Likelihoods

SDS 291 - Multiple Regression

4/22/2020

Likelihoods

$$L = \prod \hat{\pi}^{y_i} \times (1 - \hat{\pi})^{1-y_i}$$

Titanic Data

Table of Sex and Survival

```r
gmodels::CrossTable(Titanic$Sex, Titanic$Survived, prop.chisq = FALSE, prop.c = FALSE, prop.t = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1313 
## 
## 
##              | Titanic$Survived 
##  Titanic$Sex |         0 |         1 | Row Total | 
## -------------|-----------|-----------|-----------|
##       female |       154 |       308 |       462 | 
##              |     0.333 |     0.667 |     0.352 | 
## -------------|-----------|-----------|-----------|
##         male |       709 |       142 |       851 | 
##              |     0.833 |     0.167 |     0.648 | 
## -------------|-----------|-----------|-----------|
## Column Total |       863 |       450 |      1313 | 
## -------------|-----------|-----------|-----------|
## 
## 
```

Null Deviance

What if *everyone* had the same probability of survival, regardless of their sex? Calculate the π for the total sample, irrespective of sex and calculate the Log-Likelihood there.

```
TitNullL = (0.343^450) * (0.657^863)
TitNulllog_L = 450*(log(0.343)) + 863*(log(0.657))
TitNullNeg2LL = -2*(TitNulllog_L)
cbind(`-2LL` = TitNullNeg2LL, LogLikelihood = TitNulllog_L, Likelihood = TitNullL)
```

```
##          -2LL LogLikelihood Likelihood
## [1,] 1688.065     -844.0327          0
```

Residual Deviance

The alternative is that there *is* some relationship between probability of survival and sex.

```
TitL = (0.167^142) * (0.833^709) * (.667^308) * (.333^154)
Titlog_L = 142*(log(0.167)) + 709*(log(.833)) + 308*(log(.667)) + 154*(log(0.333))
TitNeg2LL = -2*(Titlog_L)
cbind(`-2LL` = TitNeg2LL, LogLikelihood = Titlog_L, Likelihood = TitL)
```

```
##          -2LL LogLikelihood    Likelihood
## [1,] 1355.531     -677.7654 4.469034e-295
```

Regression Models

Do you get the same -2LogLikelihood as the Null and Residual Deviance statistics in the regression model below?

```
m_titanic<-glm(Survived~SexCode, data=Titanic, family=binomial)
summary(m_titanic)
```

```
##
## Call:
## glm(formula = Survived ~ SexCode, family = binomial, data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4823  -0.6042  -0.6042   0.9005   1.8924
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.60803    0.09194  -17.49   <2e-16 ***
## SexCode      2.30118    0.13488   17.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1688.1  on 1312  degrees of freedom
## Residual deviance: 1355.5  on 1311  degrees of freedom
## AIC: 1359.5
##
## Number of Fisher Scoring iterations: 4
```

G Statistic or Drop-in-Deviance test

```
library(lmtest)
lrtest(m_titanic)
```

```
## Likelihood ratio test
##
## Model 1: Survived ~ SexCode
## Model 2: Survived ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -677.77
## 2   1 -844.03 -1 332.53  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Can you identify/calculate the G test statistic from the information given in the regression summary?

You can use the code above to test whether G is statistically signficant. Or the anova function:

```r
anova(m_titanic, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     1312     1688.1
## SexCode  1   332.53      1311     1355.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatively, you can use the general framework of `R` code for `1-pchisq(G,df)` to calculate the p-value where G is the G test statistic and df is the degrees of freedom in the difference between the null and alternative models (in this case, think about how many $\hat{\beta}$ there are that are different between the null and alternative models – the number is the degree of freedom).

```r
G <- TitNullNeg2LL - TitNeg2LL
1-pchisq(G,1)
```

```
## [1] 0
```

Whickham Data - Smoking and Survival

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1314
##
##
##                  | Whickham$outcome
## Whickham$smoker |     Alive |      Dead | Row Total |
## -----------------|-----------|-----------|-----------|
##               No |       502 |       230 |       732 |
##                  |     0.686 |     0.314 |     0.557 |
## -----------------|-----------|-----------|-----------|
##              Yes |       443 |       139 |       582 |
##                  |     0.761 |     0.239 |     0.443 |
## -----------------|-----------|-----------|-----------|
##     Column Total |       945 |       369 |      1314 |
## -----------------|-----------|-----------|-----------|
##
##
```

Do it by Hand

Practice with the example above to calculate the null and residual deviance from this 2-by-2 table.

Regression Modeling

Can you get the same example as the regression model?

Calculate and Test G

Calculate G by hand and evaluate that value on the χ^2 distribution with 1 df: (https://gallery.shinyapps.io/dist_calc/)

Use one of the functions above to have R calculate it for you.

Try both!