POLS 503: Advanced Quantitative Political Methodology: The Notes

Jeffrey B. Arnold

2016-04-19

Contents

Chapter 1

Introduction

hello, world!

Chapter 2

Linear Regression and the Ordinary Least Squares (OLS) Estimator

Since we will largely be concerned with using linear regression for inference, we will start by discussion the population parameter of interest (population linear regression function), then the sample statistic (sample linear regression function) and estimator (ordinary least squares).

We will then consider the properties of the OLS estimator.

2.1 Linear Regression Function

The **population linear regression function** is

$$r(x) = \mathrm{E}[Y|X = x] = \beta_0 + \sum_{k=1}^{K} \beta_k x_k.$$

The population linear regression function is defined for random variables, and will be the object to be estimated.

Names for $\boldsymbol{y}$

- dependent variable
- explained variable
- response variable
- predicted variable
- regressand
- outcome variable

Names for $\boldsymbol{X}$,

- indpendent variables
- explanatory varaibles
- treatment and control variables
- predictor variables
- covariates
- regressors

To estimate the unkonwn population linear regression, we will use the **sample linear regression function**,

$$\hat{r}(x_i) = \hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_k.$$

However, we

$\hat{Y}_i$ are the fitted or predicted value The **residuals** or **errors** are the prediction errors of the estimates

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

$\boldsymbol{\beta}$ are the parameters; β_0 is called the *intercept*, and $\beta_1, \dots, \beta_K$ are called the *slope parameters*, or *coefficients*.

The linear regression function can be written as a scalar function for each observation, $i = 1, \dots, N$,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{K,i} + \varepsilon_i \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{k,i} + \varepsilon_i \\ &= \sum_{k=0}^{K} \beta_k x_{k,i} + \varepsilon_i \end{aligned}$$

where $x_{0,i} = 1$ for all $i \in 1 : N$.

The linear regression can be more compactly written in matrix form,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{K,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,N} & x_{2,n} & \cdots & x_{K,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} .$$

More compactly, the linear regression model can be written as,

$$\underbrace{\boldsymbol{y}}_{N \times 1} = \underbrace{\boldsymbol{X}}_{N \times K} \underbrace{\boldsymbol{\beta}}_{K \times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N \times 1}.$$

The matrix $\boldsymbol{X}$ is called the *design* matrix. Its rows are each observation in the data. Its columns are the intercept, a column vector of 1's, and the values of each predictor.

2.2 Ordinary Least Squares

Ordinary least squares (OLS) is an estimator of the slope and statistic of the regression line[1]. OLS finds values of the intercept and slope coefficients by minimizing the squared errors,

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K = \mathop{\arg\min}_{b_0, b_1, \dots, b_k} \sum_{i=1}^{N} \underbrace{\left(y_i - b_0 - \sum_{k=1}^{K} b_k x_{i,k} \right)^2}_{\text{squared error}},$$

[1] Ordinary least squares is distinguished from *generalized least squares* (GLS).

or, in matrix notation,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} (y_i - \boldsymbol{b}'\boldsymbol{x}_i)^2 \\ &= \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} u_i^2 \\ &= \arg\min_{\boldsymbol{b}} \boldsymbol{u}'\boldsymbol{u}\end{aligned}$$

where $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

In most statistical models, including even genalized linear models such as logit, the solution to this minimization problem would be solved with optimization methods that require interation. One nice feature of OLS is that there is a closed form solution for $\hat{\beta}$ even in the multiple regression case, so no iterative optimization methods need to be used.

In the bivariate regression case, the OLS estimators for β_0 and β_1 are

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 7 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ &= \frac{\mathrm{Cov}(\boldsymbol{x}\boldsymbol{y})}{\mathrm{Var}\,\boldsymbol{x}} &= \frac{\text{Sample covariance betweeen } \boldsymbol{x} \text{ and } \boldsymbol{y}}{\text{Sample variance of } \boldsymbol{x}}.\end{aligned}$$

In the multiple regression case, the OLS estimator for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

The term $\boldsymbol{X}'\boldsymbol{X}$ is similar to the variance of $\boldsymbol{x}$ in the bivariate case. The term $\boldsymbol{X}'\boldsymbol{y}$ is similar to the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ in the bivariate case.

The sample linear regression function estimated by OLS has the following properties:

1. Residuals sum to zero,

$$\sum_{i=1}^{N} \hat{\epsilon}_i = 0.$$

   This implies that the mean of residuals is also 0.
2. The regression function passes through the point $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}_K})$. In other words, the following is always true,

$$\bar{\boldsymbol{y}} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k \bar{\boldsymbol{x}}_k.$$

3. The resisuals are uncorrelated with the predictor

$$\sum_{i=1}^{N} x_i \hat{\epsilon}_i = 0$$

4. The residuals are uncorrelated with the fitted values

$$\sum_{i=1}^{N} \hat{y}_i \hat{\varepsilon}_i = 0$$

2.3 Properties of the OLS Estimator

2.3.1 What makes an estimator good?

Estimators are evaluated not on how close an estimate in a given sample is to the population, but how their sampling distributions compare to the population. In other words, judge the *methodology* (estimator), not the *result* (estimate).[^ols-properties-references]

Let θ be the population parameter, and $\hat{\theta}$ be an estimator of that population parameter.

Bias The bias of an estimator is the difference between the mean of its sampling distribution and the population parameter,

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

Variance The variance of the estimator is the variance of its sampling distribution, $\mathrm{Var}(\theta)$.

Efficiency (Mean squared error) An efficient estimator is one that minimizes a given “loss function”, which is a penalty for missing the population average. The most common loss function is squared loss, which gives the *Mean Squared Error (MSE)* of an estimator.

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\left[{(\hat{\theta} - \theta)}^2\right] = (\mathrm{E}(\hat{\theta}) - \theta)^2 + \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$$

The mean squared error is a function of both the bias and variance of an estimator.

This means that some biased estimators can be more efficient : than unbiased estimators if their variance offsets their bias.[2]

\begin{table}

\caption{Examples of clocks as “estimators” of the time[3]}

	Biased	Variance
Stopped clock	Yes	High
Random clock	No	High
Clock that is ”a lot ” fast	Yes	Low
Clock that is ”a little” fast	Yes	Low
Atomic clock	No	Low

\end{table}

Another property is **consistency**. Consistency is an asymptotic property[4], that roughly states that an estimator converges to the truth as the number of obserservations grows, $\mathrm{E}(\hat{\theta} - \theta) \to 0$ as $N \to \infty$. Roughly, this means that if you had enough (infinite) data, the estimator will give you the true value of the parameter.

2.3.2 Properties of OLS

- When is OLS unbiased?
- When is OLS consistent?
- When is OLS efficient?

[2] It follows from the definition of MSE, that biased estimator, $\hat{\theta}_B$, has a lower MSE than an unbiased estimator, $\hat{\theta}_U$, if $\mathrm{Bias}(\theta_B)^2 < \mathrm{Var}(\theta_U) - \mathrm{Var}(\theta_B)$.

[3] Example from Chris Adolph

[4] As the number of observations goes to infinity.

1. **Linearity** The popluation model is
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
where $\boldsymbol{\varepsilon}$ is an unobserved random error or disturbance term with $\mathrm{E}(\varepsilon) = 0$.
2. Random/iid sample. $(y_i, \boldsymbol{x}_i')$ are a random sample from the population.
3. **No Perfect Collinearity**. There is no exact *linear* relationships among the independent variables. $\boldsymbol{X}$ is a $N \times K$ matrix with rank K.
4. Zero conditional mean. The error, ε, has an expected value of zero, conditional on the predictors.
$$\mathrm{E}(\varepsilon | X) = 0$$
5. Constant variance (Homoskedasticity). The error has the same variance conditional on the predictors, for all observations,
$$\mathrm{E}(\epsilon_i | \boldsymbol{x}_i) = \sigma^2) \text{ for all } i$$
6. Fixed $\boldsymbol{X}$ or $\boldsymbol{X}$ measured without error and independent of the error.
7. Normal disturbances: $\epsilon_i | \boldsymbol{X} \sim N(0, \sigma^2)$.

What do these assumptions give us?

- Identification of OLS: Under Assumption 1, OLS can be estimated. In other words, there is a *unique* $\hat{\beta}$ that minimizes the sum of squared errors.
- Unbiasedness of OLS: Under Assumptions 1–4, OLS is unbiased.
$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$
- Gauss-Markov theorem. Under Assumptions 1–5, OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$. *Linear* means that the estimates can be written as a linear functions of the outcomes,
$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{i,j} y_i$$
Best means that it has the smallest variance. This means for any unbiased and linear estimator, $\tilde{\beta}$, the OLS estimator, $\hat{\beta}_{OLS}$, has a smaller variance,
$$\mathrm{Var}(\tilde{\beta}) > \mathrm{Var}(\hat{\beta}_{OLS})$$
Not that this does not imply that OLS has the lowest MSE of any estimator, since a biased estimator could have a lower MSE. In fact, for any regression with three or more variables, there is a ridge estimator with a lower MSE.

Assumption	Formal statement	Consequence of violation
No (perfect) collinearity	$\mathrm{rank}(\boldsymbol{X}) = K, K < N$	Coefficients unidentified
$\boldsymbol{X}$ is exogenous	$\mathrm{E}(\boldsymbol{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
Disturbances have mean 0	$\mathrm{E}(\varepsilon) = 0$	Biased, even as $N \to \infty$
No serial correlation	$\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$	Unbiased, wrong se
Homoskedastic errors	$\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$	Unbiased, wrong se
Gaussian errors	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	Unbiased, se wrong unless $N \to \infty$

Note that these assumptions can be sometimes be written in largely equivalent, but slightly different forms.

2.4 References

- Wooldrige, Ch 3.
- Fox, Ch 6, 9.