L6. Deal with Categorical Predictors.

— Dummy variables in MLR

So far we have only dealt with Numeric variables: just through in the model directly, then $\vec{b} = (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$

can be plug in directly with the predictor values.

What if we have categorical predictors?

GENDER: male female

RACE: white, asian, black, ...

COLOR: red, green, blue.

we are collecting this kind of information from observations. Whose the values can be k different levels, which are not numeric. We need to change it to be "NUMBERS" so we can perform the analysis we have used.

- Dummy Variables. – MOST COMMON.

The most common way to "recode" categorical variable is to define dummy variables.:

$$X_i = \begin{cases} L_1 & \text{when obs } i \text{ is in category } L_1 \\ L_2 & \text{when obs } i \text{ is in category } L_2 \\ L_3 & \vdots \\ \vdots & \\ L_K & \end{cases}$$

$$X_{i\_L1} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in } L_1 \\ 0 & \text{o.w. } (L_2, L_3, \cdots, L_K) \end{cases}$$

$\Longleftrightarrow$

$$X_{i\_L2} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in } L_2 \\ 0 & \text{o.w. } (L_1, L_3, \cdots, L_K) \end{cases}$$

$\vdots$

$$X_{i\_L_{K-1}} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in } L_{K-1} \\ 0 & \text{o.w. } (L_1, \cdots, L_{K-2}, L_K) \end{cases}$$

X_{i1K} doesn't need to define: If $X_{i\_L1} = \cdots = X_{i\_L_{K-1}} = 0$

then i^{th} obs is in L_K.

For ex : eye color of 5 obs.

Obs	Color		Color_Brown	Color_Blue
1	Brown	Dummy coding →	1	0
2	Blue		0	1
3	Blue		0	1
4	Green		0	0
5	Brown		1	0

Green.

One categorical variable with k levels will be recoded as (k-1) dummy variables. then the regression fitted is

$$y_i \sim \underbrace{X_1 \alpha_1 + X_1 \alpha_2 + \cdots + X_1 \alpha_{k-1}}_{\text{original } X_1} + X_2 + X_3 + \cdots + X_{p-1}$$

1. How to read the estimated coefficient ?

Again, use the eye-color example. the fitted line would look like :

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} * \text{Color\_brown} + \hat{\beta_2} * \text{Color\_blue}$$

$\hat{\beta}_1$: when color_brown = 1.

① $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * \text{Color\_Brown}$. (Color_blue = 0 when Color_Brown = 1)

compared to

② $\hat{y}_i = \hat{\beta}_0$ (color_blue = 0 and Color_Brown = 0, AKA. Color = Blue)

① − ② = $\hat{\beta}_1$: measures the change in y when color changes from Blue (Baseline) to Brown (Color_Brown)

Summary: the coef. of X_{-Lj} measures the change caused in y when the obs is in the Level Lj compared to the baseline level.

Python: coef for fuel-type [gas] = −5.6297.

means the cars with gas fuel-type has, on average, 5.6297 less city mpg than diesel type.

2. How to read t-test ?

t test on $H_0: \beta_{4j} = 0$ v.s. $H_1: \beta_{4j} \neq 0$

Rejection indicates there is a significant difference in $y|_{4j}$ compared to $y|_{4\text{baseline}}$.

As long as one test for the ~~~~ dummy variables is significant, we can conclude the original categorical variable is significant to y.

Ex: Python: citympg ~ drive-wheels

	p-value of t test
drivewheels_fwd :	0.006
drivewheels_rwd:	0.185

Indicates there's significan (positive) change in city_mpg from baseline (4wd) to fwd.

No sig. chang from baseline (4 wd) to rwd.

In general. drive-wheels has significant impact on citympg.

3. How to read ANOVA? Both typ=1 and 2 shows the original categorical variable, instead of dummies.

type = 1.

	df	SS
Drivewheels	2	~~SS~~
fuel-type	1	
enginesize	1	
Residuals	n-5.	

Reduced: $y_i = \beta_0 + \varepsilon_i$

Full: $y_i = \beta_0 + \beta_1 * \text{Drivewheels\_fwd} + \beta_2 * \text{Drivewheels\_rwd} + \varepsilon_i$

In Full model: β_0, two coeffs with drivewheels, 1 coefficient with fuel-type, 1 coeff with enginesize. Total of 5 parameters estimated.

type = 2.

	df	SS.
Drivewheels	2	
fuel-type	1	
enginesize	1	
Residuals	n-5	

Reduced: $y_i = \beta_0 + \beta_3 * \text{fuel-type\_gas} + \beta_4 * \text{enginesize} + \varepsilon_i$

Full: $y_i = \beta_0 + \beta_1 * \text{Drivewheel\_fwd} + \beta_2 * \text{Drivewheel\_rwd} + \beta_3 * \text{fuel\_type\_gas} + \beta_4 * \text{enginesize} + \varepsilon_i$

Note: There are cases that you would want to treat the predictor as categorical, but the data was either recorded with numbers:

① Color

1
2
2
3
3
4
⋮

where 1 = "red"
2 = "green"
3 = "blue"
..

Then we need to do smf.ols('y ~ C(x)')

Change x into categorical

Otherwise it would be analysized as numeric variable

② Or for an ordinal variable, change it into dummies can give major information.

Age_group

1 <18
2 18≤ <30
3 30≤ <40
4 40≤ <50
5 ≥50

the levels comes with numerical order. Change it into dummies can see changes between levels.

Note: Potential problem with dummies.

The information was collected in one categorical variable, but in the model we need to estimate (k-1) parameters, AKA, we are adding (k-1) predictors in the model, that could cause two problems:

① Multicolinearity (Later)

② Not enough data to support a precise estimation: we all know that more data can have a more accurate estimate.

In Regression, the $\frac{\text{Number of parameters}}{\text{Number of observation}} \geq 1/n \geq 20$

(More details can be decided in "power analysis, not covered!"