PRACTICE 5: BINARY OUTCOME

## R software possibilities:

* New method: glm(Y ~ A+B\*X, family=binomial) any formula is valid, where Y is a binary response variable and A and B are factors (qualitative variables) and X a covariate
* Be careful with the default order of factor levels :.
  + Reorder to simplify interpretation: factor(variable, levels=c(nivell1, …, nivellsk))
  + If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(nivell1, …, nivellsk),labels=c(nom1,…,nomk)).
* step( ) method in R, base on AIC (*Akaike information criteria*) can be used to assess the best model consistent to data.
* Link function by default is logit(probability)=log(probability/(1-probability)).

## CASE 1: Proposed Data Set: Swiss Labor Market Participation Data

### Description

Cross-section data originating from the health survey SOMIPOPS for Switzerland in 1981. The concern is about female labor force participation for a sample of 872 women from Switzerland. The response variable is participation which is regressed on all further variables plus age square, i.e., on income, education, age, numbers of younger and older children and the factor foreign which indicates citizenship.

### Usage

data("SwissLabor")

### Format

A data frame containing 872 observations on 7 variables.

|  |  |
| --- | --- |
| participation | Factor. Did the individual participate in the labor force? |
| income | Logarithm of nonlabor income. |
| age | Age in decades (years divided by 10). |
| education | Years of formal education. |
| youngkids | Number of young children (under 7 years of age). |
| oldkids | Number of older children (over 7 years of age). |
| foreign | Factor. Is the individual a foreigner (i.e., not Swiss)? |

### Source

Journal of Applied Econometrics Data Archive. http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/

## Guided Laboratory Session:

We will be using data(“SwissLabor”) in package AER. Response variable is **participation.**

* Exploratory Analysis of data set. Determine global univariate outliers and outliers for each category of the response variable.
* Test association between pairs of variables.
* Firstly, propose and estimate with R a glm containing the gross effect of age. Is it worth to use a second order model of age; i.e., linear and quadratic terms of age included in the linear predictor? Justify your answer using a suitable statistic test (differences in deviance, for example).
* Consider all available variables to explain participation in female labor force. According to feature selection conclusions propose an overparametrized initial model consisting on no less than 2 pairs of interaction (involving factor-covariate). Compute the initial logistic model for explaining the probability of PARTICIPATE in the labor force.
* Test if interactions have statistical significance: use method anova(nestedmodel, model, test=”Chiq”).
* Apply step() on the initial model in order to simplify the model based on Akaike Information Criteria.

Diagnostic tools- Not so easy as normal models, some helpful method in R libraries are:

**library(effects)**

**plot(allEffects(m1),ask=FALSE)**

**library(car)**

**residualPlots(m1, layout=c(1, 2))**

**marginalModelPlots(m1,labels=row.names(df),id.method=abs(cooks.distance(m1)), id.n=5)**

**influenceIndexPlot(m1,label=row.names(df),vars=c("Cook", "Student","hat"), id.n=5)**

## CASE 2: - Accidents

The Highway Safety Research Center in North Carolina has studied the relationship between the severity of the damage suffered by drivers in an accident and some variables that characterize the time the accident happened. The data shown in the table pertains to 39024 accidents that occurred between 1966-72 **where only a single vehicle was involved, on rural roads with a non-drunk male driver driving within the speed limit**. We want to analyze the effect of weather conditions and the period of the day on the severity of the damage. **Positive outcome will be Severe Injury.**

|  |  |  |  |
| --- | --- | --- | --- |
| Weather  Conditions | Time of Day | Severity of Injury | |
| **Severe** | Mild |
| Good | Day | **1251** | 9026 |
| Good | Night | **2145** | 12205 |
| Bad | Day | **620** | 6621 |
| Bad | Night | **715** | 6441 |

|  |  |  |
| --- | --- | --- |
| Time of Day | Severity of Injury | |
| Severe | Mild |
| Day | *1871* | *15647* |
| Night | *2860* | *18646* |
|  | ***4731*** | ***34293*** |

P = 4731/(4731+34293)

**The logistic regression model with the Time of the Day factor has a residual deviance of 165.04. A set of models calculated on the data in the table are shown at the end, although you are supposed to calculate them using RStudio.**

1. Determine the proportion of accidents where the driver suffers severe damage. What are the odds of suffering severe damage? What is the logit of the probability of having severe damage in an accident?
2. Calculate the null logistic regression model for the positive response defined as suffering severe damage in an accident.
3. Calculate the null probit model for the positive response defined as suffering severe damage in an accident.
4. Calculate the logistic regression model with the gross effect of the ‘Time of the Day’ factor for the positive response defined as suffering severe damage in an accident.
5. Interpret the effect of adverse weather conditions in the additive logit model (m2) on the logit scale, in terms of the odds and the approximate effect on the probability scale.
6. Examine the coefficient corresponding to the dummy variable for adverse weather conditions in the m1 model. Subsequently examine the effect of bad weather on the additive m2 model. How do you interpret the fact that the coefficient for the bad weather dummy has not changed too much after introducing ‘Time of the Day’ into the model?
7. Statistically contrast whether the net effect of each of the factors is statistically significant. Order the net effect of the factors in importance.
8. Calculate the value of the generalized Pearson statistic for the additive logit model and contrast the goodness of the fit.
9. Calculate according to m2 a point estimator of the probability of severe damage in a night accident with good weather conditions. What additional information would you need to calculate a confidence interval for this probability.
10. Do you consider that there is evidence to affirm that the increase in the odds of having a severe accident when it happens at night depends on the weather conditions? Justify the answer from a contrast.
11. Calculate the confusion table corresponding to the additive model and determine from it the predictive capacity of the m2 model. Compare the figure with the predictive capacity of the null model.

**> acc$Weather<-factor(acc$Weather,levels=c("Good","Bad"))**

**> summary(acc)**

**Weather TimeDay Severity.Severe Severity.Mild**

**Good:2 Day :2 Min. : 620.0 Min. : 6441**

**Bad :2 Night:2 1st Qu.: 691.2 1st Qu.: 6576**

**Median : 983.0 Median : 7824**

**Mean :1182.8 Mean : 8573**

**3rd Qu.:1474.5 3rd Qu.: 9821**

**Max. :2145.0 Max. :12205**

**> apply(acc[,3:4],2,sum)**

**Severity.Severe Severity.Mild**

**4731 34293**

**> xtabs(cbind(Severity.Severe,Severity.Mild)~ TimeDay, data=acc)**

**TimeDay Severity.Severe Severity.Mild**

**Day 1871 15647**

**Night 2860 18646**

**>**

**> summary(m1)**

**Call: glm(formula = cbind(Severity.Severe, Severity.Mild) ~ Weather, family = binomial, data = acc)**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -1.83286 0.01848 -99.17 <2e-16 \*\*\***

**WeatherBad -0.44791 0.03416 -13.11 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Null deviance: 228.048 on 3 degrees of freedom**

**Residual deviance: 47.958 on 2 degrees of freedom**

**AIC: 86.628**

**Number of Fisher Scoring iterations: 3**

**> summary(m2)**

**Call: glm(formula = cbind(Severity.Severe, Severity.Mild) ~ Weather + TimeDay, family = binomial, data = acc)**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -1.96340 0.02705 -72.596 < 2e-16 \*\*\***

**WeatherBad -0.43000 0.03428 -12.545 < 2e-16 \*\*\***

**TimeDayNight 0.21695 0.03180 6.822 8.95e-12 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Null deviance: 228.04789 on 3 degrees of freedom**

**Residual deviance: 0.94903 on 1 degrees of freedom**

**AIC: 41.619**

**> summary(m3)**

**Call: glm(formula = cbind(Severity.Severe, Severity.Mild) ~ Weather \***

**TimeDay, family = binomial, data = acc)**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -1.97617 0.03017 -65.504 < 2e-16 \*\*\***

**WeatherBad -0.39212 0.05171 -7.583 3.38e-14 \*\*\***

**TimeDayNight 0.23746 0.03819 6.218 5.03e-10 \*\*\***

**WeatherBad:TimeDayNight -0.06733 0.06911 -0.974 0.33**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**….**

**> resid(m2,type="pearson")**

**1 2 3 4**

**-0.4230920 0.3308226 0.5980583 -0.5506987**

**> ifelse(predict(m2,type="response")>0.5,"Pred\_yes","Pred\_no")**

**1 2 3 4**

**"Pred\_no" "Pred\_no" "Pred\_no" "Pred\_no"**

# CASE 3: Interpreting odds, log-odds, odds-ratio

Agresti (2002) analyzes data related to traffic accidents according to driver’s sex, seat-belt use and urban/rural driving environment to assess the incidence of injured persons. We focus on the relationship between the use of seat-belt (Factor A) and the presence of injured persons in the car due to the accident:

|  |  |  |  |
| --- | --- | --- | --- |
| **Seat-Belt** | **Injured present**  **(positive response)** | **Otherwise** | **m** |
| ***Yes (ref)*** | 2409 | 35383 | 37792 |
| ***No*** | 3865 | 27037 | 30902 |
|  | ***6274*** | ***62420*** | ***68694*** |

1. Calculate the expected number of accidents with Injured present and Otherwise under the null hypothesis considering that ‘the probability of Injured is independet of the use of Seat-Belt’.

|  |  |  |  |
| --- | --- | --- | --- |
| **Seat-Belt**  **Factor A** | **Under H0, predicted**  ***Injured present***  **(positive response)** | **Under H0, predicted *Otherwise*** | **m** |
| ***Yes (ref)*** |  |  | 37792 |
| ***No*** |  |  | 30902 |
|  | ***6274*** | ***62420*** | ***68694*** |

P(Injured)**=*0.0913***

1. Compare observed number of injured present and the predicted number of accidents with injured people according to the null model (the one represented by H0). You have to use Generalized Pearson Statistic and Deviance Statistic, both are discrepancy measures.



Alternative formula:





1. Calculate the odds of accident with injured people (positive response) in the sample, whether using seat-belt (i=1 Yes) or not using seat-belt (i=2 No)

|  |  |  |
| --- | --- | --- |
|  | **Injured Present** | **odds** |
| ***Factor A***  ***Yes (ref)*** |  | I=1 |
| ***Factor A***  ***No*** |  | I=2 |

1. Estimate the null generalized linear model for the response Y-Injured. Use logit and probit links.

**(M0) **

**acc.l0 <-glm(cbind(ypos,yneg)~1, family=binomial(link=logit), data= dfbelt)**

**(M0) **

**acc.m0 <-glm(cbind(ypos,yneg)~1, family=binomial(link=probit), data=dfbelt)**

1. Calculate the expected probabilities for Injured Present/Otherwise using (M0) and the expected number of accidents with Injured Present/Otherwise. Repeat the process for logit and probit models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Y-1** | **Predictlogit (M0)** | | **Predict probit (M0)**  ***Injured Present*** | |
| ***Injured Present*** | **Otherwise** | ***Injured Present*:** | **Otherwise** |
| ***Factor A***  ***Yes (ref)*** |  |  |  |  |
| ***Factor A***  ***No*** |  |  |  |  |

1. Estimate the generalized linear model for the response Y-Injured when Factor A (Use of Seat-Belt) is considered. Use logit link.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Factor A** | **Sample Probabilities of Injured Present (M1)** | **Odd Injured Present (M1)** | **Log-Odd Injured Present (M1)** | **Predictions (M1)**  **(Fitted.values)** |
| ***Factor A***  ***Yes (ref)*** | **2409/37792=0.064** |  |  |  |
| ***Factor A***  ***No*** | **3865/30902=0.125** |  |  |  |

**(M1) **







**acc.l1 <-glm(cbind(ypos,yneg)~seatbelt, family=binomial(link=logit), , data=dfbelt)**

1. Estimate the generalized linear model for the response Y-Injured when Factor A (Use of Seat-Belt) is considered. Use logit link.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factor A** | **Sample Probabilities of Injured Present (M1)** | **Probit Injured Present (M1)** | **Predictions (M1)**  **(Fitted.values)** |
| ***Factor A***  ***Yes (ref)*** | **2409/37792=0.064** |  |  |
| ***Factor A***  ***No*** | **3865/30902=0.125** |  |  |

**(M1) **





**acc.m1 <-glm(cbind(ypos,yneg)~seatbelt, family=binomial(link=probit), data= dfbelt)**