Introduction to R

Mandy Vogel

University Leipzig

October 19, 2015

Overview

- Recap and Exercises
- Modelling
    - Choosing a Method
- ANOVA
    - Data
    - Sums of Squares
- Linear Model
- Generalized Linear Models
    - Overview and Data
    - Binary Response Variables
    - Binomial/Logistic Regression
    - The famous O-Ring example
    - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

filter() and select()

```
> subframe <- select(result, measurement, first_pulse, s
> subframe <- filter(result, response_time < 5000) %>%
+      select(measurement, first_pulse, subject)
```

group_by(), mutate() and summarize()

```
> sumframe <- group_by(result, subject) %>%
+     summarise(right.perc = sum(accuracy == 1)/n(),
+               mean.resp.time = mean(response_time, na.rm = T))
> head(sumframe)
Source: local data frame [6 x 3]

   subject right.perc mean.resp.time
1 00436_39  0.6388889       7974.889
2 02411_39  0.7500000       7048.104
3 02544_39  0.6354167       9079.635
4 03858_39  0.7552083       9031.745
5 04517_39  0.7916667       8727.469
6 09458_39  0.7083333       7214.573
```

ggplot

- creating a plot with ggplot involves
  - create a ggplot object (`ggplot()`)
  - mapping qualities of the plot to variables `aes()`
  - add layers consisting of a geometry `geom_specify()` and a statistic (every geom owns a default statistic, so at this time we do not care about statistics in ggplot)

ggplot

Example

```
> p <- ggplot(mtcars, aes(wt, mpg))
> p + geom_point()
```

Exercises

1. load the data and run the lines in the r file to create a new variable containing the sex of the person in the video (result$video, result$video.sex)
2. use dplyr to summarize your data per time point and per person: calculate the 1. proportion of right answers and 2. the mean response time per person and time point useing group_by() and summarize()
3. now visualize the proportion dependent on time: use ggplot() and geom_boxplot() map time to x and the proportion to y using æs() inside of ggplot()
4. repeat the exercise, but this time group additional by the sex of the person in the video
5. visualize for each of the trials (1-48) the mean time and the percentage of right answers use facet_wrap to plot separate plots for each time point

Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

Choosing the appropriate method

It is essential, therefore, that you can answer the following questions:

- Which of your variables is the response variable?
- Which are the explanatory variables?
- Are the explanatory variables continuous or categorical, or a mixture of both?
- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, or a category?

Choosing the appropriate method

Explanatory Variables are	
all continuous	Regression
all categorical	Analysis of variance (ANOVA)
both continuous and categorical	Analysis of covariance (ANCOVA)

Choosing the appropriate method

Response Variables	
(a) Continuous	Normal regression, ANOVA or ANCOVA
(b) Proportion	Logistic regression
(c) Count	Log-linear models
(d) Binary	Binary logistic analysis
(e) Time at death	Survival analysis

The best model is the model that produces the least unexplained variation (the minimal residual deviance), subject to the constraint that all the parameters in the model should be statistically significant (or there are other reasons to keep them).

Choosing the appropriate method

- It is very important to understand that there is not one model;
- there will be a large number of different, more or less plausible models that might be fitted to any given set of data.

Maximum Likelihood

We define best in terms of maximum likelihood.

- given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

We judge the model on the basis how likely the data would be if the model were correct.

Ockham’s Razor

The principle is attributed to William of Ockham, who insisted that, given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation. The most useful statement of the principle for scientists is when you have two competing theories which make exactly the same predictions, the one that is simpler is the better.

Ockham's Razor

For statistical modelling, the principle of parsimony means that:

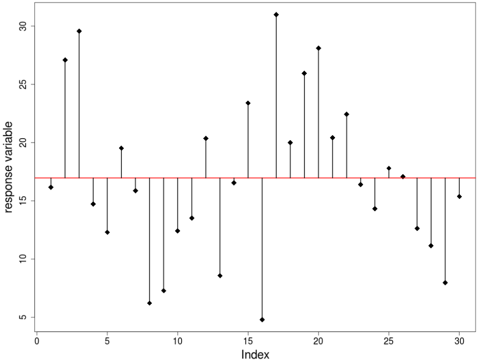
- models should have as few parameters as possible;
- linear models should be preferred to non-linear models;
- experiments relying on few assumptions should be preferred to those relying on many;
- models should be pared down until they are minimal adequate;
- simple explanations should be preferred to complex explanations.

Models

Fitting models to data is the central function of R. There are no fixed rules and no absolutes. The object is to determine a minimal adequate model from a large set of potential models.

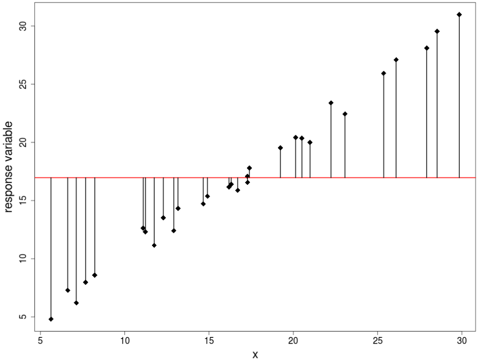
- first we look at the null model

The Null model



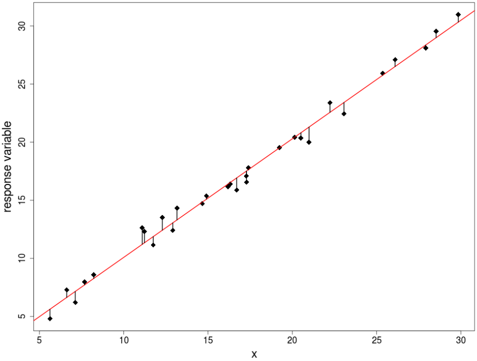
- Just one parameter, the overall mean $\bar{y}$
- Fit: none; $SSE = SSY$
- Degrees of freedom: $n - 1$
- Explanatory power of the model: none

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSR}{SSY}$

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSA}{SSY}$

How to choose...

- models are representations of reality that should be both accurate and convenient
- it is impossible to maximize a model’s realism, generality and holism simultaneously
- the principle of parsimony is a vital tool in helping to choose one model over another
- only include an explanatory variable in a model if it significantly improved the fit of the model (or if there other strong reasons)
- the fact that we went to the trouble of measuring something dœs not mean we have to have it in our model

Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

ANOVA

- a technique we use when all explanatory variables are categorical (factor)
- if there is one factor with three or more levels we use one-way ANOVA (only two levels: t-test should be preferred, would give exactly the same answer since with 2 levels $F = t^2$)
- for more factors there there is two-way, three-way anova
- central idea is to compare two or more means by comparing variances

The Garden Data

A data frame with 14 observations on 2 variables.

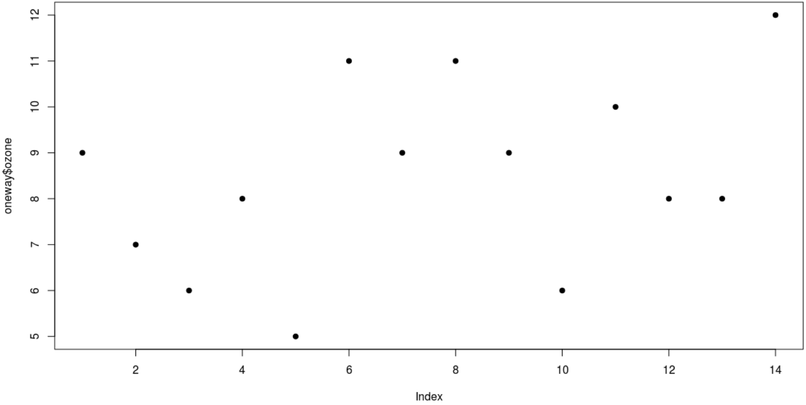
`ozone:`	athmospheric ozone concentration
`garden:`	garden id

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ozone	9	7	6	8	5	11	9	11	9	6	10	8	8	12
garden	a	a	a	b	a	b	b	b	b	a	b	a	a	b

From: Michæl Crawley, The R-Book

Total Sum of Squares

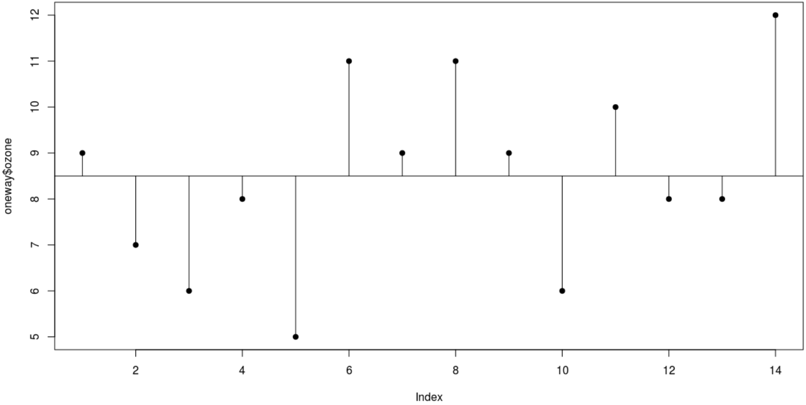
- we plot the values in order they are measured



Total Sum of Squares

- there is a lot of scatter, indicating that the variance in ozone is large
- to get a feel for the overall variance we plot the overall mean (8.5) and indicate each of the residuals by a vertical line

Total Sum of Squares



Total Sum of Squares

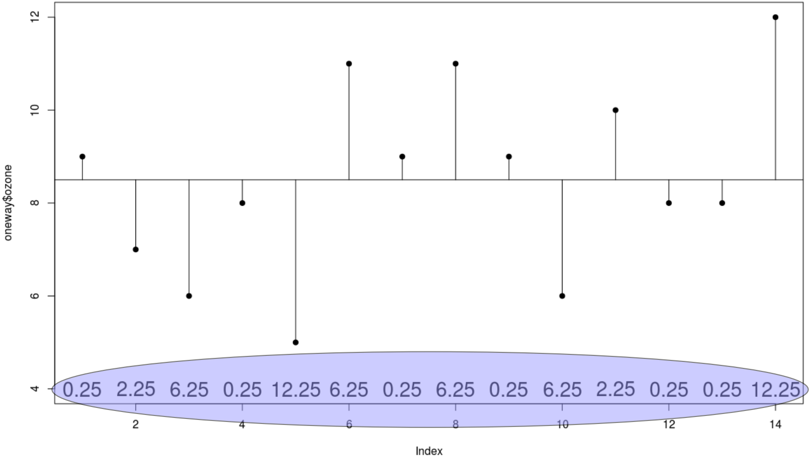
- we refer to this overall variation as the total sum of squares, SSY or TSS

$$SSY = \sum (y - \bar{y})^2$$

Total Sum of Squares

- in this case

$$SSY = 55.5$$

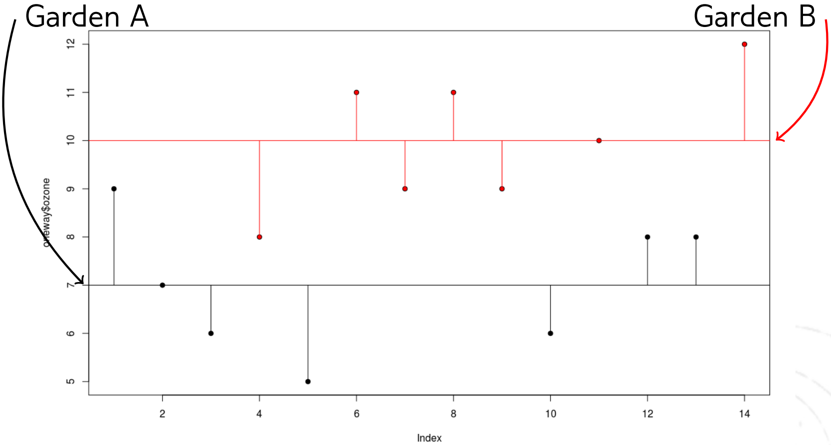


Group Means

- now instead of fitting the overall mean, let us fit the individual garden means

garden	a	b
mean	7	10

Group Means



Garden A
Garden B
oneway$ozone
Index
5
6
7
8
9
10
11
12
2
4
6
8
10
12
14

Group Means

- now we see that the mean ozone concentration is substantially higher in garden B
- the aim of ANOVA is to determine
  - whether it is significantly higher or
  - whether this kind of difference could come by chance alone

Error Sum of Squares

When the means are significantly different then the sum of squares computed from the individual garden means will be smaller than the sum of squares computed from the overall mean.

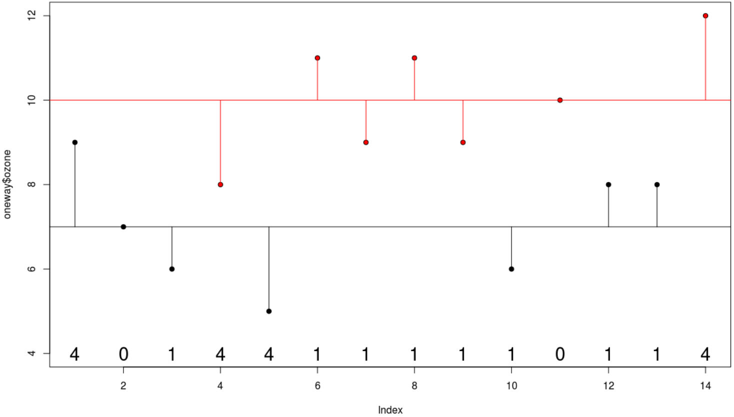
- we define the new sum of squares as the error sum of squares (error in the sense of 'residual')

$$SSE = \sum(y_{gardenA} - \bar{y}_{gardenA})^2 + \sum(y_{gardenB} - \bar{y}_{gardenB})^2$$

Total Sum of Squares

- in this case

$$SSE = 24.0$$



Treatment Sum of Squares

- then the component of the variation that is explained by the difference of the means is called the treatment sum of squares SSA
- analysis of variance is based on the notion that we break down the total sum of squares into useful and informative components

  $$SSY = SSE + SSA$$

  where

  - SSA = explained variation
  - SSE = unexplained variation

ANOVA table

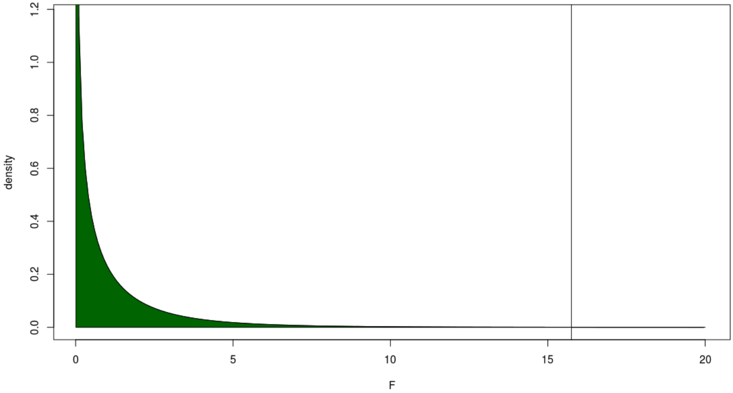
Source	Sum of squares	Degrees of freedom	Mean square	F ratio
Garden	31.5	1	31.5	15.75
Error	24.0	12	$s^2 = 2.0$	
Total	55.5	13		

ANOVA

- now we need to test whether an F ratio of 15.75 is large or small
- we can use a table or software package
- I use here software to calculate the cumulative probability

```
> 1 - pf(15.75,1,12)
[1] 0.001864103
```

ANOVA



ANOVA in R

- in R we use the `lm()` or the `aov()` command and
- the formula syntax `a ~b`
- we assign this to an variable

ANOVA in R

```
mm <- lm(ozone ~ garden, data=oneway)
mm

Call:
lm(formula = ozone ~ garden, data = oneway)

Coefficients:
(Intercept)      gardenb
          7            3
```

ANOVA in R

```
> summary(mm)

Call:
lm(formula = ozone ~ garden, data = oneway)

Residuals:
   Min     1Q Median     3Q    Max 
    -2     -1      0      1      2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.0000     0.5345  13.096 1.82e-08 ***
gardenb       3.0000     0.7559   3.969  0.00186 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 12 degrees of freedom
Multiple R-squared:  0.5676, Adjusted R-squared:  0.5315 
F-statistic: 15.75 on 1 and 12 DF,  p-value: 0.001864
```

ANOVA in R

```
> anova(mm)
Analysis of Variance Table

Response: ozone
          Df Sum Sq Mean Sq F value   Pr(>F)
garden     1   31.5    31.5   15.75 0.001864 **
Residuals 12   24.0     2.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA in R I

```
> m2 <- aov(ozone ~ garden, data=oneway)
> m2
Call:
   aov(formula = ozone ~ garden, data = oneway)

Terms:
                garden Residuals
Sum of Squares    31.5      24.0
Deg. of Freedom      1        12

Residual standard error: 1.414214
Estimated effects may be unbalanced
> summary(m2)
            Df Sum Sq Mean Sq F value  Pr(>F)
garden       1   31.5    31.5   15.75 0.00186 **
Residuals   12   24.0     2.0
---
```

ANOVA in R II

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.lm(m2)

Call:
aov(formula = ozone ~ garden, data = oneway)

Residuals:
   Min     1Q Median     3Q    Max
    -2     -1      0      1      2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0000     0.5345  13.096 1.82e-08 ***
gardenb       3.0000     0.7559   3.969  0.00186 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 12 degrees of freedom
Multiple R-squared:  0.5676, Adjusted R-squared:  0.5315
```

ANOVA in R III

```
F-statistic: 15.75 on 1 and 12 DF,  p-value: 0.001864

> summary(m2)
            Df Sum Sq Mean Sq F value  Pr(>F)   
garden       1   31.5    31.5   15.75 0.00186 **
Residuals   12   24.0     2.0                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Assumptions

Central Assumptions

- independed, normal distributed errors
- equality of variances (homogeneity)

Welch ANOVA I

- generalization of the Welch t-test
- tests whether the means of the outcome variables are different across the factor levels
- assumes sufficiently large sample (greater than 10 times the number of groups in the calculation, groups of size one are to be excluded)
- sensitive to the existence of outliers (only few are allowed)
- the r command is `oneway.test()`
- non-parametric alternative `kruskal.test()`

Exercises

1. Look at the help of the TukeyHSD function. What is its purpose?
2. Execute the code of the example near the end of the help page, interpret the results!
3. install and load the granovaGG package (a package for visualization of ANOVAs), load the `arousal` data frame and use the `stack()` command to bring the data in the long form. Do a anova analysis. Is there a difference at least 2 of the groups? If indicated do a post-hoc test.
4. Visualize your results

Exercises - Solutions I

1. Look at the help of the TukeyHSD function. What is its purpose?
2. Execute the code of the example near the end of the help page, interpret the results!
3. install and load the granovaGG package (a package for visualization of ANOVAs), load the `arousal` data frame and use the `stack()` command to bring the data in the long form. Do a anova analysis. Is there a difference at least 2 of the groups? If indicated do a post-hoc test.

```
> require(granovaGG)
> data(arousal)
> datalong <- stack(arousal)
> m1 <- aov(values ~ ind, data = datalong)
> summary(m1)
            Df Sum Sq Mean Sq F value   Pr(>F)
ind          3  273.4   91.13   10.51 4.17e-05 ***
Residuals   36  312.3    8.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

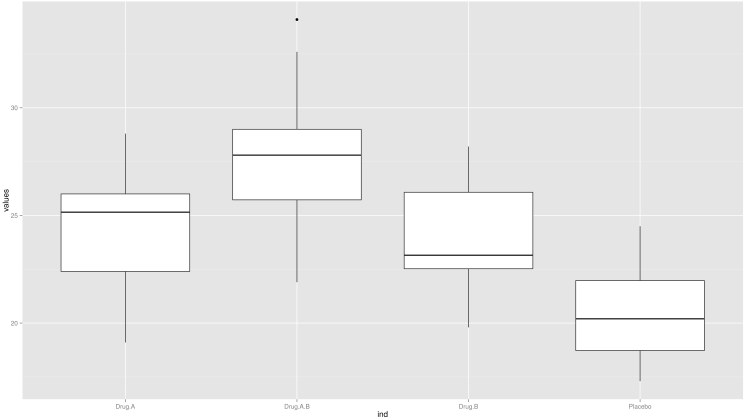
Exercises - Solutions II

```
> TukeyHSD(m1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = datalong)

$ind
                  diff          lwr        upr     p adj
Drug.A.B-Drug.A   3.54  -0.007542384  7.0875424 0.0506601
Drug.B-Drug.A    -0.45  -3.997542384  3.0975424 0.9860554
Placebo-Drug.A   -3.84  -7.387542384 -0.2924576 0.0296168
Drug.B-Drug.A.B  -3.99  -7.537542384 -0.4424576 0.0223986
Placebo-Drug.A.B -7.38 -10.927542384 -3.8324576 0.0000137
Placebo-Drug.B   -3.39  -6.937542384  0.1575424 0.0654726
```

4. Visualize your results

```
> ggplot(datalong,aes(x=ind,y=values)) +
+       geom_boxplot()
```

Exercises - Solutions III



Exercises - Solutions IV

```
> granovagg.1w(datalong$values,group = datalong$ind)

By-group summary statistics for your input data (ordered by group means)
     group group.mean trimmed.mean contrast variance standard.deviation
4  Placebo      20.43        20.30    -3.65     5.83               2.41
3   Drug.B      23.82        23.85    -0.26     7.50               2.74
1   Drug.A      24.27        24.45     0.19     7.89               2.81
2 Drug.A.B      27.81        27.52     3.73    13.49               3.67
  group.size
4         10
3         10
1         10
2         10

Below is a linear model summary of your input data

Call:
lm(formula = score ~ group, data = owp$data)

Residuals:
   Min     1Q Median     3Q    Max 
-5.910 -2.015 -0.075  1.885  6.290 
```

Exercises - Solutions V

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    24.2700     0.9314  26.057  < 2e-16 ***
groupDrug.A.B   3.5400     1.3172   2.688  0.01083 *  
groupDrug.B    -0.4500     1.3172  -0.342  0.73461    
groupPlacebo   -3.8400     1.3172  -2.915  0.00608 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.945 on 36 degrees of freedom
Multiple R-squared:  0.4668, Adjusted R-squared:  0.4223 
F-statistic:  10.5 on 3 and 36 DF,  p-value: 4.173e-05
```

Exercises - Solutions VI

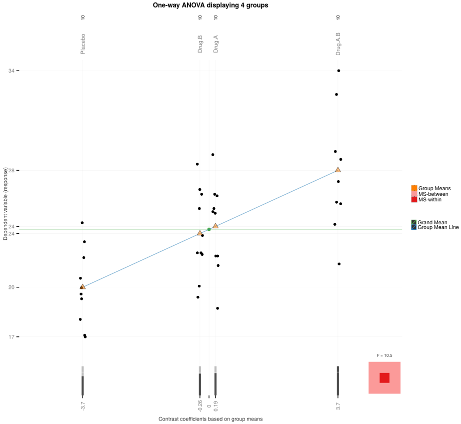


Table of Contents I

- Recap and Exercises
- Modelling
    - Choosing a Method
- ANOVA
    - Data
    - Sums of Squares
- Linear Model
- Generalized Linear Models
    - Overview and Data
    - Binary Response Variables
    - Binomial/Logistic Regression
    - The famous O-Ring example
    - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

The births data

A data frame with 500 observations on the following 8 variables.

`id:`	Identity number for mother and baby.
`bweight:`	Birth weight of baby.
`lowbw:`	Indicator for birth weight less than 2500 g.
`gestwks:`	Gestation period.
`preterm:`	Indicator for gestation period less than 37 weeks.
`matage:`	Maternal age.
`hyp:`	Indicator for maternal hypertension.
`sex:`	Sex of baby: Male, Female.

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

Variables in Models

The response variable must be numeric. Main types are

- Metric (a measurement with units); the easiest case, we will begin with this
- Binary (two values code 0/1)
- Count (aggregated data)
- Failure (dœs the subject fail at end of follow up)

Explanatory variables can be

- Numeric
- Factor

Metric Response, Numeric explanatory variable

Assuming that the relationship of `bweight` with `gestwks` is roughly linear we can find the linear effect on `bweight` of a unit increase in `gestwks` with

```
> m.1 <- lm(bweight ~ gestwks, data=births)
```

- `lm()` is the linear model function
- `bweight ~ gestwks` is the model formula
- `m` is a model object (containing all information about our model), there are certain functions to extract these information, e.g.:

```
> coef(m.1)
(Intercept)     gestwks
 -4489.1398    196.9726
```

One extra week of gestation produces an extra 197g of baby.

Extractor functions

```
> summary(m.1)

Call:
lm(formula = bweight ~ gestwks, data = births)

Residuals:
     Min       1Q   Median       3Q      Max
-1698.40  -280.14    -3.64   287.61  1382.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4489.140    340.899  -13.17   <2e-16 ***
gestwks       196.973      8.788   22.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 449.7 on 488 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.5073, Adjusted R-squared: 0.5062
F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

Extractor functions

```
> coef(m.1)
(Intercept)     gestwks
 -4489.1398    196.9726
> confint(m.1)
                 2.5 %     97.5 %
(Intercept) -5158.9503 -3819.3293
gestwks       179.7054   214.2399
```

Visualize Simple Linear Regression

- for visualization of simple linear regression ggplot can be easily used
- with `geom_smooth()` it provides a layer for smoothing

Exercise:

- create a scatter plot using ggplot the independent variable on the x-axis and the dependent variable on the y-axis
- add `geom_smooth()`
- what is the result?

Visualize Simple Linear Regression

- the change the fitting method set the argument `method` of `geom_smooth()`
- in our case set `method` to “lm”

Other Useful Functions

The model object is a list of different elements each of which can be accessed separately (see `str(m)` for the full list).

Other useful functions:

- `print(m)` simple display
- `plot(m)` produces various diagnostic plots based on residuals
- `fitted(m)` returns a vector of fitted values
- `resid(m)` returns a vector of residuals
- `predict(m, newdata)` predicts the response for new values of the explanatory variables
- `deviance(m)` residual sum of squares
- `df.residual(m)` for the residual degrees of freedom
- `vcov(m)` variance-covariance matrix

code file for examples

Explanatory Variable is a Factor

The effect of `hyp` (2-level factor) on `bweight` is obtained with

```
> m.aov <- lm(bweight ~ hyp, data=births)
> coef(m.aov)
(Intercept)    hyphyper
  3198.9042   -430.6959
```

Omitting the intercept gives the mean `bweight` at the two levels of `hyp`

```
> m.aov2 <- lm(bweight ~ -1 + hyp, data=births)
> coef(m.aov2)
hypnormal  hyphyper
 3198.904  2768.208
```

Explanatory Variable is a Factor

```
> summary(m.aov)

Call:
lm(formula = bweight ~ hyp, data = births)

Residuals:
    Min      1Q  Median      3Q     Max 
-2570.9  -286.4    69.1   383.9  1667.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3198.90      29.96 106.768  < 2e-16 ***
hyphyper     -430.70      78.95  -5.455 7.73e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 619.8 on 498 degrees of freedom
Multiple R-squared:  0.05638, Adjusted R-squared:  0.05449 
F-statistic: 29.76 on 1 and 498 DF,  p-value: 7.729e-08
```

Exercise

1. What is the appropriate plot to visualize the effect of `hyp`?
2. What is the most common test to test these effect?

A Multivariable Model

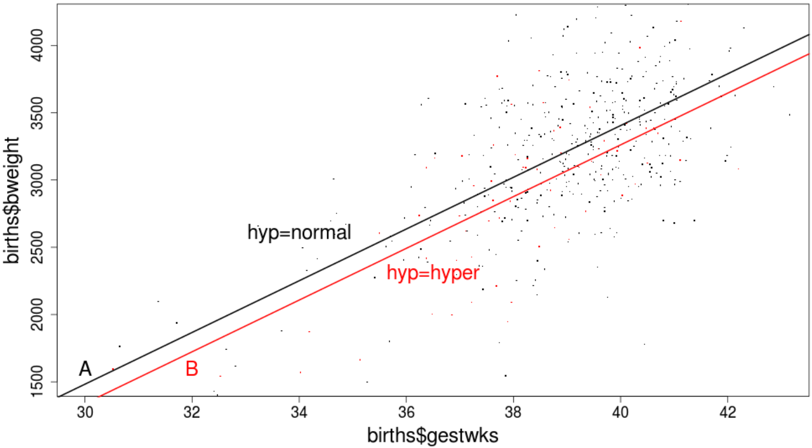
The joint effect of `hyp` and `gestwks` on `bweight` is obtained with

```
> m.3 <- lm(bweight ~ hyp + gestwks, data=births)
```

```
               Estimate
(Intercept) -4285.002
hyphyper     -143.675 (level 2 vs. level 1)
gestwks       192.238 (increase per week)
```

The effect of `hyp` is attenuated (from -430.7 to -143.7). This suggests that much of the effect of hypertension on birth weight is mediated through a shorter gestation period.

A Model With Both `gestwks` and `hyp`



The effect of `gestwks` is the slope of the lines A and B (assumed to be the same). The effect of `hyp` ist the vertical distance between them.

Interaction Models in lm

To specify an interaction term in lm, change the model formula from

Input

```
> m.3 <- lm(bweight ~ hyp + gestwks, data=births)
```

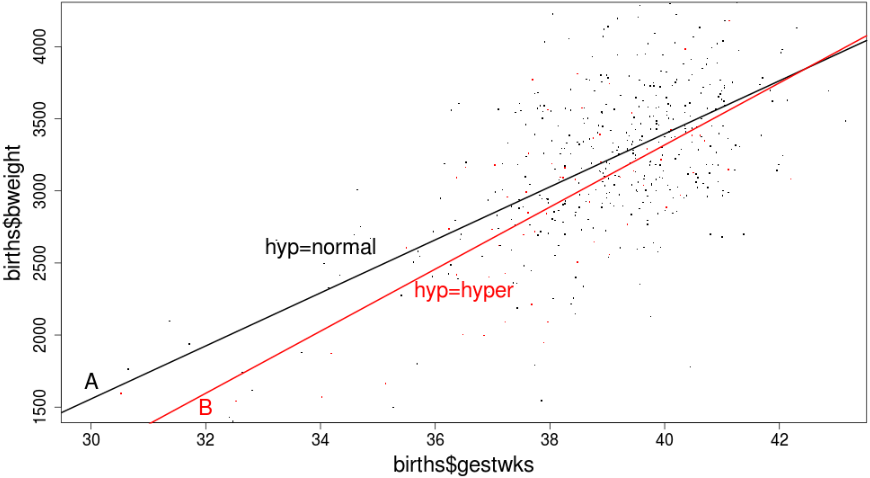
to

```
> m.5 <- lm(bweight ~ hyp + gestwks + hyp:gestwks, data=b
```

or shorter

```
> m.5 <- lm(bweight ~ hyp * gestwks, data=births)
```

Interaction Between gestwks and hyp



Interactions Models in lm

Output

```
                   Estimate
(Intercept)        -3960.82
hyphyper           -1332.66 (level 2 vs level 1 - inter
gestwks              183.91
hyphyper:gestwks      31.39 (level 2 vs level 1 - slope
```

Now the effect of `hyp` is more difficult to explain, because it is not constant. The effect of -1332 is valid on a hypothetical gestational age of 0. Which dœsn't make sense. You could scale the `gestwks` variable.

```
> births$gwsc <- births$gestwks-40
> m <- lm(bweight ~ hyp * gwsc, data=births)
```

Interactions Models in `lm`

Input/Output

```
                 Estimate
(Intercept)    3395.60329
hyphyper        -77.25215 (level 2 vs level 1 - interc
gwsc            183.91048
hyphyper:gwsc    31.38510 (level 2 vs level 1 - slope)
```

How much is explained? - aov

In the Null-Model we have seen that $SSE = SSY$ (the error sum of squares is equal to the total sum of squares in y) and therefore the Null-Model explaines nothing of the overall variance. So the fraction how much of the overall variance is explained by our model regarding to the overall variance is a first measure for the fit of the model...

- the simple model with one explanatory variable

```
> m <- lm(bweight ~ gestwks, data=births)
> anova(m)
Analysis of Variance Table

Response: bweight
           Df    Sum Sq   Mean Sq F value    Pr(>F)    
gestwks     1 101603845 101603845  502.36 < 2.2e-16 ***
Residuals 488  98698698    202251                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

How much is explained? - aov

- in the second column of the summary we see the regression sum of squares (SSR) in the first line and in the second line the error sum of squares (SSE). So the total sum of squares (SSY - a measure for the overall variation) is the sum of both:

```
> sum(anova(m)$Sum)
[1] 200302543
```

- and the fraction is

```
> anova(m)$Sum[1]/sum(anova(m)$Sum)
[1] 0.5072519
```

How much is explained? - aov

- this is r-squared

```
> summary(m)$r.squared
[1] 0.5072519
```

- which you can extract from the summary of the model

```
> summary(m)
Call:
lm(formula = bweight ~ gestwks, data = births)
Residuals:
     Min        1Q    Median        3Q       Max
-1698.40  -280.14     -3.64    287.61   1382.24
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4489.140    340.899  -13.17   <2e-16 ***
gestwks       196.973      8.788   22.41   <2e-16 ***

Residual standard error: 449.7 on 488 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.5073, Adjusted R-squared: 0.5062
F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

Exercise

The dataset `teengamb` is part of the `faraway` package and concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the `sex`, `status`, `income` and `verbal` score as predictors. Present the output.

(a) What percentage of variation in the response is explained by these predictors?

(b) Which observation has the largest (positive) residual? Give the case number.

(c) Compute the mean and median of the residuals.

(d) Compute the correlation of the residuals with the fitted values.

(e) Compute the correlation of the residuals with the income.

(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

Beyond Linear Models

- linear models are central to the practice of statistics
- the standard linear model cannot handle non-normal responses, such as counts or proportions. This motivates the development of generalized linear models that can represent categorical, binary and other response types.

Beyond Linear Models

- Some data has a grouped, nested or hierarchical structure. Repeated measures, longitudinal and multilevel data consist of several observations taken on the same individual or group. This induces a correlation structure in the error. mixed effect models allow the modeling of such data.
- non-parametric regression models: Methods such as additive models, trees and neural networks allow a more flexible regression modeling of the response that combine the predictors in a nonparametric manner.

Generalized Linear Models

Linear modeling assumes constant variance and normally distributed errors. Certain kinds of respond variables lack these constraints. GLMs are excellent at dealing with it.

Input/Output

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

give the same answer. The model formula is the same for both, but for `glm()` it is necessary to specify the family of likelihoods which will be used to fit the model.

The `glm()` function allows us to fit other models including logistic regression and Poisson regression.

Beyond Linear Models

- We begin with a binary response variable:

Bernoulli model

- $f(y;p) = p^y(1-p)^{1-y}$
- it is modelled with a logit as canonical link

$$\eta = \log(\frac{p}{1-p})$$

- i.e. our linear model looks like

$$\eta = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

with a binomial error structure

Exercise

- We need data. So...
- The excel file data.xlsx contains the worksheets mother and child containing respective parts of the births data set. Use the `read_excel()` command to read both data sets and use `merge()` to join them
- Hints:
  - the `read_excel()` function is a part of the `readxl` package
  - check which columns are contained in both data frames and use them for merging

Data Structure

The structure of the data should be of the following form:

Input/Output

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

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

Binary Response Variable

Many statistical problems involve binary response variables. For example, we often classify individuals as:

- dead or alive,
- occupied or empty,
- healthy or diseased,
- wilted or turgid,
- male or female,
- literate or illiterate,
- mature or immature,
- solvent or insolvent, or
- employed or unemployed.

Binary Response Variable

Question

Which variable in the births data set is (most) suitable to use as binary response given this data set? Why?

Binary Response Variable

In order to work with correct coded variables, we transform hyp and lowbw to categorical variables, and define normal as reference level for both of them

Input/Output

```
> summary(births$hyp)
  Length     Class      Mode
  500 character character
> births$hyp <- factor(births$hyp,levels = c("normal","hyper"))
> summary(births$hyp)
  normal  hyper
   428     72
> summary(births$lowbw)
  Length     Class      Mode
  500 character character
> births$lowbw <- factor(births$lowbw,levels = c("normal","low"))
> summary(births$lowbw)
  normal    low
  440     60
```

Predicting Low Birth Weight

- Now we are more interested in predicting birth weight under 2500g (`lowbw`).
- This requires a model where the outcome is not metric, but binary.
- For a binary response we use a `glm()` with a binomial family.
- the binomial family uses a logit link as default

Predicting Low Birth Weight

How it looks in R:

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
```

Predicting Low Birth Weight

What it looks like as a math formula:

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

Interpreting the Cœfficients

- While using a binomial family R uses a logit as link function.
- Therefore the returned estimates are log odds (Intercept) or log odds ratios (for the parameters).
- The `arm` package contains a function `invlogit()` which dœs invert the logit function.
- Alternatively you can use the formula

$$\text{logit}^{-1} = \frac{\exp(x)}{1 + \exp x}$$

Interpreting the Cœfficients

- Our example is a simple analysis of variance.
- Our model here is

$$\text{Pr(lowbw)} = \text{logit}^{-1}(-2.2721 + 1.3166 \cdot \text{hyp})$$

- We have two levels of our predictor variable `hyp`: normal and hyp.
- For the reference level normal hyp $= 0$
- in this case we get

$$\text{Pr(lowbw)} = \text{logit}^{-1}(-2.2721 + 1.3166 \cdot 0) = \text{logit}^{-1}(-2.2721)$$

which is a log odds as mentioned before, so

Input/Output

```
> invlogit(coef(m)[1])
(Intercept)
 0.09345794
```

Interpreting the Cœfficients

- The result is the probability of low birth weight within the group of moms with normal blood pressure. We can check this by using `table`:

Input/Output

```
> table(births$lowbw,births$hyp)

         normal hyper
  normal    388    52
  low        40    20
> 40/(388+40)
[1] 0.09345794
```

Interpreting the Cœfficients

- for the level hyp (i.e. hyp $= 1$) we get a difference of 1.3166 on the logit scale

$$\text{Pr(lowbw)} = \text{logit}^{-1}(-2.2721 + 1.3166 \cdot 1)$$

- which turns out to be

Input/Output

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

- so the probability for low birth weight is 27.8% in for moms with high blood pressure

Understanding the Cœfficients

- in this simple case, the response variable gives the probability for low birth weight for each of the two groups of moms (with and without high blood pressure)
- you can get the result also using (a) a proportion test:

```
> prop.test(c(20,40),c(72,428))

2-sample test for equality of proportions with continuity

data:  c(20, 40) out of c(72, 428)
X-squared = 18.121, df = 1, p-value = 2.073e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06913673 0.29950294
sample estimates:
    prop 1     prop 2
0.27777778 0.09345794
```

Understanding the Cœfficients

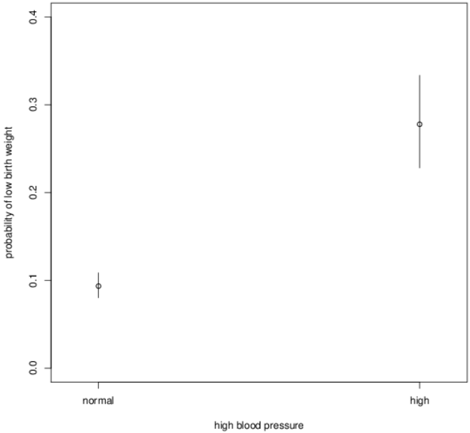
- or (b) a χ^2-test:

Input/Output

```
> chisq.test(table(births$lowbw,births$hyp))

Pearson's Chi-squared test with Yates' continuity correction

data:  table(births$lowbw, births$hyp)
X-squared = 18.121, df = 1, p-value = 2.073e-05
```

Understanding the Cœfficients

- a hand made plot

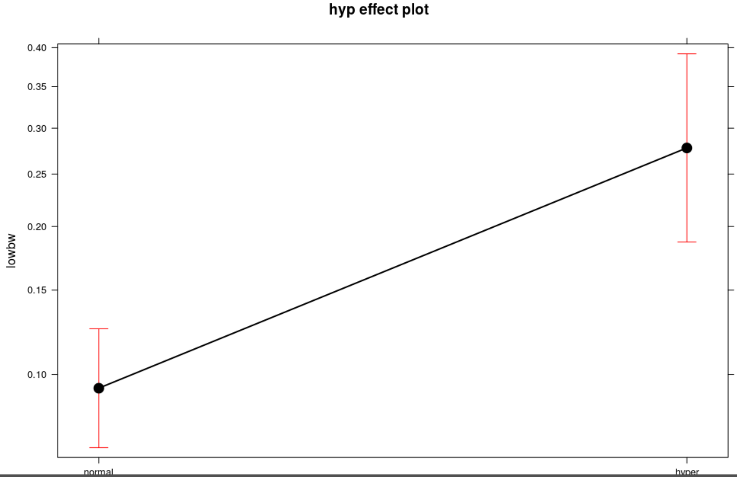


Understanding the Cœfficients

- and one effect plot (`effects` package)

Input

```
> plot(Effect("hyp",m))
```



Understanding the Cœfficients

- btw: `Effect()` gives you the probabilities without using a explicit transformation

Input/Output

```
> Effect("hyp",m)

 hyp effect
hyp
    normal      hyper
0.09345794 0.27777778
```

Controlling

Controlling the effect of `hyp` on `lowbw` for `sex`

Input/Output

```
> m2 <- glm(lowbw ~ hyp+sex, family=binomial, data=births)

             Estimate   StdErr  Pr(>|z|)
(Intercept)  -2.5088   0.2331  < 2e-16 ***
hyphyper      1.3625   0.3144 1.47e-05 *** hyp controlled for
sexF          0.4473   0.2843    0.116      sex controlled for
```

When you control for a variable you are assuming that any interaction can be ignored.

Interaction (effect modification)

- We add an interaction term to the model

Input/Output

```
> m3 <- glm(lowbw ~ hyp + sex + hyp:sex,
+           family=binomial, data=births) # or shorter
> m3 <- glm(lowbw ~ hyp*sex, family=binomial,
            data=births)
```

Interaction (effect modification)

- we have four estimates now, and to get the effects in terms of probabilites we need to type

Input/Output

```
> m3 <- glm(lowbw ~ hyp*sex, family=binomial, data=births)
> summary(m3)

Call:
glm(formula = lowbw ~ hyp * sex, family = binomial, data = births)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8090  -0.5074  -0.3749  -0.3749   2.3195

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.6198     0.2674  -9.796  < 2e-16 ***
hyphyper        1.6707     0.4326   3.862 0.000112 ***
sexF            0.6347     0.3421   1.855 0.063535 .
hyphyper:sexF  -0.6507     0.6366  -1.022 0.306694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Interaction Cœfficients

Input/Output

```
> invlogit(coef(m3)[1])
  (Intercept)
   0.0678733
> invlogit(coef(m3)[1] + coef(m3)[2])
  (Intercept)
   0.2790698
> invlogit(coef(m3)[1] + coef(m3)[3])
  (Intercept)
   0.1207729
> invlogit(coef(m3)[1] + coef(m3)[2] + coef(m3)[3] + coef(m3)[4])
  (Intercept)
   0.2758621
```

Exercises

You can calculate the effects by hand and using the `invlogit()` function, but this becomes a little annoying, the `allEffects()` function provides a nicer way to do the same.

- now you have three models, use the `Effects()`, `allEffects()` and the `plot()` function to get the following information:
  1. the estimated probability for moms with hypertension to get a baby with low birth weight for all three models
  2. is their a difference in effects between boys and girls? Which model can answer this question?

Testing for Interaction

- Do we need to keep the interaction term?

Input/Output

```
> m2 <- glm(lowbw ~ hyp+sex, family=binomial,
+                            data=births)
> m3 <- glm(lowbw ~ hyp*sex, family=binomial,
+                            data=births)
> anova(m2,m3,test="Chisq")

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       497     348.34
2       496     347.29  1   1.0561   0.3041
```

- The `anova` function conducts an analysis of variance – a test of significance between two nested models.
- The interaction term dœs not improve the fit - so we leave it out and keep the simpler model.

Stratified Effects

- When there is a strong interaction it may be best to report stratified effects.
- Omitting the main effect of `hyp` in an interaction model gives us the effect of `hyp` within strata of `sex`.

Stratified Effects

Input/Output

```
> m4 <- glm(lowbw ~ sex + sex:hyp, family=binomial, data=birth
> summary(m4)

Call:
glm(formula = lowbw ~ sex + sex:hyp, family = binomial, data =

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8090  -0.5074  -0.3749  -0.3749   2.3195

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.6198     0.2674  -9.796  < 2e-16 ***
sexF            0.6347     0.3421   1.855 0.063535 .
sexM:hyphyper   1.6707     0.4326   3.862 0.000112 ***
sexF:hyphyper   1.0200     0.4670   2.184 0.028952 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

Stratified Effects

A slightly shorter way to define the same model:

Input/Output

```
> m4 <- glm(lowbw ~ sex/hyp, family=binomial, data=births)
> m4

Call:  glm(formula = lowbw ~ sex/hyp, family = binomial, data

Coefficients:
  (Intercept)           sexF  sexM:hyphyper  sexF:hyphyper
      -2.6198         0.6347         1.6707         1.0200

Degrees of Freedom: 499 Total (i.e. Null);  496 Residual
Null Deviance:      366.9
Residual Deviance: 347.3  AIC: 355.3
```

Exercise

- compare the effects in `m3` and `m4`

Understanding the Cœfficients

```
> ftable(births$hyp,
+        births$sex,
+        births$lowbw)
          normal low

normal M     206  15
       F     182  25
hyper  M      31  12
       F      21   8
```

```
## male/normal bp
> 15/(206+15)
[1] 0.0678733
## female/normal bp
> 25/(25+182)
[1] 0.1207729
## male/high bp
> 12/(12+31)
[1] 0.2790698
## female/high bp
> 8/(8+21)
[1] 0.2758621
```

Simple Logistic Regression

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

Input

```
> m5 <- glm(lowbw ~ gestwks, family=binomial, data=births)
```

- and as math formula

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

Simple Logistic Regression

- where the output look similar to the output above

Input/Output

```
> summary(m5)

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)[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)

- the parameter for `gestwks` describes how fast the probability decreases with increasing gestation 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(m5)[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

Exercise

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

Input/Output

```
> Effect("gestwks",m5)

 gestwks effect
gestwks
        25         30         35         40 
0.99992022 0.99299324 0.61574996 0.01779725 
> Effect("gestwks",m5,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

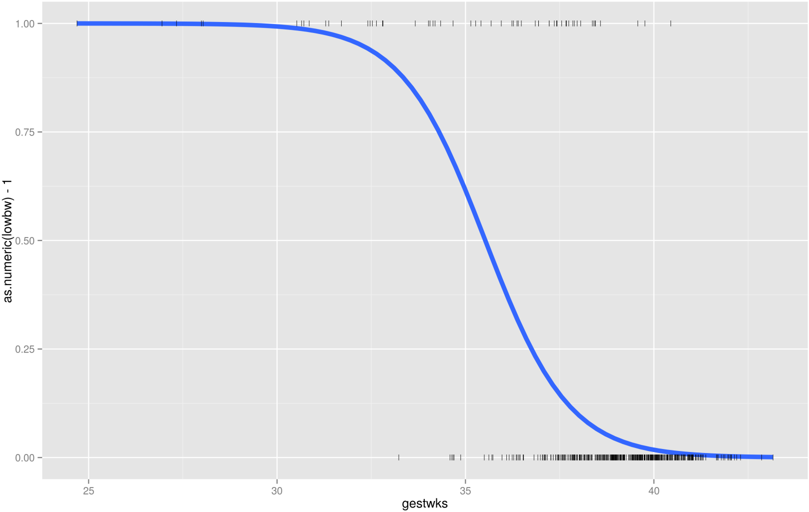
ggplot() and glm()

- 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`)

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)

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

```
> 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.

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

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

The Challenger Disaster Example

- the output looks familiar:

```
> summary(m)
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

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

- the intercept

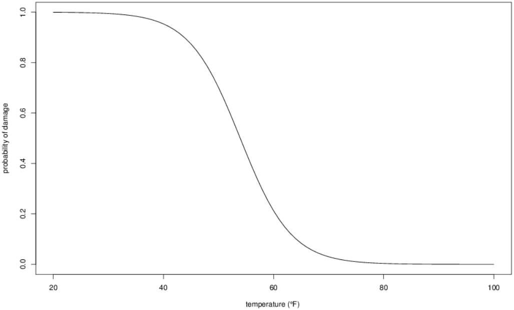
```
> invlogit(coef(m)[1])
(Intercept)
  0.9999914
```

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

- the parameter for temperature describes how fast the probability decreases with increasing temperature

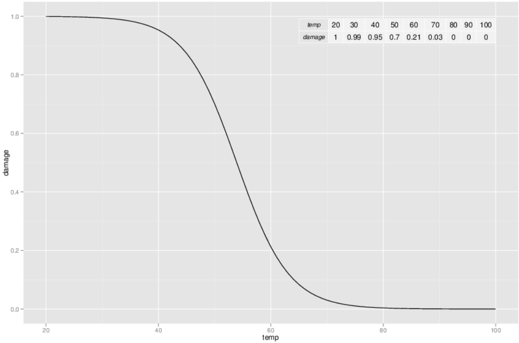
Understanding the Cœfficients

```
> tf <- 20:100
> pd <- predict(m,newdata=list(temp=tf), type="response")
> plot(tf,pd,type="l",
+  xlab=expression(paste("temperature (",degree,"F)",sep=" ")),
+  ylab="probability of damage")
```



Understanding the Cœfficients

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



temp	20	30	40	50	60	70	80	90	100
damage	1	0.99	0.95	0.7	0.21	0.03	0	0	0

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

```
> infection <- read.table("infection.txt",header=T)
> summary(infection)
    infected             age               sex       
 Min.   :0.000   Min.   :  2.00   Min.   :0.000  
 1st Qu.:0.000   1st Qu.: 46.00   1st Qu.:0.000  
 Median :0.000   Median : 84.50   Median :1.000  
 Mean   :0.324   Mean   : 93.69   Mean   :0.514  
 3rd Qu.:1.000   3rd Qu.:139.25   3rd Qu.:1.000  
 Max.   :1.000   Max.   :200.00   Max.   :1.000  
```

Parasite Infection Example

```
> m <- glm(infected~age*sex,family=binomial,
+                                data=infection)
> summary(m)
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

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

- for females is the probability at age 0

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

Parasite Infection Example

- 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

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

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

Parasite Infection Example

- 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

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

- `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

Parasite Infection Example

- 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

```
> allEffects(m)
 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
```

Parasite Infection Example

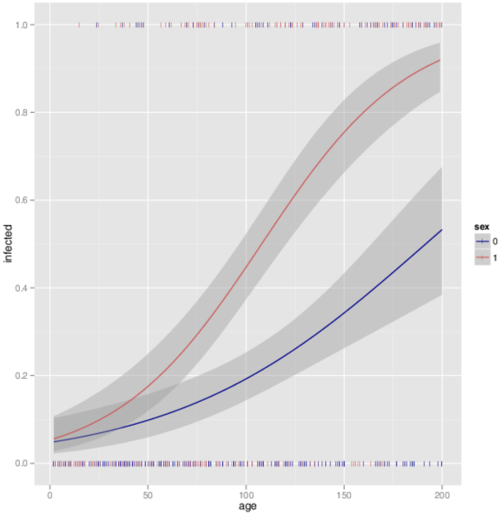


Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

Count Data

- a great deal of the data collected is in the form of counts
- for example:
  - number of individuals that died
  - number of firms going
  - bankrupt, the number of days of frost,
  - the number of red blood cells on a microscope slide, and the
  - number of craters in a sector of lunar landscape
- with count data, the number 0 often appears as a value of the response (zero inflated data)

Count Data

- we must consider a different cases in dealing with data on frequencies: cases
  - where we count how many times something happened, but we have no way of knowing how often it did not happen (e.g. lightning strikes, bankruptcies, deaths, births).
  - count data on proportions, where we know the number doing a particular thing, but also the number not doing that thing (e.g. the proportion dying, sex ratios at birth, proportions of different groups responding to a questionnaire)

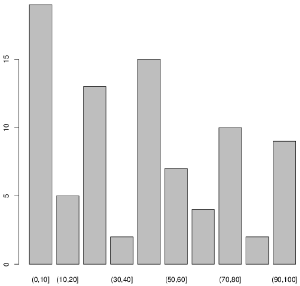
A Poisson Regression

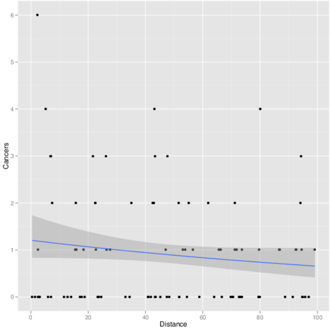
- The following example has a count (the number of reported cancer cases per year per clinic) as the response variable
- and a single continuous explanatory variable (the distance from a nuclear plant to the clinic in km).
- The question is whether or not proximity to the reactor affects the number of cancer cases.

```
> cancer <- read.table("clusters.txt",header=T)
> head(cancer)
  Cancers Distance
1       0 11.46952
2       0 66.55395
3       0 47.46230
4       0 48.38129
5       0 73.76534
6       0 70.57555
```

Count Data

- look at a barplot (cut the `Distance` variable in ten classes) and a scatter plot





Count Data

- There seems to be a downward trend in cancer cases with distance. But is the trend significant?

```
> m <- glm(Cancers~Distance,family=poisson,data=cancer)
> summary(m)
Call:
glm(formula = Cancers ~ Distance, family = poisson,
                                     data = cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5504  -1.3491  -1.1553   0.3877   3.1304
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.186865   0.188728   0.990   0.3221
Distance    -0.006138   0.003667  -1.674   0.0941 .
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
AIC: 262.41
```

Count Data

- The trend dœs not look to be significant, but look at the residual deviance:
- It is assumed that this is the same as the residual degrees of freedom (because the errors are supposed to be Poisson distributed)
- this indicates that we have overdispersion (extra, unexplained variation in the response).
- we compensate for the overdispersion by refitting the model using quasi-Poisson rather than Poisson errors

Count Data

- the refitted model

```
> m <- glm(Cancers~Distance,family=quasipoisson,data=cancer)
> summary(m)
Call:
glm(formula = Cancers ~ Distance,
               family = quasipoisson, data = cancer)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5504  -1.3491  -1.1553   0.3877   3.1304

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.186865   0.235364   0.794    0.429
Distance    -0.006138   0.004573  -1.342    0.183

(Dispersion parameter for quasipoisson family
                                 taken to be 1.555271)
    Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
```

Interpreting the Cœfficients

- the estimates remained the same, but the p-vals changed
- so there is no compelling evidence to support the existence of a trend in cancer incidence with distance from the nuclear plant (this is a completely made up example, neither considering varying population nor clinic density)

Interpreting the Cœfficients

- if you use glms with Poisson errors, the default link function is `log`
- so the parameter estimates and the predictions from the model (the 'linear predictor') are in logs, and need to be antilogged
- so we have the following following formula for our model

$$\text{count} = \exp\left(0.186865 - 0.006138 \cdot \text{Distance}\right)$$

- antilog the intercept:

```
  > exp(coef(m)[1])
(Intercept)
   1.205464
```

- get 1.2 expected cases at a distance of zero

Interpreting the Cœfficients

- the slope for `Distance` is a bit easier to interpret than with a logit link

  ```
  > exp(coef(m)[2])
   Distance
  0.9938805
  ```

  means that for every additional km distance you get 0.006 less cancer cases (it is nicer to say for every 10 km the expected count of cancer cases decreases by 6%)

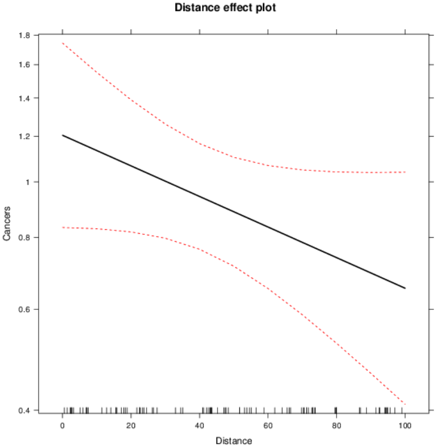
Interpreting the Cœfficients

- again, the `effects` package is very helpful to give an overview

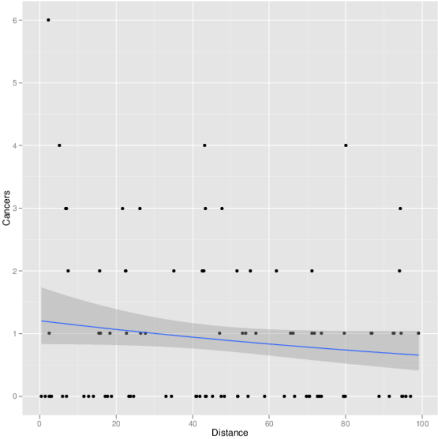
```
> allEffects(m,xlevels=list(Distance=seq(0,100,by=10))
+ )
 model: Cancers ~ Distance

 Distance effect
Distance
        0        10        20        30
1.2054642 1.1336940 1.0661968 1.0027182
       40        50        60        70
0.9430189 0.8868740 0.8340718 0.7844133
       80        90       100
0.7377114 0.6937900 0.6524835
```

Interpreting the Cœfficients

- now the effect plot and the (non-significant) fitted line can be drawn



Interpreting the Cœfficients



Anova with Count Data

- next example the response variable is a count of infected blood cells per mm^2 on microscope slides prepared from randomly selected individuals
- explanatory variables are smoker (logical, yes or no)
- and body mass score (three levels, normal, overweight, obese)
- so we fit the following model (including the interaction term)

Anova with Count Data

```
> m <- glm(cells~smoker*weight,family=poisson,data=cells)
> summary(m)
Call:
glm(formula = cells ~ smoker * weight, family = poisson, data = c
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6511  -1.1742  -0.9148   0.5533   3.6436
Coefficients:            Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.8712     0.1302  -6.692 2.20e-11 ***
smokerTRUE                0.8224     0.1833   4.486 7.27e-06 ***
weightobese               0.4993     0.1671   2.987 0.002817 **
weightover                0.2618     0.1866   1.404 0.160465
smokerTRUE:weightobese    0.8063     0.2296   3.511 0.000446 ***
smokerTRUE:weightover     0.4935     0.2546   1.939 0.052548 .

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1052.95  on 510  degrees of freedom
Residual deviance:  792.85  on 505  degrees of freedom
AIC: 1318.5
```

Anova with Count Data

- again we see overdispersion (residual deviance > degrees of freedom)
- we compensate by refitting the model using quasi-Poisson errors

Anova with Count Data

```
> m <- glm(cells~smoker*weight,family=quasipoisson,data=cells)
> summary(m)
Call:
glm(formula = cells ~ smoker * weight, family = quasipoisson,
    data = cells)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6511  -1.1742  -0.9148   0.5533   3.6436
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.8712     0.1760  -4.950 1.01e-06 ***
smokerTRUE               0.8224     0.2479   3.318 0.000973 ***
weightobese              0.4993     0.2260   2.209 0.027598 *
weightover               0.2618     0.2522   1.038 0.299723
smokerTRUE:weightobese   0.8063     0.3105   2.597 0.009675 **
smokerTRUE:weightover    0.4935     0.3442   1.434 0.152226

(Dispersion parameter for quasipoisson family taken to be 1.82792
```

Interpreting the Cœfficients

- remember poisson has log as link so

  ```
  > exp(coef(m)[1])
  (Intercept)
    0.4184397
  ```

  is the expected count of infected blood cells for a normal weighted non-smoker

- all the other estimates are interpretable as factors (because of the log link!)
- so a smoker has

  ```
  > exp(coef(m)[2])
  smokerTRUE
    2.276029
  ```

  more than twice as many infected cells which is

  ```
  > exp(coef(m)[1])*exp(coef(m)[2])
  (Intercept)
     0.952381
  ```

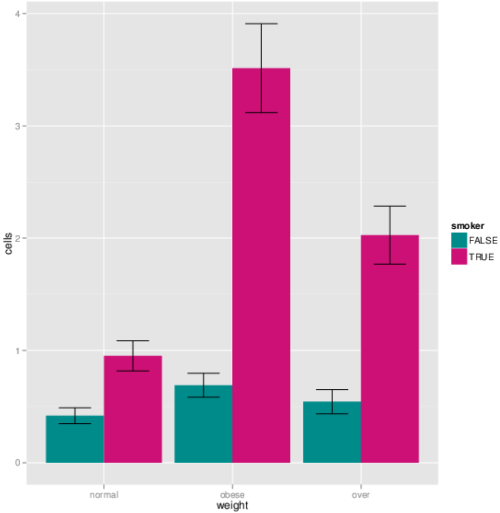
Interpreting the Cœfficients

- unfortunately `effect()` dœs not work on our model object, so we use `tapply()` (for simple models a good alternative, as soon as I remove an interaction term, or nested effects this dœs not work anymore)

```
> with(cells,tapply(cells,list(smoker,weight),mean))
         normal     obese      over
FALSE 0.4184397 0.6893939 0.5436893
TRUE  0.9523810 3.5142857 2.0270270
```

Interpreting the Cœfficients

- for visualization we use barplot with errorbars indicating the standard error



Ancova with Count Data

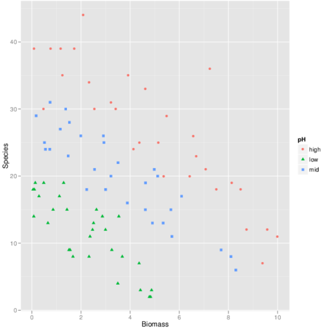
- last example: analysis of covariance
- response is a count of the number of plant species on plots
- that have different biomass (a continuous explanatory variable) and
- different soil pH (a categorical variable with three levels: high, mid and low)

```
> species<-read.table("species.txt",header=T)
> head(species)
    pH   Biomass Species
1 high 0.4692972      30
2 high 1.7308704      39
3 high 2.0897785      44
4 high 3.9257871      35
5 high 4.3667927      25
6 high 5.4819747      29
```

Ancova with Count Data

- this time we begin with a scatter plot

```
p <- ggplot(species,aes(x=Biomass,y=Species,
+                         shape=pH,colour=pH)) +
     geom_point()
```



Ancova with Count Data

- we see: number of species declines with Biomass
- soil pH has a big effect on Species
- Dœs the slope of the relationship between Species and Biomass depend on pH?

Ancova with Count Data

- define the model and look at the summary

```
> m <- glm(Species~Biomass*pH,family=poisson,data=species)
> summary(m)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
...
```

Ancova with Count Data

- test for the need for different slopes by comparing this maximal model (with six parameters) with a simpler model with different intercepts but the same slope

```
> m2 <- glm(Species~Biomass+pH,
+                  family=poisson,data=species)
> anova(m,m2,test="Chi")
Analysis of Deviance Table

Model 1: Species ~ Biomass * pH
Model 2: Species ~ Biomass + pH
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         84     83.201
2         86     99.242 -2   -16.04 0.0003288 ***
```

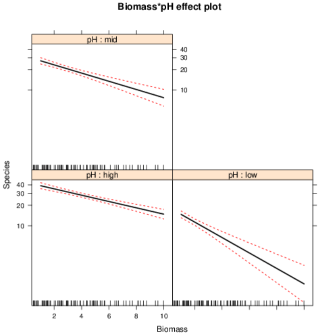
- AIC: m: 514.4; m2: 526.4

Ancova with Count Data

- slopes are very significantly different $p = 0.00033$, so it is justified to retain the more complicated model
- finally, we have a look on the effects and then draw the fitted lines through the scatterplot using the plot object p from above

```
> allEffects(m,xlevels=list(Biomass=1:10))
 model: Species ~ Biomass * pH
 Biomass*pH effect
      pH
Biomass     high       low       mid
     1  38.89998 14.737487 27.048707
     2  34.94810 11.338867 23.538030
     3  31.39769  8.724005 20.483007
     4  28.20797  6.712158 17.824498
     5  25.34229  5.164264 15.511039
     6  22.76775  3.973330 13.497847
     ....
```

Ancova with Count Data



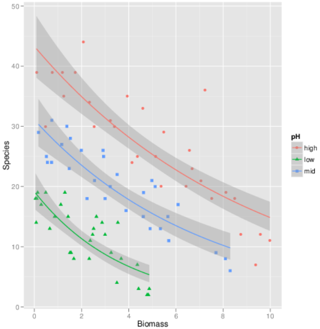


Table of Contents I

- Recap and Exercises
- Modelling
  - Choosing a Method
- ANOVA
  - Data
  - Sums of Squares
- Linear Model
- Generalized Linear Models
  - Overview and Data
  - Binary Response Variables
  - Binomial/Logistic Regression
  - The famous O-Ring example
  - Ancova with a Binary Response Variable
- GLMs and Count Data
- Count Data on Proportions

Proportion Data

- For comparisons of one binomial proportion with a constant, use `binom.test()`
- For comparison of two samples of proportion data, use `prop.test()`
- The use of GLMs on proportion data is for complex models

GLMs & Proportion Data

- uses also logit as link function and binomial error distribution
- if there is overdispersion use quasibinomial to compensate
- fitted values are counts
- we have seen one example so far: in the challenger example we have already used the responds variable in form of a proportion

GLMs & Proportion Data

- we use an example concerning sex ratios in insects as response and
- population density as explanatory variable
- so load the data and fit the model

```
> numbers <-read.table("sexratio.txt",header=T)
> head(numbers)
  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
> m <- glm(cbind(males,females)~density,
+                  family=binomial,data=numbers)
```

GLMs & Proportion Data

```
> summary(m)
Call:
glm(formula = cbind(males, females) ~ density, family = binomial,
    data = numbers)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4619  -1.2760  -0.9911   0.5742   1.8795

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0807368  0.1550376   0.521    0.603
density     0.0035101  0.0005116   6.862 6.81e-12 ***

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 71.159  on 7  degrees of freedom
Residual deviance: 22.091  on 6  degrees of freedom
AIC: 54.618
```

GLMs & Proportion Data

- the residual deviance is larger than the residual degrees of freedom
- because it is something like a growth process we try a log transformation (before using quasibinomial family)

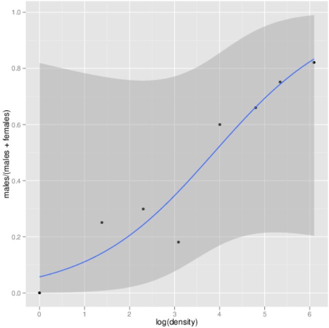
```
> m <- glm(cbind(males,females)~log(density),
+                           family=binomial,data=numbers)
> summary(m)
Call:
glm(formula = cbind(males, females) ~ log(density),
                   family = binomial, data = numbers)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9697  -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 ***
log(density)  0.69410    0.09056   7.665 1.80e-14 ***

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

GLMs & Proportion Data

- the transformation caused a welcome decrease in the residual deviance
- we conclude that the proportion of animals that are males increases significantly with increasing density, and
- that the logistic model is linearized by logarithmic transformation of the explanatory variable

GLMs & Proportion Data

```
ggplot(numbers, aes(x=log(density),y=males/(males+female
  geom_point() +
  geom_smooth(method=glm,family=binomial)
```


males/(males + females)
log(density)