Assumptions for OLS Regression and the Gauss-Markov Theorem

2021-09-06

One reason that OLS estimation is so useful is that, under a certain set of assumptions underlying the classical linear regression model, the estimators, $b_0, b_1, b_2, \ldots, b_k$, have several desirable statistical properties. These properties include:

- The least squares estimators are *linear estimators*; they are linear functions of the observations. (This property helps us derive the sampling distributions for each estimator, which allows for statistical inference.)
- The least squares estimators are *unbiased estimators* of the population coefficients.
- The least squares estimators have sampling variances and a covariance.
- Of all the linear, unbiased estimators, the least squares estimators have the smallest sampling variance (most precise/efficient).

Assumptions of the OLS Regression Model

As mentioned, there are a certain set of assumptions underlying the linear regression model (both the simple and multiple regression models) for these properties to be true. These assumptions are:

- **A.1:** The model is correctly specified.
- **A.2:** The design matrix, **X**, is of full rank.
- **A.3:** The population errors given **X** have a mean of zero.
- **A.4:** The population errors given **X** are homoscedastic.
- **A.5:** The population errors given **X** are independent.
- **A.6:** The predictor values are fixed with finite, non-zero variance.

If these six assumptions are satisfied, then the estimators will have the properties we referred to previously. This is sometimes referred to as the *weak classical regression model.*

Another assumption that is useful is:

- **A.7:** The population errors given **X** are normally distributed

If all seven assumptions are met, we refer to this as the *strong classical regression model.* When this assumption is met (in addition to the six other assumptions), the sampling distribution for the least squares estimators are also normally distributed; they are approximately normal under other conditions, especially with large sample sizes. This is useful for carrying out statistical inference. Furthermore, under the full set of seven assumptions, the least squares estimators are the maximum-likelihood estimators of the population coefficients.

We will now examine each of the assumptions underlying the linear regression model

A.1: The model is correctly specified.

When we posit or fit the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, we are assuming that there is a linear relationship between the predictor(s) and the outcome. Furthermore, we are stating that this model is correctly specified using the set of predictors included and that deviation from this model in the observed data is all due to random sampling error.

A.2: The design matrix, X, is of full rank.

This assumption indicates that there is no perfect multicollinearity in the predictor space. That is, the rows (or columns) of X are linearly independent. This is what allows us to compute $(\mathbf{X}^\top\mathbf{X})^{-1}$.

A.3: The population errors given X have a mean of zero.

This assumption states that the mean error (in the population) at a given *x*-value is zero. Using our rules of expectation:

$$
\begin{aligned}
\mathbb{E}(\epsilon|X) &= \mathbb{E}\begin{bmatrix} \epsilon_1|X \\ \epsilon_2|X \\ \epsilon_3|X \\ \vdots \\ \epsilon_n|X \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}(\epsilon_1|X) \\ \mathbb{E}(\epsilon_2|X) \\ \mathbb{E}(\epsilon_3|X) \\ \vdots \\ \mathbb{E}(\epsilon_n|X) \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}
\end{aligned}
$$

This assumption implies that $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$.

A.4: The population errors given X are homoscedastic.

This assumption indicates that residuals at a given value of X have equal variances (homoskedasticity). To show this, we make use of our rules of expectations and the fact that the variance-covariance matrix of the errors at a given X (denoted as $\Sigma(\epsilon|\mathbf{X})$) can be defined as the expected value of $\epsilon^\top\epsilon$.

$$\begin{aligned}
\Sigma(\epsilon|\mathbf{X}) &= \mathbb{E}(\epsilon\epsilon^\top|X) \\
&= \mathbb{E}\left(\begin{bmatrix} \epsilon_1|X \\ \epsilon_2|X \\ \epsilon_3|X \\ \vdots \\ \epsilon_n|X \end{bmatrix} \begin{bmatrix} \epsilon_1|X & \epsilon_2|X & \epsilon_3|X & ... & \epsilon_n|X \end{bmatrix} \right) \\
&= \mathbb{E} \begin{bmatrix} \epsilon_1^2|X & \epsilon_1\epsilon_2|X & \epsilon_1\epsilon_3|X & ... & \epsilon_1\epsilon_n|X \\ \epsilon_2\epsilon_1|X & \epsilon_2^2|X & \epsilon_2\epsilon_3|X & ... & \epsilon_2\epsilon_n|X \\ \epsilon_3\epsilon_1|X & \epsilon_3\epsilon_2|X & \epsilon_3^2|X & ... & \epsilon_3\epsilon_n|X \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1|X & \epsilon_n\epsilon_2|X & \epsilon_n\epsilon_3|X & ... & \epsilon_n^2|X \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}(\epsilon_1^2|X) & \mathbb{E}(\epsilon_1\epsilon_2|X) & \mathbb{E}(\epsilon_1\epsilon_3|X) & ... & \mathbb{E}(\epsilon_1\epsilon_n|X) \\ \mathbb{E}(\epsilon_2\epsilon_1|X) & \mathbb{E}(\epsilon_2^2|X) & \mathbb{E}(\epsilon_2\epsilon_3|X) & ... & \mathbb{E}(\epsilon_2\epsilon_n|X) \\ \mathbb{E}(\epsilon_3\epsilon_1|X) & \mathbb{E}(\epsilon_3\epsilon_2|X) & \mathbb{E}(\epsilon_3^2|X) & ... & \mathbb{E}(\epsilon_3\epsilon_n|X) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(\epsilon_n\epsilon_1|X) & \mathbb{E}(\epsilon_n\epsilon_2|X) & \mathbb{E}(\epsilon_n\epsilon_3|X) & ... & \mathbb{E}(\epsilon_n^2|X) \end{bmatrix}
\end{aligned}$$

The elements along the main diagonal are the error variances. For example $\mathbb{E}(\epsilon_i^2|X)$ is the variance of the *i*th residual. To show that this is the case we use the rules of expectations:

$$\begin{aligned}
\text{Var}(\epsilon_i|\mathbf{X}) &= \mathbb{E}(\epsilon_i^2|\mathbf{X}) - \left[\mathbb{E}(\epsilon_i|\mathbf{X})\right]^2 \\
&= \mathbb{E}(\epsilon_i^2|\mathbf{X}) - 0 \\
&= \mathbb{E}(\epsilon_i^2|\mathbf{X})
\end{aligned}$$

The homoskedasticity assumption makes each variance in the matrix equal, but unknown. Because it the value of the variance is unknow, we can denote it as such usiung the placeholder σ^2—e.g., $\mathbb{E}(\epsilon_i^2|X) = \sigma^2$. Using this to re-write our variance-covariance matrix:

$$\Sigma(\epsilon|\mathbf{X}) = \begin{bmatrix} \sigma^2 & \mathbb{E}(\epsilon_1\epsilon_2|X) & \mathbb{E}(\epsilon_1\epsilon_3|X) & ... & \mathbb{E}(\epsilon_1\epsilon_n|X) \\ \mathbb{E}(\epsilon_2\epsilon_1|X) & \sigma^2 & \mathbb{E}(\epsilon_2\epsilon_3|X) & ... & \mathbb{E}(\epsilon_2\epsilon_n|X) \\ \mathbb{E}(\epsilon_3\epsilon_1|X) & \mathbb{E}(\epsilon_3\epsilon_2|X) & \sigma^2 & ... & \mathbb{E}(\epsilon_3\epsilon_n|X) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(\epsilon_n\epsilon_1|X) & \mathbb{E}(\epsilon_n\epsilon_2|X) & \mathbb{E}(\epsilon_n\epsilon_3|X) & ... & \sigma^2 \end{bmatrix}$$

A.5: The population errors given X are independent.

The off-diagonal elements in the variance-covariance matrix of the error are the covariances between the errors. For example, $\mathbb{E}(\epsilon_i\epsilon_j|X)$ is the covariance between the *i*th and *j*th errors. We can show this using rules of expectations:

$$
\begin{aligned}
\text{Cov}(\epsilon_i, \epsilon_j) &= \mathbb{E}\Bigg[\bigg(\epsilon_i - \mathbb{E}[\epsilon_i]\bigg)\bigg(\epsilon_j - \mathbb{E}[\epsilon_j]\bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\bigg(\epsilon_i - 0\bigg)\bigg(\epsilon_j - 0\bigg)\Bigg] \\
&= \mathbb{E}(\epsilon_i\epsilon_j)
\end{aligned}
$$

The assumption of independence indicates that each of these covariances is equal to zero. Using this result and the result from the homoscedasticity assumption, we can re-write the variance-covariance matrix of the errors as:

$$
\begin{aligned}
\boldsymbol{\Sigma}(\epsilon|\mathbf{X}) &= \begin{bmatrix}\sigma^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma^2\end{bmatrix} \\
&= \sigma^2 \begin{bmatrix}1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1\end{bmatrix} \\
&= \sigma^2\mathbf{I}
\end{aligned}
$$

> The assumptions of homoskedasticity and independence can be more compactly expressed as:
>
> $$\boldsymbol{\Sigma}(\epsilon|\mathbf{X}) = \sigma^2\mathbf{I} =$$

A.6: The predictor values are fixed with finite, non-zero variance.

Fixing the predictor values implies that the values in the **X** matrix are the same under repeated sampling from the same population. **In most research in the social sciences, this is not the case.** In those cases, we instead assume that the values in X are measured without error and that they are independent (uncorrelated) with the errors; $\text{Cov}(\mathbf{X}, \epsilon) = 0$.

Gauss-Markov Theorem

The Gauss–Markov Theorem is a powerful theorem that states that under the weak classical model (**A.1—A.6**), the least squares estimators have certain desirable properties. Both estimators are: - Linear functions of the observations, y_i; - Unbiased estimators of the population coefficients; - The most efficient (smallest sampling variance) unbiased linear estimators of the population coefficients.

> Because of this theorem, we typically refer to the OLS coefficents as BLUE (Best Linear Unbiased Estimators).

Fox (2016) reminds us that the "best" in BLUE means that they have the smallest sampling variance of all the possible linear unbiased estimators. There may be a biased or non-linear estimator that produces a smaller sampling variance than the OLS estimator. It is also worth noting that if we also invoke the normality assumption (A.7), then the OLS estimators become "best" among all unbiased estimators (both linear and non-linear).

Proving this theorem is beyond the scope of the class, but an outline for this proof would entail: - Show that b_0 and b_1 are linear functions of the observations; we can express each estimator as $\sum w_i y_i$ for some w_i. - Show that b_0 and b_1 are unbiased; that $\mathbb{E}(b_0) = \beta_0$ and $\mathbb{E}(b_1) = \beta_1$ - Show that for any other unbiased linear estimator, say L_0 and L_1, that $\text{Var}(b_0) < \text{Var}(L_0)$ and $\text{Var}(b_1) < \text{Var}(L_1)$.

A.7: The population errors given X are normally distributed.

A final assumption that we make about normality is not required to prove the Gauss-Markov Theorem, but it is used to carry out hypothesis testing. This assumption states that the distribution of errors (in the population) at each *x*-value is normally distributed. If we combine this with the property that the mean error given **X** is zero, and the homoskedasticity assumption, then

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

This encapsulates the assumptions we check about the probability distribution of the model errors.

OLS Estimators are Maximum Likelihood Estimators

Given the assumptions of the strong classical model (**A.1–A.7**), we can show that the least squares estimators are also the maximum likelihood estimators. Recall that the likelihood is the probability of a set of parameters, computed as the joint density of the data, given a set of observations under a particular probability distribution.

The joint density of the errors is:

$$\prod_{i=1}^{n} f(\epsilon_i;\ 0, \sigma^2_\epsilon) = (2\pi\sigma^2_\epsilon)^{-n/2} \times e^{-\frac{1}{2\sigma^2_\epsilon}\sum \epsilon_i^2}$$

Using properties **P.3** (independence of Ys) and **P.4** (normality of **Y**), and that ϵ_i is a linear function of y_i, we can write the likelihood of the parameters given the observations and the normal probability distribution of **y** as,

$$\mathcal{L}\bigg(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2_\epsilon \mid Y, n\bigg) = (2\pi\sigma^2_\epsilon)^{-n/2} \times e^{-\frac{1}{2\sigma^2_\epsilon}\sum \big(\mathbf{y} - \mathbf{X}\beta\big)^2}$$

Or, in log-likelihood form,

$$\log \mathcal{L}\bigg(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2_\epsilon \mid \mathbf{y}, n\bigg) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2_\epsilon) - \frac{1}{2\sigma^2_\epsilon}\sum \big(\mathbf{y} - \mathbf{X}\beta\big)^2$$

We can again use calculus (beyond the scope of this course) to optimize this function.

Differentiating this expression with respect to each of the parameters, we get:

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \beta_0} &= \frac{1}{\sigma_\epsilon^2} \sum (\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial \log \mathcal{L}}{\partial \beta_1} &= \frac{1}{\sigma_\epsilon^2} \sum x_{1i}(\mathbf{y} - \mathbf{X}\beta) \\ &\vdots \\ \frac{\partial \log \mathcal{L}}{\partial \beta_k} &= \frac{1}{\sigma_\epsilon^2} \sum x_{ki}(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial \log \mathcal{L}}{\partial \sigma_\epsilon^2} &= -\frac{n}{2\sigma_\epsilon^2} + \frac{1}{2\sigma_\epsilon^4} \sum (\mathbf{y} - \mathbf{X}\beta)^2\end{aligned} \tag{1}$$

We can then set each of these equal to zero and solve.

Solving these equations, we find that the maximum likelihood estimators for the regression coefficients are equivalent to the OLS estimators of these parameters. We also find that,

$$\begin{aligned}\hat{\sigma}_\epsilon^2 &= \frac{1}{n} \sum (\mathbf{y} - \mathbf{X}\beta)^2 \\ &= \frac{1}{n} \sum (e_i)^2\end{aligned}$$

Thus the maximum likelihood estimate for the error variance is not the same as the OLS estimate for error variance (the OLS version divides the sume of the errors by $n - 2$).

References

Fox, J. (2016). *Applied regression analysis & generalized linear models* (3rd ed.). Sage.