GLMs: binomial family

Mandy Vogel

University Leipzig

August 17, 2015

Overview

Recap

Binomial/Logistic Regression

The famous O-Ring example

Ancova with a Binary Response Variable

Table of Contents I

Recap

Binomial/Logistic Regression

The famous O-Ring example

Ancova with a Binary Response Variable

R syntax - `glm()` vs `lm()`

- the `glm()` function needs
  - a model formula (like `lm`)
  - the specification of error distribution (`family=`)

Input/Output

```
> m1 <- lm(bweight ~ hyp, data=births)
> m2 <- glm(bweight ~ hyp, family=gaussian, data=births)
```

`glm()` for logistic regression

- every error family has a canonical link
- we have seen binomial error family with its canonical logit link
- common choices for link functions used with binomial errors
  - logit: $\eta = \log(p/(1-p))$
  - probit: $\eta = \Phi^{-1}(p)$
  - log-log: $\log(-\log(1-p))$

Odds

- logistic regression is more understandable if you look at the cœfficients in terms of odds where
- $\Omega(A) = \dfrac{P(A)}{1 - P(A)}$
- so what are the corresponding odds for a probability of
  - $p = 1$
  - $p = 0.99$
  - $p = 0.5$
  - $p = 0.1$
  - $p = 0.01$
  - $p = 0$

Odds

$p = 1$	$\omega = \infty$
$p = 0.99$	$\omega = 99$
$p = 0.5$	$\omega = 1$
$p = 0.1$	$\omega = 0.\bar{1}$
$p = 0.01$	$\omega = 0.\overline{01}$
$p = 0$	$\omega = 0$

Remember the Data

Input/Output

```
  > str(births)
  'data.frame': 500 obs. of  8 variables:
 $ id     : num  100 101 102 103 104 105 106 107 108 109 ...
 $ preterm: Factor w/ 2 levels "preterm","normal": 2 2 2 2 2 2
 $ gestwks: num  39.8 39 38.1 39.5 39.5 ...
 $ hyp    : Factor w/ 2 levels "normal","hyper": 1 1 1 1 2 1 2
 $ matage : num  33 32 33 38 40 29 32 40 41 39 ...
 $ bweight: num  3576 3784 2796 3226 3138 ...
 $ lowbw  : Factor w/ 2 levels "normal","low": 1 1 1 1 1 1 1 1
 $ sex    : Factor w/ 2 levels "M","F": 2 2 2 2 2 2 1 1 2 2 ..
```

Data from: Michæl Hills and Bianca De Stavola (2002). A Short Introduction to Stata 8 for Biostatistics, Timberlake Consultants Ltd URL: http://www.timberlake.co.uk

Exercises

Remember: We used hypertension of the mom to explain variation in the birth weight (in terms of low birth weight or not of course) of the kid. Without looking in the material of the last session, try to redo the model. Here are some hints:

- of course you need the `glm()` function
- you need to specify the formula which has to have the general form $y \sim x$
- additional you need to specify the data and the error family (in the case `binomial`)
- use the `summary()` function on the model
- use `Effect()` or `allEffects()` commands on the model
- how to interpret the results? Is the effect of hypertension statistically significant?

Exercise - Solution

Input/Output

```
> m <- glm(lowbw ~ hyp, family=binomial, data=births)
> summary(m)
Call:
glm(formula = lowbw ~ hyp, family = binomial, data = births)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8067  -0.4430  -0.4430  -0.4430   2.1773

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2721     0.1661 -13.682  < 2e-16 ***
hyphyper      1.3166     0.3111   4.232 2.32e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 366.92  on 499  degrees of freedom
Residual deviance: 350.84  on 498  degrees of freedom
AIC: 354.84
```

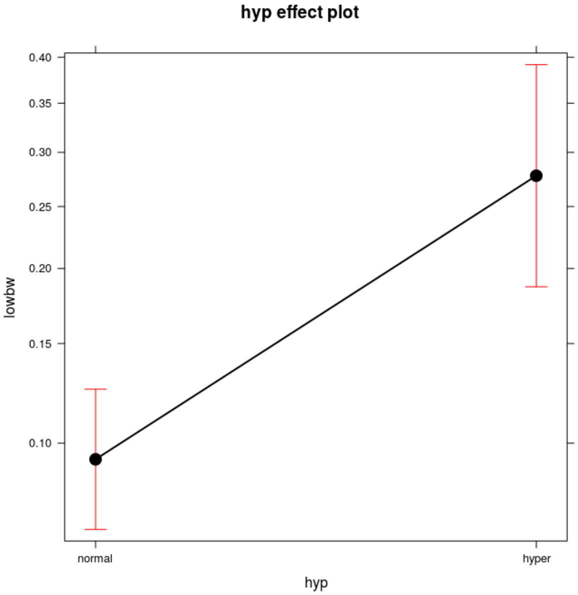
Exercise - Solution

Input/Output

```
> require(effects)
Lade nötiges Paket: effects
> allEffects(m.hyp)
 model: lowbw ~ hyp

 hyp effect
hyp
    normal      hyper
0.09345794 0.27777778
> plot(allEffects(m.hyp))
```

Exercise - Solution



Exercise - Solution

Input/Output

```
> res <- allEffects(m.hyp, se = T)
> summary(res)
 model: lowbw ~ hyp

 hyp effect
hyp
    normal      hyper
0.09345794 0.27777778

 Lower 95 Percent Confidence Limits
hyp
    normal      hyper
0.06929267 0.18675845

 Upper 95 Percent Confidence Limits
hyp
   normal     hyper
0.1249195 0.3917861
```

Exercises

What is the relationship between the cœfficients of the model (from the model summary) and the effects?

Exercises - Solutions

What is the relationship between the cœfficients of the model (from the model summary) and the effects?

- we have to use the inverse link function on the cœfficients to transform the cœfficients on the logit scale to more interpretable probabilities

Input/Output

```
> invlogit(coef(m.hyp)[1])
(Intercept)
 0.09345794
> invlogit(coef(m.hyp)[1] + coef(m.hyp)[2])
(Intercept)
  0.2777778
```

Table of Contents I

Recap

Binomial/Logistic Regression

The famous O-Ring example

Ancova with a Binary Response Variable

Simple Logistic Regression

- now we model the probability of low birth weight dependent on gestational age (numeric variable)
- so the model in R is

Input

```
> m.wks <- glm(lowbw ~ gestwks, family=binomial, data=births)
```

- and as math formula

$$\log\left(\frac{\mathsf{Pr(lowbw)}}{1-\mathsf{Pr(lowbw)}}\right)=\beta_0+\beta_1\cdot\mathsf{gestwks}+\epsilon$$

Simple Logistic Regression

- where the output look similar to the output above

Input/Output

```
> summary(m.wks)

Call:
glm(formula = lowbw ~ gestwks, family = binomial, data = births)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0873  -0.3623  -0.2223  -0.1369   2.9753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  31.8477     4.0574   7.849 4.18e-15 ***
gestwks      -0.8965     0.1084  -8.272  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 360.38  on 489  degrees of freedom
Residual deviance: 205.75  on 488  degrees of freedom
  (10 observations deleted due to missingness)
```

Understanding the Cœfficients

- this relationship is described by

$$\text{Pr(lowbw)} = \text{logit}^{-1}(31.8477 + -0.8965 \cdot \text{gestwks})$$

- the intercept

Input/Output

```
> invlogit(coef(m.wks)[1])
(Intercept)
  1
```

is interpretable as the probability for a low birth weight at a hypothetical gestational age of 0 (which makes no sense because it lies outside the range of gestational ages in our data and is nonsense anyway)

- the parameter for `gestwks` describes how fast the probability decreases with increasing gestational age

Understanding the Cœfficients

$$\text{Pr(lowbw)} = \text{logit}^{-1}(31.8477 + -0.8965 \cdot \text{gestwks})$$

- the cœfficient for `gestwks` is best interpretable if we use it as argument to the exponential function

Input/Output

```
> exp(coef(m.wks)[2])
  gestwks
0.4080114
```

this way it is interpretable as odds ratio for low birth weight for a difference of 1 week of gestational age (because we are measuring gestational in weeks as unit)

Exercise

1. here is a example for the `Effects()` command for regression

Input/Output

```
> Effect("gestwks",m.wks)

 gestwks effect
gestwks
        25         30         35         40
0.99992022 0.99299324 0.61574996 0.01779725
> Effect("gestwks",m.wks,xlevels = list(gestwks = c(20,30,40)))

 gestwks effect
gestwks
        20         30         40
0.99999910 0.99299324 0.01779725
```

2. use the command to gain the estimated probability of low birth weight for a gestational age of 27 and 36 weeks

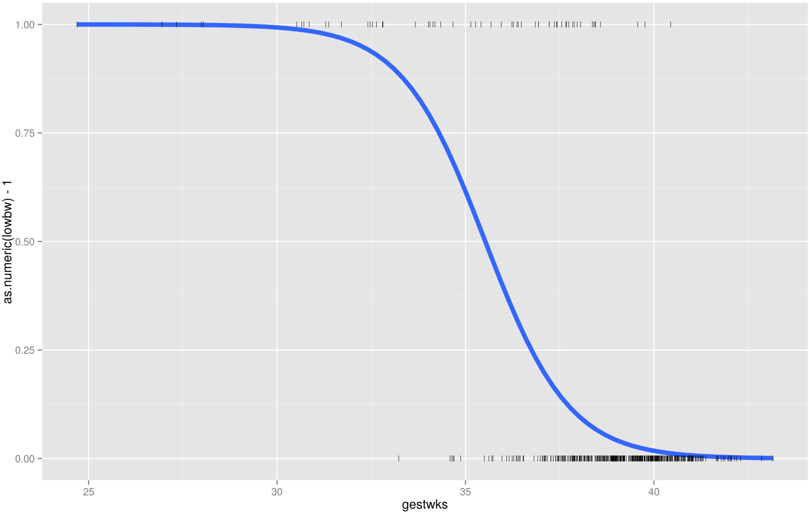
ggplot() and glm()

- ggplot2 knows also glms
- unfortunately the y-variable needs to be coded in 0s and 1s, but we can do this on the fly with `as.numeric()`

Input

```
> require(ggplot2)
> ggplot(births,aes(x = gestwks, y = as.numeric(lowbw)-1)) +
+     geom_smooth(method = "glm", family = "binomial",se = F,size = 2) +
+     geom_point(shape="|")  ## adds actual values
```

ggplot() and glm()



Exercise

Take the code producing the graph

1. try to change the axis titles (`xlab()` and `ylab()`)
2. add a title (`ggtitle()`)
3. change the colour of the function to black, set se = T
4. change the colour of the points to red for the low birth weight and green for the one with normal birth weight
5. change the position of the legend; place it somewhere near the upper right corner inside the plotting area (`legend.position`)

Table of Contents I

Recap

Binomial/Logistic Regression

The famous O-Ring example

Ancova with a Binary Response Variable

The Challenger Disaster Example

In January 1986, the space shuttle Challenger exploded shortly after launch. An investigation was launched into the cause of the crash and attention focused on the rubber O-ring seals in the rocket boosters. At lower temperatures, rubber becomes more brittle and is a less effective sealant. At the time of the launch, the temperature was 31°F. Could the failure of the O-rings have been predicted? In the 23 previous shuttle missions for which data exists, some evidence of damage due to blow by and erosion was recorded on some O-rings. Each shuttle had two boosters, each with three O-rings. For each mission, we know the number of O-rings out of six showing some damage and the launch temperature.(faraway) `http://www.history.com/topics/challenger-disaster/videos/engineering-disasters---challenger`

The Challenger Disaster Example

- the data are given in the data frame `orings` in the `faraway` package
- after loading we have a look at the first six lines

Input/Output

```
> library(faraway)
> data(orings)
> head(orings)
  temp damage
1   53      5
2   57      1
3   58      1
4   63      1
5   66      0
6   67      0
```

- we see that every shuttle mission has its own row (but not every O-ring)

The Challenger Disaster Example

- that is not a problem: one way of defining a binary response variable in a glm is to form a two-column matrix with the first column representing the number of “successes” y and the second column the number of “failures” n–y.

Input/Output

```
> m.oring <- glm(cbind(damage,6-damage) ~ temp,
+                      family=binomial, orings)
```

The Challenger Disaster Example

- the output looks familiar:

Input/Output

```
> summary(m.oring)
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp,
     family = binomial, data = orings)
Deviance Residuals:
    Min       1Q    Median       3Q      Max
-0.9529  -0.7345  -0.4393  -0.2079   1.9565
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.675
```

- remember, the response is a probability. Therefore our model describes the probability of a damaged O-ring depending on the temperature

Understanding the Cœfficients

- this relationship is described by

$$\text{Pr(damage)} = \text{logit}^{-1}(11.66299 + -0.21623 \cdot \text{temp})$$

Input/Output

```
> invlogit(coef(m.oring)[1])
(Intercept)
  0.9999914
```

- the intercept is interpretable as the probability for a damaged O-ring at a temperature of 0°F

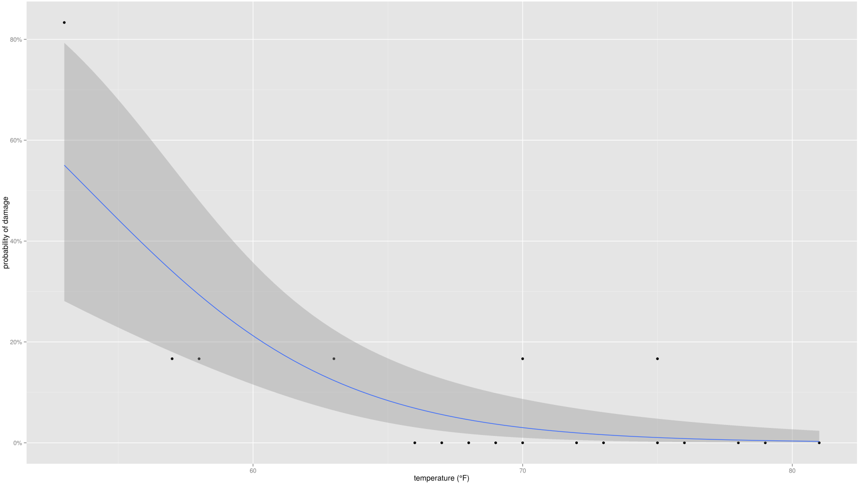
Understanding the Cœfficients

- the parameter for temperature describes how fast the probability decreases with increasing temperature and it is again best interpretable as odds ratio

Input/Output

```
> exp(coef(m.oring)[2])
     temp
0.8055471
```

Understanding the Cœfficients



Understanding the Cœfficients

and the same plot made with ggplot (incl. adding a table)

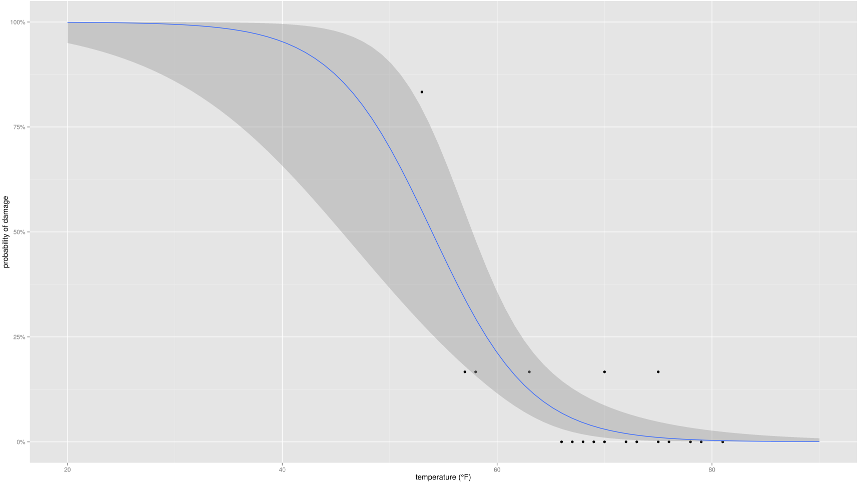


Table of Contents I

Recap

Binomial/Logistic Regression

The famous O-Ring example

Ancova with a Binary Response Variable

Parasite Infection Example

- the binary response variable is parasite infection (infected or not)
- the explanatory variables are weight and age (continuous)
- and sex (categorical)
- we want to investigate if there is a different effect of age for each of the sexes on the outcome variable

Input/Output

```
> load("infection.rdata")
> summary(infection)
         infected        age             sex
 infected    :338   Min.   :  2.00   female:243
 not infected:162   1st Qu.: 46.00   male  :257
                    Median : 84.50
                    Mean   : 93.69
                    3rd Qu.:139.25
                    Max.   :200.00
```

Parasite Infection Example

Input/Output

```
> m.inf <- glm(infected~age*sex,family=binomial,
+                               data=infection)
> summary(m.inf)
Call:
glm(formula = infected ~ age * sex, family = binomial,
                                     data = infection)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0411  -0.7307  -0.4363   0.6632   2.3215
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.000513   0.413639  -7.254 4.05e-13 ***
age          0.015657   0.003176   4.929 8.25e-07 ***
sex          0.116664   0.553956   0.211   0.8332
age:sex      0.011050   0.004612   2.396   0.0166 *

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 629.85  on 499  degrees of freedom
Residual deviance: 477.61  on 496  degrees of freedom
AIC: 485.61
```

Parasite Infection Example

- so for male at a age of 0 there is a probability of

Input/Output

```
> invlogit(coef(m.inf)[1])
(Intercept)
 0.04740269
```

- for females the probability at age 0 is

Input/Output

```
> invlogit(coef(m.inf)[1]+coef(m.inf)[3])
(Intercept)
 0.05295775
```

Compare Slopes

- so what about the slope?
- for males the underlying model is the following

$$\text{Pr(infection)} = \text{logit}^{-1}(-3.000513 + 0.015657 \cdot \text{age})$$

- for females the slope is almost twice as high

$$\text{Pr(infection)} = \text{logit}^{-1}(-2.883849 + 0.02670685 \cdot \text{age})$$

Compare Slopes

- looking at the odds ratios (which seem to be rather small)
- for males and females:

Input/Output

```
> exp(coef(m.inf)[2]) ## males
    age
1.01578
> exp(coef(m.inf)[2] + coef(m.inf)[4]) ## females
     age
1.027067
```

- these are the odds ratios for +1 time unit

Compare Slopes

- if time unit is days you get the odds ratio for +1 month by

Input/Output

```
> exp(30 * coef(m.inf)[2])
     age
1.599512
> exp(30 * (coef(m.inf)[2] + coef(m.inf)[4]))
     age
2.228225
```

- so keep in mind the scale you are measuring on

Compare Slopes

- we can also compare them by looking at the age where the probability to be infected is 50%
- this is the case when

$$-3.000513 + 0.015657 \cdot \text{age} = 0$$

respectively

$$-2.883849 + 0.02670685 \cdot \text{age} = 0$$

you can do it by hand or use R

Compare Slopes

- solve() solves systems of linear equations in the form A*x=b, where A is the matrix of cœfficients and b are the (negative) intercepts, here we have the special case with just one equation

Input/Output

```
> ## male
> solve(0.015657,3.000513)
[1] 191.6404
> ## female
> solve(0.02670685,2.883849)
[1] 107.9816
```

Compare Effects

- you can also use the allEffects() function (part of the effects package), which give you the probabilities for being infected on several ages for both sexes

Input/Output

```
> allEffects(m.inf)
 model: infected ~ age * sex

 age*sex effect
     sex
age            0          1
  2   0.04883687 0.05570148
  24  0.06756215 0.09596497
  46  0.09276694 0.16038932
  68  0.12610300 0.25582483
  90  0.16918450 0.38219715
  112 0.22322468 0.52680374
  134 0.28853152 0.66704908
  156 0.36399154 0.78286130
  178 0.44679328 0.86645480
  200 0.53265591 0.92110968
```

Compare Effects

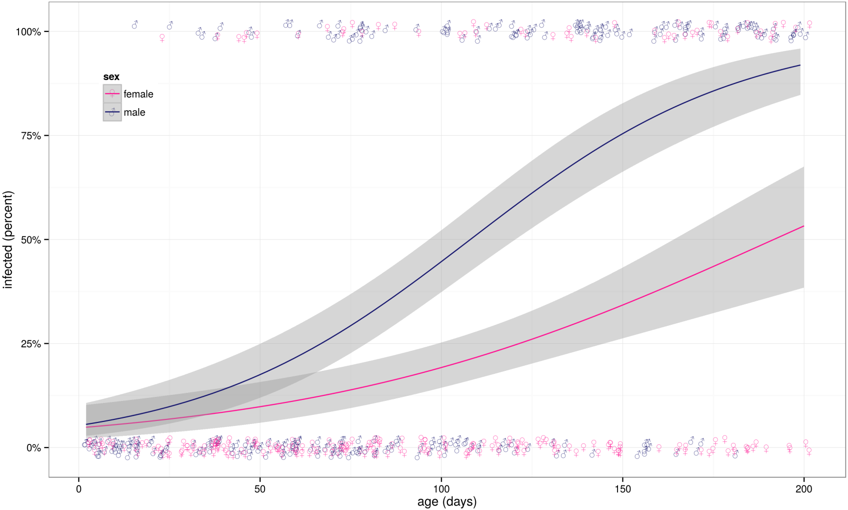
- choose values of age

Input/Output

```
> allEffects(m.inf,
+            xlevels = list(age = seq(0,200,by = 50)))
 model: infected ~ age * sex

 age*sex effect
     sex
age       female      male
  0   0.04740269 0.05295775
  50  0.09817379 0.17530204
  100 0.19234385 0.44690980
  150 0.34253427 0.75439251
  200 0.53265591 0.92110968
```

Parasite Infection graph



Exercise

Try to reproduce the plot! Hints:

1. set up a ggplot object, think about the æsthetics (`aes()`). Which quality of the graph you wanna set to which variable?
2. begin with the lines (`geom_smooth()`)
3. add the points (`geom_jitter()`; do not think about the symbols in the first place; try to adjust the width and height appropriately)
4. change the colour of the lines and points (`scale_colour_manual()`); I used midnightblue for male and deeppink for female
5. change the symbols (`scale_shape_manual()`); use

   ```
       values = c("male" = "\u2642","female" = "\u2640")
   ```

   as values
6. set the axes titles
7. change to text of the y axis to percentage
8. etc