Statistical Inference for the Regression Model

2021-09-09

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, $\mathbf{X}$, is of full rank.
- **A.3:** The population errors given $\mathbf{X}$ have a mean of zero.
- **A.4:** The population errors given $\mathbf{X}$ are homoscedastic.
- **A.5:** The population errors given $\mathbf{X}$ 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 $\mathbf{X}$ are normally distributed

Sampling Distribution of the OLS Estimators

When $\mathbf{X}$ is fixed, the $\mathbf{b}$ vector can be written as a linear transformation of the response vector $\mathbf{y}$:

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$= \mathbf{M} \mathbf{y}$$

where $\mathbf{M} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

Previously we showed that the estimator $\mathbf{b}$ are unbiased estimates of β, that is $\mathbb{E}(\mathbf{b}) = \beta$. We can also define the variance–covariance matrix of $\mathbf{b}$.

$$\mathrm{Var}(\mathbf{b}) = \mathbf{M}\,\mathrm{Var}(\mathbf{y})\mathbf{M}^\top$$

$$= [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]\sigma_\epsilon^2 \mathbf{I}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top$$

Rearranging this and using our rules of transposes and inverses, we get:

$$\mathrm{Var}(\mathbf{b}) = [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]\sigma_\epsilon^2 \mathbf{I}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top$$

$$= \sigma_\epsilon^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$$

$$= \sigma_\epsilon^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

This implies that the sampling variances and covariances of the estimators depend only on the predictor values and the error variance. This matrix is often reffered to as **V**.

Note that to derive this, we only need the assumption of linearity. The normality assumption is not used to compute the mean (expectation) nor the sampling variance, nor covariances of the estimators. However, if the errors (and hence **y**) are normally distributed (i.e., assumption **A.7**), then so are the sampling distributions of the estimators. We can express these distributions as:

$$\mathbf{b} \sim \mathcal{N}\left(\beta, \ \sigma_\epsilon^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)$$

Inference for the Estimators

An individual estimator, b_j, has a sampling distribution of:

$$b_j \sim \mathcal{N}\left(\beta_j, \ \mathbf{V}_{jj}\right)$$

where $\mathbf{V}_{jj}$ is the element in the jth row and jth column of the variance–covariance matrix for **b**. To test the hypothesis that:

$$H_0 : \beta_j = \beta_j^{(0)}$$

where $\beta_j^{(0)}$ is some value (e.g., $H_0 : \beta_j = 0$), we compute the ratio:

$$Z_0 = \frac{b_j - \beta_j^{(0)}}{\sqrt{\mathbf{V}_{jj}}}$$

which is unit-normal distributed; $\mathcal{N}(0, 1)$. Note the denominator is the standard deviation of the sampling distribution for b_j.

Unfortunately, in practice we do not know σ_ϵ to compute **V**. Instead, we substitute in the estimate for this value from our sample, the unbiased estimator s_e, where:

$$s_e^2 = \frac{\mathbf{e}^\top\mathbf{e}}{n - k - 1}$$

where n is the sample size, k is the number of predictors in the model, and $n - k - 1$ is the residual degrees of freedom for the model. Thus, we get an estimate of the variance–covariance matrix using:

$$\widehat{\mathrm{Var}}(\mathbf{b}) = s_e^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

$$= \frac{\mathbf{e}^\top\mathbf{e}}{n - k - 1}(\mathbf{X}^\top\mathbf{X})^{-1}$$

The standard error for b_j is then estimated using the jth diagonal element of this matrix, namely $\sqrt{\hat{v}_{jj}}$.

There is 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, defined as:

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

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

Since it can be shown that $(n - k - 1)\frac{s_e^2}{\sigma_\epsilon^2}$ follows a chi-squared distribution with $n - k - 1$ degrees of freedom, and also that B_j and s_e^2 are independent; it follows that:

$$t = \frac{b_j - \beta_j^{(0)}}{\sqrt{\mathbf{V}_{jj}}}$$

is t-distributed with $n - k - 1$ degrees of freedom. To evaluate the hypothesis that $H_0 : \beta_j = \beta_j^{(0)}$, we compute the test statistic:

$$t_0 = \frac{b_j - \beta_j^{(0)}}{\sqrt{\mathbf{V}_{jj}}}$$

and evaluate it within the t-distribution having $n - k - 1$ degrees of freedom (where n is the sample size, and k is the number of predictors in the model).

Confidence Intervals

A $1 - \alpha\%$ confidence interval can also be constructed for each coefficient using:

$$\text{CI}_{1-\alpha} = b_j \pm |t^*_{\alpha/2}|\big[s_e(\mathbf{X}^\top\mathbf{X})^{-1}\big]$$

$$= b_j \pm |t^*_{\alpha/2}|(\sqrt{\mathbf{V}_{jj}})$$

where $t^*_{\alpha/2}$ is the critical value demarcating the area in the $\alpha/2$ proportion of the distribution in the t-distribution with $n - k - 1$ degrees of freedom. Fo example, if we wanted to compute a 95% CI, then $\alpha = 0.05$ and t^* would be the critical value that demarcates the lowest 0.025 of the distribution.

Model-Level Inference

At the model-level we are interested in testing the hypothesis $H_0 : \rho^2 = 0$. Recall this is equivalent to testing the hypothesis that all the regression parameters (except the intercept) in the model are zero:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k$$

This is a specific form of the general linear hypothesis which can be expressed as:

$$H_0 : \mathbf{L}\beta = \mathbf{c}$$

where,

- **L** is a $q \times (k+1)$ matrix referred to as the *hypothesis matrix*, where q is the number of parameters being tested;
- β is a $(k+1) \times 1$ vector of parameters included in the model; and
- **c** is a $q \times 1$ vector of hypothesized values.

As an example, if we wanted to test the model-level null hypothesis in a two predictor model ($y_i = \beta_0 + \beta_1(x_{1i}) + \beta_2(x_{2i}) + \epsilon_i$),

$$H_0 : \beta_1 = \beta_2 = 0$$

Then the general linear hypothesis could be expressed as:

$$H_0 : \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This results in:

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We can construct a test statistic, F_0, as:

$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})^\top \left[\mathbf{L}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{L}^\top \right]^{-1} (\mathbf{Lb} - \mathbf{c})}{q(s_e^2)}$$

This follows an *F*-distribution with q and $n - k - 1$ degrees of freedom

Implications for Applied Researchers

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

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. Let's look at violation in turn:

Violating A.1: The model is not correctly specified.

This is probably the most egregious and costly violation. Violating this assumption means that the theoretical model embodied in the regression equation is wrong. If this is the case, there is no use proceeding.

Violating A.2: The design matrix, X, is not of full rank.

This would mean that we cannot compute an inverse for the $\mathbf{X}^\top \mathbf{X}$ matrix. Subsequently, we could not compute any parameter estimates nor SEs.

Violating A.3: The population errors given X do not have a mean of zero.

The estimator b_0 will be biased when this assumption is violated. However, the other regression estimators (e.g., $b_1, b_2, \ldots, b_k$) are still BLUE. Unfortunately, we can never test this assumption in practice since $\sum e_i = 0$ always holds when using OLS.

Violating A.4: The population errors given X are heteroscedastic.

If the assumption of constant variance is violated, the coefficient estimators will remain unbiased. However, the estimates for the variances/standard errors for these coefficients will be wrong. This also impacts the confidence intervals and hypothesis tests for these parameters. There are solutions to this problem, including using a different estimation method (e.g., weighted least squares) or estimating the sampling variances for these coefficients differently (e.g., sandwich estimation).

Violating A.5: The population errors given X are not independent.

If the errors are correlated (not independent), the coefficient estimators will still be unbiased. Again, however, the variances/standard errors and results from the confidence intervals and hypothesis tests will be wrong. Often the hypothesis tests will indicate statistical significance much more often then it should (increased probability of a type I error). Dealing with this problem requires using models that account for the correlation (e.g., mixed-effects models) and different estimation methods (e.g., maximum likelihood).

Violating A.6: The predictor values are not fixed.

When $\mathbf{X}$ is fixed, the predictors are uncorrelated with the error terms. This is what leads to unbiased estimates of both the estimators and the variances/standard errors. When $\mathbf{X}$ is random (which is almost always the case in observational data), we need to assume that $\mathrm{Cov}(\mathbf{X}, \epsilon) = 0$. That is we assume that the predictor values are generated by a mechanism unrelated to the errors. If that is not the case, then the OLS estimates will be biased.

Violating A.7: The population errors given X are not normally distributed.

Violation of the normality assumption causes the least number of problems. Under non-normality the regression estimators are still BLUE, and s_e^2 is still an unbiased estimator of σ_ϵ^2. 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.

References