Problem set 12: week 16

2. (This is a log-level model.) Increasing the number of packs smoked per day by 1 decreases the predicted weight of the baby by 8

3. If $E[\epsilon_i|packs] = 0$, then we say packs is an exogenous variable and the OLS estimator of β is unbiased, and we can give the parameter a causal interpretation. However, if $E[\epsilon_i|packs] \neq 0$ then we say packs is an endogenous variable and the OLS estimator of β will be biased, such that we cannot give it a causal interpretation. The important question is: Do we believe $E[\epsilon_i|packs] = 0$?

- Important factors in the error term are likely to include: health factors, life-style factors and prenatal care quality. Since all these factors are likely to affect birthweight.
- Given number of packs smoked is a choice variable (it is not fixed or random), health factors, life-style choice, ... etc are also likely to be correlated with number of packs of cigarettes smoked.
- Thus, we would expect $E[\epsilon_i|packs] \neq 0$. Number of packs smoked is picking up other life-style choices, which also affect birthweight.
- Because many of these factors are likely to be unobservable to the researcher, it is not possible to control (hold fixed) all confounding factors. An alternative to solve this endogeneity problem is IV estimation.

4.

- A possible solution is to use an IV. A good IV satisfies two criteria:
  1. instrument exogeneity: $cov(price, \epsilon) = 0$.
  2. instrument relevance: $cov(price, packs) \neq 0$.
- How does price fit these criteria?
- instrument exogeneity: likely to be satisfied, since price is unlikely to be correlated with the error term (health factors, life-style, prenatal care). Thus, we can argue price is only likely to have an indirect affect on birthweight, through its affect on number of packs smoked.
- Instrument relevance: prior to testing this, basic economic theory suggests there could be a relationship. Increase price reduce demand. However, given smoking is addictive, demand may be inelastic.

5.

- To test for the relevance condition regress: $pack_i = \pi_0 + \pi_1 price_i + \eta_i$. Then test $H_0 : \pi_1 = 0$ against the two-sided alternative.
- The relationship is highly insignificant (p-value 0.718). Thus, the relationship between price and packs is very weak. This is not that surprising, given demand for cigarettes is likely to be highly inelastic.
- Therefore, price does not satisfy the relevance criteria and is not suitable as an IV.
- However, what happens if we continue and use price as an IV?

6. The coefficient on packs is massive (2.98 which in a log level model implies increasing packs by one during pregnancy increases birthweight by 300%) and it is also in the unexpected direction. Further, the standard error is extremely large (almost 9, a factor of 3 greater than the estimated coefficient). However, the above regression is meaningless since the IV fails the relevance requirement. We essentially have a (very) weak instruments problems. Weak instrument causes two main problems. 1. for precision (variance) of an estimator:

- Estimating the regression $\ln(bght_i) = \alpha + \beta_1 packs_i + \epsilon_i$ using OLS, the variance of the OLS estimator is given by $V(b^{OLS}) = \frac{S^2}{\sum_n (X_i - \bar{X})^2}$.
- Estimating the regression $\ln(bght_i) = \alpha + \beta_1 packs_i + \epsilon_i$ using IV 2SLS, with price (Z) as an instrument for packs (X), the variance of the IV estimator is given by $V(b^{IV}) = \frac{S^2}{\sum_n (X_i - \bar{X})^2 R^2_{X,Z}}$, where $R^2_{X,Z}$ is the R^2 from the regression of $packs_i = \pi_0 + \pi_1 price_i + \epsilon_i$. In our case $R^2_{X,Z}$ is extremely low: $R^2 = 0.0001$.
- Thus, the variance of the IV estimator relative to the OLS estimator is huge. From the Stata regressions we see $se(b_1) = 0.0169$ in the OLS, while it is $se(b_1) = 8.692$ in the IV case.

2. Weak instrument also causes a problems for bias in the IV estimator (if there is marginal failure of the exogeneity IV assumption):

$$\begin{aligned}
b^{IV} =& \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} \\
=& \frac{Cov(Z_i, \alpha + \beta X_i + \epsilon_i)}{Cov(Z_i, X_i)} \\
=& \beta \frac{Cov(Z_i, X_i)}{Cov(Z_i, X_i)} + \frac{Cov(Z_i, \epsilon_i)}{Cov(Z_i, X_i)}
\end{aligned}$$

Then given $Corr(Z_i, \epsilon_i) = \frac{Cov(Z_i,\epsilon_i)}{\sigma_Z \sigma_\epsilon}$, we have:

$$\begin{aligned} b^{IV} =& \beta + \frac{Corr(Z_i, \epsilon_i)\sigma_Z \sigma_\epsilon}{Corr(Z_i, X_i)\sigma_Z \sigma_X} \\ =& \beta + \frac{Corr(Z_i, \epsilon_i)}{Corr(Z_i, X_i)} * \frac{\sigma_\epsilon}{\sigma_X} \end{aligned}$$

- Therefore, even if $Corr(Z, \epsilon)$ is very small, the bias of the IV estimator can be very large if Corr(Z,X) is also small (i.e. if you have a weak instrument).
- Thus, not only do weak instruments lead to large variances, but if $Corr(Z, \epsilon)$ is small then biases in IV estimates can be very large also.
- For example, in our case price may be higher, or lower, in regions where demand for cigarettes are very inelastic (in deprived regions for example).
- **Test for weak instruments**: an F-test of below 10 in the first stage is often seen as a indicator of weak instruments. In our case the F statistic is extremely low.

Now we consider a potentially stronger instrument used by a well known paper by Card.

7. Increasing education by one year increases wage by 7.5% on average.

8. Education is a choice variable (beyond a certain age), therefore we are concerned about the exogeneity of education. For instance, it may be correlated with ability which is in the error term. A potential solution is to use an IV for education. Card proposed proximity to college as an IV. Instrument relevance: intuitively individuals that live closer to college are more likely to go. Indeed running the first stage regression in Stata we see that those who grew up near a college in 1976 (controlling for experience, race, region, ... etc) had about a third of a year more education than those who didn't grow up near a college. Further, the t-statistic is 3.64 ($F = t^2 = 13.24 > 10$) which is very significant (a p-value=0 at 3 decimal places). Therefore, we can use college proximity as an IV for education (maintaining the assumption that it is exogenous). Instrument exogeneity: on first inspection there is no strong reason to believe that college proximity is correlated with the error term. So we will continue to use it as an IV.

9. The IV estimate is nearly twice as large as the OLS estimate. This implies the OLS estimate was downward bias (which is in the opposite direction to our usual argument that education is also picking up the ability affect). However, note the IV standard error is such that the 95% confidence interval for the IV is $(0.024, 0.239)$, which contains the OLS estimate. This represents the trade-off we face with IV estimation: a lower precision for an arguably more consistent estimator.

- Note we can also question exogeneity:
- Potentially parents with higher ability/education are more likely to locate in places where there is a college. Therefore, under the assumption that child ability is correlated with parent ability, we still may be concerned with the exogeneity assumption.
- Further, colleges are maybe in regions with a more developed labor market, implies more likely to get a higher wage job directly (as well as the indirect effect through education). This is one reason region is controlled for in the regression.