Logistic Regression Model for Determining the Sex of a Child Using Age of the Mother and Month of Conception

Ben Apau-Dadson, Iddrisu Wahab Abdul*, Martin Owusu Amoamah & Joseph Dadzie
School of Applied Science and Arts, Department of Mathematics and Statistics, Accra Polytechnic
Accra, Ghana
*E-mail of the corresponding author: perfectwahab@yahoo.com

Abstract

The issue of sex determination (before conception) is beyond the study of Science. Though current studies have developed means to determine gender, it is not until the baby has been conceived. After conception, scientific methods can be used to determine whether the baby will be a boy or a girl. This study seeks to use the age of the mother and the month in which she conceives to predict the gender of the baby, using appropriate statistical methods. The data used for this study was extracted from the delivery books of the Maternity Department of Achimota Hospital in the Amasaman District of Greater Accra Region in Ghana from 2000-2004. Logistic regression was used to analyze the data since the dependent variable has only two possible values (Male or Female). The results reveal that the odds of estimating correctly the gender of a baby improves by only 0.4% if one knows the age of the mother and by 1.1% if one knows the month within which the child was conceived.

Keywords: Sex Determination, Logistic Regression, Age, Month of Conception.

1. Introduction

Gender as they say is a biological issue. That is, it is the man who determines the gender of his child. The father determines the sex of a child because the mother gives only an X chromosome while the father gives an X or Y chromosome [9].

In the production of gametes, the sex chromosomes segregate in typical Mendelian fashion. In mammals each ovum contains an X chromosome; in males one half of the sperm contains an X chromosome and the other half contain a Y chromosome. The sex of an offspring depends upon which type of sperm fertilizes the ovum; male, XY and female, XX [10].

Though the man's chromosomes determines the gender of his baby, he does not have any direct control on the determination of the sex of the baby since he cannot determine whether he is going to produce an X or Y chromosome. It is believed that the Y chromosome moves faster during the cold season (i.e. July – August in Ghana). This implies that women who conceive in July – August are likely to receive a Y chromosome and that will result in a boy. During the warm season, the X chromosomes tend to move faster and thus fertilization around this season is likely to result in a female.

The issue of sex determination (before conception) is beyond the study of Science. Though current studies have developed means to determine gender, it is not until the baby has been conceived. After conception, scientific methods can be used to determine whether the baby will be a boy or a girl. This study therefore seeks to use the age of the mother and the month in which she conceives to predict the gender of the baby, using appropriate statistical methods.

2. Method

2.1 Data

The data used for this study was extracted from the delivery books of the Maternity Department of Achimota Hospital in the Amasaman District of Greater Accra Region in Ghana [4]. The Achimota Hospital is a district hospital which offers a twenty-four hour service to the people in the district and beyond.

The maternity department of the hospital has records on the vitals of new born babies whose birth occurred at the hospital. The records show detailed information on deliveries at the hospital. The information recorded include; the age of the mother at birth, the maturity of the baby, the partography, date and time of delivery, blood loss by the mother during delivery, duration of the placenta, gender of baby, weight, height, chest circumference and the APGAR scores [5]. The APGAR score is a system of scoring infant's physical condition in the first one minute and in the next five minutes. It is a means of quickly assessing the new born infant in 1- 5 minutes interval. This is a standardized method of evaluating and recording the condition of the baby in numerical terms. Five vital signs are assessed and each is given a score of 0, 1 or 2 points. The vitals are; colour, respiration, heart rate, muscle tone and reflex responses. A score of 2 for heart rate indicates the child is having a 100% normal heartbeat while a score of 0 in respiration shows a great defect. Only data in the period 2000 – 2004 was used for this study due to time constraints.

2.2 Logistic Regression

If the dependent variable has only two possible values, for example 0 and 1, methods such as multiple regression become invalid since predicted values of π would not be constrained to the 0 and 1 limits [1]. Discriminant analysis can be used in such circumstances. However, discriminant analysis will only produce optimal solutions if its assumptions are supported by the data [2]. An alternative approach is Logistic regression [3]. In Logistic regression, the dependent variable is the probability that an event will occur, hence π is constraint between 0 and 1.

Logistic regressions work with **odds** and not proportions. The odds are simply the ratio of the proportions for the two possible outcomes. If $\hat{P}$ is the proportion for one outcome, then $1-\hat{P}$ is the proportion for the second outcome [7]:

$$ODDS = \frac{\hat{P}}{1-\hat{P}} \tag{1}$$

A similar formula for the population odds is obtained by substituting π for $\hat{P}$ in this expression.

To use categorical variables in logistic regression, we need to use a numeric code. The usual way to do this is with an **indicator variable.** For our purpose we will use an indicator of whether a child is a male or female:

$$y = \begin{cases} 1 & if\ child\ is\ a\ male \\ 0 & if\ child\ is\ a\ female \end{cases} \tag{2}$$

In simple linear regression we model the mean μ of the response variable y as a linear function of the explanatory variable: $\mu = \beta_0 + \beta_1 X$. With logistic regression we are interested in modeling the mean of the response variable π in terms of an explanatory variable x [7]. We could try to relate π and x through the equation $\pi = \beta_0 + \beta_1 X$. Unfortunately, this is not a good model. As long as $\beta_1 = 0$, extreme values of x will give values of

$\beta_0 + \beta_1 X$ that are inconsistent with the fact that $0 \le \pi \le 1$.

The logistic regression solution to this difficulty is to transform the odds $\left(\pi/(1-\pi)\right)$ using the natural logarithm [8]. We use the term **log odds** for this transformation. We model the log odds as a linear function of the explanatory variable:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \qquad (3)$$

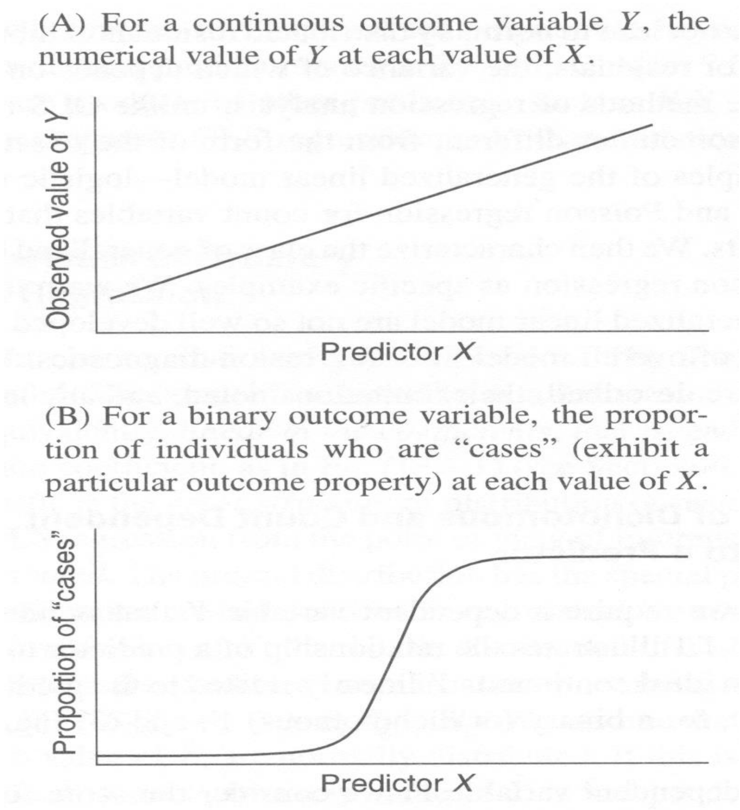

Figure 1: Logistic function

3. Results

3.1 A general descriptive analyses of data for the period 2000 – 2004

Table 1 show that a total of 5988 children were born in the five years period under consideration. As in other years, most of the women are in their twenties with the minimum age being 14 years and the maximum 50 years. The age distribution has a standard deviation of 5 years with a range of 36 years. Further, Table 1 shows that the oldest woman to give birth was at the age of 50 while the youngest was 14 years old. There was a general Skewness of about 0.332.

Table 1: Descriptive statistics for the age of mothers from 2000-2004

STATISTIC	AGE
N	5988
Mean	27
Median	26
Mode	25
Std. Deviation	5
Skewness	0.332
Std. Error of Skewness	0.032
Minimum	14
Maximum	50

Table 2 shows that for the years (2000-2004) under review there were a total of five thousand nine hundred and eighty-eight (5988) births comprising of two thousand nine hundred and ninety (2990) females and two thousand nine hundred and ninety-eight (2998) males. This shows almost an equal number of boys and girls being born in the five years under consideration. In 2004, seven hundred and seventy-three boys were born while the number of girls was eight hundred and sixteen. The least delivery was made in 2000 with four hundred and eleven boys and three hundred and eighty-six girls. In the first three years, more boys were born than girls with an average difference of twenty nine births. On the other hand, in 2003 and 2004 there were more girls born than boys.

Table 2 further shows that the number of births increases from 2000 to 2004. In 2000, the hospital recorded a total of seven hundred and ninety-seven (797) births, in 2001 nine hundred and twenty-eight (928), in 2002 one thousand two hundred and six (1206), in 2003 one thousand four hundred and sixty-eight (1,468) births, while in 2004 one thousand five hundred and eighty-nine (1589) births were recorded.

In the five years under review, the highest delivery occurred in the month of May; recording a total of seven hundred and forty-six (746) births, followed by June with six hundred and fifty-eight (658) births, then July recording six hundred and nine (609) births. April also recorded five hundred and seventy-one (571) new born children.

Table 2: Yearly and monthly summary of deliveries (2000-2004)

		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	TOT
2000	M	14	22	26	44	54	49	55	33	20	34	29	31	411
	F	20	15	35	44	49	48	47	28	26	30	18	26	386
	T	34	37	61	88	103	97	102	61	46	64	47	57	797
2001	M	32	23	39	34	72	36	37	52	33	48	43	44	493
	F	28	30	41	36	64	26	39	39	24	33	40	35	435
	T	60	53	80	70	136	72	76	91	59	81	83	79	928
2002	M	43	46	47	45	64	62	62	34	45	38	59	60	605
	F	32	40	37	47	73	77	60	54	42	58	44	37	601
	T	75	86	84	92	137	139	122	88	87	96	103	97	1206
2003	M	44	34	60	71	66	94	67	47	64	70	52	37	706
	F	44	39	56	64	94	80	92	64	57	71	58	43	762
	T	88	73	116	135	160	174	159	111	121	141	110	80	1468
2004	M	48	58	76	84	111	89	71	55	30	53	49	48	773
	F	61	61	76	102	90	87	79	61	35	49	51	55	816
	T	110	119	152	186	210	176	150	116	65	102	100	103	1589
TOT		367	368	493	571	746	658	609	467	376	484	443	416	5988
AVE		73	74	99	114	149	129	121	93	75	97	89	83	

Table 3 shows that the oldest person that gave birth was 50 years: while 14 years was the youngest female to give birth. The modal child bearing age is 25 years. On the average, women are likely to give birth at the age of 27 years. Table 3 also shows that out of the lot, only two (2) mothers were fourteen (14) years. A cumulative percentage of 6.5% are under 20 years, about 65% of the women are between the ages of 20 and 29 years. About 27% of the women were in their thirty's; while just 1% of them were 40 years and above.

Table 3: Distribution of ages of women who gave birth (2000 – 2004)

Age	Frequency	Percent	Cumulative Percent
14	2	0	0
15	17	0.3	0.3
16	22	0.4	0.7
17	68	1.1	1.8
18	135	2.3	4.0
19	149	2.5	6.5
20	289	4.8	11.4
21	198	3.3	14.7
22	328	5.5	20.1
23	363	6.1	26.2
24	444	7.4	33.6
25	543	9.1	42.7
26	474	7.9	50.6
27	456	7.6	58.2
28	458	7.7	65.9
29	350	5.8	71.7
30	419	7.0	78.7
31	205	3.4	82.2
32	279	4.7	86.8
33	172	2.9	89.7
34	167	2.8	92.5
35	138	2.3	94.8
36	105	1.8	96.5
37	71	1.2	97.7
38	54	0.9	98.6
39	30	0.5	99.1
40	23	0.4	99.5
41	7	0.11	99.61
42	12	0.21	99.82
43	4	0.1	99.92
44	2	0.03	99.95
45	2	0.03	99.98
49	1	0.01	99.99
50	1	0.01	100.0
Total	**5988**	**100.0**	

3.2 Logistic regression analysis for predicting the gender of a baby

The analysis provides information about two baseline situations; when only the constant is in the equation and when the predictor variables are in the equation. When only the constant is in the equation, it shows how well to predict the gender of the baby without knowing either the age of the mother or the month of conception. From

Table 4, 2993 babies who are boys were correctly classified as boys when one does not know the age of the mother and the month of delivery, and 2991 were wrongly classified as boys. Thus, if one simply guesses the gender of a baby, one would classify 50% of them correctly by chance.

Table 4: A Classification table for Gender

Classification Table[a,b]

Observed			Predicted: GENDER .00	Predicted: GENDER 1.00	Percentage Correct
Step 0	GENDER	.00	0	2991	.0
		1.00	0	2993	100.0
	Overall Percentage				50.0

a. Constant is included in the model.
b. The cut value is .500

Table 5: Effects of variables in the equation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.001	.026	.001	1	.979	1.001

Table 6 shows that the two variables (age and month) are individually not significant predictors of the gender of a baby. Age has a score of 0.787, 1 degree of freedom and significance of 0.375, while Month has a score of 1.907, degree of freedom of 1 and significance of 0.167. However, the month of conception does a better job than the age of the mother in predicting the gender of a baby.

Table 6: Effects of variables not in the equation

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	AGE	.787	1	.375
		MONTH	1.907	1	.167
	Overall Statistics		2.659	2	.265

The omnibus tests of model coefficients table indicates that, when both age and month are entered together, the model or the equation is still not significant (x^2 = 2.66, df = 2, N = 5986, P < 0.265). Thus when both predictor variables (age and month) are considered together, they do not significantly predict whether or not a child's gender will be male or female. This is evident from Table 7.

Table 7: Omnibus test of model coefficients

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2.659	2	.265
	Block	2.659	2	.265
	Model	2.659	2	.265

Table 8 presents the odds ratios, which suggest that the odds of estimating correctly the gender of a baby improves by only 0.4% if one knows the age of the mother and by 1.1% if one knows the month within which the child was conceived. The model summary estimates the percentage variance accounted for in the use of the model. It indicates only 0.4% or 1.1% of the variance, in whether a baby is a girl or a boy, can be predicted by the linear combination of the two variables.

Table 8: Results of logistic regression

Variable	B	SE	Odds ratio	P
Age	0.004	0.005	1.004	0.375
Month	0.011	0.008	1.011	0.167
Constant	-0.190	0.149	0.827	0.256

The final classification table (Table 9) indicates how well the combination of the variables predicts gender. Overall, 50.3% of the babies were predicted correctly. The independent variables were better at helping to determine who would be a female (53.7%) correctly than who would be a male (47.0%). The odds ratios in Table 8 indicates that the odds of estimating correctly the gender of a baby improve by 0.4% if one knows the age of the mother and improves by 1.1% if one knows the month of conception of the baby.

Table 9: Final classification table

Classification Table[a]

	Observed		Predicted		
			GENDER		Percentage Correct
			.00	1.00	
Step 1	GENDER	.00	1605	1386	53.7
		1.00	1587	1406	47.0
	Overall Percentage				50.3

a. The cut value is .500

4. Conclusion

The success story is that there is a relationship between the age of the mother, the month of conception and the gender of her baby based on the percentages of births of males and females. The research has revealed that the gender of a child is generally independent of the month of conception. Also the gender of a baby is independent of the age of the mother. On using the age of the mother and the month of conception to predict the gender of the baby, the month of conception does a better prediction than the age of the mother. Thus in predicting the gender of a baby, though both are statistically independent, the month of conception gives a better prediction than the age of the mother. Hence the gender of a child can be predicted using the month of conception with about 50% success. However, there can also be a 50% success if one guesses the gender of the baby.

References

[1] Benony, K.G. and Nathaniel, K.H. (2000). *Elements of Statistical Analysis.* The City Printers, Accra, Ghana.

[2] Bhatthcharyya G.K. and Johnson R.A. (1977). *Statistical Concept and Methods* pp. 428-444, John Wiley and Sons, USA

[3] Freund, J.E. and Simon, G.A. (1992). *Modern Elementary Statistics.* 8th edition, Prentice-Hall International Ltd., U.K

[4] Ghana Demographic and Health Survey (2000). Ghana Statistical Service

[5] Ghana Service Provision Assessment Survey (2002). Ghana Statistical Service

[6] Richard, A.J. and Dean, W.W. (1982). *Applied Multivariate Statistical Analysis* Prentice-Hall International, Inc. U.K.

[7] Sharma, S. (1996), *Applied Multivariate Techniques,* John Wiley & Sons, Inc. USA

[8] Srivastava M.S. and Carter E.M. (1983). *An Introduction to Applied Multivariate Statistics.* Elsevier Science Publishing Co., Inc., USA

[9] Sylvia, S.M. (1997), *Inquiry into Life,* 8th Edition, pp 499, McGraw-Hill Higher Education

[10] Taylor, D.L., Green, N.P.O., Stout, G..W., & Soper, R. (1998). *Biological Science.* 3rd Edition, pp 820/821, Cambridge University Press