4 main assignments.

**Preparation**

Begin by opening the Excel file “ANOVA simple linear regression – Workshop data”

You may need to add the “Data Analysis” add-in to your version of Excel.

* Click on “Data” in the header ribbon

If you do not see the “Data Analysis” option, you need to install the add-in.

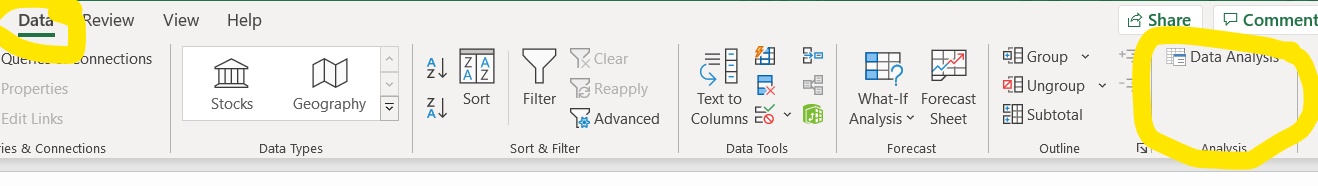
* Click on the Windows symbol in upper left-hand corner OR “File” in the header ribbon.
* Click on “Info”, and then below in the same column, click on “Options.”
* A new box or options will pop up. In the left-hand column, click “Add-ins.”
* Select “Analysis ToolPak” by clicking on it and then click the “Go” button below.
* Another box may pop up with a list of “Add-ins available.” Select “Analysis ToolPak” and click “OK.”
* You will know you are successful when back in your Excel document you click “Data” in the header ribbon and see “Data Analysis” listed in the “Analysis” box.

**Part 1: Estimate a simple linear regression model**

We are going to run a simple linear regression where housing affordability (“home\_afford” variable) is predicted by annual household income (“hh\_income” variable). The unit of analysis is people. The regression can be expressed like this:

Follow these steps:

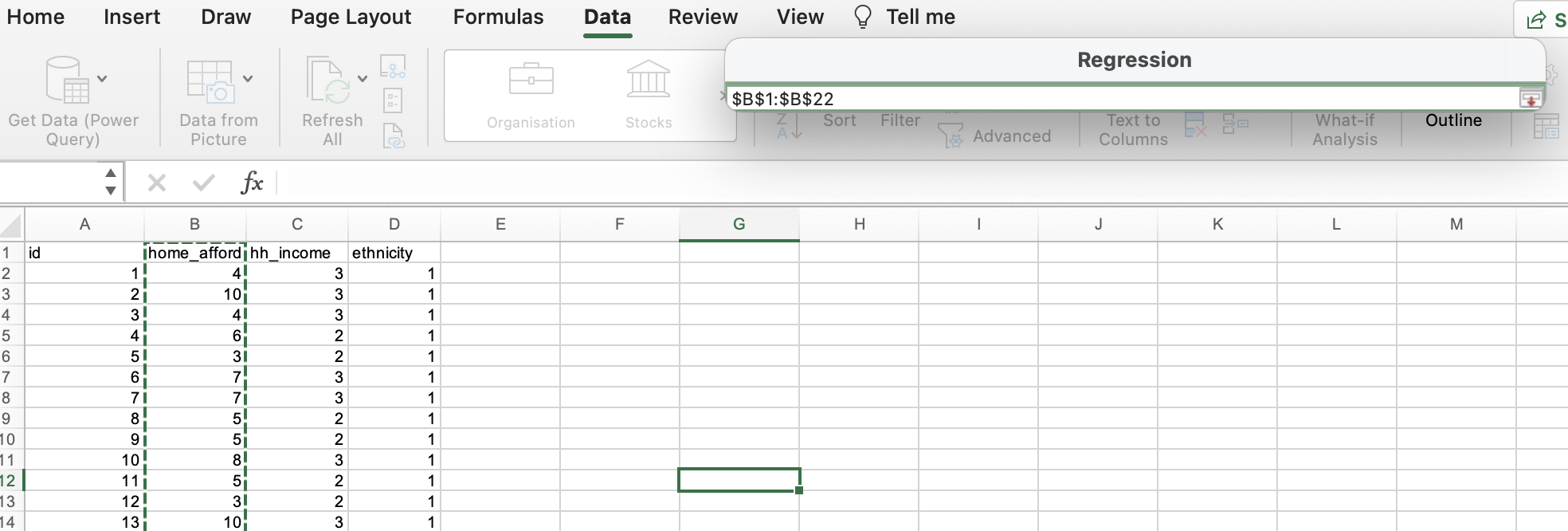
1. Click on “Data” in the header ribbon.
2. Click on “Data Analysis” in the “Analysis” box.



1. A box will pop up. Click on “Regression” in the list of “Analysis Tools”, then click “OK.”
2. Input the “Y” data—data in the “home\_afford” variable column—by clicking the little arrow button in the “Input Y Range” row.

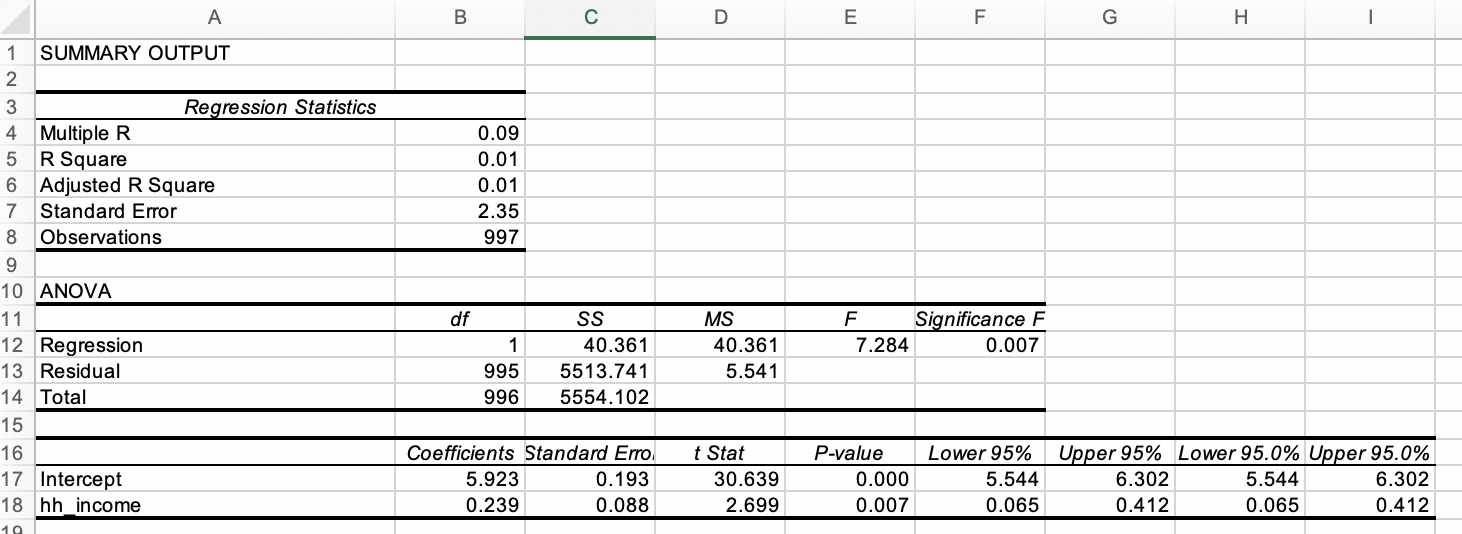


1. A new blank box titled “Regression” will appear. Select the data by clicking the variable name and dragging to the end cell in the column. In this dataset, that will be row 998.



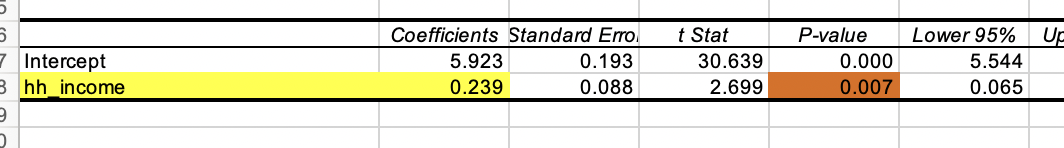
1. Once data are selected, click the little downward arrow button to the right, which will bring you back to the main “Regression” box.
2. Repeat steps 4-6 for inputting the X data—data in the “hh\_income” variable column.
3. Select “Labels” for including variable names in your output.
4. Select “New Worksheet Ply:” and type “Afford-Income Reg” in the text box to the right. This will create a new sheet to store your regression results.
   * Be consistent with decimal places with your output: try setting your coefficients and other values to three decimal places by:
     + Clicking on the sheet where your results are stored.
     + Click “Home” on the header ribbon.
     + Highlight the cells where you want to change the decimal places.
     + In the “Number” box in the header, click this button  until you get to three decimal places.

You should get this:



***Application to Module 3 main assignment***

For your main assignment in module 3, you will need to discuss the following statistics:

****

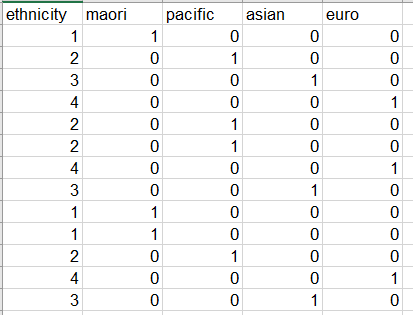
The yellow highlighted value (0.239) is the coefficient value showing the association between income and housing affordability. This statistic shows a positive association, whereby each increase in the level of income variable is associated with a 0.239 increase in the housing affordability scale.

The red highlighted value (0.007) is telling us the extent to which this association is statistically significant. The thresholds of significance are the same as you learned in Module 1. This is a statistically significant association at the *p <* .01 level.

**Part 3: Estimate a linear regression model with a categorical variable/set of dummy variables**

Often the variables we want to examine don’t lend themselves to be examined continuously, such as ethnicity. Because ethnicity is such as important variable, let’s practice turning our variable into a series of dummy variables and use them in a regression.

To begin, we need to translate our categorical variable in to a series of ‘dummy’ variables. Basically a set of indicators. We are aiming for this:



**Go to the tab “Ethnicity recode”: The data are already set up for you, including the equations.**

In essence, what we are doing with the COUNTIF statements, is asking Excel to look in the ‘ethnicity’ variable and create a “1” if the variable indicates the person is NZ European, for example, or leave as a “0” (i.e., not count the value) if that person does not identify as NZ European.

To complete the coding in this table:

* Remove the ‘ from the front of each COUNTIF statement to activate it
* In cell E2 (‘māori’ column), move the cursor lower right-hand corner until the black cross appears.
* Double-click on the lower-right corner to populate the entire column with that COUNTIF equation
* Repeat for column F, G, H, and I.

**Run a linear equation with dummy variables**

Let’s run a regression to examine whether housing affordability differs by ethnicity.

The regression can be expressed like this:

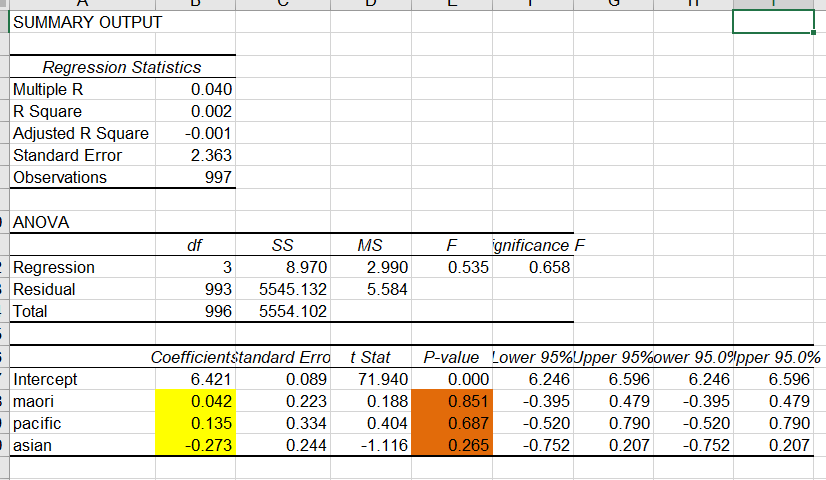
*Note how “Euro” is not in this model. “Euro” is our reference category. We’ll talk about this more in Module 4.*

Follow these steps:

1. Click on “Data” in the header ribbon.
2. Click on “Data Analysis” in the “Analysis” box.
3. A box will pop up. Click on “Regression” in the list of “Analysis Tools”, then click “OK”
4. Input the “Y” data—data in the “home\_afford” variable column—by clicking the little arrow button in the “Input Y Range” row.
5. A new blank box titled “Regression” will appear. Select the data by clicking the variable name and dragging to the end cell. In this dataset, that will be row 998.
6. Once data are selected, click the little downward arrow button to the right, which will bring you back to the main “Regression” box.
7. Input the “X” data—data in the “maori”, “pacific”, and “asian” columns. Do this by selecting the variable name from the first column, dragging down and diagonal across all four columns until you’ve selected all data in all four variable columns.
8. Select “Labels” for including variable names in your output.
9. Select “New Worksheet Ply:” and type “Afford-Eth Reg” in the text box to the right. This will create a new sheet to store your regression results.

***Application to Module 3 main assignment***

For your main assignment in module 3, you will need to discuss the following statistics:



The yellow highlighted values are the coefficient values showing the association between ethnicity and housing affordability. These coefficients should be interpreted as being the difference with the reference group.

As an example, the Māori coefficient of 0.042 indicates that Māori, on average, report a 0.043 higher housing affordability scale score compared to NZ Europeans. The red highlighted p-value of 0.851, however, indicates that this difference is not statistically significant.

As another example, the Asian coefficient of -0.273 indicates that Asian people, on average, report a 0.273 *lower* housing affordability scale score compared to NZ Europeans. The red highlighted p-value of 0.265, however, indicates that this difference is not statistically signicant.