93/100

Group Members: Kailu Wang, Kaiyue Wang

Part 1: Determinants of Weight

The following questions are based on a sample of respondents in New Jersey in 2015 from the CDC's Behavioral Risk Factor Surveillance System, which is an excellent source for individual-level health data. Note I took a 1000 person random sample from amongst the New Jersey respondents. The codebook is very important and also available to download.

Can you include all 3 income levels in your regression? Why or why not?

No, because we cannot have perfect multicollinearity. Since these three income levels are highly correlated which means we need to set one of income level as reference group. For this we use low income as the reference group.

Table: Estimated Effect of Personal Character on Weights

	(1)	(2)	(3)
Age	1.63 *** (0.47)	1.47 *** (0.48)	1.61 *** (0.52)
Age squared	-0.0145 *** (0.0046)	-0.013 *** (0.0047)	-0.014 *** (0.005)
(sex)female	-15.06 *** (2.96)	-15.88 *** (3.0)	-40.55 *** (2.20)
Middle income	-6.73 (8.21)	-7.54 (7.88)	-8.75 (8.43)
High income	-17.13 ** (8.05)	-15.42 ** (7.8)	-14.38 * (8.34)
married	-5.55 ** (2.38)	-3.65 (2.35)	-3.71 (2.45)
height	4.27 *** (0.38)	4.15 *** (0.38)	------
Black (race)		16.25 *** (4.23)	15.27 *** (4.43)
Other race only		-8.35 (5.39)	-17.37 *** (5.81)
Multiracial		7.4 (17.45)	6.69 (16.08)
Hispanic		-3.73 (4.1)	-7.76 * (4.37)
Grades 1 through 8		-19.8 (60)	-11.06 (47.95)
Grades 9 through 11		-5.92 (59.59)	-3.84 (46.73)
High school graduate		-1.3 (59.59)	2.82 (46.13)

Some college		1.78 (59.03)	5.95 (46.11)
College graduate or more		-3.98 (59.06)	0.013 (46.13)
Constant	-129.01 ***	-118. 91 *	166.83 ***
(leave blank line above # of obs)			
Number of Observations	1000	1000	1000
R^2	0.3594	0.3769	0.288

Notes: Robust standard errors shown in parentheses. *** represent significance at 0.01 level, ** represent significance at 0.05 level, * represent significance at 0.1 level. Age squared is measured in year squared. We only care about People are married or not. Middle income is income>=$15,000 and <$75,000. **High income** is income>=$75,000. Heights are measured in inches. Weights are measured in lbs. Education level is measured in highest grade completed. Omitted dummy variable categories are: low income, male, not married, White and Never attend school or only kindergarten.

1) Questions about these results (for any questions about "significance" you can default to the 5% level or state a different level that you choose in your sentence):
    a. Is there a quadratic relationship between age and weight? Is it significant? How do you know?
    Yes, there is a quadratic relationship between age and weight. It is significant, because t-statistic for age squared is |-3.12| which is greater than 1.64, so it is significant at 5% level.
    b. How do you interpret the coefficient on your gender variable? Are men significantly heavier than women?
    In average, women's weight is 15.06 lbs. less than men's weight.
    Men are significantly heavier than women because it has three stars which means P-value is less than 1%, so they all significant in 1% level.
    c. What is the 95% confidence interval for the association between being married and weight? What does this mean?
    (-5.55-1.96*2.38, -5.55+1.96*2.38) = (-10.21, -0.89)
    This means that we are 95% confidence that the true difference of married people's weight between not married people's weights are falls in interval (-10.21, -0.89). Also, it is significant at 5% level, because there is no zero in this interval.
    d. Is income significantly associated with weight? How do you know? (which test statistic did you use and why?)
    Yes, the income is significantly associated with weight, because by doing Joint hypothesis test. We can see that P-value on F-statistic testing that both middle income and high income coefficients are zero is 0.0302 which is less than 0.05 so is significant at 5% level. I use F-statistic because we cannot have perfect multicollinearity and since all income levels are highly correlated, so we need to use

F-statistic.

e. Do middle income people weigh significantly more than high income people? How do you know? (note you need to run an additional command)

Yes, middle income people weigh significantly more than high income people. Because, by doing Joint hypothesis test. We can see that P-value on F-statistic testing that both middle income and high income coefficients are zero is 0.0302 which is less than 0.05 so is significant at 5% level. So, we can reject the null hypothesis, so middle income people weight significantly more.

f. How confident can you be that height has a statistically significant association with weight.
I am fairly confident that height has a statistically significant association with weight because it's p-value is very small which is less than 2.2 e^-16 and have three stars, so we are very confident.

g. How much of the sample variation in weight can you explain with this regression?
The samples variations can explain 35.94% variation in weight.

h. Now you want to add race into your regression and education level into your regression. Run your new regression in R, include all the variables from the previous question AND these new ones. Add a second column to your table that includes these new results.
    a. Does this new regression explain more of the sample variation in weight?

The new regression explains more of the sample variation in weight because it explains 37.69% variation in weight which is greater than 35.94%.

    b. Is race significantly associated with weight? How do you know? (which test statistic did you use and why? See the Joint hypothesis testing R tutorial we did in class)

Race is significantly associated with weight, because it's p-value on F-statistic when all coefficients are zero is 3.335 e^-5 which is less than 0.05. This is at 5% level significance. I use F-statistic because we cannot have perfect multicollinearity and since all races are highly correlated, so we need to use F-statistic.

    c. Is education level significantly associated with weight? How do you know? (which test statistic did you use and why?)

-3 Education level is significantly associated with weight, because it's p-value on F-statistic when all coefficients are zero is 0.01226 which is less than 0.05. This is at 5% level significance. I use F-statistic because we cannot have perfect multicollinearity and since all education levels are highly correlated, so we need to use F-statistic.

Run the regression from 4) again, but do not include height. Include this as the 3rd column in your table.

    d. Do any of your coefficients change a lot? Choose the one that changed the

most and tell an omitted variable bias story to explain the difference in the two results

Yes, one of variables that changed the most is female in sex. In the previous one, we know everyone's heights which is a huge determinant in a people's weight. And when we compare table two on females, we can find that with variable height, in average, male is 15.88 lbs. heavier than female. However, without variable height, in average, male is 40.55 lbs. heavier than female, which is huge difference than last one.

-2

height & gender are correlated

Part 2 (30 points): Choose your own adventure

There are a lot of interesting potential associations to estimate in this dataset. One interesting dependent variable is the average number of alcoholic drinks someone has. Ask an interesting question using this data about an association between average drinking and some independent variable. Run the univariate regression which shows this association.

Create a new table (similar to the previous one) that shows this univariate relationship in Column 1. How do you interpret the coefficient on age variable? Is it statistically significant?
Additional year of age increase is associated with 0.015 decrease in average number of alcoholic drinks per day over the last 30 days.
This is significantly significant because it has three stars which based on our notes it is significance at 1% level.

What other variables might be determinants of drinking? Might they cause omitted variable bias (why or why not). Choose one variable (or set of variables) that might cause OVB that is also included in the dataset and add it to your regression. Run this new regression and add it to column 2. How do you interpret the new results?
A person's income may also determinants of drinking. Also, gender difference may also influence the quantity of drinking. They could cause omitted variable bias, for example, one person's income is correlated with his or her age (elder people may have more experience and earn more), weight (people with high income maybe lighter, lower income maybe heavier) and whether or not employed (employed people earn money, unemployed people may be earn zero). Also, the income could determinant average number of alcohols a person drink. Gender can't cause OVB, because it's not correlated with age. A person's age is not determined by gender.
Choose income as one variable that cause OVB. Set low income as reference group.
People who have high income drink 0.125 more alcohol in average number of alcoholic drinks per day over the last 30 days than low income people. Middle income people drink 0.12 more alcohol in average number of alcoholic drinks per day over the last 30 days than low income people.

one regressor

Table: Effect of personal image on average number of alcohol drinks by someone

	(1)	(2)	(3)
age	-0.015 *** (0.003)	-0.0156 *** (0.003)	
weight	0.0052 *** (0.001)	0.0053 *** (0.001)	
employed	0.047 (0.088)	0.036 (0.098)	
High income		0.125 (0.18)	
Middle income		0.12 (0.189)	
Constant	1.83 ***	1.72 ***	
(leave blank line above # of obs)			
Number of Observations	1000	1000	
R^2	0.049	0.047	

Notes: Robust standard errors shown in parentheses. *** represent significance at 0.01 level, ** represent significance at 0.05 level, * represent significance at 0.1 level. Drink alcohol is measured in average number of alcoholic drinks per day over the last 30 days. Age is measured in years. Weight is measured in lbs. We only care people are employed or not. Middle income is income>=$15,000 and <$75,000. **High income** is income>=$75,000. Omitted dummy variables are low income, male and not employed.