Logistic Regression

Linear Combinations and Qualitative Predictors

Demetris Athienitis

Linear combination of parameters

If a qualitative predictor is deemed significant, the next step is an investigation into the different levels. Which one differ from which.

This yields situations where one might we to test linear combinations of parameters.

$$\mathrm{H}_0 : \sum_{i=1}^{k} c_i \beta_i = \Delta_0$$

for constants c_i and constant null value Δ_0.

Example (Horseshoe crad continued)

Testing $\beta_2 = 0, \beta_3 = 0$ and $\beta_4 = 0$ individually amounts to testing differences between each group to the base group

Color	$\text{logit}\,[\pi(x)]$
medium light	$(\alpha + \beta_2) + \beta_1 x$
medium	$(\alpha + \beta_3) + \beta_1 x$
medium dark	$(\alpha + \beta_4) + \beta_1 x$
dark	$\alpha + \beta_1 x$

To motivate the next section consider medium light vs. medium,

$$H_0 : \beta_2 = \beta_3 \Leftrightarrow \beta_2 - \beta_3 = 0$$

a linear combination of the parameters i.e.

$$(0)\alpha + (0)\beta_1 + (1)\beta_2 + (-1)\beta_3 + (0)\beta_4$$

Linear combination of parameters

Instead of a test, create CI for $\sum c_i \beta_i$ (using the asymptotic normality property)

$$\sum_{i=1}^{k} c_i \hat{\beta}_i \mp z_{1-\alpha/2} \sqrt{\hat{V}\left(\sum_{i=1}^{k} c_i \hat{\beta}_i\right)}$$

where

$$\begin{aligned} V\left(\sum_{i=1}^{k} c_i \hat{\beta}_i\right) &= \sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) \\ &= \sum_{i=1}^{k} c_i^2 V(\hat{\beta}_i) + 2 \sum\sum_{i<j} c_i c_j \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) \end{aligned}$$

Example (continued)

To perform all 6 pairwise comparisons among color levels. Each entry is the log odds ratio comparing the two groups.

Comparison	CI on
medium light vs dark	β_2
medium vs dark	β_3
medium dark vs dark	β_4
medium light vs medium	$\beta_2 - \beta_3$
medium light vs medium dark	$\beta_2 - \beta_4$
medium vs medium dark	$\beta_3 - \beta_4$

So for $\beta_2 - \beta_3$

$$\hat{\beta}_2 - \hat{\beta}_3 \mp z_{1-\alpha/2}\sqrt{s^2_{\beta_2} + s^2_{\beta_3} - 2s_{\beta_2\beta_3}}$$

Example (continued)

- from original output $\hat{\beta}_2 = 1.2694$ and $\hat{\beta}_3 = 1.4143$
- $s^2_{\beta_2} = -0.040, s^2_{\beta_3} = 0.238$ and $2s_{\beta_2\beta_3} = 0.721$ from the output below

```
> round(vcov(fit2),3)
            (Intercept) weight colorML colorM colorMD
(Intercept)       1.008 -0.342  -0.146 -0.215  -0.254
weight           -0.342  0.151  -0.040 -0.009   0.008
colorML          -0.146 -0.040   0.721  0.238   0.233
colorM           -0.215 -0.009   0.238  0.297   0.235
colorMD          -0.254  0.008   0.233  0.235   0.346
```

Bonferroni adjustment

- Due to the *multiple comparison problem* the critical value must be adjusted
- For example, cannot put 6 inferences together, each at 95% confidence level and expect overall experimentwise confidence level to remain 95%

If there are g simultaneous inferences to be performed then use $z_{1-\alpha/(2\times g)}$ in each CI.

Exercise

Perform all the CIs mentioned in the previous example, using the Bonferroni method.

Exercise

For the sake of practice let us compare dark vs non-dark using the current model, for a fixed level of weight. Hence a CI on

$$\frac{(\alpha+\beta_2+\beta_1 x)+(\alpha+\beta_3+\beta_1 x)+(\alpha+\beta_4+\beta_1 x)}{3}-(\alpha+\beta_1 x)$$
$$=\frac{1}{3}\beta_2+\frac{1}{3}\beta_3+\frac{1}{3}\beta_4$$

Example (Florida Death Penalty continued)

Victim's Race	Defendant's Race	Death Penalty: Yes	Death Penalty: No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

```
> dp.fit1=glm(cbind(Yes,No)~Defendant+Victim,
+  family=binomial,data=dpwide)
> summary(dp.fit1)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.5961     0.5069  -7.094 1.30e-12 ***
DefendantWhite -0.8678     0.3671  -2.364   0.0181 *
VictimWhite     2.4044     0.6006   4.003 6.25e-05 ***
---
    Null deviance: 22.26591  on 3  degrees of freedom
Residual deviance:  0.37984  on 1  degrees of freedom
```

Example (continued)

```
> exp(dp.fit1$coefficients[2])
DefendantWhite
     0.4198757
```

Odds ratio of a white defendant receiveing the deasth penalty (as compared to a black defendant), controlling for victim's race is 0.42, with a 95% CI

```
> exp(dp.fit1$coefficients[2]+c(-1,1)*1.96*
+  sqrt(vcov(dp.fit1)[2,2]))
[1] 0.2044847 0.8621455
```

To test if any predictor can be removed via LRT

```
> drop1(dp.fit1,test="LRT")
Single term deletions
Model: cbind(Yes, No) ~ Defendant + Victim
          Df Deviance    AIC     LRT Pr(>Chi)
<none>         0.3798 19.300
Defendant  1   5.3940 22.314  5.0142  0.02514 *
Victim     1  20.7298 37.650 20.3499 6.45e-06 ***
```

Section 2

Ordinal variables as quantitative

For example, you can order a drink in 3 sizes: small, medium and large, and there is an inherent order of 1, 2 and 3.

Size	Score
Small	1
Medium	2
Large	3

Now, assume medium size is 50% larger than the small, and large is 250% larger than the small. More representative scores might be

Size	Score
Small	1
Medium	1.5
Large	3.5

Example (Horseshoe crab continued)

3 binary variables were created to distinguish the 4 levels of color. If "darkness" is of interest, color is ordinal.

Color	Score
Medium Light	1
Medium	2
Medium dark	3
Dark	4

and the model is

$$\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 c$$

where x is weight and c is color score.

Example (continued)

Color	logit $[\pi(x)]$ Qualitative	logit $[\pi(x)]$ Quantitative
medium light	$(\alpha+\beta_2)+\beta_1 x$	$(\alpha+\beta_2)+\beta_1 x$
medium	$(\alpha+\beta_3)+\beta_1 x$	$(\alpha+2\beta_2)+\beta_1 x$
medium dark	$(\alpha+\beta_4)+\beta_1 x$	$(\alpha+3\beta_2)+\beta_1 x$
dark	$\alpha+\beta_1 x$	$(\alpha+4\beta_2)+\beta_1 x$

Note that the qualitative model is a lot more flexible (as it has more parameters) in differentiating between groups, while the quantitative model assumes a systematic change between groups.

Example (continued)

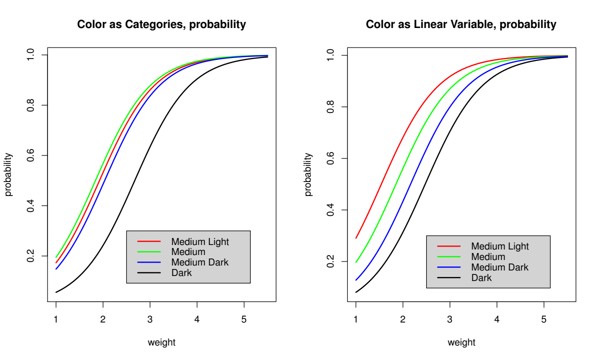
```
> linear=unclass(color)  #  convert back to integer levels
> fit2.3=glm(y ~ weight + linear, family=binomial)
> summary(fit2.3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0316     1.1161  -1.820   0.0687 .
weight        1.6531     0.3825   4.322 1.55e-05 ***
linear       -0.5142     0.2234  -2.302   0.0213 *
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 190.27  on 170  degrees of freedom
AIC: 196.27
```

Testing the significance of color via $H_0 : \beta_2 = 0$ for this model

- Via Wald test, p-value = 0.0213
- Via LRT, $G^2 = 195.74 - 190.27$ with 1 df and p-value = 0.0193637

Example (continued)

Color	df	LRT p-value
Qualitative	3	0.07
Binary (dark vs. non-dark)	1	0.01
Quantitative	1	0.02



We learned

- CI on linear combination of parameter coefficients
- Multiple comparison problem and adjusting individual confidence levels
- Treating ordinal variables as qualtitative and quantitative