Ordinary Least Squares

Professor Dan Black **PP 414:Applied Regression Analysis: Analysis of Microeconomics Data**

2018

Overview of lecture

- **OLS as a minimum distance estimator**
- OLS as a method of moments estimator
- OLS as a MLE estimator
- OLS as a weighting estimator
- OLS and the common effects assumption
- Fully saturated OLS models
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

OLS as a minimum distance estimator

- OLS is an awesome method of estimation. It is so awesome we can interpret it at least four different ways
- First, OLS is a minimum distance estimator because it solves the following problem:

$$\min_{\beta} V = \sum_{i=1}^{n} \left(y_i - x_i\beta\right)^2$$

- The solution to this problem, $\hat{\beta}$, provides the minimum distance between each realization of y_i and its predicted value, $\hat{y}_i = x_i\hat{\beta}$
- It is not too hard to show that the function V is strictly convex in β under some conditions on X so the solution is unique
- $\Big(y_i - \hat{y}_i)^2$ is just the square of the Euclidean distance between realization y_i and its predicted value, $\hat{y}_i$
- OLS is just a minimum distance estimator!

Overview of lecture

- *OLS as a minimum distance estimator*
- **OLS as a method of moments estimator**
- OLS as a MLE estimator
- OLS as a weighting estimator
- OLS and the common effects assumption
- Fully saturated OLS models
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

OLS as a method of moments estimator

- But wait! OLS is a method of moments (MOM) estimator too.
- OLS is a minimum distance estimator because it solves the following problem:

$$\min_{\beta} V = \sum_{i=1}^{n} \left(y_i - x_i \beta \right)^2$$

The necessary conditions may be written as

$$\bar{y} - \bar{x} \cdot \hat{\beta} = 0$$

$$s_{y,x_j} - \hat{\beta}_1 s_{x_1,x_j} - \hat{\beta}_2 s_{x_2,x_j} \cdots - \hat{\beta}_k s_{x_k,x_j} = 0 \;\; \forall \; x_j$$

- Thus, the solution to the necessary conditions are just a weighted sum of the various moment of the data, or $\hat{\beta} = f(\bar{y}, \bar{x}, s_{y,x_j}, s_{x_j,x_k})$
- OLS is just a method of moments estimator!

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- **OLS as a MLE estimator**
- OLS as a weighting estimator
- OLS and the common effects assumption
- Fully saturated OLS models
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

OLS as a MLE estimator

- But wait! OLS is a Maximum Likelihood Estimator (MLE). To see why, let
$$y_i = x_i\beta + \epsilon_i$$
and assume that $\epsilon_i \overset{d}{\sim} N(0, \sigma^2)$
- This would allow us to form a "likelihood function", which is
$$L(\beta) = \prod_{i=1}^{n} \phi\Big(\frac{y_i - x_i\beta}{\sigma}\Big) = \prod_{i=1}^{n} (2\pi\sigma^2)^{\frac{-1}{2}} e^{\frac{-1}{2\sigma^2}(y_i - xi\beta)^2}$$
- Now a fun fact about optimization theory. If $f(x)$ is greater than zero, then if x^* solves $\max_x f(x)$ it also solves $\max_x \ln(f(x))$
- Thus, we can work with
$$\ell(\beta) = \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

OLS as a MLE estimator

$$\ell(\beta) = \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

- But this is still ugly so other fun facts about optimization theory. If x^* is the solutions to $\max_x f(x)$, then x^* also solves $\min_x -f(x)$. If x^* is the solutions to $\min_x a + bf(x)$ for $a, b > 0$, then x^* also solves $\min_x f(x)$
- Thus, we solve
$$\tilde{\ell}(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2$$
- But this is just the standard least squares objective. Thus, OLS is an MLE estimator when ϵ is normally distributed

OLS as a MLE estimator

- This is an extremely interesting result for at least three reason
- First, there is the Cramer Rao Theorem that states: Cramer-Rao: Suppose θ^{MLE} is an unbiased estimator of θ. Then for any other unbiased estimator of θ, denoted θ^* we have $Var(\theta^*) \geq Var(\theta^{MLE})$. Thus, the MLE estimate of θ is the **M**innimum **V**ariance **U**nbiased **E**stimator (MVUE) of θ
- Second, if the model is correctly specified, MLE estimates are consistent, asymptotically normal, and invariant under monotone transformations
- Third, if the model is **in**correctly specified, MLE estimates will provide excellent fits of the data *within sample*, but it will do so by giving potentially goofy parameter estimates

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- **OLS as a weighting estimator**
- OLS and the common effects assumption
- Fully saturated OLS models
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

OLS as a weight estimator

- But wait (or should I say "weight"?)! OLS is a weighting estimator, too. To see why, consider the matrix form of the OLS estimator, where y is a $(n \times 1)$ vector, x is a $(n \times k)$ matrix, and β is a $(k \times 1)$ vector

  $$\hat{\beta} = (x'x)^{-1}x'y$$

  so that

  $$\hat{\beta} = wy$$

  where $w = (x'x)^{-1}x'$ where w is a $(k \times n)$ vector
- This means that $\hat{\beta}_i$ is formed by taking a weighted sum of the $y's$ where the weights are determined by the data
- These are four of the five ways of thinking of the OLS estimator. Which should you chose?
- The answer to that question is either "all of the above" or "it depends on the application"

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- **OLS and the common effects assumption**
- Fully saturated OLS models
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

OLS and the common effects assumption

- Consider the canonical regression equation you have seen hundreds of time
$$y_i = x_i'\beta + \delta D_i + \epsilon_i$$
- In this model, we are estimating the parameter δ as a fixed parameter. Treatment in this world raises *all* outcomes by exactly the same amount. Fix $x_i' = x^0$, the expected outcome with treatment is just $x^0\beta + \delta$ and without treatment is just $x^0\beta$
- You cannot complain about OLS as it is doing **exactly** what you told it to do
- Because you told it to estimate a single parameter, OLS will weight the data "appropriately"

OLS and the common effects assumption

- "Appropriately" sounds pretty ominous. How bad can it be? In general, trying to make sense of OLS weights is hard, but here is a special case. Suppose that X takes on a reasonably small number of values so we have $K < n$ "cells" of the data and you estimate the model
$$y_{i,j} = \delta D_{i,j} + \alpha_j + \epsilon_{i,j}$$
where j indexes the "cell" of the data in X, α_j is a fixed effect for each j.
- If we let $\Delta_j = \bar{y}_{1,j} - \bar{y}_{0,j}$ Black, Smith, Noel, and Berger (2003) show that
$$\hat{\delta} = \sum_{j=1}^{K} w_j \hat{\Delta}_j$$
where
$$w_j = \frac{r_j(1-r_j)n_j}{\sum_{i=1}^{K} r_i(1-r_i)n_i} \textit{ where } r_j = Pr(D_j = 1)$$

OLS and the common effects assumption

$$w_j = \frac{r_j(1 - r_j)n_j}{\sum_{i=1}^{K} r_i(1 - r_i)n_i}$$

- These weights are interesting because they tell us what OLS is trying to do. First, the weights provide higher weight the larger the particular data cell, n_j. This is probably what we want OLS to do: larger cells encompass more of the data
- The term $r_j(1 - r_j)$ is more problematic. This expression is maximized when $r_j = 0.5$, but this is seldom the weight we want! Thus, even in this case where we are very flexible about the functions form of our model, OLS does not give us parameter we can interpret unless $\Delta_j = \Delta \ \forall j$
- Can we use OLS to recoup more meaningful estimates? I'm glad you asked

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- *OLS and the common effects assumption*
- **Fully saturated OLS models**
- A flexible estimator of treatment effects
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

Fully saturate OLS models

- Often times we have data where the number of covariates is small relative to the number of observations, such as the Census (or US ACS) or many administrative data sets. As a result, if the X vector is all discrete, we will have many more observations (N) than data cells (K), or $K \ll N$
- In this case we can make OLS a fully nonparametric regression by estimating

$$y_{1,i,j} = \alpha_{1,j} + \epsilon_{1,i,j}$$

$$y_{0,i,j} = \alpha_{0,j} + \epsilon_{0,i,j}$$

- We can now define the treatment effect for $X = x^0$ cell as $\Delta_k = E(Y_{1,i} - Y_{1,0}|X = x^k)$
- In this case, the use of OLS makes no assumptions about functional form so is fully nonparametric
- But we can also handle the case of more parametric regression

Fully saturate OLS models

- Recall our canonical OLS model:

$$y_i = x_i'\beta + \delta D_i + \epsilon_i$$

- We could specify a slight generalization

$$y_{1,i} = x_i'\beta_1 + \epsilon_{1,i}$$

$$y_{0,i} = x_i'\beta_0 + \epsilon_{0,i}$$

- This model allows the treatment effects to vary by X. In particular, the treatment effect is just $\Delta(x^0) = x^0(\beta_1 - \beta_0)$ when $X = x^0$
- Allowing some heterogeneity in the response to treatment is of course an improvement, but both this parametric and the nonparametric give us a whole bunch of estimates. We need to provide convenient ways of summarizing the estimates

- How can we do this?

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- *OLS and the common effects assumption*
- *Fully saturated OLS models*
- **A flexible estimator of treatment effects**
- Yule's remarkable theorem
- A decomposition exploiting Yule's theorem

A flexible estimator of treatment effects

- In both the nonparametric and parametric case we have predicted values for each $X = x^0$, that we will term $\Delta(x^0)$. The key insight is that both models provide a predicted value for both $Y_{1,i}$ and $Y_{0,i}$ each value of x^0
- For those with treatment, you can construct $\hat{\Delta}_i = y_{1,i} - \hat{y}_{0,i}$ and for those without treatment, you can construct $\hat{\Delta}_i = \hat{y}_{1,i} - y_{0,i}$
- Armed with these $\hat{\Delta}_i$ we can calculate the $\hat{\Delta}^{ATE}$ by taking the mean of $\hat{\Delta}_i$ for the whole sample
- We can calculate the $\hat{\Delta}^{ATT}$ by taking the mean of $\hat{\Delta}_i$ for the treated portion of the sample
- We can calculate the $\hat{\Delta}^{ATN}$ by taking the mean of $\hat{\Delta}_i$ for the untreated portion of the sample. For standard errors, I would recommend the bootstrap

A flexible estimator of treatment effects

- For the parametric case, estimates the three standard treatment effects $(\Delta^{ATE}, \Delta^{ATT}, \Delta^{ATN})$ can be pretty simply summarized as

$$\hat{\Delta}^{ATE} = \bar{x}(\hat{\beta}_1 - \hat{\beta}_0)$$

$$\hat{\Delta}^{ATT} = \bar{x}_{D_i=1}(\hat{\beta}_1 - \hat{\beta}_0)$$

$$\hat{\Delta}^{ATN} = \bar{x}_{D_i=0}(\hat{\beta}_1 - \hat{\beta}_0)$$

- Thus, the only difference between any of treatment parameters arises from differences among the means of the sample, the treated sample, and non-treated sample
- The sample means weight the differences $(\hat{\beta}_1 - \hat{\beta}_0)$ according to their relative frequency in the population of interest
- This is clearly a generalization of canonical model, which requires that $\beta_1 = \beta_0$ except for the constant
- This restriction can be easily tested, but almost **never is** tested

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- *OLS and the common effects assumption*
- *Fully saturated OLS models*
- *A flexible estimator of treatment effects*
- **Yule's remarkable theorem**
- A decomposition exploiting Yule's theorem

Yule's remarkable theorem

- As I hope this lecture has shown you, OLS is an extremely complicated estimator. Computers hide this (but, no, we don't want to go back to doing it by hand), but it makes understanding what your doing very hard
- Enter G. Udny Yule. In 1907 (way before any computers), Yule proved that you can treat any multiple regression as a single variable regression. To see how, consider the canonical regression:

$$y_i = x_i\beta + \delta D_i + \epsilon_i$$

- Yule has you first run the regression $D_i = x_i b_D + u_{D,i}$, recover the predicted values, $\hat{D}_i$, and define the residual $\tilde{D}_i = D_i - \hat{D}_i$
- Next run the regression $y_i = x_i b_y + u_{y,i}$, recovered the predicted values, $\hat{y}_i$, and define the residual, $\tilde{y}_i = y_i - \hat{y}_i$
- The residuals $(\tilde{y}_i, \tilde{D}_i)$ are orthogonal to X by construction

Yule's remarkable theorem

- If you run our "big regression"

  $$y_i = x_i\beta + \delta D_i + \epsilon_i$$

  you will get an estimate $\hat{\delta}^{big}$

- Now run the "Yule" regression

  $$\tilde{y}_i = \delta\tilde{D}_i + \epsilon_{y,i}$$

  and you get an estimate $\hat{\delta}^{Yule}$

- Yule's theorem states that $\hat{\delta}^{big} = \hat{\delta}^{Yule}$
- This is exact relationship; the estimates are **identical**
- Economists often call this the Frisch-Waugh theorem (1933), but Yule proved it (1907). In his honor, Black and Smith (2006) refer to $(\tilde{y}_i, \tilde{D}_i)$ as "Yulized residuals"

Yule's remarkable theorem

1. The standard errors are correct subject to a degree of freedom adjustment, or

$$se(\hat{\delta}^{Yule})\left(\frac{n-1}{n-k}\right)^{\frac{1}{2}} = se(\hat{\delta}^{OLS})$$

2. To identify the parameter δ OLS relies on variation that is orthogonal to X
3. There is a rank condition that says the matrix of covariates must be of full rank. This means the R^2 of the regression $D_i = x_i b_D + u_{D,i}$ cannot be 1. $R^2 = 0.98$ is problematic, too
4. Insignificant coefficients should be interpreted with care. The R^2 from the above regression is a useful diagnostic
5. Be careful when you plot data. You only get to use the variation in D that is orthogonal to X to identify δ
6. Unless your measurement error is correlated with X, the impact of measurement error gets amplified when you have a rich covariate set

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- *OLS and the common effects assumption*
- *Fully saturated OLS models*
- *A flexible estimator of treatment effects*
- *Yule's remarkable theorem*
- **A decomposition exploiting Yule's theorem**

A decomposition exploiting Yule's theorem

- We can also use Yule's theorem to get an exact distribution of OLS estimates. Consider the regression

  $$y_i = \delta D_i + \epsilon_i$$

  where we ignored the tilde's to avoid the notational clutter. The OLS estimate is

  $$\hat{\delta} = \frac{s_{y,D}}{s_{D,D}} = \delta + \frac{s_{\epsilon,D}}{s_{D,D}}$$

  where $s_{x,y}$ is the sample covariance, (or variance if $s_{x,x}$)

- Recall that $r_{x,y} = \frac{s_{x,y}}{s_x, s_y}$, we have

  $$\hat{\delta} = \delta + r_{\epsilon,D} \; \frac{s_\epsilon}{s_D}$$

  where s_ϵ is the sample standard deviation of ϵ, s_D is the sample deviation in D, and $r_{\epsilon,D}$ is the sample correlation coefficient of (ϵ, D)

- This formula can explain much about OLS estimates

A decomposition exploiting Yule's theorem

$$\hat{\delta} = \delta + r_{\epsilon,D} \ \frac{s_\epsilon}{s_D}$$

- The error in $\hat{\delta}$ has three parts. The first part is s_ϵ, which says the more unexplained variation in the regression (the larger s_ϵ) the more the error in the estimates

- The second part, s_D, says the more variation variation in D the more precise your estimate

- The third part is just $r_{\epsilon,D}$. While $E(r_{\epsilon,D}) = 0$, in finite samples it could be positive or negative. It is the source of the "sampling" variation of the estimates. And, of course, if we violate the regression assumption $cov(\epsilon, D) = 0$, this is where the problems arise

- What happens if we add variables to X? Hint: What happens to $r_{\epsilon,D}$? What happens to s_ϵ? What happens to s_D?

Overview of lecture

- *OLS as a minimum distance estimator*
- *OLS as a method of moments estimator*
- *OLS as a MLE estimator*
- *OLS as a weighting estimator*
- *OLS and the common effects assumption*
- *Fully saturated OLS model*
- *A flexible estimator of treatment effects*
- *Yule's remarkable theorem*
- *A decomposition exploiting Yule's theorem*