An Introduction to R

3.1 A few additional statistical tools

Dan Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide
ua.edu.au/ccs/people/dan
DSTO R Workshop, 29-Apr-2015

The general linear model

GLM reminder

- Linear models:
  - Assume terms in the model combine additively
  - Assume residuals are normally distributed
  - Assume a linear relationship between the regression and the outcome variable

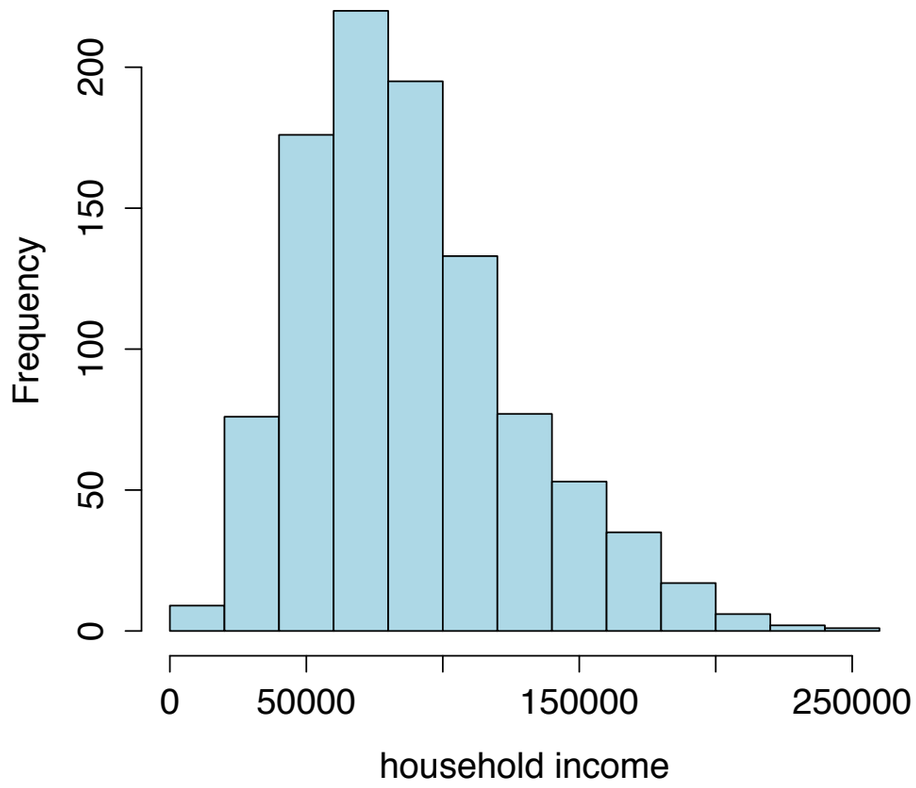
GLM reminder

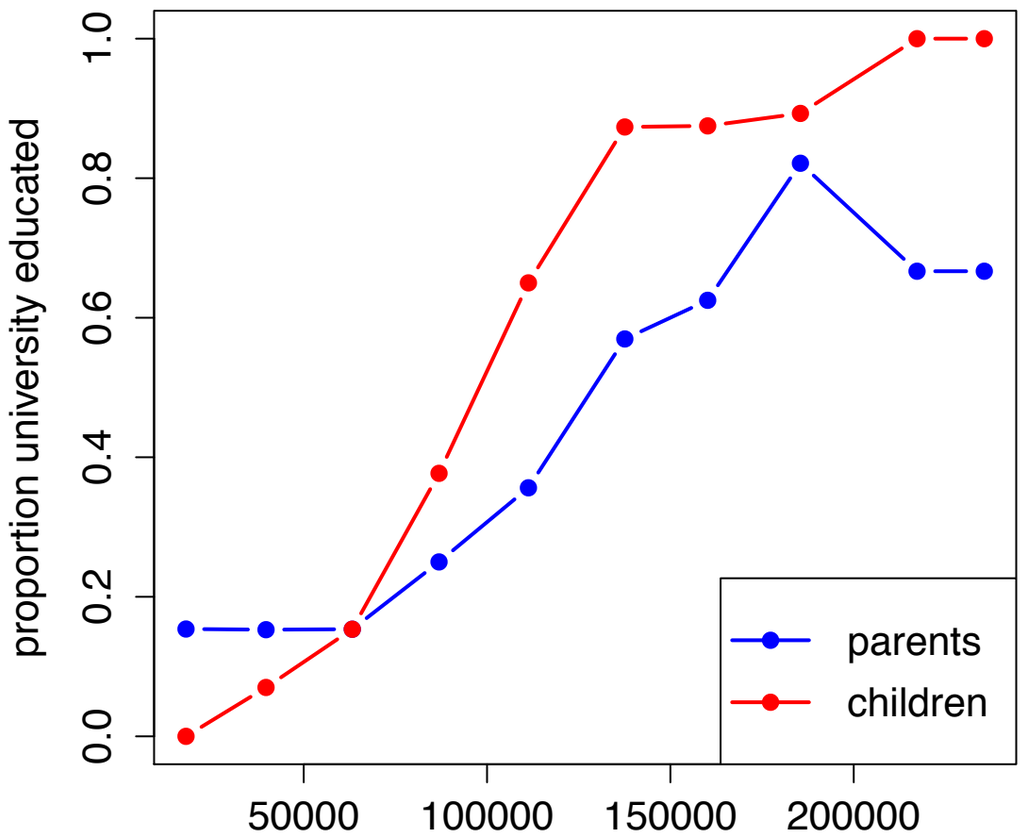
- Linear models:
  - Assume terms in the model combine additively
  - Assume residuals are normally distributed
  - Assume a linear relationship between the regression and the outcome variable

- General linear model:
  - Removes assumptions #2 and #3
  - Lots of things fall in the GLM: logistic regression, Poisson regression, probit regression, etc
  - I'll show a logistic regression example...

Data

```
> uni
  household parent child
1     40693      0     0
2    141173      0     1
3     53572      0     0
4     91917      0     1
5    151527      1     1
6     52714      0     0
```



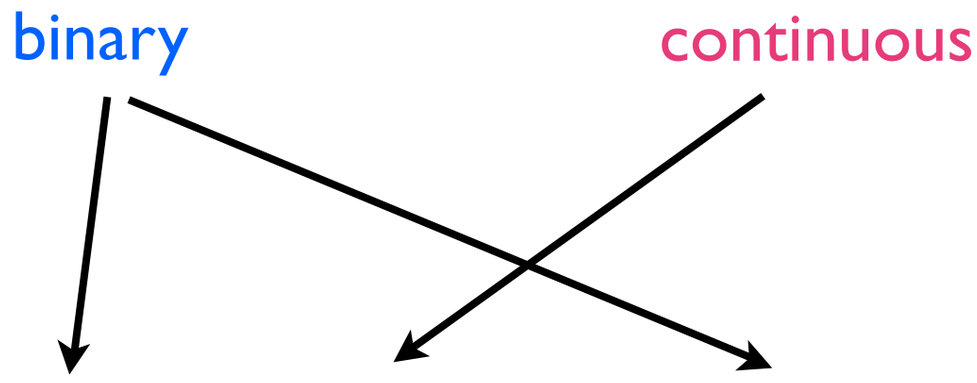



```
       parent
child    0   1
    0  533  73
    1  178 216
```

Logistic regression

```
mod <- glm(
          formula = child ~ household + parent,
          family = binomial( link=logit ),
          data = uni
        )
```

Logistic regression




```
mod <- glm(
        formula = child ~ household + parent,
        family = binomial( link=logit ),
        data = uni
      )
```

Logistic regression

```
mod <- glm(
         formula = child ~ household + parent,
         family = binomial( link=logit ),
         data = uni
       )
```

outcomes are "success or failure", so use binomial distribution rather than normal distribution

relationship between the regression model and the probability of a "success" is logistic, not linear

Logistic regression

```
mod <- glm(
         formula = child ~ household + parent,
         family = binomial( link=logit ),
         data = uni
)
```

data frame “uni” contains the variables

```
> summary(mod)

Call:
glm(formula = child ~ household + parent, family = binomial(link =
logit), data = uni)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1320  -0.6472  -0.3697   0.6251   2.5932

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.497e+00  2.910e-01 -15.454   <2e-16 ***
household    3.936e-05  3.011e-06  13.071   <2e-16 ***
parent       1.836e+00  1.895e-01   9.687   <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: 1341.0  on 999  degrees of freedom
Residual deviance:  863.5  on 997  degrees of freedom
AIC: 869.5

Number of Fisher Scoring iterations: 5
```

```
> anova( mod, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: child

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        999    1341.01
household  1   376.08        998     964.93 < 2.2e-16 ***
parent     1   101.43        997     863.50 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

Exploratory factor analysis

Some survey data...

```
> questionnaire

    A1 A2 A3 A4 A5 B1 B2 B3 B4 B5
1    4  4  5  5  4  4  3  2  4  3
2    4  5  5  5  5  2  3  3  4  3
3    4  4  4  4  5  5  3  4  3  2
4    4  4  4  3  2  5  4  4  3  2
5    5  4  7  5  6  6  6  6  6  3
6    3  2  2  3  2  2  3  3  4  4
7    4  4  4  4  6  5  5  4  4  5
8    6  5  7  5  6  6  2  2  4  4
9    5  6  6  4  5  5  1  1  4  5
10   2  3  3  2  3  4  5  4  4  5

97   4  4  3  3  2  3  1  1  1  2
98   4  5  4  4  5  3  4  5  6  5
99   5  3  3  4  2  3  3  4  4  4
100  5  5  5  5  6  3  4  5  4  4
```

Unrealistically neat correlations...

```
> correlate( questionnaire, test=TRUE )

CORRELATIONS
============
- correlation type:  pearson
- correlations shown only when both variables are numeric

      A1        A2        A3        A4        A5        B1        B2        B3        B4        B5
A1     .  0.528***  0.638***  0.608***  0.510***  0.096    -0.116    -0.040    -0.099    -0.146
A2  0.528***     .  0.627***  0.520***  0.599*** -0.062    -0.014    -0.032    -0.030    -0.104
A3  0.638***  0.627***     .  0.651***  0.564***  0.072     0.006    -0.169    -0.167    -0.116
A4  0.608***  0.520***  0.651***     .  0.671*** -0.029    -0.070     0.001    -0.063    -0.011
A5  0.510***  0.599***  0.564***  0.671***     .  0.196    -0.045     0.219     0.001     0.126
B1  0.096    -0.062     0.072    -0.029     0.196     .     0.287     0.387**   0.183     0.252
B2 -0.116    -0.014     0.006    -0.070    -0.045  0.287     .      0.387**   0.353*    0.097
B3 -0.040    -0.032    -0.169     0.001     0.219   0.387**  0.387**     .     0.245     0.239
B4 -0.099    -0.030    -0.167    -0.063     0.001  0.183     0.353*    0.245     .      0.227
B5 -0.146    -0.104    -0.116    -0.011     0.126  0.252     0.097     0.239     0.227     .

---
Signif. codes:  . = p < 1, * = p<.05, ** = p<.01, *** = p<.001
```

obvious factor is obvious!

not so obvious factor is not so obvious!

factanal()

- R has multiple tools for factor analysis
- factanal() is the basic one
  - Defaults to varimax rotation
  - Estimates model using MLE

Try a two factor model...

```
> factanal( questionnaire, factors=2 )

Call:
factanal(x = questionnaire, factors = 2)

Uniquenesses:
   A1    A2    A3    A4    A5    B1    B2    B3    B4    B5
0.455 0.471 0.313 0.351 0.310 0.734 0.794 0.412 0.836 0.848

Loadings:
   Factor1 Factor2
A1  0.731  -0.103
A2  0.724
A3  0.808  -0.183
A4  0.805
A5  0.794   0.245
B1          0.508
B2          0.451
B3          0.767
B4          0.395
B5          0.388
```

Not quite identical to how the data were generated, but close enough I guess

Try a two factor model...

```
...

                Factor1 Factor2
SS loadings       3.011   1.465
Proportion Var    0.301   0.146
Cumulative Var    0.301   0.448

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 54.17 on 26 degrees of freedom.
The p-value is 0.000967
```

Other comments

- R has multiple tools for factor analysis
  - psych package and GPArotation offers other factor rotation methods
  - nFactors allows parallel roots analysis etc
  - probably others too... I'm not much of a factor analysis guy!

Cronbach's alpha

Suppose A and B were actually pre-existing scales with 5 questions each...

- We want to validate the scales, not run exploratory factor analysis
- Typical method:
  - Compute Cronbach's alpha for each scale
  - psych package contains the alpha() function for this
  - Similar but not identical to SPSS method: it automatically reverses item key direction for those items that correlate negatively (you can manually specify keying direction though)

Cronbach's alpha:

```
> library( psych )
> alpha( questionnaire[,1:5] )

Reliability analysis
Call: alpha(x = questionnaire[, 1:5])

  raw_alpha std.alpha G6(smc) average_r mean  sd
       0.88      0.88    0.86      0.59  4.1 1.1

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
A1      0.86      0.86    0.83      0.61
A2      0.86      0.86    0.83      0.61
A3      0.84      0.84    0.81      0.57
A4      0.85      0.85    0.81      0.58
A5      0.85      0.85    0.82      0.60
```

Cronbach's alpha:

```
...

 Item statistics
     n    r r.cor r.drop mean  sd
A1 100 0.80  0.73   0.68  4.0 1.3
A2 100 0.80  0.73   0.68  4.0 1.4
A3 100 0.85  0.80   0.75  4.1 1.4
A4 100 0.84  0.80   0.74  4.2 1.2
A5 100 0.81  0.76   0.70  4.0 1.4

Non missing response frequency for each item
      1    2    3    4    5    6    7 miss
A1 0.03 0.08 0.21 0.34 0.22 0.10 0.02    0
A2 0.03 0.10 0.21 0.35 0.20 0.05 0.06    0
A3 0.02 0.11 0.20 0.27 0.25 0.09 0.06    0
A4 0.01 0.05 0.27 0.28 0.26 0.09 0.04    0
A5 0.02 0.14 0.19 0.31 0.19 0.13 0.02    0
```

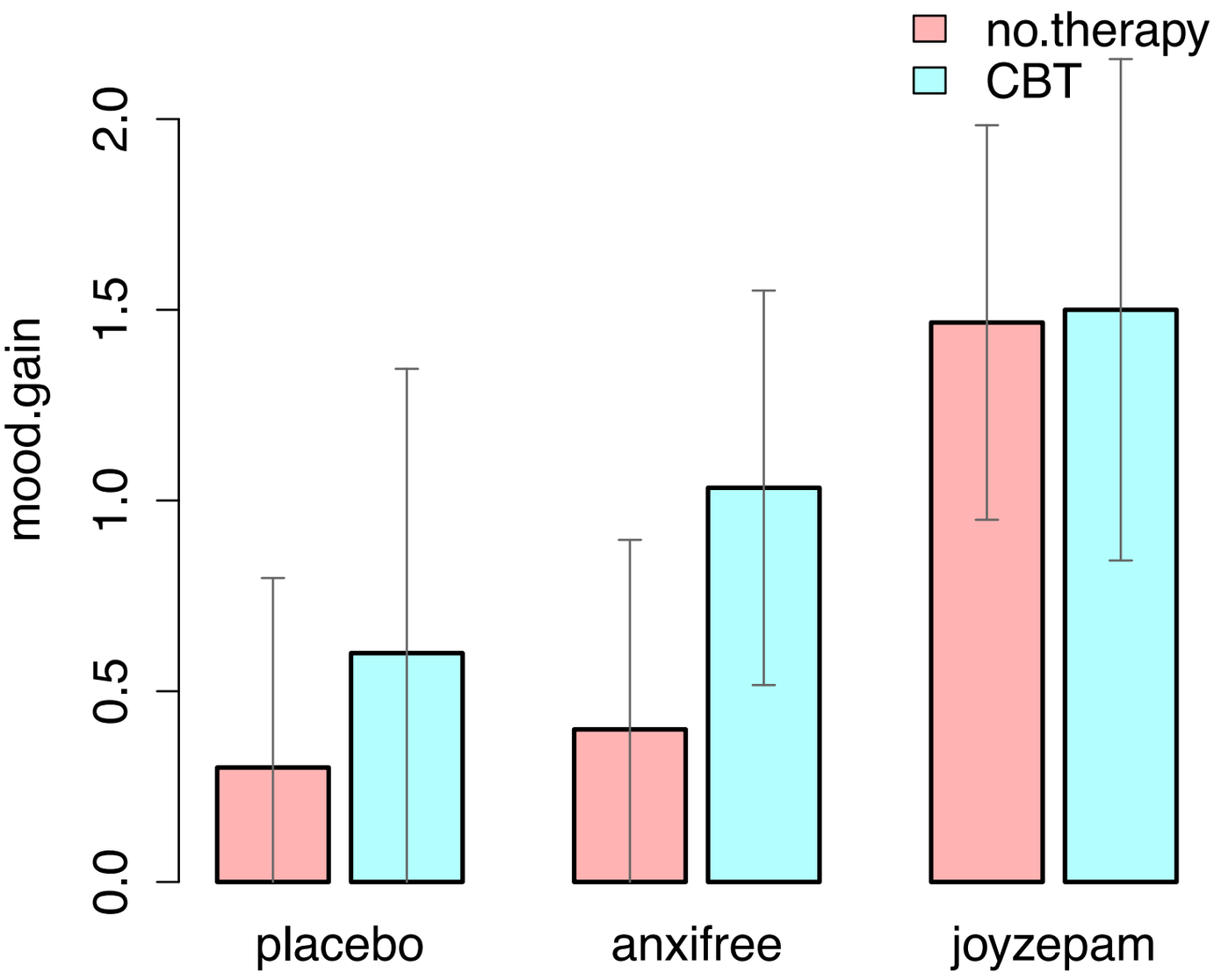
Bayesian methods

New data set

```
> clin.trial

      drug    therapy mood.gain
1   placebo no.therapy       0.5
2   placebo no.therapy       0.3
3   placebo no.therapy       0.1
4  anxifree no.therapy       0.6
5  anxifree no.therapy       0.4
6  anxifree no.therapy       0.2
7  joyzepam no.therapy       1.4
8  joyzepam no.therapy       1.7
9  joyzepam no.therapy       1.3
10  placebo         CBT       0.6
11  placebo         CBT       0.9
12  placebo         CBT       0.3
13 anxifree         CBT       1.1
14 anxifree         CBT       0.8
15 anxifree         CBT       1.2
16 joyzepam         CBT       1.8
17 joyzepam         CBT       1.3
18 joyzepam         CBT       1.4
```

What the data look like...




```
bars(
  mood.gain ~ drug + therapy,
  clin.trial
)
```

Orthodox null hypothesis tests

```
> library(car)
> Anova( aov(  mood.gain ~ drug * therapy, clin.trial  ))

Anova Table (Type II tests)

Response: mood.gain
               Sum Sq Df F value    Pr(>F)
drug           3.4533  2 31.7143 1.621e-05 ***
therapy        0.4672  1  8.5816   0.01262 *
drug:therapy   0.2711  2  2.4898   0.12460
Residuals      0.6533 12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significant effects of “drug” and “therapy”, no evidence for interaction

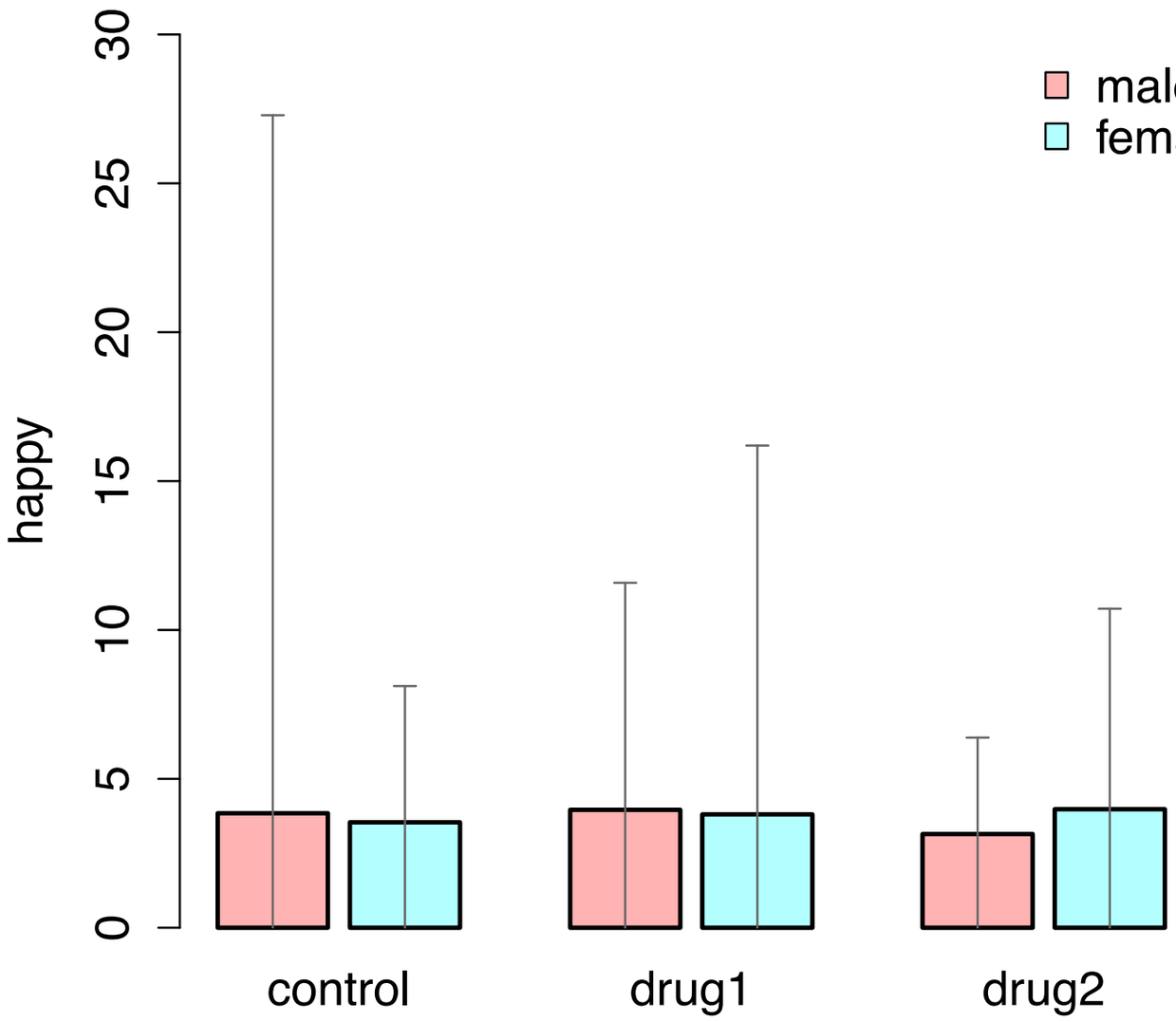
Bayesian hypothesis tests

```
> library( BayesFactor )
> anovaBF( mood.gain ~ drug * therapy, clin.trial )
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
Bayes factor analysis
--------------
[1] drug                          : 245.9026  ±0%
[2] therapy                       : 0.7316007 ±0%
[3] drug + therapy                : 692.0251  ±0.76%
[4] drug + therapy + drug:therapy : 683.1502  ±1.55%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

Best model contains effects of “drug” and “therapy” and no interaction term

A second example



Intuitively, I'm pretty confidence that there's no effect here.

I would like to quantify the evidence in favour of the null hypothesis...

```
bars(
  happy ~ treatment + gender,
  expt
)
```

Orthodox analysis?

```
> Anova( aov( happy ~ treatment * gender , expt ))

Anova Table (Type II tests)

Response: happy
                 Sum Sq Df F value Pr(>F)
treatment        0.2072  2  0.0599 0.9424
gender           0.0469  1  0.0271 0.8747
treatment:gender 0.7674  2  0.2218 0.8074
Residuals       10.3803  6
```

Intuitively, we want to be able to claim that these p-values represent strong evidence for the null, but we're not allowed to: NHST doesn't work that way.

Bayesian ANOVA is better: it tells you the amount of evidence for the null

```
> anovaBF( happy ~ treatment + gender , expt )
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
Bayes factor analysis
--------------
[1] gender                                : 0.4734681  ±0.01%
[2] treatment                             : 0.3604037  ±0.03%
[3] gender + treatment                    : 0.1652649  ±1.28%
[4] gender + treatment + gender:treatment : 0.09086368 ±1.31%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

Best performing non-null model is
"gender only" and that still has evidence
of about 2:1 in favour of a null model

End of this section