Multiple Regression Example: Smoking & Birthweight

SPSS download: http://bit.ly/iGMOgu
Stata download: http://bit.ly/lHa2G0

Source: Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* 21: 489-519.

Estimating the effect of smoking on the health of babies is difficult, since omitted (unobserved) variables are likely to be correlated with a mother's decision to smoke. In this study, Abrevaya dealt with this problem by looking at data from mothers with multiple births. This allowed the researcher to control for those omitted factors while analyzing the effect of smoking.

Here's a sample of the data collected for 8,604 births to 3,978 women:

Mom ID	Child	Birth Weight (grams)	State	Mother Age	Mother smoke?	Child male?	Mother married?	Mother High School?	Mother some college?	Mother college grad?	Black
14	1	2790	AL	16	0	0	0	0	0	0	1
14	2	2693	AL	17	0	0	0	0	0	0	1
14	3	3600	AL	20	0	0	0	0	0	0	1
5770	1	2807	IA	22	1	0	1	1	0	0	0
5770	2	2948	IA	23	1	0	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...

Are there any concerns you have about conducting a multiple regression analysis on this data?

For this example, we will analyze only the data for firstborn children. To help in writing our regression models, let's let:

Y = birthweight
X_1 = mother age?
X_2 = mother smoke?
X_3 = child male?
X_4 = mother married?
X_5 = mother high school?
X_6 = mother some college?
X_7 = mother college graduate?
X_8 = mother black?

Here are correlations among our variables:

	birwt	mage	smoke	male	married	hsgrad	somecoll	collgrad	black
birwt	1.0000								
mage	0.1244	1.0000							
smoke	-0.1991	-0.1566	1.0000						
male	0.1033	-0.0044	-0.0083	1.0000					
married	0.1599	0.3079	-0.2844	-0.0074	1.0000				
hsgrad	-0.0330	-0.2082	0.0876	0.0243	-0.0991	1.0000			
somecoll	0.0208	0.0233	-0.0494	-0.0377	0.1106	-0.3535	1.0000		
collgrad	0.1094	0.4288	-0.2418	0.0128	0.2551	-0.4774	-0.4156	1.0000	
black	-0.1460	-0.1496	0.0419	-0.0100	-0.3686	0.0839	-0.0226	-0.1467	1.0000

Conduct appropriate analyses to address the following questions. You will present your analyses to the class. Remember to consider the assumptions necessary for your chosen analyses.

1. Do mothers who smoke have lighter babies? Are girls lighter than boys at birth? Check the assumptions necessary for your analysis.

2. Describe the distribution of each of the following variables: mage, meduc, birwt, cigs, married. Create visualizations and calculate summary statistics.

3. Correlate the following variables: birwt, mage, smoke, gestat.

4. Evaluate several regression models and determine which one you would use to predict the birthweight of a firstborn child. First, compare a single predictor to a reduced model with no predictors. Then, test the added contribution of a second variable. Finally, test the added contribution of a third variable. Consider and evaluate the appropriateness of the assumptions necessary to conduct a multiple regression analysis.

5. Find the best-fitting curve to predict birthweight as a function of the mother's age.

6. Create a model to predict whether a mother smokes or not (based on a single independent variable).

.