ORIGIN := 0

Logistic Regression with Proportions

Logistic Regression can be extended to binary situations (i.e., "presence vs "absence, "yes vs "no", etc.) where multiple observations (n_i) are made for each case i of the independent variables (X_1, X_2, X_3, etc.). As a result, response variable (Y_i) consists of a proportions of "yes" vs "no" (or "yes" out of total observations) for each case i. For instance, as in Zuur et al.'s example below, proportions of deer positive for tuberculosis (Y_i) were censused in different farms (i) for which potential contributing factors (the X_i's) were also assesed. Proportions of tuberculosis is the response variable, with n_i identical values of X representing each farm i. Because there are now multiple observations n_i for each case i, Y_i is distributed according to the Binomial distribution ~ $B(n_i,\pi_i)$ and contrasts with the Bernoulli distribution of standard Logistic Regression ~ $B(1,\pi_i)$ where $n_i = 1$. Examples here are drawn from Zuur et al. 2009 (Za) in *Mixed Effects Models and Extensions in Ecology with R*, and Crawley (Cr) 2007, *The R Book*.

Assumptions:

Regression depends on specifying in Response & Independent variables in advance:

Y = vector of binary response variable (0 or 1), each row indicated by index i.
X= matrix independent variables (columns) with observations of X_i (rows) matched to Y_i (rows of Y).
β = vector of linear coefficients and $X\beta$ is the linear predictor (systematic component) of the model.
Cases Y_i are independent.

Model:

$Y_i \sim B(n_i,\pi_i)$ **< Y's are Binomial distributed with probability π_i for each case i.**

$E(Y_i) = \pi_i$, and $var(Y_i) = \pi_i(1-\pi_i)$ **< mean and variance defined.**

$logit(\pi_i) = X\beta$ where $logit(\pi_i) = \pi_i/(1-\pi_i)$

alternatively: $\pi_i = e^{(X\beta)}/(1+e^{(X\beta)})$ **< for logit link to π_i for Logistic Regression**

Estimation of Regression Coefficients:

Estimation is based on determining the maximum likelihood function given the data. Since a closed-form solution doesn't exit, this requires interative computation, here using glm() in the {nlme} package in R.

```
#GLM 030 LOGISTIC REGRESSION WITH PROPORTIONS
library(nlme)
setwd("c:/DATA/Models")

S=read.table("sexratio.txt",header=T)
S
Y=cbind(S$males,S$females)
logdensity=log(S$density)
FM1=glm(Y~logdensity,family=binomial)
summary(FM1)
```

Crawley's Sex Ratio Example Ch 16

```
> S
  density females males
1       1       1     0
2       4       3     1
3      10       7     3
4      22      18     4
5      55      22    33
6     121      41    80
7     210      52   158
8     444      79   365
```

Proportion of males/famales in this insect appears to be related to population density. A preliminary graph (not shown here) indicates that log(density) would be an appropriate independent variable.

```
Results=cbind(Y=Y,logdensity=logdensity,Fitted=fitted(FM1,type="response"))
Results
```

```
> summary(FM1)

Call:
glm(formula = Y ~ logdensity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9697  -0.3411   0.1499   0.4019   1.0372

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.65927    0.48758  -5.454 4.92e-08 ***
logdensity   0.69410    0.09056   7.665 1.80e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71.1593  on 7  degrees of freedom
Residual deviance:  5.6739  on 6  degrees of freedom
AIC: 38.201

Number of Fisher Scoring iterations: 4
```

Fitted Values:

$Y_4 := \frac{4}{4 + 18}$ $Y_4 = 0.182$

^ response variable as proportion

$X_4 := 3.091042$ **< observed values of X & Y for case 4**

$\beta_0 := -2.65927$ $\beta_1 := 0.69410$ **< regression coefficients from above**

$\pi_4 := \frac{e^{(\beta_0+\beta_1 \cdot X_4)}}{1 + e^{(\beta_0+\beta_1 \cdot X_4)}}$ $\pi_4 = 0.3743$ **< fitted value for the first case π_1**

```
> Results
  Y  N logdensity    Fitted
1   0  1   0.000000 0.0654202
2   1  3   1.386294 0.1548524
3   3  7   2.302585 0.2571105
4   4 18   3.091042 0.3743095
5  33 22   4.007333 0.5305182
6  80 41   4.795791 0.6613896
7 158 52   5.347108 0.7411899
8 365 79   6.095825 0.8280467
```

Odds & Odds Ratio:

$Odds_4 := \frac{\pi_4}{1 - \pi_4}$ $Odds_4 = 0.598$ **= Odds = ratio of probability π_4 (probability Y=1)and its complement (1=π_4) (probability Y=0).**

$X_n := X_4 + 1$ **< unit increase in independent variable X**

$p_n := \frac{e^{(\beta_0+\beta_1 \cdot X_n)}}{1 + e^{(\beta_0+\beta_1 \cdot X_n)}}$ $p_n = 0.545$ **< probability at X+1**

$Odds_n := \frac{p_n}{1 - p_n}$ $Odds_n = 1.198$ **= Odds at X+1**

$\frac{Odds_n}{Odds_4} = 2.002$ $e^{\beta_1} = 2.002$ **= Odds Ratio for case 4**

^ A unit increase in the independent variable X results in a two-fold increase in the odds of Y (proportions of males in this example).

```
Pmales=S$males/(S$males+S$females)
Residuals=cbind(logdensity=logdensity,Pmales=Pmales,
        Pearson=resid(FM1,type="pearson"),
        Deviance=resid(FM1,type="deviance"))
Residuals
```

```
> Residuals
  logdensity    Pmales    Pearson     Deviance
1   0.000000 0.0000000 -0.2645744 -0.367853955
2   1.386294 0.2500000  0.5260203  0.491272300
3   2.302585 0.3000000  0.3103336  0.305157947
4   3.091042 0.1818182 -1.8656374 -1.969677045
5   4.007333 0.6000000  1.0325072  1.037167990
6   4.795791 0.6611570 -0.0054063 -0.005405998
7   5.347108 0.7523810  0.3702766  0.372054745
8   6.095825 0.8220721 -0.3336343 -0.332127534
```

Pearson Residuals:

$n_4 := 4 + 18$

$$r_P := \frac{\sqrt{n_4} \cdot (Y_4 - \pi_4)}{\sqrt{\pi_4 \cdot (1 - \pi_4)}} \qquad r_P = -1.8656$$

< pearson residual for case 4

Deviance Residuals:

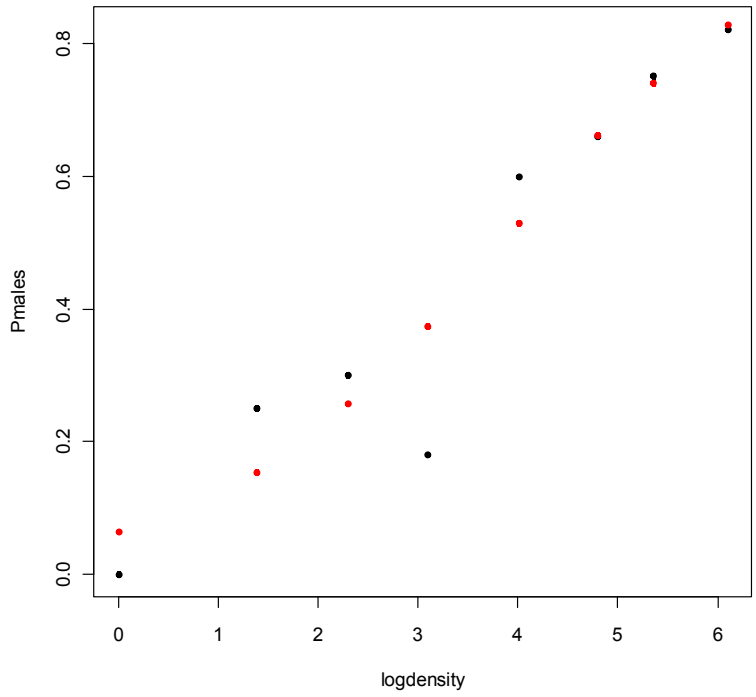
I haven't found a verified formula for calculating deviance residuals as reported by R for proportion data yet. Note, however, that the sum of squares of deviance residuals produces the overall residual deviance reported in summary(FM1) above.

```
sum((resid(FM1,type='deviance')^2))
```

```
[1] 5.673894
```

Plotting Regression Fit:

```
#PLOTTING REGRESSION FIT:
Pmales=S$males/(S$males+S$females)
plot(logdensity,Pmales,pch=20,col="black")
points(logdensity,predict(FM1,type="response"),pch=20,col='red')
```



Multiple Logistic Regression with Proportions:

Finding an optimal model with proportions follows the same format seen in standard Linear models. However, with proportion data, one must check for overdispersion and employ a "quasi-binomial" corrective measure. Za caution that although R software refers to quasi-binomial (and quasi-poisson) using what appears to be a call to an error distribution - i.e., family="quasibinomial" within glm() - there is no such distribution. The correction actually involves use of binomial GLM with correction factor ϕ for overdispersion using a quasi-binomial model.

```
#ZUUR ET AL. 10.3 DEER TB DATA:
T=read.table("TBdeer.txt",header=T)
T
T$fFenced=factor(T$Fenced)
pos=T$DeerPosCervi
neg=T$DeerSampledCervi-T$DeerPosCervi
```

Zuur et al.'s deer TB data in Section 10.3. The dataset is large with many columns of data only some of which are utilized in this example...

The response variable consists of two-columns bound together by c() with counts of alternative binary responses represented in each column: e.g., c(pos,neg), c(yes, no), etc.

Detecting Overdispersion:

```
FM2=glm(cbind(pos,neg)~
    OpenLand+ScrubLand+QuercusPlants+QuercusTrees+
    ReedDeerIndex+EstateSize+fFenced,
    family=binomial,data=T)
summary(FM2)
```

```
> summary(FM2)
Call:
glm(formula = cbind(pos, neg) ~ OpenLand + ScrubLand + QuercusPlants +
    QuercusTrees + ReedDeerIndex + EstateSize + fFenced, family = binomial,
    data = T)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.0590  -0.8956   0.1093   1.8354   3.7123

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.843e+00  7.772e-01   4.945 7.61e-07 ***
OpenLand      -3.950e+00  6.383e-01  -6.187 6.12e-10 ***
ScrubLand     -7.696e-01  6.140e-01  -1.253 0.210042
QuercusPlants -3.633e-04  2.308e-02  -0.016 0.987439
QuercusTrees   2.290e-03  5.326e-02   0.043 0.965707
ReedDeerIndex  6.689e-02  2.097e-02   3.191 0.001419 **
EstateSize    -8.218e-05  2.478e-05  -3.316 0.000913 ***
fFenced1      -2.273e+00  5.954e-01  -3.819 0.000134 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.85  on 22  degrees of freedom
Residual deviance: 152.79  on 15  degrees of freedom
  (9 observations deleted due to missingness)
AIC: 227.87

Number of Fisher Scoring iterations: 4
```

Checking for overdispersion involves comparing residual deviance with residual degrees of freedom found in summary(). The model is not overdispersed if the ratio is 1:1. Here the ratio is approx 10:1, indicating the need to correct for overdispersion. Note that the first example, above, showed no evidence of overdispersion, so is correct as it stands.

Correcting for Overdispersion: using a quasi-binomial model:

```
FM3=update(FM2,family=quasibinomial)
summary(FM3)
```

Output of quasi-Binomial fit:

GLM t-Test of single β:

Hypotheses:

H_0: β = 0

H_1: β ≠ 0

Test Statistic:

t = Estimate/std.error

Example calculations:

$\beta_1 := -3.950$ $\quad s_{\beta_1} := 2.026$

$t_1 := \frac{\beta_1}{s_{\beta_1}}$ $\quad t_1 = -1.9497$

$\beta_5 := 6.689 \cdot 10^{-2}$ $\quad s_{\beta_5} := 6.655 \cdot 10^{-2}$

$t_5 := \frac{\beta_5}{s_{\beta_5}}$ $\quad t_5 = 1.005$

```
> summary(FM3)
Call:
glm(formula = cbind(pos, neg) ~ OpenLand + ScrubLand + QuercusPlants
 + QuercusTrees + ReedDeerIndex + EstateSize + fFenced, family =
 quasibinomial, data = T)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.0590  -0.8956   0.1093   1.8354   3.7123

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.843e+00  2.467e+00   1.558   0.1401
OpenLand      -3.950e+00  2.026e+00  -1.949   0.0702 .
ScrubLand     -7.696e-01  1.949e+00  -0.395   0.6985
QuercusPlants -3.633e-04  7.325e-02  -0.005   0.9961
QuercusTrees   2.290e-03  1.691e-01   0.014   0.9894
ReedDeerIndex  6.689e-02  6.655e-02   1.005   0.3308
EstateSize    -8.218e-05  7.866e-05  -1.045   0.3127
fFenced1      -2.273e+00  1.890e+00  -1.203   0.2476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 10.07626)

    Null deviance: 234.85  on 22  degrees of freedom
Residual deviance: 152.79  on 15  degrees of freedom
  (9 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 4
```

Sampling Distribution:

If Assumptions hold and H_0 is true,

then t ~ t(0,df) with df = Residual Deviance df

Probability:

df := 15 **< from residual deviance df reported in summary()**

$P_1 := \min\left[\left(2 \cdot \text{pt}\left(t_1, \text{df}\right)\right), 2 \cdot \left(1 - \text{pt}\left(t_1, \text{df}\right)\right)\right]$ $\quad P_1 = 0.0702$

$P_5 := \min\left[\left(2 \cdot \text{pt}\left(t_5, \text{df}\right)\right), 2 \cdot \left(1 - \text{pt}\left(t_5, \text{df}\right)\right)\right]$ $\quad P_5 = 0.3308$

Decision Rule:

IF P < α THEN REJECT H_0, OTHERWISE ACCEPT H_0

Confidence Interval for β:

α := 0.05

$C := \left|\text{qt}\left(1 - \frac{\alpha}{2}, \text{df}\right)\right|$ $\quad C = 2.131$

	estimate	confidence interval
$CI_1 := \left(\beta_1 - C \cdot s_{\beta_1} \quad \beta_1 + C \cdot s_{\beta_1}\right)$	$\beta_1 = -3.95$	$CI_1 = (-8.268 \quad 0.368)$
$CI_5 := \left(\beta_5 - C \cdot s_{\beta_5} \quad \beta_5 + C \cdot s_{\beta_5}\right)$	$\beta_5 = 0.06689$	$CI_5 = (-0.07496 \quad 0.20874)$

Note: This version of the t-Test is a marginal test and depends on obtaining appropriate values from the summary() wrapper of object made by glm().

GLM F-Test:

This test is the GLM version of Full Model (FM) versus Reduced Model (RM) is conducted by consulting anova(RM,FM) for a serial report or drop1(FM) for a set of marginal tests droping one element of FM at a time but keeping all other independent variables in the RM. According to R documentation in ?anova.glm, the F-test is appropriate for the Gaussian link, and for tests involving quasi-Binomial & quasi-Poisson where overdispersion is estimated. Otherwise one should use likelihood ratio tests and chi-square for tests of this sort.

drop1(FM3,test="F")

quasi-Binomial tests

marginal F-tests for a single factor :

```
> drop1(FM3,test="F")
Single term deletions

Model:
cbind(pos, neg) ~ OpenLand + ScrubLand + QuercusPlants + QuercusTrees +
    ReedDeerIndex + EstateSize + fFenced
              Df Deviance F value   Pr(F)
<none>             152.79
OpenLand       1   198.29  4.4669 0.05172 .
ScrubLand      1   154.37  0.1547 0.69959
QuercusPlants  1   152.79  0.0000 0.99613
QuercusTrees   1   152.79  0.0002 0.98943
ReedDeerIndex  1   163.69  1.0697 0.31740
EstateSize     1   163.52  1.0537 0.32093
fFenced        1   171.62  1.8487 0.19403
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

anova(FM3,test="F")

serial F-tests for factors entering into model formula from left to right:

```
> anova(FM3,test="F")
Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: cbind(pos, neg)

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                             22     234.85
OpenLand       1   51.681        21     183.17 5.1290 0.03877 *
ScrubLand      1    2.693        20     180.47 0.2673 0.61271
QuercusPlants  1    2.568        19     177.91 0.2548 0.62102
QuercusTrees   1    0.000        18     177.91 0.0000 0.99828
ReedDeerIndex  1    5.948        17     171.96 0.5903 0.45422
EstateSize     1    0.337        16     171.62 0.0334 0.85744
fFenced        1   18.831        15     152.79 1.8688 0.19175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypotheses:

H_0: a specified subset of β's = 0

H_1: same subset of β's <> 0

Test Statistic:

$$F = [SSE_R - SSE_F / (df_R - df_F)] / MSE_F$$ **< Same formula as with standard Linear models.**

Sampling Distribution:

If Assumptions hold and H_0 is true,

then $F \sim F(df_R - df_F, df_F)$ **< where df_F & df_R are degrees of freedom of FM & RM respectively**

Probability:

$$P = 1 - pF(df_R - df_F, df_F)$$

Decision Rule:

IF $P < \alpha$ THEN REJECT H_0, OTHERWISE ACCEPT H_0

Optimal Model found by Za:

```
RM1=glm(cbind(pos,neg)~OpenLand,family=quasibinomial,data=T)
summary(RM1)
drop1(RM1,test="F")
anova(FM3,RM1)
```

```
> drop1(RM1,test="F")
Single term deletions

Model:
cbind(pos, neg) ~ OpenLand
         Df Deviance F value   Pr(F)
<none>        183.80
OpenLand  1   235.58  6.1976 0.02084 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
#PLOTTING FIT:
M=data.frame(OpenLand=seq(from = min(T$OpenLand),to = max(T$OpenLand),by=0.01))

Pred=predict(RM1,newdata=M,type="response",se=TRUE)
plot(M$OpenLand,Pred$fit,
   type="l",ylim=c(0,1),col='red',
   xlab="Percentage open land",
   ylab="E. cervi Probability")
PROPORTION=T$DeerPosCervi/T$DeerSampledCervi
points(T$OpenLand,PROPORTION,pch=20)
lines(M$OpenLand,Pred$fit+1.96*Pred$se.fit,lty=2,col='blue')
lines(M$OpenLand,Pred$fit-1.96*Pred$se.fit,lty=2,col='blue')
```