AREC422 Notes

Youpei Yan

March 26, 2019

V. CI for predictions.

We have a model $y = \beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k+u$, we are interested in a point with $\tilde{x} = (c_1, c_2, ..., c_k)$
To get the predicted value is simple, we just need to plug in the numbers into the fitted line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k$. We get:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + ... + \hat{\beta}_k c_k$$

But, what is the standard error of $\hat{\theta}_0$ (and the corresponding CI)?

Step 1: Re-write $\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - ... - \beta_k c_k$
Step 2: Plug into $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$, now we have a new model:

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + ... + \beta_k(x_k - c_k) + u$$

The intercept is θ_0, the value we care about. We have a group of new regressors as well.

Step 3: Re-run the regression, we can get the estimate of $\hat{\theta}_0$ and its corresponding *se*.

e.g.:

$$\hat{y} = \underset{(0.1)}{5} + \underset{(0.2)}{2} x_1 - \underset{(0.15)}{3} x_2 - \underset{(0.7)}{5} x_3$$

The point we are interested in is $(1, 2, 3)$.
$\hat{\theta}_0 = 5 + 2 * 1 + 3 * 2 + 5 * 3 = 28$

We can also run the following regression to get the same estimates:
$lm(y \sim I(x_1 - 1) + I(x_2 - 2) + I(x_3 - 3))$

Say, the result we get from R is

$$\hat{y} = \underset{(0.1)}{28} + \underset{(0.2)}{2}(x_1 - 1) - \underset{(0.15)}{3}(x_2 - 2) - \underset{(0.7)}{5}(x_3 - 3)$$

Therefore, $se(\hat{\theta}_0) = 0.05$, and the 95% CI for the expected value at $(1, 2, 3)$ is $[28 \pm 0.05 \cdot 1.96]$

e.g.: With more complicated regressors by including interaction terms or quadratic terms:

$$\hat{y} = \underset{(0.075)}{1.493} + \underset{(0.00007)}{0.00149} x_1 - \underset{(0.00056)}{0.01382} x_2 - \underset{(0.01650)}{0.06088} x_3 + \underset{(0.00227)}{0.00546} x_3^2$$

The point we are interested in is $x_1 = 1200$, $x_2 = 30$, $x_3 = 5$
We run the following regression: $lm(y \sim I(x_1 - 1200) + I(x_2 - 30) + I(x_3 - 5) + I(x_3^2 - 25))$
Note that we have $(x_3^2 - 25)$, rather than $(x_3 - 5)^2$

We get the following results:

$$\hat{y} = \underset{(0.020)}{2.700} + \underset{(0.00007)}{0.00149}(x_1 - 1200) - \underset{(0.00056)}{0.01382}(x_2 - 30) - \underset{(0.01650)}{0.06088}(x_3 - 5) + \underset{(0.00227)}{0.00546}(x_3^2 - 25)$$

So, the 95% CI for the expected value under the interested point is $[2.700 \pm 1.96 \cdot 0.020]$

e.g.: If we also know the residual standard error $\hat{\sigma} = 0.560$ in the above example, which is reported in R. We can also find the prediction interval for the predicted value IN THE FUTURE.

The idea of this "in the future" expression can be explained in this example: if we have the model $yield \sim prec + temp + radiation$
Before: we'll find the 95% CI for the average yield given a specific characteristics of Xs.
Now: We'll find the 95% CI for any particular yield with these characteristics of Xs in the future.

The major difference is that the latter should be wider because it will include unobserved characteristics in the error term that also affect *yield*. Basically, the latter one is harder to be precisely predicted.

The way to get this wider CI is to include RSE:
Now $se(\hat{e}) = \sqrt{se(\theta_0)^2 + \hat{\sigma}^2} = \sqrt{(0.020)^2 + (0.560)^2} \approx 0.561$

Therefore, the 95% CI for the predicted value in the future is $[2.70 \pm 0.561 \cdot 1.96]$

Qualitative Information in MRA

I. Binary variable/dummy variable/0-1 variable

It represents an on-off "switch" type of situation with 2 values. For instance: gender, employ or not, graduate or not, rain or not, install a technology or not, etc.

For instance, we may have the following dataset:

firm	*pollution*	*production*	*near_water*	*near_resi*	*_abate*
1	3.10	7.13	0	1	1
2	3.24	10.25	1	1	1
3	6.00	18.92	1	0	0
4	5.30	15.74	0	0	1

....

These dummy variables (the last three) describe a relative/raw comparison. We can treat them the same as regressors, but they have different coefficient interpretations.

II. Dummy Independent Variable

* Single Dummy Indep. Var.

e.g.1: $deer\_density = \beta_0 + \beta_1 forest\_area + \delta_0 hunt\_area + u$. $hunt\_area = 1$ if this is a hunting area, $hunt\_area = 0$ otherwise.

δ_0: Difference in the density of deer between a hunting area and a non-hunting area, given the same area of forest.