Another Example, from Raw data

I dug up the following very raw data from a data archive. I would like to study these data, finding interesting questions, and manipulating the data to answer the questions. The subjects were 478 students in grades 4–6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer (Chase, M. A., and Dummer, G. M. (1992), *The Role of Sports as a Social Determinant for Children*, Research Quarterly for Exercise and Sport, 63, 418-424) stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.

The variable names are

1. Gender: Boy or girl
2. Grade: 4, 5 or 6
3. Age: Age in years
4. Race: White, Other
5. Urban/Rural: Rural, Suburban, or Urban school district
6. School: Brentwood Elementary, Brentwood Middle, Ridge, Sand, Eureka, Brown, Main, Portage, Westdale Middle
7. Goals: Student's choice in the personal goals question where options were 1 = Make Good Grades, 2 = Be Popular, 3 = Be Good in Sports
8. Grades: Rank of "make good grades" (1=most important for popularity, 4=least important)
9. Sports: Rank of "being good at sports" (1=most important for popularity, 4=least important)
10. Looks: Rank of "being handsome or pretty" (1=most important for popularity, 4=least important)
11. Money: Rank of "having lots of money" (1=most important for popularity, 4=least important)

The first few lines of the data

```
Gender  Grade   Age     Race    Urban/Rural School  Goals   Grades  Sports  Looks   Money
boy     5       11      White   Rural       Elm     Sports  1       2       4       3
boy     5       10      White   Rural       Elm     Popular 2       1       4       3
girl    5       11      White   Rural       Elm     Popular 4       3       1       2
girl    5       11      White   Rural       Elm     Popular 2       3       4       1
girl    5       10      White   Rural       Elm     Popular 4       2       1       3
girl    5       11      White   Rural       Elm     Popular 4       2       1       3
girl    5       10      White   Rural       Elm     Popular 3       4       1       2
girl    5       10      White   Rural       Elm     Grades  3       4       2       1
girl    5       10      White   Rural       Elm     Sports  3       2       1       4
girl    5       10      White   Rural       Elm     Sports  4       3       2       1
girl    5       11      White   Rural       Elm     Sports  2       3       1       4
girl    4       10      White   Rural       Elm     Grades  2       3       4       1
```

So, now what do we do. This doesn't look like a table at all. There seems to be one line of data for each subject. So, at the least we have to tabulate, but first we have to decide what we want tablulated.

There are many variables to consider, but for now I want to concentrate on just 5 of them, Gender, Grade, Race, Urban.Rural, and Goals. One of these (Urban.Rural) has to be considered an explanatory variable (recall the earlier discussion about how the data were collected). One can debate whether each of the others is explanatory or response, but there is at least no danger considering the rest as response variables.

```
Start by reading in the raw data......
> kids <- read.table("ex.dat", head=T)
> kids[1:10,]
   Gender Grade Age  Race Urban.Rural School   Goals Grades Sports Looks Money
 1    boy     5  11 White       Rural    Elm  Sports      1      2     4     3
 2    boy     5  10 White       Rural    Elm Popular      2      1     4     3
 3   girl     5  11 White       Rural    Elm Popular      4      3     1     2
 4   girl     5  11 White       Rural    Elm Popular      2      3     4     1
 5   girl     5  10 White       Rural    Elm Popular      4      2     1     3
 6   girl     5  11 White       Rural    Elm Popular      4      2     1     3
 7   girl     5  10 White       Rural    Elm Popular      3      4     1     2
 8   girl     5  10 White       Rural    Elm  Grades      3      4     2     1
 9   girl     5  10 White       Rural    Elm  Sports      3      2     1     4
10   girl     5  10 White       Rural    Elm  Sports      4      3     2     1
>
```

We still only have the raw data, and we have no good ideas on how to tabulate. **There has to be a better way of doing this, but right now I can't seem to figure it out. I'm open to suggestions.** What I did was use the Splus `tabulate` command, and then use some tricks from there.

```
> attach(kids,pos=1)
> table(Gender,Grade,Race,Urban.Rural,Goals)

, , Other, Rural, Grades
     4 5 6
 boy 0 0 0
girl 0 1 2

, , White, Rural, Grades
      4 5  6
 boy  7 6  8
girl 12 9 12

, , Other, Suburban, Grades
     4 5 6
 boy 0 2 2
girl 0 1 0
```

This has clearly produced something that looks like a matrix (or table), but that is not the form that we need for the `glm` command. (It is what the `loglinear` command needs, but we won't ever use that.

As it isn't the format we need, I created a little bit of code to turn this into the proper form. If I were smarter, I would clean it up and make it more general, but I'm not.

```
doit <- function(){
        fofo <-table(Gender,Grade,Race,Urban.Rural,Goals)
        dimd <- attributes(fofo)$dim
        name <- attributes(fofo)$dimnames
        fnames<-list(Gender=name[[1]],Grade=name[[2]],
             Race=name[[3]],Urban=name[[4]],Goals=name[[5]])
        ans <- fac.design(dimd,fnames)
        ans$Resp <- as.vector(fofo)
        ans
}
```

It doesn't much matter if you follow this or not, the chief idea is that it takes a table and converts it into the correct type of data frame.

```
> kids.f <- doit()
> kids.f[1:10,]
   Gender Grade  Race Urban  Goals Resp
 1    boy     4 Other Rural Grades    0
 2   girl     4 Other Rural Grades    0
 3    boy     5 Other Rural Grades    0
 4   girl     5 Other Rural Grades    1
 5    boy     6 Other Rural Grades    0
 6   girl     6 Other Rural Grades    2
 7    boy     4 White Rural Grades    7
 8   girl     4 White Rural Grades   12
 9    boy     5 White Rural Grades    6
10   girl     5 White Rural Grades    9

> assign("kids.f", kids.f,w=2)
> detach(1,save=F)
```

Whew!, now at least I have the data in a form that I can use, so maybe I can get on with the modelling. Now, try to look at the data as a loglinear model.

```
> attach(kids.f,pos=1)
> ls()
[1] ".Last.value" "Gender"      "Goals"       "Grade"       "Race"
[6] "Resp"        "Urban"
```

What follows is not a particularly directed analysis, but we will end up in a reasonable place.

```
> mymod <- glm(Resp ~ Urban*Gender*Goals*Grade*Race -
         Urban:Gender:Goals:Grade:Race, family=poisson)
```

```
Warning messages:
  linear convergence not obtained in  10  iterations. in: glm.fitter(x = X, y =
      Y, w = w, start = start, offset = offset, family = family, ....
```

Note the error messages about lack of convergence. Why might that happen, and why didn't it worry me (yet).

```
> anova(mymod)
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
                         Df Deviance Resid. Df Resid. Dev
                    NULL                   107   679.9861
                   Urban  2   3.2349       105   676.7511
                  Gender  1   1.2055       104   675.5456
                   Goals  2  79.2794       102   596.2662
                   Grade  2  16.2392       100   580.0271
                    Race  1 407.2321        99   172.7949
            Urban:Gender  2  12.1266        97   160.6683
             Urban:Goals  4  18.5706        93   142.0976
            Gender:Goals  2  23.3472        91   118.7504
             Urban:Grade  4  30.9144        87    87.8360
            Gender:Grade  2   0.8923        85    86.9437
             Goals:Grade  4   1.3379        81    85.6058
              Urban:Race  2  10.2736        79    75.3322
             Gender:Race  1   0.3164        78    75.0158
              Goals:Race  2   0.6479        76    74.3679
              Grade:Race  2   1.0471        74    73.3208
      Urban:Gender:Goals  4   3.7229        70    69.5979
      Urban:Gender:Grade  4   4.3112        66    65.2867
       Urban:Goals:Grade  8   6.4401        58    58.8466
      Gender:Goals:Grade  4   5.2033        54    53.6433
       Urban:Gender:Race  2   1.2544        52    52.3889
        Urban:Goals:Race  4   4.9217        48    47.4672
       Gender:Goals:Race  2   2.5781        46    44.8892
        Urban:Grade:Race  4   4.5668        42    40.3224
       Gender:Grade:Race  2   1.2655        40    39.0568
        Goals:Grade:Race  4   4.9344        36    34.1224
Urban:Gender:Goals:Grade  8  10.4653        28    23.6572
 Urban:Gender:Goals:Race  4  11.2730        24    12.3841
 Urban:Gender:Grade:Race  4   2.2919        20    10.0922
  Urban:Goals:Grade:Race  8   3.5111        12     6.5811
 Gender:Goals:Grade:Race  4   6.5796         8     0.0015
Warning messages:
```

```
(like before)
```

That was actually quite informative. (The model fits well, it might fit better if converged). The 4-way interactions all seem pretty useless, so lets remove them all.

```
> mymod <- update(mymod, . ~ . - Gender:Goals:Grade:Race -
              Urban:Goals:Grade:Race - Urban:Gender:Grade:Race -
              Urban:Gender:Goals:Race - Urban:Gender:Goals:Grade)
> anova(mymod)
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
                  Df Deviance Resid. Df Resid. Dev
             NULL                   107   679.9861
            Urban  2   3.2349       105   676.7511
           Gender  1   1.2055       104   675.5456
            Goals  2  79.2794       102   596.2662
            Grade  2  16.2392       100   580.0271
             Race  1 407.2321        99   172.7949
      Urban:Gender  2  12.1266        97   160.6683
       Urban:Goals  4  18.5706        93   142.0976
      Gender:Goals  2  23.3472        91   118.7504
       Urban:Grade  4  30.9144        87    87.8360
      Gender:Grade  2   0.8923        85    86.9437
       Goals:Grade  4   1.3379        81    85.6058
        Urban:Race  2  10.2736        79    75.3322
       Gender:Race  1   0.3164        78    75.0158
        Goals:Race  2   0.6479        76    74.3679
        Grade:Race  2   1.0471        74    73.3208
Urban:Gender:Goals  4   3.7229        70    69.5979
Urban:Gender:Grade  4   4.3112        66    65.2867
 Urban:Goals:Grade  8   6.4401        58    58.8466
Gender:Goals:Grade  4   5.2033        54    53.6433
 Urban:Gender:Race  2   1.2544        52    52.3889
  Urban:Goals:Race  4   4.9217        48    47.4672
 Gender:Goals:Race  2   2.5781        46    44.8892
  Urban:Grade:Race  4   4.5668        42    40.3224
 Gender:Grade:Race  2   1.2655        40    39.0568
  Goals:Grade:Race  4   4.9344        36    34.1224
```

Again, the high order interactions are quite pointless, so be cavalier and remove them all.

```
> mymod <- update(mymod, . ~
  (Urban+Gender+Goals+Grade+Race)*
  (Urban+Gender+Goals+Grade+Race))
```

```
> anova(mymod)
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev
        NULL                   107   679.9861
       Urban  2   3.2349       105   676.7511
      Gender  1   1.2055       104   675.5456
       Goals  2  79.2794       102   596.2662
       Grade  2  16.2392       100   580.0271
        Race  1 407.2321        99   172.7949
Urban:Gender  2  12.1266        97   160.6683
 Urban:Goals  4  18.5706        93   142.0976
 Urban:Grade  4  30.9144        89   111.1832
  Urban:Race  2  10.2736        87   100.9097
Gender:Goals  2  23.3472        85    77.5624
Gender:Grade  2   0.8923        83    76.6701
 Gender:Race  1   0.3164        82    76.3537
 Goals:Grade  4   1.3379        78    75.0158
  Goals:Race  2   0.6479        76    74.3679
  Grade:Race  2   1.0471        74    73.3208
```

There are still some two-way interactions that are very small, so remove them.

```
> anova(mymod <- update(mymod, . ~ . - Grade:Race - Goals:Race - Goals:Grade))
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev
        NULL                   107   679.9861
       Urban  2   3.2349       105   676.7511
      Gender  1   1.2055       104   675.5456
       Goals  2  79.2794       102   596.2662
       Grade  2  16.2392       100   580.0271
        Race  1 407.2321        99   172.7949
Urban:Gender  2  12.1266        97   160.6683
 Urban:Goals  4  18.5706        93   142.0976
 Urban:Grade  4  30.9144        89   111.1832
  Urban:Race  2  10.2736        87   100.9097
Gender:Goals  2  23.3472        85    77.5624
```

```
Gender:Grade  2   0.8923        83    76.6701
 Gender:Race  1   0.3164        82    76.3537
>
> anova(mymod <- update(mymod, . ~ . - Gender:Race - Gender:Grade))
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev
        NULL                   107   679.9861
       Urban  2   3.2349       105   676.7511
      Gender  1   1.2055       104   675.5456
       Goals  2  79.2794       102   596.2662
       Grade  2  16.2392       100   580.0271
        Race  1 407.2321        99   172.7949
Urban:Gender  2  12.1266        97   160.6683
 Urban:Goals  4  18.5706        93   142.0976
 Urban:Grade  4  30.9144        89   111.1832
  Urban:Race  2  10.2736        87   100.9097
Gender:Goals  2  23.3472        85    77.5624
> qchisq(0.95,85)
[1] 107.5217
> qchisq(0.95,2)
[1] 5.991465
> qchisq(0.95,4)
[1] 9.487729
```

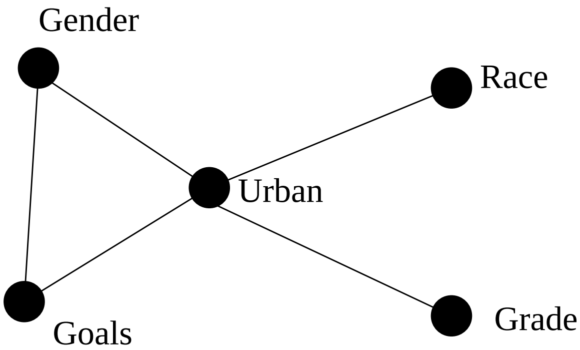
OK, now just about everything looks significant, and most surprising the overall model fits quite well! For reasons that will be much clearer later, I prefer to add a term to this model—a term which by itself is not very significant.

```
> anova(mymod <- update(mymod, . ~ . + Urban*Gender*Goals))
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
                   Df Deviance Resid. Df Resid. Dev
              NULL                   107   679.9861
             Urban  2   3.2349       105   676.7511
            Gender  1   1.2055       104   675.5456
             Goals  2  79.2794       102   596.2662
             Grade  2  16.2392       100   580.0271
```

```
               Race  1 407.2321        99   172.7949
       Urban:Gender  2  12.1266        97   160.6683
        Urban:Goals  4  18.5706        93   142.0976
        Urban:Grade  4  30.9144        89   111.1832
         Urban:Race  2  10.2736        87   100.9097
       Gender:Goals  2  23.3472        85    77.5624
Urban:Gender:Goals  4   3.7406        81    73.8219
>
```

For this model the plots look OK, I can interpret the model and no other terms seem to be important.

But what is the interpretation of the model? I can only answer that by drawing a picture.



It isn't obvious what this picture means, so let me "read" it to you. Remember that `Urban` is really an explanatory variable, but for now lets pretend it is a response. The picture says that Gender and Goals are conditionally independent of Race and Grade give the Urban Status. Even more, Gender and Goals are related (that is not independent), but Race is independent of Grade, given Urban Status.

Eventually you will learn how to read these sort of graphs, and you will also understand why I wanted the three-way interaction that I did.

More models and some notation

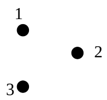
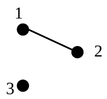
In order to interpret our results we intend to restrict our approach to a hierarchical family of models. The family we will consider restricts us to the class where if we want to fit an interaction term we *must* also fit the corresponding lower order interactions and main effects.

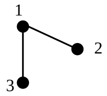
Now consider a 3-way table.

$x_{1,1,1}$	$x_{1,1,2}$	$x_{1,1,3}$
$x_{1,2,1}$	$x_{1,2,2}$	$x_{1,2,3}$

$x_{2,1,1}$	$x_{2,1,2}$	$x_{2,1,3}$
$x_{2,2,1}$	$x_{2,2,2}$	$x_{2,2,3}$

We will consider the following models, abbreviations, and corresponding pictures.

Model	Sufficient Stats	Picture	Meaning
$u + u_1 + u_2 + u_3$	[1][2][3]		1 indep 2 indep 3
$u + u_1 + u_2 + u_3 + u_{12}$	[12][3]		(1 and 2) indep 3

$u + u_1 + u_2 + u_3 +$ $u_{12} + u_{13}$	[12][23]		1 indep 3 given 2
$u + u_1 + u_2 + u_3 +$ $u_{12} + u_{13} + u_{23}$	[12][23][23]	none	I don't know

The [$\cdot$] are the indicate the margins of the table that make up the sufficient statistics—thus [1] is the margin formed by summing over the 2 and 3 indices of the table and [12] is the margin formed by summing over the 3 index.

Pictures (like the above) are really only appropriate when all the variables are responses, but they are very useful in all cases.