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 $\beta_1, \ldots, \beta_K$ are called the *slope parameters*, or *coefficients*.

We will then consider the properties of the OLS estimator.

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,

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

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 generalized linear models such as logit, the solution to this minimization problem would be solved with optimization methods that require interation. One nice feature

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 7 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$= \frac{\operatorname{Cov}(\boldsymbol{xy})}{\operatorname{Var} \boldsymbol{x}} \quad = \frac{\text{Sample covariance betweeen } \boldsymbol{x} \text{ and } \boldsymbol{y}}{\text{Sample variance of } \boldsymbol{x}}.$$

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 $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}_K})$. 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
$$\sum_{i=1}^{N} \hat{y}_i \hat{\epsilon}_i = 0$$

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

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

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.

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

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]

Consistency is an asymptotic property[3], that roughly states that an estimator converges to the truth as the number of observations grows, $\text{E}(\hat{\theta} - \theta) \to 0$ 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?

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

[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 $\text{Bias}(\theta_B)^2 < \text{Var}(\theta_U) - \text{Var}(\theta_B)$.

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

Chapter 3

OLS Troubleshooting and Diagnostics

3.1 Multi-Collinearity

3.2 Omitted Variable Bias

3.3 Measurement Error

3.4 Non-linearity

3.5 Heteroskedasticity and Auto-correlation

Note, that OLS assumes that the variance of the the disturbances is constant $\hat{Y} - Y = \varepsilon = \sigma^2$. What happens if it isn't?

$$\mathbf{\Sigma} = \begin{bmatrix} \text{Var}(\varepsilon_1) & \text{Cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{Cov}(\varepsilon_1, \varepsilon_N) \\ \text{Var}(\varepsilon_2, \varepsilon_1) & \text{Var}(\varepsilon_2) & \cdots & \text{Cov}(\varepsilon_2, \varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\varepsilon_N, \varepsilon_1) & \text{Cov}(\varepsilon_N, \varepsilon_2) & \cdots & \text{Cov}(\varepsilon_N) \end{bmatrix} \Sigma = \begin{bmatrix} \text{E}(\varepsilon_1^2) & \text{E}(\varepsilon_1\varepsilon_2) & \cdots & \text{E}(\varepsilon_1\varepsilon_N) \\ \text{E}(\varepsilon_2\varepsilon_1) & \text{E}(\varepsilon_2^2) & \cdots & \text{E}(\varepsilon_2\varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}(\varepsilon_N\varepsilon_1) & \text{E}(\varepsilon_N\varepsilon_2) & \cdots & \text{E}(\varepsilon_N^2) \end{bmatrix}$$

The matrix can be written more compactly as,

$$\mathbf{\Sigma} = \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$$

An assumption is that errors are independent, $\text{E}(\epsilon_i\epsilon_j)$ for all $i \neq j$. This means that all off-diagonal elements of $\mathbf{\Sigma}$ are 0$. Additionally, all ϵ_i are assumed to have the same variance, σ^2. Thus, the variance-covariance matrix of the errors is a assumed to have a diagonal matrix with the form,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_N$$

If these assumptions of the errors do not hold, then Σ does not take this form, and more complicated models than OLS need to be used to get correct standard errors.

3.6 Non-constant variance (Heteroskedasticity)

The homoskedastic case assumes that each eror term has its own variance. In the heteroskedastic case, each distrurbance may have its own variance, but they are still uncorrelated ($\mathbf{\Sigma}$ is diagonal)

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}$$

The problem is that now there are N variance parameters to estimate, in addition to the K slope coefficients. Now, there are more parameters than we can estimate. With heteroskedasticity, OLS with be unbiased, but the standard errors will be incorrect.

More general case allows for heteroskedasticity, and autocorrelation ($\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$),

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,N} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N,1} & \sigma_{N,2} & \cdots & \sigma_N^2 \end{bmatrix}$$

As with heteroskedasticity, OLS with be unbiased, but the standard errors will be incorrect.

Chapter 4

Appendix

4.1 Multivariate Normal Distribution

The multivariate normal distribution is the generalization of the univariate normal distribution to more than one dimension.[1] The random variable, $\boldsymbol{x}$, is a length k vector. The k length vector $\boldsymbol{\mu}$ are the means of $\boldsymbol{x}$, and the $k \times k$ matrix, $\boldsymbol{\Sigma}$, is the variance-covariance matrix,

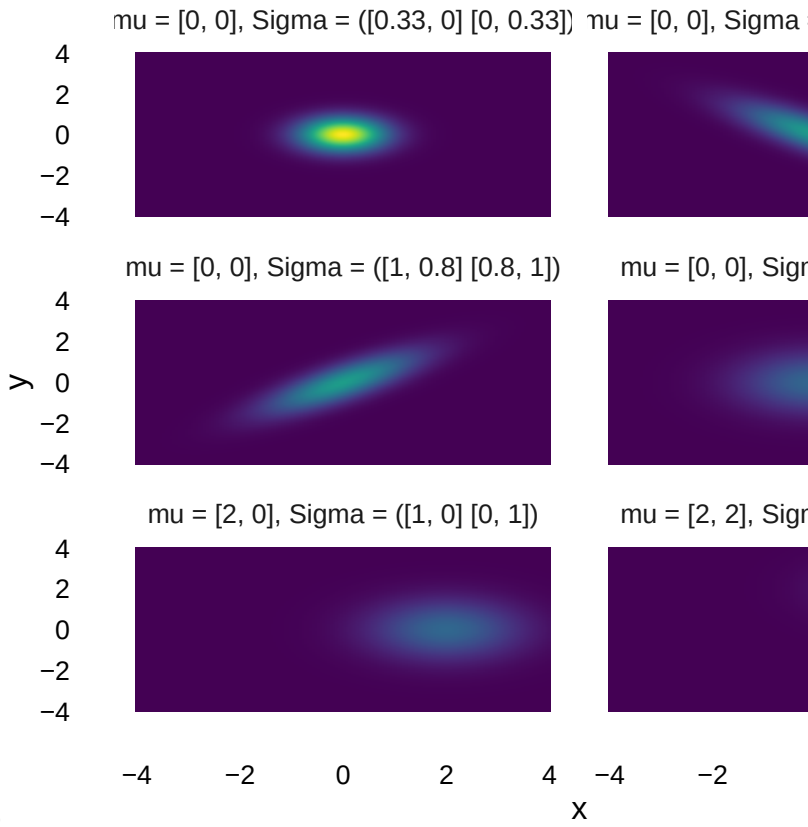
$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} &\sim \mathcal{N}_k \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k,1} & \sigma_{k,2} & \cdots & \sigma_{k,k} \end{bmatrix} \right) \end{aligned}$$

The density function of the multivariate normal is,

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2k)^{-\frac{k}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

You can sample from and calculate the density for the multivariate normal distribution with the functions `dmvnorm` and `rmvnorm` from the package **mvtnorm**.

[1] See Multivariate normal distribution and references therein.



Density plots of different bivariate normal distributions,