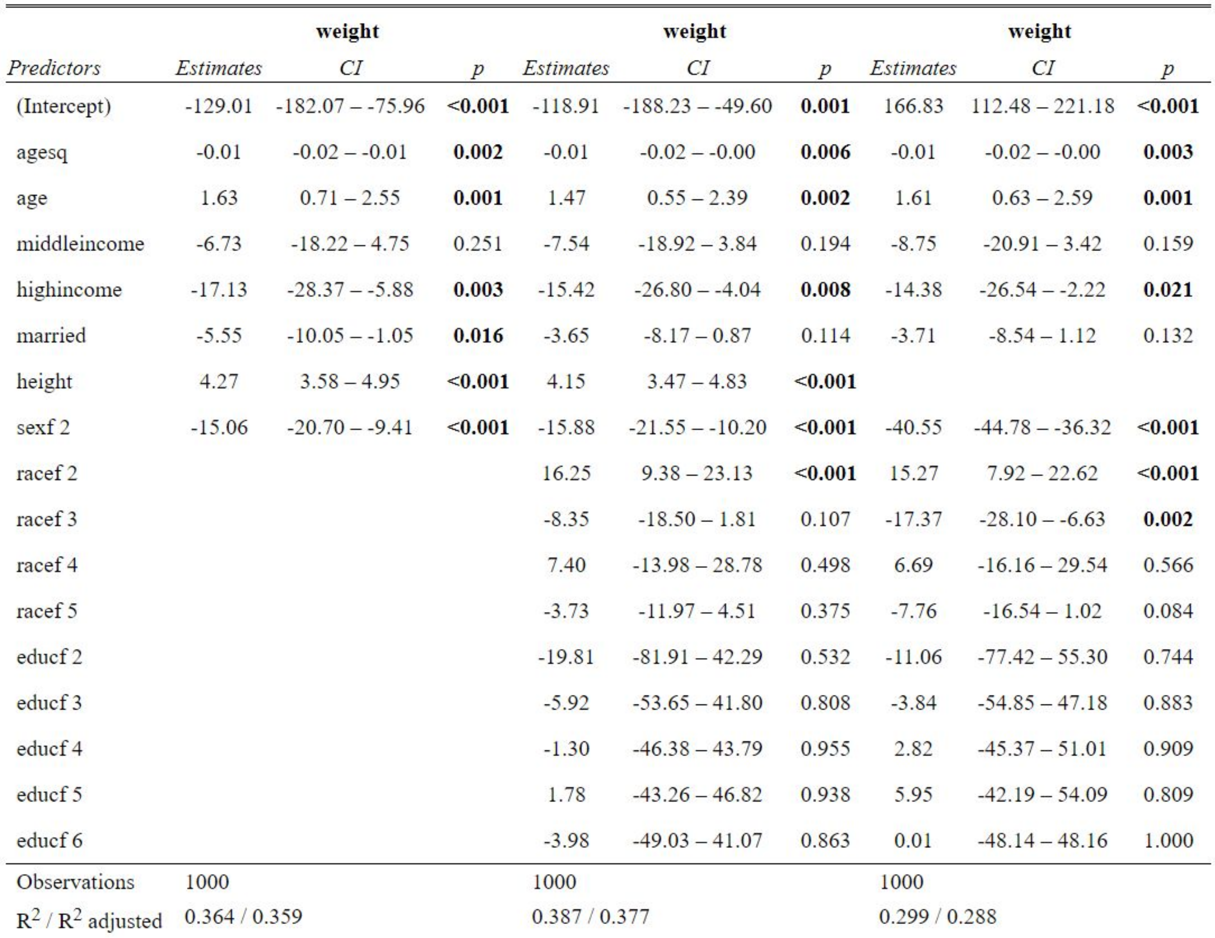
Ryan Shah, Ahmed, Zil
Professor Yu
Econometrics
13 August 2019

Analysis of Weight



		weight			**weight**			**weight**	
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-129.01	-182.07 – -75.96	**<0.001**	-118.91	-188.23 – -49.60	**0.001**	166.83	112.48 – 221.18	**<0.001**
agesq	-0.01	-0.02 – -0.01	**0.002**	-0.01	-0.02 – -0.00	**0.006**	-0.01	-0.02 – -0.00	**0.003**
age	1.63	0.71 – 2.55	**0.001**	1.47	0.55 – 2.39	**0.002**	1.61	0.63 – 2.59	**0.001**
middleincome	-6.73	-18.22 – 4.75	0.251	-7.54	-18.92 – 3.84	0.194	-8.75	-20.91 – 3.42	0.159
highincome	-17.13	-28.37 – -5.88	**0.003**	-15.42	-26.80 – -4.04	**0.008**	-14.38	-26.54 – -2.22	**0.021**
married	-5.55	-10.05 – -1.05	**0.016**	-3.65	-8.17 – 0.87	0.114	-3.71	-8.54 – 1.12	0.132
height	4.27	3.58 – 4.95	**<0.001**	4.15	3.47 – 4.83	**<0.001**			
sexf 2	-15.06	-20.70 – -9.41	**<0.001**	-15.88	-21.55 – -10.20	**<0.001**	-40.55	-44.78 – -36.32	**<0.001**
racef 2				16.25	9.38 – 23.13	**<0.001**	15.27	7.92 – 22.62	**<0.001**
racef 3				-8.35	-18.50 – 1.81	0.107	-17.37	-28.10 – -6.63	**0.002**
racef 4				7.40	-13.98 – 28.78	0.498	6.69	-16.16 – 29.54	0.566
racef 5				-3.73	-11.97 – 4.51	0.375	-7.76	-16.54 – 1.02	0.084
educf 2				-19.81	-81.91 – 42.29	0.532	-11.06	-77.42 – 55.30	0.744
educf 3				-5.92	-53.65 – 41.80	0.808	-3.84	-54.85 – 47.18	0.883
educf 4				-1.30	-46.38 – 43.79	0.955	2.82	-45.37 – 51.01	0.909
educf 5				1.78	-43.26 – 46.82	0.938	5.95	-42.19 – 54.09	0.809
educf 6				-3.98	-49.03 – 41.07	0.863	0.01	-48.14 – 48.16	1.000
Observations	1000			1000			1000		
R^2 / R^2 adjusted	0.364 / 0.359			0.387 / 0.377			0.299 / 0.288		

Notes: The left section is regression 1, the middle section is regression 2, and the right section is regression 3. "Sexf 2" is a binary variable where if the person is female, we set it equal to one. "racef 2-5" represents a categorical variable where each race was labeled between 1-5. We exclude "racef 1" to avoid perfect multicollinearity. "educf 2-6", similar to racef, is a categorical variable assigning numbers 1-6 to the level of education. Education: 1=never attended school or only kindergarten, 2=grades 1 through 8, 3=grades 9 through 11 (some high school), 4=high school graduate, 5=some college, 6= college graduate or more. We excluded educf 1 to avoid perfect multicollinearity. Under "estimates" for each regression, we have the coefficients for the regressors. CI refers to the 95% confidence intervals. "p" is the p-value.
age: age (restricted to be below 80)

racef: 1=white (non-hispanic), 2=black (non-hispanic), 3=other race only (non-hispanic), 4-multi-racial, 5=hispanic

educf: highest grade completed - 1=never attended school or only kindergarten, 2=grades 1 through 8, 3=grades 9 through 11 (some high school), 4=high school graduate, 5=some college, 6= college graduate or more

married: =1 if married, =0 otherwise

lowincome: income<$15,000
middleincome: income>=$15,000 and $75,000
highincome: income>=$75,000

We could not include all three incomes as that would cause perfect multicollinearity.

First regression (weight on age, gender, income, married, height)

There is a quadratic relationship on age as we cannot fail to reject the null hypothesis that age is significant. The p values for age and age squared are 0.0006069 and 0.01800 which are both at least significant at the 95% level. If we say the null hypothesis is that are not significant, we will reject the null hypothesis. Thus, there is a quadratic relationship on age. We interpret the gender variable (sexf) as follows: if the person is female, they will weigh 15.06 pounds lighter assuming other independent variables are all equal. This is significant at the .01 level, so we can say that men weigh much more than women. The 95% confidence interval for the association between being married and weight is [-10.05 – -1.05]. This means we are 95% confident that the population parameter (being married and its relationship to weight) falls between these values. In order to see if income is significantly associated with weight, we must use the F-test because this is a joint hypothesis about regression coefficients. Running the F-test (joint hypothesis about regression coefficients (Ho: coefficients=0, Ha: coefficients ≠0) we get an F value of F=8.534, which is significant at the 5% level. Thus, we can say that income is significantly associated with weight. We need to use another F-statistic (single restriction on multiple coefficients). Using R and heteroskedasticity-robust errors, the F-statistic for this hypothesis is F=14.1, which is significant at the 5% level. So, we can reject the null hypothesis at 5% significance, so with 95% certainty, middle income people weigh more than high income people.

We can say that height is associated with weight on a 99% level with a t value of 11.2039 and a p-value of 2.2e-16. Our R^2 / R^2 adjusted are 0.364 / 0.359. Thus, we can say that the variables can explain 35.9% of the variation of weight for individuals. We are using the adjusted R^2 because adjusted R^2 is preferable to R^2 in that adding a new variable always leads to an increase in R^2 while adjusted R^2 will only increase if the new variable is somewhat significant. By somewhat significant I mean that the coefficient has a t-stat greater than 1 in absolute value.

Second Regression (weight on age, gender, income, married, height, race, education)

The adjusted R^2 is 0.377. Thus, this regression explains 37.7% of the variation for weight for individuals based on the variables that existed and the ones that were added. That is 1.8% more than the first regression. To see if race is associated with weight, we will use the F-test. We get an F value of 6.9661 and a p-value of 1.572e-05. This is significant at the 99% level. To see if education is significantly associated with weight, we use the F-test again (joint hypothesis about regression coefficients (Ho: coefficients=0, Ha: coefficients ≠0). We get an F=5.2012, which is significant at the 5% level. So, we can reject the null hypothesis at 5% significance, so with 95% certainty, race is significantly associated with weight.

Third Regression (weight on age, gender, income, married, race, education)

The gender coefficient changed a lot after excluding height. This signals omitted variable bias. The reason this bias exists is because height is correlated with gender, but we do not include height in this regression. Thus, a bias exists.