Econ 270 Extra Credit Assignment

This assignment is due Friday, May 9th at midnight. You may submit answers in any reasonable format - just make sure both your answers and calculations can be clearly followed. You may do calculations in either a statistical language like R or python or in Excel. AI-generated answers will receive no credit. If you want to use AI to either write the code or tell you how to generate the pivot tables then that is fine.

NLSY

The National Longitudinal Survey is a series of surveys in the US where they survey an initial group of around 5000 individuals in the US and track these individuals with follow-up surveys throughout their lives. The NLSY97 tracks individuals who were aged 12-18 in 1997 and contains thousands of variables related to things like school results, earnings, and attitudes/beliefs. The publicly available data contains only a broad geographic indicator (4 major US regions), but there is also a restricted data set that can be applied for that contains more granular geographic data. For this assignment I have NLSY97 data containing a small subset of columns that have been transformed to included their decoded value that are included in the data dictionary provided. More information on the data is available at https://www.nlsinfo.org/content/getting-started.

The data contains the sex and race of each individual, their census region code (1-4), an urban indicator (1 if urban, 0 if rural), a gifted indicator (1 if gifted, 0 if not), their favorite ice cream flavor, and their earnings at age 25. If unemployed, their earnings at age 25 is recorded as 0.

Question 1

Calculate the mean earnings at age 25 for each category of ice cream (e.g. mean earnings for vanilla, chocolate, etc), along with the number of individuals in that category and the standard deviation of earnings. If using R this is easily done using the dplyr package with the group_by function, and pandas in python has similar dataframe functionality. In excel, you can use a pivot table to easily aggregate data like this.

Which flavors are associated with the highest and lowest earnings? Do you think it makes sense for these categories in particular to have different associated incomes?

Question 2

Let's focus on Butter Pecan. Calculate the mean earnings at age 25 for people whose favorite ice cream was butter pecan vs all other individuals (so only 2 categories this time). Now, conduct a formal hypothesis test to determine whether this income difference is statistically significant. Do the calculations as we did them in class rather than calling any built in t-test functions (you can always check your work against the built-in t-test functionality if you want).

Question 3

Suppose we obtain a significant result for butter pecan in question 2. How should we interpret this? Does this mean imply a causal link between ice cream preference and later earnings? Does it imply that ice cream is an independent predictor of future earnings?

Question 4

We want to know why ice cream flavor might predict future earnings. We have other covariates in our dataset that may also explain differences in income. In econometrics you'll have a more powerful way of finding joint patterns of correlation using multiple regression analysis. Here, we can still find what's driving our results by doing a two-way table. Rather than finding mean earnings by a binary indicator of whether their favorite flavor is butter pecan, we can do the same tabulation by the cartesian product of ice cream flavor and any covariates (e.g. gifted). If being gifted explains the difference in both ice cream preferences and earnings at age 25, then it is the case that when filtering to gifted=0, we will see mean earnings for butter pecan and non-butter pecan will be similar, and similarly when filtering to gifted=1.

We can formalize this by creating a two-way table and conducting a t-test. We can calculate the mean within-group difference in earnings (the weighted average of $earnings_{butterpecan;gifted} - earnings_{nonbutterpecan;gifted}$ and $earnings_{butterpecan;nongifted} - earnings_{nonbutterpecan;nongifted}$), and similarly we can calculate the standard error as we did with differences-in-means for t-tests. This will approximately give us the same t-statistic as would be obtained in multiple regression. If the difference remains significant even within-gifted comparisons, it implies that the difference in earnings explained by ice cream preferences is not explained by giftedness, whereas if the result is no longer significant, it implies that the difference was likely due to giftedness explaining the results and simply reflecting a joint correlation (confounding variable).

Conduct this t-test using the binary indicators of butter pecan preference and gifted. What do you conclude from this test?

Question 5

Why do kids who prefer butter pecan have lower earnings at age 25? This will likely require a decent amount of trial and error. Note that this is an extremely common real world scenario, as you'll often be asked to explain relationships in data.