Lab 9

Wiktor Soral

May 8th 2017

Regression with categorical variables as predictors

- None of the regression assumptions states that predictors have to be continuous.
- On the contrary, categorical predictors are frequently used in regression models, to account for, e.g. gender, level of education, ethnicity.
- Evenmore, when it comes to analysis of experimental data, regression analysis is a powerful alternative to ANOVA.
- However, using categorical predictors requires some additional attention.

Regression with binary predictor

- The simplest case is a regression with one binary predictor, e.g. gender.
- In such a case we have to code such a variable as numeric, but specific values are somewhat arbitrary, e.g. we could code males as 1 and females as 2.
- However, proper coding can make it easier to interpret the results.
- E.g. we can code one gender (lets say males) as 0, and another (females) as 1.

Regression with one binary predictor - dummy coding

$BMI_i = \beta_0 + \beta_1 * gender_i$; 0 = males; 1 = females

- For males:

$BMI = \beta_0 + \beta_1 * 0$

$BMI = \beta_0$

- For females:

$BMI = \beta_0 + \beta_1 * 1$

$BMI = \beta_0 + \beta_1$

Regression with one binary predictor - effect coding

$BMI_i = \beta_0 + \beta_1 * gender_i$; -1 = males; 1 = females

- For males:

$BMI = \beta_0 + \beta_1 * -1$

$BMI = \beta_0 - \beta_1$

- For females:

$BMI = \beta_0 + \beta_1 * 1$

$BMI = \beta_0 + \beta_1$

$\beta_0 = grand\ mean$

Regression with one binary predictor - weighted effect coding

- $\beta_0 = grand\ mean$ only if number of observations - n - in category 1 is equal to n in category 2
- With unequal category sizes, e.g. $n_1 = 150$ and $n_1 = 100$ one can use weighted effect coding scheme
- Category 1 is coded as $-n_2/n_1 = -100/150 = -0.6667$ and category 2 is coded as 1
- With this coding scheme, β_0 indicates overall sample mean, and β_1 indicates deviance of group mean from overall sample mean
- This coding accounts for unequal sample size

Regression with nominal predictors with n levels

- Variables with more than 2 levels require additional attention
- We just cannot put nominal variables in our regression equation, as if it were continuous predictor - it doesn't make sense.
- However we can code variable with $n > 2$ levels, with $n - 1$ instrumental variables

Regression with nominal predictors with n levels - dummy coding

	$d1$	$d2$
$level\ 1$	0	0
$level\ 2$	1	0
$level\ 3$	0	1

E.g. BMI and education (lowest, middle, highest)

$BMI_i = \beta_0 + \beta_1 * d1_i + \beta_2 * d2_i;$

Regression with nominal predictors with n levels - dummy coding

Level 1: $BMI = \beta_0 + \beta_1 * 0 + \beta_2 * 0$

$BMI = \beta_0$

Level 2: $BMI = \beta_0 + \beta_1 * 1 + \beta_2 * 0$

$BMI = \beta_0 + \beta_1$

Level 3: $BMI = \beta_0 + \beta_1 * 0 + \beta_2 * 1$

$BMI = \beta_0 + \beta_2$

Regression with nominal predictors with n levels - effect coding

	$e1$	$e2$
$level\ 1$	-1	-1
$level\ 2$	1	0
$level\ 3$	0	1

E.g. BMI and education (lowest, middle, highest)

$BMI_i = \beta_0 + \beta_1 * e1_i + \beta_2 * e2_i;$

Regression with nominal predictors with n levels - effect coding coding

Level 1: $BMI = \beta_0 + \beta_1 * -1 + \beta_2 * -1$

$BMI = \beta_0 - \beta_1 - \beta_2$

Level 2: $BMI = \beta_0 + \beta_1 * 1 + \beta_2 * 0$

$BMI = \beta_0 + \beta_1$

Level 3: $BMI = \beta_0 + \beta_1 * 0 + \beta_2 * 1$

$BMI = \beta_0 + \beta_2$

Regression with nominal predictors with n levels - weighted effect coding

	$we1$	$we2$
$level\ 1$	$-n_2/n_1$	$-n_3/n_1$
$level\ 2$	1	0
$level\ 3$	0	1

E.g. BMI and education (lowest, middle, highest)

$BMI_i = \beta_0 + \beta_1 * we1_i + \beta_2 * we2_i;$

Regression with nominal predictors with n levels - weighted effect coding coding

Level 1: $BMI = \beta_0 - \beta_1 * n_2/n_1 - \beta_2 * n_3/n_1$

Level 2: $BMI = \beta_0 + \beta_1 * 1 + \beta_2 * 0$

$BMI = \beta_0 + \beta_1$

Level 3: $BMI = \beta_0 + \beta_1 * 0 + \beta_2 * 1$

$BMI = \beta_0 + \beta_2$

Regression with nominal predictors with n levels - orthogonal (contrast coding)

	$o1$	$o2$
$level\ 1$	-0.5	0
$level\ 2$	0.25	-0.5
$level\ 3$	0.25	0.5

E.g. BMI and education (lowest, middle, highest)

$BMI_i = \beta_0 + \beta_1 * o1_i + \beta_2 * o2_i;$

Regression with nominal predictors with n levels - weighted effect coding coding

Level 1: $BMI = \beta_0 - \beta_1 * -0.5 - \beta_2 * 0$

$BMI = \beta_0 - 0.5\beta_1$

Level 2: $BMI = \beta_0 + \beta_1 * 0.25 + \beta_2 * -0.5$

$BMI = \beta_0 + 0.25\beta_1 - 0.5\beta_2$

Level 3: $BMI = \beta_0 + \beta_1 * 0.25 + \beta_2 * 0.5$

$BMI = \beta_0 + 0.25\beta_1 + 0.5\beta_2$