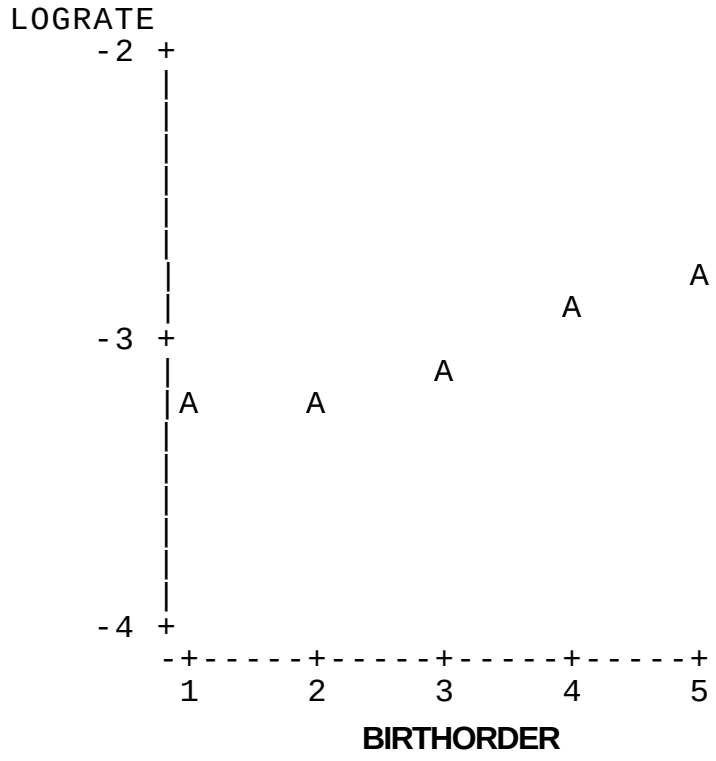
```
DATA a;
 INFILE 'downs.dat' ;
 INPUT  AgeL   AgeU  BirthOrd   Cases Births        ;
 MidAge = (AgeL + AgeU)/2 ;
 Rate = 1000*Cases/Births;  (epidemiologically correct: a prevalence rate)
 LogRate = Log10( (Cases+0.5)/Births );
           log10 for familiarity:   log10[1/1000] = -3
                                    log10[1/ 100] = -2
     Since π is small, log[π] & logit[π] virtually identical
 LogDenom = Log(Births);     For Poisson Regression [later]
 age_c    = MidAge - 30;            'CENTERED'
 age_c_sq = age_c * age_c;          X² less correlated with X
 age_c_cu = age_c * age_c * age_c;  if X already 'centered'
PROC PRINT;
```

OBS	AGEL	AGEU	BIRTHORD	CASES	BIRTHS	MIDAGE	RATE	LOGRATE	LOGDENOM	AGE_C	AGE_CSQ	AGE_CCU
1	15	19	1	107	230061	17	0.46509	-3.33043	12.3461	-13	169	-2197
2	15	19	2	25	72202	17	0.34625	-3.45201	11.1872	-13	169	-2197
3	15	19	3	3	15050	17	0.19934	-3.63347	9.6191	-13	169	-2197
4	15	19	4	1	2293	17	0.43611	-3.18431	7.7376	-13	169	-2197
5	15	19	5	0	327	17	0.00000	-2.81558	5.7900	-13	169	-2197
6	20	24	1	141	329449	22	0.42799	-3.36703	12.7052	-8	64	-512
7	20	24	2	150	326701	22	0.45914	-3.33661	12.6968	-8	64	-512
8	20	24	3	71	175702	22	0.40409	-3.39047	12.0765	-8	64	-512
9	20	24	4	26	68800	22	0.37791	-3.41434	11.1390	-8	64	-512
10	20	24	5	8	30666	22	0.26088	-3.55724	10.3309	-8	64	-512
11	25	29	1	60	114920	27	0.52210	-3.27864	11.6520	-3	9	-27
12	25	29	2	110	208667	27	0.52716	-3.27609	12.2485	-3	9	-27
13	25	29	3	114	207081	27	0.55051	-3.25733	12.2409	-3	9	-27
14	25	29	4	64	132424	27	0.48330	-3.31241	11.7938	-3	9	-27
15	25	29	5	63	123419	27	0.51046	-3.28861	11.7233	-3	9	-27
16	30	34	1	40	39487	32	1.01299	-2.98900	10.5837	2	4	8
17	30	34	2	84	83228	32	1.00928	-2.99341	11.3293	2	4	8
18	30	34	3	103	117300	32	0.87809	-3.05436	11.6725	2	4	8
19	30	34	4	89	98301	32	0.90538	-3.04073	11.4958	2	4	8
20	30	34	5	112	149919	32	0.74707	-3.12470	11.9179	2	4	8
21	35	39	1	39	14208	37	2.74493	-2.55594	9.5616	7	49	343
22	35	39	2	82	28466	37	2.88063	-2.53787	10.2565	7	49	343
23	35	39	3	108	45026	37	2.39861	-2.61803	10.7150	7	49	343
24	35	39	4	137	46075	37	2.97341	-2.52516	10.7380	7	49	343
25	35	39	5	262	104088	37	2.51710	-2.59827	11.5530	7	49	343
26	40	50	1	25	3052	45	8.19135	-2.07804	8.0236	15	225	3375
27	40	50	2	39	5375	45	7.25581	-2.13378	8.5895	15	225	3375
28	40	50	3	75	8660	45	8.66051	-2.05957	9.0665	15	225	3375
29	40	50	4	96	9834	45	9.76205	-2.00820	9.1936	15	225	3375
30	40	50	5	295	34392	45	8.57758	-2.06590	10.4456	15	225	3375

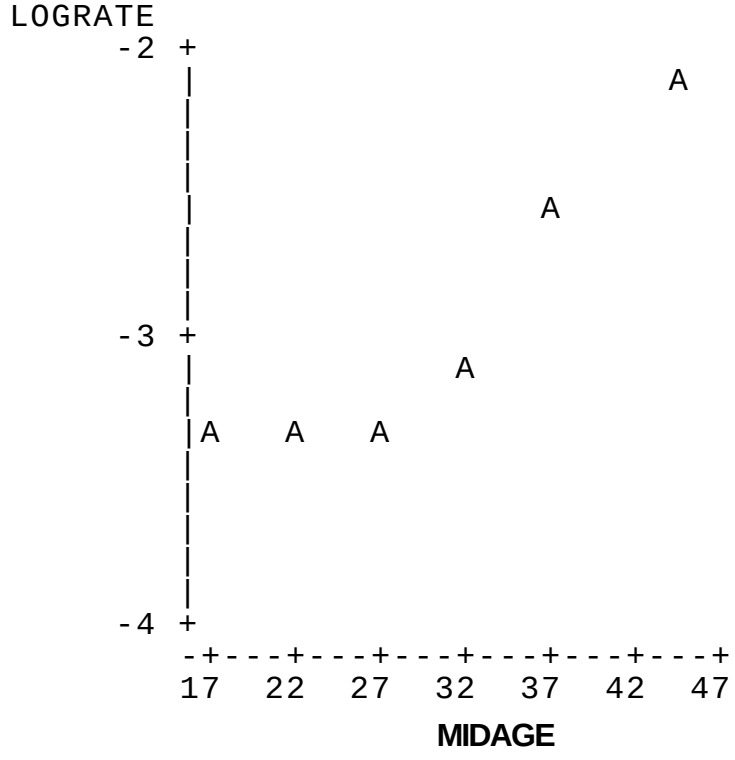
```
PROC MEANS DATA=a SUM NOPRINT;
CLASS BirthOrd;
 Var Cases Births;
 OUTPUT OUT=sums SUM= Cases Births; RUN;
DATA b_order; DROP _TYPE_ _FREQ_ ;
SET sums;
  Rate = 1000*Cases/Births;
  LogRate = Log10( (Cases+0.5)/Births );
  IF BirthOrd ne .; PROC PRINT; RUN;
PROC PLOT DATA=b_order ;
  PLOT LogRATE*BirthOrd /
   HPOS=25 VPOS=20 VAXIS= -4 -3 -2;
```

BIRTHORD	CASES	BIRTHS	RATE	LOGRATE
1	412	731177	0.56	-3.24
2	490	724639	0.67	-3.16
3	474	568819	0.83	-3.07
4	413	357727	1.15	-2.93
5	740	442811	1.67	-2.77



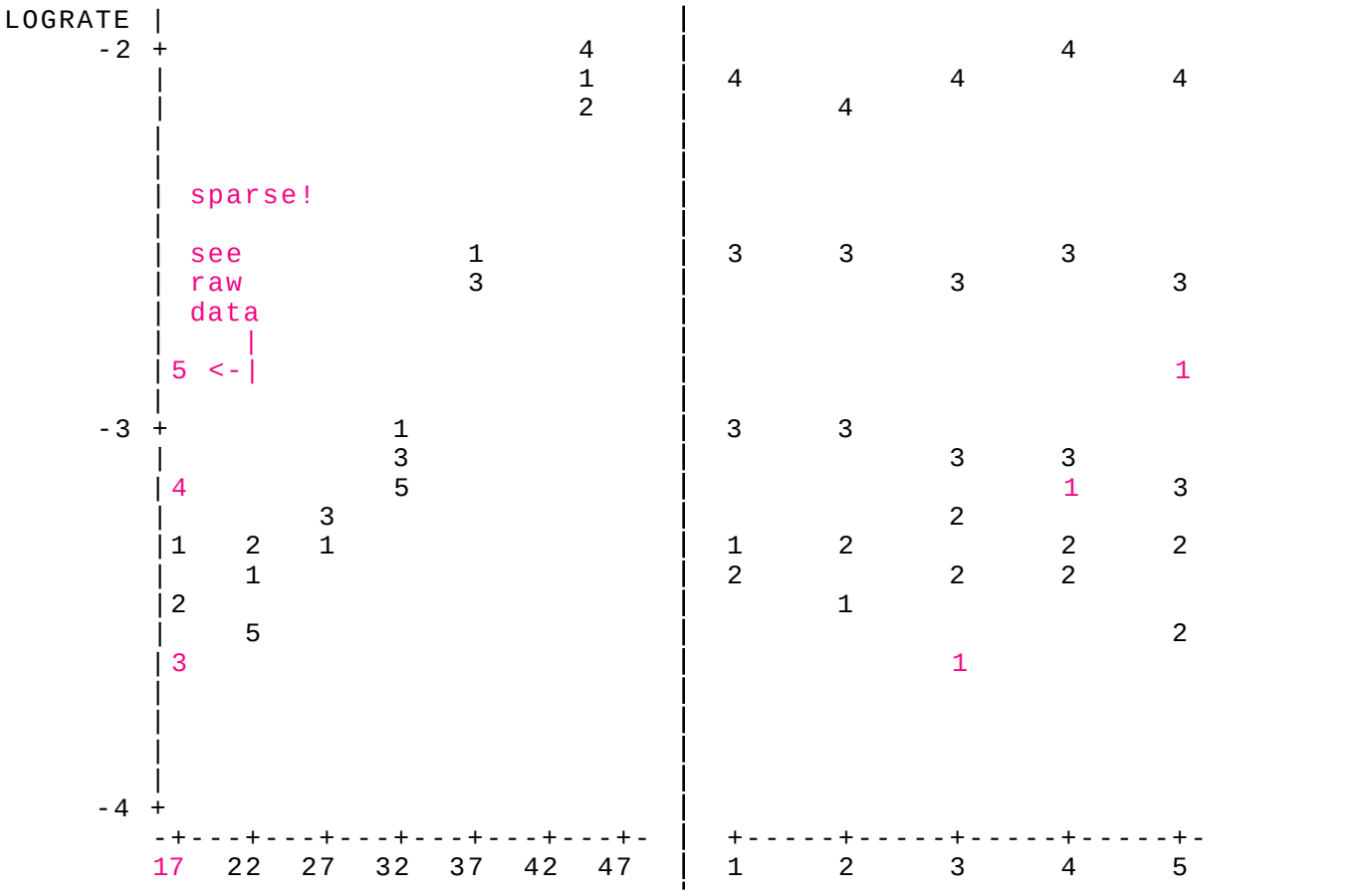

```
PROC MEANS DATA=a SUM NOPRINT;
CLASS MidAge;
 Var Cases Births;
 OUTPUT OUT=sums SUM= Cases Births; RUN;
DATA MidAge; DROP _TYPE_ _FREQ_ ;
SET sums;
  Rate = 1000*Cases/Births;
  LogRate = Log10( (Cases+0.5)/Births );
  IF MidAge ne .; PROC PRINT; RUN;
PROC PLOT DATA=MidAge ;
  PLOT LogRATE*MidAge / HPOS=25 VPOS=20
VAXIS= -4 -3 -2 ;
;
RUN;
```

MIDAGE	CASES	BIRTHS	RATE	LOGRATE
17	136	319933	0.425	-3.36
22	396	931318	0.425	-3.37
27	411	786511	0.522	-3.28
32	428	488235	0.876	-3.05
37	628	237863	2.640	-2.57
45	530	61313	8.644	-2.06



```
PROC PLOT DATA=a HPERCENT=50;
  PLOT LogRATE*MidAge=BirthOrd  LogRATE*BirthOrd=MidAge /
       HPOS=40 VPOS=28 VAXIS= -4 -3 -2;
```

uses birthorder as plotting symbol | *uses leftmost digit of age as plotting symbol*



Number of observations in data set = 30 [**unweighted***]

```
PROC GLM; MODEL LogRate = BIRTHORD;
```

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|.....	SE[b]
INTERCEPT	-6.83	-13.61	0.0001	0.50
BIRTHORD	**0.02**	0.14	0.8917	0.15

```
PROC GLM;
  MODEL LogRate = MIDAGE;
```

Parameter	Estimate	T for H0	Pr > |T|	SE[b]
INTERCEPT	-10.06	-31.91	0.0001	0.31
MIDAGE	**0.10**	10.92	0.0001	0.01

```
PROC GLM;
  MODEL LogRate = MIDAGE BIRTHORD;
```

Parameter	Estimate	T for H0	Pr > |T|	SE[b]
INTERCEPT	-10.12	-26.74	0.0001	0.37
MIDAGE	**0.10**	10.74	0.0001	0.01
BIRTHORD	**0.02**	0.31	0.7591	0.06

SAS GLM = "General Lin. Model: Identity Link g[μ|X] = μ|X = B.X; y|μ ~ Gaussian(0, σ)
Stata GLM = Generalized Lin Model: specify link & distrn [default: Identity Gaussian]
SAS REG and SAS GLM, and Stata regress all similar -- and relevant to c621;

* could[should!] weight the observations: since Var[logRate] is 1/#cases - 1/#births, and is dominated by 1/#cases, might use its reciprocal, #cases, as weights.

PROC GLM; MODEL LogRate = BIRTHORD; WEIGHT Cases; RUN;

30 observations in dataset; 1 has weight of 0 -> only 29 observations used.

Source	DF	SS	MS	F Value	Pr > F
Model	1	137.7	137.7	7.80	*0.0095*
Error	27	476.6	17.6		

R-Square: 0.22

Parameter	Estimate	SE	T	Pr > |T|
INTERCEPT	-3.35	0.203	-16.48	0.0001
BIRTHORD	**0.16**	0.057	2.79	*0.0095*

PROC Logistic DATA=a; MODEL Cases/Births = BirthOrd ; * NOTE: log is now to base *e*

Response Variable (Events): CASES EVENT 2529
Response Variable (Trials): BIRTHS NO EVENT 2822644 Nmbr Obsns:30 Link:Logit

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept & Covariates	Chi-Square for Covariates
-2 LOG L	40555.313	40154.036	401.277 with 1 DF (p=0.0001)
Score	.	.	414.851 with 1 DF (p=0.0001)

Parameter	Estimate	SE	Wald Chi-sq	Pr > Chi-sq	OddsRatio
INTERCPT	-7.84*	0.049	25308	0.0001	
BIRTHORD	**0.28**	0.014	400	0.0001	1.323

PROC Logistic;MODEL Cases/Births = MidAge BirthOrd / LACKFIT;

Criterion	Intercept Only	Intercept & Covariates	Chi-Square for Covariates
-2 LOG L	40555.313	38238.491	2316.822 with 2 DF (p=0.0001)
Score	.	.	2811.079 with 2 DF (p=0.0001)

Parameter	Estimate	SE	Wald Chi-sq	Pr > Chi-sq	OddsRatio
INTERCPT	**-10.87**	0.088	15079	0.0001	
MIDAGE	**0.14**	0.003	2185	0.0001	1.148
BIRTHORD	**-0.06**	0.015	16.8	0.0001	0.938

*[**Remark**: compare 0.14/year of age & age 'span' > 20 years, vs. 0.28/paroty & parity 'span' 1 to 5]*

Hosmer and Lemeshow Goodness-of-Fit Test

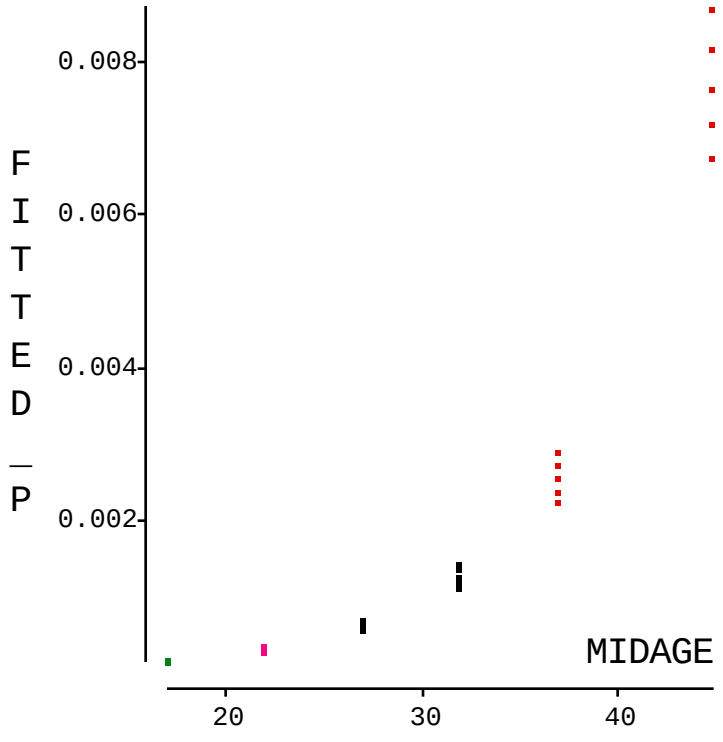
		EVENT		NO EVENT	
Group*	Total	Obs.	Exp.	Observed	Expected
1	319933	136	45.9	319797	319887.1
2	931318	396	266.4	930922	931051.6
3	462924	241	263.7	462683	462660.3
4	323587	170	223.6	323417	323363.4
5	248220	201	281.5	248019	247938.4
6	240015	227	316.6	239788	239698.4
7	299176	1158	1021.9	298018	298154.0

GoodnessOfFitStatistic=321, 5df (p=0.0001)

* Cf. Ch 5: groups formed by combining covariate patterns with similar predicted (fitted) probabilities

One can check, using the cell frequencies listed on the first page, and the totals in the 7 "Groups", which cells were combined together. Here the 10 rightmost cells were aggregated to form group 7, the 5 leftmost ones (coincident on graph) form group 1, the next five ie early 20's group 2, the next 3 (older 20's, parity 3-5 form group 3, etc.

From visual comparison of observed and expected in the table, not a good fit.



Maternal age as categorical variable

(rather than Midage as interval var.)

Make your own indicator terms..

(in SAS ver 8, CLASS* statement in PROC LOGISTIC makes terms. BUT be careful* and watch/override, default coding)

SAS

```
i_a_1519 = (ageL = 15); i_a_2024 = (ageL = 20);
i_a_2529 = (ageL = 25); i_a_3034 = (ageL = 30);
i_a_3539 = (ageL = 35); i_a_4049 = (ageL = 40);
```

Stata

```
gen i_a_1519 = (ageL == 15)
etc..
```

```
PROC Logistic DATA=a;
 MODEL Cases/Births = i_a_2024 i_a_2529
                      i_a_3034 i_a_3539 i_a_4049 BirthOrd / LACKFIT;
 OUTPUT OUT=fitted predicted=fitted_p ;
```

Model Fitting Info; Testing Global Null H BETA=0

Criterion	Intercept	&Covariates	Chi-Square
-2 LOG L	40555.3	38037.6	2517 6 DF (p=0.0001)
Score	.		5361 6 DF (p=0.0001)

Maximum Likelihood Estimates

Parameter	Est	SE	Chi-sq	P-value	OddsRatio
INTERCPT 1	-7.72	0.088	7615.4	0.0001	.
I_A_2024 1	0.02	0.100	0.0	0.8287	1.02
I_A_2529 1	0.25	0.102	6.1	0.0134	1.3
I_A_3034 1	0.78	0.104	56.9	0.0001	2.2
I_A_3539 1	1.90	0.102	342.4	0.0001	6.7
I_A_4049 1	3.10	0.106	856.1	0.0001	22.2
BIRTHORD 1	-0.02	0.016	3.2	0.0713	0.97

Hosmer and Lemeshow Goodness-of-Fit Test... ***see*** --->

	EVENT		NO EVENT	
Group Total	Observed	Expected	Observed	Expected
1 1251251	532	477.47	1250719	1250774
2 786511	411	386.33	786100	786124.7
3 448748	388	375.91	448360	448372.1
4 338663	1198	1189.41	337465	337473.6

GoodnessOfFitStatistic = 8.3 with 2 DF (p=0.0161)

*Given how little birth order, & how much age, matters (see earlier **Remark** re age and parity ranges), and fact that by using indicator terms for age categories, **we are effectively using the differences in empirical logits as the regression coefficients** (close to saturated model)*

Here are empirical logits, and the differences thereof along with the estimated coefficients from 'model' with 5 indicator terms for 6 age categories and using 'logit is linear in age' model (birthorder omitted)

Raw data

AGE	CASES/ BIRTHS	RATE /K	EMPIRICAL LOGIT	Diff- erence	Coeff. Cat.	Model*	Linear Logits
17	136/320K	0.42	**-7.76**	Ref 0	INTRCPT	-7.76	**-8.65**
22	396/931K	0.42	**-7.76**	+0.00	I_A_2024	0.000	**-7.99**
27	411/787K	0.52	**-7.56**	+0.20	I_A_2529	0.206	**-7.33**
32	428/488K	0.88	**-7.04**	+0.72	I_A_3034	0.724	**-6.67**
37	628/238K	2.64	**-5.93**	+1.83	I_A_3539	1.828	**-6.00**
45	530/ 61K	8.64	**-4.74**	+3.02	I_A_4049	3.020	**-4.95**

```
data fit;set fitted; [ model directly opposite]
 expected = ROUND(births*fitted_p, 0.1);
 chi_sq=ROUND((cases - expected)**2/expected, .1);
 PROC PRINT data=fit NOOBS; WHERE (expected > 5);
  VAR Midage BirthOrd cases expected chi_sq;
  SUM                 cases expected chi_sq;
```

MIDAGE	BIRTHORD	CASES	EXPECTED	CHI_SQ
17	1	107	98.8	0.7
17	2	25	30.1	0.9
17	3	3	6.1	1.6
22	1	141	144.6	0.1
22	2	150	139.2	0.8
22	3	71	72.7	0.0
22	4	26	27.6	0.1
22	5	8	12.0	1.3
27	1	60	63.5	0.2
27	2	110	112.0	0.0
27	3	114	107.9	0.3
27	4	64	67.0	0.1
27	5	63	60.6	0.1
32	1	40	37.2	0.2
32	2	84	76.2	0.8
32	3	103	104.2	0.0
32	4	89	84.8	0.2
32	5	112	125.6	1.5
37	1	39	40.7	0.1
37	2	82	79.3	0.1
37	3	108	121.7	1.5
37	4	137	121.0	2.1
37	5	262	265.3	0.0
45	1	25	28.9	0.5
45	2	39	49.4	2.2
45	3	75	77.2	0.1
45	4	96	85.2	1.4
45	5	295	289.3	0.1
		2528	2528.1	17.0

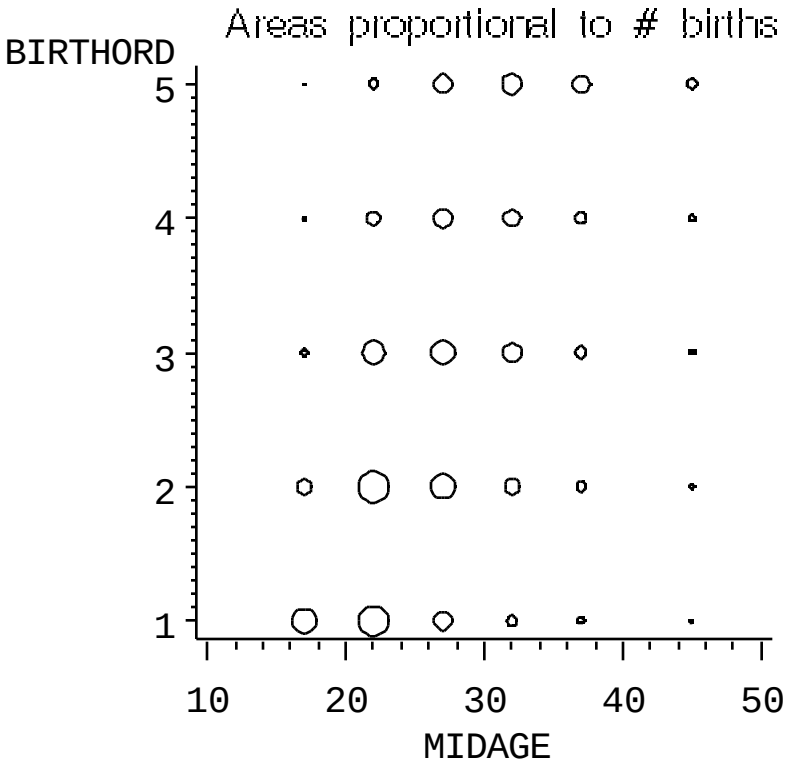
I have omitted neglibible $(O–E)^2/E$ contributions from the "NO EVENT" frequencies, and the (unstable) ones where E<5. The discrepancies are fairly small, and their sum well below critical chi-sq for > 20 degrees of freedom. Hosmer & Lemeshow test has too few categories for an adequate test.

```
*PROC Logistic; MODEL Cases/Births  = i_a_2024 i_a_2529 i_a_3034 i_a_3539 i_a_4049;
                MODEL Cases/Births  =                                    MIDAGE;
                                -10.8988          + 0.1323 × MIDAGE    -2 LOG L  38255.2
```

Comment:
Without maternal age in single years, and with the problem of the open-ended upper age interval, and with the very steep rise at older ages, we cannot adequately model the rates as a function of age. We could fit other functional forms (e.g. quadratic, cubic, etc..), but with effectively only 6 datapoints on age, we would quickly come close to a saturated model -- and thus have no way to test if our model has some stability.

Collinearity..
Maternal age and parity do not have a 'pathologically' collinear distribution: thus, we are able to estimate the two separate contributions reasonably accurately. For an extreme case of collinearity, and the instability it can create, see the 'resting on a knife-edge' spreadsheet on the c678 website).

Distribution of maternal age and parity (called BirthOrd here)



----------footnotes for previous page----------
See Onlinedoc [http://v8doc.sas.com/sashtml/ }

OPTIONS for CLASS Statement
PARAM=keyword
specifies the parameterization method for the classification variable or variables. Design matrix columns are created from CLASS variables according to the following coding schemes. **The default is PARAM=EFFECT**. If PARAM=ORTHPOLY or PARAM=POLY, and the CLASS levels are numeric, then the ORDER= option in the CLASS statement is ignored, and the internal, unformatted values are used.

EFFECT specifies effect coding
GLM specifies less than full rank, reference cell coding; this option can only be used as a global option
ORTHPOLY specifies orthogonal polynomial coding
POLYNOMIAL | POLY specifies polynomial coding
REFERENCE | REF specifies **reference cell coding**