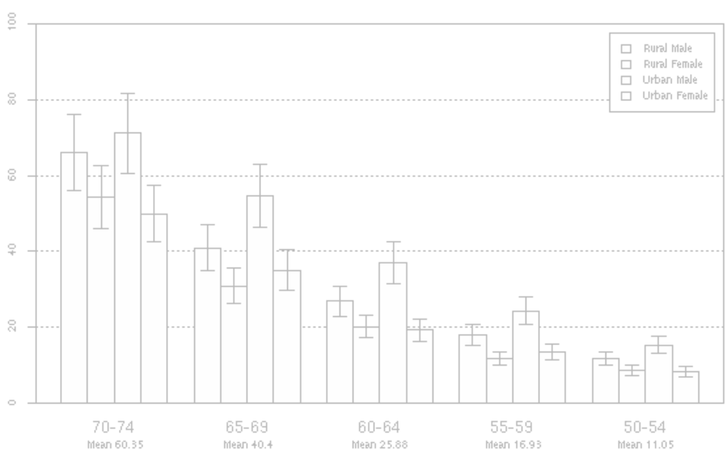
It's a Boy!
An Analysis of 96 Million Birth Records Using R



Susan I. Ranney

UseR! 2011

Overview

- Objective
- The U.S. birth records data sets
- Importing and cleaning the data
- Visualization of the percentage of boys born by a variety of factors
- Logistic regression controlling for many factors

REVOLUTION ANALYTICS

Objective: Use a "typical" big, public data set for analysis in R using RevoScaleR

- Lots of observations
- Typical issues of data collected over time
- Appropriate for basic "life cycle of analysis"
  - Import data
  - Check and clean data
  - Basic variable summaries
  - Big data logistic/linear regression

United States Birth Records
1985 - 2008

Revolution Analytics

The U.S. Birth Data

- Public-use data sets containing information on all births in the United States for each year from 1985 to 2008 are available to download: http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm
- *"These natality files are gigantic; they're approximately 3.1 GB uncompressed. That's a little larger than R can easily process"* – Joseph Adler**, R in a Nutshell**

The U.S. Birth Data (continued)

- Data for each year are contained in a compressed, fixed-format, text files
- Typically 3 to 4 million records per file
- Variables and structure of the files sometimes change from year to year, with birth certificates revised in1989 and 2003. Warnings:

  NOTE: THE RECORD LAYOUT OF THIS FILE HAS CHANGED SUBSTANTIALLY. USERS SHOULD READ THIS DOCUMENT CAREFULLY.

The Challenge

Import variables for all observations for all 24 different years into one common file for use in R

Application: More Boys than Girls Born

For the years 1985 – 2008 in the United States:

- 96,534,205 babies born
- 51.2% were boys
- 2,293,649 "excess" boys
- Slight downward trend in % boys observed in the U.S. since 1971

Examine Factors Associated with the Sex Ratio at Birth

There is a substantial literature on the sex ratio at birth. Basic demographic factors considered include:

- Age of mother and father
- Race/ethnicity of mother and father
- Birth order
- Multiple births
- Gestational age

Importing the U.S. Birth Data for Use in R

Importing Data: Basic Strategy

- Use *RevoScaleR*'s `rxImport` function to import specific variables for each year into an *.xdf* file, applying data transformations
- Append to a "master" file for all years with common variables

Note: `rxImport` also supports delimited text, SPSS, SAS, and ODBC import

Importing the Data for Each Year

`rxImport` takes a `colInfo` list as an argument. For each variable you want to import from a fixed format file, specify:

- Type of data (e.g., factor, integer, numeric
- Starting position
- Width (number of characters)
- Variable description
- For factors, levels in the data and, if desired, new levels to use after importing

REVOLUTION ANALYTICS

Example of Differences for Different Years

To create common variables across years, use common names and new factor levels. For example:

- For 1985:

```
SEX = list(type="factor", start=35, width=1,
    levels=c("1", "2"),
    newLevels = c("Male", "Female"),
    description = "Sex of Infant“)
```

- For 2003:

```
SEX = list(type="factor", start=436, width=1,
    levels=c("M", "F"),
    newLevels = c("Male", "Female"),
    description = "Sex of Infant”)
```

REVOLUTION ANALYTICS

Using Transformations to Create New Variables

In RevoScaleR, you can use a list of R "transforms" at various stages of your analysis.

- `rxImport`: for importing data to .xdf or data frame
- `rxDataStep`: for subselecting and transforming data
- `rxSummary, rxCube, rxLinMod, rxLogit`, etc.: for "on-the-fly" data transformations

Creating Transformed Variables on Import

Use a list of R "transforms" when importing each year.

- For example, create a factor for Live Birth Order using the imported LBO_REC integer variable:

```
LBO4 = cut(LBO_REC, breaks=c(0:3, 8, 9),
       labels = c("1", "2", "3",
       "4 or more", "NA"))
```

- Create binary variable for "It's a boy"

```
ItsaBoy = SEX == 'Male'
```

- Create variables with all "missings" for data not collected for that year (e.g., Hispanic ethnicity before 1989)

REVOLUTION ANALYTICS

Steps for Complete Import

- Lists for column information and transforms are created for 3 base years: 1985, 1989, 2003 when there were very large changes in the structure of the input files
- Changes to these lists are made where appropriate for in-between years
- A test script is run, importing only 1000 observations per year for a subset of years
- Full script is run, importing each year and appending to a master .xdf file

Examining and Cleaning the Big Data File

Examining Basic Information

- Basic file information

```
>rxGetInfo(birthAll)
File name: C:\Revolution\Data\CDC\BirthUS.xdf
Number of observations: 96534205
Number of variables: 43
Number of blocks: 206
```

- Use rxSummary to compute summary statistics for continuous data and category counts for each of the factors (under 4 minutes on my laptop)

```
rxSummary(~., data=birthAll, blocksPerRead = 10)
```

REVOLUTION ANALYTICS

Example of Summary Statistics

MomAgeR7	Counts
Under 20	11503530
20-24	24968335
25-29	27532554
30-34	21384197
35-39	9313706
Over 39	1831883
Missing	0

DadAgeR8	Counts
Under 20	3089189
20-24	14715683
25-29	22877854
30-34	22237430
35-39	12689871
40-44	4736792
Over 44	2018450
Missing	14168936

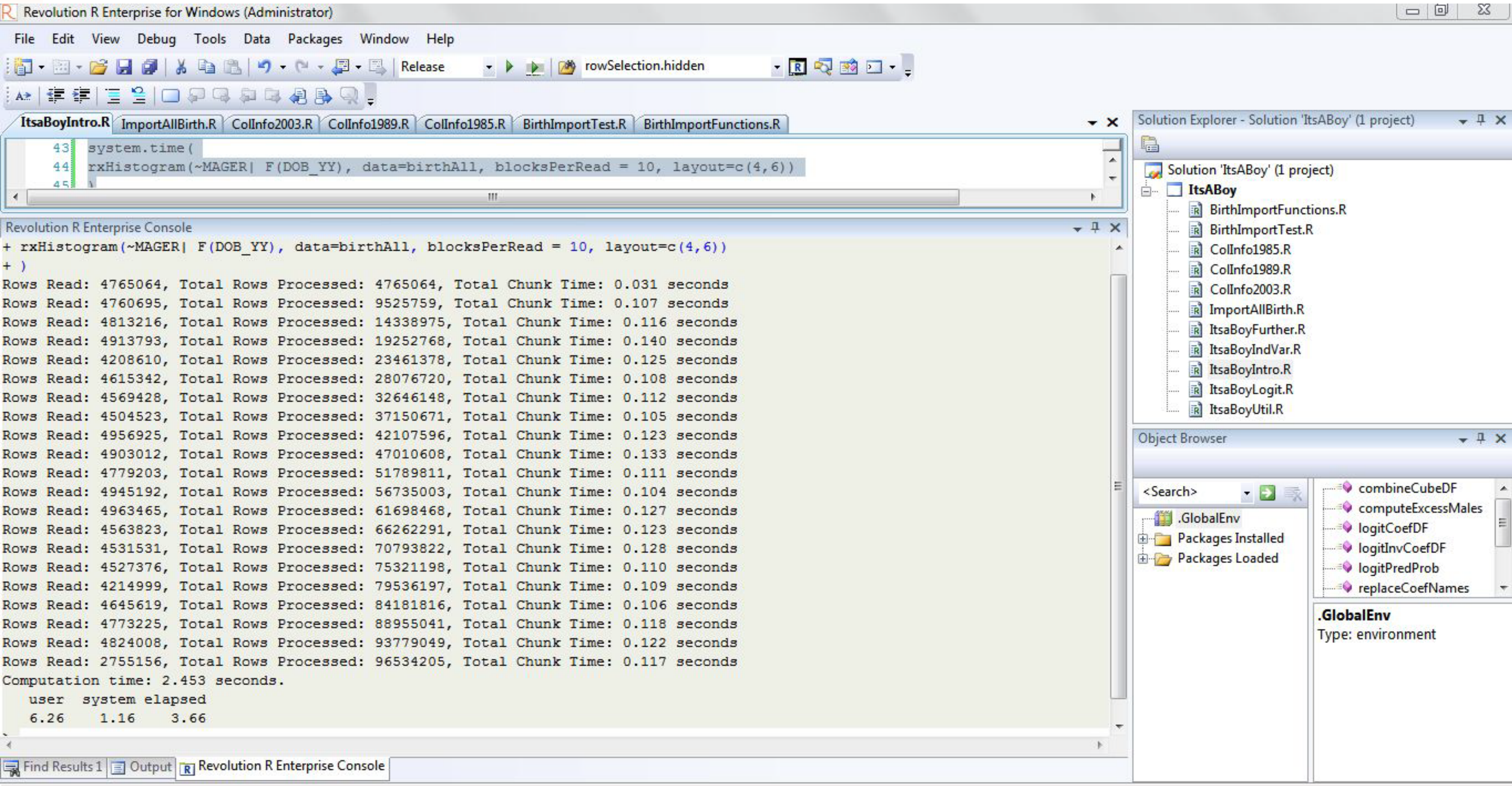
Histograms by Year

Easily check for basic errors in data import (e.g. wrong position in file) by creating histograms by year – very fast (just seconds on my laptop)

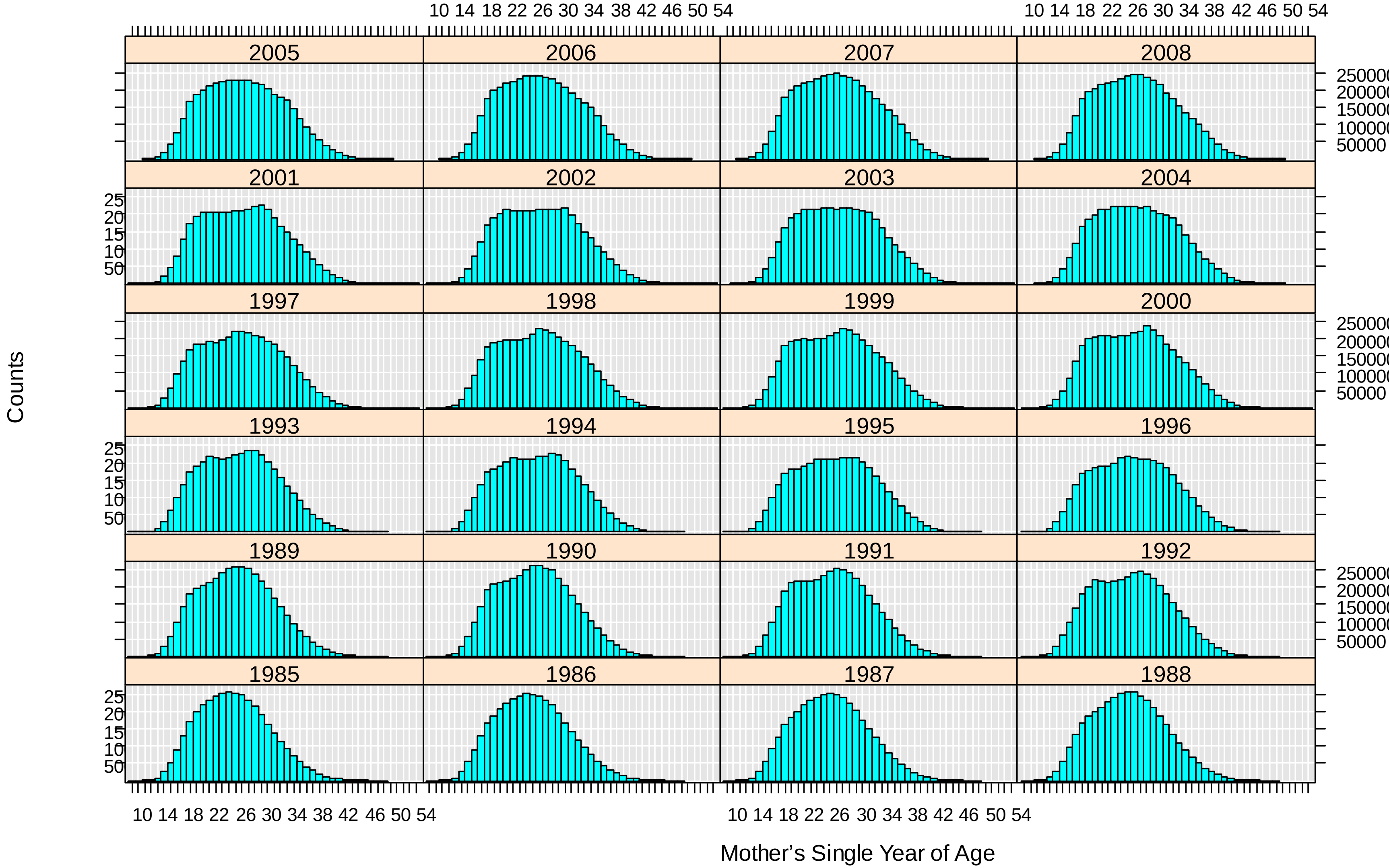
- Example: Distribution of mother's age by year. Use F() to have the integer year treated as a factor.

```
rxHistogram(~MAGER| F(DOB_YY),
  data=birthAll, blocksPerRead = 10,
  layout=c(4,6))
```

Screenshot from RPE



Age of Mother Over Time



REVOLUTION ANALYTICS

Drill Down and Extract Subsamples

- Take a quick look at "older" fathers:

```
rxSummary(~F(UFAGECOMB),
  data=birthAll,
  blocksPerRead = 10)
```

- What's going on with 89-year old Dads? Extract a data frame:

```
dad89 <- rxDataStep(
  inData = birthAll,
  rowSelection = UFAGECOMB == 89,
  varsToKeep = c("DOB_YY", "MAGER",
  "MAR", "STATENAT", "FRACEREC"),
  blocksPerRead = 10)
```

Dad's Age	Counts
80	139
81	103
82	78
83	71
84	54
85	43
86	42
87	26
88	26
89	3327

REVOLUTION ANALYTICS

Year and State for 89-Year-Old Fathers

```
rxCube(~F(DOB_YY):STATENAT,
  data=dad89, removeZeroCounts=TRUE)
```

F_DOB_YY	STATENAT	Counts
1990	California	1
1999	California	1
2000	California	1
1996	Hawaii	1
1997	Louisiana	1
1986	New Jersey	1
1995	New Jersey	1
1996	Ohio	1
1989	Texas	3316
1990	Texas	1
2001	Texas	1
1985	Washington	1

REVOLUTION ANALYTICS

89-Year-Old Fathers in Texas in 1989: Race and Mother's Age & Marital Status

```
dadTexas89 <- subset(dad89,
    STATENAT == 'Texas' & DOB_YY == 1989)
```

```
>head(dadTexas89)
   DOB_YY MAGER MAR STATENAT            FRACEREC
3    1989    23  No    Texas Unknown or not stated
4    1989    17  No    Texas Unknown or not stated
5    1989    16  No    Texas Unknown or not stated
6    1989    23  No    Texas Unknown or not stated
7    1989    16  No    Texas Unknown or not stated
8    1989    26  No    Texas Unknown or not stated
> tail(dadTexas89)
      DOB_YY MAGER MAR STATENAT            FRACEREC
3313    1989    18  No    Texas Unknown or not stated
3314    1989    21  No    Texas Unknown or not stated
3315    1989    30  No    Texas Unknown or not stated
3316    1989    18  No    Texas Unknown or not stated
3317    1989    18  No    Texas Unknown or not stated
3318    1989    24  No    Texas Unknown or not stated
```

REVOLUTION ANALYTICS

Strategy for Handling Suspicious Data

Use `transforms` to create factor variables.

When creating a factor variable for Dad's age, put all ages 89 and over in the "missing" category.

```
DadAgeR8 = cut(DadAgeR8, breaks =
c(0, 19, 24, 29, 34, 39, 44, 88, 99 ),
labels = c("Under 20", "20-24", "25-29",
  "30-34", "35-39", "40-44", "Over 44",
  "Missing"))
```

REVOLUTION ANALYTICS

Basic Computations Using Full Data Set: Percent Baby Boys by Basic Demographic Characteristics

Parental Age

- Contradictory results are found in the literature for the effects of paternal age, maternal age, and birth order *[Jacobsen 1999]*
- Unprecedented increases in births to older mothers in U. S. during 1981 – 2006 *[Branum 2009]*

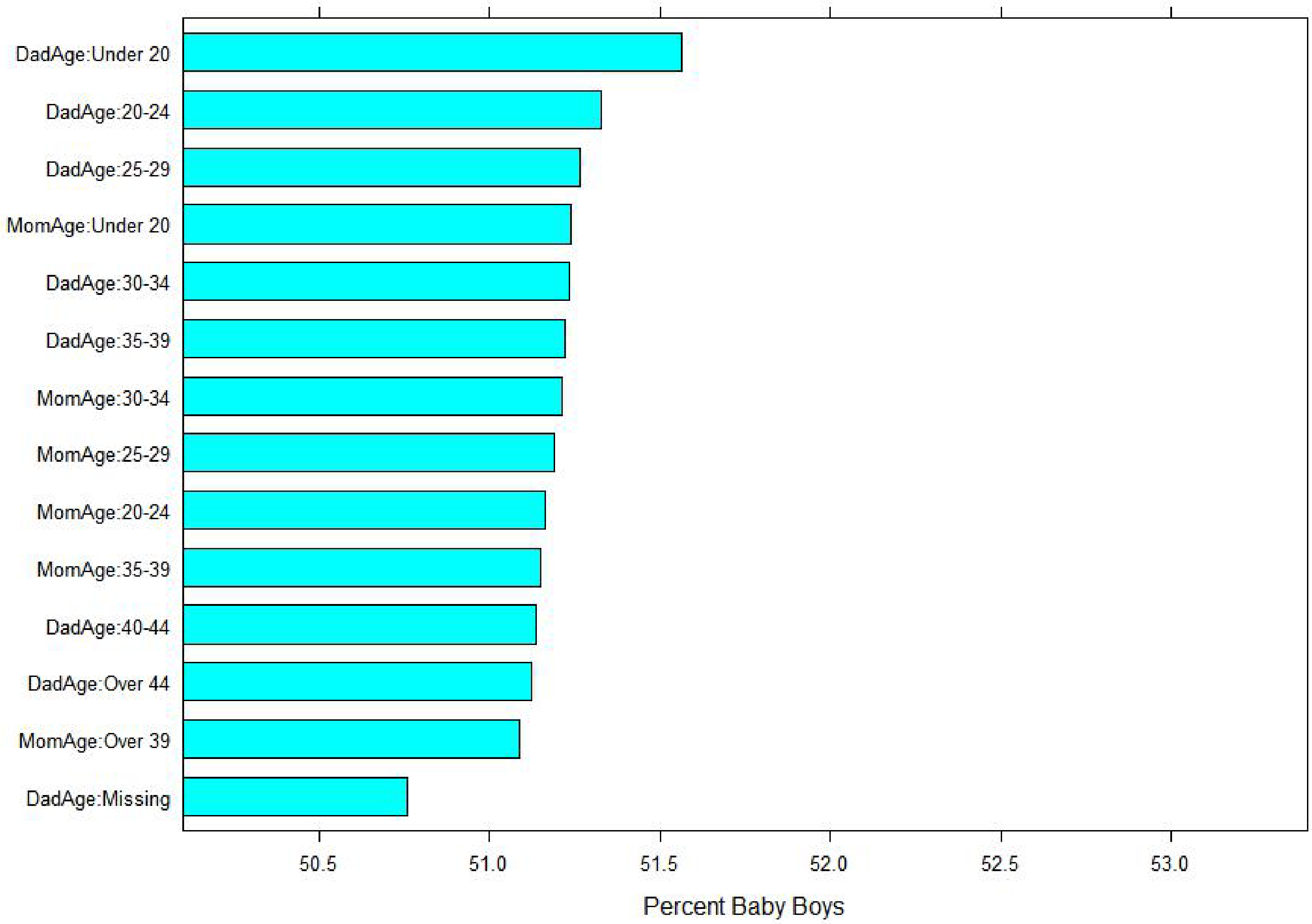
Percent Baby Boy by Parental Age

- Use rxCube to extract percentages by group for both mother's and father's age (independently)

```
rxCube(ItsaBoy~MomAgeR7, data=birthAll,
  blocksPerRead = 10)
rxCube(ItsaBoy~DadAgeR8, data=birthAll,
  blocksPerRead = 10)
```

- Combine results, sort, and plot. For comparison with other factors, fix the x-axis range to include (50.1, 53.4)

Percent Baby Boy by Parental Age



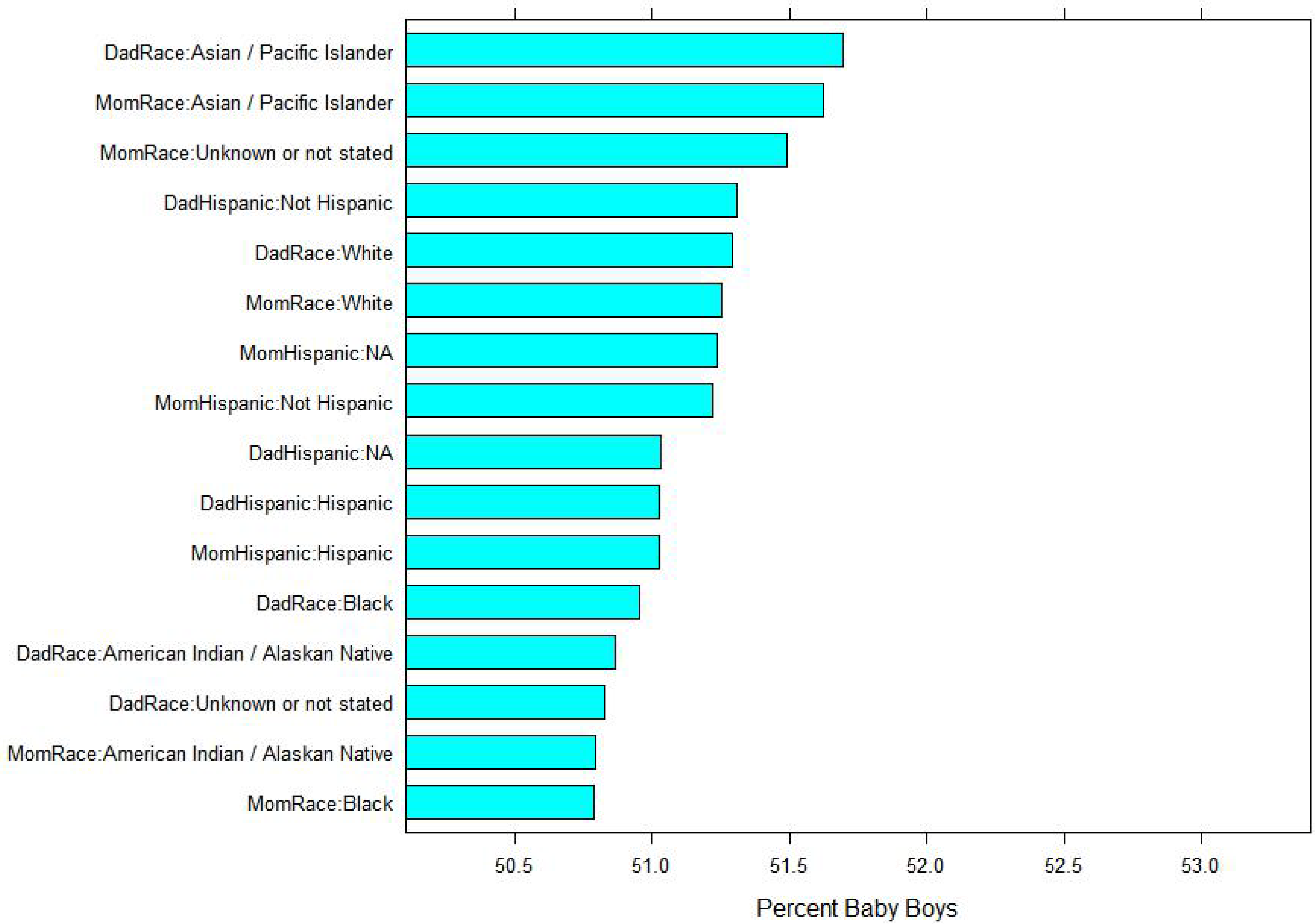
Summary of Parental Age

- Highest percentage of boys for young fathers (51.565%)
- Lowest percentage of boys for fathers of unknown age (50.759%) and older mothers (51.087%)

Actual Excess Boys (Percent = 51.188 %)	2,293,649
Excess Boys (Old Mothers (51.087%))	2,097,803
Excess Boys (Young Fathers (51.565%))	3,020.814

REVOLUTION ANALYTICS

Race and Ethnicity of Parents



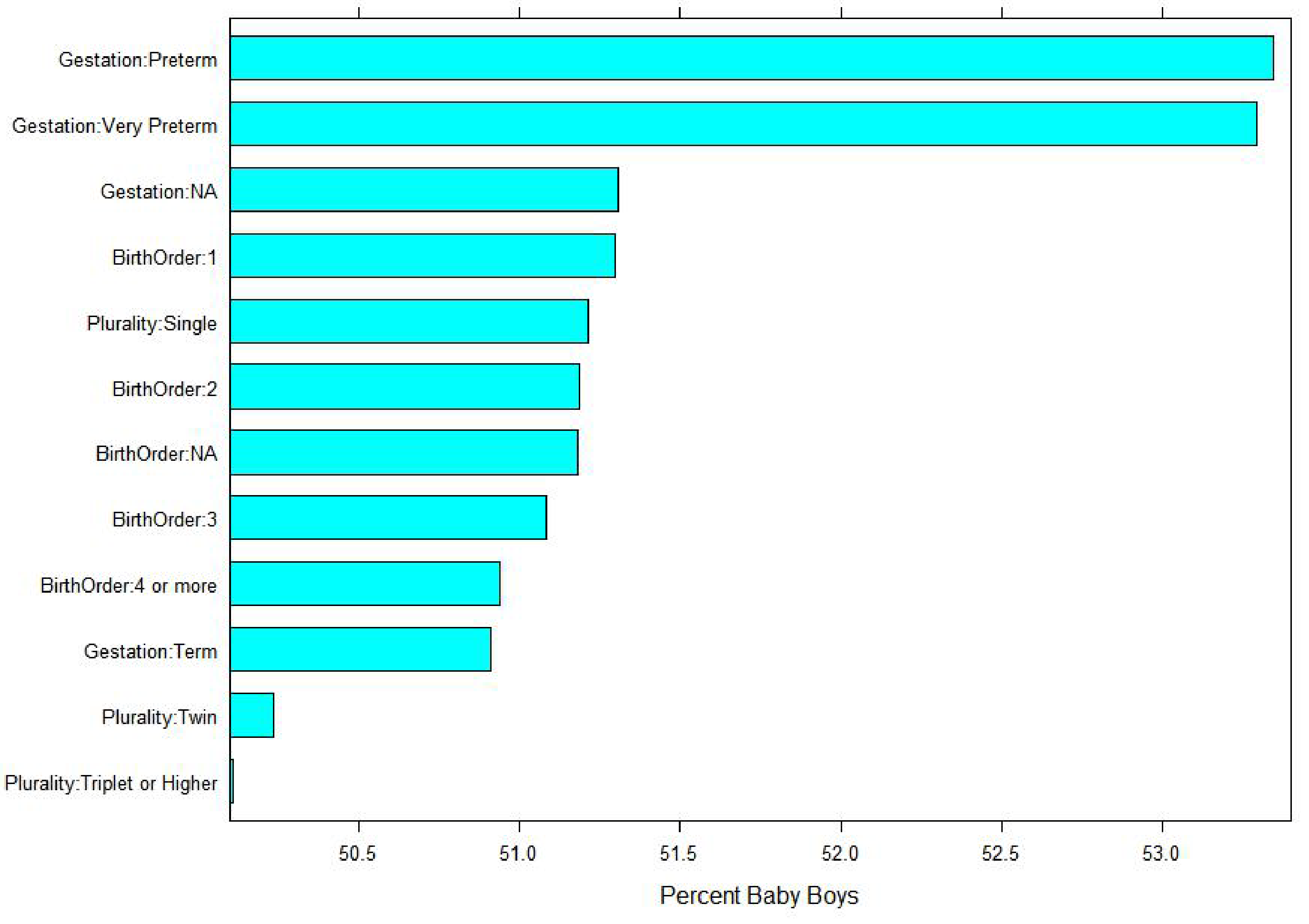
Summary of Race/Ethnicity

- Highest percentage of boys for Asian fathers (51.697%)
- Lowest percentage of boys for Black mothers (50.789%)

Actual Excess Boys (Percent = 51.188 %)	2,293,649
Excess Boys (Black Moms (50.789%))	1,522,559
Excess Boys (Asian Dads (51.697%))	3,276,349

Revolution Analytics

Birth Order, Plurality, and Gestation



REVOLUTION ANALYTICS

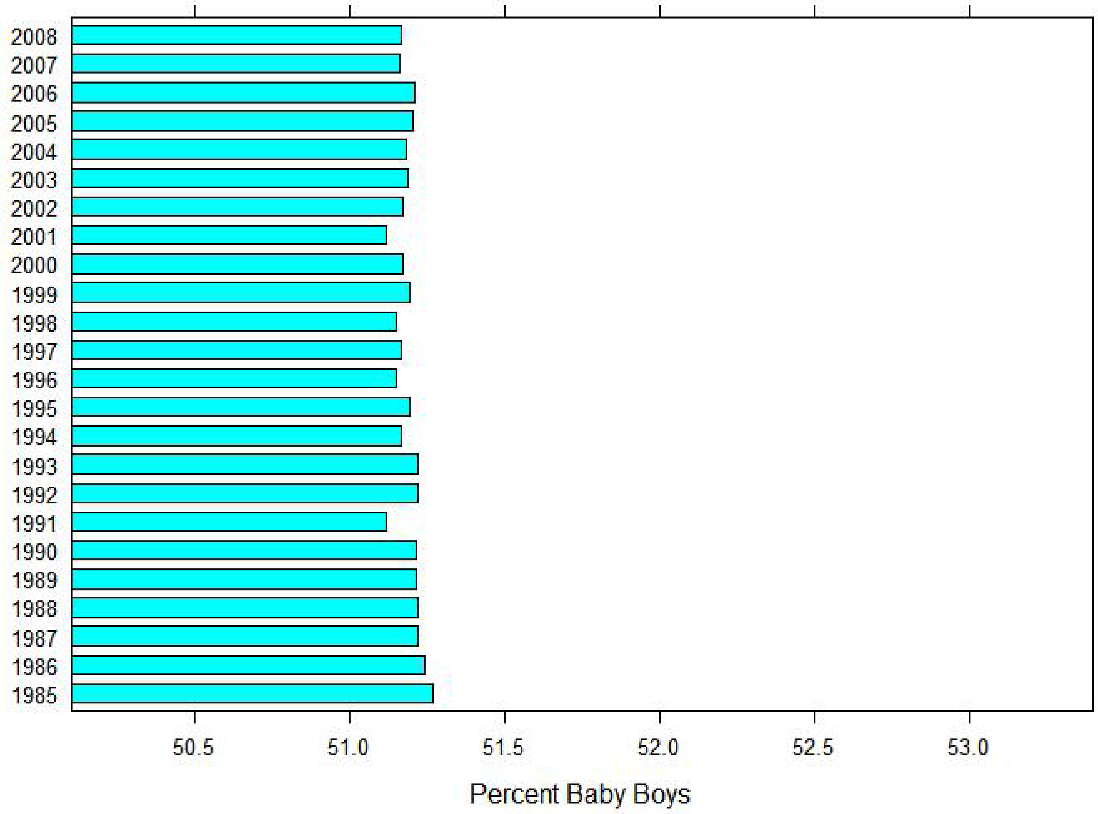
Summary of Other Baby Characteristics

- Highest percentage of boys for Preterm babies (53.342 %)
- Lowest percentage of boys for Triplets (or more) (50.109%)

Actual Excess Boys (Percent = 51.188 %)	2,293,649
Excess Boys (Triplets (50.109%))	210,765
Excess Boys (Preterm (53.342%))	6,452,011

REVOLUTION ANALYTICS

Year of Birth



Summary of Variation over Years

- Highest percentage of boys in 1985 (51.268 %)
- Lowest percentage of boys in 2001 (51.116 %)

Actual Excess Boys (Percent = 51.188 %)	2,293,649
Excess Boys (2001 (51.116%))	2,154,817
Excess Boys (1985 (51.268%))	2,448,979

REVOLUTION ANALYTICS

It's a Boy Multivariate Analysis

Multivariate Analysis

- Can we separate out the effects? Older mother's often paired with older fathers, parents are typically of the same race, higher birth order born to older parents?
- Do we see a statistically significant change in the percentage of boys born associated with years?

Other Studies Using this Data

- Some similar studies:
  - *Trend Analysis of the Sex Ratio at Birth in the United States* (Mathews 2005). Looks at data from 1940 to 2002.
  - *Trends in US sex ratio by plurality, gestational age and race/ethnicity* (Branum 2009). Looks at data from 1981 – 2006.
  - *Change in Composition versus Variable Force as Influences on the Downward Trend in the Sex Ratio at Birth in the U.S.,* 1971-2006 (Reeder 2010)
- In all cases (as far as I can tell), tables of data are extracted for each year, then "binned" data are analyzed.

Binning Approach

- Data can be binned very quickly using a combined file. Let's use some of the categories we've looking at:

```
bigCube <-
  rxCube(ItsaBoy~DadAgeR8:MomAgeR7:FRACEREC:
  MRACEREC:FHISP_REC:MHISP_REC:LBO4:DPLURAL_REC,
  data = birthAll, blocksPerRead = 10,
  returnDataFrame=TRUE)
```

Not enough data!

- Produces a total of 189,000 cells
- Only 48,697 cells have any observations
- Only 820 cells have more than 10,000 observations
- Alternative, run a logistic regression using the individual level data for all 96 million observations – takes about 5 minutes on my laptop

Logistic Regression Results

```
Call:
rxLogit(formula = ItsaBoy ~ DadAgeR8 + MomAgeR7 + FRACEREC +
    FHISP_REC + MRACEREC + MHISP_REC + LBO4 + DPLURAL_REC + Gestation +
    F(DOB_YY), data = birthAll, blocksPerRead = 10, dropFirst = TRUE)

Logistic Regression Results for: ItsaBoy ~ DadAgeR8 + MomAgeR7 + FRACEREC +
   FHISP_REC + MRACEREC + MHISP_REC + LBO4
    + DPLURAL_REC + Gestation + F(DOB_YY)
File name: C:\Revolution\Data\CDC\BirthUS.xdf
Dependent variable(s): ItsaBoy
Total independent variables: 68 (Including number dropped: 11)
Number of valid observations: 96534205
Number of missing observations: 0
-2*LogLikelihood: 133737824.1939 (Residual deviance on 96534148 degrees of
   freedom)
```

REVOLUTION ANALYTICS

Logistic Regression Results (con't)

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.0622737  0.0025238  24.674 2.22e-16 ***
DadAgeR8=Under 20              Dropped    Dropped Dropped  Dropped
DadAgeR8=20-24              -0.0061692  0.0013173  -4.683 2.83e-06 ***
DadAgeR8=25-29              -0.0080745  0.0013535  -5.966 2.44e-09 ***
DadAgeR8=30-34              -0.0094569  0.0013982  -6.764 1.34e-11 ***
DadAgeR8=35-39              -0.0095198  0.0014769  -6.446 1.15e-10 ***
DadAgeR8=40-44              -0.0124339  0.0016672  -7.458 2.22e-16 ***
DadAgeR8=Over 44            -0.0126309  0.0019898  -6.348 2.18e-10 ***
DadAgeR8=Missing            -0.0236538  0.0017614 -13.429 2.22e-16 ***
MomAgeR7=Under 20              Dropped    Dropped Dropped  Dropped
MomAgeR7=20-24              -0.0023200  0.0007839  -2.959 0.003082 **
MomAgeR7=25-29              -0.0018819  0.0008691  -2.165 0.030353 *
MomAgeR7=30-34              -0.0012946  0.0009746  -1.328 0.184073
MomAgeR7=35-39              -0.0027734  0.0010750  -2.580 0.009880 **
MomAgeR7=Over 39            -0.0058270  0.0017977  -3.241 0.001190 **
MomAgeR7=Missing               Dropped    Dropped Dropped  Dropped
```

Logistic Regression Results (con't)

FRACEREC=White	Dropped	Dropped	Dropped	Dropped	
FRACEREC=Black	-0.0069161	0.0010278	-6.729	1.71e-11	***
FRACEREC=American Indian / Alaskan Native	-0.0147689	0.0027099	-5.450	5.04e-08	***
FRACEREC=Asian / Pacific Islander	0.0097634	0.0019540	4.997	5.83e-07	***
FRACEREC=Unknown or not stated	-0.0050019	0.0013185	-3.794	0.000148	***
FHISP_REC=Not Hispanic	Dropped	Dropped	Dropped	Dropped	
FHISP_REC=Hispanic	-0.0123212	0.0009562	-12.886	2.22e-16	***
FHISP_REC=NA	-0.0058123	0.0012641	-4.598	4.27e-06	***
MRACEREC=White	Dropped	Dropped	Dropped	Dropped	
MRACEREC=Black	-0.0183213	0.0009090	-20.155	2.22e-16	***
MRACEREC=American Indian / Alaskan Native	-0.0128220	0.0022885	-5.603	2.11e-08	***
MRACEREC=Asian / Pacific Islander	0.0048036	0.0017915	2.681	0.007334	**
MRACEREC=Unknown or not stated	0.0098108	0.0021389	4.587	4.50e-06	***
MHISP_REC=Not Hispanic	Dropped	Dropped	Dropped	Dropped	
MHISP_REC=Hispanic	-0.0031037	0.0008926	-3.477	0.000506	***
MHISP_REC=NA	0.0034600	0.0021365	1.619	0.105340	

REVOLUTION ANALYTICS

Logistic Regression Results (con't)

```
LBO4=1                                      Dropped    Dropped Dropped  Dropped
LBO4=2                                   -0.0028028  0.0004975  -5.634 1.76e-08 ***
LBO4=3                                   -0.0053614  0.0006249  -8.579 2.22e-16 ***
LBO4=4 or more                           -0.0098031  0.0007528 -13.022 2.22e-16 ***
LBO4=NA                                  -0.0035381  0.0029063  -1.217 0.223449
DPLURAL_REC=Single                          Dropped    Dropped Dropped  Dropped
DPLURAL_REC=Twin                         -0.0869872  0.0012982 -67.005 2.22e-16 ***
DPLURAL_REC=Triplet or Higher            -0.1390442  0.0056857 -24.455 2.22e-16 ***
Gestation=Term                              Dropped    Dropped Dropped  Dropped
Gestation=Preterm                         0.1110529  0.0007413 149.815 2.22e-16 ***
Gestation=Very Preterm                    0.1185777  0.0013339  88.897 2.22e-16 ***
Gestation=NA                              0.0189284  0.0017299  10.942 2.22e-16 ***
```

REVOLUTION ANALYTICS

Logistic Regression Results (con't)

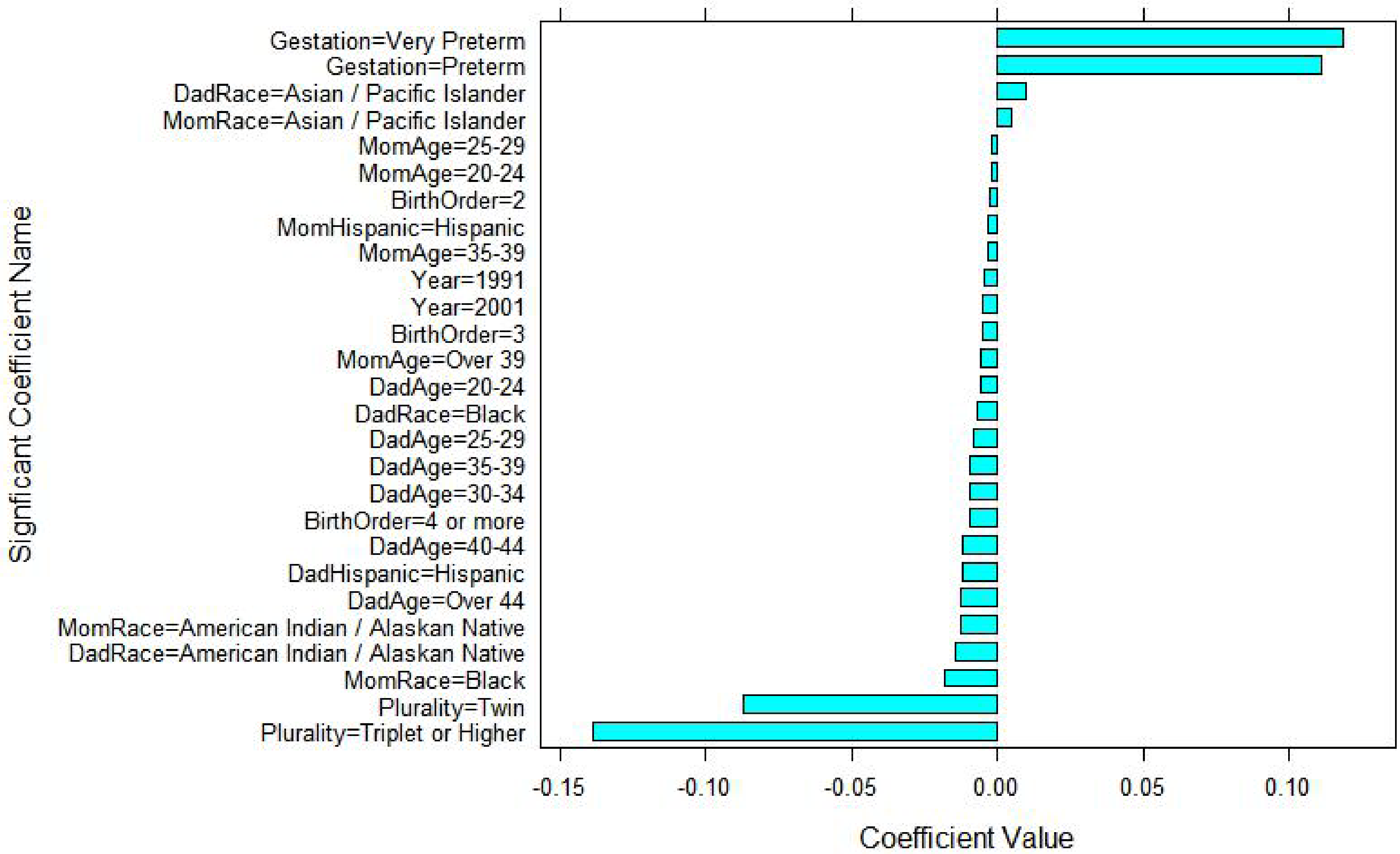
F_DOB_YY=1985	Dropped	Dropped	Dropped	Dropped	
F_DOB_YY=1986	-0.0010805	0.0014588	-0.741	0.458904	
F_DOB_YY=1987	-0.0019242	0.0014539	-1.323	0.185679	
F_DOB_YY=1988	-0.0017670	0.0014447	-1.223	0.221308	
F_DOB_YY=1989	-0.0012260	0.0023675	-0.518	0.604566	
F_DOB_YY=1990	-0.0011653	0.0024034	-0.485	0.627779	
F_DOB_YY=1991	-0.0047808	0.0024247	-1.972	0.048642	*
F_DOB_YY=1992	-0.0003573	0.0024264	-0.147	0.882935	
F_DOB_YY=1993	-0.0006009	0.0024312	-0.247	0.804798	
F_DOB_YY=1994	-0.0027435	0.0024367	-1.126	0.260199	
F_DOB_YY=1995	-0.0018579	0.0024329	-0.764	0.445066	
F_DOB_YY=1996	-0.0035393	0.0024345	-1.454	0.145994	
F_DOB_YY=1997	-0.0031564	0.0024354	-1.296	0.194958	
F_DOB_YY=1998	-0.0036406	0.0024360	-1.494	0.135047	
F_DOB_YY=1999	-0.0022566	0.0024355	-0.927	0.354177	
F_DOB_YY=2000	-0.0027696	0.0024316	-1.139	0.254707	
F_DOB_YY=2001	-0.0051636	0.0024395	-2.117	0.034289	*
F_DOB_YY=2002	-0.0029790	0.0024395	-1.221	0.222031	
F_DOB_YY=2003	-0.0027329	0.0024370	-1.121	0.262120	
F_DOB_YY=2004	-0.0027669	0.0024367	-1.136	0.256157	
F_DOB_YY=2005	-0.0018732	0.0024374	-0.768	0.442194	
F_DOB_YY=2006	-0.0014927	0.0024329	-0.614	0.539503	
F_DOB_YY=2007	-0.0033361	0.0024316	-1.372	0.170066	
F_DOB_YY=2008	-0.0025908	0.0024341	-1.064	0.287158	

- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-
- Condition number of final variance-covariance matrix: 12427.47
- Number of iterations: 2

Logistic Regression Results Summary

- Control group is:
  - Dad: under 20, white, non-Hispanic
  - Mom: under 20, white, non-Hispanic
  - Baby: first child for Mom, full-term, singleton
  - Born in 1985
- Almost everything highly significant, except for “Year” variables
- What are relative sizes of coefficients?

Logistic Regression on 'It's a Boy!' with 96,534,205 Observations



REVOLUTION ANALYTICS

Summary of Logistic Regression Results

- Gestation period and plurality have by far the biggest effects
- Race and ethnicity important. Low sex ratio for black mothers, in particular, needs further investigation.
- Can separate out effects of parents ages and birth order:
  - Effect of Mom's age is small
  - Dad's age matters more
  - Birth order still significant when controlling for parent's age

Further Research

- Data management
  - Further cleaning of data
  - Import more variables
  - Import more years (use weights)
  - Distributed data import
- *It's a Boy!* Analysis
  - More variables (e.g., education)
  - Investigation of sub-groups (e.g., missing Dad)
- Other analyses with birth data

Summary of Approach

- Small variation in sex ratio requires large data set to have the power to capture effects
- Significant challenges in importing and cleaning the data – using R and .xdf files makes it possible
- Even with a huge data, “cells” of tables looking at multiple factors can be small
- Using combined.xdf file, we can use regression with individual-level data to examine conditional effects of a variety of factors

References

- Branum, Amy M., Jennifer D. Parker, and Kenneth C. Schoendorf. 2009. "Trends in US sex ratio by plurality, gestational age and race/ethnicity." *Human Reproduction 24 (11):2936*- 2944.
- Matthews, T.J. and Brady E. Hamilton, Ph.D. 2005. "Trend Analysis of the Sex Ratio at Birth in the United States." *National Vital Statistics Reports 53(20):1-20.*
- Reeder, Rebecca A. "Change in Composition versus Variable Force as Influences on the Downward Trend in the Sex Ratio at Birth in the U.S., 1971-2006." University of Cincinnati, 2010.

Thank you!

- R-Core Team
- R Package Developers
- R Community
- Revolution R Enterprise Customers and Beta Testers

Contact:

sue@revolutionanalytics.com