Logistic Regression

NCP Seminar with Rstudio

and Andy Field‘s DSUR Logistic Regression

interaction terms in regression models

- adding interaction terms expand the model and the understanding

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 + X_{1i} * X_{2i}$$

- multiplication of predictors creates interaction term
- in R:

```
motion_mod2 = lm(LnThresh ~ Mtype + Sex + LnAge + Mtype:Sex
                         Mtype:LnAge + Sex:LnAge, data = motion)
coef(summary(motion_mod2))
```

- interpretation of sig. interaction terms is a chapter for itself!

what is logistic regression?

- continuous outcome:
- $Y_i = b_0 + b_1 X_{1i} + \varepsilon_i$

linear regression model

- $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i$
- categorical outcome:
- $\boxed{P(Y)} = \frac{1}{1+e^{-(b_0 + b_1 X_{1i})}}$

Logistic regression model

- $\boxed{P(Y)} = \frac{1}{1+e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni})}}$

GOAL:
predict outcome (yes/no) based on predictors

WHY:
categorical outcome violates linearity assumption

probability:
classifier, threshold
convergence between 0 & 1

→

sigmoid function, ML training, accuracy

methods of logistic regression

- forced entry method
  - all predictors in one block and revealing their estimated parameter
- stepwise methods
  - forward / backward integration
  - forward: constant + predictor inclusion with AIC / BIC (must improve model)
  - backward: exclusion of predictors that tune information criterion (AIC / BIC)
  - hybrid: forward/backward – more dynamic on each step
- method selection
  - theoretical background / data exploration
  - stepwise is a good approach to fit data (not causality driven!)

assumptions and background

- linearity
- independence of errors
- multicollinearity
- MLE (Maximum-likelihood estimation)

$$\chi^2 = (-2LL(\text{baseline})) - (-2LL(\text{new}))$$
$$= 2LL(\text{new}) - 2LL(\text{baseline})$$
$$df = k_{\text{new}} - k_{\text{baseline}}$$

log-likelihood = $\sum_{i=1}^{N}[Y_i ln(P(Y_i)) + (1 - Y_i)\ln(1 -$

deviance = $-2LL$ (follows a χ^2 distribution)

- complete separation

```
Warning messages:
1: glm.fit: algorithm did not
converge
```

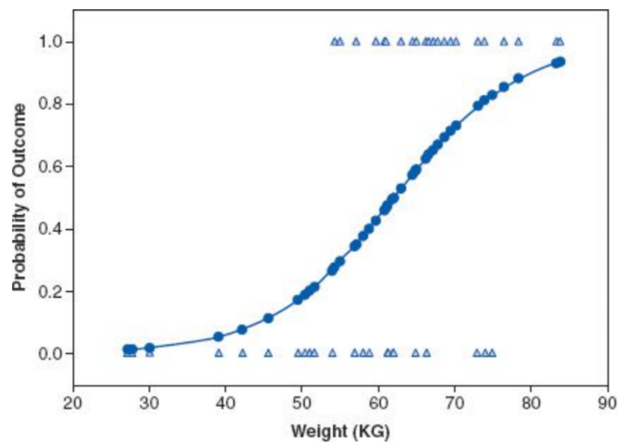

FIGURE 8.3 An example of the relationship between weight (*x*-axis) and a dichotomous outcome variable (*y*-axis, 1 = Burglar, 0 = Teenager) – note that the weights in the two groups overlap

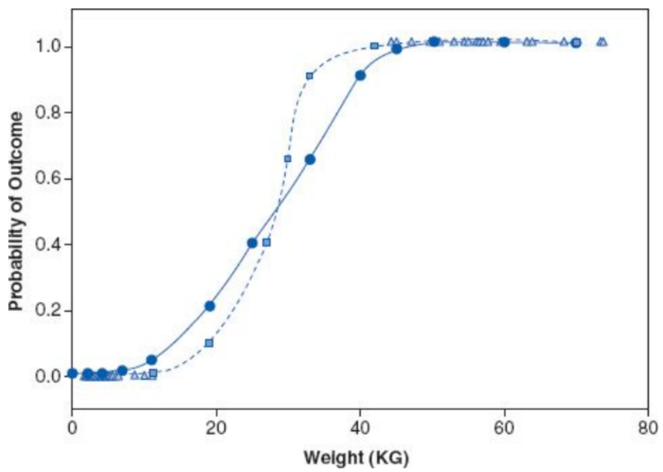

FIGURE 8.4 An example of complete separation – note that the weights (*x*-axis) of the two categories in the dichotomous outcome variable (*y*-axis, 1 = Burglar, 0 = Cat) do not overlap

assumptions and background

- Akaike information criterion AIC = $-2LL + 2k$
- Bayes information criterion BIC = $-2LL + 2k \ * \log(n)$

- contribution of the predictors: the *z*-statistic $z = \frac{b}{SE_b}$

- the odds ratio $odds = \frac{P(event)}{P(no\ event)}$

$P(event\ of\ Y) = \frac{1}{1 + e^{-(b_0 + b_1 Y_1)}}$ $P(no\ event\ of\ Y) = 1 - P(event\ of\ Y)$

exercise one

- use the functions vif() and log(), dataset = ‘penalty.dat‘
- evaluate the multicollinearity of the variables: Previous, Anxious, PSWQ
- is the assumption violated? If so what can you do?
  - omit variable (is there a relevant criterion?)
  - test more subjects
  - factor analysis, general load of predictors combined as a factor
  - accept the unreliable model
- Evaluate the linearity of the variables. Interaction: X * log(X)
  - PSWQ – logPSWQInt; Previous – logPrevInt; Anxious – logAnxInt
  - What do you know about logarithmic any special cases?

exercise one

PSWQ	Anxious	Previous	Scored	logPSWQInt	logAnxInt	logPrevInt
18	21	56	Scored Penalty	52.02669	63.93497	226.41087
17	32	35	Scored Penalty	48.16463	110.90355	125.42316
16	34	35	Scored Penalty	44.36142	119.89626	125.42316
14	40	15	Scored Penalty	36.94680	147.55518	41.58883
5	24	47	Scored Penalty	8.04719	76.27329	181.94645
1	15	67	Scored Penalty	0.00000	40.62075	282.70702

etc.

model output

TODO: What does family mean and how many families are there?

```
Call:
glm(formula = Cured ~ Intervention, family = binomial(), data = eelData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5940  -1.0579   0.8118   0.8118   1.3018

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.2877     0.2700  -1.065  0.28671
InterventionIntervention      1.2287     0.3998   3.074  0.00212 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 111  degrees of freedom
AIC: 148.16

Number of Fisher Scoring iterations: 4
```

```
family(object, ...)

binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

but is the model significant?

our model improves when adding intervention

model output

but is the model significant?

```
Call:
glm(formula = Cured ~ Intervention, family = binomial(), data = eelData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5940  -1.0579   0.8118   0.8118   1.3018

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.2877     0.2700  -1.065  0.28671
InterventionIntervention       1.2287     0.3998   3.074  0.00212 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 111  degrees of freedom
AIC: 148.16

Number of Fisher Scoring iterations: 4
```

follows a chi-square statistic

χ^2 = null deviance – deviance

$\chi^2 = 154.08 - 144.16 = 9.926$

Degrees of freedom

$df = 112 - 111 = 1$

p-value:

chi.sq.prob = 1 – pchisq(χ^2 , *df*)

$\chi^2 = 9.926$, $p = 0.002$

exercise two

- pick a data set from the Drive folder
- run a logistic regression for variables, interactions, …
- present and explain

estimates for R^2

```
logisticPseudoR2s <- function(LogModel) {
      dev <- LogModel$deviance
      nullDev <- LogModel$null.deviance
      modelN <- length(LogModel$fitted.values)
      R.l <-  1 -  dev / nullDev
      R.cs <- 1- exp ( -(nullDev - dev) / modelN)
      R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
      cat("Pseudo R^2 for logistic regression\n")
      cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
      cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
      cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
    }
```

```
Pseudo R^2 for logistic regression
Hosmer and Lemeshow R^2    0.064
Cox and Snell R^2          0.084
Nagelkerke R^2             0.113
```

equivalents to R^2

running multinomial logistic regression

- install.packages(“mlogit”)
- select a baseline (relevel() function)
- use function mlogit()
  - read documentary - ?mlogit
  - reference is included in mlogit function
- log-likelihood ratio:
  - how much unexplained data is in the model

Table 8.3 How to report multinomial logistic regression

		95% CI for odds ratio		
	B (SE)	Lower	Odds Ratio	Upper
Phone number vs. no response				
Intercept	−1.78 (0.67)**			
Good Mate	0.13 (0.05)*	1.03	1.14	1.27
Funny	0.14 (0.11)	0.93	1.15	1.43
Female	−1.65 (0.80)*	0.04	0.19	0.92
Sexual Content	0.28 (0.09)**	1.11	1.32	1.57
Female × Funny	0.49 (0.14)***	1.24	1.64	2.15
Female × Sex	−0.35 (0.11)*	0.57	0.71	0.87
Going home vs. no response				
Intercept	−4.29 (0.94)***			
Good Mate	0.13 (0.08)	0.97	1.14	1.34
Funny	0.32 (0.13)*	1.08	1.38	1.76
Female	−5.63 (1.33)***	0.00	0.00	0.05
Sexual Content	0.42 (0.12)**	1.20	1.52	1.93
Female × Funny	1.17 (0.20)***	2.19	3.23	4.77
Female × Sex	−0.48 (0.16)**	0.45	0.62	0.86

```
Log-Likelihood: -868.74
McFadden R^2:   0.13816
Likelihood ratio test : chisq = 278.52 (p.value=< 2.22e-16)
```

McFadden R^2 and LL relationship to χ^2