Joel Cabrera

Econometrics 322

Professor Agan

9 February 2019

**R Data Analysis - HW #1**

Part 1: Determinants of Weight

1. We run a regression of weight on age (as a quadratic term), gender, income level, marriage status, and height, based on the NJHealth1000.csv dataset. For income level, only the middle- and high-income level variables will be included in the regression model. The low-income level variable will be omitted for it to be used as the reference group for income levels. In other words, not all 3 income levels can be used in the regression. This is because doing so would result in the violation of the fourth assumption of multivariate regression: no perfect multicollinearity. Having the low-income variable also in the model would make the model perfectly multicollinear, as shown by: *low-income + middle-income + high-income =* 1 (each variable takes on a value of 0 or 1). To avoid perfect multicollinearity, the multivariate regression model, then, takes the following form instead:

*Weight*  = *a* + B1\**age +* B2\**age2 +* B3\**sex +* B4\**middle-income +* B5\**high-income +* B6\**married +* B7\**height*

2. The regression model above is then performed on R. The following table indicates the results.

**Table 1:** Estimated Effects of Demographic Variables on (Individual) Weight

|  |  |  |  |
| --- | --- | --- | --- |
| **Regressor** | **(1)** | **(2)** | **(3)** |
| Age | 1.63\*\*\*  (0.47) | 1.63\*\*\*  (0.48) | 1.81\*\*\*  (0.51) |
| Age squared | -0.01\*\*  (0.00) | -0.01\*\*  (0.00) | -0.02\*\*\*  (0.00) |
| Sex | -15.06\*\*\*  (2.96) | -15.01\*\*\*  (3.02) | -39.99\*\*\*  (2.20) |
| Middle-income | -6.73  (8.21) | -6.88  (8.22) | -8.36  (8.77) |
| High-income | -17.13\*  (8.05) | -16.37  (8.21)\* | -15.79 (. = 0.05)  (8.74) |
| Married | -5.55\*  (2.38) | -5.62\*  (2.38) | -5.85\*  (2.47) |
| Height | 4.27\*\*\*  (0.38) | 4.23\*\*\*  (0.39) |  |
| Race |  | -0.66  (0.97) | -2.05\*  (1.02) |
| Education Level |  | -1.88  (1.24) | -1.99  (1.29) |
| Intercept | -129.01\*\*\*  (28.08) | -116.04\*\*\*  (29.59) | 180.93\*\*\*  (15.69) |
|  |  |  |  |
| Number of Observations | 1000 | 1000 | 1000 |
| R2 | 0.3594 | 0.3600 | 0.2668 |

**Note:** Heteroskedasticity-robust standard errors in parentheses. \*\*\* indicates p < 0.01. \*\* indicates p < 0.05. \* indicates p < 0.10. Numbers, except for R2, are rounded to 2 digits. Income levels are measured in tens of thousands. R2 is based on the adjusted R-squared computed result. Data is from the NJHealth1000.csv dataset. Low-income level is the reference group (omitted variable category).

3. (see column 1 of the table above for details on answers below)

3a. Performing a linear hypothesis test in R, where the coefficients on age and age-squared = 0, it is found that its F-statistic is 7.85, which is greater than 1.96 (critical value of a = 0.05). Due to this, it can be said that we reject the null hypothesis, and, thus, age and age-squared are significantly associated with weight. As a side note, the F-statistic was used because it not only was quick to determine statistical significance, but also the sample size that we have, 1000, is normally distributed according to the number of restrictions being test.

3b. The interpretation of the coefficient on my gender variable is: Being male is associated with a 15.06 pound decrease in (individual) weight (when sex = 1), holding all other independent variables constant. Moreover, being female is associated with a 30.12 pound decrease in (individual) weight (when sex = 2), holding all other independent variables constant. Performing a t-test on gender (joint hypothesis testing cannot be used due to the twofold levels that the variable contains), we find the mean weight, in pounds, on men and women are 195.32 and 156.28, respectively. However, the t-statistic is 17.95, which is greater than 1.96 (critical value of a = 0.05). Due to this statistical phenomenon, it must be said that the null hypothesis should be rejected, and thus, in terms of weight, men are significantly associated with women (they are significantly heavier).

3c. Performing the 95% confidence interval command on the regression model, it is found that the 95% confidence interval for the association between married and weight is (-10.05, -1.04). Interpreting said 95% confidence interval: The rate of change of the conditional mean of weight with respect to being married is estimated to be between -10.05 and -1.04.

3d. Performing a linear hypothesis test in R, where the coefficients on middle-income and high-income levels = 0, it is found that its F-statistic is 11.72, which is greater than 1.96 (critical value of a = 0.05). Due to this, it can be said that we reject the null hypothesis, and, thus, income – particularly, at both the middle and high levels - is significantly associated with weight. As a side note, the F-statistic was used because it not only was quick to determine statistical significance, but also the sample size that we have, 1000, is normally distributed according to the number of restrictions being test.

3e. Performing two linear hypothesis tests in R, separately for where the coefficients on middle-income and high-income levels = 0, it is found that their F-statistics are 1.32 and 8.91, respectively. Only the latter is greater than 1.96 (critical value of a = 0.05). Due to this, it can be said that we reject the null hypothesis for the latter but fail to reject the null hypothesis for the former. Thus, having a high-income level is significantly associated with weight. This, then, implies that middle-income people do not weight significantly more than high-income people.

3f. Performing the 95% confidence interval command on the regression model, it is found that the 95% confidence interval for the association between married and weight is (3.58, 4.95) Interpreting said 95% confidence interval: I am 95% confident that the rate of change of the conditional mean of weight with respect to height is estimated to be between 3.58 and 4.95.

3g. The independent variables in this regression explain about 35.94% of the sample variation in weight.

4. (See column 2 of the table above for details on answers below).

4a. The new regression does explain more of the sample variation in weight, but only very slightly. Instead of the independent variables explaining 35.94% of the variation in weight, they now explain 36% of the variation in weight (assuming we do not round the percentage numbers).

4b. Performing a linear hypothesis test in R, where the coefficient on race is 0, it is found that its F-statistic is 0.45, which is less than 1.96 (critical value of a = 0.05). Due to this, it can be said that we fail to reject the null hypothesis, and, thus, race is not significantly associated with weight. As a side note, the F-statistic was used because it not only was quick to determine statistical significance, but also the sample size that we have, 1000, is normally distributed according to the number of restrictions being test.

4c. Performing a linear hypothesis test in R, where the coefficient on race is 0, it is found that its F-statistic is 2.41, which is greater than 1.96 (critical value of a = 0.05). Due to this, it can be said that we reject the null hypothesis, and, thus, education is significantly associated with weight. As a side note, the F-statistic was used because it not only was quick to determine statistical significance, but also the sample size that we have, 1000, is normally distributed according to the number of restrictions being test.

5. (See column 3 of the table above for details on answer below).

5a. After performing and seeing the results of the regression without the “height” variable, I found that the coefficients on my independent variables did change by a lot. The change in the coefficient on gender was, in particular, of prominence to me – decreasing from approximately -15.01 to -39.99. A possible OVB story to explain the difference in the two results would be: height would be positively correlated with gender, since, on average, men tend to be taller than women. Furthermore, height would be a determinant of weight, since, the taller one is, the more weight that they will gain (e.g. larger bones, bigger body structure, etc.). Omitting this variable, then, probably affected the overall value of the regression – significantly lowering it in the process.

Part 2: Choosing Your Own Adventure

1. Considering the types of variables that are contained in the NJHealth1000.csv, I would like to pose and answer the following question: Is possession of a low-income level associated with a higher average number of alcoholic drink intake per month? Such a research question is of interest to me since it has been said that “money makes you happy”. The inverse of that axiom, then, would be “not having money makes you unhappy”. In the context of this investigation, an insufficient amount of money would be less than $15k, while unhappiness would be measured by the average amount of alcohol one drinks per month. As part of this research, I propose the following hypothesis (and its accompanying inverse).

Null Hypothesis (H0): In a comparison of individuals, those who earn an income of less than $15k are not more likely to have a higher average number of alcoholic drink intake per month than those who do not earn an income of less than $15. (That is, μd ≠ 0, or μi – μa ≠ 0, where i = income of less than $15k and a = average number of alcoholic drink intake).

Alternative Hypothesis (HA): In a comparison of individuals, those who earn an income of less than $15k are more likely to have a higher average number of alcoholic drink intake per month than those who do not earn an income of less than $15. (That is, μd ≠ 0, or μi – μa ≠ 0, where i = income of less than $15k and a = average number of alcoholic drink intake).

Using the provided “NJHealth1000.csv” dataset in R, I perform a regression of the monthly average alcohol intake variable, “avgDrinksPerDay”, on the income variable, “lowincome”. The table at the end of this paper illustrates the results of the computed univariate regression:

The coefficient on those who earn a low-income level (less than $15k) is -0.06. This means that: Having an income of less than $15k is associated with, on average, drinking 6 less alcoholic drinks per day over the last 30 days. Since the p-value for the coefficient is 0.804, which is greater than a = 0.05, we fail to reject H0. Furthermore, because the absolute value of the t-statistic, 0.25, for the coefficient on low-income is less than its critical value, 1.96, we still fail to reject H0. This analysis, then, indicates that having a low-income level is not significantly associated with average monthly alcoholic drink intake.

Despite the results for the univariate regression performed earlier, it must still be said that there are other variables that might be determinants of alcoholic drinking, and, thus, cause omitted variable bias (OVB). Such another variable could be the marriage status of an individual. An individual who earns a low income and is not married, for example, could be more prone to alcoholic drinking due to feelings of loneliness (being single) and unlikability (not having a partner to care for or love you). Furthermore, the marriage status of said individual could be correlated with a low income in the first place. Without having a partner, for example, that individual is responsible for all expenses – for materialistic desires, household responsibilities, paying off student debt/mortgages, etc., all of which is done completely on his or her own. To test this relationship, I perform a regression of the monthly average alcohol intake variable, “avgDrinksPerDay”, on the income variable, “lowincome”, and marriage variable “married”. Again, the table at the end of this paper illustrates the results of the computed multivariate regression:

The coefficients on those who earn a low-income level (less than $15k) and who are married are -0.09 and -0.08, respectively. This means that: Having an income of less than $15k is associated with, on average, drinking 9 less alcoholic drinks per day over the last 30 days. holding being married constant. Moreover, being married is associated with, on average, drinking 8 less alcoholic drinks per day over the last 30 days, holding having an income of less than $15k constant. Performing a summary function of the regression in R: Since the F-statistic is 0.52, and that it is less than 1.96 (critical value of a = 0.05),we fail to reject H0. This analysis, then, indicates that a having a low-income level and being married are not significantly associated with average monthly alcoholic drink intake.

**Table 2:** Estimated Effects of Low-Income & Marriage Status on Monthly Average Alcohol Intake

|  |  |  |
| --- | --- | --- |
|  | (1) | (2) |
| Low-Income | -0.06  (0.16) | -0.09  (0.17) |
| Married |  | -0.08  (0.09) |
| Intercept | 1.97\*\*\*  (0.04) | 2.02\*\*\*  (0.07) |
|  |  |  |
| Number of Observations | 1000 | 1000 |
| R2 | -0.0009400 | -0.0009553 |

**Note:** Heteroskedasticity-robust standard errors in parentheses. \*\*\* indicates p < 0.01. \*\* indicates p < 0.05. \* indicates p < 0.10. Numbers, except for R2, are rounded to 2 digits. Low-income level is measured in tens of thousands. R2 is based on the adjusted R-squared computed result. Data is from the NJHealth1000.csv dataset.