Statistics 2

Regression Modeling with Categorical Predictors: Code Variables

Casper Albers & Jorge Tendeiro

Lecture 8, 2019 – 2020

university of groningen

Overview

Literature for this lecture

Read:

Agresti, Section 12.1

Context

- So far, we only used continuous predictors in multiple regression.
- The multiple regression model can be extended to incorporate categorical predictors.

Example

- Gender (e.g., male, female, other).
- Political party (democrat, independent, republican).
- Clinical trial (new treatment, standard treatment, placebo).

Today we learn how to incorporate this type of predictors into multiple regression models.

Example – Preventing flashbacks

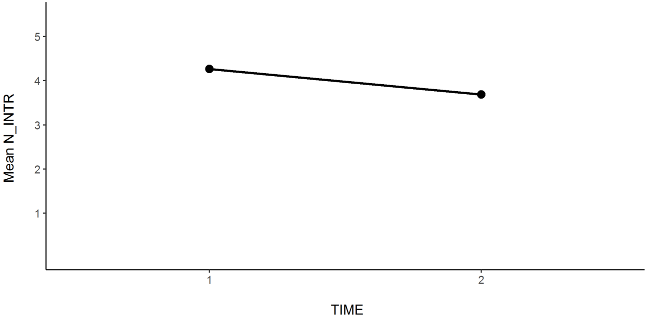
James et al. (2015) studied whether playing a computer game (Tetris) could prevent intrusive memories (flashbacks) related to a traumatic event from occurring, via a reactivation-reconsolidation mechanism[1].

- Dependent variable
  $y = \text{N\_INTR} =$ Number of intrusive memories over the next seven days
- Independent variables
  - TIME, with two levels:
    - 1 = morning
    - 2 = afternoon
  - CONDITION, with four levels:
    - 1 = No-task control
    - 2 = Reactivation + Tetris
    - 3 = Tetris only
    - 4 = Reactivation only

[1] James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., & Holmes, E. A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological Science*, *26*, 1201-1215. doi: 10.1177/0956797615583071

Example – Preventing flashbacks (TIME)

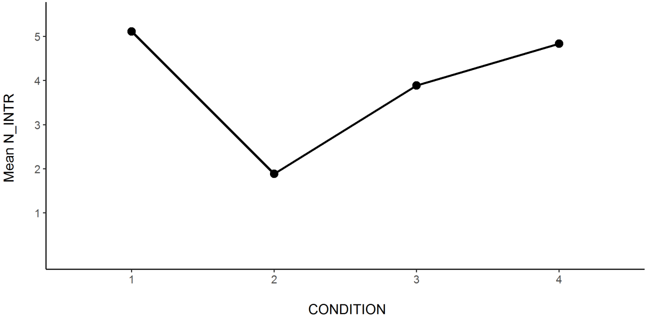
	TIME	
	1	2
Valid	30	42
Mean	4.267	3.690
Std. Deviation	3.352	3.382



Question: Is there an effect of TIME on N_INTR?
Equivalently, how can we regress N_INTR on TIME?

Example – Preventing flashbacks (N_INTR)

	N_INTR			
	1	2	3	4
Valid	18	18	18	18
Mean	5.111	1.889	3.889	4.833
Std. Deviation	4.227	1.745	2.888	3.330



Question: Is there an effect of CONDITION on N_INTR?

Equivalently, how can we regress N_INTR on CONDITION?

Code variables – Two groups

Code variable = An *artificial* variable indicating group membership.

A categorical variable with two levels requires one code variable with two possible values.

Example (coding provided in data set):

$$z_i = \begin{cases} 1 & \text{if TIME} = \text{MORNING} \\ 2 & \text{if TIME} = \text{AFTERNOON} \end{cases}$$

- Any two different values could be used as codes:
  The test of the effect is invariant to the coding used.
- However, the interpretation of the effect (e.g., via regression coefficients or confidence intervals) does depend on the coding used.
- The most common coding system uses 0s and 1s:
  Dummy coding system.

$$z_i = \begin{cases} 0 & \text{if TIME} = \text{MORNING} \\ 1 & \text{if TIME} = \text{AFTERNOON} \end{cases}$$

Let's entertain the idea of regressing the DV on the code variable z and see where that takes us:

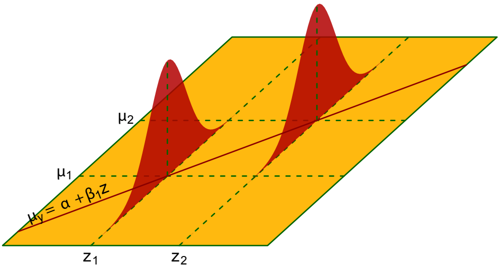
$$\text{Population regression line: } \mu_y = \alpha + \beta_1 z.$$

- The regression model describes how the population mean of y, μ_y, depends on the values of z (say, z_1 and z_2).
- The values of z define two subpopulations (e.g., morning and afternoon for predictor TIME):
  - In both groups y is normally distributed.
  - Means: μ_1 and μ_2.
  - Constant σ.

Code variables – Two groups

Continuous IV: Many subpopulations defined by the IV.

Code variable: Two subpopulations.



Code variables – Two groups

$$\boxed{\mu_y = \alpha + \beta_1 z}$$

Two groups defined by the values of z:

$$z_i = \begin{cases} z_1 & \text{if person } i \text{ in group 1} \\ z_2 & \text{if person } i \text{ in group 2.} \end{cases}$$

Plugging this into the regression equation:

$$\text{Group 1: } \mu_1 = \alpha + \beta_1 \times z_1$$

$$\text{Group 2: } \mu_2 = \alpha + \beta_1 \times z_2$$

The parameters α and β_1, together with z, define the mean values in each group (i.e., μ_1 and μ_2).

Code variables – Two groups

Interpret regression parameters α, β_1

$$\begin{cases} \mu_1 = \alpha + \beta_1 \times z_1 \\ \mu_2 = \alpha + \beta_1 \times z_2 \end{cases}$$

Solve system of equations with respect to α, β_1.

- This allows interpreting α and β_1 in terms of group mean values.

The mathematical solution is not beautiful
(note: You DON'T need to know how to manually derive this!):

$$\begin{cases} \alpha = \frac{z_2\mu_1 - z_1\mu_2}{z_2 - z_1} \\ \beta_1 = \frac{\mu_2 - \mu_1}{z_2 - z_1} \end{cases}$$

Bahhh, this looks horrible.

Code variables – Two groups

It pays off to use simple codes, say, $z_1 = 0$, $z_2 = 1$.
This is why the dummy coding system is so popular:

$$z_i = \begin{cases} 0 & \text{if person } i \text{ in group 1} \\ 1 & \text{if person } i \text{ in group 2.} \end{cases}$$

For the dummy coding system,

$$\begin{cases} \alpha = \frac{z_2\mu_1 - z_1\mu_2}{z_2 - z_1} \\ \beta_1 = \frac{\mu_2 - \mu_1}{z_2 - z_1} \end{cases} \quad \underset{(z_1=0, z_2=1)}{\longrightarrow} \quad \begin{cases} \alpha = \mu_1 \\ \beta_1 = \mu_2 - \mu_1 \end{cases}$$

Hey, this looks neat!

- $\alpha =$ Mean of group 1 (i.e., the group coded with 0s: Reference group).
- $\beta_1 =$ Difference between the mean of group 2 and the mean of the reference group.

Estimate the parameters using the corresponding sample quantities:

$$\widehat{y} = a + b_1 z, \text{ with } a = \overline{y}_1, b = \overline{y}_2 - \overline{y}_1.$$

Example – Preventing flashbacks (TIME)

	TIME	
	1	2
Valid	30	42
Mean	4.267	3.690
Std. Deviation	3.352	3.382

Using dummy codes, thus $0 = \text{morning}$, $1 = \text{afternoon}$ (predictor TIME01):

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	4.267	0.615		6.935	$< .001$
	TIME01	-0.576	0.806	-0.085	-0.715	0.477

- $a = \overline{y}_1 = 4.267$.
- $b_1 = \overline{y}_2 - \overline{y}_1 = 3.690 - 4.267 = -0.577$.

Example – Preventing flashbacks (TIME)

What would happen had we used the original predictor TIME
(1 = morning, 2 = afternoon)?

$$\begin{cases} \alpha = \frac{z_2\mu_1 - z_1\mu_2}{z_2 - z_1} \\ \beta_1 = \frac{\mu_2 - \mu_1}{z_2 - z_1} \end{cases} \xrightarrow[(z_1=1, z_2=2)]{} \begin{cases} \alpha = 2\mu_1 - \mu_2 \\ \beta_1 = \mu_2 - \mu_1 \end{cases}$$

	TIME	
	1	2
Valid	30	42
Mean	4.267	3.690
Std. Deviation	3.352	3.382

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	4.843	1.336		3.625	$<$ .001
	TIME	-0.576	0.806	-0.085	-0.715	0.477

- $a = 2\overline{y}_1 - \overline{y}_2 = 2 \times 4.267 - 3.690 = 4.844$.
- $b_1 = \overline{y}_2 - \overline{y}_1 = 3.690 - 4.267 = -0.577$.

Example – Preventing flashbacks (TIME)

Different coding system $\Longrightarrow$ Different parameters $\Longrightarrow$ Different interpretation.

But: Test (not C.I.!!) of the 'time' effect remains unaffected!

Codes 0, 1

Model		Unstandardized	Standard Error	Standardized	*t*	*p*
1	(Intercept)	4.267	0.615		6.935	$< .001$
	TIME01	-0.576	0.806	-0.085	-0.715	0.477

Codes 1, 2

Model		Unstandardized	Standard Error	Standardized	*t*	*p*
1	(Intercept)	4.843	1.336		3.625	$< .001$
	TIME	-0.576	0.806	-0.085	-0.715	0.477

Codes -3, 7

Model		Unstandardized	Standard Error	Standardized	*t*	*p*
1	(Intercept)	4.094	0.458		8.938	$< .001$
	TIME-3+7	-0.058	0.081	-0.085	-0.715	0.477

What *t*-test is this?

Code variables – Two groups

Recall that

$$\beta_1 = \frac{\mu_2 - \mu_1}{z_2 - z_1}.$$

Note that

$$\underbrace{\beta_1 = 0} \iff \frac{\mu_2 - \mu_1}{z_2 - z_1} = 0 \iff \mu_2 - \mu_1 = 0 \iff \underbrace{\mu_2 = \mu_1}.$$

Therefore, testing $\mathcal{H}_0 : \beta_1 = 0$ is equivalent to testing $\mathcal{H}_0 : \mu_1 = \mu_2$.

- But this is the independent samples t-test!!

Thus, the independent samples t-test and regression with a binary code variable are linked.

	t	df	p
N_INTR	0.715	70	0.477

Independent samples t-test as a regression model

Test the difference between two groups: $\mathcal{H}_0$: $\mu_1 = \mu_2$.

Independent samples t-test:

- Compute s_p and
$$t = \frac{\overline{y}_2 - \overline{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$
- Compare t with $t^*_{0.05(2);n_1+n_2-2}$.

Regression:

- Create a code variable with values 0 and 1 (or any other pair of values!).
- Compute $b_1 = \overline{y}_2 - \overline{y}_1$ and SE_{b_1}.
- Test $\mathcal{H}_0$ through $t = b_1/SE_{b_1}$.
- Compare t with $t^*_{0.05(2);n_1+n_2-2}$.

Both statistical procedures are equivalent.

Code variables – Two groups

Q: What about CIs for regression parameters?

A: CIs, just like the associated regression parameters, must be interpreted in light of the coding variable used.

$$\begin{cases} \alpha = \frac{z_2\mu_1 - z_1\mu_2}{z_2 - z_1} \\ \beta_1 = \frac{\mu_2 - \mu_1}{z_2 - z_1} \end{cases} \xrightarrow[(z_1=0, z_2=1)]{} \begin{cases} \alpha = \mu_1 \\ \beta_1 = \mu_2 - \mu_1 \end{cases}$$

Codes 0, 1

						95% CI	
	Unstd. Coef.	SE	Std. Coef.	t	p	Lower	Upper
(Intercept)	4.267	0.615		6.935	$< .001$	3.040	5.494
TIME01	−0.576	0.806	−0.085	−0.715	0.477	−2.183	1.030

- $\alpha = 4.267$, $95\% CI = (3.040, 5.494)$:
  We are 95% confident that μ_1 lies between 3.040 and 5.494.
- $\beta_1 = -0.576$, $95\% CI = (-2.183, 1.030)$: Estimate and inference for $(\mu_2 - \mu_1)$:
  We are 95% confident that $(\mu_2 - \mu_1)$ lies between −2.183 and 1.030.

Code variables – More than two groups

- In general, a categorical variable with g levels ($g \geq 2$) requires $(g-1)$ code variables.
- There are many ways of choosing a set of $(g-1)$ code variables.
- Just like we saw when $g = 2$, the following two rules-of-thumb apply:
  - Testing the 'effect' of the categorical predictor on y does not depend on the coding system.
  - Interpreting this effect (by means of regression coefficients and CIs) does depend on the coding system.

We will focus on the dummy coding system.

Example – Preventing flashbacks (CONDITION)

CONDITION

1 = No-task control

2 = Reactivation + Tetris

3 = Tetris only

4 = Reactivation only

Group	z_1	z_2	z_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

The code variables function as group identifiers:

- z_1: Identifier for subjects in Group 1.
- z_2: Identifier for subjects in Group 2.
- z_3: Identifier for subjects in Group 3.

By exclusion of parts, all subjects scoring 0 on all code variables belong to the last group (reference group).

Code variables – More than two groups

Population regression equation:

$$\mu_y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3.$$

Each set of values $\{z_1, z_2, z_3\}$ (i.e., each row in the table) defines one subpopulation of y values, normally distributed around μ_y with constant σ.

For the dummy coding system:

Group	z_1	z_2	z_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

$$\longrightarrow \begin{cases} \mu_1 = \alpha + 1\beta_1 + 0\beta_2 + 0\beta_3 = \alpha + \beta_1 \\ \mu_2 = \alpha + 0\beta_1 + 1\beta_2 + 0\beta_3 = \alpha + \beta_2 \\ \mu_3 = \alpha + 0\beta_1 + 0\beta_2 + 1\beta_3 = \alpha + \beta_3 \\ \mu_4 = \alpha + 0\beta_1 + 0\beta_2 + 0\beta_3 = \alpha \end{cases}$$

Code variables – More than two groups

Interpret regression parameters α, β_i $(i = 1, \ldots, g-1)$

Using the dummy coding system, solve the system of equations with respect to α and β_i $(i = 1, \ldots, g-1)$.

- This allows interpreting α and the β_i's in terms of group mean values.

$$\begin{cases} \mu_1 = \alpha + \beta_1 \\ \mu_2 = \alpha + \beta_2 \\ \mu_3 = \alpha + \beta_3 \\ \mu_4 = \alpha \end{cases} \longrightarrow \begin{cases} \beta_1 = \mu_1 - \mu_4 \\ \beta_2 = \mu_2 - \mu_4 \\ \beta_3 = \mu_3 - \mu_4 \\ \alpha = \mu_4 \end{cases}$$

- $\alpha =$ Mean of group 4 (reference group).
- $\beta_i =$ Difference between the mean of group i and the mean of the reference group.

Example – Preventing flashbacks (CONDITION)

CONDITION

1 = No-task control

2 = Reactivation + Tetris

3 = Tetris only

4 = Reactivation only

	N_INTR			
	1	2	3	4
Valid	18	18	18	18
Mean	5.111	1.889	3.889	4.833
Std. Deviation	4.227	1.745	2.888	3.330

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	4.833	0.749		6.457	$< .001$
	z1	0.278	1.059	0.036	0.262	0.794
	z2	−2.944	1.059	−0.382	−2.781	0.007
	z3	−0.944	1.059	−0.123	−0.892	0.375

- $a = \overline{y}_4 = 4.833$.
- $b_1 = \overline{y}_1 - \overline{y}_4 = 5.111 - 4.833 = 0.278$.
- $b_2 = \overline{y}_2 - \overline{y}_4 = 1.889 - 4.833 = -2.944$.
- $b_3 = \overline{y}_3 - \overline{y}_4 = 3.889 - 4.833 = -0.944$.

Example – Preventing flashbacks (CONDITION)

How to retrieve the group means from the regression equation?

	N_INTR 1	2	3	4
Valid	18	18	18	18
Mean	5.111	1.889	3.889	4.833
Std. Deviation	4.227	1.745	2.888	3.330

Level	z_1	z_2	z_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	**4.833**	0.749		6.457	$<$ .001
	z1	**0.278**	1.059	0.036	0.262	0.794
	z2	$-$**2.944**	1.059	-0.382	-2.781	0.007
	z3	$-$**0.944**	1.059	-0.123	-0.892	0.375

$$\widehat{y} = 4.833 + 0.278z_1 - 2.944z_2 - 0.944z_3$$

- $\overline{y}_1 = 4.833 + 0.278 \times 1 - 2.944 \times 0 - 0.944 \times 0 = 5.111$
- $\overline{y}_2 = 4.833 + 0.278 \times 0 - 2.944 \times 1 - 0.944 \times 0 = 1.889$
- $\overline{y}_3 = 4.833 + 0.278 \times 0 - 2.944 \times 0 - 0.944 \times 1 = 3.889$
- $\overline{y}_4 = 4.833 + 0.278 \times 0 - 2.944 \times 0 - 0.944 \times 0 = 4.833$

Example – Preventing flashbacks (CONDITION)

Different coding system $\Longrightarrow$ Different parameters $\Longrightarrow$ Different interpretation

Level	z_1	z_2	z_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Level	zz_1	zz_2	zz_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

						95% CI	
	Unstd. Coef.	SE	Std. Coef	t	p	Lower	Upper
(Intercept)	4.833	0.749		6.457	$<$ .001	3.340	6.327
z1	0.278	1.059	0.036	0.262	0.794	−1.835	2.390
z2	−2.944	1.059	−0.382	−2.781	0.007	−5.057	−0.832
z3	−0.944	1.059	−0.123	−0.892	0.375	−3.057	1.168

						95% CI	
	Unstd. Coef.	SE	Std. Coef.	t	p	Lower	Upper
(Intercept)	5.111	0.749		6.828	$<$ .001	3.617	6.605
zz1	−3.222	1.059	−0.418	−3.044	0.003	−5.335	−1.110
zz2	−1.222	1.059	−0.159	−1.155	0.252	−3.335	0.890
zz3	−0.278	1.059	−0.036	−0.262	0.794	−2.390	1.835

Code variables – More than two groups

(Exercise: Interpret the regression coefficients based on code variables zz_1, zz_2, and zz_3!)

So interpretation depends on the coding system.
However, testing the effect of the categorical predictor on y does not!

Similarly to the two-group situation, it can be shown that testing

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

is equivalent to testing

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{g-1} = 0,$$

which is also equivalent to (see lecture 4)

$$\mathcal{H}_0 : R^2 = 0,$$

irrespective of the coding system used.

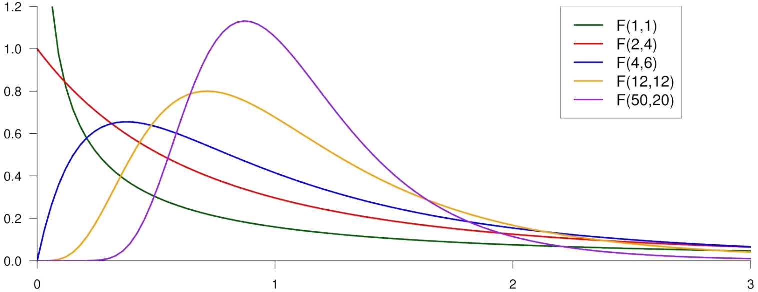
Code variables – More than two groups

An F-test (recall lecture 4) is used to test this effect:

$$F = \frac{MSR}{MSE} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \underset{\mathcal{H}_0}{\sim} F(p, n-p-1),$$

where $p =$ number of predictors $= g - 1$.

This is the infamous (as you'll see later in the course) omnibus ANOVA F test.



Example – Preventing flashbacks (CONDITION)

For any coding system, for example,

Level	z_1	z_2	z_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

or

Level	zz_1	zz_2	zz_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

or any other,

the same F test applies:

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	114.8	3	38.27	3.795	0.014
	Residual	685.8	68	10.09		
	Total	800.7	71			

$F(p, n-p-1) = 3.795$, where:

- $p = g - 1 = 3$;
- $n - p - 1 = 72 - 3 - 1 = 68$.

In this case, $F(3, 68) = 3.795$, $p = .014$, thus we reject $\mathcal{H}_0$ at 5% significance level.

Beyond the F test

The null hypothesis tested by the F test is quite general ('omnibus'):

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_g \quad \Longleftrightarrow \quad \mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{g-1} = 0.$$

Rejecting $\mathcal{H}_0$ means actually very little:

There is evidence that not all population group means are equal...

(how surprising is that?!)

But we are using regression! So, more focused tests of effects are possible:

$$\mathcal{H}_0 : \beta_i = 0 \quad \text{versus} \quad \mathcal{H}_a : \beta_i \neq 0.$$

When the predictors are code variables like today, each β_i can be expressed in terms of group means. Therefore, testing $\mathcal{H}_0 : \beta_i = 0$ can be translated into testing special relations between population group means. These tests are known as contrasts.

Studying contrasts is the next lecture's topic!

Next lecture – Regression Modeling with Categorical Predictors: Multiple comparisons, contrasts

Agresti, Section 12.2