Central Limit Theorem for Linear Models

Fall 2014

The CLT you learn in Math Stat applies to sums or averages of independent R.V.s. We need a slightly different version to apply to coefficient estimates, which are much like means.
From Math Stat:
Lindeberg CLT (Arnold, 1981) says that if $Y_1, Y_2, \ldots \sim (0, \sigma^2)$ with $\sigma^2 < \infty$, and $c_1, c_2, \ldots$ is a sequence of constants such that $\frac{\max c_j^2}{\sum c_j^2} \longrightarrow 0$ as $n \to \infty$, then $\frac{\sum_{i=1}^n c_j Y_j}{\sqrt{\sum_{j=1}^n c_j^2}}$ converges in distribution to $N(0, \sigma^2)$. The condition keeps the c_j's from increasing so fast that the last term dominates the sum.

Simple case:
Suppose we have a balanced ANOVA model (1, 2, or multiway) with n observations per cell, $Y_{ij} - \mu_i \sim$ iid$(0, \sigma^2)$. Take the mean in group i to be $\frac{1}{n}\sum(Y_{ij} - \mu_i)$, so let $c_j = n^{-1}$. The condition is easily satisfied because $\max(c_j^2) = n^{-2}$ and $\sum c_j^2 = n \times n^{-2} = n^{-1}$ so the fraction is n^{-1} which goes to 0 nicely as $n \to \infty$. We conclude

$$\frac{n^{-1}\sum(Y_{ij} - \mu_i)}{n^{-1/2}} = n^{1/2}(\overline{Y}_i - \mu_i) \xrightarrow{\mathcal{D}} N(0, \sigma^2) \text{ or } n^{1/2}\overline{Y}_i \xrightarrow{\mathcal{D}} N(\mu_i, \sigma^2).$$

For any n, the cell means are independent of each other. Any linear combination of cell means will converge in distribution to a normal distribution.

Full linear model (Sen and Srivastava, 1997)
Change the condition for the CLT slightly, letting $a_i = c_i/\sqrt{\sum c_i^2}$ and require $\max|a_{n_i}| \longrightarrow 0$ and $\sum_i a_{n_i}^2 \longrightarrow 1$. This allows the use of a triangular array of constants instead of reusing the same values (Gnedenko and Kolmogorov, 1954). The conclusion is that $\sum a_{n_i} Y_i \xrightarrow{\mathcal{D}} N(0, \sigma^2)$ as $n \to \infty$.
Linear model:
Take the simple case: $\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ with $\boldsymbol{X}$ of full column rank, and consider the coefficient vector $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, or actually the distribution of

$$\begin{aligned}
\sigma^{-2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \sigma^{-2}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{\beta}]^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{\beta}] \\
&= \sigma^{-2}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta})]^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta})] \\
&= \sigma^{-2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}) \\
&= \sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \sigma^{-2}\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\epsilon} \\
&= \sigma^{-2}\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{H}\boldsymbol{\epsilon} = \sigma^{-2}\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{P}_{\text{ppo}}\boldsymbol{\epsilon}
\end{aligned}$$

Use the full rank SVD or eigenvalue decomposition on $\boldsymbol{H}_n$ to obtain $\boldsymbol{H} = \boldsymbol{L}_n{}^\mathsf{T}\boldsymbol{L}_n$ with $\boldsymbol{L}_n\boldsymbol{L}_n{}^\mathsf{T} = \boldsymbol{I}_r$. Take a vector $\boldsymbol{b}$ s.t. $\boldsymbol{b}^\mathsf{T}\boldsymbol{b} = 1 = \sum b_i^2$ and let $a_n^{(i)} = \boldsymbol{b}^\mathsf{T}\boldsymbol{L}_n^{(i)}$ (a scalar), then look at $\boldsymbol{b}^\mathsf{T}\boldsymbol{L}_n\boldsymbol{\epsilon} = \sum_{i=1}^n a_n^{(i)}\epsilon_i$.

$$|a_n^{(i)}| = |\boldsymbol{b}^\mathsf{T}\boldsymbol{L}_n^{(i)}| \le (\boldsymbol{b}^\mathsf{T}\boldsymbol{b})^{1/2}(\boldsymbol{L}_n^{(i)\mathsf{T}}\boldsymbol{L}_n^{(i)})^{1/2} = 1 \times h_{ii}$$

By Hölder's inequality, where h_{ii} is the ith diagonal of the hat matrix. We need: $\max h_{ii} = \max \boldsymbol{L}_n^{(i)\mathsf{T}}\boldsymbol{L}_n^{(i)} \longrightarrow 0$, so the condition needed is that diagonals of the hat matrix, the leverages, must go to zero as $n \longrightarrow 0$. Then $\sum_i (a_n^{(i)})^2 = \boldsymbol{b}^\mathsf{T}\boldsymbol{L}_n\boldsymbol{L}_n{}^\mathsf{T}\boldsymbol{b} = \boldsymbol{b}^\mathsf{T}\boldsymbol{b} = 1$ so we conclude that

$\boldsymbol{b}^\mathsf{T}\boldsymbol{L}_n\boldsymbol{\epsilon} \longrightarrow N(0,\sigma^2)$ for every norm one vector $\boldsymbol{b}$. By properties of MVN, $\boldsymbol{L}_n\boldsymbol{\epsilon}$ has an asymptotic $N(\boldsymbol{0},\sigma^2\boldsymbol{I})$ distribution. Hence

$$\sigma^{-2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \chi^2_r.$$

In practice, we use t and F tests, so we need to ensure that these distributions also hold. We know that s^2 converges to σ^2 in probability so

$$s^{-2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \chi^2_r.$$

In general if $\boldsymbol{C}$ is a p by q matrix of rank q and $\boldsymbol{C}^\mathsf{T}\boldsymbol{\beta}$ is estimable, then

$$s^{-2}(\boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\boldsymbol{C}^\mathsf{T}\boldsymbol{\beta})^\mathsf{T}[\boldsymbol{C}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^g\boldsymbol{C}]^{-1}(\boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\boldsymbol{C}^\mathsf{T}\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \chi^2_q.$$

However, the above assumes that s^2 is a "perfect" estimate of the variance. We generally prefer to not make that assumption and use the F distribution instead.

$$\frac{1}{qs^2}(\boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\boldsymbol{C}^\mathsf{T}\boldsymbol{\beta})^\mathsf{T}[\boldsymbol{C}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^g\boldsymbol{C}]^{-1}(\boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\boldsymbol{C}^\mathsf{T}\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} F_{q,n-r}.$$

The critical assumption is that no single point is entirely "influential", but the leverage of each is going to zero. We shall see that $\sum h_{ii} = \text{trace}(\boldsymbol{H}) = \text{rank}(\boldsymbol{X})$ which does not increase as n does.

More generally, the above also works with $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$ for known $\boldsymbol{V}$, we just need to do a little transformation. Take $\boldsymbol{L}$ to be the Cholesky decomposition of $\boldsymbol{V}^{-1}$. That means $\boldsymbol{V}^{-1} = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$, with $\text{rank}(\boldsymbol{V}^{-1}) = \text{rank}(\boldsymbol{V}) = \text{rank}(\boldsymbol{L})$, since we cannot increase rank through multiplication. $\boldsymbol{L}$ and $\boldsymbol{L}^\mathsf{T}$ are therefore non–singular, and $\boldsymbol{L}^{\mathsf{T}-1}\boldsymbol{L}^{-1} = (\boldsymbol{L}\boldsymbol{L}^\mathsf{T})^{-1} = \boldsymbol{V}$

Premultiply our linear model by $\boldsymbol{L}^\mathsf{T}$ to get:

$$\boldsymbol{L}^\mathsf{T}\boldsymbol{y} = \boldsymbol{L}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{L}^\mathsf{T}\boldsymbol{\epsilon} = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

The mean of $\boldsymbol{\epsilon}^*$ is still $\boldsymbol{0}$, and it has variance-covariance matrix $\sigma^2\boldsymbol{L}^\mathsf{T}\boldsymbol{V}\boldsymbol{L} = \sigma^2\boldsymbol{L}^\mathsf{T}\boldsymbol{L}^{\mathsf{T}-1}\boldsymbol{L}^{-1}\boldsymbol{L} = \sigma^2\boldsymbol{I}$, so the second formulation of our model fits the conditions for the CLT and as n gets big, the sampling distribution of $\widehat{\boldsymbol{\beta}}$ converges to a normal distribution.

In practice, we never know exactly how big n needs to be to make the sampling distribution close enough to normality so that our inferences based on t and F distributions are valid. Problems in convergence occur when we have skewed residuals and/or outliers in the data. If residuals have short tails, the CLT will work well even if sample sizes are only moderate ($n - r = 30$?), otherwise we might need several hundred points to appy the CLT.

References

Arnold, S. *The Theory of Linear Models and Multivariate Analysis*. Wiley (1981).

Sen, A. K. and Srivastava, M. S. *Regression Analysis: Theory, Methods, and Applications*. Springer-Verlag Inc (1997).