Negative Binomial Model for Count Data
Log-linear Models for Contingency Tables - Introduction

Statistics 149

Spring 2006

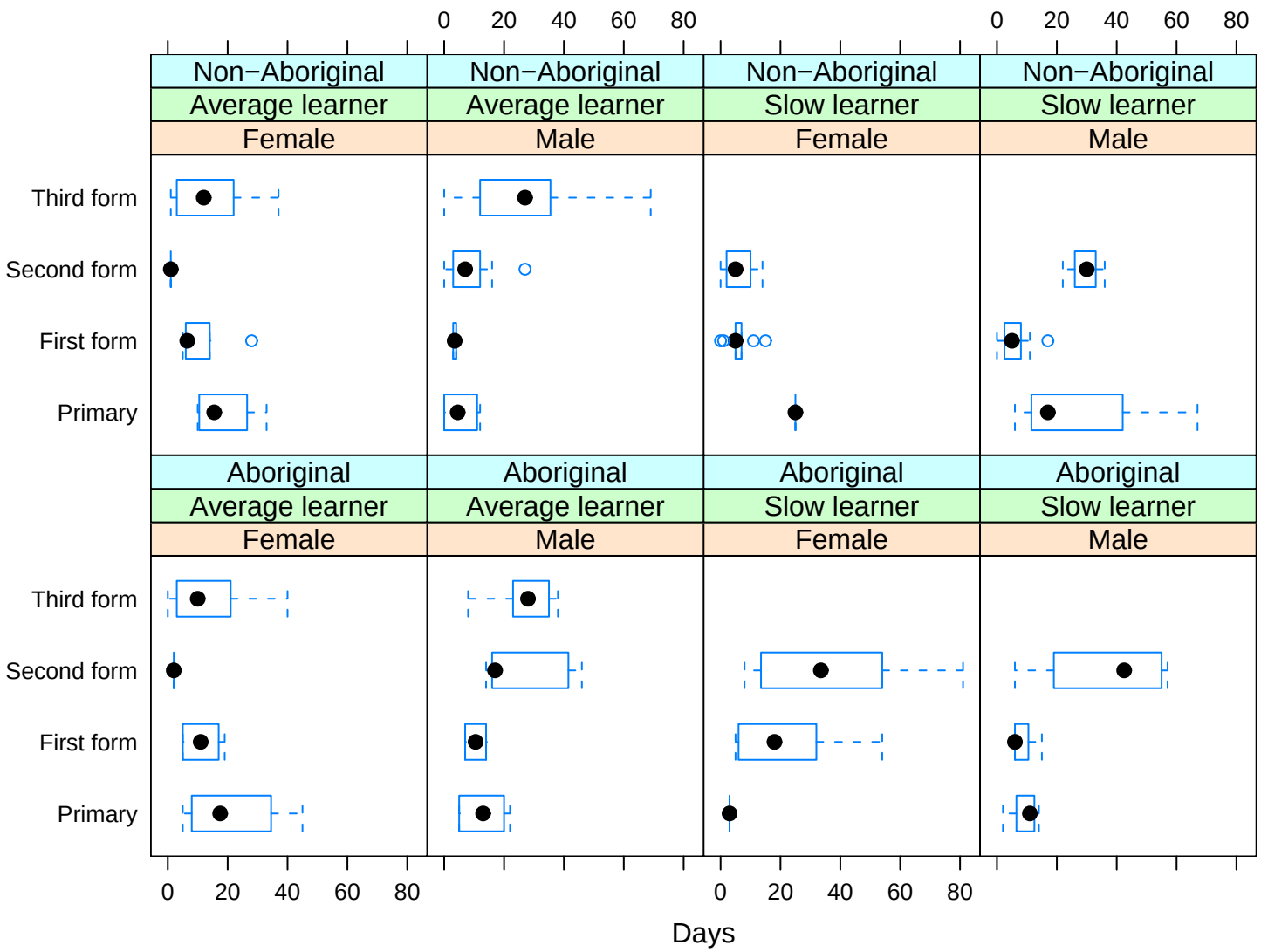
Negative Binomial Family

Example: Absenteeism from School in Rural New South Wales

The 'quine' data frame in the `MASS` package has 146 observations on 5 variables. Children from Walgett, New South Wales, Australia, were classified by

- Culture: aboriginal vs non-aboriginal
- Age: primary, first, second, or third form (like grade)
- Sex
- Learner status: average vs slow learner

For each child the number of days absent from school in a particular school year was recorded.

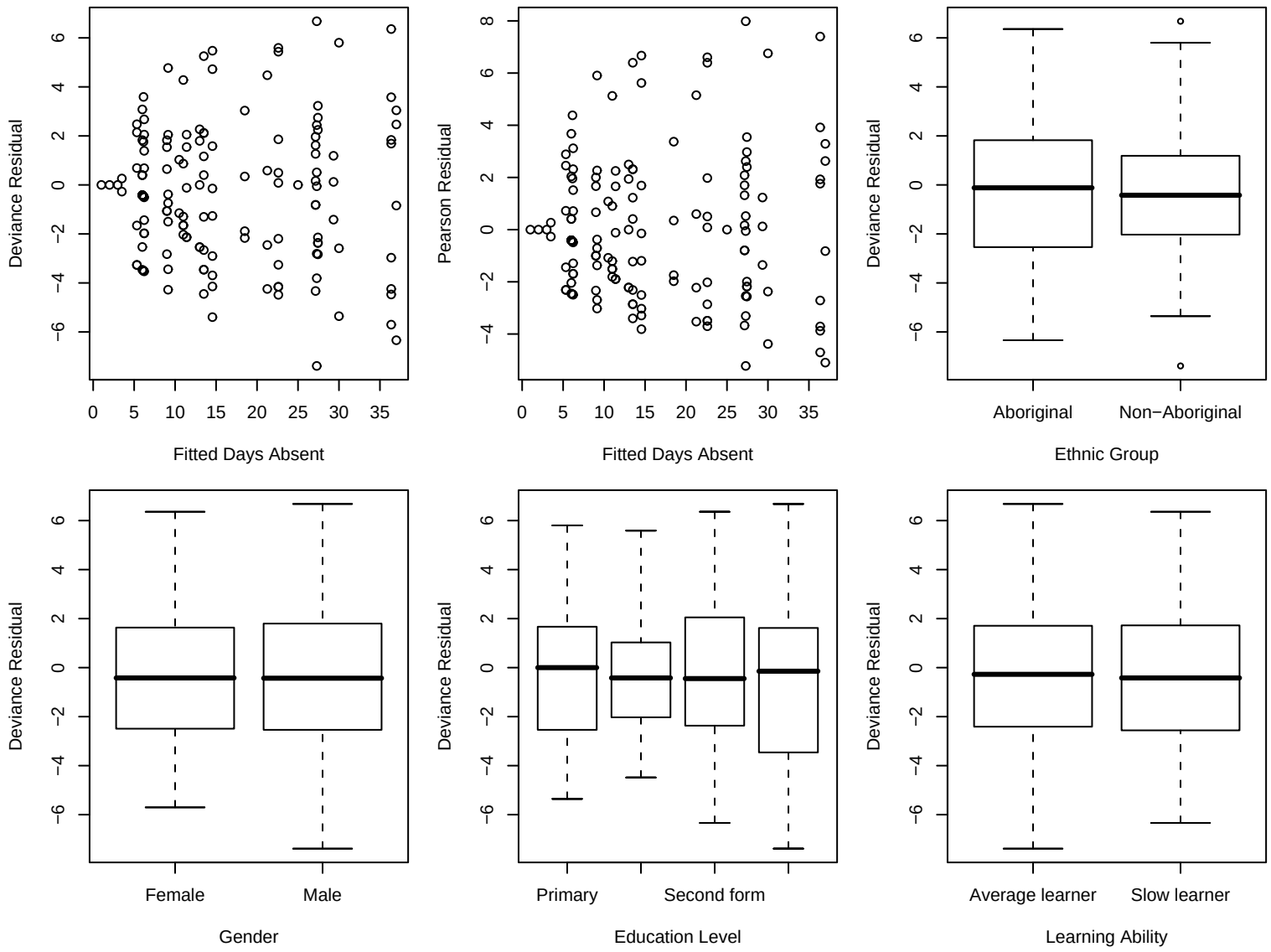

Non−Aboriginal
Average learner
Female
Non−Aboriginal
Average learner
Male
Non−Aboriginal
Slow learner
Female
Non−Aboriginal
Slow learner
Male
Aboriginal
Average learner
Female
Aboriginal
Average learner
Male
Aboriginal
Slow learner
Female
Aboriginal
Slow learner
Male
Third form
Second form
First form
Primary
0
20
40
60
80
Days

```
> summary(quine.qglm)

Call:
glm(formula = Days ~ .^4, family = quasipoisson(), data = quine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3872  -2.5129  -0.4205   1.7424   6.6783

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.0564     0.3346   9.135 2.22e-15 ***
EthN                  -0.1386     0.4904  -0.283   0.7780
SexM                  -0.4914     0.5082  -0.967   0.3356
AgeF1                 -0.6227     0.5281  -1.179   0.2406
AgeF2                 -2.3632     2.2066  -1.071   0.2864
AgeF3                 -0.3784     0.4296  -0.881   0.3802
LrnSL                 -1.9577     1.8120  -1.080   0.2822
 . . .
```

```
EthN:SexM:AgeF1:LrnSL   2.1711     2.7527   0.789   0.4319
EthN:SexM:AgeF2:LrnSL   2.1029     4.4203   0.476   0.6351
EthN:SexM:AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 9.51)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1173.9  on 118  degrees of freedom
```

So there is some suggestion of overdispersion, which is supported by the following residual plots.

Note that this is the largest model that can be fit with these 4 categorical predictors, not necessarily the best model.


Deviance Residual
Fitted Days Absent
Pearson Residual
Fitted Days Absent
Deviance Residual
Aboriginal
Non−Aboriginal
Ethnic Group
Deviance Residual
Female
Male
Gender
Deviance Residual
Primary
Second form
Education Level
Deviance Residual
Average learner
Slow learner
Learning Ability

An alternative approach to the quasi-likelihood model is to build a hierarchical model for count data along the lines of the Beta-Binomial distribution for binary data.

$$
\begin{aligned}
Y_i|E_i &\overset{ind}{\sim} Poisson(\mu_i E_i) \\
g(\mu_i) &= X_i\beta \\
E_i &\overset{iid}{\sim} Gamma(\theta, \theta) \\
E[E_i] &= 1 \\
\mathrm{Var}(E_i) &= \frac{1}{\theta}
\end{aligned}
$$

Then the marginal distribution of Y_i is negative binomial with density

$$
f(y; \theta, \mu_i) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \frac{\mu_i^y \theta^\theta}{(\mu_i + \theta)^{y+\theta}}; \qquad y = 0, 1, 2, \ldots
$$

and moments

$$E[Y_i] = E[E[Y_i|E_i]] = E[\mu_i E_i] = \mu_i$$

$$\begin{aligned}
\mathrm{Var}(Y_i) &= E[\mathrm{Var}(Y_i|E_i)] + \mathrm{Var}(E[Y_i|E_i]) \\
&= E[\mu_i E_i] + \mathrm{Var}(\mu_i E_i) \\
&= \mu_i + \mu_i^2 \mathrm{Var}(E_i) \\
&= \mu_i + \frac{\mu_i^2}{\theta}
\end{aligned}$$

In this case, the bigger θ is, the less overdispersion. Note that this model doesn't fit into the $\mathrm{Var}(Y) = \psi V(\mu)$ framework, exhibiting that other possibilities exist.

Note that this is not the parametrization often seen for the negative binomial model, which has density

$$f(y; p, \theta) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} p^{\theta} (1 - p)^{y}; \qquad y = 0, 1, 2, \ldots$$

This can be made to match by setting

$$p = \frac{\theta}{\mu + \theta}$$

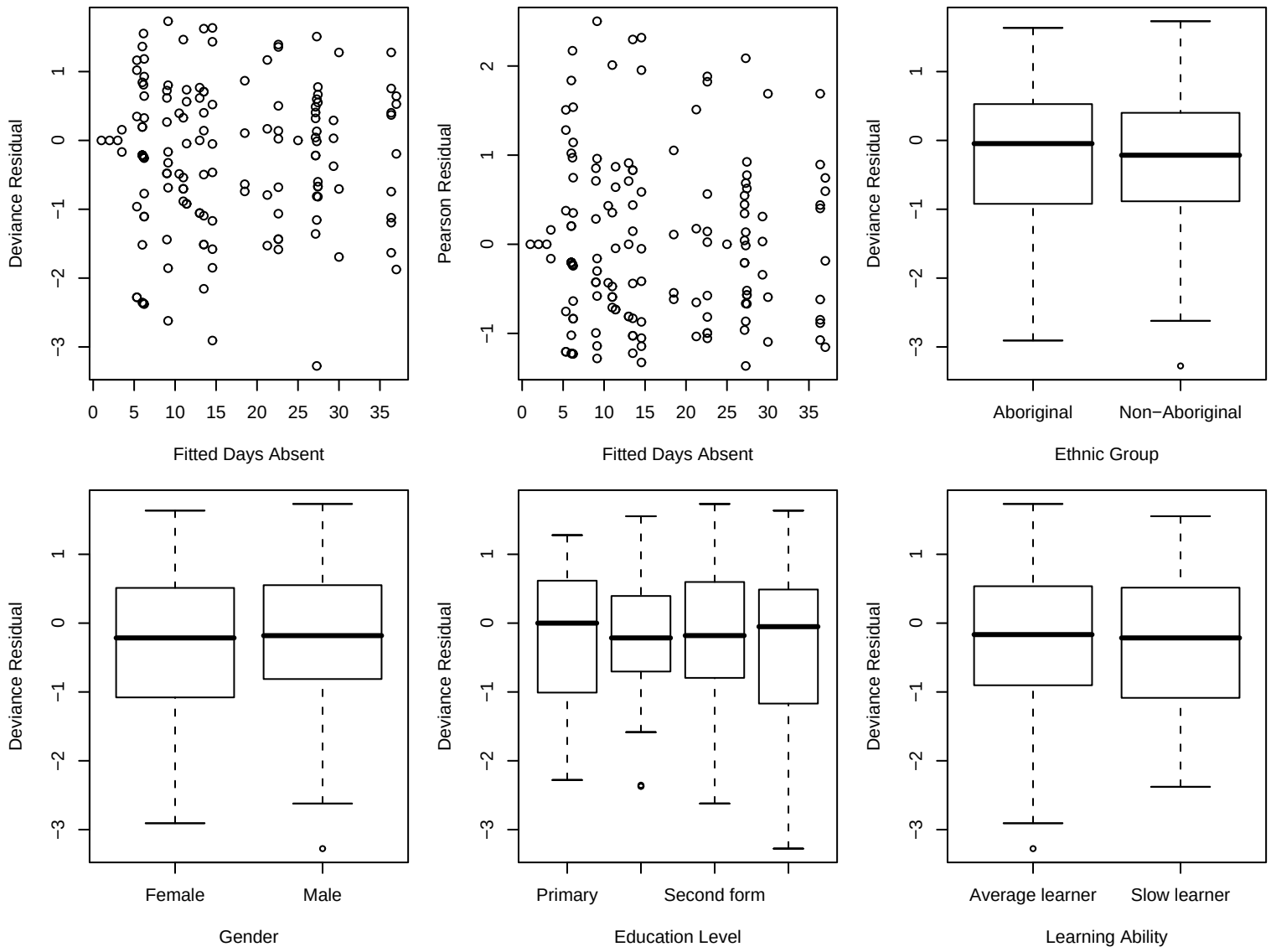
If θ is known, y is a member of the exponential family, and thus can be fit by the methods already discussed. In the `MASS` package, the additional code needed to fit these models is done with the `negative.binomial family` function. The first argument of the function is the value of `theta` and second value is the link, which takes values `log` (default), `identity`, and `sqrt`, the same link functions as for the Poisson.

An earlier analysis suggested that for the Quine example, $\theta \approx 2$. Lets fit the full interaction model in this case.

```
> summary(quine.glm)

Call:
glm(formula = Days ~ .^4, family = negative.binomial(2),
       data = quine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2766  -0.9214  -0.2050   0.5263   1.7314

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.0564     0.3807   8.027 8.32e-13 ***
EthN                  -0.1386     0.5402  -0.257  0.79797
SexM                  -0.4914     0.5170  -0.951  0.34380
AgeF1                 -0.6227     0.5192  -1.199  0.23277
```

```
AgeF2                      -2.3632     1.0977  -2.153  0.03337 *
AgeF3                      -0.3784     0.4604  -0.822  0.41280
LrnSL                      -1.9577     1.0141  -1.931  0.05593 .
 . . .
SexM:AgeF3:LrnSL               NA         NA      NA       NA
EthN:SexM:AgeF1:LrnSL       2.1711     1.9480   1.114  0.26734
EthN:SexM:AgeF2:LrnSL       2.1029     2.3865   0.881  0.38001
EthN:SexM:AgeF3:LrnSL           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2) family
     taken to be 1.060021)

    Null deviance: 280.18  on 145  degrees of freedom
Residual deviance: 171.98  on 118  degrees of freedom
AIC: 1095.4
```

Things look better here. The increasing variance has disappeared as can be seen in the following plots. Also based on the Pearson based measure of overdispersion, the negative binomial model seems to have accounted for much of the overdispersion.


Deviance Residual
Fitted Days Absent
Pearson Residual
Fitted Days Absent
Deviance Residual
Aboriginal
Non−Aboriginal
Ethnic Group
Deviance Residual
Female
Male
Gender
Deviance Residual
Primary
Second form
Education Level
Deviance Residual
Average learner
Slow learner
Learning Ability

One slight problem with this approach is that θ needs to be specified. This isn't required as we can estimate it along with β.

`MASS` has a function `glm.nb` for getting the maximum likelihood estimate of β and θ jointly. It works similarly to the `glm` function, but only works the negative binomial model. Thus it doesn't take a `family` option. Instead it takes a `link` options, with possibilities `log` (default), `identity`, and `sqrt`. There are `summary` and `anova` methods available for this function.

For the full interaction model

```
> quine.nb <- glm.nb(Days ~ .^4, data = quine)

> c(theta = quine.nb$theta, SE = quine.nb$SE)
    theta        SE
1.9283601 0.2688968

> summary(quine.nb)

Call:
```

```
glm.nb(formula = Days ~ .^4, data = quine, init.theta = 1.92836014
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2377  -0.9079  -0.2019   0.5173   1.7043

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.0564     0.3760   8.128 4.38e-16 ***
EthN                  -0.1386     0.5334  -0.260 0.795023
SexM                  -0.4914     0.5104  -0.963 0.335653
AgeF1                 -0.6227     0.5125  -1.215 0.224334
AgeF2                 -2.3632     1.0770  -2.194 0.028221 *
AgeF3                 -0.3784     0.4546  -0.832 0.405215
LrnSL                 -1.9577     0.9967  -1.964 0.049493 *

 . . .
```

```
EthN:SexM:AgeF2:LrnSL   2.1029     2.3444   0.897 0.369718
EthN:SexM:AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.9284) family
   taken to be 1)

    Null deviance: 272.29  on 145  degrees of freedom
Residual deviance: 167.45  on 118  degrees of freedom
AIC: 1097.3

Number of Fisher Scoring iterations: 1

Correlation of Coefficients:
                     (Intercept) EthN  SexM  AgeF1 AgeF2 AgeF3
EthN                 -0.70
SexM                 -0.74        0.52
AgeF1                -0.73        0.52  0.54
```

```
AgeF2                      -0.35         0.25  0.26  0.26

  . . .

EthN:SexM:AgeF1:LrnSL -0.43
EthN:SexM:AgeF2:LrnSL -0.69                0.52


              Theta:  1.928
         Std. Err.:  0.269

 2 x log-likelihood:  -1039.324
```

A more reasonable model in this situation, is to eliminate the `Eth:Sex:Age:Lrn` and `Eth:Sex:Lrn` interactions. This can be seen with

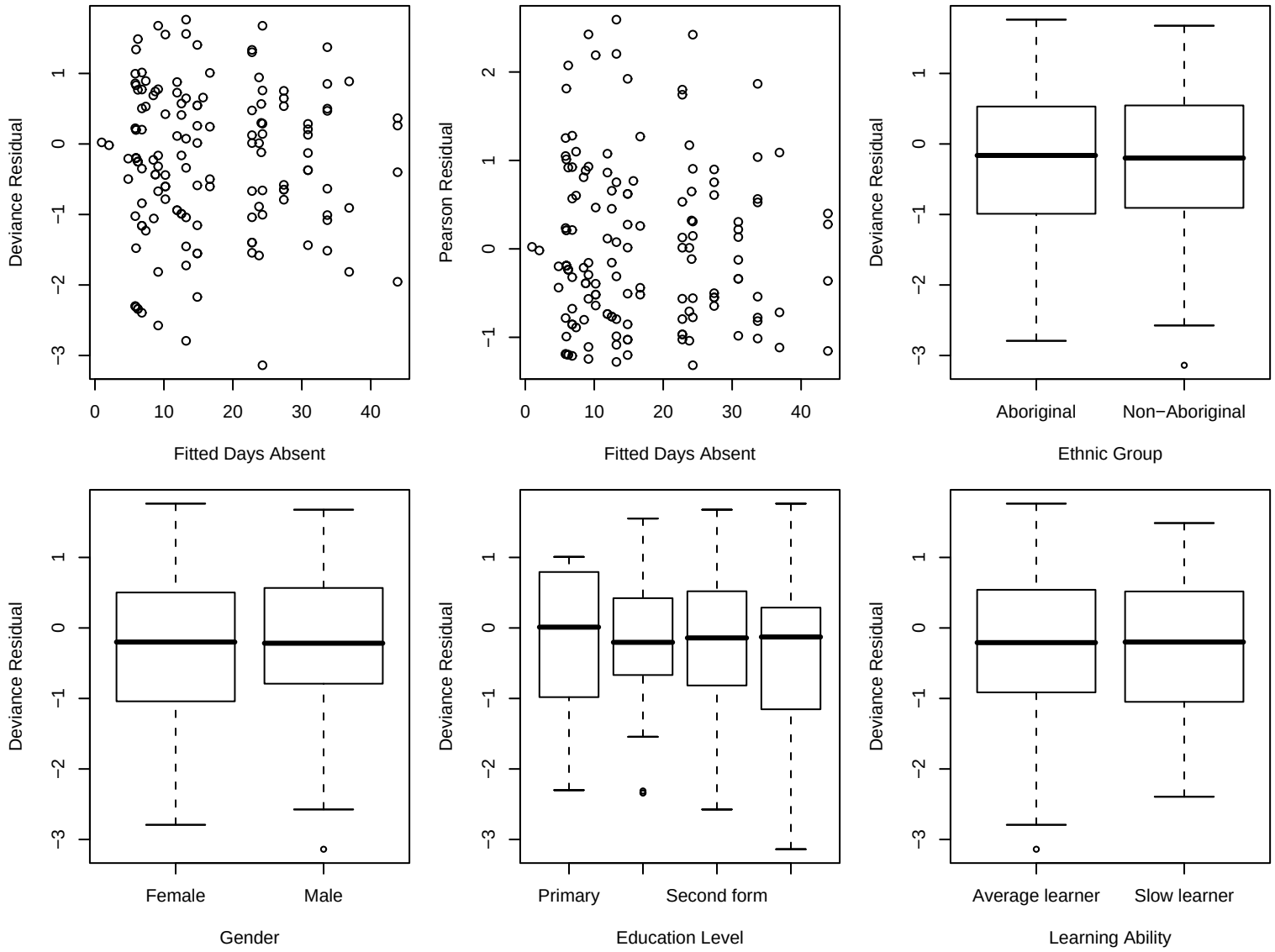
```
> quine2.nb <- glm.nb(Days ~ Lrn/(Age + Eth + Sex)^2, data=quine)

> anova(quine2.nb, quine.nb)
Likelihood ratio tests of Negative Binomial Models

Response: Days
                      Model    theta Resid. df    2 x log-lik.   Test
1   Lrn/(Age + Eth + Sex)^2 1.865343       123       -1043.409
2 (Eth + Sex + Age + Lrn)^4 1.928360       118       -1039.324 1 vs 2
     df LR stat.   Pr(Chi)
1
2     5 4.084768 0.5372772
```

The test performed here is a likelihood ratio test, assuming the estimated θ from the full model. The log-likelihood is calculated for the reduced model, under the θ calculated for the full model.

It ends up for the deviance tests to be applicable, the θ parameter needs to be held constant for all fitted models.

The residual plots do not suggest any serious problems with the smaller

model, as seen in the following plot



Log-linear Models for Two-way Contingency Tables

Consider the case where two categorical variables are of interest, X with r possible levels and Y with c possible levels.

For now, consider both as response variables (we'll consider other sampling schemes later)

Lets form the $r \times c$ table, with the (i, j)th entry equal to the number of observations with $X = x_i$ and $Y = y_j$, denoted by n_{ij}

Example: Business Administration Majors and Gender

A study of the career plans of young men and women sent questionaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen.

Major	Women	Men
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Lets assume that this data was generated under Poisson sampling. We want to come up with a model on how the cell counts depend on the levels of X and Y.

The nature of dependence relates to the association and the interaction structure among the variables.

Model for the data

- The joint PDF of (X, Y): $P[X = x_i, Y = y_i] = \pi_{ij}$
- Marginal PDF of X: $P[X = x_i] = \pi_{i+}$
- Marginal PDF of Y: $P[Y = Y_j] = \pi_{+j}$
- Expected cell counts: $\mu_{ij} = n\pi_{ij}$

  where $n = n_{++}$ is the total count.
- $N = rc$ is the effective sample size (number of observations).
- Poisson rate: π_{ij}
- Log-linear model on $\log \mu_{ij}$

Independence Model for Two-way Table

If X and Y are independent, then

$$P[X = x_i, Y = y_i] = P[X = x_i] \times P[Y = y_i] = \pi_{i+}\pi_{+j}$$

and the expected count is

$$\mu_{ij} = n\pi_{ij} = N\pi_{i+}\pi_{+j}$$

This implies that the log-linear model satisfies

$$\begin{aligned} \log \mu_{ij} &= \log N + \log \pi_{i+} + \log \pi_{+j} \\ &= \lambda + \lambda_i^X + \lambda_j^Y \end{aligned}$$

The estimates for the marginal probabilities are

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n} \qquad \hat{\pi}_{+j} = \frac{n_{+j}}{n}$$

The fitted values for this model are

$$\mu_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$$

In **R**, the model can be fit by

```
> business.ind <- glm(n ~ major + gender, family=poisson(),
    data=business)
```

```
> summary(business.ind)

Call:
glm(formula = n ~ major + gender, family = poisson(),
  data = business)

Deviance Residuals:
      1         2         3         4         5         6         7
-0.5085   0.5872   1.6257  -2.0806  -0.5802   0.6291  -1.0940   1.2

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.28054    0.09959  42.981  < 2e-16 ***
majorAdministration  0.05492    0.12529   0.438  0.66117
majorEconomics      -2.42239    0.31460  -7.700 1.36e-14 ***
majorFinance        -0.03279    0.12805  -0.256  0.79790
genderMale          -0.33470    0.10323  -3.242  0.00119 **
---
(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 168.473  on 7  degrees of freedom
Residual deviance:  11.017  on 3  degrees of freedom
AIC: 63.832

Number of Fisher Scoring iterations: 4

> anova(business.ind, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log
Response: n

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       7    168.473
major   3  146.796          4     21.677 1.294e-31
gender  1   10.661          3     11.017     0.001
```

We can check for goodness of fit with either the deviance or Pearson GOF tests.

For this example, the independence model doesn't seems to fit properly. The deviance test gives

```
> pchisq(deviance(business.ind),df.residual(business.ind),
   lower.tail=F)
[1] 0.01163662
```

The Pearson test for two way tables can be calculated by

```
> business.tab
                gender
major            Female Male
  Accounting         68   56
  Administration     91   40
  Economics           5    6
  Finance            61   59
```

```
> chisq.test(business.tab)

        Pearson's Chi-squared test

data:  business.tab X-squared = 10.8267, df = 3, p-value = 0.0127

Warning message: Chi-squared approximation may be incorrect in:
chisq.test(business.tab)
```

where `business.tab` is the 2-way table of counts.