AREC422 Notes

Youpei Yan

March 15, 2019

On March 12, we went through the midterm questions plus the first two parts of some further issues under the multiple regression analysis (MRA). On March 14, we covered the middle two parts of these issues.

MRA: Further issues.

I. Scaling:

$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$, c is some constant number.

change	β_j	other β_i	t/F-test	$se(\beta_j)$	R^2
$x_j \cdot c$	β_j/c	β_i	same	$se(\beta_j)/c$	same
$y \cdot c$	$\beta_j \cdot c$	$\beta_i \cdot c$	same	$se(\beta_j) \cdot c$	same

Basically, "playing" data will not change the REAL impact.

Note that in R, we need to use the $I()$ function.
For instance, if we want to scale x_2 by 10, the old model is $lm(y \sim x_1 + x_2)$, and the new model is $lm(y \sim x_1 + I(x_2 * 10))$.

II. Beta Coefficients:

Standardize a variable is to re-scale it to the z-score, which has a mean of 0 & a sd of 1.

$$z = \frac{x - \bar{x}}{sd(x)}$$

If the old model is $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i$, standardize each of the variable, the regression becomes

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\beta_1 \frac{x_{i1} - \bar{x}_1}{\hat{\sigma}_1} + ... + \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\beta_k \frac{x_{ik} - \bar{x}_k}{\hat{\sigma}_k} + \frac{u_i}{\hat{\sigma}_y}$$

$$i.e. : z_y = \hat{b}_1 z_1 + ... + \hat{b}_k z_k + e$$

$\hat{b}_j$ is called the beta/standard coefficients. The interpretation is that: x_j increases by 1 sd, $\hat{y}$ changes by $\hat{b}_j$ sd. So we can compare coefficients:

	previous model	beta model
measure effects	in original units	in sd units
magnitudes	large coeff. $\neq$ large impact	large coeff. represents large impact

e.g., $z(\widehat{pollution}) = -0.2 z_{fine} - 0.3 z_{inv} + 0.15 z_{prod} + ...$ then:
pollution fine increases by 1 sd, pollution level is decreasing by 0.2 sd; abatment investment increases by 1 sd, pollution level is decreasing by 0.3 sd. So, same <u>relative movement</u> of "inv" has a larger effect on pollution reduction than "fine".

However, it is hard to compare variables' impacts using the old model or the traditional way, because we can easily re-scale variables and change their units.

III. Functional forms

· **log-level model**.

$\widehat{log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1$: In this case, if x_1 increase by a units, y will change by
approximately $100\hat{\beta}_1 a\%$, or
exactly $100[exp(\hat{\beta}_1 a) - 1]\%$

For instance, if $\widehat{log(y)} = 9.23 - 0.052x$, then x increases by 1 unit, y is changing by
approximately $100\hat{\beta}_1 a\% = 100 \cdot (-0.052)\%$ = -5.2% , or
exactly $100[exp(\hat{\beta}_1 a) - 1]\% = 100[exp(-0.052) - 1]\%$= -5.1%

If x increases by 5 units, y is changing by
approximately $100\hat{\beta}_1 a\% = 100 \cdot (-0.052) * 5\%$ = -26% , or
exactly $100[exp(\hat{\beta}_1 a) - 1]\% = 100[exp(-0.052 * 5) - 1]\%$= -22.9%

If x decreases by 5 units, y is changing by
approximately $100\hat{\beta}_1 a\% = 100 \cdot (-0.052) * (-5)\% = 26\%$, or
exactly $100[exp(\hat{\beta}_1 a) - 1]\% = 100[exp(-0.052 * (-5)) - 1]\%$= 29.7%

Note, if Δx is larger, the difference between the approximated and the exact changes is larger.
Also, increasing x and decreasing x bring different absolute values in the exact form.

· **quadratic model**.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$. If $\hat{\beta}_1$ and $\hat{\beta}_2$ have different signs, we can have a turning point in x, which is:

$$x^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|$$

For instance, if $\widehat{yield} = 3 + 0.3irrigation - 0.05irrigation^2$, then $irrigation^* = 0.3/(2 * 0.05) = 3$ (kg). The corresponding interpretation will be: above 3kg, increasing *irrigation* will decrease *yield*, below 3kg, increasing *irrigation* will increase *yield*.

The other thing to notice is the non-constant marginal effect of x:

$$\frac{dy}{dx} = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

which is a linear function of x.

Similarly, if we have a log-log model with the quadratic term in, $log(price) = \beta_0 + \beta_1 log(pollution) + \beta_2 log(pollution)^2 + u$, then the elasticity of price with respect to pollution is also a linear function of $log(pollution)$:

$$\frac{dlog(price)}{dlog(pollution)} = \beta_1 + 2\beta_2 log(pollution)$$

· **interaction terms**.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$. Note that the partial effect of x_1 (or x_2) is a function of another variable:

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

Interpretation: one unit increase in x_1 gives a higher (if $\beta_3 > 0$) increase in y for higher x_2.

e.g.: $conversion = \beta_0 + \beta_1 subsidy + \beta_2 erosion + \beta_3 (erosion \cdot subsidy) + u$. We expect that all the βs are positive.

$$\frac{\partial conversion}{\partial subsidy} = \beta_1 + \beta_3 \cdot erosion$$

Interpretation: If $\beta_3 > 0$, an additional unit of payment gives a higher increase in land conversion for higher erosion level.

To get a general impression of the partial effect (rather than knowing it as a function of variables), we use the average partial effects (APE) to represent this average effect. We use the mean of the variable(s) in the partial effect function to calculate APE.
e.g.: $APE_{subsidy} = \hat{\beta}_1 + \hat{\beta}_3 \overline{erosion}$

IV. Adjusted R^2: $\bar{R}^2$

$\bar{R}^2$ is the corrected R^2, which imposes a penalty for adding more control variables.

Recall that R^2 is higher by adding a new independent variable, but $\bar{R}^2$ increases if and only if the $|t - stat|$ of the new variable is greater than 1.

Also, $\bar{R}^2$ can be negative (R^2 is between 0 to 1) if the model is a poor fit, and is higher if the model fits better.

We use $\bar{R}^2$ to compare non-nested models like $y \sim x_1 + x_2$ v.s. $y \sim x_1 + x_3$, or $y \sim log(x)$ v.s. $y \sim x + x^2$. (Recall that we use F-test to compare nested models like $y \sim x_1 + x_2$ vs. $y \sim x_1 + x_2 + x_3$)

V. CI for predictions. After the break.