Analysis of Deviance and Hypothesis Tests for Logistic Regression

STAT 340: Applied Regression

Example: Crab species identification

We will work with a data set about Leptograpsus crabs originally presented in

Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.

They have also been discussed previously in

Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

There are two species of this crab; we will examine the effect of certain physical dimenions on crab species. The data we are working with contains 5 morphological measurements on 200 crabs, 100 each of two species of Leptograpsus crabs collected at Fremantle, W. Australia.

The variables in this data set are as follows:

- `sp`: species - "B" or "O" for blue or orange.
- `sex`: the crab's sex
- `FL`: frontal lobe size (mm).
- `RW`: rear width (mm).
- `CL`: carapace length (mm).
- `CW`: carapace width (mm).
- `BD`: body depth (mm).

Hypothesis Testing for a Single Coefficient

```
crabs_full <- glm(sp_01 ~ sex + FL + CL, data=crabs, family = binomial)
summary(crabs_full)
```

```
##
## Call:
## glm(formula = sp_01 ~ sex + FL + CL, family = binomial, data = crabs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.59086  -0.06773  -0.00006   0.01268   2.06787
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.5800     3.6551  -4.262 2.02e-05 ***
## sexM          2.3899     0.9934   2.406   0.0161 *
## FL           13.0992     2.8309   4.627 3.71e-06 ***
## CL           -5.8564     1.2697  -4.612 3.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.259  on 199  degrees of freedom
## Residual deviance:  37.751  on 196  degrees of freedom
## AIC: 45.751
##
## Number of Fisher Scoring iterations: 9
```

Recall from your reading, the hypothesis test for a single coefficient is:

$H_0 : \beta_j = \beta_j^{(0)}$ versus $H_A : \beta_j \neq \beta_j^{(0)}$, and the corresponding test statistic is

$$Z_0 = \frac{\hat{\beta}_j - \beta_j^{(0)}}{SE(\hat{\beta}_j)}$$

where $Z_0 \sim Normal(0, 1)$. This is a Wald statistic.

The corresponding confidence interval is

$$\hat{\beta}_j - z_{\alpha/2} \times SE(\hat{\beta}_j),$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution with probability $\alpha/2$ to the right.

Hypothesis test 1: $H_0 : \beta_{sex} = 0$ **vs.** $H_A : \beta_{sex} \neq 0$

Hypothesis test 2: $H_0 : \beta_{sex} = 1$ **vs.** $H_A : \beta_{sex} \neq 1$

Model Comparisons and Sequential Tests

Likelihood ratio test

Model 1 (Full model)

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k$$

Model 2 (Reduced model)

$$logit(\pi) = \beta_0 + 0x_1 + 0x_2 + \cdots + 0x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k = \beta_0 + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k$$

- Likelihood ratio test statistic:

$$G_0^2 = 2[\log(L_1) - \log(L_0)]$$

where L_1 is the maximized likelihood for the full model and L_0 is the maximized likelihood for the reduced model.

Under the null hypothesis ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$), $G_0^2 \approx \chi_q^2$ for large n.

```
## Fit null (intercept only) model
crabs_null <- glm(sp_01 ~ 1, data=crabs, family = binomial)
## Print summary
summary(crabs_null)
```

```
##
## Call:
## glm(formula = sp_01 ~ 1, family = binomial, data = crabs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.177  -1.177   0.000   1.177   1.177
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0000     0.1414       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 277.26  on 199  degrees of freedom
## AIC: 279.26
##
## Number of Fisher Scoring iterations: 2
```

```
## Calculate log likelihood for the full model
logL1 <- logLik(crabs_full)

## Calculate the log likelihood for the null model
logL0 <- logLik(crabs_null)

## Calculate the likelihood ratio test statistic (this is G_0^2)
2*(logL1-logL0)
```

```
## 'log Lik.' 239.5077 (df=4)
```

Analysis of Deviance

```
## Compare full model and null (intercept only) model
anova(crabs_null, crabs_full)
```

```
## Analysis of Deviance Table
##
## Model 1: sp_01 ~ 1
## Model 2: sp_01 ~ sex + FL + CL
##   Resid. Df Resid. Dev Df Deviance
## 1       199    277.259
## 2       196     37.751  3   239.51
```

```
## Fit a reduced model (that is not the intercept only model, but is still nested)
crabs_red <- glm(sp_01 ~ FL, data=crabs, family=binomial)

## Compare full model and reduced model (nested models)
anova(crabs_red, crabs_full)
```

```
## Analysis of Deviance Table
##
## Model 1: sp_01 ~ FL
## Model 2: sp_01 ~ sex + FL + CL
##   Resid. Df Resid. Dev Df Deviance
## 1       198    235.485
## 2       196     37.751  2   197.73
```

```
## Compare three models
anova(crabs_null, crabs_red, crabs_full)
```

```
## Analysis of Deviance Table
##
## Model 1: sp_01 ~ 1
## Model 2: sp_01 ~ FL
## Model 3: sp_01 ~ sex + FL + CL
##   Resid. Df Resid. Dev Df Deviance
## 1       199    277.259
## 2       198    235.485  1   41.774
## 3       196     37.751  2  197.733
```

Model Selection

- We can use analysis of deviance (similar to analysis of variance) from above to compare *nested* models (one model is simplified version of the other).
- We can also use various information criterion, which are computed as functions of the likelihood. These can be used to compare non-nested models.
  - $AIC = -2(\log L) + 2k$, where k is the number of parameters in the model
  - $BIC = -2(\log L) + k \times \log(n)$, where k is the number of parameters in the model

In both cases, smaller is better, and an AIC or BIC value has no meaning unless it is compared to that from another model.

```
## print summary for reduced model
summary(crabs_red)
```

```
##
## Call:
## glm(formula = sp_01 ~ FL, family = binomial, data = crabs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93977  -1.02847   0.02585   0.94443   2.03821
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.67278    0.83671  -5.585 2.34e-08 ***
## FL           0.29994    0.05278   5.683 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.26  on 199  degrees of freedom
## Residual deviance: 235.48  on 198  degrees of freedom
## AIC: 239.48
##
## Number of Fisher Scoring iterations: 3
```

```
## calculate AIC
AIC(crabs_red)
```

```
## [1] 239.4846
```

```
## calculate BIC
BIC(crabs_red)
```

```
## [1] 246.0812
```

References:

N.A. Campbell and R.J. Mahon (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.

J. Fox. 2016. *Applied Regression Analysis and Generalized Linear Models*, 3rd Edition. Sage.

J. Fox and S. Weisberg. 2019. *An R Companion to Applied Regression*, 3rd Edition. Sage.

F. Ramsey and D. Schafer. 2013. *The Statistical Sleuth: A Course in Methods of Data Analysis*, 3rd Edition. Cengage.