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:

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

Looking at the data a little:

```r
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
```

```r
head(bb,10)
```
<table>
<thead>
<tr>
<th>race</th>
<th>age</th>
<th>edu</th>
<th>year</th>
<th>wed</th>
<th>unwed</th>
</tr>
</thead>
<tbody>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>0-8</td>
<td>1990</td>
<td>10863</td>
<td>13637</td>
</tr>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>9-11</td>
<td>1990</td>
<td>18148</td>
<td>31985</td>
</tr>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>12</td>
<td>1990</td>
<td>8648</td>
<td>10421</td>
</tr>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>13-15</td>
<td>1990</td>
<td>939</td>
<td>968</td>
</tr>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>16+</td>
<td>1990</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hispanic</td>
<td>15-19</td>
<td>unknown</td>
<td>1990</td>
<td>1502</td>
<td>2920</td>
</tr>
<tr>
<td>hispanic</td>
<td>20-24</td>
<td>0-8</td>
<td>1990</td>
<td>25963</td>
<td>18065</td>
</tr>
<tr>
<td>hispanic</td>
<td>20-24</td>
<td>9-11</td>
<td>1990</td>
<td>28297</td>
<td>24625</td>
</tr>
<tr>
<td>hispanic</td>
<td>20-24</td>
<td>12</td>
<td>1990</td>
<td>39408</td>
<td>22065</td>
</tr>
<tr>
<td>hispanic</td>
<td>20-24</td>
<td>13-15</td>
<td>1990</td>
<td>13979</td>
<td>5886</td>
</tr>
</tbody>
</table>

Convert character variables to factors:

```r
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:

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

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

```r
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.

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

```
    , , year = 1990
race
marriage    asian  white  hispanic  black
  unwed   35338   443012  218515  441634
  wed    133878  2183488  4680558  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:

$$\log\left(\frac{p}{1-p}\right) = \text{Linear function of numerical codings of race, age, education and year.}$$

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:

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

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

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

```r
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:
\[
\text{glmbb} \leftarrow \text{glm} \left( \text{cbind}(\text{unwed}+1, \text{wed}+1) \sim \right. \\
\left. \text{race} + \text{age} + \text{edu} + \text{year}, \text{family} = \text{binomial}, \text{data} = \text{bb} \right)
\]

Use zero sum constraints on the coefficients for each factor, to simplify interpretation of the coefficients:
\[
\text{coef} \leftarrow \text{glmZS} \left( \text{fdata} = \text{bb}, \text{cbind}(\text{unwed}+1, \text{wed}+1) \sim \right. \\
\left. \text{race} + \text{age} + \text{edu} + \text{year}, \text{family} = \text{binomial} \right) \text{coef}
\]
\[
\text{coef} \leftarrow \text{data.frame}(\text{coef}[\text{, c(1,3)}])
\]
\[
\text{coef}$\text{relrate} \leftarrow \exp(\text{coef}[\text{, 1}])
\]
\[
\text{round(coef,2)}
\]

<table>
<thead>
<tr>
<th>Estimate</th>
<th>z.value</th>
<th>relrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.27</td>
<td>-173.44</td>
</tr>
<tr>
<td>age.15-19</td>
<td>1.37</td>
<td>636.36</td>
</tr>
<tr>
<td>age.20-24</td>
<td>0.30</td>
<td>207.63</td>
</tr>
<tr>
<td>age.25-29</td>
<td>-0.59</td>
<td>-390.92</td>
</tr>
<tr>
<td>age.30-34</td>
<td>-1.08</td>
<td>-671.12</td>
</tr>
<tr>
<td>edu.0-8</td>
<td>0.46</td>
<td>123.33</td>
</tr>
<tr>
<td>edu.9-11</td>
<td>0.77</td>
<td>296.40</td>
</tr>
<tr>
<td>edu.12</td>
<td>0.22</td>
<td>109.79</td>
</tr>
<tr>
<td>edu.13-15</td>
<td>-0.18</td>
<td>-73.43</td>
</tr>
<tr>
<td>edu.16+</td>
<td>-1.23</td>
<td>-335.02</td>
</tr>
<tr>
<td>edu.unknown</td>
<td>-0.04</td>
<td>-17.80</td>
</tr>
<tr>
<td>race.asian</td>
<td>-0.64</td>
<td>-210.74</td>
</tr>
<tr>
<td>race.white</td>
<td>-0.72</td>
<td>-470.64</td>
</tr>
<tr>
<td>race.hispanic</td>
<td>0.07</td>
<td>37.07</td>
</tr>
<tr>
<td>race.black</td>
<td>1.29</td>
<td>644.04</td>
</tr>
<tr>
<td>year.1990</td>
<td>-0.52</td>
<td>-421.60</td>
</tr>
<tr>
<td>year.2009</td>
<td>0.52</td>
<td>421.60</td>
</tr>
</tbody>
</table>

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 0.59 for 1990 and 1.7 for 2009 means that the unwed/wed birth ratio increased by the factor of 3 = 1.7/0.59 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 ≈ 3.64/0.49).
6 Binomial Plots for main effects model

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

```r
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(glmmbb$fit, bb$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$unwedp ~glmmbb$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

\[ \text{glm} \left( \text{cbind}(\text{bb}$unwed$, \text{bb}$wed)$\sim (\text{race} + \text{age} + \text{edu} + \text{year})^2, \text{data}=\text{bb}, \text{family}=\text{binomial} \right) \]

Include this function to give simpler anova display:
\[ \text{FixDeviance}\left\{ \begin{array}{c} \text{a}[,3] <- (\text{a}[,2]/\text{a}[,1])/(\text{a}[\text{nrow}(\text{a}),4]/\text{a}[\text{nrow}(\text{a}),3]) \\
\text{names}(\text{a})[3] <- "F" \\
\text{return}(\text{a}[,,-4]) \\
\end{array} \right. \]

\[ \text{print( round(FixDeviance(anova(glmbb))) ) } \]

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>3</td>
<td>1125123</td>
</tr>
<tr>
<td>age</td>
<td>3</td>
<td>1102002</td>
</tr>
<tr>
<td>edu</td>
<td>5</td>
<td>247211</td>
</tr>
<tr>
<td>year</td>
<td>1</td>
<td>185230</td>
</tr>
<tr>
<td>race:age</td>
<td>9</td>
<td>27676</td>
</tr>
<tr>
<td>race:edu</td>
<td>15</td>
<td>13061</td>
</tr>
<tr>
<td>race:year</td>
<td>3</td>
<td>4974</td>
</tr>
<tr>
<td>age:edu</td>
<td>14</td>
<td>54535</td>
</tr>
<tr>
<td>age:year</td>
<td>3</td>
<td>8574</td>
</tr>
<tr>
<td>edu:year</td>
<td>5</td>
<td>9324</td>
</tr>
</tbody>
</table>

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(glmmb$fit, bb$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$unwedp~glmmb$fit,w=bb$total),lwd=3)

dev.off()
Compute the correlation between fit and proportion:

```r
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.
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.

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

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

<table>
<thead>
<tr>
<th>Estimate</th>
<th>z.value</th>
<th>relrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>race.asian</td>
<td>-0.55</td>
<td>-116.34</td>
</tr>
<tr>
<td>race.white</td>
<td>-0.71</td>
<td>-284.50</td>
</tr>
<tr>
<td>race.black</td>
<td>1.31</td>
<td>372.81</td>
</tr>
<tr>
<td>year.1990</td>
<td>-0.45</td>
<td>-222.36</td>
</tr>
<tr>
<td>year.2009</td>
<td>0.45</td>
<td>222.36</td>
</tr>
<tr>
<td>age.30-34:year.1990</td>
<td>0.19</td>
<td>87.46</td>
</tr>
<tr>
<td>age.30-34:year.2009</td>
<td>-0.19</td>
<td>-87.46</td>
</tr>
<tr>
<td>edu.12:year.1990</td>
<td>-0.12</td>
<td>-50.04</td>
</tr>
<tr>
<td>edu.unknown:year.1990</td>
<td>0.18</td>
<td>68.06</td>
</tr>
<tr>
<td>edu.12:year.2009</td>
<td>0.12</td>
<td>50.04</td>
</tr>
<tr>
<td>edu.unknown:year.2009</td>
<td>-0.18</td>
<td>-68.06</td>
</tr>
<tr>
<td>race.hispanic:age.15-19</td>
<td>-0.42</td>
<td>-87.60</td>
</tr>
<tr>
<td>race.hispanic:age.25-29</td>
<td>0.18</td>
<td>55.45</td>
</tr>
<tr>
<td>race.white:year.1990</td>
<td>-0.12</td>
<td>-54.73</td>
</tr>
<tr>
<td>race.hispanic:age.30-34</td>
<td>0.38</td>
<td>117.07</td>
</tr>
<tr>
<td>race.white:year.2009</td>
<td>0.12</td>
<td>54.73</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>3</td>
<td>1125123</td>
</tr>
<tr>
<td>age</td>
<td>3</td>
<td>1102002</td>
</tr>
<tr>
<td>edu</td>
<td>5</td>
<td>247211</td>
</tr>
<tr>
<td>year</td>
<td>1</td>
<td>185230</td>
</tr>
<tr>
<td>race:age</td>
<td>9</td>
<td>27676</td>
</tr>
<tr>
<td>race:edu</td>
<td>15</td>
<td>13061</td>
</tr>
<tr>
<td>race:year</td>
<td>3</td>
<td>4974</td>
</tr>
<tr>
<td>age:edu</td>
<td>14</td>
<td>54535</td>
</tr>
<tr>
<td>age:year</td>
<td>3</td>
<td>8574</td>
</tr>
<tr>
<td>edu:year</td>
<td>5</td>
<td>9324</td>
</tr>
<tr>
<td>race:age:edu</td>
<td>42</td>
<td>6260</td>
</tr>
<tr>
<td>race:age:year</td>
<td>9</td>
<td>905</td>
</tr>
<tr>
<td>race:edu:year</td>
<td>15</td>
<td>1706</td>
</tr>
<tr>
<td>age:edu:year</td>
<td>14</td>
<td>996</td>
</tr>
</tbody>
</table>

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.
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 compared 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)
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


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:

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

<table>
<thead>
<tr>
<th>X.1</th>
<th>X.2</th>
<th>X.3</th>
<th>X.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MAR</td>
<td>Total</td>
<td>Married</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Meduc_rec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Total</td>
<td>4158212</td>
<td>2992828</td>
</tr>
<tr>
<td>5</td>
<td>0-8 years</td>
<td>246481</td>
<td>143981</td>
</tr>
<tr>
<td>6</td>
<td>9-11 years</td>
<td>672563</td>
<td>293755</td>
</tr>
</tbody>
</table>

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

```r
b1 <- read.csv("data/marriedbirths1990.csv", header=T, as.is=T, skip=4)
names(b1)<-c("race", "age", "edu", "total", "wed", "unwed")
b1$year <- "1990"
print(head(b1), 13)
```
Read 2009 table:
`b2 <- read.csv("data/marriedbirths2009.csv", header=F, as.is=T, skip=4)`
`b2$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:
```r
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))
```

Replace - by 0, make the counts numeric:
```r
b[b == "-"] <- 0
b[, 4:6] <- as.numeric(unlist(b[, 4:6]))
```
print(head(b))

<table>
<thead>
<tr>
<th>race age</th>
<th>edu</th>
<th>total</th>
<th>wed</th>
<th>unwed</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-8 years</td>
<td>246481</td>
<td>143981</td>
<td>102500</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>9-11 years</td>
<td>672563</td>
<td>293755</td>
<td>378808</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>12 years</td>
<td>1479183</td>
<td>1043366</td>
<td>435817</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>13-15 years</td>
<td>783275</td>
<td>647540</td>
<td>135735</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>16 years and over</td>
<td>674043</td>
<td>640456</td>
<td>33587</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>Not stated</td>
<td>61038</td>
<td>38792</td>
<td>22246</td>
<td>1990</td>
<td></td>
</tr>
</tbody>
</table>

Fix Race: Consolidate racial and ethnic categories:
print(table(b$race))

<table>
<thead>
<tr>
<th>Central or South American</th>
<th>152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuban</td>
<td>152</td>
</tr>
<tr>
<td>Mexican</td>
<td>152</td>
</tr>
<tr>
<td>Non-Hispanic other race</td>
<td>152</td>
</tr>
<tr>
<td>Non-Hispanic white</td>
<td>152</td>
</tr>
<tr>
<td>Not stated Other and unknown Hispanic</td>
<td>152</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>152</td>
</tr>
<tr>
<td>Non-Hispanic total</td>
<td>152</td>
</tr>
<tr>
<td>Not on certificate</td>
<td>72</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

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$race != "hispanic"] <- "Total"
```r
print(table(b$race))

<table>
<thead>
<tr>
<th></th>
<th>asian</th>
<th>black</th>
<th>hispanic</th>
<th>Total</th>
<th>white</th>
</tr>
</thead>
<tbody>
<tr>
<td>152</td>
<td>152</td>
<td>152</td>
<td></td>
<td>1287</td>
<td>152</td>
</tr>
</tbody>
</table>
```

Fix Education:
```r
print(table(b$edu))

<table>
<thead>
<tr>
<th></th>
<th>0-8 years</th>
<th>12 years</th>
<th>13-15 years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>237</td>
<td>237</td>
<td>237</td>
</tr>
<tr>
<td>16 years and over</td>
<td>237</td>
<td>237</td>
<td>237</td>
</tr>
<tr>
<td>Not stated</td>
<td>237</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>237</td>
<td>237</td>
<td></td>
</tr>
</tbody>
</table>
```

Combine "Not on certificate" with "Not stated"; then remove "Not stated":
```r
use <- b$edu == "Not on certificate"
b$edu[b$edu == "Not stated"] <- "Total"
```

Remove years and other extra words from b$edu:
```r
b$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))

<table>
<thead>
<tr>
<th></th>
<th>0-8</th>
<th>12</th>
<th>13-15</th>
<th>16+</th>
<th>9-11</th>
<th>Total</th>
<th>unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>237</td>
<td>237</td>
<td>237</td>
<td>237</td>
<td>237</td>
<td>473</td>
<td>237</td>
</tr>
</tbody>
</table>
```

Fix age, dropping 50-54, missing from 1990:
```r
print(table(b$age))

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>192</td>
<td>200</td>
</tr>
</tbody>
</table>
```
\begin{verbatim}
# Set up some example data
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
[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
[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$wed[w30+12] + b$wed[w30+18]
                    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$wed[c(w30-24, w30+6, w30+12, w30+18)] <- NA
\end{verbatim}
```r
b <- na.omit(b)
print(table(b$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)

<table>
<thead>
<tr>
<th>race</th>
<th>age</th>
<th>edu</th>
<th>year</th>
<th>marriage</th>
<th>births</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>hispanic</td>
<td>15-19</td>
<td>0-8</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>90</td>
<td>hispanic</td>
<td>15-19</td>
<td>9-11</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>91</td>
<td>hispanic</td>
<td>15-19</td>
<td>12</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>92</td>
<td>hispanic</td>
<td>15-19</td>
<td>13-15</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>93</td>
<td>hispanic</td>
<td>15-19</td>
<td>16+</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>95</td>
<td>hispanic</td>
<td>15-19</td>
<td>unknown</td>
<td>1990</td>
<td>wed</td>
</tr>
<tr>
<td>97</td>
<td>hispanic</td>
<td>20-24</td>
<td>0-8</td>
<td>1990</td>
<td>wed</td>
</tr>
</tbody>
</table>
```
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(counts), 1), c(0, 1+0.5*max(counts) ),
     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]
               ccount[col]
    x[3] <- x[1]-ccount[col]

    y[3] <- 1-rcum[row]
  }
}

# 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
# couldn't 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)