

Chapter 1

General Introduction

Birth Order Data Analysis	1-3
Exploratory Data Analysis.....	1-3
Fitting a General Linear Model that Ignores Nesting	1-8

Birth Order Data Analysis

The data for this demonstration were generously provided by the authors of the manuscript by Wichman, Rodgers, and MacCallum *A Multilevel Approach to the Relationship between Birth Order and Intelligence* (2006, *Personality and Social Psychology*, 32, 117-127). The data was originally drawn from the National Longitudinal Study of Youth (NLSY). The sample we will explore consists of 3312 children nested within 2207 families. Of the 3312 children, 1623 were 7-8 years of age when assessed (young cohort) and 1689 were 13-14 years of age when assessed (old cohort). The primary outcome of interest was an age-standardized measure of the child's intelligence as assessed by the Peabody Individual Achievement Test. Here we will examine the mathematics subtest of the PIAT as the criterion variable. For this demonstration we use the `birth.sas7bdat` data file. Accompanying SAS code is in the `Birth_GLM.sas` file. Variables in the data are:

<code>mom_id</code>	A unique numeric identifier for each mother.
<code>kid_id</code>	A unique numeric identifier for each child.
<code>brthordr</code>	Indicates first, second, third, or fourth born as 1, 2, 3, or 4, respectively.
<code>math</code>	Scores on the math IQ test.
<code>cohort</code>	Indicates whether the child is in the young (7-8 year old) or old (13-14 year old) cohort. The young cohort is coded as <code>cohort=0</code> , the old cohort is coded as <code>cohort=1</code> .
<code>brthage</code>	The age of the child's mother when she delivered her first child.

Exploratory Data Analysis

Our analysis of the birth order data begins by exploring the characteristics of the data. We start by printing the first 20 records of data:

```
*looking at data;
proc print data=demo.birth (obs=20);
run;
```

Obs	mom_id	kid_id	brthordr	math	cohort	brthage
1	4	401	1	105	0	18
2	20	2001	1	112	0	30
3	25	2501	1	105	0	30
4	58	5801	1	107	0	24
5	97	9701	1	109	0	29
6	99	9901	1	104	0	29
7	128	12801	1	116	0	24
8	137	13701	1	131	0	26
9	138	13801	1	128	0	26
10	143	14301	1	111	0	24
11	149	14901	1	107	0	27
12	157	15701	1	105	0	22
13	158	15801	1	99	0	26

1-4 | Chapter 1 General Introduction

14	160	16001	1	102	0	30
15	165	16501	1	100	0	27
16	180	18001	1	112	0	21
17	186	18601	1	93	0	31
18	188	18801	1	88	0	26
19	192	19201	1	112	0	27
20	205	20501	1	95	0	21

We next wish to determine how many times multiple siblings from the same family appear in the data. To do this, we first determine the frequency of each mother ID number. Then we determine how often mother ID numbers occur once, twice, etc.

```
proc freq data=demo.birth noprint;
  table mom_id / out=count;
run;
proc freq data=count;
  table count;
run;
```

Results are shown here:

Frequency Count				
COUNT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1304	59.08	1304	59.08
2	719	32.58	2023	91.66
3	165	7.48	2188	99.14
4	19	0.86	2207	100.00

This indicates that the observations within the data are likely not independent – Many siblings are present. We next consider how many observations for each birth order are available within each cohort:

```
proc freq data=demo.birth;
  table brthordr*cohort/nocol norow nopercent;
run;
```

Table of brthordr by cohort				
brthordr	cohort			
Frequency	0	1	Total	
1	650	976	1626	
2	559	507	1066	
3	312	158	470	
4	102	49	151	
Total	1623	1690	3313	

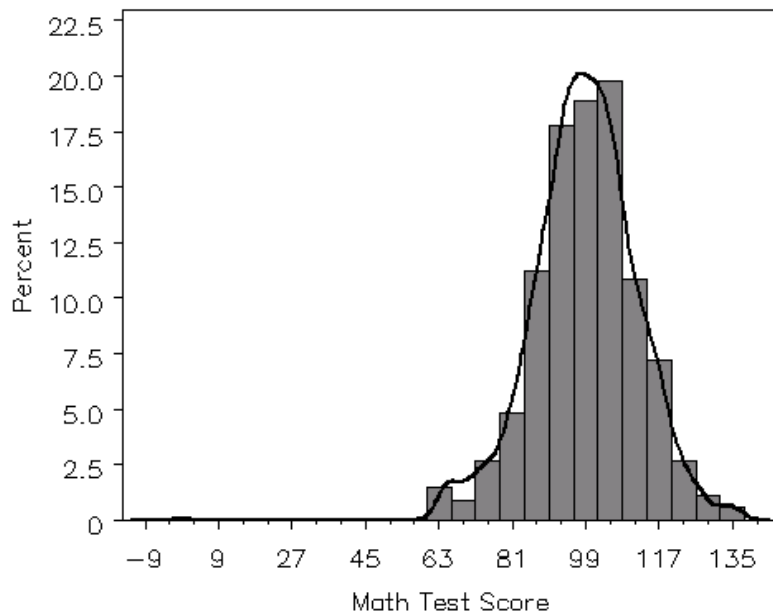
Note that, as one would expect, there are progressively fewer later born children in the sample.

We next examine the distribution of math IQ scores, first by calculating some simple statistics and generating a histogram.

```
goptions hsize=5 vsize=4;
proc univariate data=demo.birth;
  var math;
  histogram/ cfill=red kernel(color=black w=2) haxis=axis1;
  axis1 label=("Math Test Score");
run;
```

Selected results are shown here:

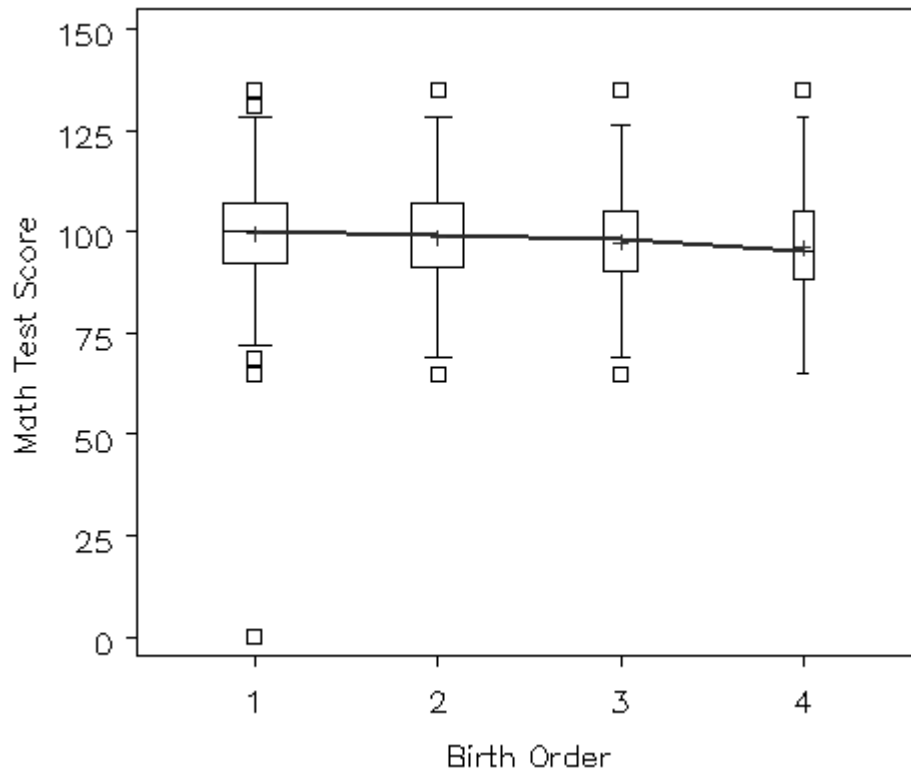
Moments			
N	3313	Sum Weights	3313
Mean	98.7986719	Sum Observations	327320
Std Deviation	12.5744488	Variance	158.116762
Skewness	-0.2676893	Kurtosis	1.35413657
Basic Statistical Measures			
Location		Variability	
Mean	98.7987	Std Deviation	12.57445
Median	99.0000	Variance	158.11676
Mode	100.0000	Range	135.00000
		Interquartile Range	16.00000
Extreme Observations			
----Lowest----		----Highest---	
Value	Obs	Value	Obs
0	1447	135	2676
65	3288	135	3027



Note the extreme observation with a math test score of zero.

Our initial graphical examination of birth order effects is via a boxplot. To do a side-by-side boxplot in SAS one must first sort on the classification variable (in this case, birth order).

```
proc sort data=demo.birth; by brthordr; run;
proc boxplot data=demo.birth;
  plot math*brthordr / boxstyle=schematic boxwidthscale=.5 cboxes=black
  cboxfill=red boxconnect=median vaxis=axis1 haxis=axis2;
  axis1 label=("Math Test Score");
  axis2 label=("Birth Order");
run;
```

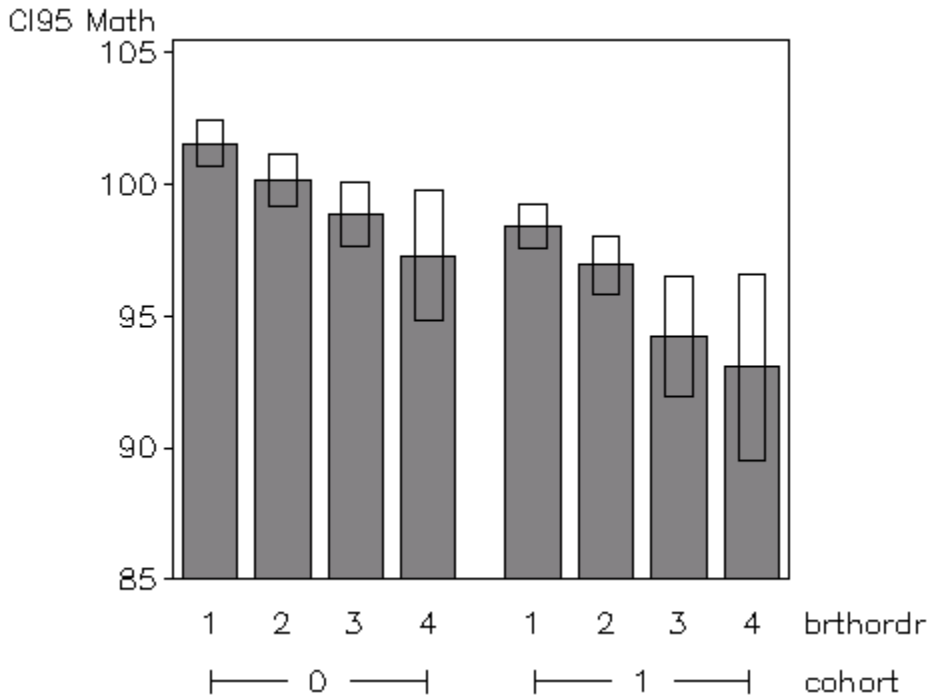


The boxplot shows a slight decline in median math test scores with birth order. The box widths here are proportional to sample size, once again indicating the smaller numbers of observations at later birth orders. The outlier is also shown clearly on the boxplot. We will now remove the outlier.

```
*removing outlier;
data birth;
  set demo.birth;
  where math ne 0;
run;
```

As a last preliminary analysis, we plot the mean math IQ score as a function of birth order and cohort. We seek to model these mean differences, determining if they are statistically significant.

```
proc gchart data=birth;
  vbar brthordr / discrete type=mean sumvar=math group=cohort
  axis=axis1 minor=0 coutline=black errorbars=bars;
  pattern value=solid color=red;
  axis1 label=("CI95 Math") order=(85 to 105 by 5);
run;quit;
```



Notice that children in the older cohort score lower, on average, than children in the younger cohort. This trend may be due to the sampling design -- children in the older cohort were typically born to younger mothers.

Fitting a General Linear Model that Ignores Nesting

We will now fit a GLM that fails to account for nesting to the data. The model we will fit is:

$$\text{Math}_i = \beta_0 + \beta_1 \text{Second}_i + \beta_2 \text{Third}_i + \beta_3 \text{Fourth}_i + \beta_4 \text{Old}_i + \beta_5 \text{Second}_i \times \text{Old}_i + \beta_6 \text{Third}_i \times \text{Old}_i + \beta_7 \text{Fourth}_i \times \text{Old}_i + r_i \quad r_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

We are assuming that the errors are independent and identically distributed (normal, with constant variance).

By default, PROC GLM will generate binary coding variables for any nominal predictor that is declared in the CLASS statement. By default, it will automatically make the last category the reference category. For birth order, the reference category would then be 4th born children. To get SAS to make 1st born children the reference category, we first sort the data by birth order in descending order (4, 3, 2, 1). Then in PROC GLM we use the ORDER=DATA option so that it will use the last value appearing in the data set (1) as the reference category.

```
proc sort data=birth; by descending brthordr; run;
proc glm data=birth order=data;
  class brthordr;
  model math=brthordr cohort brthordr*cohort/ solution;
run;quit;
```

Selected results are shown here:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	12950.7724	1850.1103	12.20	<.0001
Error	3304	500967.8170	151.6246		
Corrected Total	3311	513918.5894			
R-Square	Coeff Var	Root MSE	math Mean		
0.025200	12.45956	12.31360	98.82850		
Source	DF	Type III SS	Mean Square	F Value	Pr > F
brthordr	3	2562.271875	854.090625	5.63	0.0008
cohort	1	5039.262508	5039.262508	33.24	<.0001
cohort*brthordr	3	216.831731	72.277244	0.48	0.6985
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	101.5415385 B	0.48297896	210.24	<.0001	
brthordr 4	-4.2376169 B	1.31140565	-3.23	0.0012	
brthordr 3	-2.6697436 B	0.84808319	-3.15	0.0017	
brthordr 2	-1.3644365 B	0.71028961	-1.92	0.0548	
brthordr 1	0.0000000 B	.	.	.	
cohort	-3.1425641 B	0.62352316	-5.04	<.0001	
cohort*brthordr 4	-1.1001330 B	2.22927739	-0.49	0.6217	
cohort*brthordr 3	-1.4887244 B	1.35440289	-1.10	0.2718	
cohort*brthordr 2	-0.0779304 B	0.97932946	-0.08	0.9366	
cohort*brthordr 1	0.0000000 B	.	.	.	

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The note at the end of the output is normal and is not cause for concern (it merely indicates that for a 4-level classification variable, only 3 dummy variables are required to capture the effects). The row entry of 0.0000 indicates the reference category.

We see that the birth order main effect is significant and that the expected math IQ decreases almost linearly with birth order. No interaction with cohort is apparent, though children in the older cohort score about 3 points lower.