Study on the Correlation Characteristics Between Fetal Y Chromosome Concentration and Maternal Indicators Based on a Multiple Linear Regression Model

Jian Wang*, Geqi Zhang, Zhangjunyan Li

Army Engineering University of PLA, Shijiazhuang, China, 050003

* Corresponding Author Email: wj1581777836@163.com

Abstract. This study investigates the association between fetal Y chromosome concentration and maternal indicators in non-invasive prenatal testing (NIPT). It details the modeling process for quantitative relationships between fetal Y chromosome concentration and maternal factors including gestational age, BMI, and age. The research first standardized raw data and removed outliers based on the 3 sigma criterion to ensure foundational data quality. Following data visualization and Shapiro-Wilk normality tests confirming non-normal distribution, Spearman correlation coefficients were applied. Results indicated extremely weak correlations between gestational age, BMI, and age with Y chromosome concentration. A multiple linear regression model was then constructed using OLS estimation for regression coefficients, with significance assessed via t-tests and F-tests. The test results clearly indicated that gestational age and BMI exerted highly significant effects on Y chromosome concentration (p<0.05), while the influence of age was insignificant. Given the initial model's low coefficient of determination R^2 of only 0.0522, this study successfully optimized the model by clustering participants based on age and parity. This approach improved the R^2 value to between 0.1 and 0.3, significantly enhancing the model's explanatory power and validating the rationale for refining the basic model through detailed subgrouping.

Keywords: Multiple linear regression; Y chromosome concentration; Spearman correlation coefficient.

1. Introduction

Non-invasive prenatal testing (NIPT) has become the core technology for screening fetal chromosomal abnormalities [1]. The accuracy of NIPT, particularly for detecting male fetuses, is typically determined by the concentration of fetal Y chromosome in maternal blood. However, this concentration is not constant but closely correlates with the gestational age and body mass index (BMI) of the pregnant woman. Traditional empirical BMI grouping methods struggle to accommodate significant individual variations among pregnant women, potentially leading to suboptimal timing of testing and increasing the risk of missing early treatment windows. To address the clinical demand for “precision and personalization” in NIPT technology, this study utilizes NIPT testing data from a regional cohort of high-BMI pregnant women. It aims to resolve the core issues of optimizing BMI grouping and determining the optimal testing window. This paper first focuses on investigating the associative characteristics between fetal Y chromosome concentration and maternal physiological indicators such as gestational age and BMI. The approach involves preprocessing raw data and conducting exploratory analyses to establish multivariate regression models quantifying the influence of these indicators on Y chromosome concentration. Rigorous significance testing will be performed to provide scientifically grounded quantitative relationships supporting subsequent BMI grouping optimization based on risk minimization [2-3]. This study must first address the foundational issue of constructing and validating a quantitative association model between fetal Y chromosome concentration and maternal clinical indicators [4-5]: Specifically, how should NIPT data from high-BMI pregnant women be cleaned and transformed? How can correlation analysis reveal the associative patterns between fetal Y chromosome concentration and maternal factors such as gestational age, BMI, and age? And how can a statistically significant multiple linear regression model be constructed to accurately quantify the independent effects of these maternal indicators on Y chromosome concentration?

2. Model Establishment and Solution

2.1. Model Analysis

This problem aims to explore the correlation characteristics between the fetal Y-chromosome concentration and indicators such as the pregnant woman's gestational week and BMI, construct a relational model, and test its significance. By building a multiple regression model, evaluating the significance of individual variables and the overall effectiveness of the model, we ensure that the model can truly reflect the correlation characteristics between the Y-chromosome concentration and indicators like gestational week and BMI, thereby providing support for subsequent research.

2.2. Descriptive Statistics

The data source for this study is: https://www.mcm.edu.cn/. Descriptive statistics is a basic method in statistics used to summarize, organize, and present the basic characteristics of data. Through numerical indicators (such as mean, standard deviation, and quantiles), raw data is converted into intuitively interpretable information, serving as a preliminary step for in-depth analysis (e.g., correlation analysis). The purposes of conducting descriptive statistics are to fully grasp the basic situation of the data and lay the foundation for subsequent analysis; identify outliers and unreasonable information in the data to ensure data quality; and intuitively present the distribution characteristics of variables to assist in initially forming research hypotheses [6-7].

For the 4 key variables, the main detection objectives include data completeness (for all variables), central tendency (BMI, Y-chromosome concentration), dispersion degree (gestational week, Y-chromosome concentration), and extreme values and abnormal ranges (minimum, maximum, and quantiles of all variables).

Table 1. Distribution of Descriptive Statistics.

Statistic	Maternal BMI	Gestational Week	Y-Chromosome Concentration	Gestational Days
Count	1012	1012	1012	1012
Mean	32.33	117.67	0.08	117.67
Standard Deviation	2.99	28.43	0.03	28.43
Minimum	20.70	77.00	0.01	77.00
25th Percentile	30.28	93.00	0.05	93.00
Median	31.94	112.00	0.07	112.00
75th Percentile	33.95	140.00	0.10	140.00
Maximum	46.87	203.00	0.23	203.00

By analyzing the above table 1, it is found that there are no missing values in variables such as "maternal BMI, gestational week, Y-chromosome concentration, and gestational days" in the table, indicating that the dataset used for analysis has good completeness. The BMI characteristics of the samples are consistent with the background of the problem, with a prominent proportion of the high BMI group, and the gestational week distribution covers the detection window period [8].

2.3. Data Visualization

This study conducts visualization processing of the Y-chromosome concentration and key influencing factors (gestational week, gestational days, age, BMI) to quickly verify the "variable correlation hypothesis", thereby determining the direction for subsequent quantitative analysis, and further revealing "information that cannot be reflected in numerical values". For example, a sample with a gestational age of 100 days but a Y-chromosome concentration of only 1% can be quickly identified as a case of "too early detection time" or "data error". The specific visualization results are shown in the following figure 1:

Through analyzing the above figure, it can be concluded that the Y-chromosome concentration generally shows an upward trend with the increase in gestational days; as the gestational days increase, the number of points with high concentrations gradually increases, indicating that the longer the gestational days, the higher the tendency of the Y-chromosome concentration to be. The distribution of Y-chromosome concentration among different age groups is relatively scattered without obvious changes, indicating that age has no significant impact on the Y-chromosome concentration. With the increase in BMI, the Y-chromosome concentration tends to decrease, indicating that the higher the BMI, the lower the Y-chromosome concentration may be [9-10].

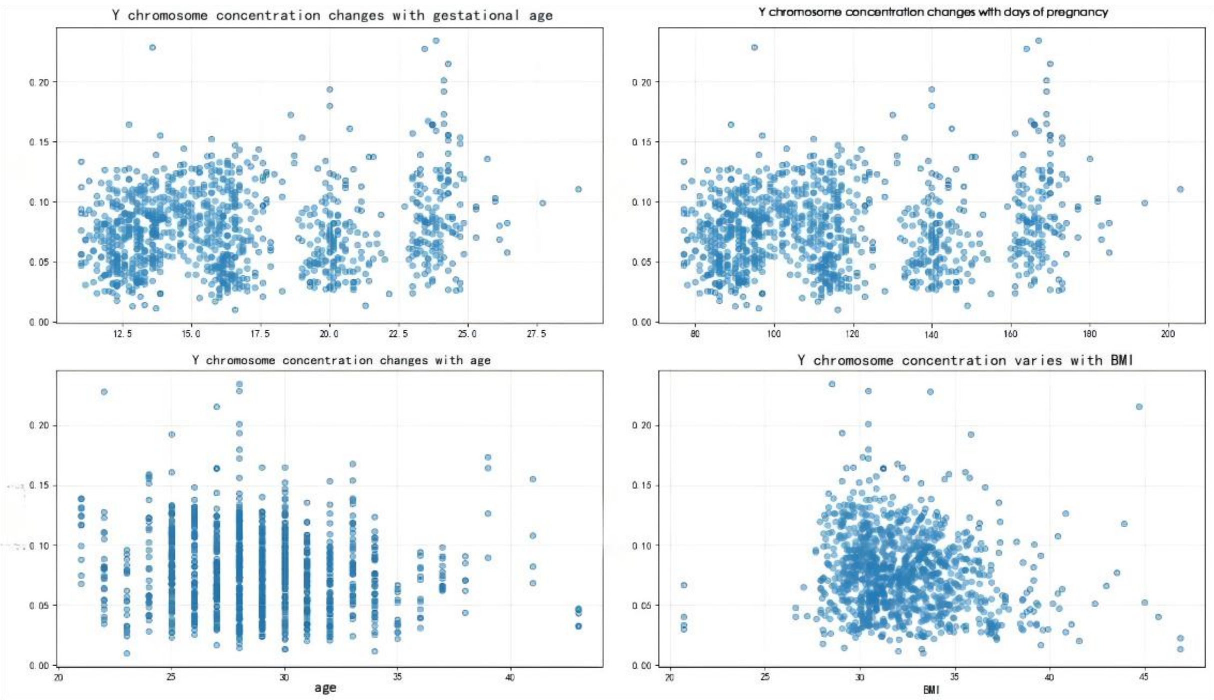

Figure 1. Scatter Plots of Y-Chromosome Concentration and Various Factors.

2.4. Standardization / Normalization Processing

Standardization / normalization is a commonly used method in data preprocessing. It mainly maps data to a specific range through linear transformation, aiming to eliminate the impact of differences in dimensions and magnitudes among different features, make the data comparable, and facilitate more stable training and analysis of subsequent models. Table 2 shows the normalization of some factors.

Table 2. Examples of Partial Standardization and Normalization.

t	Standardized t	...	Standardized B	Normalized B
83.0	-1.22	...	-1.40	0.28
141.0	0.82	...	-1.27	0.30
160.0	1.49	...	-1.14	0.31
97.0	-0.73	...	0.33	0.48
117.0	-0.02	...	0.63	0.51

Table 2 shows the results of variables t and B after standardization and normalization.

2.5. Outlier Removal

This study screens outliers for the Y-chromosome concentration and BMI values and adopts a method combining boxplots and the 3σ criterion. A boxplot displays the data distribution by plotting the minimum value, first quartile (Q1), median (Q2), third quartile (Q3), and maximum value of the data. If the data follows a normal distribution, approximately 99.7% of the data will fall within the range of the mean μ plus or minus 3 times the standard deviation σ (i.e., $\mu \pm 3\sigma$), and data beyond this range can be regarded as outliers. For the Y-chromosome concentration and BMI data, we first assume they approximately follow a normal distribution, calculate the mean and standard deviation, and then determine the points beyond the range of $\mu \pm 3\sigma$ as outliers for processing. The processing situation is shown in the following figure 2:

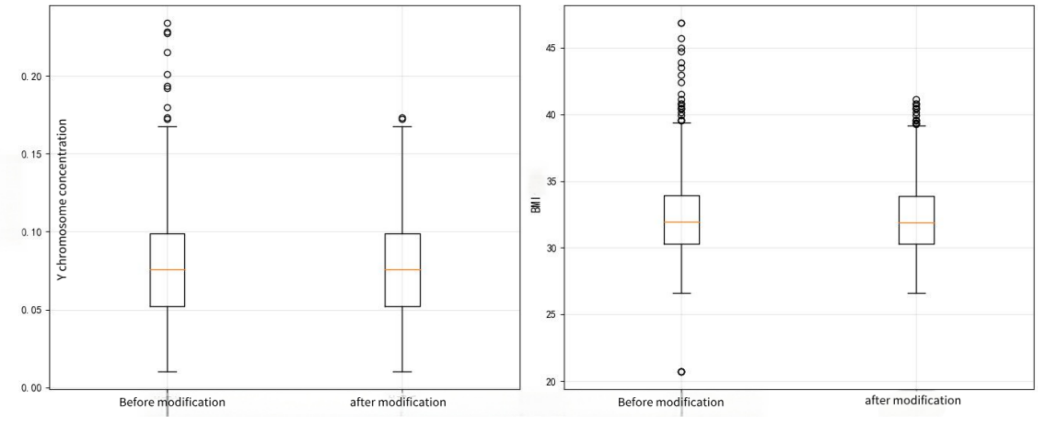

Figure 2. Outlier Processing for Y-Chromosome Concentration (Left) and BMI (Right).

Since problem focuses on analyzing the correlation characteristics between the fetal Y-chromosome concentration and indicators such as the pregnant woman's gestational week and BMI, only data related to male fetuses is retained for the analysis of this problem. To further ensure data quality and the accuracy of the analysis results, additional key steps are added for the research: first, normalize the Y-chromosome concentration, BMI value, gestational days, and maternal age to eliminate the interference of different indicators on the correlation analysis; then, use the 3σ test and boxplot test to detect outliers in the normalized Y values and BMI data, and remove the two types of abnormal data according to the test results, laying a data foundation for constructing a reliable relational model between the Y-chromosome concentration and related indicators. Data retention rate after the 3σ test: 97.92%.

2.6. Correlation Test

The normal distribution is a common continuous probability distribution, and its probability density function presents a symmetric bell-shaped curve. Its core characteristics are that the mean (μ) determines the central position of the curve, and the standard deviation (σ) determines the degree of dispersion of the curve. Most data are concentrated near the meaning, and the data on both sides decreases symmetrically. This study uses normal distribution to provide a prerequisite basis for subsequent statistical analysis; the Shapiro-Wilk test is a method commonly used to test whether data follows a normal distribution and performs well in the normal test of small-sample data (usually with a sample size $n \leq 50$). Its test steps are as follows: first, propose a hypothesis, where the null hypothesis (H_0) is that the data follows a normal distribution, and the alternative hypothesis (H_1) is that the data does not follow a normal distribution; then calculate the W statistic, determine the p-value, and finally make a decision by comparing the p-value with the pre-set significance level α: if the p-value is greater than α, the null hypothesis is not rejected, that is, the data is considered to follow a normal distribution; if the p-value is less than α, the null hypothesis is rejected, that is, the data is considered not to follow a normal distribution. The test results are shown in Table 3:

Table 3. Shapiro-Wilk Test Results.

Variable	W Statistic	p-Value	Distribution Type
Y	0.9833	0.00001	Non-normal Distribution
t	0.9171	0.00001	Non-normal Distribution
days	0.9171	0.00001	Non-normal Distribution
B	0.9657	0.00001	Non-normal Distribution
A	0.9722	0.00001	Non-normal Distribution

Typically, the Pearson product-moment correlation coefficient is used when there is a linear correlation between two continuous variables, and the Spearman correlation coefficient is used to describe the correlation when the applicable conditions of the product-moment correlation analysis are not met. It can be observed from the above table that the test results do not follow a normal distribution, so the Spearman statistical method is selected. The results are shown in the following table 4:

Table 4. Relationship Between Each Variable and Y-Chromosome Concentration.

Variable	Correlation Direction	Spearman Correlation Coefficient	p-Value
t	Very Weak Positive Correlation	0.0743	0.0194
days	Very Weak Positive Correlation	0.0743	0.0194
B	Very Weak Negative Correlation	-0.1556	0.0000
A	Very Weak Negative Correlation	-0.1047	0.0010

Among them, variables t and days have a very weak positive correlation with the Y-chromosome concentration, while variables B and A have a very weak negative correlation with the Y-chromosome concentration.

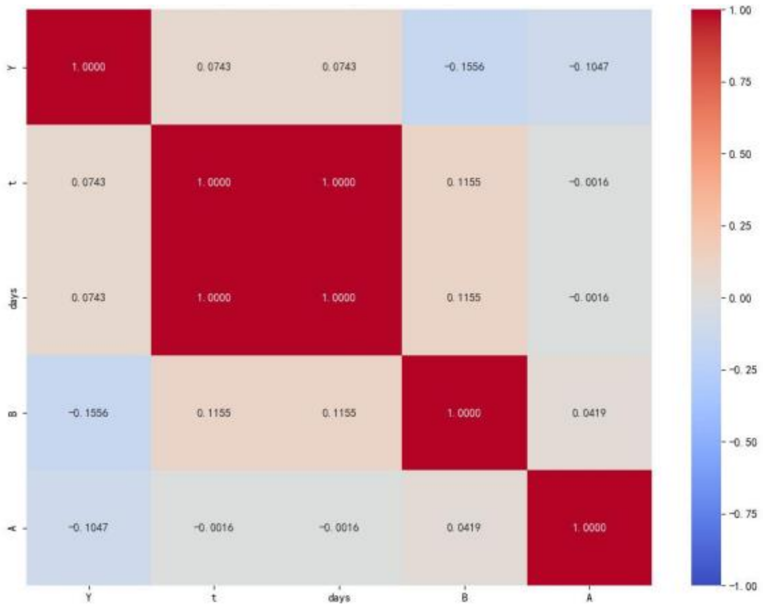

Figure 3. Heatmap of Spearman Correlation Coefficients Among Variables.

In figure 3, red represents a positive correlation and blue represents a negative correlation. It can be analyzed from the above figure that variables t, days, B, and A all have a certain Spearman correlation with the Y-chromosome concentration.

2.7. Construction of Multiple Linear Regression Model

Let Y be the Y-chromosome concentration (after normalization), and the gestational week value, maternal BMI, and age be independent variables. It is assumed that there is a linear relationship between the Y-chromosome concentration and the gestational week, BMI, and age. The model estimates each regression coefficient through the least squares method, reflecting the independent impact of each independent variable on the Y-chromosome concentration. Then the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 \times t + \beta_2 \times B + \beta_3 \times A + \varepsilon \tag{1}$$

Where Y is the Y-chromosome concentration, t is the gestational week value, B is the maternal BMI value, A is the age, β_0 is the intercept term, $\beta_1, \dots, \beta_p$ are the regression coefficients (indicating the impact intensity of the independent variables on Y), and ε is the random error term (satisfying $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$).

2.7.1. Model Solution

Estimating regression coefficients: Based on the sample data, the least squares method (OLS) is used to solve the coefficients. The core is to minimize the sum of the squares of the error terms, that is, to minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (where $\hat{y}_i$ is the predicted value). Through matrix operations or partial derivative calculation and derivation, the estimated values of the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are obtained, forming the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$. The final solving formula is:

2.7.2. Significance Test

After constructing the relational model between the fetal Y-chromosome concentration and indicators such as the pregnant woman's gestational week and BMI, in order to verify the statistical significance of the model and each variable in the model, the T-test and F-test are used to conduct parameter significance test and overall model significance test respectively.

The core purpose of the T-test is to determine whether the regression coefficient of a single independent variable (such as gestational week, BMI, age) in the model is significantly non-zero, that is, whether the variable has a statistically significant impact on the Y-chromosome concentration. According to the multiple linear regression theory, the formula for the T-statistic of the regression coefficient β_i is:

$$t_i = \frac{\beta_i - 0}{s_{\beta_i}} \tag{2}$$

Where β_i is the estimated value of the regression coefficient of the i-th variable, and s_{β_i} is the standard error of β_i, which can be calculated through the sample data and the model residuals.

The F-test is used to determine whether the constructed multiple regression model is significant overall, that is, whether there is a linear relationship between all independent variables combined and the Y-chromosome concentration. The formula for the F-statistic based on the "regression sum of squares (SSR)" and "residual sum of squares (SSE)" of the model is:

$$F = \frac{SSR/k}{SSE/(n-k-1)} \tag{3}$$

Where SSR reflects the degree of explanation of the variation of the dependent variable by the independent variables, and SSE reflects the impact of random errors on the variation of the dependent variable; k is the number of independent variables, and n is the sample size.

Table 5. Test Results.

Variable	Coefficient	Standard Error	t-Value	p-Value	95% Confidence Interval	Significance
Constant Term	0.1576	0.017	9.321	0.000	[0.124, 0.191]	Highly Significant
Gestational Week	0.0001	4.17×10^{-5}	2.798	0.005	[3.58×10^{-5}, 0.000]	Significant
BMI	-0.0024	0.000	-5.448	0.000	[-0.003, -0.002]	Highly Significant
Age	-0.0006	0.000	-1.845	0.065	[-0.001, 3.82×10^{-5}]	Not Significant

It can be seen from the above table 5 that the p-values of the constant term, gestational week, and BMI are all less than 0.05, which are highly significant or significant, indicating that in the basic linear model, the constant term, gestational week, and BMI have a significant impact on the dependent variable; while the p-value of age is 0.065, which is not significant, and its impact on the dependent variable is not obvious.

2.7.3. Model Goodness-of-Fit Evaluation

Goodness-of-fit is a key indicator for evaluating the fitting effect of the model. The closer the R^2 value is to 1, the better the fitting effect of the model. Its calculation formula is $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$. The calculated result is 0.0522, indicating that the fitting effect is not good. This may be due to excessive individual differences leading to excessive discretization of the sample data, making it difficult to achieve a good fitting effect.

2.7.4. Model Residual Verification

After constructing the multiple linear regression model, the quadratic polynomial regression model with interaction terms, and the exponential multiple nonlinear regression model between the fetal Y-chromosome concentration and indicators such as the pregnant woman's gestational week, BMI, and age, it is necessary to evaluate the fitting effect and generalization ability of each model through multi-dimensional verification methods, and select the optimal model that best fits the data characteristics and can accurately reflect the correlation between variables, providing reliable support for subsequent result interpretation and practical application. Combined with the research objectives and data characteristics, the residual analysis method is selected for model verification.

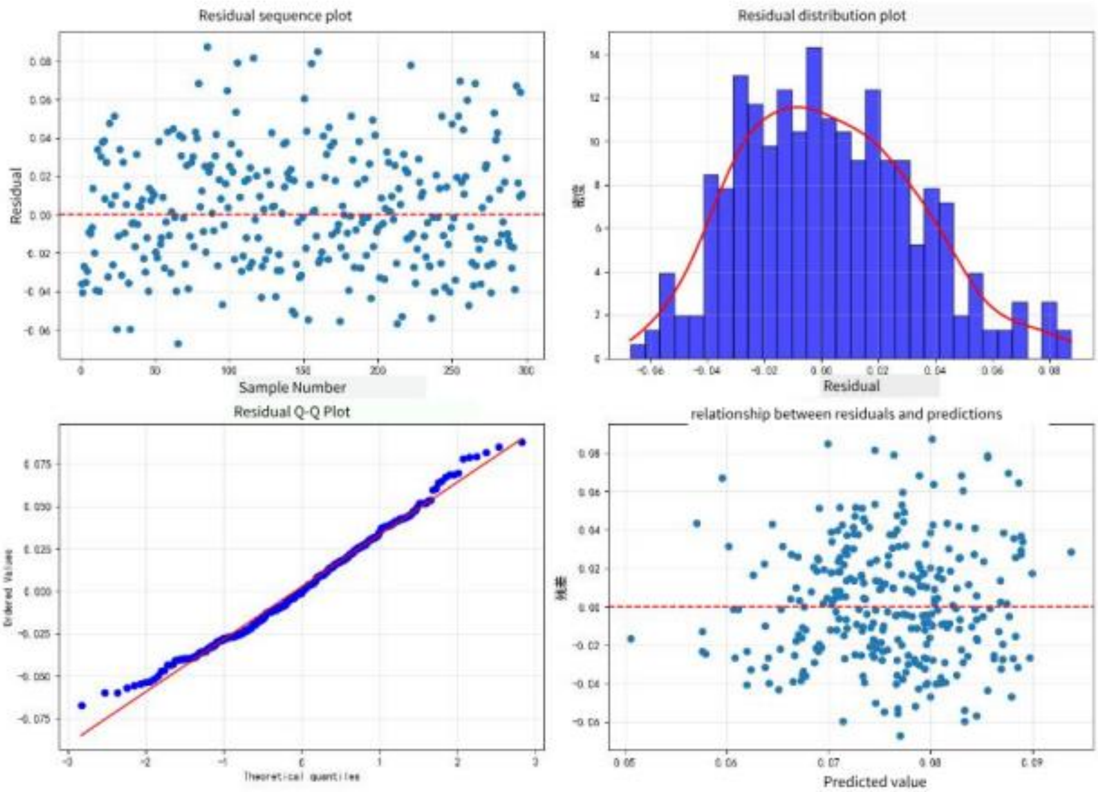

Figure 4. Residual Analysis.

The analysis of the four residual figure 4 is as follows: in the residual sequence graph, the residual points are randomly distributed around 0 without obvious regular trends; the residual distribution graph is approximately normally distributed, and the peak is near 0; in the residual Q-Q graph, most points are close to the theoretical quantile reference line, confirming the normality of the residuals; in the graph of residuals versus predicted values, the residual points are evenly distributed on both sides of 0, showing no trend related to the predicted values. Overall, the residuals meet the assumptions of randomness, normality, and homoscedasticity of linear regression, indicating that the fitting effect of the basic multiple linear models on the Y-chromosome concentration basically meets the requirements.

2.7.5. Result Interpretation

Through the above complete solution process, the entire process from raw data preprocessing to final model selection and application is realized, and it is considered that the basic model, namely the multiple linear regression equation, is the best model. However, the goodness-of-fit and residual graphs show that the fitting effect of this model is relatively poor. Therefore, this study adopts a grouping method to further analyze its correlation and relational model.

2.8. Grouped Fitting Optimization

Considering the excessive individual differences among pregnant women, cluster analysis is now performed based on the pregnant women's age and number of deliveries. The grouping situation is shown in figure 5:

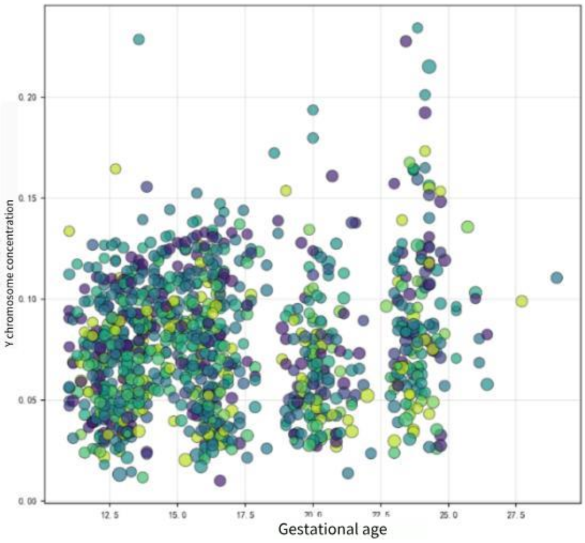

Figure 5. Schematic Diagram of Cluster Grouping.

After fitting regression, significance test, and goodness-of-fit evaluation, the following results are obtained in figure 6:

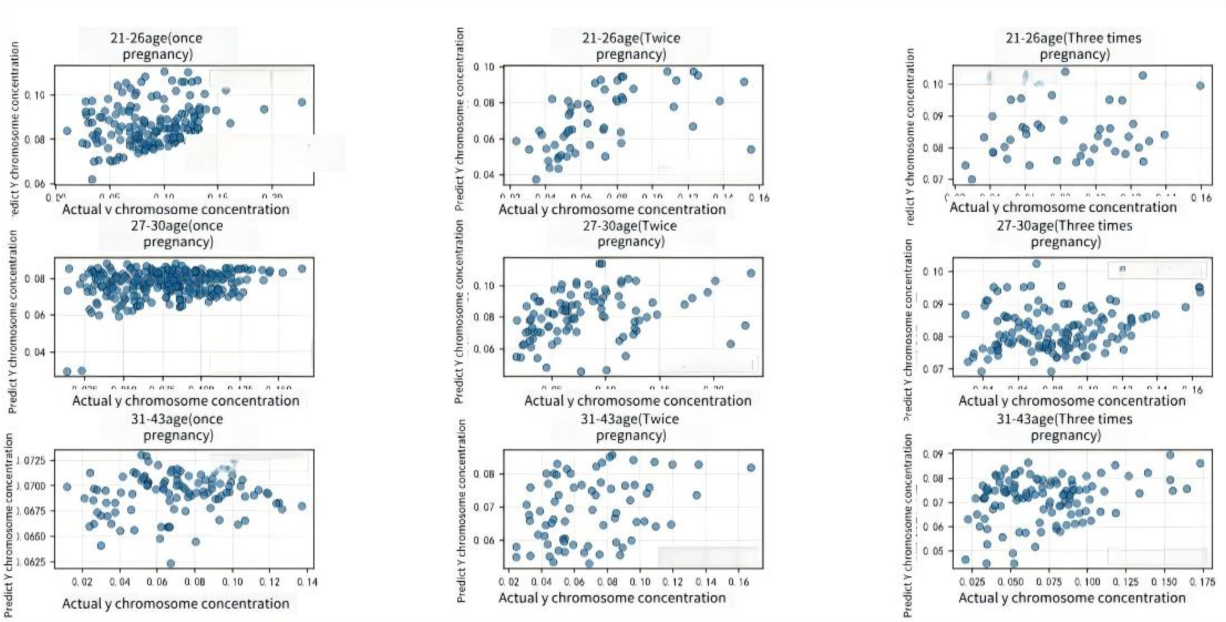

Figure 6. Schematic Diagram of Fitting Results of Each Group.

Through cluster grouping, it is found that the goodness-of-fit within some groups is significantly improved. Among them, the $R^2 = 0.296$ for women aged 21-26 with 2 pregnancies, indicating that the model's prediction of the Y-chromosome concentration is enhanced under the grouping of different ages and numbers of pregnancies.

3. Conclusions

This study successfully constructed a multiple linear regression model with gestational age, BMI, and maternal age as independent variables to quantify their effects on fetal Y chromosome concentration. Although the model exhibited low initial goodness-of-fit, residual analysis validated its structural compliance with basic linear regression assumptions. A key finding was the statistically highly significant impact of gestational age and BMI on Y chromosome concentration, providing clear focus variables for subsequent research. To overcome insufficient fit due to data dispersion, this study adopted an optimized strategy of clustering by age and parity. This significantly enhanced the model's explanatory power, demonstrating that the underlying model structure, when combined with refined grouping, better addresses individual variations among high-BMI pregnant women. This establishes a reliable quantitative foundation for determining the optimal NIPT testing timing to minimize risks.

References

[1] Pan X, Weng L, Pan Y, et al. Borderline Z-scores in non-invasive prenatal screening: does its presence hold clinical significance? [J]. Archives of Gynecology and Obstetrics,2025, (prepublish):1-9.

[2] Yamamoto K, Suzumori N, Miura K, et al. Clinical Implications of Low Cell-Free DNA Fetal Fraction in Non-Invasive Prenatal Testing: A Retrospective Cohort Study of 40,716 Pregnancies. [J]. Prenatal diagnosis,2025.

[3] Wang C, Zhang H, Liu Y, et al. Results and indications changes by prenatal diagnosis of 3,458 pregnant women: a 10-year retrospective study. [J]. Annals of medicine,2025,57(1):2563753.

[4] Yamashita Y, Shirato N, Ishii T, et al. Perinatal outcomes of fetal CNVs detected by genome-wide non-invasive prenatal testing in Japan[J]. Journal of Human Genetics,2025, (prepublish):1-9.

[5] Sun F, Zhou Y, Zhao X, et al. Application of optimized non-invasive prenatal testing for thalassemia based on change of haplotype doses. [J]. Journal of genetics and genomics = Yi chuan xue bao,2025.

[6] Stevens C J S, Dijk V W, Souren Y N, et al. Clinically Irrelevant Terminal 16q21 Deletion Detected by NIPT Is Attributable to Inherited Fragility at FRA16B.[J]. American journal of medical genetics. Part A,2025, e64271.

[7] Zhou D Z, Zhou Y G, Russell M A, et al. Disorder–Order Phase Transformation in Ni50Pt50 and Ni20Pt45Al35 (at %) Alloys[J]. Physics of Metals and Metallography,2025,126(6):651-656.

[8] Ryan H, Lambert M, Loingsigh N S, et al. Evaluation of 6 years of fetal RHD screening in Ireland: From implementation to practice. [J]. Vox sanguinis,2025.

[9] Eekhout V A C J, Bax J C, Schuurman P V L, et al. Performance of non-invasive prenatal testing in vanishing-twin and multiple pregnancies: results of TRIDENT-2 study. [J]. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology,2025.

[10] Li S, Yue F, Tan H, et al. Tailoring the Effects of Spin State and Intermediate Hydrogen Adsorption on NiPt/Ni Bridge Sites toward Robust Acidic Water Electrolysis. [J]. Nano letters,2025.