LING82100: null hypothesis testing for categorical variables in regression

Kyle Gorman

1 Introduction

Categorical independent variables—in particular those with more than two levels—pose particular challenges for hypothesis testing. Under standard dummy-coding schemes, fitting a regression with an categorical independent variable of n "levels" produces $n - 1$ coefficients (i.e., βs) and $n - 1$ standard errors. Each of these coefficients has a t-distribution, parameterized by the associated degrees of freedom and standard error, but the null hypothesis—that the population-level coefficient is zero—is rarely one we're interested in, and its precise interpretation depends on how the factor has been coded:

- In *treatment coding* (the default in R), if the null hypothesis is true then the relevant level has the same mean value of the dependent variable as the reference level.
- In *sum coding*, if the null hypothesis true then the relevant level is not significantly different from the overall category mean of the dependent variable.

While these are not trivial, one is usually interested in different questions, such as:

- Is the categorical independent variable *overall* correlated with the dependent variable? I.e., are changes in the value of the IV associated with changes in the value of the DV?
- Which of the levels of the categorical independent variable are significantly different from any other level? I.e., are different levels of the IV associated with changes in the value of the DV?

We will refer to the first as *tests for effects of group* and the second as *post-hoc tests*.

As a running example, we will use (simulated) data based on Presmanes Hill et al. (2015). This study examined verbal and non-verbal memory in three groups of children between the ages of 5 and 8: children with autism spectrum disorder (ASD) and no language impairment (ALN; $n = 20$); children with ASD and language impairment (ALI; $n = 22$), and group of children with specific language impairment (SLI; $n = 18$). The dependent variable here is `CLS`, the core language score on the CELF (Clinical Evaluation of Language Fundamentals), a widely used clinical instrument. Sum coding is used for the independent variable `DX` ("diagnosis").

```
> r <- lm(CLS ~ DX, data = d)
```

2 Tests for effect of group

As expected, the model includes three coefficients: one for the intercept and two for the dummy-coded IV:

```
> summary(r)

Call:
lm(formula = CELF.CLS ~ DX, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-28.324  -6.510   1.358   6.929  30.207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.531      1.526  54.726  < 2e-16 ***
DX1          -15.574      2.106  -7.397 6.94e-10 ***
DX2           23.445      2.155  10.879 1.56e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.78 on 57 degrees of freedom
Multiple R-squared:  0.6879,    Adjusted R-squared:  0.6769
F-statistic:  62.8 on 2 and 57 DF,  p-value: 3.883e-15
```

Both of the dummy-coding coefficients are significant, but is diagnosis correlated with the dependent variable? One way to frame this as a null hypothesis test is the *likelihood ratio test*, which we define below.

The *likelihood* $\mathcal{L}$ of a model is simply the probability the model assigns to some data. For instance, if the model assumes that a coin is fair, and the data consists of two heads, the likelihood is $.5^2 = .25$, assuming the coin flips are independent. For large data sets, the likelihood is often vanishingly small, so we instead work with log-likelihood, denoted by ℓ; Suppose we have two statistical models, one defined by the parameters/coefficients θ and another by a subset, θ_0. In this configuration we say that θ *nests*θ_0. It must be the case, then, that the (log-)likelihood of θ with respect to some data is greater or equal to the (log-)likehood of θ_0, because θ is more expressive than θ_0. Thus:

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta)} \in [0, 1]$$

That is, the ratio the two likelihoods is between 0 and 1, inclusive. Or, equivalently:

$$\ell(\theta_0) - \ell(\theta) \leq 0$$

That is, the difference in log-likelihoods is less than or equal to zero. In the *likelihood-ratio test* the null hypothesis is that the difference in likelihoods is exactly zero, and the (one-sided) alternative hypothesis is that it is less than zero. Let us define a test statistic $LR = -2\left[\ell(\theta_0) - \ell(\theta)\right]$.

Then, according to a well-known theorem, LR has an asymptotic χ^2 distribution under the null hypothesis; the degrees of freedom are given by $|\theta| - |\theta_0|$, the difference between the number of parameters in θ and in θ_0. When one uses this test to test for the significance of a categorical independent variable in a regression, θ is the full model; θ_0 is the model with that categorical variable removed; and the degrees of freedom are $n - 1$. In R, one can compute this test statistic manually by fitting the θ_0 model, a model without the `DX` independent variable, and then computing the test statistic and associated p-value, as follows:

```
> r0 <- lm(CELF.CLS ~ 1, data = d)
> ell.r0 <- logLik(r0)[1]
> ell.r <- logLik(r)[1]
> LR <- -2 * (ell.r0 - ell.r)
> p <- 1 - pchisq(LR, df = nlevels(d$DX) - 1)
> c("LR" = LR, "p" = p)
          LR            p
6.985713e+01 6.661338e-16
```

R also provide an automatic method to apply this test, `drop1`, though it does not report the test statistic itself:[1]

```
> drop1(r, test = "Chisq")

Single term deletions

Model:
CELF.CLS ~ DX
       Df Sum of Sq     RSS    AIC  Pr(>Chi)
<none>               7914.4 298.93
DX      2     17440 25354.9 364.78 6.772e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One might report this as follows:

> There was a significant effect of diagnostic group on CELF CLS ($\chi^2(2) = 69.857$, $p < .001$).

The only major limitation of this test is that it is asymptotically valid and thus inappropriate for very small samples. Also note that the test is not valid if θ does not nest θ_0 as defined above.

3 Post-hoc tests

Post-hoc tests allow us to test the null hypothesis that any two pairs of levels from the same categorical variable are associated with the same value of the IV. Since there can be a huge number

[1] We also observe a small difference in the estimated p-value.

of such pairs when a categorical variable has many levels—and there can be many such categorical variables—it is necessary to apply a *correction for multiple comparison*. For linear models, this is handled using the Tukey HSD ("Honestly Significant Difference") test. In lecture, I gave an example of using the built-in `TukeyHSD` function; here, I instead use functions from the `multcomp` package, which produces easier-to-read output.[2]

```
> pairs <- glht(r, linfct = mcp(DX = "Tukey"))
> summary(pairs)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = CELF.CLS ~ DX, data = d)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
ALN - ALI == 0   39.019      3.641  10.718   <1e-04 ***
SLI - ALI == 0    7.704      3.745   2.057    0.108
SLI - ALN == 0  -31.315      3.828  -8.180   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Here, the `Estimate` column gives the difference between the mean DV for the two levels; the remaining columns report the associated standard error, the t-statistic, and the p-value (adjusted for multiple comparisons). There are significant differences (at $\alpha = .05$) between ALN and ALI, and between ALN and SLI; in both, the ALN group has a higher mean value for the CELF CLS. (This is unsurprisingly because CELF CLS was used to diagnose children for language impairment.) Informally, one might write this as "[SLI = ALI] $<$ ALN".

References

Presmanes Hill, Alison, Jan van Santen, Kyle Gorman, Beth Hoover Langhorst, and Eric Fombonne. 2015. Memory in language-impaired children with and without autism. *Journal of Neurodevelopmental Disorders* 2015:7–19.

[2]This package does not ship with R; to install it, execute `install.packages("multcomp")` in R.