Confidence Intervals and Sets

Throughout we adopt the normal-error model, and wish to say some things about the construction of confidence intervals [and sets] for the parameters $\beta_0, \ldots, \beta_{p-1}$.

1. Confidence intervals for one parameter

Suppose we want a confidence intervals for the ith parameter β_i, where $1 \le i \le p$ is fixed. Recall that

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}} \sim \mathrm{N}(0\,,1),$$

and

$$S^2 := \frac{\mathrm{RSS}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)_i\right)^2 \sim \frac{\sigma^2\chi^2_{n-p}}{n-p},$$

and is independent of $\hat{\boldsymbol{\beta}}$. Therefore,

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{S^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}} \sim \frac{\mathrm{N}(0\,,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} = t_{n-p},$$

where the normal and χ^2 random variables are independent. Therefore,

$$\hat{\beta}_i \pm t^{(\alpha/2)}_{n-p} S \sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}$$

is a $(1-\alpha)\times 100\%$ confidence interval for β_i. This yields a complete analysis of confidence intervals for a univariate parameter; these confidence intervals can also be used for testing, of course.

2. Confidence ellipsoids

The situation is more interesting if we wish to say something about more than one parameter at the same time. For example, suppose we want to know about (β_0, β_1) in the *overly-simplified case that $\sigma = 1$*. The general philosophy of confidence intervals [for univariate parameters] suggests that we look for a random set Ω such that

$$\mathrm{P}\left\{ \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \in \Omega \right\} = 1 - \alpha.$$

And we know that

$$\begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\Sigma} \right),$$

where $\boldsymbol{\Sigma}$ is a 2×2 matrix with

$$\Sigma_{i,j} = \left[(\mathbf{X}'\mathbf{X})^{-1} \right]_{i,j}.$$

Here is a possible method: Let

$$\hat{\boldsymbol{\gamma}} := \begin{pmatrix} \hat\beta_1 - \beta_1 \\ \hat\beta_2 - \beta_2 \end{pmatrix}, \quad \text{so that} \quad \hat{\boldsymbol{\gamma}} \sim \mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma}).$$

Also, recall that

$$\hat{\boldsymbol{\gamma}}' \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\gamma}} \sim \chi^2_2.$$

Therefore, one natural choice for Ω is

$$\Omega := \left\{ \boldsymbol{x} \in \mathbf{R}^2 : \ \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \le \chi^2_2(\alpha/2) \right\}.$$

What does Ω look like? In order to answer this, let us apply the spectral theorem:

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}' \qquad \Longrightarrow \qquad \boldsymbol{\Sigma}^{-1} = \boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}'.$$

Then we can represent Ω as follows:

$$\begin{aligned} \Omega &= \left\{ \boldsymbol{x} \in \mathbf{R}^2 : \ \boldsymbol{x}'\boldsymbol{P}\boldsymbol{D}^{-1}\boldsymbol{P}'\boldsymbol{x} \le \chi^2_2(\alpha/2) \right\} \\ &= \left\{ \boldsymbol{x} \in \mathbf{R}^2 : \ (\boldsymbol{P}'\boldsymbol{x})\boldsymbol{D}^{-1}(\boldsymbol{P}'\boldsymbol{x}) \le \chi^2_2(\alpha/2) \right\} \\ &= \left\{ \boldsymbol{P}\boldsymbol{y} \in \mathbf{R}^2 : \ \boldsymbol{y}'\boldsymbol{D}^{-1}\boldsymbol{y} \le \chi^2_2(\alpha/2) \right\} \\ &= \left\{ \boldsymbol{P}\boldsymbol{y} \in \mathbf{R}^2 : \ \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} \le \chi^2_2(\alpha/2) \right\}. \end{aligned}$$

Consider

$$\mathcal{E} := \left\{ \boldsymbol{y} \in \mathbf{R}^2 : \ \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} \le \chi_2^2(\alpha/2) \right\}. \tag{1}$$

This is the interior of an ellipsoid, and $\Omega = \boldsymbol{P}\mathcal{E}$ is the image of the ellipsoid under the "linear orthogonal map" $\boldsymbol{P}$. Such sets are called "generalized ellipsoids," and we have found a $(1-\alpha)\times 100\%$ confidence [generalized] ellipsoid for $(\beta_1, \beta_2)'$.

The preceding can be generalized to any number of the parameters $\beta_{i_1}, \ldots, \beta_{i_k}$, but is hard to work with, as the geometry of Ω can be complicated [particularly if $k \gg 2$]. Therefore, instead we might wish to look for approximate confidence sets that are easier to work with. Before we move on though, let me mention that if you want to know whether or not $H_0 : \ \beta_1 = \beta_{1,0},\ \beta_2 = \beta_{2,0}$, then we can use these confidence bounds fairly easily, since it is not hard to check whether or not $(\beta_{1,0}, \beta_{2,0})'$ is in Ω: You simply compute the scalar quantity

$$(\beta_{1,0}, \beta_{2,0})\, \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \beta_{1,0} \\ \beta_{2,0} \end{pmatrix},$$

and check to see if it is $\le \chi_2^2(\alpha/2)$! But if you really need to imagine or see the confidence set[s], then this exact method can be unwieldy [particularly in higher dimensions than 2].

3. Bonferonni bounds

Our approximate confidence intervals are based on a fact from general probability theory.

Proposition 1 (Bonferonni's inequality). *Let $E_1, \ldots, E_k$ be k events. Then,*

$$\mathrm{P}\left(E_1 \cap \cdots \cap E_k\right) \ge 1 - \sum_{j=1}^{k} \mathrm{P}(E_j^c) = 1 - \sum_{j=1}^{k} \left(1 - \mathrm{P}(E_j)\right).$$

Proof. The event $E_1^c \cup \cdots \cup E_k^c$ is the complement of $E_1 \cap \cdots \cap E_k$. Therefore,

$$\mathrm{P}\left(E_1 \cap \cdots \cap E_k\right) = 1 - \mathrm{P}\left(E_1^c \cup \cdots \cup E_k^c\right),$$

and this is $\ge 1 - \sum_{j=1}^{k} \mathrm{P}(E_j^c)$ because the probability of a union is at most the sum of the individual probabilities. □

Here is how we can use Bonferonni's inequality. Define

$$C_j := \hat{\beta}_j \pm t_{n-p}^{(\alpha/4)} S \sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{j,j}} \qquad (j = 1, 2).$$

We have seen already that

$$\mathrm{P}\{\beta_j \in C_j\} = 1 - \frac{\alpha}{2}.$$

[This is why we used $\alpha/4$ in the definition of C_j.] Therefore, Bonferonni's inequality implies that

$$\mathrm{P}\{\beta_1 \in C_1 \ , \ \beta_2 \in C_2\} \geq 1 - \left[\frac{\alpha}{2} + \frac{\alpha}{2}\right] = 1 - \alpha.$$

In other words, $(C_1 \, , C_2)$ forms a "conservative" $(1-\alpha) \times 100\%$ simultaneous confidence interval for $(\beta_1 \, , \beta_2)$. This method becomes very inaccurate quickly as the number of parameters of interest grows. For instance, if you want Bonferonni confidence sets for $(\beta_1 \, , \beta_2 \, , \beta_3)$, then you need to use individual confidence intervals with confidence level $1 - (\alpha/3)$ each. And for k parameters you need individual confidence level $1 - (\alpha/k)$, which can yield bad performance when k is large.

This method is easy to implement, but usually **very** conservative.[1]

4. Scheffé's simultaneous conservative confidence bounds

There is a lovely method, due to Scheffé, that works in a similar fashion to the Bonferonni method; but has also the advantage of being often [far] less conservative! [We are now talking strictly about our linear model.] The starting point of this discussion is a general fact from matrix analysis.

Proposition 2 (The Rayleigh—Ritz inequality). *If $\boldsymbol{L}_{p\times p}$ is positive definite, then for every p-vector $\boldsymbol{b}$,*

$$\boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b} = \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{(\boldsymbol{h}'\boldsymbol{b})^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}}\right].$$

The preceding is called an inequality because it says that

$$\boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b} \leq \frac{(\boldsymbol{h}'\boldsymbol{b})^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}} \qquad \text{for all } \boldsymbol{h}_{p\times 1},$$

and it also tells us that the inequality is achieved for some $\boldsymbol{h} \in \mathbf{R}^p$.

[1] On the other hand, the Bonferonni method can be applied to a wide range of statistical problems that involve simultaneous confidence intervals [and is not limited to the theory of linear models. So it is well worth your while to understand this important method.

Proof of Rayleigh–Ritz inequality. Recall the Cauchy–Schwarz inequality from your linear algebra course: $(\boldsymbol{u}'\boldsymbol{v})^2 \leq \|\boldsymbol{u}\|^2 \cdot \|\boldsymbol{v}\|^2$ with equality if and only if $\boldsymbol{v} = a\boldsymbol{u}$ for some $a \in \mathbf{R}$. It follows then that

$$\max_{\boldsymbol{v}\neq\boldsymbol{0}} \left[\frac{(\boldsymbol{u}'\boldsymbol{v})^2}{\boldsymbol{v}'\boldsymbol{v}} \right] = \|\boldsymbol{u}\|^2.$$

Now write $\boldsymbol{L} := \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, in spectral form, and change variables:

$$\boldsymbol{v} := \left(\boldsymbol{L}^{1/2}\right)' \boldsymbol{h},$$

in order to see that

$$\boldsymbol{u}'\boldsymbol{v} = \boldsymbol{u}' \left(\boldsymbol{L}^{1/2}\right)' \boldsymbol{h}, \qquad \boldsymbol{v}'\boldsymbol{v} = \boldsymbol{h}'\boldsymbol{L}\boldsymbol{h},$$

and so

$$\|\boldsymbol{u}\|^2 = \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left(\boldsymbol{u}' \left(\boldsymbol{L}^{1/2}\right)' \boldsymbol{h}\right)^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}} \right].$$

Change variables again: $\boldsymbol{b} := \boldsymbol{L}^{1/2}\boldsymbol{u}$ to see that

$$\|\boldsymbol{u}\|^2 = \left\| \boldsymbol{L}^{-1/2}\boldsymbol{b} \right\|^2 = \boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b}.$$

This does the job. □

Now let us return to Scheffé's method. Recall that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathrm{N}_p \left(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \right),$$

so that

$$\frac{1}{\sigma^2} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2_p,$$

and hence

$$\frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{pS^2} = \frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)/p}{S^2} \sim F_{p,n-p}.$$

We apply the Rayleigh–Ritz inequality with $\boldsymbol{b} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $\boldsymbol{L} := (\boldsymbol{X}'\boldsymbol{X})^{-1}$ in order to find that

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left\{\boldsymbol{h}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\}^2}{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}} \right].$$

Therefore,

$$\mathrm{P}\left\{ \frac{1}{pS^2} \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left\{\boldsymbol{h}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\}^2}{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}} \right] \leq F_{p,n-p}(\alpha) \right\} = 1 - \alpha.$$

Equivalently,

$$\mathrm{P}\left\{\frac{\left|\boldsymbol{h}'\hat{\boldsymbol{\beta}}-\boldsymbol{h}'\boldsymbol{\beta}\right|}{\sqrt{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}}}\le\sqrt{pS^2\,F_{p,n-p}(\alpha)}\text{ for all }\boldsymbol{h}\in\mathbf{R}^p\right\}=1-\alpha.$$

If we restrict attention to a subcollection of $\boldsymbol{h}$'s then the probability is even more. In particular, consider only $\boldsymbol{h}$'s that are the standard basis vectors of $\mathbf{R}^p$, in order to deduce from the preceding that

$$\mathrm{P}\left\{\frac{\left|\hat{\beta}_j-\beta_j\right|}{\sqrt{[(\boldsymbol{X}'\boldsymbol{X})]_{j,j}}}\le\sqrt{pS^2\,F_{p,n-p}(\alpha)}\text{ for all }1\le j\le p\right\}\ge 1-\alpha.$$

In other words, we have demonstrated the following.

Theorem 3 (Scheffé). *The following are conservative* $(1-\alpha)\times 100\%$ *simultaneous confidence bounds for* $(\beta_1,\ldots,\beta_p)'$*:*

$$\hat{\beta}_j\pm\sqrt{pS^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{j,j}F_{p,n-p}(\alpha)}\qquad(1\le j\le p).$$

When the sample size n is very large, the preceding yields an asymptotic simplification. Recall that $\chi^2_{n-p}/(n-p)\to 1$ as $n\to\infty$. Therefore, $F_{p,n-p}(\alpha)\approx\chi^2_p(\alpha)/p$ for $n\gg 1$. Therefore, the following are asymptotically-conservative simultaneous confidence bounds:

$$\hat{\beta}_j\pm S\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{j,j}\chi^2_p(\alpha)}\qquad(1\le j\le p)\text{ for }n\gg 1.$$

5. Confidence bounds for the regression surface

Given a vector $\boldsymbol{x}\in\mathbf{R}^p$ of predictor variables, our linear model yields $\mathrm{E}\boldsymbol{y}=\boldsymbol{x}'\boldsymbol{\beta}$. In other words, we can view our efforts as one about trying to understand the unknown function

$$f(\boldsymbol{x}):=\boldsymbol{x}'\boldsymbol{\beta}.$$

And among other things, we have found the following estimator for f:

$$\hat{f}(\boldsymbol{x}):=\boldsymbol{x}'\hat{\boldsymbol{\beta}}.$$

Note that if $\boldsymbol{x}\in\mathbf{R}^p$ is held fixed, then

$$\hat{f}(\boldsymbol{x})\sim\mathrm{N}\left(f(\boldsymbol{x}),\sigma^2\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\right).$$

Therefore, a $(1-\alpha)\times 100\%$ confidence interval for $f(\boldsymbol{x})$ [for a fixed predictor variable $\boldsymbol{x}$] is

$$\hat{f}(\boldsymbol{x})\pm S\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}\,t^{(\alpha/2)}_{n-p}.$$

That is, if we are interested in a confidence interval for $f(\boldsymbol{x})$ for a fixed $\boldsymbol{x}$, then we have

$$\mathrm{P}\left\{f(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) \pm S\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}\, t_{n-p}^{(\alpha/2)}\right\} = 1-\alpha.$$

On the other hand, we can also apply Scheffé's method and obtain the following simultaneous $(1-\alpha)\times 100\%$ confidence set:

$$\mathrm{P}\left\{f(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) \pm S\sqrt{p\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\, F_{p,n-p}(\alpha)} \text{ for all } \boldsymbol{x}\in\mathbf{R}^p\right\} \geq 1-\alpha.$$

Example 4 (Simple linear regression)**.** Consider the basic regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1\leq i\leq n).$$

If $\boldsymbol{w} = (1\,, w)'$, then a quick computation yields

$$\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w} = \frac{\left[\overline{x^2} - 2w\bar{x} + w^2\right]}{ns_x^2},$$

where we recall $s_x^2 := \sum_{i=1}^p (x_i-\bar{x})^2$. Therefore, a simultaneous confidence interval for all of $\alpha+\beta w$'s, as w ranges over $\mathbf{R}$, is

$$\hat{\alpha}+\hat{\beta}w \pm S\sqrt{\frac{2}{ns_x^2}\left[\overline{x^2} - 2w\bar{x} + w^2\right]F_{2,n-2}(\alpha)}.$$

This expression can be simplified further because:

$$\begin{aligned}\overline{x^2} - 2w\bar{x} + w^2 &= \overline{x^2} - (\bar{x})^2 + (\bar{x})^2 - 2w\bar{x} + w^2\\ &= s_x^2 + (\bar{x}-w)^2.\end{aligned}$$

Therefore, with probability $1-\alpha$, we have

$$\alpha+\beta w = \hat{\alpha}+\hat{\beta}w \pm S\sqrt{2\left(\frac{1}{n}+\frac{(\bar{x}-w)^2}{ns_x^2}\right)F_{2,n-2}(\alpha)} \qquad \text{for all } w.$$

On the other hand, if we want a confidence interval for $\alpha+\beta w$ for a fixed w, then we can do better using our t-test:

$$\alpha+\beta w = \hat{\alpha}+\hat{\beta}w \pm S\sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-w)^2}{ns_x^2}\right)}\, t_{n-2}^{(\alpha/2)}.$$

You should check the details of this computation. □

6. Prediction intervals

The difference between confidence and prediction intervals is this: For a confidence interval we try to find an interval around the **parameter** $x'\beta$. For a prediction interval we do so for the **random variable** $y_0 := x_0'\beta + \varepsilon_0$, where x_0 is known and fixed and ε_0 is the "noise," which is hitherto unobserved (i.e., independent of the vector Y of observations).

It is not hard to construct a good prediction interval of this type: Note that

$$\hat{y}_0 := x_0'\hat{\beta}$$

satisfies

$$\hat{y}_0 - y_0 = x_0'\left(\hat{\beta} - \beta\right) - \varepsilon_0 \sim \mathrm{N}\left(0\,, \sigma^2\left[x_0'(X'X)^{-1}x_0 + 1\right]\right).$$

Therefore, a prediction interval is

$$\hat{y}_0 \pm S\, t_{n-p}^{(\alpha/2)} \sqrt{x_0'(X'X)^{-1}x_0 + 1}.$$

This means that

$$\mathrm{P}\left\{ y_0 \in \hat{y}_0 \pm S\, t_{n-p}^{(\alpha/2)} \sqrt{x_0'(X'X)^{-1}x_0 + 1} \right\} = 1 - \alpha,$$

but note that both y_0 and the prediction interval are now random.