Categorical Variables Transformations Interactions

GR 5205 / GU 4205
Section 3

Columbia University
Xiaofei Shi

Categorical predictors

- Different states, different groups categories factor
- Different level of treatment ordered categories ordered

Simplest case: binary categories $X_1 \in \{0, 1\}$

- Also sometimes called indicator variables or dummy variables.
- Usually we code them as qualitative categories with 0 or 1.

LRM with categorical predictors

- Different intercept
- Different slope
- Different slope and different intercept

Categorical variables with more than two levels

- One-hot representation vs 1,2,...k

  - design the value based on your needs

  - usually the one-hot representation is preferable

  - make sure only introduce k-1 columns to avoid multicollinearity

- Other thoughts?

Simple factor models

- k-number of levels
- n-number of experiments repeated in each level

Summary of one-way ANOVA

- k-number of levels
- n-number of experiments repeated in each level

Source	Sum of Squares	df	Test statistics
Between	SS(B)=SS(full)	k-1	$F = \frac{SS(B)/(k-1)}{SS(W)/k(n-1)}$
Within	SS(W)	k(n-1)	
Total	SS(B)+SS(W)=SS(reduced)	nk-1	

Single factor and one-way ANOVA

Data Table

Drug Dose	Libido					Sample Size	Sample Means	Sample Variance
Placebo (k_1)	3	2	1	1	4	5 (n_1)	2.2 ($\bar{x}_1$)	1.7 (s_1^2)
Low (k_2)	5	2	4	2	3	5 (n_2)	3.2 ($\bar{x}_2$)	1.7 (s_2^2)
High (k_3)	7	4	5	3	6	5 (n_3)	5.0 ($\bar{x}_3$)	2.5 (s_3^2)
Total ($k = 3$)						15 (n_T)	3.5 ($\bar{x}$)	3.1 (s^2)

The model with the same slopes

$$Y = \beta_0 + x^{(1)}\beta_1 + x^{(2)}\beta_2 + \epsilon$$

- For $x^{(1)} = 0, \qquad Y = \beta_0 + x^{(2)}\beta_2 + \epsilon$

- For $x^{(1)} = 1, \qquad Y = \beta_0 + \beta_1 + x^{(2)}\beta_2 + \epsilon$

Two parallel regression lines for different category

Two factor models

- k-number of levels for factor A
- s-number of levels for factor B
- n-number of experiments repeated in each level

Summary of two-way ANOVA

- k-number of levels for factor A
- s-number of levels for factor B
- n-number of experiments repeated in each level

<table>
<tr><th>Source</th><th>Sum of Squares</th><th>df</th><th>Test statistics</th></tr>
<tr><td>Factor A</td><td>SS(A)</td><td>k-1</td><td>$F_A = \frac{SS(A)/(k-1)}{SS(R)/ks(n-1)}$</td></tr>
<tr><td>Factor B</td><td>SS(B)</td><td>s-1</td><td>$F_B = \frac{SS(B)/(s-1)}{SS(R)/ks(n-1)}$</td></tr>
<tr><td>Factor A&B</td><td>SS(A&B)</td><td>(k-1)(s-1)</td><td rowspan="3">$F_{A\&B} = \frac{SS(A\&B)/(k-1)(s-1)}{SS(R)/ks(n-1)}$</td></tr>
<tr><td>Residual</td><td>SS(R)</td><td>sk(n-1)</td></tr>
<tr><td>Total</td><td>SS(Total)</td><td>nks-1</td></tr>
</table>

Two factors and two-way ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Seed	512.8667	2	256.4333	28.283	0.000008
Fertilizer	449.4667	4	112.3667	12.393	0.000119
Interaction	143.1333	8	17.8917	1.973	0.122090
Within	136.0000	15	9.0667		
Total	**1241.4667**	**29**			

The model with the same intercept

$$Y = \beta_0 + x^{(1)} x^{(2)} \beta_1 + x^{(2)} \beta_2 + \epsilon$$

- For $x^{(1)} = 0,$ $\qquad Y = \beta_0 + x^{(2)} \beta_2 + \epsilon$
- For $x^{(1)} = 1,$ $\qquad Y = \beta_0 + x^{(2)} (\beta_2 + \beta_1) + \epsilon$

Two regression lines with different slopes but same intercept for different categories

R & Python

- R: factor(x); anova(full_model, reduced_model)

- Python: design your own design matrix x

  use ols and anova_lm() function in statsmodel

Any other thoughts?

- Higher order effects: the model gets more complicated easily!

  For example, in a model with 3 factors, the full model can be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \epsilon.$$

- Want both different slope and different intercept: separate the data!

Takeaways

- ANOVA test:

  Reduced model vs Full model

- How to model the linear dependence wrt categorical variables