

**Elab 14.21**

[F2007/F2007-136]

Fitting a model to a given sample of data means finding a coefficient on each model vector. If we were to collect a new sample of data, the fitted coefficient would likely be somewhat different. The **standard error** quantifies the meaning of “somewhat.” If the standard error is small, the coefficient for the new sample is likely to be close to the coefficient for the original sample. If the standard error is large, the coefficients for the new sample and the original sample are likely to be very different.

The size of the standard error is affected by several aspects of the data and the model:

- The number of cases in the sample,  $n$ . The more cases, the smaller the standard error.
- The size of the residual. The larger the residual, the larger the standard error.
- Collinearity among the model vectors. The more collinear are the model vectors, the larger the standard error. The smallest standard error occurs when the model vectors are mutually perpendicular.

We’re going to explore how the standard error depends on these aspects by using a simulation. The program `explore.collinear` generates two random variables,  $A$  and  $B$ , that are related by  $A \sim 1+B$ . In generating  $A$  and  $B$ , the `explore.collinear` program takes the values you specify of

**n** the number of cases.

**sd.resid** the standard deviation of the residual vector.

**theta** the angle, in degrees, between  $B$  and the  $\mathbf{1}$  vector. This angle describes the collinearity of the two explanatory vectors.

Your job is to see how the standard error depends on  $N$ , `sd.resid` and `theta`. You can do this by generating some data, fitting the model  $A \sim 1+B$ . For example, here is a very small data set generated with  $n=4$ ,  $\theta=90$  degrees, and `sd.resid=1`.

```
> f = explore.collinear(n=4, theta=90, sd.resid=1)
```

```
> f
      A      B
1 -0.4092258 -1.2565542
2  4.6994583  1.2667682
3  2.6245651 -0.6439801
4  5.0852025  0.6337660
```

You can see directly that  $n=4$ . To see that  $B$  is perpendicular to the  $\mathbf{1}$  vector (that is, that the angle is  $\theta=90$ ), try

```
> angle(f$B, c(1,1,1,1), degrees=TRUE)
[1] 90
```

By fitting a model you can see the standard error and the size of the residuals:

```
> mod = lm( A ~ 1+B, data=f)
> sd(mod$resid)
[1] 1
> summary(mod)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000     0.6124   4.899  0.0392 *
B              2.0000     0.6124   3.266  0.0823 .
```

**Dependence of standard error on sample size N.** Using the same values of  $\theta=90$  and  $\text{sd.resid}=1$ , change the value of  $N$  and observe how the standard error changes. Try  $n=100$  and  $n=400$  and  $n=2500$ . Use your observations to answer the following questions:

Comparing the standard error for  $n=400$  to that for  $n=100$ , the standard error for  $n=400$  is how much bigger than for  $n=100$ ?

- A Same size
- B About half as big
- C About one quarter as big
- D About twice as big
- E About four times as big

Elab 14.21-1

Comparing the standard error for  $n=2500$  to that for  $n=100$ , the standard error for  $n=2500$  is how much bigger than for  $n=100$ .

- A Same size
- B About one fifth as big
- C About one twenty-fifth as big
- D About five times as big
- E About twenty-five times as big

Elab 14.21-2

Now let's generalize. Suppose we are comparing the standard errors for two different size samples,  $n_1$  and  $n_2$ . The standard error for the sample of size  $n_2$  compared to the standard error for the sample of size  $n_1$  will be what? (Note, pay careful attention to the order of the subscripts.)

- A Same size
- B About  $\frac{\sqrt{n_2}}{\sqrt{n_1}}$  as big
- C About  $\frac{n_2}{n_1}$  as big
- D About  $\frac{\sqrt{n_1}}{\sqrt{n_2}}$  as big
- E About  $\frac{n_1}{n_2}$  as big

Elab 14.21-3

In words, which statement is correct?

- A The standard error doesn't depend much on n.
- B The standard error gets bigger as n gets bigger.
- C The standard error get smaller as n gets bigger.

Elab 14.21-4

Which of these statements is closest to being correct:

- A The standard error is independent of n.
- B The standard error scales with n.
- C The standard error scales with  $\frac{1}{n}$ .
- D The standard error scales with  $\sqrt{n}$ .
- E The standard error scales with  $\frac{1}{\sqrt{n}}$ .

Elab 14.21-5

**Dependence of standard error on residual size.** Now you are to explore how the size of the standard error depends on the size of the residual. As a point of comparison, use  $n=100$ ,  $\theta=90$ , and  $sd.resid=1$ :

```
> f = explore.collinear(n=100, theta=90, sd.resid=1)
> mod = lm( A ~ 1+B, data=f)
> summary(mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.0000	0.1005	29.85	<2e-16 ***
B	2.0000	0.1005	19.90	<2e-16 ***

Repeat this with  $sd.resid=2$ ,  $sd.resid=4$ , and  $sd.resid=10$ . Based on your observations which statement is true?

- A The standard error is independent of  $sd.resid$ .
- B The standard error increases proportionally to  $sd.resid$ .
- C The standard error is inversely proportional to  $sd.resid$ .
- D The standard error increases proportional to the square root of  $sd.resid$ .
- E The standard error is inversely proportional to the square root of  $sd.resid$ .

Elab 14.21-6

**Dependence of standard error on collinearity.** When there are two model vectors, 1 and B in the output of `explore.collinear`, the standard error depends on the angle  $\theta$  between them. When  $\theta$  is small, that is, when the two model vectors are closely aligned, the standard error is big. When  $\theta$  is near 90 degrees, that is, when the two model vectors are almost perpendicular, the standard error is small.

As basis for comparison, use the standard error when  $n=100$ ,  $\theta=90$ , and  $sd.resid=1$ . Compare this to the standard error when  $\theta = 45$ ,  $\theta = 20$ ,  $\theta = 10$ , and  $\theta=5$ . Based on your observations, which of the following statements is true?

**A** The standard error changes only a little with theta.

**B** The standard error is greatly inflated for small theta.

Elab 14.21-7

**C** The standard error is greatly inflated for theta near 45 degrees.

If you are mathematically inclined, you can verify that the standard error is proportional to  $\frac{1}{\sin(\theta)}$  for any given  $n$  and  $\text{sd.resid}$ . (Remember that the `sin` function in R expects the angle to be in radians. To convert an angle  $\theta$  in degrees to an angle in radians, multiply  $\theta$  by the conversion factor  $\pi/180$ .)

**The complete expression** We can put all of the above together into one formula. There are two standard errors, one for each of the two explanatory model vectors. For each of the two vectors, the standard error involves the lengths of the residual and the vector itself, the number of cases  $n$ , and the angle between the two explanatory vectors.

$$\frac{1}{\sqrt{n-2}} \frac{|\text{resid}|}{|\text{vector}|} \frac{1}{\sin(\theta)}$$

Remember that  $|\text{resid}|$  is the square-root of the sum of squares of the residual, and  $|\text{vector}|$  is the square-root of the sum of squares of the explanatory model vector.

Notice that the standard error scales as  $\frac{1}{\sqrt{n-2}}$ . The 2 is just the count of model vectors.