Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Rui Zhou

Spring 2023

目录

6 MLR Analysis with Dummy

A way to incorporate qualitative information is to use dummy variables. They may appear as the dependent (Linear Probability Model) or independent variables.

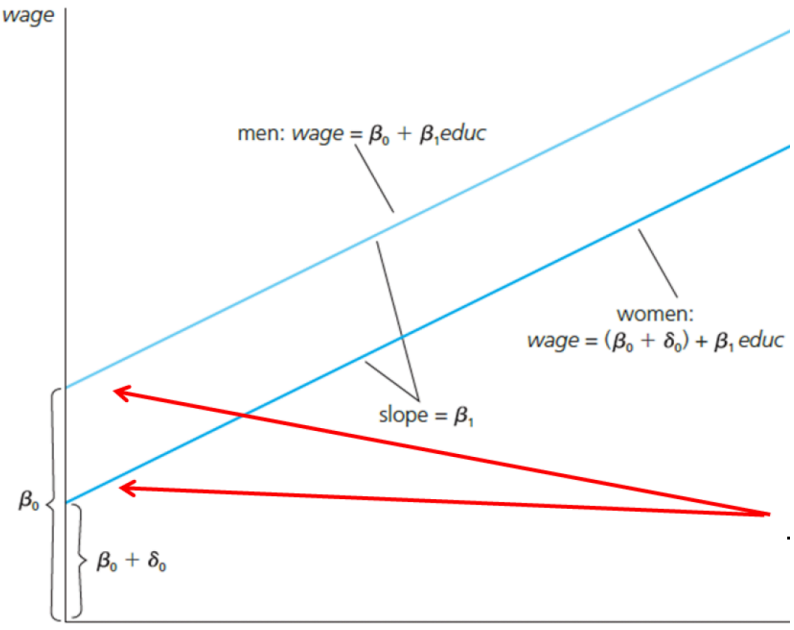
6.1 Different Intercepts: Dummy as Regressors

A Dummy is also called a 0-1 variable, or a binary variable. It does not matter whether to code the dummy variable under certain state as 0 or 1. Anyway, bear in mind what the base group (or, benchmark group) is!

Start with a simple example where $d_0 = 0$ implies an observation belongs to the base group, $d_0 = 1$ otherwise.

$$y = \beta_0 + \delta_0 d_0 + \beta_1 x_1 + u$$

For the base group, its intercept is β_0; for its counterpart, the intercept is $\beta_0 + \delta_0$. δ_0 represents an intercept shift on graphical illustration.



Generally speaking, the slope coefficient of the dummy variable indicates the difference in the mean of dependent variable across the two states, holding others factors constant.

6.1.1 Dummy Variable Trap

Mind yourself of Dummy Variable Trap! If the number of dummy variables describing one attribute equals or even exceeds its levels, then perfect collinearity creeps in!

Alternatively, one could include all dummies representing all levels and **omit the intercept term**,

$$y = \gamma_0 d_0 + \gamma_1 d_1 + \beta_1 x_1 + u$$

In this case, the estimated difference between groups is $|\gamma_1 - \gamma_0|$. However, in this method, we can only get the correct estimation for each coefficient, but the differences between the parameters (groups) are harder to be tested. Moreover, R^2 is invalid, since the intercept term is removed.

6.1.2 Incorporating Ordinal Information Using Dummies

To study ordinal variables, start with a case where we are interested in the relationship between city credit ratings (CR, integer ranging from 0 to 4) and municipal bond interest rates (MBR). One possible and straightforward way is to estimate

$$MBR = \beta_0 + \beta_1 CR + o.f.$$

which is the fixed partial effect model.

However, the differences between any two adjacent levels may not hold constant across all levels. They only stands for ordinal meanings, instead of interval meanings. Since most of the time ordinal variable only takes limited values, one better solution is to define a dummy for each level of the ordinal variable (Pay attention to dummy variable trap meanwhile!)

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_1 CR_2 + \delta_1 CR_3 + \delta_1 CR_4 + o.f.$$

Note that $CR = 0$ is the base group here. After introducing dummies to represent every possible levels of the ordinal variable, we allow for more possibilities of the effect of level change.

Fixed partial effect model is a special case. The constraint is $\delta_2 = 2\delta_1, \delta_3 = 3\delta_1, \delta_4 = 4\delta_1$. Plug in the constraint back to get the restricted model as

$$\begin{aligned} MBR &= \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + o.f. \\ &= \beta_o + \delta_1 CR + o.f. \\ \Longleftarrow \quad & H_0 : \delta_2 = 2\delta_1, \delta_3 = 3\delta_1, \delta_4 = 4\delta_1 \end{aligned}$$

Sometimes, the ordinal variable of interest takes too many values to include each level as a dummy. In this case, we can classifiy the ordinal variables into several categories by dividing them according to some ranges.

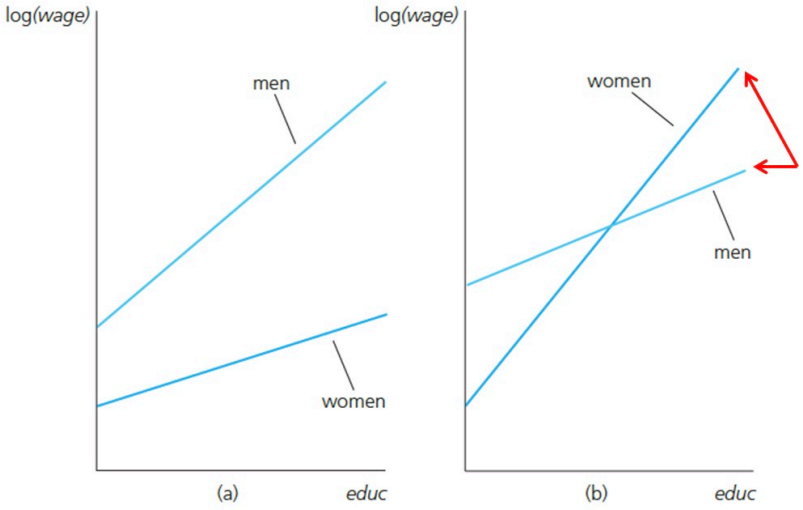
6.2 Different Slopes: Interaction with Dummy

6.2.1 Graphical Illustration

An interaction term of dummy variable(s) allows for different slopes between groups

$$y = \beta_0 + \beta_1 x_1 + \delta_0 d_0 + \delta_1 (d_0 \cdot x_1) + u$$

where $\beta_0 + \delta_0$. δ_0 represents an intercept shift on graphical illustration.



where δ_1 means the partial effect of x_1 on y in the base group and its counterpart differs for δ_1. If $\delta_1 = 0$, this means partial effect of x_1 is the same for both groups. If $\delta_0 = \delta_1 = 0$, this means there is no difference between groups, and the whole equation is the same! Also note that, if an interaction term is introduced into the model, be careful about the interpretation of β_1! β_1 equals to x_1's partial effect if, and only if $d_1 = 0$. Pay attention to the same issue with δ_0. In order to make δ_0 practically meaningful and reasonable, the common trick is to centralize x_1 to its mean.

6.2.2 Chow Test

From the simple case above, we have grasped the overall picture of detecting differences between groups using dummies and even interaction terms with dummies. In order to prove that the dummy classification has no effect (or difference between groups), we need to conduct a joint test about all slope coefficients for regressors with dummies. Take s step further, if we aim to test if the regression function differs for different groups, we need to introduce all interaction terms into the model (including interaction with intercept, i.e., dummy itself). This is practical if explanatory variables are not that much; however, if we have a lot of explanatory variables, we can turn to an alternative way to compute F-statistic. Chow-test is such a convenient way. Steps for Chow test follows as

1. Run separate regressions for the base group and its counterpart; the unrestricted SSR is given by the sum of the SSR of these two regressions. ($SSR_{ur} = SSR_1 + SSR_2$)

2. Run regression for the restricted model and store restricted SSR_p (**pooled** data this time).

3. Compute F statistics as

$$F = \frac{(SSR_p - SSR_{ur})/(k+1)}{SSR_{ur}/(n-2\cdot(k+1))} = \frac{[SSR_p - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{n-2(k+1)}{k+1}$$

   Note that the test assumes a constant error variances across groups.

The hypothesis for Chow Test is that, there is no difference between groups, including all the partial effects and the intercept. However, if we allow for the differences in intercepts across groups and then test the differences in slope coefficients, such test would be more meaningful. In this way, the number of constraints is k, instead of $k+1$.

Moreover, Chow Test can be applied to test structural change over time (including intercept change here). The principle is still the same as before, but this time we turn to dummies that represents time periods. First, for each period in the T periods of interest, conduct a regression and obtain its SSR. The unrestricted SSR_{ur} is then $SSR_{ur} = SSR_1 + SSR_2 + ... + SSR_T$. If there are k explanatory variables and T periods, we have to test for $(T-1)(k+1)$-many constraints. In the unrestricted model, there are $T(k+1)$ parameters to estimate. Suppose $n = n_1 + n_2 + ... + n_T$ is the total number of observations, then degrees of freedom for F statistic is $(T-1)(k_+1)$ and $n - T(k+1)$. Note again that Chow Test cannot tolerate with heteroskedasticity. If with heteroskadasticity, we have to actually construct all the interaction terms and conduct a regression (i.e., the direct method).

6.3 Binary Regressand: Linear Probability Model

When y is binary, our model turns out to be a linear probability model (LPM).

$$\begin{aligned}
& y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\
\Longrightarrow \quad & E(y|\vec{x}) = 1 \cdot P(y=1|\vec{x}) + 0 \cdot P(y=0|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \\
\Longrightarrow \quad & P(y=1|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \\
\Longrightarrow \quad & \beta_j = \frac{\Delta P(y=1|x)}{\Delta x_j}
\end{aligned}$$

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that $y = 1$.

One of the most obvious drawback of the linear probability model is that, all observed factors in the model are considered of ***linear*** effect. However, actually most of the influence may be nonlinear but based on its level. Other disadvantages of the linear probability model include

- Predicted probabilities may be larger than one or smaller than zero.
- Marginal probability effects sometimes logically impossible.
- The linear probability model is necessarily heteroskedastic. Thus, heteroskedasticity consistent standard errors need to be computed.

$$Var(y|\vec{x}) = P(y = 1|\vec{x})[1 - P(y = 1|\vec{x})]$$

Clearly, the variance of y depends on x. However, in the classical linear model's assumptions (CLM), the variance should have nothing to do with $\vec{x}$.

6.4 Problem of Self-Selection

Consider a PRF as

$$E(y|w, x) = \alpha + \tau w + \gamma_1 x_1 + ... + \gamma_k x_k$$

where w is a treatment indicator. $w = 1$ when the treatment has been applied. Denote $y(0)$ as the outcome of y when $w = 0$, $y(1)$ when $w = 1$, then

$$y = (1 - w)y(0) + wy(1)$$

We include x_1 through x_k to account for the possibility that the treatment w is not randomly assigned and clean of self-selection problem (to get closer to random assignment into the treament and control group).

Now we need to make strong assumption that w is independent of $[y(0), y(1)]$ conditional upon $x_1, ..., x_k$. This is known as **regression adjustment** and allows us to adjust for differences across units in estimating the causal effect of the treatment.

We can relax the assumption of a constant treatment effect. We allow the treatment effect to vary across observations and estimate the average treatment effect (ATE).

$$y_i = \alpha + \tau w_i + \gamma_1 x_{i1} + ... + \gamma_k x_{ik} + \delta_1 w_i(x_{i1} - \bar{x}_1) + ...{}_+ \delta_k w_i(x_{ik} - \bar{x}_k) + u$$

where the estimated coefficient on w will be the ATE. The regression that allows individual treatment effects to vary is known as the unrestricted regression adjustment (URA). URA more is closer to the reality, since the individual effects won't be homogenious for all. By contrast, a restricted regression adjustment (RRA) forces the treatment effect to be identical across individuals.

Note that tt doesn't matter whether $\gamma_j x_{ij}$ is centralized, but it matters whether interaction terms are centralized, otherwise the interpretation of τ is affected.

Here is an alternative method for obtaining the URA ATE. For control group and treatment group, estimate the following equation respectively

$$
\begin{aligned}
& y_i = \alpha + \gamma_1 x_{i1} + ... + \gamma_k x_{ik} \\
\Longrightarrow \quad & \begin{cases} \text{Control: } \hat{y}_i^{(0)} = \hat{\alpha} + \hat{\gamma}_{0,1} x_{i1} + ... + \hat{\gamma}_{0,k} x_{ik}, \text{ using control observations} \\ \text{Treatment: } \hat{y}_i^{(1)} = \hat{\alpha} + \hat{\gamma}_{1,1} x_{i1} + ... + \hat{\gamma}_{1,k} x_{ik}, \text{ using treatment observations} \end{cases}
\end{aligned}
$$

Now for every unit in the sample, predict $y_i(0)$ and $y_i(1)$ regardless of whether the unit is in the control of treatment groups. Then use these predicted values to compute the ATE as

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{(1)} - \hat{y}_i^{(0)}))$$

Though this yields the same ATE as running the regression with interaction terms, computing a standard error by hand can be tricky.