International Journal of Public Health and Medical Research
ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 6, Number 2, Year 2026
DOI: https://doi.org/10.62051/ijphmr.v6n2.05
Journal homepage: https://wepub.org/index.php/IJPHMR/index

Quantitative Study on Factors Influencing Male Fetal Y Chromosome Concentration Based on Linear Mixed-Effects Models

Xin Zhao *

School of Mathematics and Artificial Intelligence, Chongqing University of Arts and Sciences, Chongqing, 402160, China

*Corresponding Author: jyzx959595@163.com

ABSTRACT

This study aims to quantify the association between fetal Y chromosome concentration and key indicators such as maternal gestational age and body mass index to inform non-invasive prenatal testing strategies. First, fetal testing data underwent rigorous preprocessing, including exclusion of concentrations below 4%, removal of missing and outlier values, and conversion of gestational age formats to continuous weeks suitable for continuous numerical variables. Exploratory data analysis revealed a wide range of fetal Y chromosome concentration values at identical or similar gestational ages, indicating stratified clustering and nested structures—significant inter-individual variation among pregnant women. Consequently, a linear mixed-effects model with random intercepts was constructed, incorporating gestational age, BMI, and their interaction term as fixed effects, while treating individual variation as random intercepts. The model fitted 541 observations from 236 pregnant women. The random intercept variance was significantly non-zero, confirming substantial individual variation influencing concentration variability. Significance tests for fixed effects revealed a significant negative main effect of BMI on Y chromosome concentration, along with a significant interaction effect between gestational week and BMI, confirming that BMI amplifies the influence of gestational week on Y chromosome concentration. The model achieved a conditional coefficient of determination of 0.725, but the marginal coefficient of determination was only 0.117, suggesting the need to incorporate additional important fixed effect factors.

KEYWORDS

Linear mixed-effects model; Y chromosome concentration; Repeated measures data

1. INTRODUCTION

Non-invasive prenatal testing (NIPT) serves as a critical prenatal screening technique, assessing fetal health by detecting cell-free fetal DNA in maternal blood, thereby playing a vital role in safeguarding newborn health. For male fetuses, the accuracy of NIPT testing largely depends on whether the Y chromosome concentration reaches or exceeds the 4% threshold [1]. Extensive clinical observations indicate significant correlations between maternal indicators—such as Body Mass Index (BMI), gestational age, and maternal age—and fetal DNA concentration in maternal blood. Particularly in high-BMI pregnancies, insufficient fetal DNA levels may cause testing failure, leading to missed intervention windows and increased pregnancy risks [2-4].

Although these associations are widely recognized, existing studies exhibit notable limitations in quantitatively analyzing the relationship between these factors and Y chromosome concentration. First, most research focuses on the independent influence of a single factor (e.g., BMI or gestational

age), failing to adequately account for potential interactions between factors (e.g., whether BMI modulates the effect of gestational age on Y chromosome concentration) [5, 6]. Second, and more critically, such studies typically involve multiple repeated measurements from the same pregnant woman, yielding hierarchical data (where repeated measurement time points are nested within individuals). Traditional linear regression models require independence between data points when handling such repeated measures, which clearly contradicts the reality of clinical data. Ignoring the correlation between measurements within individuals and baseline differences between individuals (i.e., random effects) may lead to biased estimates of fixed effects (such as the effects of gestational age or BMI) and reduce the accuracy of statistical inference [7,8]. In other words, existing analytical methods fail to effectively isolate and quantify the variability stemming from individual differences among pregnant women. Consequently, they struggle to accurately reveal the net effect of fixed factors like gestational age and BMI on Y chromosome concentration.

This study aims to address these methodological shortcomings by refining modeling strategies, thereby more precisely quantifying the statistical association between male fetal Y chromosome concentration and key indicators such as gestational age and BMI. The innovation of this section manifests in three primary aspects:

First, during data preprocessing, we systematically and rigorously cleaned male fetal detection data. This included excluding data with Y chromosome concentrations below 4%, handling missing and outlier values, and uniformly converting gestational age formats to weeks suitable for continuous variables, thereby establishing a high-quality data foundation for subsequent modeling.

Second, through exploratory data analysis (including scatterplot visualization and distribution examination), we clearly identified hierarchical clustering and nested structures within the data. Specifically, Y chromosome concentrations exhibited significant variability within the same gestational age or BMI level, with data points showing pronounced intra-individual clustering. This provided intuitive and robust justification for selecting the Linear Mixed-Effects Model (LMM).

Third, we constructed a linear mixed-effects model incorporating random intercepts. This model innovatively treated gestational age, BMI, and their interaction term as fixed effects while including individual variation among pregnant women as random effects. This approach effectively controls for interference from inter-individual differences, enabling a purer estimation of fixed effects and resolving estimation biases caused by traditional models' neglect of hierarchical data structures [9, 10].

Based on this, the study follows the technical workflow of "data preprocessing → exploratory analysis → model building and estimation." First, data cleaning and variable transformations were performed. Subsequently, exploratory analysis confirmed the data distribution patterns and hierarchical structure. Finally, the linear mixed-effects model was established and estimated, with significance analysis conducted via t-tests and F-tests, followed by a comprehensive model evaluation. This study not only provides more reliable quantitative results for understanding the factors influencing Y chromosome concentration but also offers methodological references for handling similar clinical data with repeated measurements.

2. MODEL ESTABLISHMENT AND SOLUTION

2.1. Data Preprocessing

2.1.1. Removal of Missing Values

The data is from https://www.mcm.edu.cn/. Only male fetal data is considered here. In data cleaning, we first delete data with missing values.

2.1.2. Outlier Handling

Secondly, as described in the problem, the NIPT result is considered basically accurate only if the Y-chromosome concentration of male fetuses is 4% or higher. Therefore, we only retain data where the Y-chromosome concentration is greater than or equal to 4%.

GC content is an important indicator for evaluating sequencing data quality. GC content lower than 0.4 or higher than 0.6 is regarded as sequencing failure, so we retain data with GC content between 0.4 and 0.6.

2.1.3. Conversion and Processing of Gestational Week Units

For the gestational weeks of pregnant women, the original format is "weeks + days". In clinical practice and related research of obstetrics, it is customary to record gestational weeks as the unit, but this is not conducive to quantitative modeling. To convert it into a linear model more suitable for continuous numerical variables, we use the following model for conversion:

$$\text{Continuous gestational week} = \text{Weeks} + \frac{\text{Days}}{7} \tag{1}$$

Since pregnant women can have the fetal chromosome concentration detected between 10 and 25 weeks of gestation, we process the gestational week data and exclude data less than 10 weeks and greater than 25 weeks [5-6].

2.1.4. Outlier Detection Using IQR Method

For numerical variables, we use the IQR method to identify outliers. Here, Q1 is the first quartile (25th percentile), Q3 is the third quartile (75th percentile), with the upper limit calculated as $Q1 - 1.5 \times IQR$, the lower limit as $Q3 + 1.5 \times IQR$, and $IQR = Q3 - Q1$. Data outside this range is considered an outlier.

For Y-chromosome concentration:

Lower limit of outliers: $Q1 - 1.5 \times IQR = 0.0006024 - 0.00059985 = 0.00000255$

Upper limit of outliers: $Q3 + 1.5 \times IQR = 0.0010023 + 0.00059985 = 0.00160215$

$$IQR = Q3 - Q1 = 0.000399$$

For gestational weeks:

Lower limit of outliers: $Q1 - 1.5 \times IQR = 13.29 - 10.065 = 3.225$ weeks

Upper limit of outliers: $Q3 + 1.5 \times IQR = 20.00 + 10.065 = 30.065$ weeks

$IQR = Q3 - Q1 = 6.71$ weeks

For BMI:

Lower limit of outliers: $Q1 - 1.5 \times IQR = 30.18 - 4.92 = 25.26$

Upper limit of outliers: $Q3 + 1.5 \times IQR = 33.46 + 4.92 = 38.38$

$$IQR = Q3 - Q1 = 3.28$$

Boxplot drawing in figure 1:

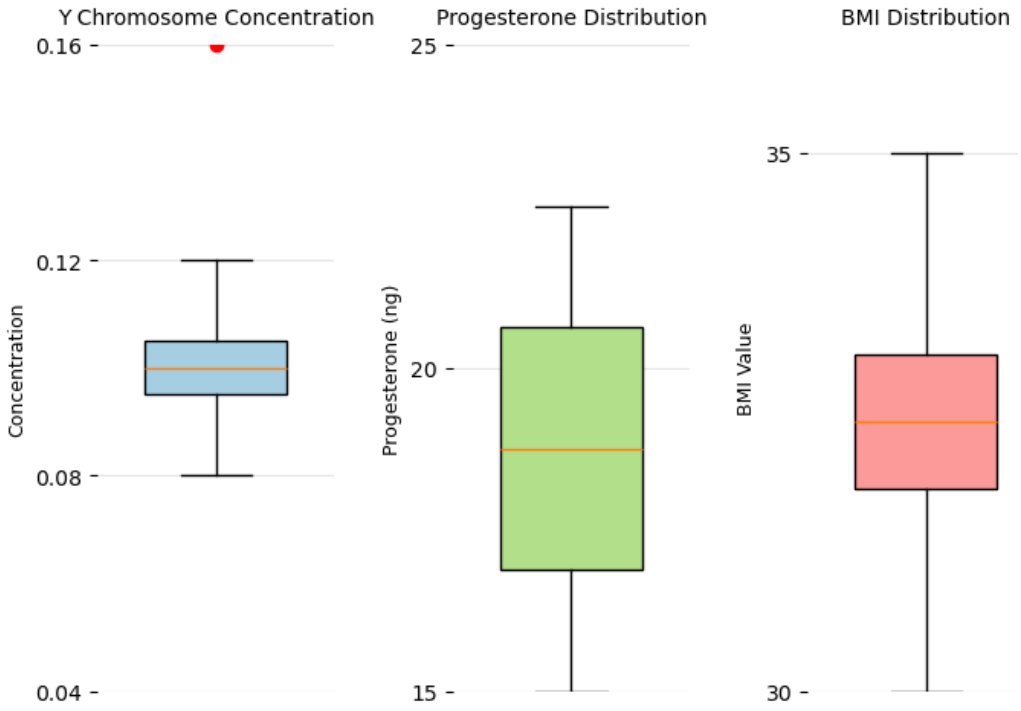

Figure 1. Boxplot of Y-chromosome concentration, gestational week, and BMI distribution

2.2. Exploratory Data Analysis and Modeling Basis

2.2.1. Global Linear Correlation Analysis

To identify the dependent variable (fetal Y-chromosome concentration), find independent variables with strong correlation, and select more appropriate predictor variables for model establishment, we also need to exclude the problem of reduced model stability and interpretability caused by excessively high correlation between two independent variables. To more intuitively view the relationships between numerical variables in the dataset, we calculate the correlation coefficients of numerical data in the male fetal detection data and draw a heatmap of the correlation relationships [7-8]:

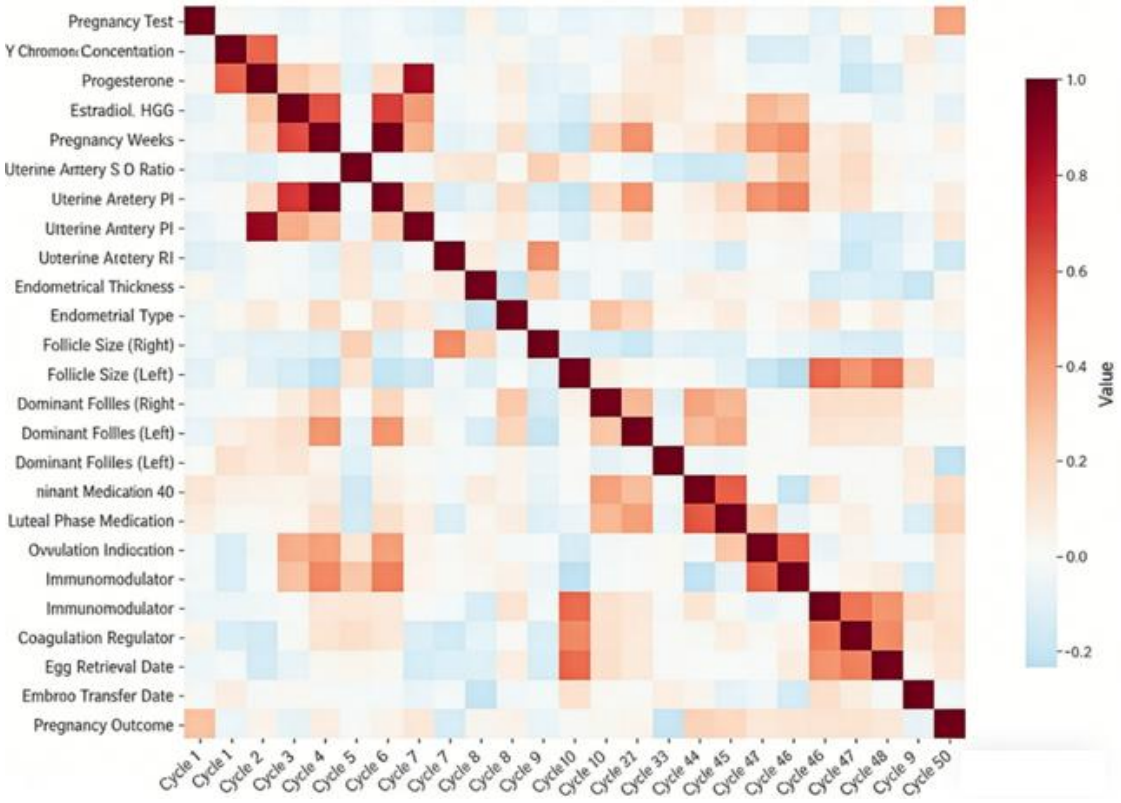

Figure 2. original heatmap data

From the figure 2, we can see that the data most correlated with Y-chromosome concentration is X-chromosome concentration (0.57), followed by the number of detection blood draws (0.37). In the analysis of independent variable correlation data, we can see that there is a very strong correlation between a pregnant woman's BMI and weight (0.78), which is related to the BMI calculation formula: $\mathrm{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}$.

The correlation coefficients between Y-chromosome concentration (the key variable we focus on exploring) and gestational weeks, as well as BMI, are 0.37 and 0.05 respectively. Although there is a positive correlation trend, the relationship is not strong, and changes in Y-chromosome concentration cannot be explained by these two variables alone.

It is worth noting that the correlation coefficients between the number of detection blood draws and a pregnant woman's BMI, as well as Y-chromosome concentration, are 0.16 and 0.3 respectively. This indicates that the number of blood draws will affect a pregnant woman's BMI and Y-chromosome concentration to a certain extent. Therefore, the number of blood draws will have a certain impact on the correlation between Y-chromosome concentration and a pregnant woman's gestational weeks and BMI that we aim to explore [9-10].

2.2.2. Scatter Plots and Analysis of Core Relationships

To further explore the correlation between Y-chromosome concentration and gestational weeks, as well as BMI, and then conduct reasonable modeling, we use R to draw scatter plots of Y-chromosome concentration versus detection gestational weeks and Y-chromosome concentration versus pregnant women's BMI:

From the figures, we can see that the scatter plots of Y-chromosome concentration versus gestational weeks and BMI do not show obvious linear or non-linear trends, and the data distribution is relatively scattered. This indicates that there may be other potential factors influencing the relationship between Y-chromosome concentration and these two variables.

Moreover, according to the information in the figures, the data points are not completely randomly distributed. Among them, there are multiple vertical data clusters in Figure 3 (Scatter plot of Y-chromosome concentration versus gestational weeks), meaning that within the same or similar gestational weeks, the range of fetal Y-chromosome concentration values is relatively wide. This implies that at the same detection week, the fetal Y-chromosome concentration will vary greatly due to individual differences.

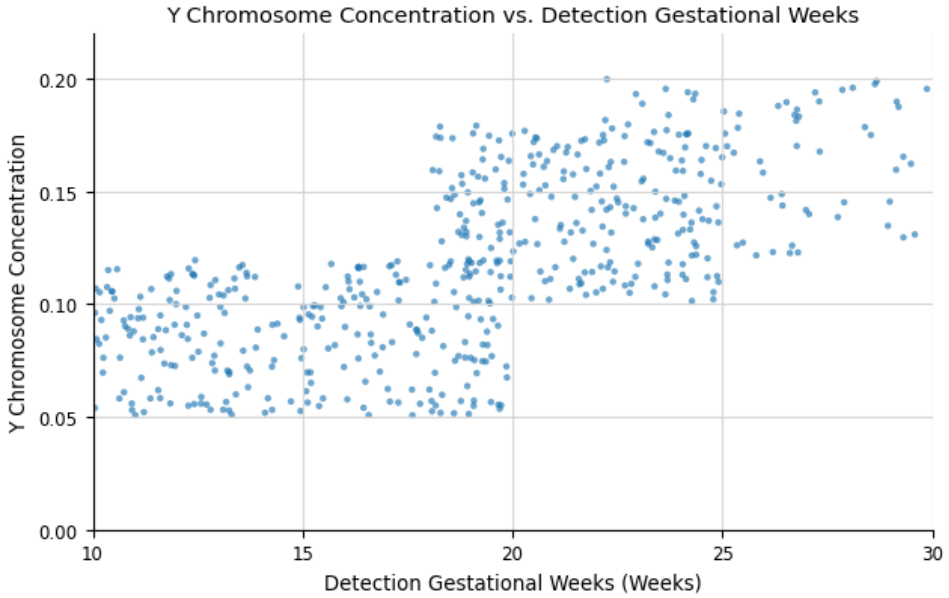

Figure 3. Scatter plot of Y-chromosome concentration versus gestational weeks

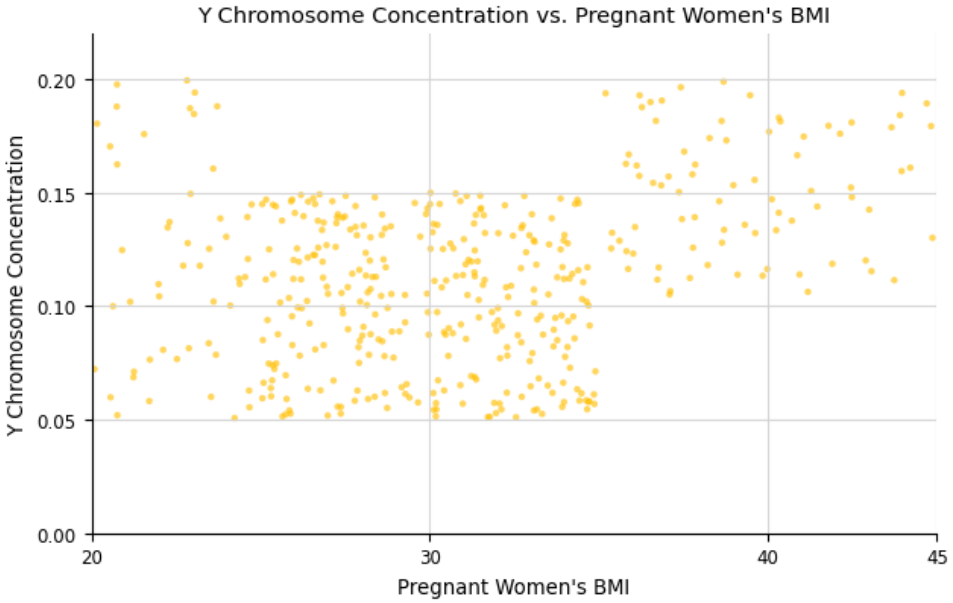

Figure 4. Scatter plot of Y-chromosome concentration versus pregnant women's BMI

Figure 4 (Scatter plot of Y-chromosome concentration versus pregnant women's BMI) also shows certain data clustering. Similar to Figure 3, at the same or adjacent BMI values, there are significant differences in Y-chromosome concentration.

This indicates that the data has a hierarchical clustering phenomenon, which in turn suggests that the data has a nested structure. That is, repeated measurement time points are nested within individual pregnant women, and there are significant inter-individual differences among different pregnant women.

In summary, since the data in this study comes from repeated measurements of multiple different pregnant women at different time points, the data has a certain hierarchical structure. Due to multiple measurements of the same individual in the data, the results are somewhat correlated. Therefore, the data is not suitable for traditional linear regression models. A linear mixed-effects model that can handle both fixed effects (gestational weeks, BMI, etc.) and random effects (differences among individual pregnant women, differences among groups of pregnant women) should be established for analysis.

2.3. Establishment of Linear Mixed-Effects Model

2.3.1. Model Explanation

Based on the above analysis, we choose to use a linear mixed-effects model to fit the data. A linear mixed-effects model with random intercepts, also known as a hierarchical linear model, is mainly used for longitudinal data analysis, multi-level data analysis, and repeated measurement data analysis. The basic formula of the model is:

$$Y = X \times \beta + Z \times \mu + \varepsilon \tag{2}$$

Where Y is the response vector, $X \times \beta$ represents fixed effects, $Z \times \mu$ represents random effects, and ε denotes residual errors.

The basic idea of this model is to introduce random effects in addition to fixed effects (gestational weeks, BMI) to address the impact of random variations among individual pregnant women. This model can better simulate the hierarchical or grouped structure in the data and fit the nested structure in the data well.

Fixed effects represent the average trend between independent variables and the dependent variable. In this problem, they represent the fixed impact of gestational weeks and BMI on fetal Y-chromosome concentration.

Random effects refer to the variations caused by the randomness of the sample units themselves in the sample. In this problem, since there are differences in the baseline Y-chromosome concentration among different pregnant women, each pregnant woman will have an estimated random intercept.

The mathematical expression of the linear mixed-effects model established based on this problem is:

$$Y_{ij} = \beta_0 + \beta_1 \times W_{ij} + \beta_2 \times BMI_{ij} + \beta_3 \times W_{ij} \times BMI_{ij} + \mu_{0_i} + \varepsilon_{ij} \tag{3}$$

Where:

Y_{ij} represents the fetal Y-chromosome concentration measured for the j-th time in the i-th pregnant woman.

β_0 represents the fixed effect intercept.

β_1, β_2, and β_3 represent fixed effect slopes, which are the coefficients of gestational weeks, BMI, and the interaction term between gestational weeks and BMI, respectively.

μ_{0_i} is the random intercept of the i-th pregnant woman, representing individual differences, i.e., the deviation of the Y-chromosome concentration of the individual pregnant woman from the average Y-chromosome concentration of all pregnant women.

ε_{ij} represents the residual error during the j-th measurement of the i-th pregnant woman, indicating intra-individual random errors.

Additionally, the random intercept μ_{0_i} and residual error ε_{ij} follow a normal distribution $N(0, \sigma^2)$.

2.3.2. Model Results

The model is fitted using the maximum likelihood estimation method. The data includes 541 observations from 236 pregnant women.

(1) Random Effects Results

The variance of the random intercept is 0.0006174, and the residual variance is 0.0002793. The non-zero variance of the random intercept is significant, indicating that there are significant differences in Y-chromosome concentration among individual pregnant women. Based on this analysis, the use of a mixed-effects model is reasonable. The small residual variance indicates that the model captures most of the random errors.

(2) Fixed Effects Results

Using a linear mixed-effects model (LMM) with Y-chromosome concentration as the dependent variable, detection gestational weeks, pregnant women's BMI, and their interaction term as fixed effects, and pregnant woman code as the random effect (random intercept), the relationship model is obtained as follows:

$$Y_{ij} = 0.1608 - 0.002941 \times W_{ij} - 0.003811 \times BMI_{ij} + 0.0001738 \times W_{ij} \times BMI_{ij} + \mu_{0_i} + \varepsilon_{ij} \quad (4)$$

From this, we can initially analyze that the coefficient of BMI is negative and significant at the 0.05 level, indicating that the higher the BMI, the lower the Y-chromosome concentration.

Table 1 presents the fixed effect coefficients, standard errors, degrees of freedom, t-values, and p-values.

Table 1. Fixed effect estimation results

Effect term	Estimate	Standard error	Degrees of freedom	t-value	p-value
Intercept	0.1608	0.05084	516.8	3.164	0.00165
Detection gestational week	-0.002941	0.002566	412.5	-1.146	0.25241
Pregnant woman's BMI	-0.003811	0.001605	523.4	-2.374	0.01796
Detection gestational week & Pregnant woman's BMI	0.0001738	0.00008015	420.7	2.169	0.03067

2.3.3. Significance Test Analysis (t-test and F-test)

(1) t-test

The t-test is used to determine whether each independent variable in the model has a significant impact on the dependent variable. The expression is:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim t(n - k - 1) \quad (5)$$

The null hypothesis (H_0) is: the coefficient of the variable = 0. According to Table 1, the hypothesis test of fixed effects shows:

The t-value of the intercept is 3.164, and the p-value is 0.00165 ($p < 0.00165$), which passes the significance test, indicating that the set value of the baseline is reasonable.

The p-value of the main effect of detection gestational weeks is 0.25241 (> 0.05), and the t-value is -1.146, which fails to pass the significance test. The impact of detection gestational weeks on Y-chromosome concentration is not significant, indicating that no independent linear impact of gestational weeks has been found yet.

The t-value of the pregnant woman's BMI is -2.374, and the coefficient p-value is 0.01796 ($p < 0.05$), indicating that under the condition of controlling for gestational weeks, the negative main effect of BMI on Y-chromosome concentration is statistically significant.

The t-value of the interaction term between detection gestational weeks and pregnant woman's BMI is 2.169, and the coefficient p-value is 0.03067 ($p < 0.05$), which indicates that there is a significant interaction effect between gestational weeks and BMI. BMI amplifies the impact of gestational weeks on Y-chromosome concentration.

(2) F-test

The F-test is used to determine the overall significance of the model and test whether the independent variables together have a significant impact on the dependent variable. The expression is:

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1) \tag{6}$$

Analysis of Variance (ANOVA) verification table 2:

Table 2. ANOVA results of fixed effects

Effect term	Sum of squares	Mean square	NumDF	DenDF	F-value	p-value
Detection gestational week	0.0003669	0.0003669	1	412.53	1.3136	0.25241
Pregnant woman's BMI	0.0015741	0.0015741	1	523.37	5.6356	0.01796
Detection gestational week & Pregnant woman's BMI	0.0013135	0.0013135	1	420.7	4.7028	0.03067

The consistent results obtained from the t-test and F-test confirm the significance of the main effect and interaction effect of BMI again, indicating that after adding the interaction term, the model fit has a statistically significant improvement. There is a significant interaction between a pregnant woman's gestational weeks and BMI on the impact of fetal Y-chromosome concentration ($p < 0.05$). In addition, even if both main effects are not significant, as long as the interaction effect is significant, we must retain the interaction term in the model and cannot view these two factors in isolation.

2.3.4. Model Evaluation

(1) Conditional R^2

The conditional R^2 of the model is 0.725, indicating that the entire model (fixed effects + random effects) has high explanatory power. Among them, the large explanatory power comes from random effects, meaning that there are huge and stable differences in the baseline Y-chromosome concentration among different pregnant women, and the model well reflects the individual differences of pregnant women.

(2) Marginal R^2

The marginal R^2 of the model is 0.117, meaning that the proportion of Y-chromosome concentration variation explained only by fixed effects (detection gestational weeks, pregnant woman's BMI, and their interaction term) is 11.7%. This value is relatively low, and there may be other important fixed factors not considered in the model.

2.3.5. Model Diagnosis

(1) Normal Q-Q Plot of Random Effects

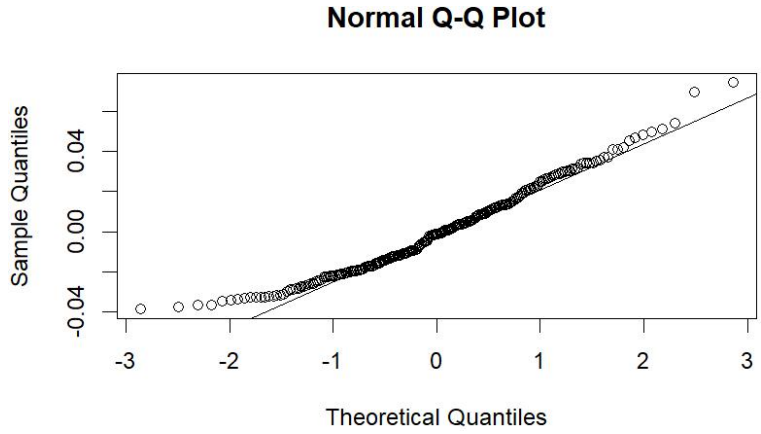

Figure 5. Normal Q-Q plot of random effects

According to the Figure 5, the residuals of the random effects approximately lie on a straight line, indicating that the random effects conform to the normal distribution assumption.

(2) Residual Distribution Plot

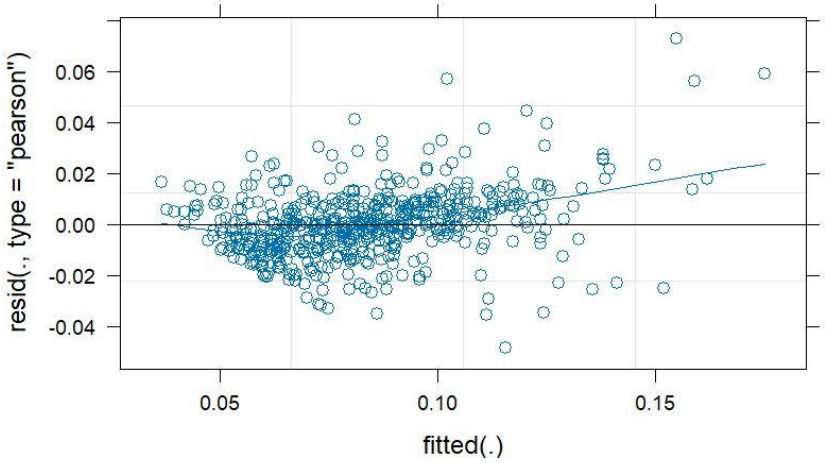

Figure 6. Residual distribution plot

According to the Figure 6, the residual distribution is roughly symmetric and randomly distributed around the zero line, but there are some outliers in certain areas.

3. CONCLUSION

This study successfully established a linear mixed-effects model (LMM) to quantify the statistical association between fetal Y-chromosome concentration and maternal gestational age and BMI.

Prior to model construction, meticulous data preprocessing was performed, including retaining data with Y chromosome concentration ≥4% and GC content between 0.4 and 0.6, and converting gestational age to a continuous variable using the model "continuous gestational age = weeks + days/7". Exploratory data analysis revealed a wide range of Y chromosome concentrations at the same or similar gestational ages, confirming the presence of hierarchical clustering and nested structures in the data—indicating significant inter-individual variation among pregnant women. Model results showed significant random intercept variance (0.0006174), indicating the model adequately captured individual differences among pregnant women. Fixed-effect analysis revealed a significant main effect of maternal BMI on Y chromosome concentration at the 0.05 level ($p=0.01796$), exhibiting a negative correlation: higher BMI correlated with lower Y chromosome concentration. Interaction analysis revealed a significant interaction term between gestational age and BMI ($p=0.03067$), indicating that BMI amplifies the effect of gestational age on Y chromosome

concentration. The model's conditional R^2 was 0.725, demonstrating high overall explanatory power for the fixed-plus-random effects model.

Limitations of the Model and Future Research Directions

The LMM established in this study adequately explains individual variation but has limitations. The marginal R^2 is only 0.117, indicating that the proportion of Y chromosome concentration variation explained solely by fixed effects (gestational age, BMI, and their interaction) is relatively low. The main effect of gestational age failed to pass the significance test at the 0.05 level ($p=0.25241$), indicating its independent linear influence is not significant. This may suggest: First, other important fixed factors such as age, height, or number of blood draws may exist but were not included in the model, yet exert a fixed effect on Y chromosome concentration; Second, the relationship between gestational age and Y chromosome concentration may not be a simple linear one. Future research should focus on identifying and incorporating additional potential key fixed-effect variables to enhance the model's marginal explanatory power. Concurrently, exploring nonlinear mixed-effects models or generalized additive mixed models could better capture the potential nonlinear patterns in the relationship between gestational age and Y chromosome concentration.

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] 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.

[3] Zhang Le, Wei Jie, Zhang Jinhua, et al. Clinical value analysis of expanded non-invasive prenatal testing for screening fetal chromosomal copy number variations [J]. Journal of Practical Obstetrics and Gynecology, 2025, 41(06): 514-519.

[4] Zhang L, Han X, Li N, et al. Detection efficacy of non-invasive prenatal testing for copy number variations in the 17p12 region [J]. Journal of Shanghai Jiao Tong University (Medicine Edition), 2025, 45(03): 310-316.

[5] Qian WJ, Zhang L, Yang T, et al. Application and Efficacy Analysis of Non-Invasive Prenatal Testing in Screening for Fetal Sex Chromosome Aneuploidy [J]. Chinese Journal of Maternal and Child Health, 2025, 40(03):389-393.

[6] Ren Daoju, Li Xiaowei, Li Cuiying. Research Progress of Non-invasive Fetal Cell-free DNA Blood Type Testing in Prenatal Diagnosis [J]. Clinical Transfusion and Laboratory Medicine, 2024, 26(06): 835-842.

[7] Tang Daili, Zheng Huiling, Jin Yaqing. Comparative Efficacy of Non-invasive DNA Testing and Amniocentesis in Prenatal Diagnosis of Fetal Chromosomal Abnormalities [J]. Jilin Medicine, 2024, 45(12): 2900-2903.

[8] Li Na, Li Yipei, Yang Weixian, et al. Analysis of the Application Value of Non-Invasive Prenatal Testing Technology in Prenatal Diagnosis of Fetal Sex Chromosome Aneuploidy [J]. Chinese Journal of Maternal and Child Health, 2024, 39(24): 4963-4966.

[9] Li Shanshan, Zhang Meng, Lü Wei, et al. Prenatal Diagnosis and Genetic Analysis of Chromosome 9 Abnormalities Detected by Non-Invasive Prenatal Testing [J]. Laboratory Medicine and Clinical Diagnosis, 2024, 21(22): 3380-3387.

[10] Dai Peng, Kong Xiangdong. Application of Non-Invasive Prenatal Testing Technology in Screening Chromosomal Abnormalities in Fetal Choroid Plexus Cysts [J]. Chinese Journal of Medical Engineering, 2024, 32(10):1-6.