Pre-Class Response for Lecture 16:

Earlier in the quarter, we talked about *omitted variable bias*; that is, what might go wrong if we don't observe an important confounding variable.

Now, let's think about what happens if we have *too many* observed variables (if there is such a thing). Suppose that we have a data set with, say, 1000 people, and we want to estimate the returns to education from this dataset. We have information about all the stuff that seems important -- years of education, where they went to school, their complete financial history, lots of info about their parents, etc -- as well as lots of stuff that may or may not seem important -- eye color, mother's maiden name, information on childhood pets, zodiac sign, etc.

(Of course, we don't know ahead of time what's important and what's not! According to some sources, your zodiac sign is a great predictor of all sorts of personality traits; and even if you aren't into astrology, zodiac sign tells us your approximate month of birth, which tells us, say, if you were old or young for your grade, which might be important for education outcomes...).

Altogether, say we observe about 2000 variables about each of these 1000 people.

Suppose we run a regression to figure out the effect of education on income, controlling for all of the variables that we observe. That is, we write down the regression

$$Y = \beta \cdot X + \gamma_1 \cdot W_1 + \gamma_2 \cdot W_2 + \cdots + \gamma_{2000} \cdot W_{2000} + \text{noise},$$

where Y is ln(earnings), X is years of education, and $W_1, \ldots, W_{2000}$ are our 2000 other observables.

Do things only get better when you get more variables? Could something go wrong? How would you interpret a $\hat{\beta}$ that you estimated from this regression?