Lab07 - Multiple Logistic Regression Inference

Rosamonde Jiang

Loading Packages

Run the code chunk below to load packages needed for this lab.

```
library(readr)
library(dplyr)
```

```
## Registered S3 method overwritten by 'dplyr':
##   method               from
##   as.data.frame.tbl_df tibble

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
library(caret)
```

```
## Loading required package: lattice
```

Refugees

In this lab we will examine data originally presented in

Greene and Shaffer (1992). Leave to appeal and leave to commence judicial review in Canada's refugee-determination system: Is the process fair? *International Journal of Refugee Law*, 4:71-83.

The data were discussed again in

Fox (1997). Applied Regression Analysis, Linear Models, and Related Methods. Sage Publications, London.

The following description of the data is from Fox (1997).

"Greene and Shaffer (1992) analyzed decisions by the Canadian Federal Court of a Appeal on cases filed by refugee applicants who had been turned down by the Immigration and refugee Board. . . . Restricting our attention to the 10 (of 23) judges who were present on the court during the entire period of the study, and to countries of origin that produced at least 20 appeals during this period, we shall elaborate Green and Shaffer's analysis using a logistic regression. The dependent variable is whether or not leave was granted to appeal the decision of the Refugee Board. We shall examine a random subsample of cases for which an independent expert rted the merit of the case. (The judge does not decide whether the applicant is granted refugee status; if the case has any merit, an appeal should be granted.) ... The principle object of the analysis is to determine whether the substantial differences among the judges in their rates of granitng leave to appeal can be explained by differences in characteristics of the cases [they heard]. [T]he cases were assigned to the judges not at random, but on a rotating basis."

The following R code reads the data in and does some minimal pre-processing. The variables in the data set are as follows:

- `case_id`: a unique identifier for each case
- `judge`: the name of the judge who heard the case
- `origin`: the country of origin of the refugee applicant
- `independent_decision`: the recommendation made by the independent expert as to whether the case merits appeal
- `judge_decision`: the judge's decision as to whether to grant an appeal
- `case_language`: the language in which the case was heard
- `claim_location`: the location of the court in which the case was heard
- `logit_success`: The logit of the success rate for all cases from the applicant's nation decided during the period of the study (i.e., log(number of leaves granted / number of leaves denied))

```
# read_table is provided by the readr package and can be used to read files
# where columns are separated by whitespace
refugees <- read_table("http://www.evanlray.com/data/fox/Greene.dat", col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character(),
##   X6 = col_character(),
##   X7 = col_character(),
##   X8 = col_double()
## )
```

```
# set column names in refugees data frame
colnames(refugees) <- c("case_id", "judge", "origin", "independent_decision", "judge_decision", "case_la

refugees <- refugees %>%
  mutate(
    judge = factor(judge),
    origin = factor(origin),
    independent_decision = factor(independent_decision),
    judge_decision = factor(judge_decision),
    case_language = factor(case_language)
  )

head(refugees)
```

```
## # A tibble: 6 x 8
##   case_id judge origin independent_dec~ judge_decision case_language
##     <int> <fct> <fct>  <fct>            <fct>          <fct>
## 1      13 Heald Leban~ no               no             English
## 2      15 Heald Sri_L~ no               no             English
## 3      19 Heald El_Sa~ no               yes            English
## 4      30 MacG~ Czech~ no               yes            French
## 5      36 Desj~ Leban~ yes              yes            French
```

```
## 6      42 Stone Leban~ yes              yes            English
## # ... with 2 more variables: claim_location <chr>, logit_success <dbl>
```

Problem 1: Fit a model with `judge_decision` as the response variable and `judge`, `independent_decision`, `case_language`, `claim_location`, and `logit_success` as explanatory variables. Examine a summary of the model fit. Based on two separate hypothesis tests, does it seem like the `claim_location` variable is important for predicting the judge's decision?

```r
fit1 <- train(
  form=judge_decision ~ judge + independent_decision + case_language + claim_location+logit_success,
  data=refugees,
  family="binomial",
  method="glm",
  trControl=trainControl(method="none")
  )

summary(fit1)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9155  -0.7098  -0.3854   0.7401   2.7117
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.51916    0.68266   0.760 0.446962
## judgeHeald              -1.36324    0.53655  -2.541 0.011062 *
## judgeHugessen           -1.49779    0.52893  -2.832 0.004630 **
## judgeIacobucci          -2.70031    0.72730  -3.713 0.000205 ***
## judgeMacGuigan          -1.28781    0.46167  -2.789 0.005280 **
## judgeMahoney            -0.84209    0.53489  -1.574 0.115413
## judgeMarceau             1.07194    0.59673   1.796 0.072435 .
## judgePratte             -2.00107    0.59556  -3.360 0.000779 ***
## judgeStone              -1.66145    0.55652  -2.985 0.002832 **
## judgeUrie               -0.07157    0.75393  -0.095 0.924373
## independent_decisionyes  1.40494    0.27475   5.114 3.16e-07 ***
## case_languageFrench     -0.19384    0.60281  -0.322 0.747785
## claim_locationother      1.19430    0.67761   1.763 0.077981 .
## claim_locationToronto    0.94914    0.60813   1.561 0.118579
## logit_success            1.60878    0.30155   5.335 9.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 467.09  on 383  degrees of freedom
## Residual deviance: 355.83  on 369  degrees of freedom
## AIC: 385.83
```

```
##
## Number of Fisher Scoring iterations: 5
```

based on the result, it shows that claim_location is not important for prediction ## Problem 2: The real way to answer the question posed above is with a single test that compares the full model fit above with a reduced model that does not include the `claim_location` variable. Perform this test now. What is your conclusion?

```r
fit_reduced <- train(
  form=judge_decision ~ judge + independent_decision + case_language + logit_success,
  data=refugees,
  family="binomial",
  method="glm",
  trControl=trainControl(method="none")
  )
anova(fit_reduced$finalModel, fit1$finalModel, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: .outcome ~ judgeHeald + judgeHugessen + judgeIacobucci + judgeMacGuigan +
##     judgeMahoney + judgeMarceau + judgePratte + judgeStone +
##     judgeUrie + independent_decisionyes + case_languageFrench +
##     logit_success
## Model 2: .outcome ~ judgeHeald + judgeHugessen + judgeIacobucci + judgeMacGuigan +
##     judgeMahoney + judgeMarceau + judgePratte + judgeStone +
##     judgeUrie + independent_decisionyes + case_languageFrench +
##     claim_locationother + claim_locationToronto + logit_success
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       371     359.01
## 2       369     355.83  2   3.1787   0.2041
```

p-value is 0.2041. We fail to reject the null hypothesis. claim_location is important for prediction ## Problem 3: After controlling for an independent expert's recommendation, the language the case was heard in, and the overall success rate for all cases from the applicant's origin nation, are there statistically significant differences in the chances of granting an appeal for different judges? To answer this question, fit a reduced model that includes only `independent_decision`, `case_language`, and `logit_success` as explanatory variables, then conduct a hypothesis test comparing this model to the one from problem 2 that also includes `judge`. What is your conclusion?

```r
fit_reduced2 <- train(
  form = judge_decision ~ independent_decision + case_language + logit_success,
  data = refugees,
  family="binomial",
  method="glm",
  trControl=trainControl(method="none")
)
anova(fit_reduced2$finalModel,fit_reduced$finalModel, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: .outcome ~ independent_decisionyes + case_languageFrench + logit_success
## Model 2: .outcome ~ judgeHeald + judgeHugessen + judgeIacobucci + judgeMacGuigan +
##     judgeMahoney + judgeMarceau + judgePratte + judgeStone +
##     judgeUrie + independent_decisionyes + case_languageFrench +
##     logit_success
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1       380     406.54
## 2       371     359.01  9   47.534 3.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(fit_reduced2)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6589  -0.8362  -0.5264   1.0393   2.5949
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.1438     0.3045   0.472  0.63684
## independent_decisionyes   1.1705     0.2456   4.765 1.89e-06 ***
## case_languageFrench      -0.7566     0.2822  -2.681  0.00733 **
## logit_success             1.3005     0.2634   4.936 7.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 467.09  on 383  degrees of freedom
## Residual deviance: 406.54  on 380  degrees of freedom
## AIC: 414.54
##
## Number of Fisher Scoring iterations: 4
```

p-value is 3.12*10^(-7), this shows that the variable "judge" is important for prediction. ## Problem 4: In your final model fit (whichever seems best based on the hypothesis tests you conducted above), what is the interpretation of the estimated coefficient for `logit_success`?

Don't worry about the units of `logit_success` (your answer can start with "If logit_success increases by one unit…").

If logit_success increases by one unit, the odds of the judge to grant an appeal will be e^1.3005=3.67 times higher.

Problem 5: If you were an immigrant applying for refugee status, would you want your case to be heard by the judge named Iacobucci? Explain by interpreting one of the coefficients in your final model fit.

The hypothesis test shows that this judge is statistically significant for prediction. However, the coefficient is negative. This means, the odds of the judge to grant an apeal for this judge will be e^-2.77=0.063 times higher than the odds of the judge to grant an apeal for the other judges.