

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. It's 100% free, no registration required.

[Sign up](#) [x](#)

Significance of categorical predictor in logistic regression

I am having trouble interpreting the z values for categorical variables in logistic regression. In the example below I have a categorical variable with 3 classes and according to the z value, CLASS2 might be relevant while the others are not.

But now what does this mean?

That I could merge the other classes to one?

That the whole variable might not be a good predictor?

This is just an example and the actual z values here are not from a real problem, I just have difficulties about their interpretation.

	Estimate	Std. Error	z value	Pr(> z)
CLASS0	6.069e-02	1.564e-01	0.388	0.6979
CLASS1	1.734e-01	2.630e-01	0.659	0.5098
CLASS2	1.597e+00	6.354e-01	2.514	0.0119 *

r | [logistic](#) | [feature-selection](#)

edited Jun 4 '13 at 7:59



Nick Cox

22.3k

3

35

66

asked Jun 4 '13 at 7:21



user695652

169

1

2

12

2 Answers

The following explanation is **not limited to logistic regression** but applies equally in normal linear regression and other GLMs. Usually, `R` excludes one level of the categorical and the coefficients denote **the difference of each class to this reference class (or sometimes called baseline class)** (this is called dummy coding or treatment contrasts in `R`, see [here](#) and [here](#) for an excellent overview of the different contrast options). To see the current contrasts in `R`, type `options("contrasts")`. Normally, `R` orders the levels of the categorical variable alphabetically and takes the first as reference class. This is not always optimal and can be changed by typing (here, we would set the reference class to "c" in the new variable) `new.variable <- relevel(old.variable, ref="c")`. For each coefficient of every level of the categorical variable, a **Wald test** is performed to **test whether the pairwise difference between the coefficient of the reference class and the other class is different from zero** or not. This is what the z and p -values in the regression table are. If only one categorical class is significant, this does *not* imply that the whole variable is meaningless and should be removed from the model. You can check the overall effect of the variable by performing a **likelihood ratio test**: fit two models, one with and one without the variable and type `anova(model1, model2, test="LRT")` in `R` (see example below). Here is an example:

```
mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
mydata$rank <- factor(mydata$rank)
my.mod <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
summary(my.mod)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979    1.139951  -3.500  0.000465 ***
gre           0.002264    0.001094   2.070  0.038465 *
gpa           0.804038    0.331819   2.423  0.015388 *
rank2        -0.675443    0.316490  -2.134  0.032829 *
rank3        -1.340204    0.345306  -3.881  0.000104 ***
rank4        -1.551464    0.417832  -3.713  0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The level `rank1` has been omitted and each coefficient of `rank` denotes the difference between the coefficient of `rank1` and the corresponding `rank` level. So the difference between the coefficient of `rank1` and `rank2` would be -0.675 . **The coefficient of `rank1` is simply the intercept.** So the true coefficient of `rank2` would be $-3.99 - 0.675 = -4.67$. The Wald tests here test whether the difference between the coefficient of the reference class (here `rank1`) and the corresponding levels differ from zero. In this case, we have evidence that the coefficients of all classes differ from the coefficient of `rank1`. You could also fit the model without an intercept by adding `- 1` to the model formula to see all coefficients directly:

```
my.mod2 <- glm(admit ~ gre + gpa + rank - 1, data = mydata, family = "binomial")
summary(my.mod2) # no intercept model
```

```
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
gre    0.002264  0.001094   2.070 0.038465 *
gpa    0.804038  0.331819   2.423 0.015388 *
rank1 -3.989979  1.139951  -3.500 0.000465 ***
rank2 -4.665422  1.109370  -4.205 2.61e-05 ***
rank3 -5.330183  1.149538  -4.637 3.54e-06 ***
rank4 -5.541443  1.138072  -4.869 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the intercept is gone now and that the coefficient of `rank1` is exactly the intercept of the first model. Here, the Wald test checks not the pairwise difference between coefficients but the hypothesis that **each individual coefficient is zero**. Again, we have evidence that every coefficient of `rank` differs from zero. Finally, to check whether the whole variable `rank` improves the model fit, we fit one model with (`my.mod1`) and one without the variable `rank` (`my.mod2`) and conduct a likelihood ratio test. This tests the hypothesis that all coefficients of `rank` are zero:

```
my.mod1 <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial") # with
rank
my.mod2 <- glm(admit ~ gre + gpa, data = mydata, family = "binomial") # without rank

anova(my.mod1, my.mod2, test="LRT")

Analysis of Deviance Table

Model 1: admit ~ gre + gpa + rank
Model 2: admit ~ gre + gpa
      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1          394      458.52
2          397      480.34 -3    -21.826 7.088e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test is highly significant and we would conclude that the variable `rank` should remain in the model.

This post is also very interesting.

edited Jun 6 '13 at 12:55

answered Jun 4 '13 at 7:49

 **COOLSerdash**
8,328 3 26 51


The *z*-value is just the test-statistic for a statistical test, so if you have trouble interpreting it your first step is to find out what the null hypothesis is. The null hypothesis for the test for CLASS0 is that its coefficient is 0. The coefficient for CLASS0 is the difference in log(odds) between CLASS0 and the reference class (CLASS3?) is zero, or equivalently, that the ratio of the odds for CLASS0 and the reference class is 1. In other words that there is no difference in the odds of success between CLASS0 and the reference class.


So does a non-significant coefficient mean you can merge categories? No. First, non-significant means that we cannot reject the hypothesis that there is no difference, but that does not mean that no such differences exist. An absence of evidence is not the same thing as evidence of absence. Second, merging categories, especially the reference category, changes the interpretation of all other coefficients. Whether or not that makes sense depends on what those different classes stand for.

Does that mean that the entire categorical variable is a "bad" (non-significant) predictor? No, for that you would need to perform a simultaneous test for all CLASS terms.

edited Jun 4 '13 at 7:58

answered Jun 4 '13 at 7:51

 **Nick Cox**
22.3k 3 35 66

 **Maarten Buis**
9,433 6 30