U.S Births in 1990 and 2009

by marriage status, race, age, and education

J.A.Hartigan Yale University
September 2013

1 Unwed births

In 2009, for the first time since the census began in 1790, there were more unwed births in the United States than wed births. We will study the changing rates of unwed births in the United States as a function of race, age, and education using data obtained from: http://www.cdc.gov/nchs/VitalStats.htm.

See the data preparation section for extracting the data from the website, and converting it to R friendly form in the data frames b and bb:

```
b <- read.csv("data/Births19902009.csv", header=T)
bb <- read.csv("data/bBirths19902009.csv",
header=T)
```

Looking at the data a little:

```
head(b,10)
```

```
      race   age     edu year marriage births
1  hispanic 15-19     0-8 1990      wed  10863
2  hispanic 15-19    9-11 1990      wed  18148
3  hispanic 15-19      12 1990      wed   8648
4  hispanic 15-19   13-15 1990      wed    939
5  hispanic 15-19     16+ 1990      wed      0
6  hispanic 15-19 unknown 1990      wed   1502
7  hispanic 20-24     0-8 1990      wed  25963
8  hispanic 20-24    9-11 1990      wed  28297
9  hispanic 20-24      12 1990      wed  39408
10 hispanic 20-24   13-15 1990      wed  13979
```

```
head(bb,10)
```

```
      race   age     edu year   wed unwed
1  hispanic 15-19     0-8 1990 10863 13637
2  hispanic 15-19    9-11 1990 18148 31985
3  hispanic 15-19      12 1990  8648 10421
4  hispanic 15-19   13-15 1990   939   968
5  hispanic 15-19     16+ 1990     0     0
6  hispanic 15-19 unknown 1990  1502  2920
7  hispanic 20-24     0-8 1990 25963 18065
8  hispanic 20-24    9-11 1990 28297 24625
9  hispanic 20-24      12 1990 39408 22065
10 hispanic 20-24   13-15 1990 13979  5886
```

Convert character variables to factors:

```
b$edu <- factor(b$edu,
levels(b$edu) <- c("0-8", "9-11", "12", "13-15",
"16+", "unknown"))
b$race <- factor(b$race,

levels=c("asian","white","hispanic","black"))
b$age <- factor(b$age,
      levels= c("15-19", "20-24", "25-29",
"30-34"))
b$year <- factor(b$year)
bb$edu <- factor(bb$edu,
levels(bb$edu) <- c("0-8", "9-11", "12", "13-15",
"16+", "unknown"))
bb$race <- factor(bb$race,
    levels=c("asian","white","hispanic","black"))
bb$age <- factor(bb$age,
      levels= c("15-19", "20-24", "25-29",
"30-34"))
bb$year <- factor(bb$year)
```

The levels for race are chosen in order of proportions of wed mothers in the different ethnic groups.

2 Births to unmarried women

Cross tabulate births by marriage status and race:

```
print(xtabs(births ~ marriage + race, data=b))
```

```
       race
marriage   asian   white hispanic   black
   unwed  100129 1084203   749945  885654
   wed    348229 3754495   844627  385578
```

```
print(xtabs(births ~ marriage + year, data=b))
```

```
       year
marriage    1990    2009
   unwed 1138499 1681432
   wed   2913991 2418938
```

The proportion of unmarried births increased from 28% to 41% between 1990 and 2009.

```
print(xtabs(births ~ marriage + race + year,
data=b))
```

```
, , year = 1990
        race
marriage   asian   white hispanic   black
   unwed   35338  443012   218515  441634
   wed    133878 2183488   376558  220067
, , year = 2009
        race
marriage   asian   white hispanic   black
   unwed   64791  641191   531430  444020
   wed    214351 1571007   468069  165511
```

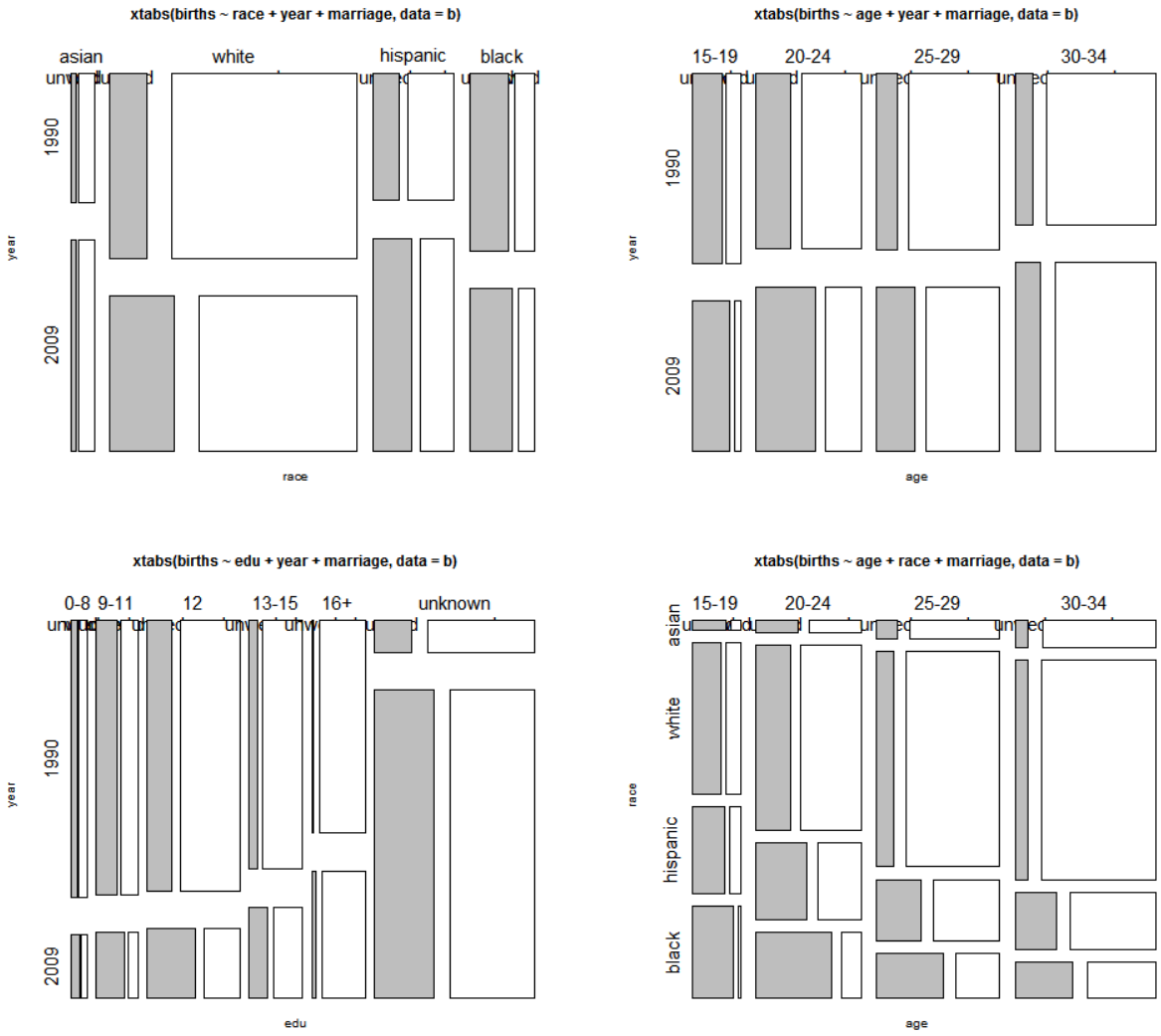
Looking in more detail, we see the largest increases in births to unmarried women in the white and Hispanic groups between 1990 and 2009. The proportions of unmarried births increased substantially in all groups.

3 Mosaics

Mosaics display many-way contingency tables. Each rectangle in the mosaic has an area proportional to the count in one of the cells of the table. The side of a rectangle is proportional to the conditional probability of some event. Order the variables in xtabs so that the most important variables are last.

3.1 Three way mosaics

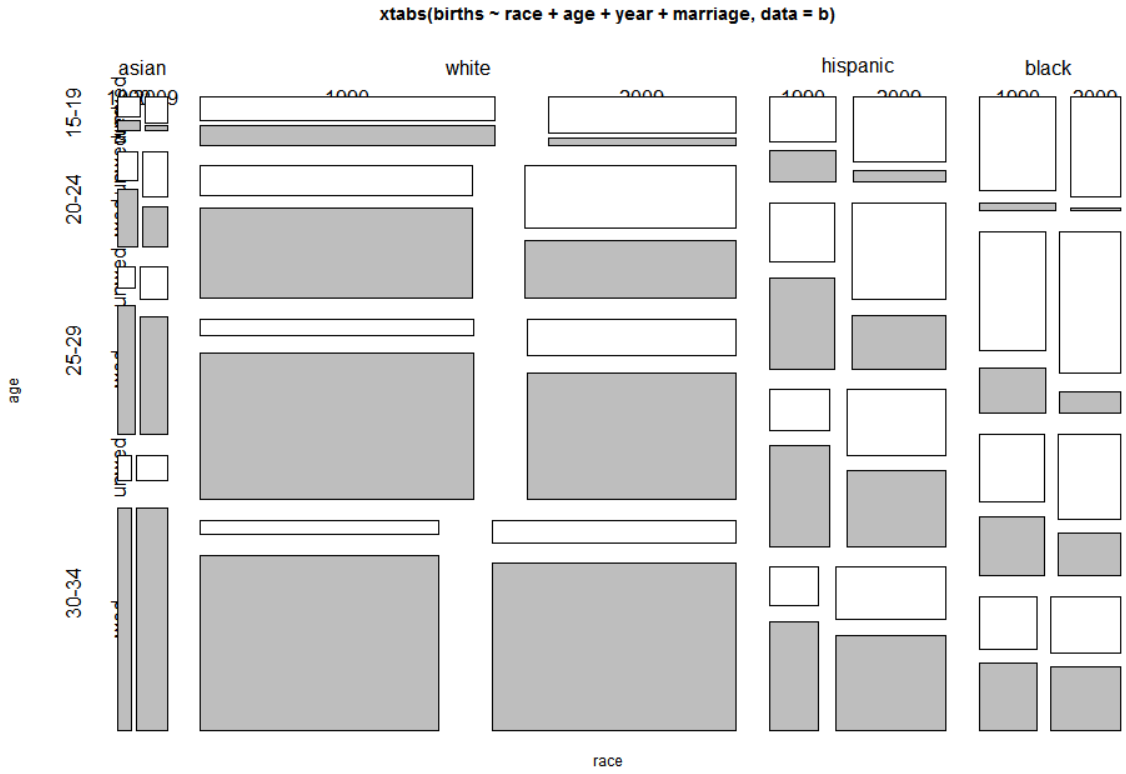
```
tiff("pictures/three way mosaics.tif", w=1000,
h=880)
par(mfrow=c(2,2))
mosaicplot(xtabs(births ~ race + year + marriage,
data=b), color= c("grey","white"), cex.axis=1.5)

mosaicplot(xtabs(births ~ age + year + marriage,
data=b), color= c("grey","white"), cex.axis=1.5)

mosaicplot(xtabs(births ~ edu + year + marriage,
data=b), color= c("grey","white"), cex.axis=1.5)

mosaicplot(xtabs(births ~ age + race + marriage,
data=b), color= c("grey","white"), cex.axis=1.5)

dev.off()
```



For example, in the first mosaic on marriage by race and year, the grey rectangle in the upper left corner has area proportional to the number of unwed Asian births in 1990. The white rectangle immediately to its right is the number of wed Asian births in 1990. The common vertical edge of these two rectangles is proportional to the probability of birth being in 1990 rather than in 2009, given that it is an Asian birth. The grey areas correspond to unmarried births, so we see the increase in unmarried births between 1990 and 2009. There are increased births overall for Asians and Hispanics. The unmarried rate for blacks is high in both years, but it increases markedly for whites and Hispanics between the years.

3.2 Five way mosaics

```
tiff("pictures/Four way mosaic.tif", w=1000, h=680)

par(mfrow=c(1,1))
mosaicplot(xtabs(births ~ race + age +  year +
marriage,

data=b),color=c("white","grey"),cex.axis=1.3)

dev.off()
```



From the large mosaic, we see that the unwed births (white blocks) occur mainly in people under age 25 across all racial groups. These numbers have increased dramatically between 1990 to 2009, especially for whites and Hispanics. The unwed births are high for blacks in all age groups in both periods, but have not increased so much relative to other racial groups.

4 Binomial model for wed vs unwed births

In the "logit" link function, we assume that the probability p of a birth being from a married woman , as race, age, education and year vary, satisfies:

log(p/(1-p)) = Linear function of numerical codings of race, age, education and year.

We use data previously saved in bb, containing columns for wed and unwed births.

The numerical codings are determined to optimally predict the probabilities, but are not uniquely specified. As a default, R chooses the first category for each factor to have zero numerical code. Usually, it is best to make that first category the most numerous, because the coefficients that we see represent differences in effect between the given category and the base category.

Relevel factors to select most numerous category:

```
print(xtabs(births~edu, data=b))
```

Rather than enter lots of levels, use relevel to position the desired category of age first:

```
totalbirths <- bb$wed + bb$unwed
print(xtabs(totalbirths~ bb$age))
```

```
bb$age
  15-19   20-24   25-29   30-34
 931982 2065240 2403936 2751702
```

5 Main effects predicting probability of unwed births:

Compute model and determine coefficients and fit:

```
glmbb <-glm(cbind(unwed+1, wed+1)~
race + age + edu + year, family=binomial, data=bb)
```

Use zero sum constraints on the coefficients for each factor, to simplify interpretation of the coefficients:

```
coef <- glmZS(fdata=bb, cbind(unwed+1, wed+1)~
     race + age + edu + year, family=binomial)$coef
coef <- data.frame(coef[, c(1,3)])
coef$relrate <- exp(coef[, 1])
round(coef,2)
```

	Estimate	z.value	relrate
(Intercept)	-0.27	-173.44	0.76
age.15-19	1.37	636.36	3.95
age.20-24	0.30	207.63	1.35
age.25-29	-0.59	-390.92	0.55
age.30-34	-1.08	-671.12	0.34
edu.0-8	0.46	123.33	1.59
edu.9-11	0.77	296.40	2.15
edu.12	0.22	109.79	1.24
edu.13-15	-0.18	-73.43	0.83
edu.16+	-1.23	-335.02	0.29
edu.unknown	-0.04	-17.80	0.96
race.asian	-0.64	-210.74	0.53
race.white	-0.72	-470.64	0.49
race.hispanic	0.07	37.07	1.07
race.black	1.29	644.04	3.64
year.1990	-0.52	-421.60	0.59
year.2009	0.52	421.60	1.69

The relative rate is the ratio, of the ratio of unwed to wed births for the specified category, to the ratio of unwed to wed births in general. For example, the relative rates of .59 for 1990 and 1.7 for 2009 means that the unwed/wed birth ratio increased by the factor of 3 = 1.7/.59 between 1990 and 2009. There is a huge increase in unwed birth odds for the under 15 age category. Blacks have seven times the rate of unwed births compared to whites ($7 \approx 3.64/.49$).

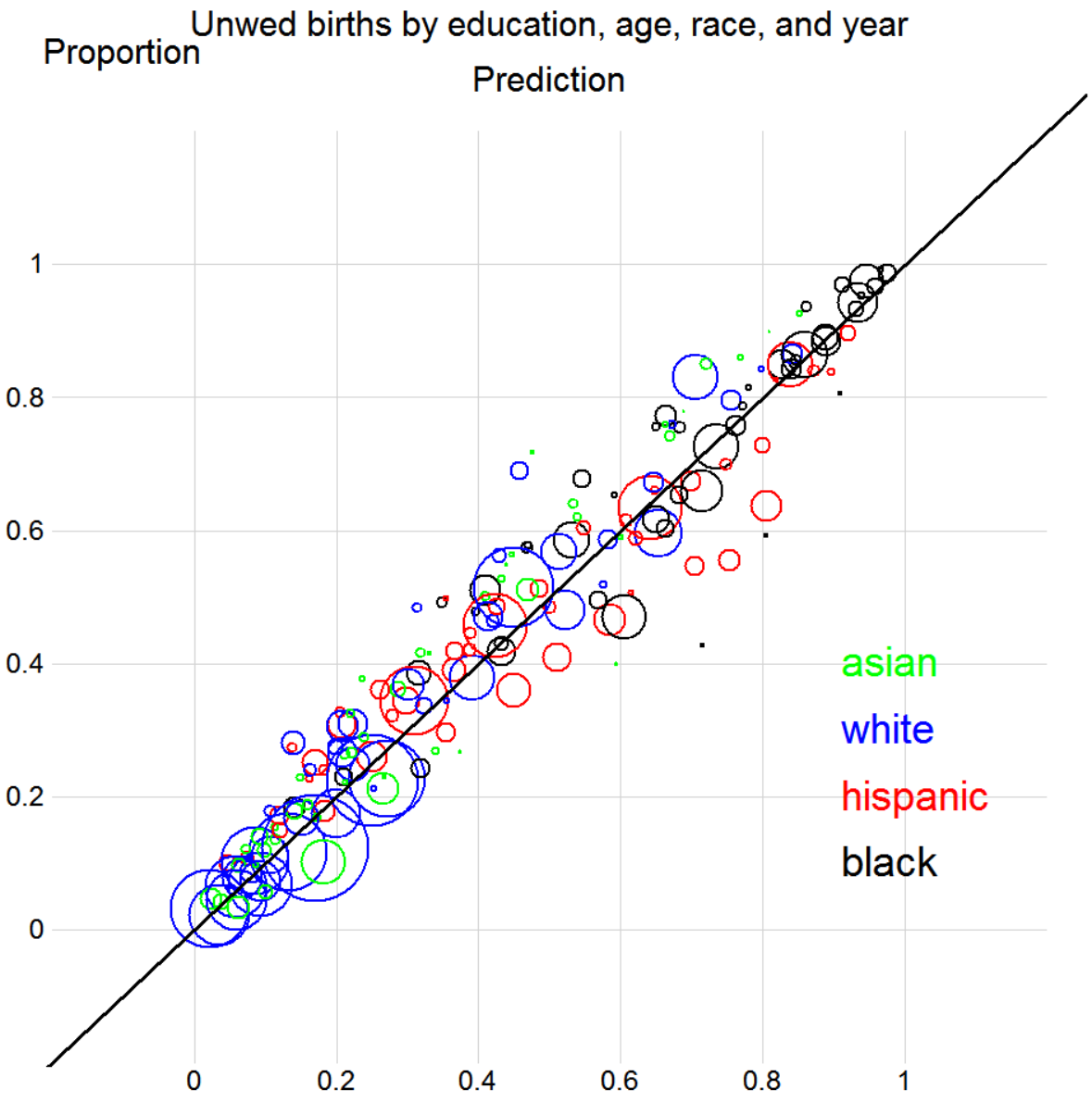
6 **Binomial Plots for main effects model**

The fit estimates the probabilities of unwed births in each row of the table by exp(M)/(1+exp(M)) where M is the estimated linear combination of category codings for that row.

```
tiff("pictures/binomial main effects.tif", w=1000,
h=1000)
bb$total <- bb$unwed + bb$wed
```

Add 1 to unwed and wed counts to handle cells with no births:

```
bb$unwedp <-(bb$unwed+1) / (bb$total + 2)

# plot actual against fit
color <- rep(1,dim(bb)[1])
color[which(bb$race == "asian")] <- "green"
color[which(bb$race == "white")] <- "blue"
color[which(bb$race == "hispanic")] <- "red"
color[which(bb$race == "black")] <- "black"

Grid(c(seq(-0.2, 1.2, 0.2)), c(seq(-0.2, 1.2,
0.2)),
ylab="Unwed births by education, age, race, and
year/Proportion/Prediction", at=c(0.5, -0.1, 0.5),
cex=2.5)

points(glmbb$fit, bb$unwedp, cex
=sqrt(bb$total/2000), col=color, lwd=2)

text(0.9,0.4, "asian", col="green", pos=4,cex=3)
text(0.9,0.3, "white", col="blue", pos=4, cex=3)
text(0.9,0.2, "hispanic", col="red", pos=4,cex=3)
text(0.9,0.1, "black", col="black", pos=4, cex=3)

abline(lm(bb$unwedp ~glmbb$fit, weight=bb$total),
lwd=3)

dev.off()
```




```
w <- bb$total / ( glmbb$fit*(1-glmbb$fit) )
corr <-sqrt(summary(lm(bb$unwedp ~ glmbb$fit,
w=w))$r.sq)
cat("Correlation", round(corr, 4), "")
```

```
Correlation 0.9903
```

The plot shows that the simple main effects logit model predicts the actual proportions fairly well. There are notable deviations from the line in some hispanic groups. Note also high unwed birth rates in the various black cells.

7 Binomial model Interactions

```
glmbb <-glm(cbind(bb$unwed, bb$wed)~
(race + age + edu + year)^2, data=bb,
family=binomial)
```

Include this function to give simpler anova display:

```
FixDeviance<- function(a){
a[,3] <- (a[,2]/a[,1])/(a[nrow(a),4]/a[nrow(a),3])
names(a)[3] <- "F"
return(a[,-4])
}
```

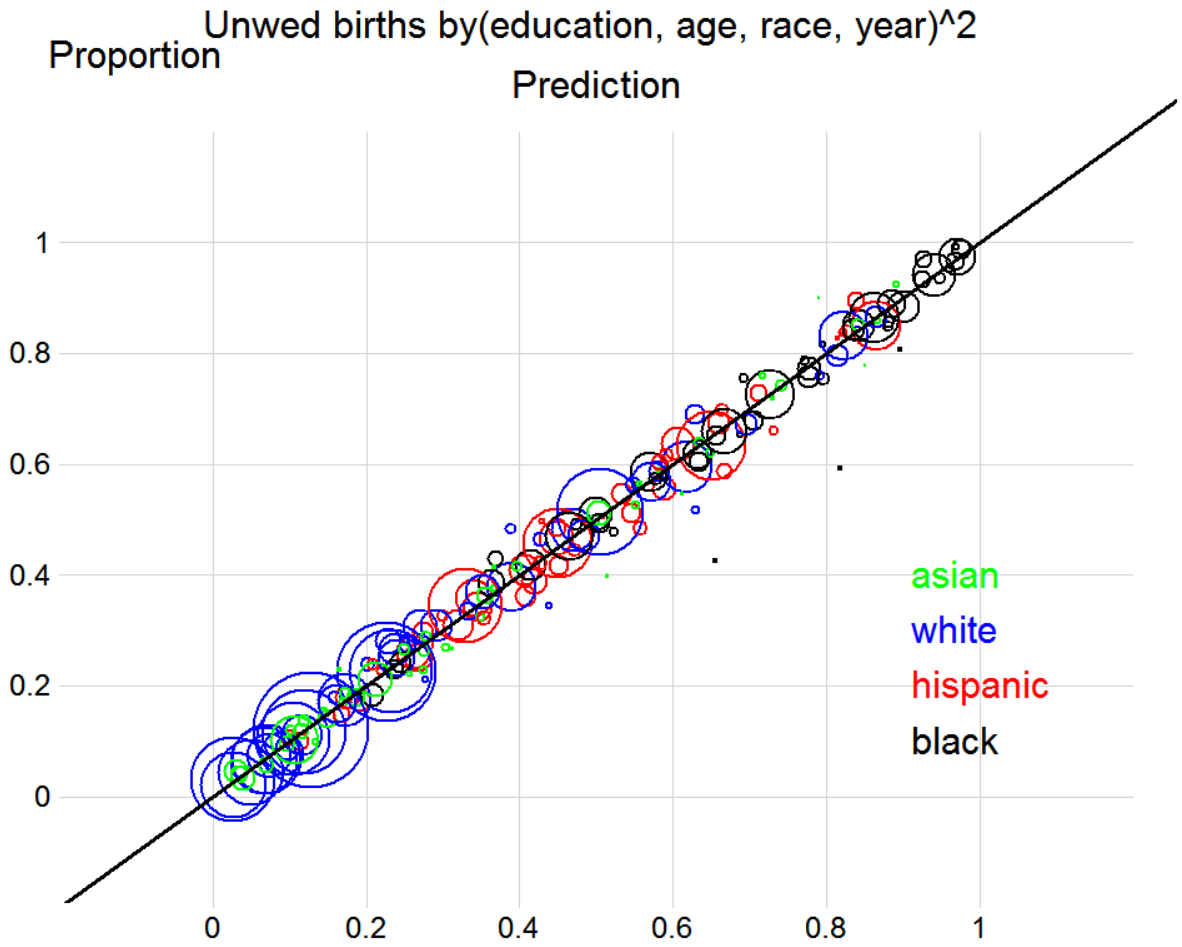
```
print( round(FixDeviance(anova(glmbb))) )
```

```
          Df Deviance    F
NULL
race       3  1125123 4453
age        3  1102002 4361
edu        5   247211  587
year       1   185230 2199
race:age   9    27676   37
race:edu  15    13061   10
race:year  3     4974   20
age:edu   14    54535   46
age:year   3     8574   34
edu:year   5     9324   22
```

The deviance is the increase in 2 log likelihood of the data for each additional term included in the fit, analogous to decrease in sums of squares for terms added to a linear regression model. And similarly, the F is the ratio of the deviance divided by the number of additional parameters for that term, to the residual deviance divided by the residual number of parameters. The F is a guide to the overall importance of the term. An F of about 1 would be achieved for a factor that had no predictive value at all. The second order F's are all small compared to the main order terms, but actually the more interesting effects lie in the interactions. For example, how do racial differences in unwed birth ratios change over time? This is the race:year interaction.

8 Plotting unwed proportions against fitted model using interactions

```
tiff("pictures/Interactions.tif", w=1000, h=800)

Grid(c(seq(-0.2,1.2,0.2)), c(seq(-0.2, 1.2, 0.2)),
ylab="Unwed births by(education, age, race, year)^2
/Proportion/Prediction", at=c(0.5,-0.1, 0.5),
cex=2.6)
points(glmbb$fit, bb$unwedp, cex
=sqrt(bb$total/2000), col=color, lwd=2)

text(0.9,0.4, "asian", col="green", pos=4, cex=2.5)
text(0.9,0.3, "white", col="blue", pos=4, cex=2.5)
text(0.9,0.2, "hispanic", col="red", pos=4,
cex=2.5)
text(0.9,0.1, "black", col="black", pos=4, cex=2.5)

abline(lm(bb$unwedp~glmbb$fit,w=bb$total),lwd=3)

dev.off()
```



Compute the correlation between fit and proportion:

```
w <- bb$total / ( glmbb$fit*(1-glmbb$fit) )
corr <-sqrt(summary(lm(bb$unwedp ~ glmbb$fit,
w=w))$r.sq)
cat("Correlation", round(corr, 4), "")
```

```
Correlation 0.9993
```

Very accurate prediction with the second order model suggests using it rather than the main effects model.

9 Interpreting Interaction terms

We look only at coefficients with z-values exceeding 50, to limit the number of effects we examine. Almost all second order effects are "statistically significant" , even while the effect sizes are small, because we have so much count.

```
coef <-glmZS(fdata=bb, cbind(unwed+1, wed+1)~
(race+age+edu+year)^2, family=binomial)$coef

coef <- data.frame(coef[, c(1,3)])
coef$relrate <- exp(coef[, 1])
print(round(coef[abs(coef[,2])> 50,], 2))
```

	Estimate	z.value	relrate
race.asian	-0.55	-116.34	0.58
race.white	-0.71	-284.50	0.49
race.black	1.31	372.81	3.69
year.1990	-0.45	-222.36	0.64
year.2009	0.45	222.36	1.57
age.30-34:year.1990	0.19	87.46	1.21
age.30-34:year.2009	-0.19	-87.46	0.82
edu.12:year.1990	-0.12	-50.04	0.89
edu.unknown:year.1990	0.18	68.06	1.20
edu.12:year.2009	0.12	50.04	1.12
edu.unknown:year.2009	-0.18	-68.06	0.84
race.hispanic:age.15-19	-0.42	-87.60	0.66
race.hispanic:age.25-29	0.18	55.45	1.19
race.white:year.1990	-0.12	-54.73	0.89
race.hispanic:age.30-34	0.38	117.07	1.46
race.white:year.2009	0.12	54.73	1.12

Note that the main effects coefficients are similar to their values in the main effects only model. This is one of the benefits of requiring the coefficients for each factor term to sum to zero, because then the correlations between the main effect predictors and the interaction predictors are nearly zero, and so adding interactions does not much affect the main effect estimates.

The large hispanic interactions show that young hispanic mothers are relatively more likely to be married, and older hispanic mothers relatively less likely to be married.

10 Binomial model Third order interactions

```
glmbb <-glm(cbind(bb$unwed, bb$wed)
~ (race + age + edu + year)^3, data=bb,
family=binomial)
print( round( FixDeviance(anova(glmbb)) ) )
```

```
              Df Deviance     F
NULL
race           3  1125123 38457
age            3  1102002 37667
edu            5   247211  5070
year           1   185230 18994
race:age       9    27676   315
race:edu      15    13061    89
race:year      3     4974   170
age:edu       14    54535   399
age:year       3     8574   293
edu:year       5     9324   191
race:age:edu  42     6260    15
race:age:year  9      905    10
race:edu:year 15     1706    12
age:edu:year  14      996     7
```

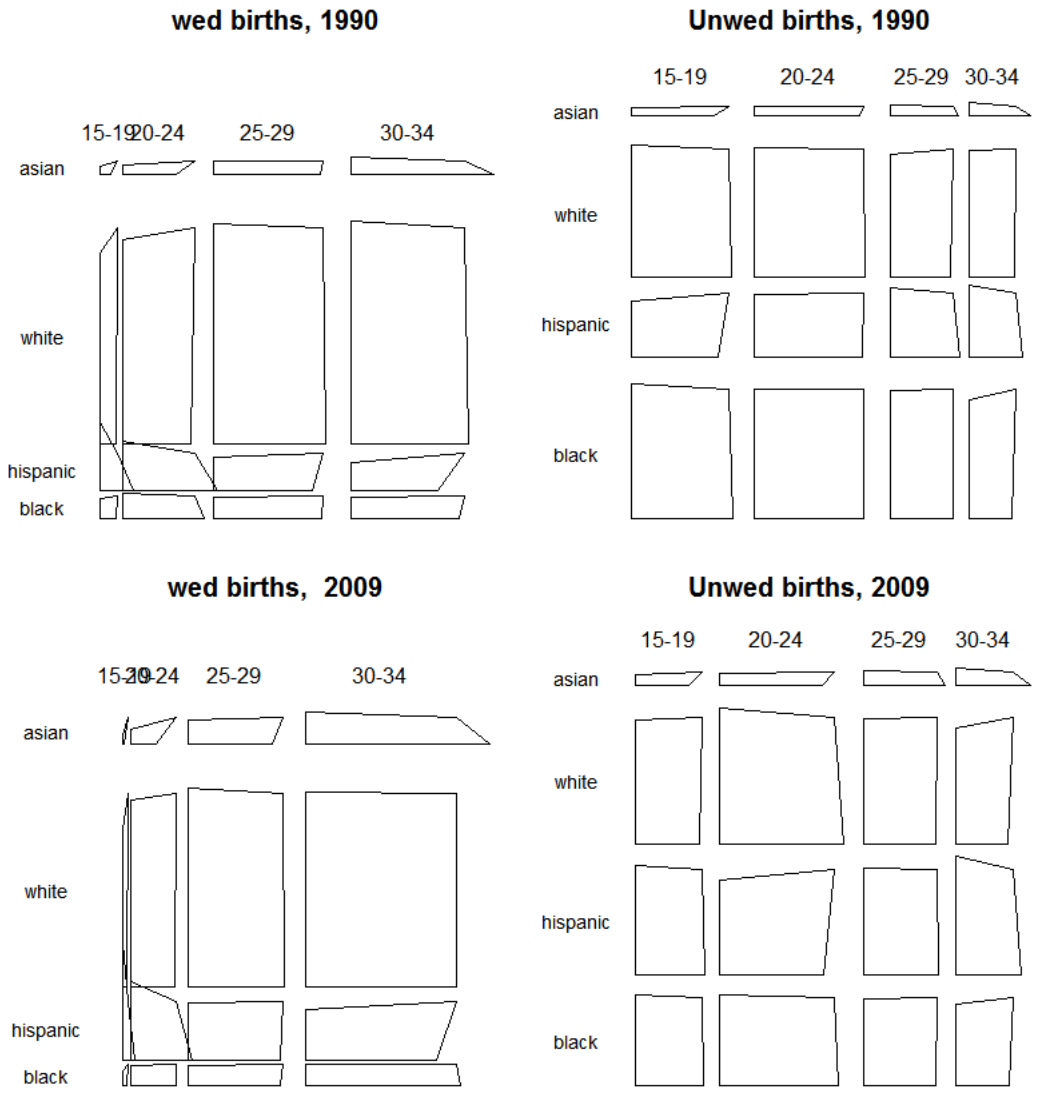
The second order F"s are small compared to the first order F's and the third order F"s are small compared to the second order F's. If we included the fourth order terms we would have perfect prediction. In the third order model, the residual deviance would be the contribution from the fourth order terms. In the second order model it is the sum of the contributions from third and fourth order terms. Maybe race:edu:year is worth taking notice of, implying that there is a changing pattern in the race by education effects over time. We have enough trouble already explaining the second order effects, so we will be content with the second order model.

11 Quad display for marriage status and year, by race and age

Quad displays are alternatives to mosaics for displaying 2 way tables. The advantage is that conditional probabilities of rows given columns, and of columns given rows are displayed, and are well aligned for comparisons. The disadvantage is that only two variables can be seen in any one plot. Each count is represented by a quadrilateral with area proportional to the count. The bottom left angle is a rightangle. The left vertical side is the conditional probability of row given column. The bottom horizontal side is the conditional probability of column given row. Quads with one acute angle have high probability compared to independence. Quads with two acute angles have low probability compare to independence.Rectangular quads conform to independence.

```
tiff("pictures/quads.tif", w=800, h=850)

par(mfrow=c(2,2))

quad(xtabs(wed ~ race + age,
data=bb[bb$year=="1990", ]), rlab=levels(bb$race),
clab=levels(bb$age), border=1, size=0.8)
title("wed births, 1990", cex.main=2)
quad(xtabs(unwed ~ race+age,
data=bb[bb$year=="1990", ]), rlab=levels(bb$race),
clab=levels(bb$age), border=1, size=0.8)
title("Unwed births, 1990", cex.main=2)

quad(xtabs(wed ~ race+age,data=bb[bb$year=="2009",
]),
rlab=levels(bb$race), clab=levels(bb$age),
border=1, size=0.8)
title("wed births,  2009", cex.main=2)

quad(xtabs(unwed ~ race+age,
data=bb[bb$year=="2009", ]),
rlab=levels(bb$race), clab=levels(bb$age),
border=1, size=0.8)
title("Unwed births, 2009", cex.main=2)
```

```
dev.off()
```



Note that the quadrilaterals are scaled differently in each of the four pictures, so that we can"t conclude, say, that whites aged 30-34 have lower unwed births than wed births in 2009.

Looking at wed births in 1990, notice the two acute angles for young Hispanics, and correspondingly high rates of wed births for young Hispanics. This pattern persists in 2009.. In the unwed births in 1990, the quads being nearly rectangular indicates no relationship between race and age in the unwed births.. the pattern is the same across all ages, the blacks and Hispanics tended to have large numbers of unwed births compared to whites, and compared to wed births. In 2009, the patterns of wed births are similar, except for an

increase in Hispanic and Asian births. The largest effect is in the increase of unwed births in the whites, again across all age groups.

14 Conclusions

Overall, between 1990 and 2090 there has been a remarkable increase in births to unwed mothers, especially in the less educated and younger age groups, across all racial categories. The ratio of unwed to wed births increased more for non-black groups.

15 Data Preparation

The site

http://www.cdc.gov/nchs/VitalStats.htm.

provides, after some tricky choices, tables of births by race, age, education of mother, and marital status, in 1990 and 2009. The tables have been saved into an excel file, and then saved in csv format in the two files:

"data/marriedbirths1990.csv"
"data/marriedbirths2009.csv"

15.1 Read the two tables

Read 1990 table:

```
b1 <- read.csv("data/marriedbirths1990.csv",
                  header=T, as.is=T)
print(head(b1[, -(1:2)]))
```

```
        X.1     X.2     X.3       X.4
1
2        MAR   Total Married Unmarried
3  MEDUC_REC
4      Total 4158212 2992828   1165384
5  0-8 years  246481  143981    102500
6 9-11 years  672563  293755    378808
```

The first four lines are headers not accessible to R. Skip them, put in new names:

```
b1 <- read.csv("data/marriedbirths1990.csv",
header=T,
                  as.is=T, skip=4)
names(b1)<-c("race", "age", "edu",
            "total", "wed", "unwed")
b1$year <- "1990"
print(head(b1), 13)
```

```
 race age                edu   total     wed  unwed year
1                  0-8 years  246481  143981 102500 1990
2                 9-11 years  672563  293755 378808 1990
3                   12 years 1479183 1043366 435817 1990
4                13-15 years  783275  647540 135735 1990
5          16 years and over  674043  640456  33587 1990
6                 Not stated   61038   38792  22246 1990
```

Read 2009 table:

```
b2 <- read.csv("data/marriedbirths2009.csv",
header=F,
                as.is=T, skip=4)
b2$year <- "2009"
names(b2) <- names(b1)
```

Combine the two tables:

```
b <- rbind(b1, b2)
```

15.3 Shape the data into R data frames:

Fill in some blank entries for race and age:

```
for (cols in 1:2){
for (rows in 2:dim(b)[1]){
if (b[rows, cols] == "")
  b[rows, cols] <- b[rows - 1, cols]
}
}
print(head(b))
```

```
 race age                edu   total     wed  unwed year
1                  0-8 years  246481  143981 102500 1990
2                 9-11 years  672563  293755 378808 1990
3                   12 years 1479183 1043366 435817 1990
4                13-15 years  783275  647540 135735 1990
5          16 years and over  674043  640456  33587 1990
6                 Not stated   61038   38792  22246 1990
```

Replace - by 0, make the counts numeric:

```
b[b == "-"] <- 0
b[, 4:6] <- as.numeric(unlist(b[, 4:6]))
```

```
print(head(b))
```

```
 race age                edu   total     wed  unwed year
1                  0-8 years  246481  143981 102500 1990
2                 9-11 years  672563  293755 378808 1990
3                   12 years 1479183 1043366 435817 1990
4                13-15 years  783275  647540 135735 1990
5          16 years and over  674043  640456  33587 1990
6                  Not stated   61038   38792  22246 1990
```

Fix Race: Consolidate racial and ethnic categories:

```
print(table(b$race))
```

```
                            Central or South American
                    71                            152
                Cuban                Hispanic total
                   152                            152
             Mexican             Non-Hispanic black
                   152                            152
Non-Hispanic other race             Non-Hispanic total
                   152                            152
      Non-Hispanic white             Not on certificate
                   152                             72
            Not stated Other and unknown Hispanic
                   152                            152
          Puerto Rican                          Total
                   152                             80
```

Make a single hispanic category; "other race" includes american indian and pacific islanders, but is mainly asian and will be so classified; remove all the other categories. This will handle entries in the table that remain blank, corresponding to various subtotals. These superfluous entries are entered as "Total", and will be removed eventually by replacing "Total" with NA.

```
b$race <- sub("Non-Hispanic ", "", b$race)
b$race <- sub("Hispanic total", "hispanic", b$race)
b$race <- sub("other race", "asian", b$race)
b$race[b$race != "black" & b$race != "white" &
b$race != "asian" & b$race != "hispanic"] <- "Total"
```

```
print(table(b$race))
```

```
  asian    black hispanic    Total    white
    152      152      152     1287      152
```

Fix Education:

```
print(table(b$edu))
```

```
       0-8 years          12 years       13-15 years
             237               237               237
16 years and over       9-11 years Not on certificate
             237               237               237
      Not stated             Total
             237               236
```

Combine "Not on certificate" with "Not stated"; then remove "Not stated":

```
use <- b$edu == "Not on certificate"
b[use,4:6] <- b[use, 4:6] + b[b$edu == "Not stated",
4:6]
b$edu[b$edu == "Not stated"] <- "Total"
```

Remove years and other extra words from b$edu:

```
b$edu <- sub(" years", "", b$edu)
b$edu <- sub(" and over", "+", b$edu)
b$edu <- sub("Not on certificate", "unknown", b$edu)
print(table(b$edu))
```

```
  0-8      12   13-15     16+    9-11   Total unknown
  237     237     237     237     237     473     237
```

Fix age, dropping 50-54, missing from 1990:

```
print(table(b$age))
```

```
          15-19    20-24    25-29    30-34    35-39
      7     200      200      200      200      200
  40-44    45-49    50-54    Total Under 15
    200      200       96      192      200
```

```
b$age <- sub("Under 15", "-15", b$age)
b$age[b$age == "50-54"] <- "Total"
```

Drop all totals:

```
b[b == "Total"] <- NA
```

Remove the total column, column 6:

```
b <- b[, c(1:3, 7, 5:6)]
b <- na.omit(b)
```

Consolidate the older age groups, tending to behave the same:

```
w30 <- which(b$age == "30-34")
print(w30)
```

```
[1]  25  26  27  28  29  30  73  74  75  76  77  78 121
122
[15] 123 124 125 126 169 170 171 172 173 174 217 218 219 220
[29] 221 222 265 266 267 268 269 270 313 314 315 316 317 318
[43] 361 362 363 364 365 366
```

```
print(which(b$age=="35-39"))
```

```
[1]  31  32  33  34  35  36  79  80  81  82  83  84 127
128
[15] 129 130 131 132 175 176 177 178 179 180 223 224 225 226
[29] 227 228 271 272 273 274 275 276 319 320 321 322 323 324
[43] 367 368 369 370 371 372
```

Add over ages at locations 0, 6, 12, and 18 higher than 30-34:

```
b$wed[w30] <- b$wed[w30] + b$wed[w30+6] +
              b$wed[w30+12] + b$wed[w30+18]
b$unwed[w30] <- b$unwed[w30] + b$unwed[w30+6] +
                b$unwed[w30+12] + b$unwed[w30+18]
```

Likewise, the two youngest ages:

```
b$wed[w30-18] <- b$wed[w30-18] + b$wed[w30-24]
b$unwed[w30-18] <- b$unwed[w30-18]+b$unwed[w30-24]
```

Now drop the extra ages:

```
b$wed[c(w30-24, w30+6, w30+12, w30+18)] <- NA
```

```
b <- na.omit(b)
print(table(b$age))
```

```
-19 20-24 25-29 30-34
   48    48    48    48
```

Save a copy of b for later binomial regressions:

```
bb <- b
```

Rearrange array so that marriage becomes a column:

```
b1 <- b[, -6]
b2 <- b[, -5]
names(b1) <- c("race", "age", "edu", "year",
"births")
names(b2) <- names(b1)

b <- rbind(b1, b2)
b$marriage <- c( rep("wed", nrow(b1)), rep("unwed",
nrow(b1)) )

b$marriage <- factor(b$marriage)
b <- b[, c(1:4, 6, 5)]
```

15.4 Save final tables

Show final birth table:

```
head(b,10)
```

```
       race   age     edu year marriage births
89  hispanic 15-19     0-8 1990      wed  10863
90  hispanic 15-19    9-11 1990      wed  18148
91  hispanic 15-19      12 1990      wed   8648
92  hispanic 15-19   13-15 1990      wed    939
93  hispanic 15-19     16+ 1990      wed      0
95  hispanic 15-19 unknown 1990      wed   1502
97  hispanic 20-24     0-8 1990      wed  25963
```

```
98  hispanic 20-24    9-11 1990      wed  28297
99  hispanic 20-24      12 1990      wed  39408
100 hispanic 20-24   13-15 1990      wed  13979
```

```
head(bb,10)
```

```
       race   age      edu year   wed unwed
89  hispanic 15-19     0-8 1990 10863 13637
90  hispanic 15-19    9-11 1990 18148 31985
91  hispanic 15-19      12 1990  8648 10421
92  hispanic 15-19   13-15 1990   939   968
93  hispanic 15-19     16+ 1990     0     0
95  hispanic 15-19 unknown 1990  1502  2920
97  hispanic 20-24     0-8 1990 25963 18065
98  hispanic 20-24    9-11 1990 28297 24625
99  hispanic 20-24      12 1990 39408 22065
100 hispanic 20-24   13-15 1990 13979  5886
```

We are pleasantly surprised to find no hispanic mothers age 15-19 with 16+ years of education!

Save clean versions:

```
write.csv(b,"data/Births19902009.csv",
row.names=F)
write.csv(bb,"data/bBirths19902009.csv",
row.names=F)
```

16 Functions

```
quad <- function(counts,
                 rlab="",clab="",size=0.7,
border=1){

# function draws a quadrilateral indicating
# conditional probs and count
# check validity, construct labels

if (!is.matrix(counts)) return(" counts not
matrix")
ylab <- deparse(substitute(rlab))
```

```
xlab <- deparse(substitute(clab))
if (rlab[1]=="") ylab <- "rows"
if (clab[1]=="") xlab <- "cols"
nrows <- dim(counts)[1]
ncols <- dim(counts)[2]
if (rlab[1]=="") rlab <- as.character(1:nrows)
if (clab[1]=="") clab <- as.character(1:ncols)

# initialize marginal counts
rcount <- apply(counts, 1, sum)
ccount <- apply(counts, 2, sum)
allsum <- sum(rcount)
rcount <- rcount/allsum
ccount <- ccount/allsum
counts <- counts/allsum
rcum <- cumsum(rcount)
ccum <- cumsum(ccount)

par(mar=c(1,1,1,1))
plot(c(-0.5*max(ccount), 1), c(0,
1+0.5*max(rcount) ),
   pch="", axes=F, xlab= "",  ylab="")

# go through all counts drawing quads
x <- 1:4
y <- x
for ( row in 1:nrows){
for ( col in 1:ncols){
  x[1] <- ccum[col]
  x[2] <- x[1]+counts[row,col]/rcount[row]-
          ccount[col]
  x[3] <- x[1]-ccount[col]
  x[4] <- x[3]

   y[3] <- 1-rcum[row]
   y[4] <- y[3]+counts[row,col]/ccount[col]
   y[1] <- y[3]+rcount[row]
   y[2] <- y[3]

# rescale all quads by  size, just big enough to
avoid overlap
   x <- x[3] + size*( x-x[3])
```

```
   y <- y[3] + size*( y-y[3])

   polygon(x,y,border=border)
}
}

# insert row and column labels
text(rep(-0.4*max(ccount),nrows),
   1+0.5*size*rcount-rcum,rlab[1:nrows], cex=1.4)
text(0.5*size*ccount+c(0,ccum[-ncols]),
   rep(1+0.2*max(ccount),ncols), clab, cex=1.6)

par(mar=c(5, 4, 4, 2)+.01)

invisible()
}

Grid <- function(xticks, yticks, ylab="",
      at=(min(xticks)+ mean(xticks))/2, cex=2.5){
# background for plot using grid of light grey lines
par(mar=c(3,3,6,2))

plot(1, 1,  xlim=range(xticks),  ylim =
range(yticks),
         xlab="", ylab="", axes=F, pch="")

# use only interior values of tick ranges in plots
usey <- rep( T, length(yticks) )
usey[c( 1, length(yticks) )] <- F
usex <- rep( T, length(xticks) )
usex[c( 1, length(xticks) )] <- F

# grey lines in both directions
for ( row in yticks[usey] )
lines(range(xticks), c(row, row), col="light grey")
for ( col in xticks[usex] )
lines(c(col, col), range(yticks), col="light grey")

# put ylab on left top, using / to split long
expressions
ylabs <- unlist(strsplit(ylab,"/"))
```

```
# identify tick marks on both axes
if (length(yticks) > 2)
text(pos=2, rep(min(xticks), length(yticks)-2 ),
     yticks[usey], yticks[usey], cex=2, xpd=T)
if (length(xticks)>2)
text(pos=1, xticks[usex],  rep(min(yticks),
    length(xticks)-2), xticks[usex], cex=2, xpd=T)
lylabs <- min(5, length(ylabs))
if(lylabs > 0){
mtext(ylabs, side=3, line =
(5/lylabs)*(lylabs-1):0,
      at = at, cex=cex)
}
par(mar=c(5, 4, 4, 2))

invisible()
}
```

```
glmZS <-  function(fdata, ...)
{
# fixes up labelling and missing terms in categorical
models to handle contrast sum
# fdata is a data matrix including all variables in
the regression, corresponding to data =

# use options contrast so that effects sum to zero
options(contrasts = c("contr.sum","contr.sum"))

data <- fdata
ncol <- dim(data)[2]

# pick out factors in data
fl <- rep(F, ncol)
for ( col in  1:ncol) fl[col] <- is.factor(data[,
col])

n <- sum(fl)
if( n==0) return(" no factors in data")

fl <- which(fl)

# run over 2^n choices of factor level patterns to
be omitted
binmat <- matrix(0, nrow =2^n, ncol=n)
for(i in 2:2^n) {
binmat[i, ] <- binmat[i-1, ]
for (j in 1:n) {
  if (binmat[i, j] == 0) {
    binmat[i, 1:j] <- 0
```

```
    binmat[i, j] <- 1


      break
}
}
}

# construct initial levels for factors
llevels <- list(1:n)
for (i in 1:n) {
llevels[[i]] <- levels(data[, fl[i]])
}

# define different factors for each pattern of
missing levels and iterate through each choice
for (iter in 1:2^n) {
for(i in 1:n) {
  nlevels <- length(llevels[[i]])

# first case return to original levels
  if(binmat[iter, i] == 0)
    data[, fl[i]] <-
    factor(data[, fl[i]],llevels[[i]])

# second case interchange last two levels
  if(binmat[iter,i] ==1 ){
    if(nlevels == 2)
      data[, fl[i]] <-
      factor(data[, fl[i]], llevels[[i]][2:1])
    if(nlevels > 2)
      data[, fl[i]] <- factor(data[, fl[i]],
      llevels[[i]][c(1:(nlevels-2), nlevels,
      nlevels-1)])
   }
}


# run regression with this choice of missing levels
 fn <- names(data)[fl]

 lm.f <- glm(data=data, ...)
```

```
 sc=summary(lm.f)$coef

if(sum(is.na(lm.f$coef)))
  return(" cant handle Na's in coef")

# get level names for these missing levels
for (i in 1:length(fn)) {
  levelnames <- levels(data[,fl[i]])
  nlevels <- length(levelnames)
  for (j in 1:nlevels ) {
    newname <-
    paste(fn[i], levelnames[j], sep = ".")
    oldname <-
    paste(fn[i], as.character(j), sep = "")

# substitute meaningful newname for obscure oldname
row.names(sc) <-
gsub(oldname,newname,row.names(sc))
  }
}

# combine all the lists of coefficients
if(iter == 1) coef <- sc
if(iter > 1){
  use=!row.names(sc) %in% row.names(coef)
  if(sum(use)>0){
    rn <- c(row.names(coef),row.names(sc)[use])
    coef <-  rbind(coef, sc[use,])
  row.names(coef)=rn

  }
}
}

# order by main variables
rn <- row.names(coef)
rn <- gsub(" ", "", rn)
rn1 <- rn[-1]
lv <- rep(0, length(rn1))

# pick out variable names before the dot, if a dot
```

```
# couldnt figure out how to use regexpr to find the
"."
for( i in 1:length(rn1)){
wheredot <-
which(unlist(strsplit(rn1[i],""))==".")
if(length(wheredot) == 0) lv[i] <- nchar(rn[i])+1
if(length(wheredot) > 0) lv[i] <- min(wheredot)
}

use <- c("", substr(rn1, 1, lv-1))
coef <- coef[order(use),]
rn <-  row.names(coef)

# make sure lower order interactions come first
low <- rep(0, length(rn))
for( i in 1:length(rn))
low[i]<-sum(unlist(strsplit(rn[i],split=""))=="."
)
coef <- coef[order(low),]

ss <- summary(lm.f)
ss$coef <- coef

return(ss)
}
```