Lab05 - Multiple Logistic Regression Inference

Solutions

Loading Packages

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

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

```
##
## 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)
```

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))

```r
# 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_double(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character(),
##   X6 = col_character(),
##   X7 = col_character(),
##   X8 = col_double()
## )
```

```r
# 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
##     <dbl> <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
fit <- glm(judge_decision ~ judge + independent_decision + case_language + claim_location + logit_succes
  data = refugees,
  family = binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = judge_decision ~ judge + independent_decision +
##     case_language + claim_location + logit_success, family = binomial,
##     data = refugees)
##
## 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
```

```r
str(refugees)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 384 obs. of  8 variables:
##  $ case_id             : num  13 15 19 30 36 42 45 46 51 52 ...
##  $ judge               : Factor w/ 10 levels "Desjardins","Heald",..: 2 2 2 5 1 9 8 5 5 8 ...
```

```
##  $ origin               : Factor w/ 17 levels "Argentina","Bulgaria",..: 11 17 5 4 11 11 7 16 16 3 ..
##  $ independent_decision: Factor w/ 2 levels "no","yes": 1 1 1 1 2 2 1 1 2 1 ...
##  $ judge_decision       : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 1 1 ...
##  $ case_language        : Factor w/ 2 levels "English","French": 1 1 1 2 2 1 1 1 2 1 ...
##  $ claim_location       : chr  "Toronto" "Toronto" "Toronto" "Montreal" ...
##  $ logit_success        : num  -1.099 -0.754 -1.046 0.405 -1.099 ...
```

```r
refugees %>% distinct(claim_location)
```

```
## # A tibble: 3 x 1
##   claim_location
##   <chr>
## 1 Toronto
## 2 Montreal
## 3 other
```

In this full model fit, let's use β_{12} and β_{13} to denote the coefficients labelled in the R output as `claim_locationother` and `claim_locationToronto`, respectively. Since the `claim_location` variable is a categorical variable, these coefficients are for indicator variables indicating whether a particular observation's claim location was "other" or "Toronto". For example,

$$\texttt{claim\_locationToronto} = \begin{cases} 1 & \text{if the claim location for observation number } i \text{ is Toronto} \\ 0 & \text{otherwise} \end{cases}$$

From the R output above, we can see that the `claim_location` variable has three levels: "Montreal", "Toronto", and "other". The baseline category is therefore "Montreal". The two coefficients here are describing a difference in odds of an appeal being granted between a location of "other" and "Montreal", and between "Toronto" and "Montreal".

We might conduct two separate hypothesis tests about these differences.

For our first hypothesis test, the hypotheses are:

H_0: There is no difference in odds of approving an appeal for claims in Montreal and claims in another location (other than Toronto) after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_{12} = 0$.

H_A: There is a difference in odds of approving an appeal for claims in Montreal and claims in another location (other than Toronto) after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_{12} \neq 0$.

From the R output, the p-value for this test is 0.077981. This is larger than a typical significance level threshold such as $\alpha = 0.05$. We do not find statistically significant evidence that there is a difference in odds of granting appeals for cases heard in Montreal or in another location (other than Toronto).

For our second hypothesis test, the hypotheses are:

H_0: There is no difference in odds of approving an appeal for claims in Montreal and claims in Toronto after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_{13} = 0$.

H_A: There is a difference in odds of approving an appeal for claims in Montreal and claims in Toronto after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_{13} \neq 0$.

From the R output, the p-value for this test is 0.118579. This is larger than a typical significance level threshold such as $\alpha = 0.05$. We do not find statistically significant evidence that there is a difference in odds of granting appeals for cases heard in Montreal or in Toronto.

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?

```
fit_nolocation <-  glm(judge_decision ~ judge + independent_decision + case_language + logit_success,
  data = refugees,
  family = binomial)

summary(fit_nolocation)
```

```
##
## Call:
## glm(formula = judge_decision ~ judge + independent_decision +
##     case_language + logit_success, family = binomial, data = refugees)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8761  -0.6953  -0.4094   0.7553   2.7802
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.34831    0.47646   2.830 0.004657 **
## judgeHeald              -1.26080    0.53197  -2.370 0.017786 *
## judgeHugessen           -1.52170    0.52766  -2.884 0.003929 **
## judgeIacobucci          -2.56776    0.72211  -3.556 0.000377 ***
## judgeMacGuigan          -1.25640    0.45739  -2.747 0.006017 **
## judgeMahoney            -0.75633    0.52921  -1.429 0.152957
## judgeMarceau             0.97150    0.58518   1.660 0.096878 .
## judgePratte             -2.02570    0.59596  -3.399 0.000676 ***
## judgeStone              -1.55679    0.55261  -2.817 0.004845 **
## judgeUrie               -0.06602    0.74595  -0.089 0.929474
## independent_decisionyes  1.36421    0.27204   5.015 5.31e-07 ***
## case_languageFrench     -0.98117    0.36915  -2.658 0.007861 **
## logit_success            1.51438    0.29324   5.164 2.41e-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: 359.01  on 371  degrees of freedom
## AIC: 385.01
##
## Number of Fisher Scoring iterations: 5
```

```
anova(fit_nolocation, fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: judge_decision ~ judge + independent_decision + case_language +
```

```
##     logit_success
## Model 2: judge_decision ~ judge + independent_decision + case_language +
##     claim_location + logit_success
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       371     359.01
## 2       369     355.83  2   3.1787   0.2041
```

Now we conduct a simultaneous hypothesis test about the two coefficients:

H_0: There is no difference in odds of granting an appeal based on the location where the claim is heard after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_{12} = \beta_{13} = 0$.

H_A: There is a difference in odds of granting an appeal based on the location where the claim is heard after accounting for the judge who hears the case, the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. At least one of β_{12} or β_{13} is not equal to 0.

From the R output above, the p-value for this test is 0.2041. We fail to reject the null hypothesis; the data do not provide statistically significant evidence that there is a difference in the odds that an appeal will be granted for claims heard in different locations.

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_nojudge <- glm(judge_decision ~ independent_decision + case_language + logit_success,
  data = refugees,
  family = binomial)

summary(fit_nojudge)
```

```
##
## Call:
## glm(formula = judge_decision ~ independent_decision + case_language +
##     logit_success, family = binomial, data = refugees)
##
## 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
```

```r
anova(fit_nojudge, fit_nolocation, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: judge_decision ~ independent_decision + case_language + logit_success
## Model 2: judge_decision ~ judge + independent_decision + case_language +
##     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
```

Based on the model fit from problem 2 (without claim location, but with judge as an explanatory variable), let's label the judge-specific coefficients as $\beta_1, \dots, \beta_9$.

Our hypotheses are:

H_0: there is no association between the judge hearing a case and the odds that an appeal will be granted after accounting for the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. $\beta_1 = \cdots = \beta_9 = 0$

H_A: There is an association between the judge hearing a case and the odds that an appeal will be granted after accounting for the recommendation of an independent expert, the language in which the case is heard, and the overall success rate for cases from the applicant's country of origin. At least one of β_1, ..., β_9 is not equal to 0.

From the R output above, the p-value for this test is 3.12×10^{-07}. We reject the null hypothesis; the data provide statistically significant evidence that the odds of an appeal being granted are different for different judges, after accounting for an independent expert's recommendation, the language the case was heard in, and the overall success rate for cases from the applicant's country of origin.

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…").

I will use the model from Problem 2, which includes the judge as an explanatory variable, but not the claim location. The coefficient estimate for `logit_success` in that model was 1.51438. In order to interpret this in terms of odds, we exponentiate this:

```r
exp(1.51438)
```

```
## [1] 4.546601
```

If logit_success increases while one unit while holding all other covariates in the model fixed, the estimated odds that an appeal will be granted increase by a multiplicative factor of 4.5.

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.

```r
refugees %>% distinct(judge)
```

```
## # A tibble: 10 x 1
##    judge
##    <fct>
##  1 Heald
##  2 MacGuigan
##  3 Desjardins
##  4 Stone
##  5 Pratte
##  6 Hugessen
##  7 Iacobucci
##  8 Mahoney
##  9 Urie
## 10 Marceau
```

The first judge alphabetically is Desjardins; this is used as the reference category for this variable since we did not specify an ordering (note that all other judges' names appear labelling coefficient estimates in the model summary output). Therefore, the estimated coefficients describe differences in the odds of granting an appeal for other judges relative to Desjardins.

The estimated coefficient for Iacobucci is -2.56776; exponentiating, we find that $\exp(-2.56776) = 0.0767$. We estimate that if two cases with the same recommendation by an independent expert, the same language, and same overall success rate from the applicant's country of origin are heard by Desjardins and by Iacobucci, the odds that an appeal will be granted in the case heard by Iacobucci are only about 7% of the odds that an appeal will be granted in the case heard by Desjardins. If I were an immigrant, I would definitely not want my case to be heard by Iacobucci.