Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 6
GEMFE 2026

Study on Factors Influencing Fetal Y Chromosome Concentration Based on Beta Regression and Rank Correlation Analysis

Fangyi Wang, Hao An, Quanying Ye, ChunXiang Liu, Haohan Li, Hui Zhou *

School of Electrical and Electronic Engineering, ShanDong University of Technology, Zibo, China, 255000

* Corresponding Author Email: zhouh0118@163.com

Abstract. This study conducted an association analysis and regression modeling of factors influencing fetal Y chromosome concentration in male fetuses, aiming to quantify the complex relationship between fetal Y chromosome concentration (proportional data) and physiological indicators such as maternal gestational age and BMI. First, the attached data underwent Shapiro-Wilk tests, confirming that all variables—including fetal Y chromosome concentration—failed to meet the assumption of Gaussian distribution. Consequently, the nonparametric Spearman rank correlation coefficient was selected for association analysis. The correlation heatmap revealed a positive correlation between Y chromosome concentration and gestational age (correlation coefficient: 0.188) and a negative correlation with maternal BMI (correlation coefficient: -0.155). Considering that Y chromosome concentration is a ratio-type variable within an interval, a beta regression model was established, incorporating nonlinear terms and mixed effects to enhance model accuracy. Comparing the beta regression model with various nonlinear regression models, the beta regression model demonstrated superior parameter significance and an MAE value of 0.026146, confirming strong correlation among model indicators. In terms of research significance, the established Beta regression model outperforms traditional nonlinear models in both parameter significance and predictive accuracy (MAE: 0.026146). This model systematically reveals the influence of various factors on Y chromosome concentration, holding significant value for clinically precise determination of optimal NIPT testing timing and personalized risk assessment.

Keywords: Beta regression model; Spearman rank correlation coefficient; non-invasive prenatal testing.

1. Introduction

The rapid advancement of Non-Invasive Prenatal Testing (NIPT) technology has provided an efficient and safe means for screening fetal chromosomal abnormalities. For male fetuses, the fetal Y chromosome concentration is a critical indicator for assessing the quality and reliability of the test, as it directly impacts the accuracy of detecting sex chromosome aneuploidies (e.g., 47, XYY syndrome). However, as a typical type of proportional data, the fetal Y chromosome concentration is strictly bounded within the (0,1) interval. Its value is influenced by complex interactions with various maternal physiological indicators, such as gestational age (GA), body mass index (BMI), and age. These relationships often exhibit non-linear characteristics that are difficult to accurately characterize using traditional linear models based on the assumption of a Gaussian distribution [1-2].

Although existing research has acknowledged the impact of maternal factors on the effectiveness of NIPT, there remains a relative scarcity of studies systematically quantifying the relationship between this specific proportional variable—fetal Y chromosome concentration—and key physiological indicators. Most traditional analytical methods, such as Pearson correlation and linear regression, suffer from reduced robustness and explanatory power when applied to non-Gaussian distributed data [3]. Consequently, there is a pressing need for a statistical modeling approach that aligns with the properties of proportional data while being flexible enough to capture complex non-linear relationships.

This study aims to address this gap by conducting an in-depth investigation into the key factors influencing fetal Y chromosome concentration, based on a Beta regression model and Spearman rank correlation analysis. First, we will perform Shapiro-Wilk normality tests on the attached dataset to verify the necessity of employing non-parametric methods [4-5]. Subsequently, a heatmap of Spearman correlation coefficients will be used to visually display the strength and direction of associations between Y chromosome concentration and various indicators. Finally, a Beta regression model incorporating non-linear terms and mixed effects will be constructed and compared against various non-linear regression models to evaluate its superiority in parameter significance and predictive accuracy. This research seeks to provide a quantitative theoretical basis and practical guidance for optimizing the timing of NIPT and improving risk assessment accuracy.

2. Correlation Analysis Based on Spearman Correlation Coefficient

On the basis of data preprocessing, this section will characterize the correlation between various indicators in the appendix and the fetal Y-chromosome concentration of pregnant women with male fetuses by calculating the Spearman correlation coefficient, and draw a corresponding heatmap to intuitively analyze their correlation characteristics [6-7].

2.1. Spearman Correlation Coefficient

Both Spearman correlation coefficient and Pearson correlation coefficient are often used for correlation analysis. Compared with Pearson correlation coefficient, Spearman correlation coefficient does not require variables to follow a Gaussian distribution, nor does it require a linear relationship between variables. It has strong robustness to outliers and is not affected by data dimensions.

First, a Gaussian distribution test is performed on the data in the appendix. Since the table contains approximately 1000 rows of data, which are small-sample data, the Shapiro-Wilk test is used to test each variable in the table to determine whether it satisfies the Gaussian distribution. The final results are shown in Table 1:

Table 1. Normality Test Results

Variable	W Statistic	p Value	Significance	Obeys Gaussian Distribution (p>0.05)
Age	0.9690	0.0000	***	No
Height	0.9904	0.0000	***	No
Weight	0.9620	0.0000	***	No
Number of Blood Draws for Testing	0.8614	0.0000	***	No
Maternal BMI	0.9429	0.0000	***	No
Raw Read Count	0.9718	0.0000	***	No
Proportion Mapped to Reference Genome	0.8373	0.0000	***	No
Proportion of Duplicate Reads	0.9583	0.0000	***	No
Number of Uniquely Mapped Reads	0.9752	0.0000	***	No
GC Content	0.9700	0.0000	***	No
Z-Value of Chromosome 13	0.9924	0.0000	***	No
Z-Value of Chromosome 18	0.9899	0.0000	***	No
Z-Value of Chromosome X	0.9753	0.0000	***	No
Z-Value of Chromosome Y	0.9592	0.0000	***	No
Y-Chromosome Concentration	0.9676	0.0000	***	No
X-Chromosome Concentration	0.9879	0.0000	***	No
GC Content of Chromosome 13	0.9439	0.0000	***	No
GC Content of Chromosome 18	0.9572	0.0000	***	No
GC Content of Chromosome 21	0.9725	0.0000	***	No
Proportion of Filtered Reads	0.9762	0.0000	***	No
Number of Deliveries	0.6248	0.0000	***	No

It can be observed from the table that the significance p-values of all variables are less than 0.05, rejecting the null hypothesis. That is, indicator variables including fetal Y-chromosome concentration, maternal gestational weeks, and BMI do not satisfy the Gaussian distribution. Therefore, the Spearman correlation coefficient is selected to analyze the correlation characteristics between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI [8-9].

Spearman correlation coefficient is a non-parametric method suitable for non-Gaussian distributed data, also known as rank correlation coefficient. It is generally used to test whether there is a correlation between two variables. By comparing the ranks of two variables rather than their original values, it measures their monotonic relationship (as one variable increases, the other variable either increases or decreases). It can not only effectively reflect the monotonic relationship between two variables but also maintain strong robustness when facing outliers. The Pearson correlation coefficient does not conform to the normal distribution, so parametric tests based on the Gaussian distribution cannot be applied. At this time, the Spearman method is consistent with the data characteristics, which can not only provide robust results but also reduce the analysis error caused by deviation from the Gaussian distribution.

Its expression is as follows:

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n} \tag{1}$$

In the formula, d is the rank difference of variables, and n is the sample size.

ρ is the Spearman correlation coefficient, which characterizes the correlation between two variables. Its value ranges from [-1, 1]. When $\rho = 1$, it indicates a perfect positive correlation between the two variables; when $\rho = -1$, it indicates a perfect negative correlation between the two variables; when $\rho = 0$, it indicates no correlation between the two variables [10].

2.2. Correlation Analysis Between Fetal Y-Chromosome Concentration and Various Indicators

A heatmap of the correlation between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI is shown in Figure 1:

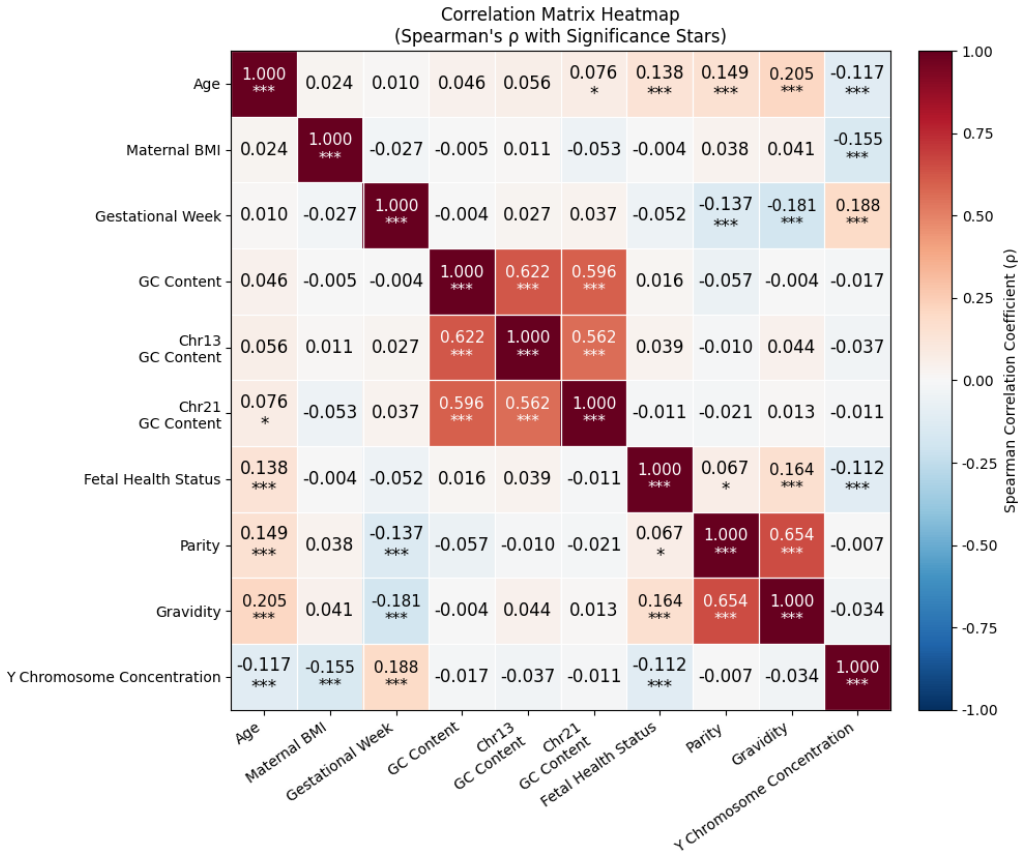

Figure 1. Heatmap of Correlation Matrix

Among them, *, **, and *** are significance markers, and the higher the number of *, the more important the correlation.

In this study, the Spearman correlation coefficient is used to measure the relationship between fetal Y-chromosome concentration and factors such as maternal gestational weeks and BMI. The Spearman correlation heatmap analysis is shown in the above figure, which displays these ten non-linear correlations. The correlation coefficient between Y-chromosome concentration and gestational weeks is 0.188, indicating a positive correlation between them; the correlation coefficient between Y-chromosome concentration and maternal BMI is -0.155, and the correlation coefficient between gestational weeks and maternal BMI is -0.027, indicating a negative correlation. This indicates that the lower the maternal BMI, the higher the Y-chromosome concentration, and the earlier the optimal

NIPT test time, thereby reducing potential risks, extending the treatment window, laying a systematic and quantifiable foundation for the future theoretical deepening and technological innovation of medicine, and determining the fetal health status as early as possible for active treatment.

3. Establishment of Fetal Y-Chromosome Concentration Regression Model

Preliminary tests on the data show that the linearity between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI is not obvious. Therefore, this section will establish a non-linear regression model to describe the correlation characteristics between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI. This study prioritizes the establishment of a Beta regression model to describe the relationship between fetal Y-chromosome concentration and other indicators, and conducts a comparative analysis of its significance in combination with a non-linear binomial regression model.

3.1. Beta Regression Model

Beta regression is often used to handle continuous proportional data and is widely applied in fields such as medicine, biostatistics, and social sciences. Compared with traditional linear regression and non-linear regression methods, the Beta regression model does not rely on the assumption of residual normality. Instead, it estimates parameters through the likelihood function. Therefore, it is still applicable even when the variable distribution deviates from the Gaussian distribution. Moreover, due to the flexible shape of the Beta distribution, it has low requirements on the shape of the data and can characterize a variety of proportional data.

Fetal Y-chromosome concentration is a typical type of proportional data. In the previous analysis, the Shapiro-Wilk test shows that it does not satisfy the Gaussian distribution, and it does not show a strong linear relationship with maternal indicators such as gestational weeks and BMI. Therefore, the Beta regression model is very suitable for characterizing its relationship with maternal indicators such as gestational weeks and BMI.

3.2. Establishment and Solution of the Beta Regression Model for Fetal Y-Chromosome Concentration

Furthermore, this section will establish a Beta regression model between fetal Y-chromosome concentration and maternal gestational weeks, BMI, and age. From the heatmap in Section 2.1, it is found that in addition to maternal gestational weeks and BMI, maternal age also has a relatively high correlation with fetal Y-chromosome concentration. Therefore, maternal gestational weeks, BMI, and age are selected as independent variables, and interaction terms are screened, and mixed effects are considered to improve the model accuracy. The specific model establishment steps are as follows:

Step 1: Data Standardization

Before establishing the Beta regression model, it is first necessary to standardize maternal gestational weeks, BMI, and age, and map them to eliminate the influence of dimensions between different variables. The specific transformation formula is as follows:

$$z = \frac{x-\mu}{\sigma} \tag{2}$$

Where x is the data point, μ is the mean value of the dataset, and σ is the standard deviation of the dataset.

Step 2: Establishment of the Mean Sub-Model

Assume that the fetal Y-chromosome concentration C_Y follows a Beta distribution, u_c is the mean value of C_Y, and the logit link is adopted. The transformation formula is:

$$\eta_i = \log \frac{u_i}{1-u_i} \tag{3}$$

η_i is the log-odds, which maps the data in the range (0, 1). The complete mean sub-model is as follows:

$$\eta_i = \beta_0 + \beta_1 x_{GA} + \beta_2 x_{BMI} + \beta_3 x_{AGE} \tag{4}$$

Where x_{GA} is the maternal gestational weeks, x_{BMI} is the maternal body mass index, and x_{AGE} is the maternal age.

β_0 is the intercept term, β_1 is the correlation coefficient of gestational weeks on Y-chromosome concentration, β_2 is the correlation coefficient of BMI on Y-chromosome concentration, and β_3 is the correlation coefficient of age on Y-chromosome concentration.

Step 3: Introduction of Non-Linear Terms

To explore the non-linear relationships between data and improve the accuracy of the regression model, interaction terms are introduced into the original model, and their significance and fitting effects are tested respectively. In this study, x_{GA}^2, x_{BMI}^2, x_{AGE}^2, as well as $x_{GA} \cdot x_{BMI}$, $x_{AGE} \cdot x_{BMI}$, and $x_{GA} \cdot x_{AGE}$ are constructed as non-linear terms.

Step 4: Mixed Effects Optimization

Since some pregnant women may undergo multiple blood draws and tests, there is a possibility of repeated tests in the appendix data. Therefore, random disturbances are introduced to characterize the scenario of repeated tests. The final regression model is:

$$\eta_i = \beta_0 + \beta_1 x_{GA}^* + \beta_2 x_{BMI}^* + \beta_3 x_{AGE}^* + \beta_4 x_{add}^* + u_j \tag{5}$$

Where x_{add} is an additional term used to explore the non-linear relationships in the model, and u_j is a disturbance term that characterizes the repeated test scenario of the pregnant woman with the corresponding ID and follows a Gaussian distribution.

Step 5: Parameter Estimation and Solution

In Beta regression, unlike traditional linear regression models and non-linear regression models, parameters are generally solved using maximum likelihood estimation instead of the least squares method. Subsequently, parameters can be estimated through numerical optimization methods and solved with the help of certain solving tools.

3.3. Model Significance Test

3.3.1. Wald Test.

To test the significance of the above regression model, the Wald test is used for relevant significance testing in this study.

Wald Test: The Wald test is a statistical method used to evaluate the significance of each parameter (such as regression coefficients) in the regression model. Its basic principle is: if the difference between the estimated value of a parameter and its assumed value is large enough, and this difference exceeds the range of estimation error, then it can be considered that the parameter has statistical significance in the model. Its explicit expression is as follows:

$$W = \frac{(\hat{\beta} - \beta_0)^2}{\mathrm{Var}(\hat{\beta})} \tag{6}$$

3.3.2. Test of Non-Linear Terms and Interaction Terms.

First, non-linear terms such as x_{GA}^2, x_{BMI}^2, x_{AGE}^2, and interaction terms such as $x_{GA} \cdot x_{BMI}$, $x_{AGE} \cdot x_{BMI}$, $x_{GA} \cdot x_{AGE}$ are respectively introduced into the model, and MAE is used to test their accuracy to screen the optimal model. The specific results are shown in Table 2:

Table 2. Fitting Results of Interaction Term Test

Added Term	Beta_MAE	Number of Features	OLS_MAE
x_{GA}^2	0.026396	4	0.026417
x_{BMI}^2	0.026146	4	0.026287
x_{AGE}^2	0.026434	4	0.026481
$x_{GA} \cdot x_{BMI}$	0.026423	4	0.026471
$x_{AGE} \cdot x_{BMI}$	0.026364	4	0.026382
$x_{GA} \cdot x_{AGE}$	0.026414	4	0.026457

It can be seen from the table that $x_{GA} \cdot x_{AGE}$ as a non-linear term can better reflect the non-linear relationships between variables, with the smallest MAE. Other models are less effective. Therefore, in the subsequent analysis, $x_{GA} \cdot x_{AGE}$ is selected as the non-linear term to describe the relationship between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI. At this time, the values of each variable are shown in Table 3:

Table 3. Variable Values

Definition	β_0	β_1	β_2	β_3	β_4	Logarithm of Beta Distribution Precision
Regression Coefficient	-2.50	0.06	-0.079	0.052	-0.05	4.17

3.3.3. Significance Comparison Between Beta Regression Model and Non-Linear Polynomial Regression.

Based on theory and our cross-validation results, the Beta regression model is more robust and more consistent with the generation mechanism of proportional data in terms of "parameter significance" and "generalization accuracy" compared with non-linear polynomial regression. As shown in Table 4:

Table 4. Significance Comparison

Added Term	Regression Coefficient	Standard Error	Significance Level	Beta Significant	OLS Significant
x_{GA}^2	0.033	0.033	0.001	TRUE	FALSE
x_{BMI}^2	-0.030	-0.030	0.001	TRUE	TRUE
x_{AGE}^2	2.060	2.060	0.999	FALSE	FALSE
$x_{GA} \cdot x_{BMI}$	0.007	0.007	0.599	FALSE	FALSE
$x_{AGE} \cdot x_{BMI}$	0.013	0.013	0.427	FALSE	FALSE
$x_{GA} \cdot x_{AGE}$	-0.024	-0.023	0.145	FALSE	FALSE

Furthermore, to evaluate the modeling effect of fetal Y-chromosome concentration, Beta regression and ordinary least squares (OLS) are respectively used for fitting. The figure below shows the residual-fitted value and observed-predicted scatter plots of the two models (top: Beta, bottom: OLS), which are used to compare the residual structure, heteroscedasticity, and calibration, thereby serving as the basis for model selection.

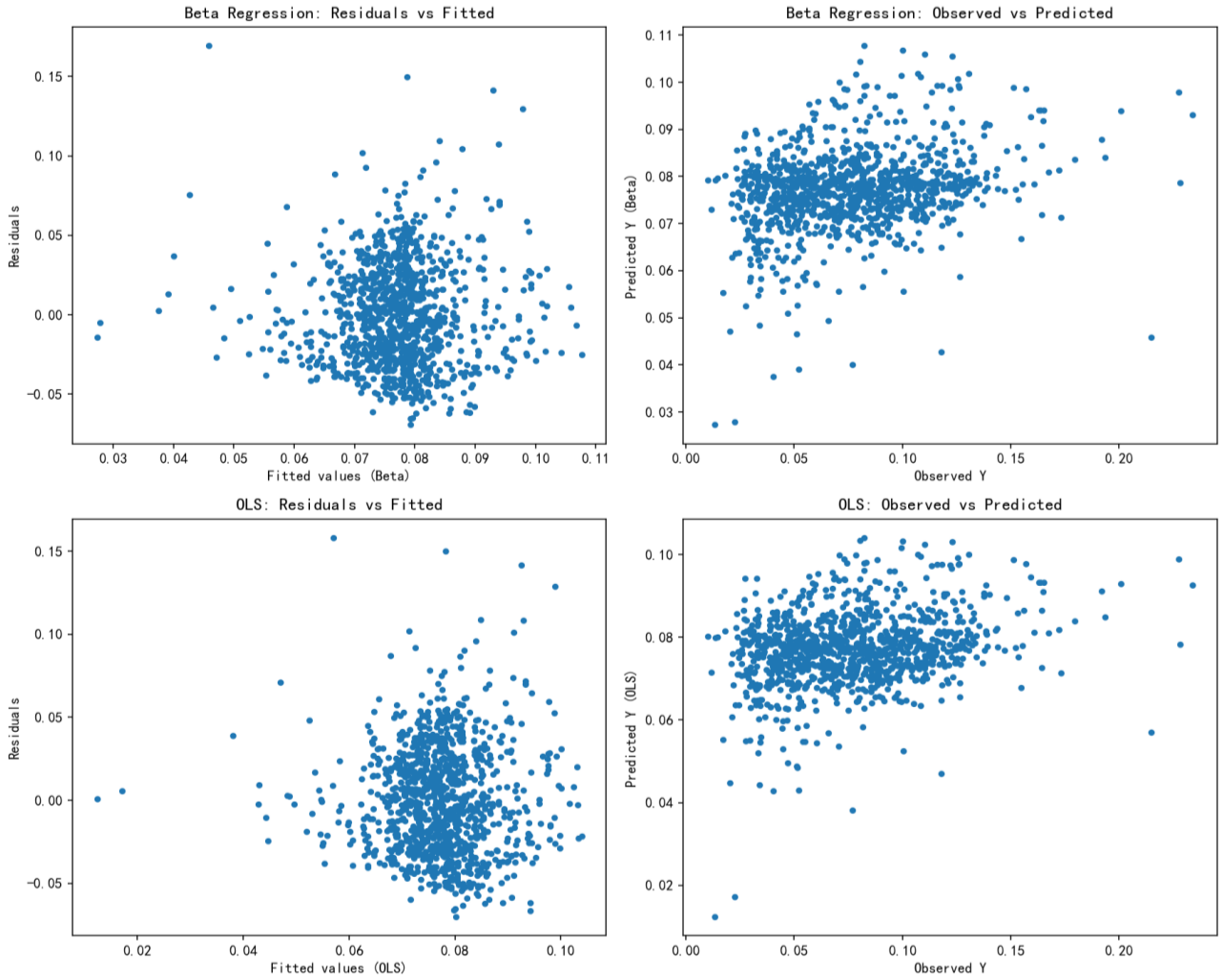

Figure 2. Comparison of Fitting Curves of Multiple Regression Models

It can be seen from the figure 2 that the MAE values of the two models are similar, and their prediction accuracy is comparable. However, the Beta regression model has better parameter significance and is more suitable for describing the relationship between fetal Y-chromosome concentration and maternal indicators such as gestational weeks and BMI. In addition, since Y-chromosome concentration is proportional data, the Beta regression model is more practical.

In addition, it can be found that even if maternal gestational weeks, BMI, age, and other non-linear terms are added as dependent variables, the effect is still not obvious. This indicates that in addition to these three indicators, there are still potential factors that have an important impact on fetal Y-chromosome concentration. It also suggests that in addition to the maternal physical indicators themselves, there may be other external factors that have a certain impact on the fetal Y-chromosome concentration of pregnant women.

4. Conclusion

This study investigated the factors influencing fetal Y chromosome concentration based on the Beta regression model and Spearman rank correlation analysis. The Shapiro-Wilk test confirmed that all variables, including fetal Y chromosome concentration, deviated from the Gaussian distribution,

justifying the use of non-parametric methods. Spearman correlation analysis revealed a positive correlation between Y chromosome concentration and gestational age (0.188) and a negative correlation with maternal BMI (-0.155). A Beta regression model was subsequently established, incorporating non-linear terms and mixed effects to accommodate the proportional nature of the data. Comparative analysis demonstrated that the Beta regression model outperformed traditional non-linear polynomial models in terms of parameter significance and predictive accuracy, achieving an MAE of 0.026146. These findings systematically quantify the impact of maternal physiological indicators on Y chromosome concentration, providing a robust theoretical basis for optimizing NIPT testing timing and facilitating personalized clinical risk assessment.

References

[1] Ju Aiping, Meng Xiangrong, Qin Yanling, et al. Application Value of Non-Invasive Prenatal Testing in Screening Fetal Chromosomal Copy Number Variations [J]. Practical Electrocardiography and Clinical Diagnosis and Treatment, 2025, 34 (05): 665 - 671.

[2] Zhu Sijing, Zhen Shuai, Niu Hui, et al. Current Status and Future Prospects of Whole Exome Sequencing in Prenatal Diagnosis [J]. Chinese Journal of Maternal and Child Health Research, 2025, 36 (09): 61 - 67.

[3] Shi Weihui, Xu Chenming. Application Value of Non-Invasive Prenatal Testing in Diagnosing Obstetric Maternal Complications and Comorbidities [J]. Journal of Practical Obstetrics and Gynecology, 2025, 41 (08): 617 - 620.

[4] Zhou Jing, Ji Xiuqing, Li, et al. Analysis of Seven Cases of Y Chromosome Abnormalities in Amniotic Fluid Cells Detected by Prenatal Diagnosis [J]. International Journal of Obstetrics and Gynecology, 2025, 52 (04): 394 - 401.

[5] Jiang Liyia, Lu Shaokan, Du Jia'en, et al. Development and Application of Non-Invasive Prenatal Testing Technology [J]. Clinical Medical Research and Practice, 2025, 10 (23): 191 - 194.

[6] Zhang Peng, Mo Weiying, Meng Minghui, et al. Impact of Non-Invasive Prenatal Testing on Detection of Sex Chromosome Aneuploidy and Related Ethical Considerations [J]. Chinese Journal of Clinical Medicine, 2025, 18 (06): 690 - 695.

[7] Zhang Yumei, Han Ying, Qiu Huiguo. Application Value of Non-Invasive Prenatal Screening Technology in Detecting Chromosomal Abnormalities and Microdeletions [J]. Journal of Yanbian University of Science and Technology (Medical Science), 2025, 48 (06): 59 - 61. DOI:

[8] Huang Chunyuan, Xu Xueqing, Shao Congwen. A Reassuring “Birth Plan” for Your Baby: Non-Invasive Prenatal Testing and Scientific Reproductive Guidance After Miscarriage [C]//Guangdong Tumor Rehabilitation Association. 2025 South China Health Management Forum Health Science Popularization Collection. Precision Medicine Clinical Laboratory, Shenzhen Hospital of Southern Medical University; 2025: 315 - 317.

[9] Tan Lingjun, Wang Bin, Huang Zhiqiong, et al. Application of Expanded Non-Invasive Prenatal Testing in Prenatal Diagnosis in Plateau Regions [J]. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2025, 17 (02): 7 - 12.

[10] Jiang Yulin. Unexpected Findings in Non-Invasive Aneuploidy Screening: To Report or Not to Report? [J]. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2025, 17 (02): 12.