Lecture 8 - GLM IV Differences between more groups

Caspar J. van Lissa

2023-08-13

Categorical predictors

More than 2 categories

Imagine we measured SES as Low, Medium and High
We want to predict fathers' involvement in child rearing
How can we model this using regression?

Two+ categories

Remember: we can use bivariate linear regression to model two categories
$\hat{Y}_i = a + b * X_i$

- a: Mean of group 1
- b: Mean difference between groups 1 and 2
- X_i is a dummy variable (codes for membership of group 2)

For **three+** categories, we can **expand** the model:
$\hat{Y}_i = a + b_1 * X_{1i} + b_2 * X_{2i}$

- a: Mean of group 1
- b_1: Mean difference between groups 1 and 2
- b_2: Mean difference between groups 1 and 3

Dummy coding

We make dummy variables for two of the categories of SES:

SES	X_1	X_2
High	0	0
Medium	1	0
Low	0	1

Regression formula:

$\hat{Y}_i = a + b_1 * X_{1i} + b_2 * X_{2i}$

Dummy variables

Regression formula:
$\hat{Y}_i = a + b_1 * X_{1i} + b_2 * X_{2i}$
Complete for High:

- $\hat{Y}_i = a + b_1 * 0 + b_2 * 0 = a$
- The mean value for high SES is the intercept (a)

Complete for Medium:

- $\hat{Y}_i = a + b_1 * 1 + b_2 * 0 = a + b_1$
- The mean value for Medium SES is a plus the difference between High and Medium SES (b_1)

Complete for Low:

- $\hat{Y}_i = a + b_1 * 0 + b_2 * 1 = a + b_2$
- The mean value for Low SES is a plus the difference between High and Low SES (b_2)

Dummy variables

Intercept (a): Mean of the “reference category”
b-coefficients: Mean difference relative to the “reference category”

SES	X_1	X_2
High	0	0
Medium	1	0
Low	0	1

Reference category

If you want to change the reference category, choose different dummies:

SES	X_1	X_2
High	1	0
Medium	0	1
Low	0	0

Degrees of freedom?

Three categories, two dummy variables, how many parameters?
$\hat{Y}_i = a + b_1 * X_{1i} + b_2 * X_{2i}$?

- df_1: p-1
- df_2: n-p

p: 3 (a, b_1, b_2)

Some output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.459[a]	.211	.210	12.430

a. Predictors: (Constant), drural, dsuburban

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87964.439	2	43982.220	284.645	<.001[b]
	Residual	329119.280	2130	154.516		
	Total	417083.720	2132			

a. Dependent Variable: Post-test

b. Predictors: (Constant), drural, dsuburban

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	61.748	.413		149.521	.000
	dsuburban	14.289	.621	.483	22.998	<.001
	drural	2.303	.688	.070	3.346	<.001

a. Dependent Variable: Post-test

ANOVA

What is ANOVA

ANalysis **O**f **VA**riance

One-way ANOVA has one categorical predictor and a continuous outcome

- It provides an omnibus (overall) test of differences between group means
- So… exactly what we've learned so far
- For historical reasons, some fields call regression with dummies a "one-way ANOVA", and present the results in a particular way
- It's important that you recognize this and interpret it appropriately

Some output

ANOVA

Post-test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	87964.439	2	43982.220	284.645	<.001
Within Groups	329119.280	2130	154.516		
Total	417083.720	2132			

ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval Lower	Upper
Post-test	Eta-squared	.211	.182	.240
	Epsilon-squared	.210	.181	.239
	Omega-squared Fixed-effect	.210	.181	.239
	Omega-squared Random-effect	.117	.099	.136

Compare to regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.459[a]	.211	.210	12.430

a. Predictors: (Constant), drural, dsuburban

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87964.439	2	43982.220	284.645	<.001[b]
	Residual	329119.280	2130	154.516		
	Total	417083.720	2132			

a. Dependent Variable: Post-test

b. Predictors: (Constant), drural, dsuburban

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	61.748	.413		149.521	.000
	dsuburban	14.289	.621	.483	22.998	<.001
	drural	2.303	.688	.070	3.346	<.001

a. Dependent Variable: Post-test

Step 1. Hypothesis

ANOVA: comparing means of two or more groups

- Common default null hypothesis:
  - $H_0 : \mu_1 = \mu_2 = \mu_3, H_a : \text{not } H_0$
  - H_0 is rejected if there are sufficient differences between ANY of the means
  - So H_0 is also rejected if we find that $\mu_1 = \mu_2; (\mu_1, \mu_2) < \mu_3$

  Significant ANOVA means that the means of at least two groups differ
- We can do follow-up analyses (e.g., pairwise comparisons) to understand which groups differ significantly

Step 2. Test statistic

- Why called “Analysis of Variance” when we are comparing group means?
  - Because we test how large the variance in group means is, relative to the error variance
  - Of course, since we compare two sources of variance, we’ll use an F-test

$$F = \frac{MS_B}{MS_W} = \frac{SSB/df_b}{SSW/df_w} = \frac{SSR/df_r}{SSE/df_e}$$

How large are the between-group differences relative to the error variance?

Sums of Squares in ANOVA

Sum	Formula	Meaning
SST	$\sum (Y_i - \bar{Y})^2$	Variation of individual observations wrt grand mean
SSB (SSR)	$\sum (\hat{Y}_i - \bar{Y})^2$	Variation of group means wrt grand mean
SSW (SSE)	$\sum (Y_i - \hat{Y}_i)^2$	Variation of individuals wrt group means

Step 3. p-value

The F-distribution has two degrees of freedom parameters:

- Numerator df: df_b = g−1 (g = number of groups)
- Denominator df: df_w = n-g (n = number of observations)

How to calculate the p-value?

- Statistical software
  - Excel: `=F.DIST.RT(x, degree_freedom1, degree_freedom2)`
  - R: `pf(F_statistic, dataframe1, dataframe2, lower.tail = FALSE)`
- Table in book

Step 4. Conclusion

- If $p < \alpha$, the test is significant
- It is very unlikely to observe differences between group means at least as large as you observed, if H_0 were true

Effect size in ANOVA

It's common to call R^2 for ANOVA models "eta squared", η^2

- Note: It is exactly the same as R^2
  - $\eta^2 = \frac{SSB}{SST}, R^2 = \frac{SSR}{SST}$
- There are rules of thumb like (Cohen, 1988)
- But they differ from rules of thumb for R^2, which is confusing

Small	0.01
Medium	0.06

Small	0.01
Large	0.14

Interpretation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.459[a]	.211	.210	12.430

a. Predictors: (Constant), drural, dsuburban

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87964.439	2	43982.220	284.645	<.001[b]
	Residual	329119.280	2130	154.516		
	Total	417083.720	2132			

a. Dependent Variable: Post-test

b. Predictors: (Constant), drural, dsuburban

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	61.748	.413		149.521	.000
	dsuburban	14.289	.621	.483	22.998	<.001
	drural	2.303	.688	.070	3.346	<.001

a. Dependent Variable: Post-test

Interpretation

Environment (urban vs. suburban vs. rural) had a significant effect on test scores, $R^2 = .21, F(2, 2130) = 284.65, p < .001$.
The average test score for urban participants was $M = 61.75$, and this score differed significantly from zero, $t(2130) = 149.52, p < .001$.
Compared to urban participants, both suburban participants ($M = 76.04$) and rural participants ($M = 64.05$) scored significantly higher, with
$t(2130) = 23.00, p < .001$ and $t(2130) = 3.35, p < .001$, respectively.

Interpretation

Important caveats:

- Standardized regression coefficients do not make sense for dummies!
  - “one SD difference in the dummy for rural living” is nonsense.
- Significance tests for dummies only make sense if you want to compare them to the reference group
  - If you want to compare to another group, you have to use different dummies