UCLA Department of Statistics
Statistical Consulting Center

Advanced Regression in R

Tiffany Himmel
tiffany@stat.ucla.edu

August 19th, 2010



Outline

1. Generalized Linear Models
2. Mixed Effects Models
3. Resources

Objective

Give an overview of different generalized linear models as well mixed models. On the way, we will also discuss

- Statistical significance
- Diagnostics
- Interpretation in context

R will be used, however most statistical softwares contain comparable packages (STATA, SAS, SPS)

1. Generalized Linear Models
   - Poisson Regression
   - Logistic Regression
   - Multinomial Logistic
2. Mixed Effects Models
3. Resources

Generalized Linear Models

Objective

To model the relationship between our predictors $X_1, X_2, \ldots, X_p$ and a binary, count or other non-normal variable Y which cannot be modeled with standard regression.

Possible Applications

- Medical Data
- Internet Traffic
- Survival Data

Poisson Regression

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

- Outcome is in the form of counts rather than a continuos numeric variable
- Explore the relationship between two categorical variables

Treatment	Outcome 1	Outcome 2	Outcome 3
1	18	17	15
2	20	10	20
3	25	13	12

Poisson Regression

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

```
counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
print(d.AD <-data.frame(treatment, outcome,
    counts))
```

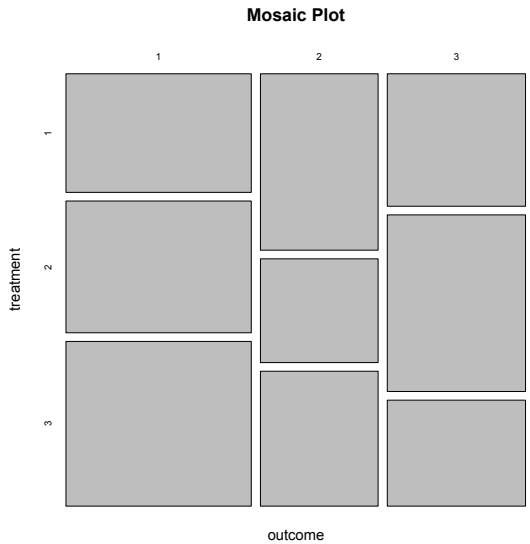
```
  treatment outcome counts
1         1       1     18
2         1       2     17
3         1       3     15
4         2       1     20
5         2       2     10
6         2       3     20
7         3       1     25
8         3       2     13
9         3       3     12
```

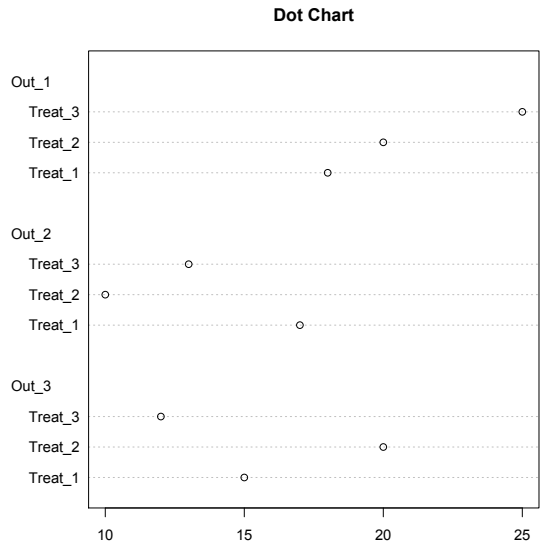


Exploratory Plots

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

```
1 contin.table<-xtabs(counts~treatment+outcome)
2 mosaicplot(contin.table,main="Mosaic Plot")
3 dimnames(contin.table)$outcome<-c("Out_1","Out
    _2","Out_3")
4 dimnames(contin.table)$treatment<-c("Treat_1",
    "Treat_2","Treat_3")
5 dotchart(contin.table,main="Dot Chart")
```





Fitting a GLM

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

```
1 glm.D93 <- glm(counts ~ outcome + treatment,
      family=poisson())
2 anova(glm.D93)
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: counts

Terms added sequentially (first to last)

          Df  Deviance Resid. Df Resid. Dev
NULL                          8    10.5814
outcome    2    5.4523         6     5.1291
treatment  2 8.882e-15         4     5.1291
```

Fitting a GLM II

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

```
1 summary(glm.D93)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.045e+00  1.709e-01  17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01  -2.247   0.0246 *
outcome3    -2.930e-01  1.927e-01  -1.520   0.1285
treatment2   1.338e-15  2.000e-01 6.69e-15  1.0000
treatment3   1.421e-15  2.000e-01 7.11e-15  1.0000
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom
AIC: 56.761
```



Chi-Squared Test

Example: Randomized Controlled Trial (Taken from Dobson, 1990)

```
1 anova(glm.D93, test = "Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: counts

Terms added sequentially (first to last)

          Df  Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          8    10.5814
outcome    2    5.4523         6     5.1291    0.0655
treatment  2 8.882e-15         4     5.1291    1.0000
```

Logistic Regression

- Like Poisson, outcomes are not continuous
- Outcome is in the form of binary variables
  - yes/no, dead/alive, male/female
  - Makes estimates based on % yes/no
  - x-covariate continuous (usually)
- Very popular in medical research

Logistic Regression

Example: Insecticide Concentration (Taken from Faraway, 2006)

```
1 library(faraway)
2 data(bliss)
3 bliss
```

```
  dead alive conc
1    2    28    0
2    8    22    1
3   15    15    2
4   23     7    3
5   27     3    4
```

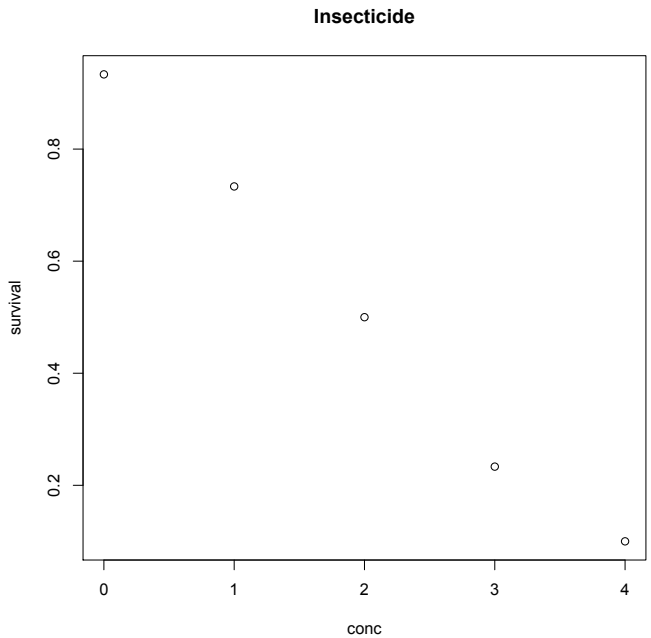


Exploratory Plots

Example: Insecticide Concentration (Taken from Faraway, 2006)

Thinking percentages...

```
1 attach(bliss)
2 survival<-(alive/(alive+dead))
3 bliss<-data.frame(bliss,survival)
4 plot(survival~conc,main="Insecticide")
```



Fitting a Logistic Model

Example: Insecticide Concentration (Taken from Faraway, 2006)

```
1 modl<-glm(cbind(dead,alive)~conc,family=
    binomial, data=bliss)
2 summary(modl)
```

```
Call:
glm(formula = cbind(dead, alive) ~ conc, family = binomial, data = bliss)

Deviance Residuals:
      1       2       3       4       5
-0.4510  0.3597  0.0000  0.0643 -0.2045

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3238     0.4179  -5.561 2.69e-08 ***
conc          1.1619     0.1814   6.405 1.51e-10 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64.76327  on 4  degrees of freedom
Residual deviance:  0.37875  on 3  degrees of freedom
AIC: 20.854

Number of Fisher Scoring iterations: 4
```



Other Link Functions

Example: Insecticide Concentration (Taken from Faraway, 2006)

```
1 modp<-glm(cbind(dead,alive)~conc,family=
    binomial(link=probit),data=bliss)
2 modc<-glm(cbind(dead,alive)~conc,family=
    binomial(link=cloglog),data=bliss)
```

Other Link Functions

Example: Insecticide Concentration (Taken from Faraway, 2006)

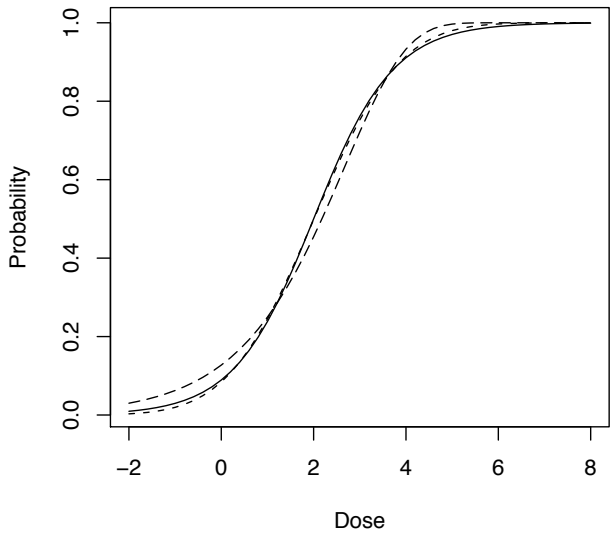
```
1 cbind(fitted(modl),fitted(modp),fitted(modc))
```

```
        [,1]       [,2]      [,3]
1 0.08917177 0.08424186 0.1272700
2 0.23832314 0.24487335 0.2496909
3 0.50000000 0.49827210 0.4545910
4 0.76167686 0.75239612 0.7217655
5 0.91082823 0.91441122 0.9327715
```

Other Link Functions

Example: Insecticide Concentration (Taken from Faraway, 2006)

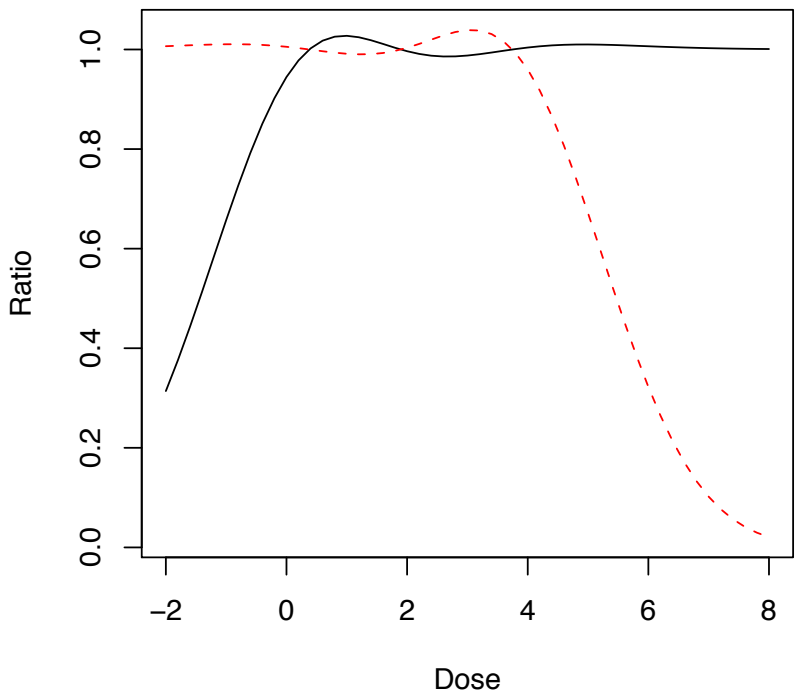
```
x<-seq(-2,8,.2)
pl<-ilogit(modl$coef[1]+modl$coef[2]*x)
pp<-pnorm(modp$coef[1]+modp$coef[2]*x)
pc<-1-exp(-exp((modc$coef[1]+modc$coef[2]*x)))
plot(x,pl,type="l",ylab="Probability",xlab="
    Dose")
lines(x,pp,lty=2)
lines(x,pc,lty=5)
```



But choose carefully!

Example: Insecticide Concentration (Taken from Faraway, 2006)

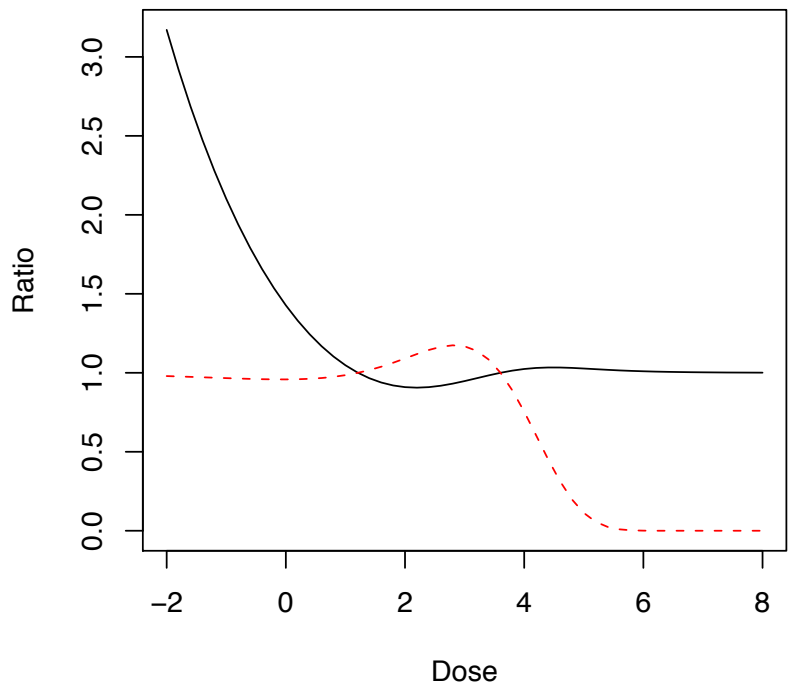
```
1  matplot(x,cbind(pp/pl,(1-pp)/(1-pl)),type="l",
       xlab="Dose",ylab="Ratio")
```



But choose carefully!

Example: Insecticide Concentration (Taken from Faraway, 2006)

```
1 matplot(x,cbind(pc/pl,(1-pc)/(1-pl)),type="l",
      xlab="Dose",ylab="Ratio")
```



Multinomial Logistic

- Like Logistic+Poisson, outcomes are not continuous
- Outcome is in the form of a categorical variable with 3 or more options
  - black/white/other, married/unmarried/divorced
  - Our example, politics: Democrat/Republican/Independent

Multinomial Logistic

Example: American National Election Study (Taken from Faraway, 2006)

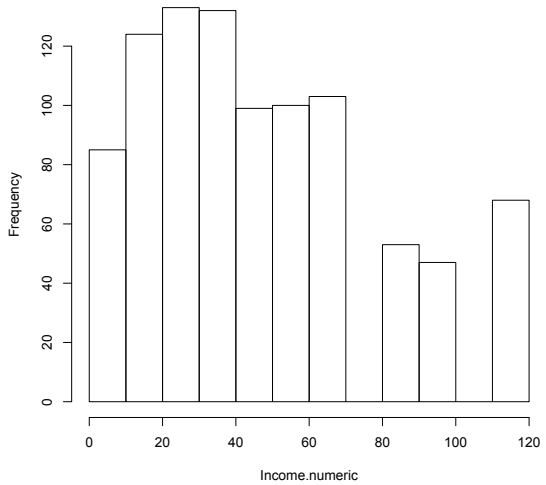
```
1 data(nes96)
2 head(nes96)
```

```
  popul TVnews selfLR ClinLR DoleLR     PID age  educ   income    vote
1     0      7 extCon extLib    Con  strRep  36    HS $3Kminus    Dole
2   190      1 sliLib sliLib sliCon weakDem  20  Coll $3Kminus Clinton
3    31      7    Lib    Lib    Con weakDem  24 BAdeg $3Kminus Clinton
4    83      4 sliLib    Mod sliCon weakDem  28 BAdeg $3Kminus Clinton
5   640      7 sliCon    Con    Mod  strDem  68 BAdeg $3Kminus Clinton
6   110      3 sliLib    Mod    Con weakDem  21  Coll $3Kminus Clinton
```

Exploratory Plots

Example: American National Election Study (Taken from Faraway, 2006)

```
1 PartyAffil<-nes96$PID
2 levels(PartyAffil)<-c("Dem","Dem","Ind","Ind",
    "Ind","Rep","Rep")
3 income.categories<-c(1.5, 4, 6, 8, 9., 10.5,
    11.5, 12.5, 13.5, 14.5, 16, 18.5, 21, 23.5,
     27.5, 32.5, 37.5, 42.5, 47.5, 55, 67.5,
    82.5, 97.5, 115)
4 Income.numeric<-income.categories[unclass(
    nes96$income)]
5 hist(Income.numeric,main="")
```



Useful Tables

Example: American National Election Study (Taken from Faraway, 2006)

```
1 Income.Split<-cut(Income.numeric,7)
2 PartyProp.by.Income<-prop.table(table(Income.
    Split,PartyAffil),1)
```

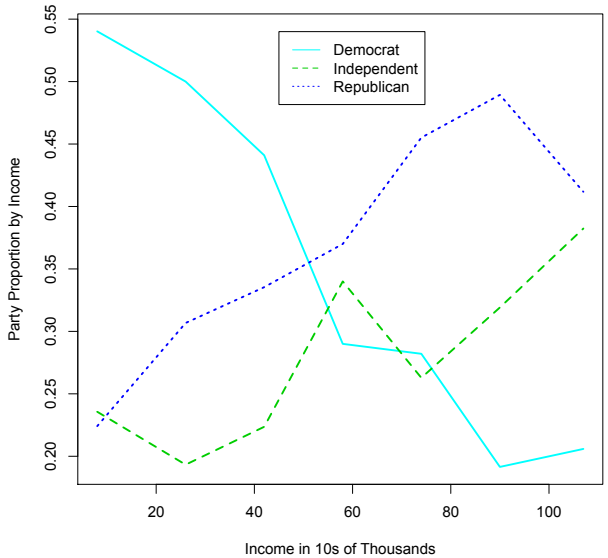
```
                PartyAffil
Income.Split          Dem       Ind       Rep
  (1.39,17.6] 0.5402299 0.2356322 0.2241379
  (17.6,33.9] 0.5000000 0.1932773 0.3067227
  (33.9,50.1] 0.4409938 0.2236025 0.3354037
  (50.1,66.4] 0.2900000 0.3400000 0.3700000
  (66.4,82.6] 0.2820513 0.2628205 0.4551282
  (82.6,98.9] 0.1914894 0.3191489 0.4893617
  (98.9,115]  0.2058824 0.3823529 0.4117647
```



Exploratory Plots II

Example: American National Election Study (Taken from Faraway, 2006)

```
1 Income.Split.Mids<-c(8,26,42,58,74,90,107)
2 matplot(Income.Split.Mids,PartyProp.by.Income,
      lty=c(1,2,3),col=c(5,3,4),lwd=2,type="l",
      xlab="Income in 10s of Thousands",ylab="
      Party Proportion by Income")
3 legend(45,.54,c("Democrat","Independent","
      Republican"),col=c(5,3,4),lty=c(1,2,3))
```



Fitting a Multinomial Logistic Model

Example: American National Election Study (Taken from Faraway, 2006)

```
1 library(nnet)
2 multinomial.model<-multinom(PartyAffil~Income.
    numeric)
3 summary(multinomial.model)
```

```
Coefficients:
    (Intercept) Income.numeric
Ind   -1.173592     0.01606080
Rep   -0.949674     0.01765140

Std. Errors:
    (Intercept) Income.numeric
Ind   0.1535317    0.002848357
Rep   0.1416261    0.002651135

Residual Deviance: 1985.485
AIC: 1993.485
```



Making Predictions

Example: American National Election Study (Taken from Faraway, 2006)

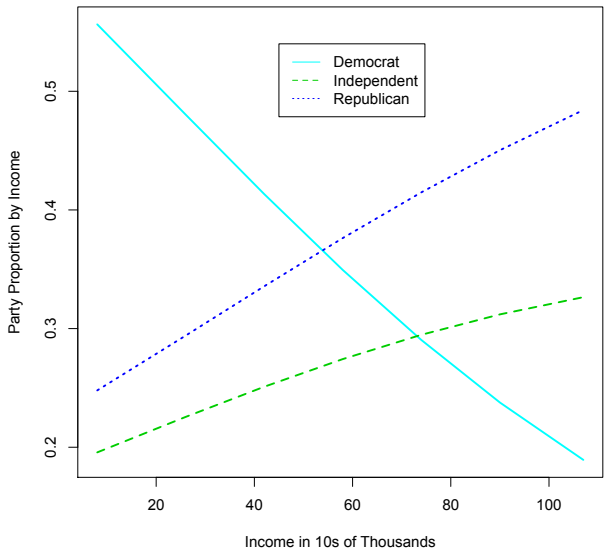
```
predict(multinomial.model,data.frame(Income.
    numeric=Income.Split.Mids))
Party.Prob.Predictions<-predict(multinomial.
    model,data.frame(Income.numeric=Income.
    Split.Mids),type="probs")
```

```
        Dem       Ind       Rep
1 0.5564220 0.1956684 0.2479096
2 0.4803742 0.2255528 0.2940730
3 0.4133828 0.2509706 0.3356467
4 0.3494149 0.2742923 0.3762928
5 0.2904121 0.2947737 0.4148142
6 0.2377033 0.3119688 0.4503278
7 0.1893241 0.3264816 0.4841943
```

Visualizing Our Model Predictions

Example: American National Election Study (Taken from Faraway, 2006)

```
1 matplot(Income.Split.Mids,Party.Prob.
      Predictions,lty=c(1,2,3),col=c(5,3,4),lwd
      =2,type="l",xlab="Income in 10s of
      Thousands",ylab="Party Proportion by Income
      ")
2 legend(45,.54,c("Democrat","Independent","
      Republican"),col=c(5,3,4),lty=c(1,2,3))
```



Second Order Terms

Example: American National Election Study (Taken from Faraway, 2006)

```
multinomial.model.II<-multinom(PartyAffil
    ~Income.numeric+I(Income.numeric^2))
```

```
Call:
multinom(formula = PartyAffil ~ Income.numeric + I(Income.numeric^2))

Coefficients:
    (Intercept) Income.numeric I(Income.numeric^2)
Ind   -1.038220     0.00993433        4.752663e-05
Rep   -1.181813     0.02888725       -9.569798e-05

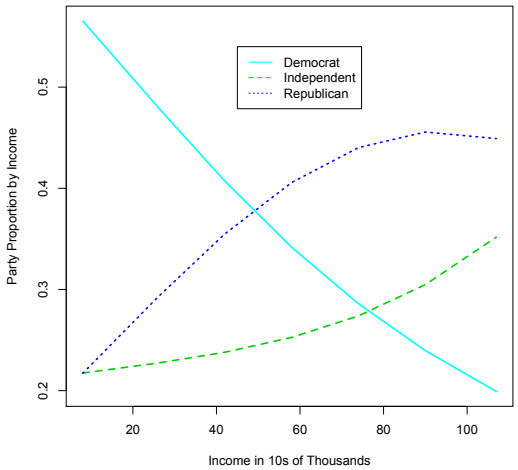
Std. Errors:
     (Intercept) Income.numeric I(Income.numeric^2)
Ind 0.0001707178    0.004581055        5.391936e-05
Rep 0.0001556944    0.004204987        5.067708e-05

Residual Deviance: 1982.148
AIC: 1994.148
```

Visualizing Our Model Predictions II

Example: American National Election Study (Taken from Faraway, 2006)

```
predict(multinomial.model.II,data.frame(Income
    .numeric=Income.Split.Mids))
Party.Prob.Predictions.II<-predict(multinomial
    .model.II, data.frame(Income.numeric=Income
    .Split.Mids), type="probs")
matplot(Income.Split.Mids,Party.Prob.
    Predictions.II,lty=c(1,2,3),col=c(5,3,4),
    lwd=2,type="l",xlab="Income in 10s of
    Thousands",ylab="Party Proportion by Income
    ")
legend(45,.54,c("Democrat","Independent","
    Republican"),col=c(5,3,4),lty=c(1,2,3))
```



Multinomial Logistic ANOVA

Example: American National Election Study (Taken from Faraway, 2006)

```
1  anova(multinomial.model,multinomial.model.II)
```

```
                  Model  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1   Income.numeric           1884   1985.485             NA       NA        NA
2   Income.numeric +         1882   1982.148 1 vs 2      2 3.336716 0.1885564
  I(Income.numeric^2)
```



1 Generalized Linear Models

2 Mixed Effects Models
- Definitions
- Model Fitting
- Assumption Validation

3 Resources

What is a Mixed Effects Model?

- Fixed effect: β is the same over all groups/patients
- Random Effect: $\nu_i \sim N(0, \sigma_\nu^2)$ where each group has their own coefficient

Often used in medical or survey data

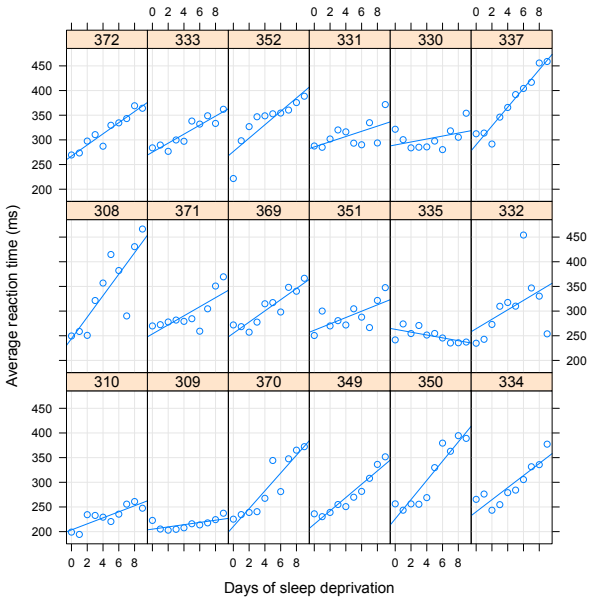
$$y_{ij} = \underbrace{\beta_0 + \ldots + \beta_l}_{\text{fixed effects}} + \underbrace{\nu_{1i} + \ldots + \nu_{mi}}_{\text{random effects}} + \underbrace{\epsilon_{ij}}_{\text{residual}}$$

subject $i = 1 \ldots n \leftarrow \#$ of subjects
observation $j = 1 \ldots n_i \leftarrow \#$ of observations for that subject

Exploratory Plots

Example: Sleep Study (Taken from Belenky et al., 2003)

```
library(lme4)
data(sleepstudy)
xyplot(Reaction ~ Days | Subject, sleepstudy,
    type = c("g","p","r"),index = function(x,y)
     coef(lm(y ~ x))[1], xlab = "Days of sleep
    deprivation", ylab = "Average reaction time
     (ms)", aspect = "xy")
```



Fitting a Mixed Effects Model

Example: Sleep Study (Taken from Belenky et al., 2003)

```
1 (fm1 <- lmer(Reaction ~ Days + (Days|Subject),
      sleepstudy))
```

```
  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.092  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```



Fitting a Mixed Effects Model II

Example: Sleep Study (Taken from Belenky et al., 2003)

```
1 (fm2 <- lmer(Reaction ~ Days + (1|Subject) +
      (0+Days|Subject), sleepstudy))
```

```
  AIC  BIC logLik deviance REMLdev
 1754 1770 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.568  25.0513
 Subject  Days         35.858   5.9882
 Residual             653.584  25.5653
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885   36.51
Days          10.467      1.559    6.71

Correlation of Fixed Effects:
     (Intr)
Days -0.184
```

Anova with Nested Models

Example: Sleep Study (Taken from Belenky et al., 2003)

```
1 anova(fm1, fm2)
```

```
    Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm2  5 1762.0 1778.0 -876.02
fm1  6 1764.0 1783.1 -875.99 0.0609      1      0.805
```

Intra-Class Correlation

Example: Sleep Study (Taken from Belenky et al., 2003)

```
1  ICC_1<-(612.092+35.072)
       /(612.092+35.072+654.941)
```

```
[1] 0.497
```

```
1  ICC_2<-(627.568+35.858)
       /(627.568+35.858+653.584)
```

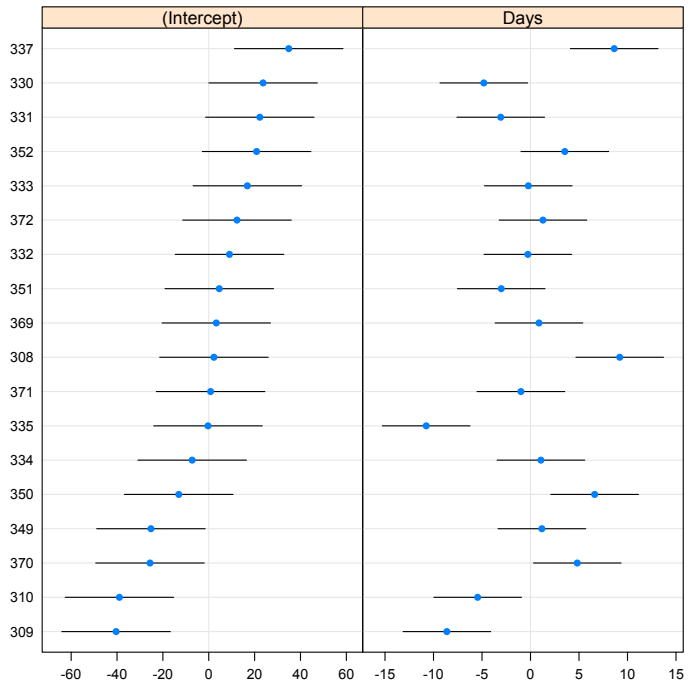
```
[1] 0.5037
```



Random Effects by Subject with Errors

Example: Sleep Study (Taken from Belenky et al., 2003)

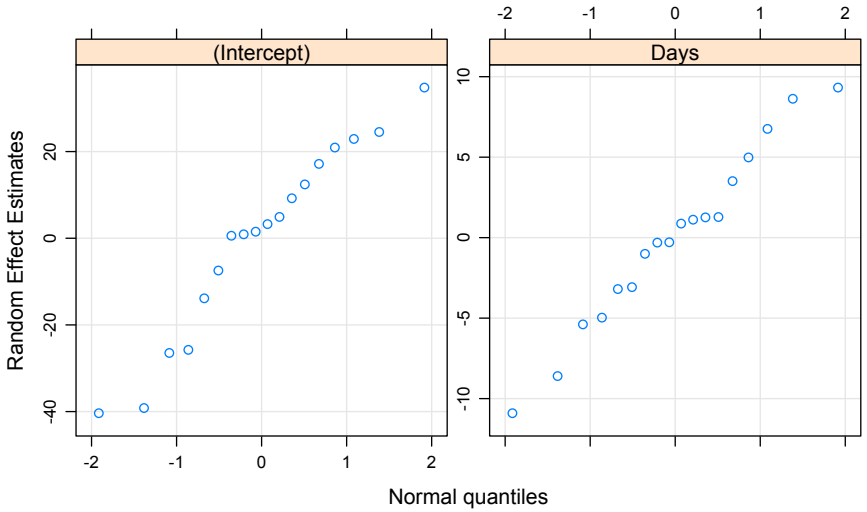
```
1 rr1 <- ranef(fm1,postVar=TRUE)
2 dotplot(rr1,scales = list(x = list(relation =
    'free')))[["Subject"]]
```



Random Effects QQ Plot

Example: Sleep Study (Taken from Belenky et al., 2003)

```
rand_eff <- ranef(fm2, drop=TRUE)[[1]]
qqmath(~ `(Intercept)` + Days,rand_eff, xlab="
    Normal quantiles", ylab="Random Effect
    Estimates",auto.key=T,outer = TRUE, scales
    = list(y = list(relation = "free")),
    layout = c(2,1), aspect = 1, type = c("g",
    "p"))
```



Simulated Data

Example: Sleep Study (Taken from Belenky et al., 2003)

```
set.seed(483)
simfe <-data.frame(t(apply(simulate(fm2, nsim
    = 100), 2, function(y) fixef(refit(fm2, y))
    )), check.names = FALSE)
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.187      8.261   36.51
Days          10.239      1.346    6.71
```



1 Generalized Linear Models

2 Mixed Effects Models

3 Resources
- Online Resources for R
- References

Online Resources for R

Download R: `http://cran.stat.ucla.edu`

Search Engine for R: `rseek.org`

R Reference Card: `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`

UCLA Statistics Information Portal: `http://info.stat.ucla.edu/grad/`

UCLA Statistical Consulting Center `http://scc.stat.ucla.edu`

References I


P. Daalgard
Introductory Statistics with R,
Statistics and Computing, Springer-Verlag, NY, 2002.


B.S. Everitt and T. Hothorn
A Handbook of Statistical Analysis using R,
Chapman & Hall/CRC, 2006.


J.J. Faraway
Practical Regression and Anova using R,
www.stat.lsa.umich.edu/~faraway/book

References II


J. Maindonald and J. Braun
Data Analysis and Graphics using R – An Example-Based Approach,
Second Edition, Cambridge University Press, 2007.


[Sheather, 2009] S.J. Sheather
A Modern Approach to Regression with R,
DOI: 10.1007/978-0-387-09608-7-3,
Springer Science + Business Media LCC 2009.