Lecture 06
Finite-Sample Properties of OLS

21 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Goals for today

1. Linear models assumptions
2. OLS Finite sample properties

Linear models assumptions

I. Linearity

We observe a pair of random variables (y, X), which have the following relationship for some random vector ϵ and fixed vector β:

$$y = X\beta + \epsilon$$

I. Linearity

We observe a pair of random variables (y, X), which have the following relationship for some random vector ϵ and fixed vector β:

$$y = X\beta + \epsilon$$

We assume that the following dimensions hold.

$$y \in \mathbb{R}^n$$
$$X \in \mathbb{R}^{n \times p}$$
$$\beta \in \mathbb{R}^p$$
$$\epsilon \in \mathbb{R}^n$$

II. Strict exogeneity

For all X, we have:

$$\mathbb{E}\left(\epsilon|X\right) = 0 \tag{1}$$

II. Strict exogeneity

For all X, we have:

$$\mathbb{E}\left(\epsilon|X\right) = 0 \tag{1}$$

The 'strict' part comes from the conditional on all of X.

II. Strict exogeneity

For all X, we have:

$$\mathbb{E}\left(\epsilon|X\right) = 0 \tag{1}$$

The 'strict' part comes from the conditional on all of X.

Notice that this implies the weaker assumption we used with simple linear models:

$$\begin{aligned}\mathbb{E}\left(\epsilon\right) &= \mathbb{E}\left\{\mathbb{E}\left(\epsilon|X\right)\right\} && (2)\\ &= \mathbb{E}\left\{0\right\} && (3)\\ &= 0 && (4)\end{aligned}$$

III. No multicollinearity

We have:

$$\mathbb{P}\left[\operatorname{rank}(X) = p\right] = 1$$

III. No multicollinearity

We have:

$$\mathbb{P}\left[\text{rank}(X) = p\right] = 1$$

When broken, it is impossible to do inference on β without additional assumptions.

IV. Spherical errors

The variance of the errors is given by:

$$\mathbb{V}\left(\epsilon|X\right) = \sigma^2 \mathbb{I}_n$$

Recall that $\mathbb{V}u = \mathbb{E}(uu^t)$.

IV. Spherical errors

The variance of the errors is given by:

$$\mathbb{V}\left(\epsilon|X\right) = \sigma^2\mathbb{I}_n$$

Recall that $\mathbb{V}u = \mathbb{E}(uu^t)$.

We can break this assumption into two parts; the *homoscedasticity* assumption:

$$\mathbb{E}(\epsilon_i^2) = \sigma^2$$

and *no autocorrelation* assumption:

$$\mathbb{E}(\epsilon_i\epsilon_j) = 0 \quad i \neq j$$

V. Normality

The final, most restrictive assumption, is that the errors follow a multivariate normal distribution:

$$\epsilon | X \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$$

Classical linear model assumptions

I. Linearity $Y = X\beta + \epsilon$

II. Strict exogeneity $\mathbb{E}\left(\epsilon|X\right) = 0$

III. No multicollinearity $\mathbb{P}\left[\text{rank}(X) = p\right] = 1$

IV. Spherical errors $\mathbb{V}\left(\epsilon|X\right) = \sigma^2\mathbb{I}_n$

V. Normality $\epsilon|X \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$

Finite sample properties

Ordinary least squares

We have already derived the ordinary least square estimator:

$$\widehat{\beta} = (X^t X)^{-1} X^t y$$

If we define the following values:

$$
\begin{aligned}
S_{xx} &= \frac{1}{n} X^t X \\
s_{xy} &= \frac{1}{n} X^t y
\end{aligned}
$$

The ordinary least squares estimator can also be written:

$$\widehat{\beta} = S_{xx}^{-1} s_{xy}$$

A form that will be useful for large sample theory.

Special matricies

Last time we defined the following matricies:

$$P = X(X^t X)^{-1} X^t$$
$$M = \mathbb{I}_n - P$$

Special matricies

Last time we defined the following matricies:

$$
\begin{aligned}
P &= X(X^t X)^{-1} X^t \\
M &= \mathbb{I}_n - P
\end{aligned}
$$

Today we have one more matrix A that does not have a direct geometric interpretation but is nonetheless very useful:

$$
\begin{aligned}
A &= (X^t X)^{-1} X^t \\
Ay &= \widehat{\beta}
\end{aligned}
$$

Last time we showed that:

$$
\begin{aligned}
P^2 &= P^t = P \\
M^2 &= M^t = M \\
PX &= X \\
MX &= 0 \\
Py &= X\beta \\
My &= M\epsilon = r
\end{aligned}
$$

The matrix A is not square, but the outer product has a nice property:

$$\begin{aligned} AA^t &= (X^tX)^{-1}X^tX(X^tX)^{-1} \\ &= (X^tX)^{-1} \end{aligned}$$

Two final definitions

The residuals and estimate of the σ^2 parameter are given as:

$$r = y - X\widehat{\beta}$$
$$s^2 = \frac{1}{n-p} r^t r$$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^t X)^{-1}$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$

(C) $\widehat{\beta}$ is the best linear unbiased estimator (Gauss-Markov)

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$

(C) $\widehat{\beta}$ is the best linear unbiased estimator (Gauss-Markov)

(D) $Cov(\widehat{\beta}, r|X) = 0$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$

(C) $\widehat{\beta}$ is the best linear unbiased estimator (Gauss-Markov)

(D) $Cov(\widehat{\beta}, r|X) = 0$

(E) $\mathbb{E}s^2 = \sigma^2$

Finite sample properties

Under assumptions I-III:

(A) $\mathbb{E}(\widehat{\beta}|X) = \beta$

Under assumptions I-IV:

(B) $\mathbb{V}(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$

(C) $\widehat{\beta}$ is the best linear unbiased estimator (Gauss-Markov)

(D) $Cov(\widehat{\beta}, r|X) = 0$

(E) $\mathbb{E}s^2 = \sigma^2$

Under assumptions I-V:

(F) $\widehat{\beta}$ achieves the Cramér–Rao lower bound

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\widehat{\beta} - \beta \quad = \quad (X^t X)^{-1} X^t y - \beta$$

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^t X)^{-1} X^t y - \beta \\
&= (X^t X)^{-1} X^t (X\beta + \epsilon) - \beta
\end{aligned}$$

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^t X)^{-1} X^t y - \beta \\
&= (X^t X)^{-1} X^t (X\beta + \epsilon) - \beta \\
&= (X^t X)^{-1} X^t X\beta + (X^t X)^{-1} X^t \epsilon - \beta
\end{aligned}$$

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^tX)^{-1}X^ty - \beta \\
&= (X^tX)^{-1}X^t(X\beta + \epsilon) - \beta \\
&= (X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= \beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= A\epsilon
\end{aligned}$$

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^tX)^{-1}X^ty - \beta \\
&= (X^tX)^{-1}X^t(X\beta + \epsilon) - \beta \\
&= (X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= \beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= A\epsilon
\end{aligned}$$

From here, we can derive the unbiased result easily:

$$\begin{aligned}
\mathbb{E}(\widehat{\beta} - \beta | X) &= \mathbb{E}(A\epsilon | X) \\
&= A \cdot \mathbb{E}(\epsilon | X) \\
&= 0
\end{aligned}$$

(A) Unbiased regression estimate $\widehat{\beta}$

Notice that the error in our estimate can be re-written in terms of the matrix A:

$$\begin{aligned}
\widehat{\beta} - \beta &= (X^tX)^{-1}X^ty - \beta \\
&= (X^tX)^{-1}X^t(X\beta + \epsilon) - \beta \\
&= (X^tX)^{-1}X^tX\beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= \beta + (X^tX)^{-1}X^t\epsilon - \beta \\
&= A\epsilon
\end{aligned}$$

From here, we can derive the unbiased result easily:

$$\begin{aligned}
\mathbb{E}(\widehat{\beta} - \beta|X) &= \mathbb{E}(A\epsilon|X) \\
&= A \cdot \mathbb{E}(\epsilon|X) \\
&= 0
\end{aligned}$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$\mathbb{V}(\widehat{\beta}|X) \quad = \quad \mathbb{V}(\widehat{\beta} - \beta|X)$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$
\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X)
\end{aligned}
$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X) \\
&= A\mathbb{V}(\epsilon|X)A^t
\end{aligned}$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X) \\
&= A\mathbb{V}(\epsilon|X)A^t \\
&= A\mathbb{E}(\epsilon\epsilon^t|X)A^t
\end{aligned}$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X) \\
&= A\mathbb{V}(\epsilon|X)A^t \\
&= A\mathbb{E}(\epsilon\epsilon^t|X)A^t \\
&= A(\sigma^2\mathbb{I}_n)A^t
\end{aligned}$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$
\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X) \\
&= A\mathbb{V}(\epsilon|X)A^t \\
&= A\mathbb{E}(\epsilon\epsilon^t|X)A^t \\
&= A(\sigma^2\mathbb{I}_n)A^t \\
&= \sigma^2 AA^t
\end{aligned}
$$

(B) Form of the variance

The formula for the variance of the ordinary least squares estimator can be derived from our assumptions and prior results.

$$
\begin{aligned}
\mathbb{V}(\widehat{\beta}|X) &= \mathbb{V}(\widehat{\beta} - \beta|X) \\
&= \mathbb{V}(A\epsilon|X) \\
&= A\mathbb{V}(\epsilon|X)A^t \\
&= A\mathbb{E}(\epsilon\epsilon^t|X)A^t \\
&= A(\sigma^2\mathbb{I}_n)A^t \\
&= \sigma^2 AA^t \\
&= \sigma^2(X^tX)^{-1}
\end{aligned}
$$

(E) Unbiased s^2

We have already established that $r^t r = \epsilon^t M \epsilon$, so all we need to do is show that $\mathbb{E}(\epsilon^t M \epsilon | X) = \sigma^2 (n - p)$.

(E) Unbiased s^2

We have already established that $r^t r = \epsilon^t M\epsilon$, so all we need to do is show that $\mathbb{E}(\epsilon^t M\epsilon|X) = \sigma^2(n-p)$.

We can write this expected value in terms of the trace of M:

$$\mathbb{E}(\epsilon^t M\epsilon|X) \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j}\mathbb{E}(\epsilon_i\epsilon_j|X)$$

(E) Unbiased s^2

We have already established that $r^t r = \epsilon^t M \epsilon$, so all we need to do is show that $\mathbb{E}(\epsilon^t M \epsilon | X) = \sigma^2 (n - p)$.

We can write this expected value in terms of the trace of M:

$$
\begin{aligned}
\mathbb{E}(\epsilon^t M \epsilon | X) &= \sum_{i=1}^{n} \sum_{j=1}^{n} m_{i,j} \mathbb{E}(\epsilon_i \epsilon_j | X) \\
&= \sum_{i=1}^{n} m_{i,i} \sigma^2
\end{aligned}
$$

(E) Unbiased s^2

We have already established that $r^t r = \epsilon^t M\epsilon$, so all we need to do is show that $\mathbb{E}(\epsilon^t M\epsilon|X) = \sigma^2(n-p)$.

We can write this expected value in terms of the trace of M:

$$\begin{aligned}
\mathbb{E}(\epsilon^t M\epsilon|X) &= \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j}\mathbb{E}(\epsilon_i\epsilon_j|X) \\
&= \sum_{i=1}^{n} m_{i,i}\sigma^2 \\
&= \sigma^2\sum_{i=1}^{n} m_{i,i}
\end{aligned}$$

(E) Unbiased s^2

We have already established that $r^t r = \epsilon^t M \epsilon$, so all we need to do is show that $\mathbb{E}(\epsilon^t M \epsilon | X) = \sigma^2 (n - p)$.

We can write this expected value in terms of the trace of M:

$$\begin{aligned}
\mathbb{E}(\epsilon^t M \epsilon | X) &= \sum_{i=1}^{n} \sum_{j=1}^{n} m_{i,j} \mathbb{E}(\epsilon_i \epsilon_j | X) \\
&= \sum_{i=1}^{n} m_{i,i} \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} m_{i,i} \\
&= \sigma^2 tr(M)
\end{aligned}$$

Now we simply need to calculate the trace of M:

$$tr(M) = tr(\mathbb{I}_n - P)$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \end{aligned}$$

Now we simply need to calculate the trace of M:

$$
\begin{aligned}
tr(M) &= tr(\mathbb{I}_n - P) \\
&= tr(\mathbb{I}_n) - tr(P) \\
&= n - tr(P)
\end{aligned}
$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \\ &= n - tr(P) \end{aligned}$$

And then,

$$tr(P) = tr(X(X^tX)^{-1}X^t)$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \\ &= n - tr(P) \end{aligned}$$

And then,

$$\begin{aligned} tr(P) &= tr(X(X^tX)^{-1}X^t) \\ &= tr((X^tX)^{-1}X^tX) \end{aligned}$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \\ &= n - tr(P) \end{aligned}$$

And then,

$$\begin{aligned} tr(P) &= tr(X(X^t X)^{-1} X^t) \\ &= tr((X^t X)^{-1} X^t X) \\ &= tr(\mathbb{I}_p) \end{aligned}$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \\ &= n - tr(P) \end{aligned}$$

And then,

$$\begin{aligned} tr(P) &= tr(X(X^tX)^{-1}X^t) \\ &= tr((X^tX)^{-1}X^tX) \\ &= tr(\mathbb{I}_p) \\ &= p \end{aligned}$$

Now we simply need to calculate the trace of M:

$$\begin{aligned} tr(M) &= tr(\mathbb{I}_n - P) \\ &= tr(\mathbb{I}_n) - tr(P) \\ &= n - tr(P) \end{aligned}$$

And then,

$$\begin{aligned} tr(P) &= tr(X(X^tX)^{-1}X^t) \\ &= tr((X^tX)^{-1}X^tX) \\ &= tr(\mathbb{I}_p) \\ &= p \end{aligned}$$

Plugging back into the original yields the result.