Assumptions for OLS Regression and the Gauss-Markov Theorem

2021-09-06

Recall that there are a certain set of assumptions underlying the linear regression model (both the simple and multiple regression models). 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.

Another assumption that is useful for inference is:

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

Sampling Variance and Covariance of the Estimators

The sampling variances and covariance for the OLS estimators are:

$$\text{Var}(B_1) = \frac{\sigma_e^2}{\sum(x_i - \bar{x})^2}$$

$$\text{Var}(B_0) = \frac{\sigma_e^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

$$\text{Cov}(B_0, B_1) = \frac{\sigma_\epsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

To derive these quantities, we will have to take advantage of the assumptions that the conditional variances are equal (**A.4**) and independence of errors/observations conditional on X (**A.5**).

Sampling Variance of the Estimators

We will start by deriving the sampling variance for b_1 in a simple regression model.

Since $b_1 = \sum w_i y_i$, then

$$\begin{aligned}
\text{Var}(b_1) &= \text{Var}(\sum w_i y_i) \\
&= \sum \text{Var}(w_i y_i) \\
&= \sum w_i^2 \text{Var}(\mathbf{y})
\end{aligned}$$

Now, recall that the variance of y is conditional on **X**, then

$$\begin{aligned}
\text{Var}(y_i|x_i) &= \text{Var}(\hat{y}_i + e_i|x_i) \\
&= \text{Var}(\hat{y}_i|x_i) + \text{Var}(e_i|x_i) + 2\text{Cov}(\hat{y}_i, e_1|x_i) \\
&= 0 + \text{Var}(e_i|x_i) + 2(0) \\
&= \text{Var}(e_i|x_i) = \sigma_\epsilon^2
\end{aligned}$$

So,

$$\begin{aligned}
\text{Var}(b_1) &= \sum w_i^2 \sigma_\epsilon^2 \\
&= \sigma_\epsilon^2 \sum w_i^2 \\
&= \sigma_\epsilon^2 \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right)^2 \\
&= \sigma_\epsilon^2 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 \sum (x_i - \bar{x})^2} \\
&= \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

We can re-write this as

$$\text{Var}(b_1) = \frac{\sigma_\epsilon^2}{(n-1)s_x^2} \qquad \text{SE}(b_1) = \sqrt{\frac{\sigma_\epsilon^2}{(n-1)s_x^2}}$$

This helps us think about when the precision of b_1 will be high (low variance and SE):

- When the error variance, σ_ϵ^2, is small;
- When the sample size, *n*, is large; and
- When the variance in the predictor values, s_x^2, is large.

We can perform a similar derivation to obtain the sampling variance of b_0 (not shown). The formula for the sampling variance for b_0, also offers us insight for when the precision of b_0 will be high (low variance and SE):

$$\text{Var}(b_0) = \frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \qquad \text{SE}(b_0) = \sqrt{\frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

Because the denominator is similar to that of $\text{Var}(b_1)$, we will have high precision around b_0 when:

- The error variance, σ_ϵ^2, is small;
- The sample size, *n*, is large; and
- The variance in the predictor values, s_x^2, is large.

But, because of the added sum term in the numerator, when $\bar{x} \approx 0$, then that will essentially cancel out the sum terms in the numerator and denominator, which will lead to higher precision, so also when

- The X-values are centered near 0.

Covariance between the Estimators

The covariance between b_0 and b_1 is defined as:

$$\text{Cov}(b_0, b_1) = \frac{\sigma_\epsilon^2 \bar{x}}{\sum (x_i - \bar{x})^2}$$

We can derive this using the covariance and expectations rules.

$$\text{Cov}(b_0, b_1) = \mathbb{E}\Big[(b_0 - \mathbb{E}[b_0])(b_1 - \mathbb{E}[b_1])\Big]$$

$$\text{Cov}(b_0, b_1) = \mathbb{E}\Big[(b_0 - \beta_0)(b_1 - \beta_1)\Big]$$

Now we use the result that $b_0 - \beta_0 = -\bar{x}(b_1 - \beta_1)$. You can show this by using the result that $\bar{y} = b_0 + b_1(\bar{x})$, which implies $b_0 = \bar{y} - b_1(\bar{x})$. Substituting $-\bar{x}(b_1 - \beta_1)$ in for $b_0 - \beta_0$, we get:

$$\begin{aligned}
\text{Cov}(b_0, b_1) &= \mathbb{E}\Big[-\bar{x}(b_1 - \beta_1)^2\Big] \\
&= -\bar{x} \times \mathbb{E}\Big[(b_1 - \beta_1)^2\Big] \\
&= -\bar{x} \times \text{Var}(b_1) \\
&= -\bar{x} \times \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

This formula provides insight into the covariance between the regression estimators. Since both σ_ϵ^2 and $\sum (x_i - \bar{x})^2$ are values greater than zero, the covariance between b_0 and b_1 depends on the sign of $\bar{x}$.

- If $\bar{x} > 0$, then $\text{Cov}(b_0, b_1) < 0$. This implies that the sampling errors $(b_0 - \beta_0)$ and $(b_1 - \beta_1)$ have opposite signs.
- If $\bar{x} < 0$, then $\text{Cov}(b_0, b_1) > 0$. This implies that the sampling errors $(b_0 - \beta_0)$ and $(b_1 - \beta_1)$ have the same signs.

Sampling Distributions of the Estimators

To derive the sampling distributions for the estimators, which are the basis for statistical inference, we need to also take advantage of the normality assumption (**A.7**). Recall that we used the *t*-distribution with $n - 2$ degrees of freedom to test hypotheses about the slope and intercept. The general form of a hypotheses test for a regression coefficient, is

$$H_0 : \beta_j = k \quad \text{where k is the tested value}$$

To test this, we create a test statistic (T) by studentizing the regression estimator (b_j) as:

$$T = \frac{B_j - k}{\text{SE}(B_j)}$$

In the simple regression model this statistic follows a t-distribution with $n - 2$ degrees of freedom.

Recall from introductory statistics if we could assume that the population was normally distributed, then the distribution of $T = \frac{\bar{y}-\mu}{\text{SE}(\bar{y})}$ was *t*-distributed with $n - 1$ degrees of freedom. Since $\bar{y}$ is a linear combination of the observations, the distribution of $\bar{y}$ is also normally distributed, and estimating the SE of $\bar{y}$ introduced additional error; making the distribution of T follow a t-distribution.

This comes from a theorem which says that (1) if Z is a standard normal variable and W is chi-squared distributed with ν degrees of freedom, and (2) Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}}$$

will follow a t-distribution with ν degrees of freedom.

In the case of the test of the mean, we can write $T = \frac{\bar{y}-\mu}{\text{SE}(\bar{y})}$ in this form as:

$$T = \frac{\sqrt{n}(\bar{y} - \mu)/\sigma_y}{\sqrt{\left[(n-1)s_y^2/\sigma_y^2\right]/(n-1)}}$$

Thus, the distribution of T will be t-distributed with $n - 1$ degrees of freedom.

For the regression estimators, under the assumption of normality, the least squares estimators are also normally distributed. This is true since B_j is a linear combination of the observations, and we are assuming the observations to be normally distributed (linear shifts do not change the distribution). Thus, we have,

$$T = \frac{B_j - k}{\text{SE}(B_j)}$$

$$= \frac{\frac{B_j - b}{\sigma_{B_j}}}{\frac{\text{SE}(B_j)}{\sigma_{B_j}}}$$

From this it is clear that the numerator of T is a standard normal variable. The denominator is

$$\begin{aligned}
\frac{\text{SE}(B_j)}{\sigma_{B_j}} &= \sqrt{\frac{\text{Var}(B_j)}{\sigma^2_{B_j}}} \\
&= \sqrt{\frac{\frac{\text{MSE}}{\sum(x_i-\bar{x})^2}}{\frac{\sigma^2_\epsilon}{\sum(x_i-\bar{x})^2}}} \\
&= \sqrt{\frac{\text{MSE}}{\sigma^2_\epsilon}} \\
&= \sqrt{\frac{\frac{\text{SSE}}{n-2}}{\sigma^2_\epsilon}} \\
&= \sqrt{\frac{\text{SSE}}{\sigma^2_\epsilon(n-2)}}
\end{aligned}$$

At this point we rely on a common theorem from regression theory which says that $\frac{\text{SSE}}{\sigma^2_\epsilon}$ is distributed as χ^2 with $n-2$ degrees of freedom and is independent of both b_0 and b_1. Relying on this,

$$T = \frac{B_j - k}{\text{SE}(B_j)} = \frac{z}{\sqrt{\frac{\chi^2(n-2)}{(n-2)}}}$$

Since z is a function of B_0 and B_1, then z and χ^2 are also independent, and it follows that $\frac{B_j-k}{\text{SE}(B_j)}$ is t-distributed with $n-2$ degrees of freedom.

> The maximum likelihood estimators for the coefficients (which are the same as the OLS estimators) possess several asymptotic (large sample) properties; most importantly normality. Because of this, with a large sample size, the sampling distribution for the estimates will be approximately t-distributed with $n-2$ degrees of freedom.

Inference in the Multiple Regression Model

Implications for Applied Researchers

If the assumptions underlying the strong classical regression model (**A.1**–**A.7**) are all valid, then the OLS estimators $b_0, b_1, \ldots, b_k$ are good estimators of $\beta_0, \beta_1, \ldots, \beta_k$. They are unbiased and efficient and have accurate sampling variances and covariance.

Of course, any of the assumptions may be challenged either on *a priori* substantive grounds, or *post hoc*, via empirical examination of the sample residuals. If one (or more) of the assumptions are violated, then some of the properties may be compromised. **If this is the case, one can often transform the data in some way or use an alternative estimation technique.**

Violation of the normality assumption (**A.7**) causes the least number of problems. Under non-normality the regression estimators are still BLUE, and s^2_e is still an unbiased estimator of σ^2_ϵ. However, the under non-normality, the use of the *F*-

and t-distributions for inference is questionable especially if the sample size is small. If the sample size is large, the sampling distributions of the coefficients are approximately normal and subsequently the use of the F- and t-distributions for inference is justified.

If Assumptions **A.1**–**A.6** are violated, then the OLS estimators are no longer BLUE. Moreover, the formulas used to compute the sampling variances will also be incorrect (which also affects inference).

References