PREDICTING THE FERTILITY OF A SPERM USING LOGISTIC REGRESSION

Jyothi Pranavi Devineni
School of Informatics and Computing
Indiana University

ABSTRACT

Fertility of human sperm is assessed using various factors like morphology, concentration and motility. It is the ability of the male sperm to fertilize the female eggs. The result of the diagnosis performed to assess the fertility of sperm might be normal (N) or altered (A). In this paper, an efficient method has been proposed to predict the result of the diagnosis based on various factors affecting the fertility of sperm like the person's smoking and drinking habits, if he has suffered any traumas and has met with any accidents and so on using logistic regression. As there are many explanatory variables, feature selection has been performed to select the most important variables. Then, using the most important or influential variables, different models have been fitted with different interactions to find the best fit for the data.

***Index Terms*—** Logistic regression, best model, fertility, explanatory variables, response variables.

1. INTRODUCTION

In today's competitive world, every individual is busy with his/her professional life, competing with each other and barely have any time for their personal life. However, at some point in life, every couple plans to have kids. A family is incomplete without children. Hence, having kids is an important part in the life of every couple. Fertility of both the male and female is essential for the female to conceive. In more than 50% of the cases, infertility is due to male sperm. The fertility of the male sperm is assessed using three factors:

Concentration: Number of sperms in each ml of serum.
Motility: What percentage of them are swimming forward.
Morphology: What percentage of them are normally shaped.

All these factors are very important for the female eggs to be fertilized. In the absence of any of these factors, some males may have children if they are lucky, else they have to go for medical techniques like ICSI (Intracytoplasmic Sperm Injection) to have their partner conceived. Hence, it is very important to know whether a man's sperm is fertile or capable of fertilizing the female's eggs so that the couple can go for some alternate methods like ICSI to have children. Diagnosis tests are available to know the fertility of the sperm. The result of the test would be either normal or altered. If the result is normal, there is no need of going for any unnatural techniques for having kids, else the couple has to take necessary treatment to have kids.

Instead of performing the diagnosis every time, we can also predict the result of the diagnosis, based on various factors which affect the fertility of a male. There are nine such factors in the given data set, from which an appropriate model can be built so that, given the factors affecting the fertility, fertility of a sperm can be assessed. Here, the result of the diagnosis is our response variable and there are nine explanatory variables. Logistic regression has been used to fit the data, as there are only two levels in the response variable, normal (1) or altered (0). Best model has been selected based on the AIC of the model and the goodness of fit of the model has been evaluated using the chi-squared P-value.

The remaining of this paper has been organized as follows: Section 2 describes the data set in detail. Section 3 gives an overview of the proposed model for the data. Section 4 discusses the experiments conducted to arrive at the best model and the respective results. Section 5 gives a conclusion to the paper, giving an overview of what has been discussed in the paper.

2. DATA DESCRIPTION

The data set is made from the analysis of semen samples collected from hundred volunteers and analyzed by the World Health Organization (WHO). The data set consists of nine attributes and one class variable. The nine attributes are different factors affecting the fertility:

1) Season in which the analysis was performed: This is one of the explanatory variables, which has four levels (-1. - 0.33, 0.33, 1) corresponding to the seasons winter, spring, summer and fall.

2) Age at the time of analysis: The age variable has been normalized in the given data set to represent 18-36 years with 0 to 1. It has been converted back to original age, for convenience in building the model.

3) Childhood diseases: This variable describes if the person from whom the sample has been collected has suffered from any childhood diseases such as chicken pox, measles, mumps, polio using 0 for “yes” and 1 for “no”.

4) Accident or serious trauma: This attribute gives the detail about whether the person has suffered from any trauma or met with any accident in the past. It gives 0 for “yes” and 1 for “no”.

5) Surgical intervention: This variable also gives a 0 if the person has undergone a surgery and 1 if he didn’t.

6) High fevers in the last year: It gives the detail when the participant has had a high fever with three levels (-1, 0, 1) representing less than three months ago, more than three months ago and no fever at all.

7) Frequency of alcohol consumption: This variable explains how frequently the participant consumes alcohol, with five levels from 0 to 1 (0.2, 0.4, 0.6, 0.8, 1) representing several times a day, every day, several times a week, once a week, hardly ever or never.

8) Smoking habit: This variable gives information about the smoking habit of the participant if it is never, occasional or daily represented by (-1, 0, 1)

9) Number of hours spent sitting per day: It is in normalized form (0,1) for hours ranging from 1 to 16 hours. This variable has been converted back to 1 to 16 hours, for convenience.

10) Diagnosis: This is the class variable that should be predicted by the proposed model. It has two levels, Normal and Altered, represented by 1 and 0 respectively.

There are no NAs in the data, hence there was no need for data cleaning. The only adjustments that have been made are in the representation of the data, as discussed above and the column names of the data set. Instead of taking the long column names given in the data set, their short forms (S,A,CD,T,O,F,AC,C,H,D)have been taken for convenience in addressing.

3. PROPOSED MODEL

As discussed earlier, the response variable has two classes represented by 0 or 1, hence binary data. So, logistic regression can be used to estimate the result of the diagnosis. We can fit a logistic model to the given data. We can convert the data into count data and fit a log-linear model as well, but handling a ten way table would be cumbersome. Hence, logit model seemed to be appropriate for the given data.

The most popular model for binary data is logistic regression. Suppose there is a single explanatory variable X, which is quantitative. For a binary response variable Y, recall that π(x) denotes the “success” probability at value x. This probability is the parameter for the binomial distribution. The logistic regression model has linear form for the logit of this probability,

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

The formula implies that π(x) increases or decreases as an S-shaped function of x. The logistic regression formula implies the following formula for the probability π(x), using the exponential function exp(α + βx) ,

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

The logistic regression formula indicates that the logit increases by β for every 1 cm increase in x. The parameter β determines the rate of increase or decrease of the S-shaped curve for π(x). The sign of β indicates whether the curve ascends (β > 0) or descends (β < 0), and the rate of change increases as |β| increases. When β = 0, π(x) is identical at all x, so the curve becomes a horizontal straight line. The binary response Y is then independent of X, i.e., Y is independent of X.

There are multiple logit models possible with different interactions between the nine explanatory variables. It’s very important to extract the most contributing variables and fit the best model, which is illustrated in the next section.

4. EXPERIMENTS AND RESULTS

Having decided to use a logistic model, it is important to select the best model among all the possible logit models. For this purpose, feature selection has to be performed. All the independent models have been fitted, starting from nine variables and gradually eliminating the less influential variables using back propagation algorithm. By doing so, nine variables have been reduced to three variables. The most important variables have been selected to be “Season in which the semen was collected, Childhood diseases and Accidents or Traumas undergone.”

The complete independence model using these three variables seemed to be the best independent model with the most influential variables, with an AIC of 73.33. There are other variables which might logically seem to be important and influential such as the Age of the participant at the time of the sample collection. It might be a little weird that Age is not included in the most influential variables, but if we look at the data, all the participants are between the age of 18 to 36 years of age. There might not be much of a difference in the fertility of a 28 year old and 36 year old. This is a very short range. Hence, age is not taken into consideration here.

Now, after performing the feature reduction, there are three variables. We cannot conclude that the independent model is the best fit without considering various models which involve different interactions between the three variables, namely joint independent models, conditionally independent models, homogeneous models and the saturated model. After fitting all these models, the results have been observed, to select for the best fitted model, with the desired fit statistics and degrees of freedom. The results of all the possible logistic models using three explanatory variables be “Season in which the semen was collected, Childhood diseases and Accidents or Traumas undergone” are as follows:

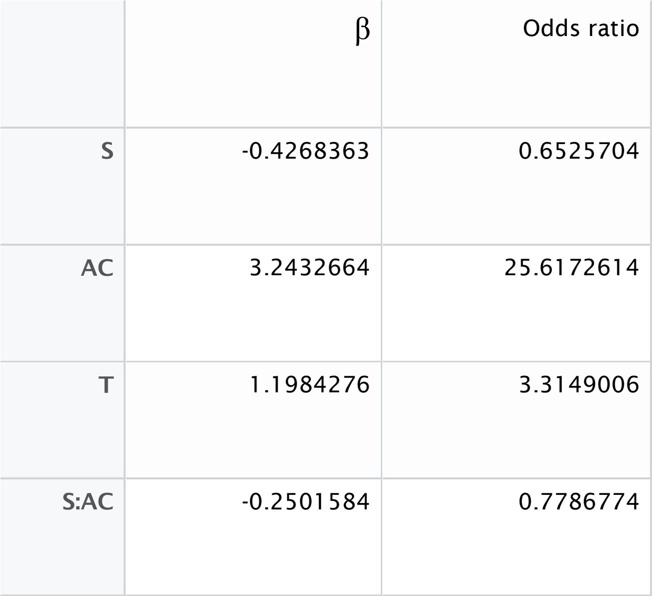
Model	df	G2	p	AIC
(S, T, AC)	96	65.32731	0.9929890	73.32731
(S, TAC)	95	64.61556	0.9927834	74.61556
(ST, AC)	95	65.26351	0.9914722	75.26351
(SAC, T)	95	65.31497	0.9913600	75.31497
(SAC, TAC)	94	64.61521	0.9910760	76.61521
(ST, ACT)	94	64.59070	0.9911314	76.59070
(ST, SAC)	94	65.26208	0.9895057	77.26208
(ST, SAC, TAC)	93	64.59070	0.9890816	78.59070

From the above results, we can see that the model with the lowest AIC is the complete independence model itself, with an AIC of 73.33 and a G^2 value of 65.33. But, when we look at the P-value for assessing the goodness of fit, it is 0.993, which is near ideal value. Hence, to avoid over fitting, an alternate model with a tradeoff between the AIC and goodness of fit has been selected. The model is (SAC, T), i.e., there exists interaction between the Season and the frequency of alcohol consumption of the participant and Trauma is independent of the other two.

The results for this model are an AIC of 75.3 with 95 degrees of freedom and G^2 = 65.3. The model can be represented by,

$$\mathrm{logit}(\pi_{ijk}) = \alpha + \beta_i^{S} + \beta_j^{AC} + \beta_i^{T} + \beta_{ij}^{SAC}$$

The β values for respective variables are,



	β	Odds ratio
S	-0.4268363	0.6525704
AC	3.2432664	25.6172614
T	1.1984276	3.3149006
S:AC	-0.2501584	0.7786774

Goodness of fit: Pearson’s Chi-squared test has been performed to evaluate the goodness of fit of the fitted model. The p-value is promising. It is not low. Hence, we don’t have any evidence against the null hypothesis. We cannot reject the null hypothesis. Hence we can say that our model is a best fit for the given data. The residuals also seemed to be pretty good.

5. CONCLUSION

The fertility of a male sperm may be affected by many factors including the age of the person, his drinking and smoking habits and so on. The model proposed in this paper takes into account such factors and predicts the result of a diagnosis to be normal or altered. As the response variable is found to be binary, logistic regression has been used and the best logistic model is found to be one of the joint independence models, (SAC, T).

REFERENCES

[1]Alan Agresti, *An Introduction to Categorical Data Analysis*, John Wiley and sons, Inc. publication, second edition.
[2] http://www.advancedfertility.com/malefactor.htm
[3] https://archive.ics.uci.edu/ml/datasets/Fertility#