Course: Multivariate statistics (AUT23)

Chapter 6: Logistic regression

6.12 Time to practice on your own

6.11.1 Exercise 1: probability of hiring a consultant according to campaign personalization

For instance, we are interested in measuring the likelihood of hiring a consultant (Y) explained by personalized style of campaigning (X). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

We can look at the likelihood of hiring of consultant (B11) by the level of campaign personalization (where B6 is recoded as 0=attention to the party and 10=attention to the candidate):

- Show the code

Calculate the odds of hiring a consultant for a very personalized campaign (personalization = 10):

- Interpretation

Now, calculate the odds of hiring a consultant for a very low personalized campaign (personalization = 0):

- Interpretation

The logit of the dependent variable (Y) is estimated by the following equation:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$$

The logit does not indicate the probability that an event occurs. Apply the necessary transformation to know this probability (prob(Y=1)):

- Answer

Let's go back to our example and run the logistic regression:

- Show the code

Coefficients in the above output are log odds: 0.22 means that by augmenting the personalization of one point, log odds change by 0.22.

Now, assess the odds of hiring a consultant for a very personalized campaign (personalization=10):

- Interpretation

6.11.2 Exercise 2: predict the reliance of social media as campaigning tool

Using the same dataset, let's investigate the following question: how does the level of campaign personalization and the fact of being affiliated to a governmental party, and being an incumbent affect the reliance of social media as campaigning tool?

In this scenario, the binary outcome is whether politicians rely on social media (combination of B4m and B4p) and the predictors are personalization (B6), being affiliated to a governmental party (based on T9), and being an incumbent (T11c).

Let's prepare the data, including the selection and recoding of the relevant variables:

- Show the code

Now, we can conduct logistic regression and interpret the findings. Recall that, for log odds, we interpret only the sign of the coefficients (positive/negative). Coefficients smaller than 1 suggests a negative effect (negative log odds) and coefficients larger than 1 suggest positive effect (positive log odds). You can also transform to percentages using the formula 100*(OR-1):

- Show the code

Coefficients smaller than 1 suggests a negative effect (negative log odds) and coefficients larger than 1 suggest positive effect (positive log odds). You can also transform to percentages using the formula 100*(OR-1).

- Show the code
- Interpretation

The marginal effects indicate a change in predicted probability as X increases by 1. For categorical predictors, you have to take the predicted probability of the group A minus the predicted probability of the group B.

There are different ways of calculating predicted probabilities. In the social sciences, the most commonly used are Adjusted Predictions at the Means (APMs). For instance, we can assess the predicted probabilities of using social media for political incumbents, when the personalization level is at the mean and for incumbent not affiliated to a party in government.

- Show the code

Nota bene: Marginal Effects at the Means (MEMs) are calculated by taking the difference of two APMs. Let's also calculate the predicted probabilities of using social media for political non-incumbents, when the personalization level is at the mean and for politicians not affiliated to a party in government. Then, calculate the difference between both predicted probabilities:

- Show the code

In logistic regressions, there is no such R-squared value for general linear models. Instead, we can calculate a metric known as McFadden's R-Squared, which ranges from 0 to just under 1, with higher values indicating a better model fit. We use the following formula to calculate McFadden's R-Squared:

- Show the code

Chapter 6: Logistic regression (answers)

6.12 Time to practice on your own

6.11.1 Exercise 1: probability of hiring a consultant according to campaign personalization

For instance, we are interested in measuring the likelihood of hiring a consultant (Y) explained by personalized style of campaigning (X). To do so, we will rely on the data covering the Swiss part of the Comparative Candidate Survey. We will be using the Selects 2019 Candidate Survey.

We can look at the likelihood of hiring of consultant (B11) by the level of campaign personalization (where B6 is recoded as 0=attention to the party and 10=attention to the candidate):

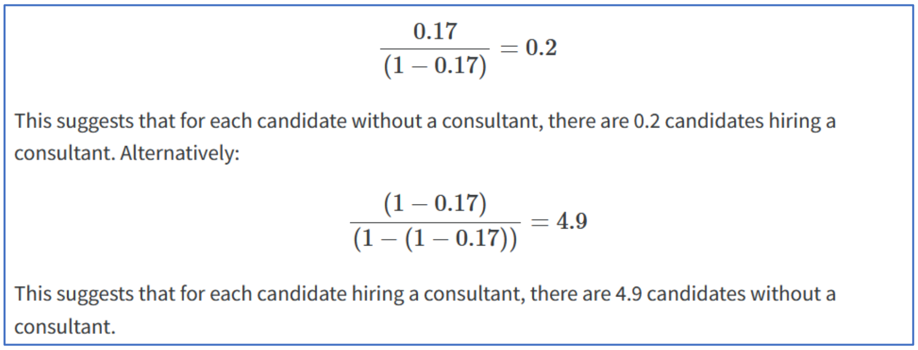
➢ Show the code

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
        "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),
        use.value.labels = F,
        to.data.frame = T)
sel <- db |>
  dplyr::select(B11,B12,B6) |>
  stats::na.omit() |>
  dplyr::rename("consultant"="B11",
        "budget"="B12",
        "personalization"="B6") |>
  plyr::mutate(budget=as.numeric(as.character(as.character(budget))))
sel$consultant <- ifelse(sel$consultant==1,1,0)
# keep candidates with a budget<100'000
sel <- sel[sel$budget<100000,]
# reverse the scale: higher values = higher personaliz.
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-11)*(-1)
# mean by level of personalization
p = aggregate(sel$consultant, by=list(sel$personalization), FUN=mean)
colnames(p) = c("personalization","mean")
```

```
p

##   personalization       mean

## 1               1 0.035398230

## 2               2 0.006896552

## 3               3 0.031690141

## 4               4 0.107279693

## 5               5 0.086956522

## 6               6 0.117391304

## 7               7 0.120000000

## 8               8 0.119047619

## 9               9 0.186440678

## 10             10 0.173913043

## 11             11 0.200000000
```

Calculate the odds of hiring a consultant for a very personalized campaign (personalization = 10):

- ➢ Interpretation



$$\frac{0.17}{(1-0.17)} = 0.2$$

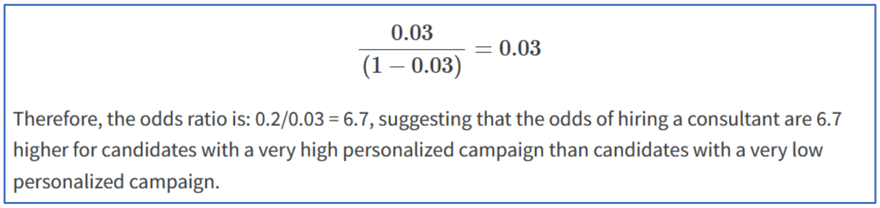
This suggests that for each candidate without a consultant, there are 0.2 candidates hiring a consultant. Alternatively:

$$\frac{(1-0.17)}{(1-(1-0.17))} = 4.9$$

This suggests that for each candidate hiring a consultant, there are 4.9 candidates without a consultant.

Now, calculate the odds of hiring a consultant for a very low personalized campaign (personalization = 0):

- ➢ Interpretation



$$\frac{0.03}{(1-0.03)} = 0.03$$

Therefore, the odds ratio is: 0.2/0.03 = 6.7, suggesting that the odds of hiring a consultant are 6.7 higher for candidates with a very high personalized campaign than candidates with a very low personalized campaign.

The logit of the dependent variable (Y) is estimated by the following equation:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$$

The logit does not indicate the probability that an event occurs. Apply the necessary transformation to know this probability (prob(Y=1)):

- ➢ Answer

$$proba = \frac{exp^{logit}}{1 + exp^{logit}}$$

Let's go back to our example and run the logistic regression:

- ➢ Show code

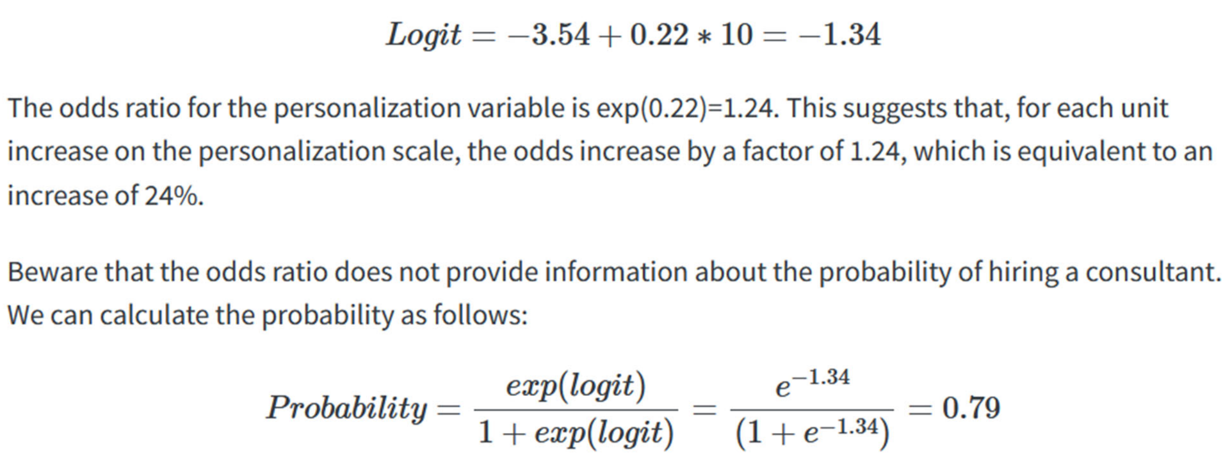
```
model2 <- glm(consultant ~ personalization,
              data=sel,
              family="binomial")
summary(model2)
##
## Call:
## glm(formula = consultant ~ personalization, family = "binomial",
##     data = sel)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.54005    0.19321  -18.32  < 2e-16 ***
## personalization  0.22052    0.03132    7.04 1.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1007.64  on 1854  degrees of freedom
## Residual deviance:  957.86  on 1853  degrees of freedom
## AIC: 961.86
##
```

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

Coefficients in the above output are log odds: 0.22 means that by augmenting the personalization of one point, log odds change by 0.22.

Now, assess the odds of hiring a consultant for a very personalized campaign (personalization=10):

➢ Interpretation



$$Logit = -3.54 + 0.22 * 10 = -1.34$$

The odds ratio for the personalization variable is exp(0.22)=1.24. This suggests that, for each unit increase on the personalization scale, the odds increase by a factor of 1.24, which is equivalent to an increase of 24%.

Beware that the odds ratio does not provide information about the probability of hiring a consultant. We can calculate the probability as follows:

$$Probability = \frac{exp(logit)}{1 + exp(logit)} = \frac{e^{-1.34}}{(1 + e^{-1.34})} = 0.79$$

6.11.2 Exercise 2: predict the reliance of social media as campaigning tool

Unsing the same dataset, let's investigate the following question: how does the level of campaign personalization and the fact of being affiliated to a governmental party, and being an incumbent affect the reliance of social media as campaigning tool?

In this scenario, the binary outcome is whether politicians rely on social media (combination of B4m and B4p) and the predictors are personalization (B6), being affiliated to a governmental party (based on T9), and being an incumbent (T11g).

Let's prepare the data, including the selection and recoding of the relevant variables:

➢ Show the code

```
library(foreign)

db <- read.spss(file=paste0(getwd(),

          "/data/1186_Selects2019_CandidateSurvey_Data_v1.1.0.sav"),

         use.value.labels = F,

         to.data.frame = T)

sel <- db |>

 dplyr::select(B4m,B4p,T9,B6,T11c) |>

 stats::na.omit() |>
```

```
  dplyr::rename("facebook"="B4m",
                "twitter"="B4p",
                "party"="T9",
                "personalization"="B6",
                "incumbentNC"="T11c")
# reliance on social media
sel$twitter=ifelse(sel$twitter>0,1,0)
sel$facebook=ifelse(sel$facebook>0,1,0)
sel$SMuse=ifelse(sel$facebook==1 | sel$twitter==1, 1, 0)
sel$SMuse=as.factor(sel$SMuse)
# party in government
sel$in_gov=ifelse(sel$party %in% c(1,2,3,4,7), 1, 0)
sel$in_gov=as.factor(sel$in_gov)
# personalization (invert scale)
sel$personalization <- as.numeric(as.character(sel$personalization))
sel$personalization <- (sel$personalization-10)*(-1)
# incumbent
sel$incumbentNC <- as.factor(sel$incumbentNC)
# head
head(sel[,c(3:ncol(sel))])
##   party personalization incumbentNC SMuse in_gov
## 1    11               0           0     0      0
## 2    11               5           0     1      0
## 3    11               3           1     1      0
## 4    11               5           0     1      0
## 5    11               0           0     0      0
## 6    11               0           0     0      0
```

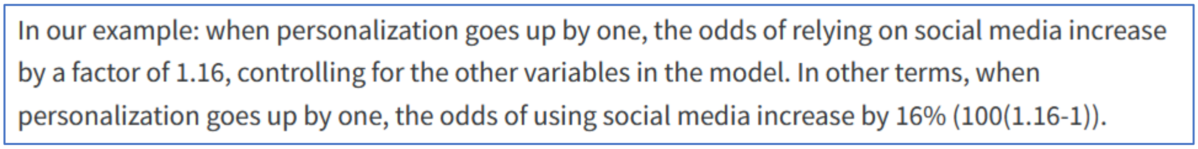
Now, we can conduct logistic regression and interpret the findings. Recall that, for log odds, we interpret only the sign of the coefficients (positive/negative). Coefficients smaller than 1 suggests a negative effect (negative log odds) and coefficients larger than 1 suggest positive effect (positive log odds). You can also transform to percentages using the formula 100*(OR-1):

- ➢ Show the code

```
mod <- glm(SMuse ~
           personalization +
           in_gov +
           incumbentNC,
         data=sel,
         family = "binomial")
summary(mod)
##
## Call:
## glm(formula = SMuse ~ personalization + in_gov + incumbentNC,
##     family = "binomial", data = sel)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.26700    0.08104   3.295 0.000986 ***
## personalization  0.16940    0.02060   8.223  < 2e-16 ***
## in_gov1          0.08525    0.10015   0.851 0.394636
## incumbentNC1     0.82037    0.38953   2.106 0.035200 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2561.7  on 2073  degrees of freedom
## Residual deviance: 2469.7  on 2070  degrees of freedom
## AIC: 2477.7
```

```
##
## Number of Fisher Scoring iterations: 4
# transformation
exp(coef(mod))
##    (Intercept) personalization       in_gov1    incumbentNC1
##       1.306044        1.184593       1.088992       2.271331
```

- ➢ Interpretation



In our example: when personalization goes up by one, the odds of relying on social media increase by a factor of 1.16, controlling for the other variables in the model. In other terms, when personalization goes up by one, the odds of using social media increase by 16% (100(1.16-1)).

The marginal effects indicate a change in predicted probability as X increases by 1. For categorical predictors, you have to take the predicted probability of the group A minus the predicted probability of the group B.

There are different ways of calculating predicted probabilities. In the social sciences, the most commonly used are Adjusted Predictions at the Means (APMs). For instance, we can assess the predicted probabilities of using social media for political incumbents, when the personalization level is at the mean and for incumbent not affiliated to a party in government.

- ➢ Show the code

```
newdata = data.frame(personalization=5,
                  in_gov="0", incumbentNC="1")
predict(mod, newdata, type="response")
##       1
## 0.8737317
```

Nota bene: Marginal Effects at the Means (MEMs) are calculated by taking the difference of two APMs. Let’s also calculate the predicted probabilities of using social media for political non-incumbents, when the personalization level is at the mean and for politicians not affiliated to a party in government. Then, calculate the difference between both predicted probabilities:

- ➢ Show the code

```
newdata2 = data.frame(personalization=5,
                   in_gov="0", incumbentNC="0")
print(paste0("for incumbents: ",
      round(predict(mod, newdata, type="response"),2),

      "; for non-incumbents: ",

      round(predict(mod, newdata2, type="response")),2))

## [1] "for incumbents: 0.87; for non-incumbents: 0.75"
```

In logistic regressions, there is no such R-squared value for general linear models. Instead, we can calculate a metric known as McFadden's R-Squared, which ranges from 0 to just under 1, with higher values indicating a better model fit. We use the following formula to calculate McFadden's R-Squared:

➢ Show the code

```
with(summary(mod), 1 - deviance/null.deviance)

## [1] 0.03592477
```