

Study on the Influencing Factors and Correlation of Fetal Y Chromosome Concentration in NIPT Based on Ordered Clustering and Linear Regression

Weichuan Xue¹, Zixin Gong², Yuebing Tao¹, Dongdong Pan^{1,*}

¹ School of Mathematics and Statistics, Yunnan University, Kunming, China, 650500

² School of Information Science and Engineering, Yunnan University, Kunming, China, 650500

* Corresponding Author Email: ddpan@ynu.edu.cn

Abstract. To address the issue of individual differences affecting the accuracy of non-invasive prenatal testing (NIPT), especially the pain point of low free DNA concentration in the fetus of pregnant women with high BMI, this study analyzed the correlation between the concentration of the Y chromosome in the fetus and the gestational age and BMI of pregnant women. First, samples with abnormal GC content (not 40% to 60%) were excluded through data preprocessing. Then, histograms and scatter plots were drawn to initially observe the distribution and correlation characteristics of the variables. It was found that the scatter plot was difficult to visually determine the correlation. Subsequently, a grouped regression strategy was adopted: BMI was grouped by equal frequency, gestational weeks were determined based on the Jenks natural breakpoint method, combined with variance goodness-of-fit (GVF), pseudo-F statistics, Bayesian information criterion (BIC), and entropy weight method - TOPSIS method to determine the optimal 15 groups. After taking the mean of each group of variables, a univariate linear regression model was constructed, and the significance was verified through the F-test. The results showed that the concentration of the Y chromosome was significantly positively correlated with the gestational weeks ($r=0.6338$, $P=0.0112$), and the degree of correlation was moderate. It was extremely significantly negatively correlated with BMI ($r=-0.8658$, $P=0.0054$), and the degree of correlation was relatively strong. This study provides data support for optimizing the timing of NIPT detection and improving the accuracy of detection in pregnant women with high BMI.

Keywords: Non-invasive prenatal testing (NIPT), Y chromosome concentration, Jenks natural breakpoint method, linear regression, significance test.

1. Introduction

Non-invasive prenatal testing (NIPT), as a core prenatal screening technology for ensuring the health of both mothers and infants and reducing the birth rate of malformed fetuses, its core principle is to collect the mother's blood and detect the free DNA fragments of the fetus [1-3]. Determine the abnormal conditions of chromosome 21 (associated with Down syndrome), chromosome 18 (associated with Edwards syndrome), and chromosome 13 (associated with Patau syndrome). In current clinical practice, NIPT is confronted with two core pain points [4-7]: the first is the contradiction between individual differences and the accuracy of detection. The accuracy of NIPT depends on the concentration of sex chromosomes in the fetus (Y chromosome for male fetuses and X chromosome for female fetuses), among which the concentration of Y chromosome in male fetuses must be $\geq 4\%$ to ensure the reliability of the results. However, individual differences such as BMI, age, and gestational weeks of pregnant women can significantly affect the concentration of the Y chromosome. Especially for the "high BMI pregnant women group" that the attached data focuses on, due to the generally low concentration of free DNA, the traditional simple grouping of BMI based on experience and the unified detection time point can easily lead to insufficient detection accuracy for some pregnant women. The second is the balance between the timing of detection and risk control. The time of fetal abnormality detection is directly related to the treatment window period - detection within 12 weeks is of low risk, 13-27 weeks is of high risk, and after 28 weeks is of extremely high risk. However, in actual detection, there are data complexities such as sequencing failure (such as too

early timing) and repeated detection (to enhance reliability). How to "detect as early as possible" while ensuring accuracy has become an urgent problem to be solved in clinical practice.

Therefore, this study mainly addresses the following issues: (1) Analyzing the correlation characteristics between the concentration of the Y chromosome in the fetus and the gestational age and BMI of the pregnant woman; (2) Establish a relationship model between Y chromosome concentration and the above indicators; (3) Test the significance of the relational model.

First, draw the histograms of the two covariates (gestational weeks and BMI) respectively, as well as the scatter plot between the dependent variable (Y chromosome concentration) and the covariates, observe the distribution pattern of the covariates, and preliminarily determine the correlation between Y chromosome concentration and gestational weeks and BMI. Based on the results of the initial judgment, the two covariates were respectively subjected to univariate linear regression with the dependent variable, and two regression equations were given. Finally, determine the direction and degree of correlation between the dependent variable and the covariate, and test the significance and fitting effect of the regression model.

2. Data preprocessing

Since the normal range of GC content is 40% to 60%, if the GC content is too high, too low, or abnormally distributed, it may indicate problems with sequencing quality. Therefore, to ensure the reliability of the data input into the model, samples with GC content not ranging from 40% to 60% were excluded in this paper.

3. Model establishment and solution

3.1. Analysis of Correlation

To intuitively understand the distribution of gestational weeks and BMI, distribution histograms of gestational weeks and BMI were respectively plotted. As shown in Figures 1 and 2, the frequency distribution of gestational weeks generally presents a multi-peak pattern. The frequency of gestational weeks detected shows relatively high values in multiple intervals, without a distinct single concentrated trend. The frequency distribution of BMI in pregnant women generally shows a unimodal pattern. As the BMI value increases, the frequency gradually rises first, reaching the highest in a certain BMI range in the middle. The distribution at both ends (especially at the high BMI end) is very sparse.

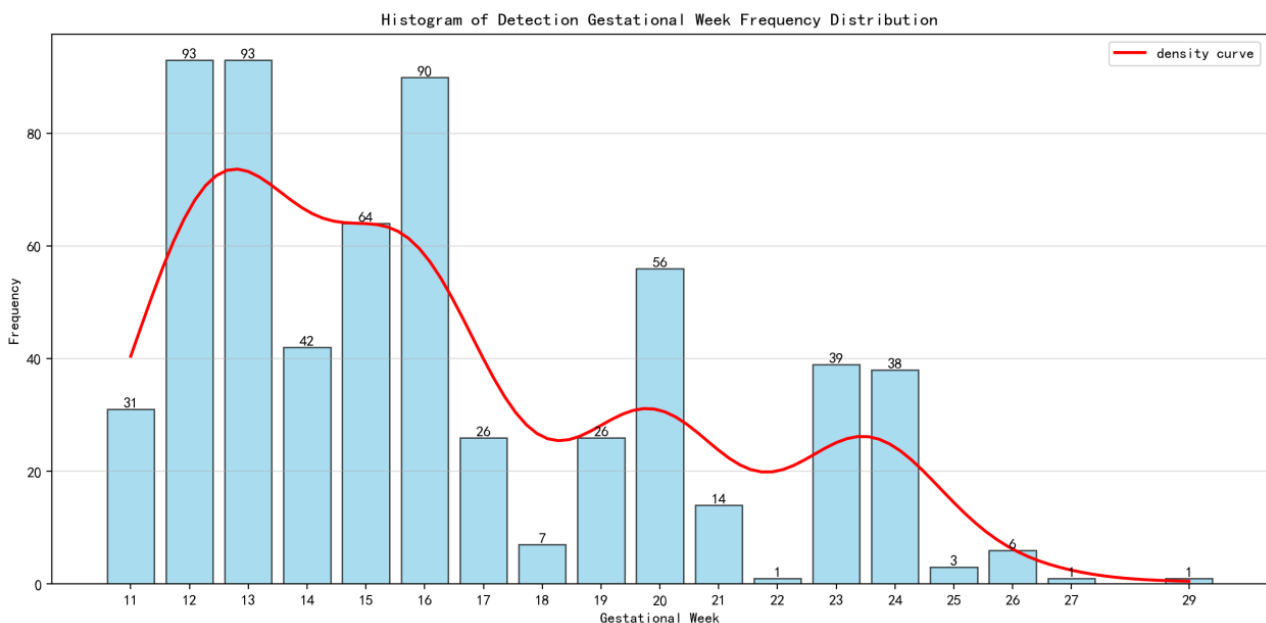


Figure 1. Distribution histogram of gestational weeks

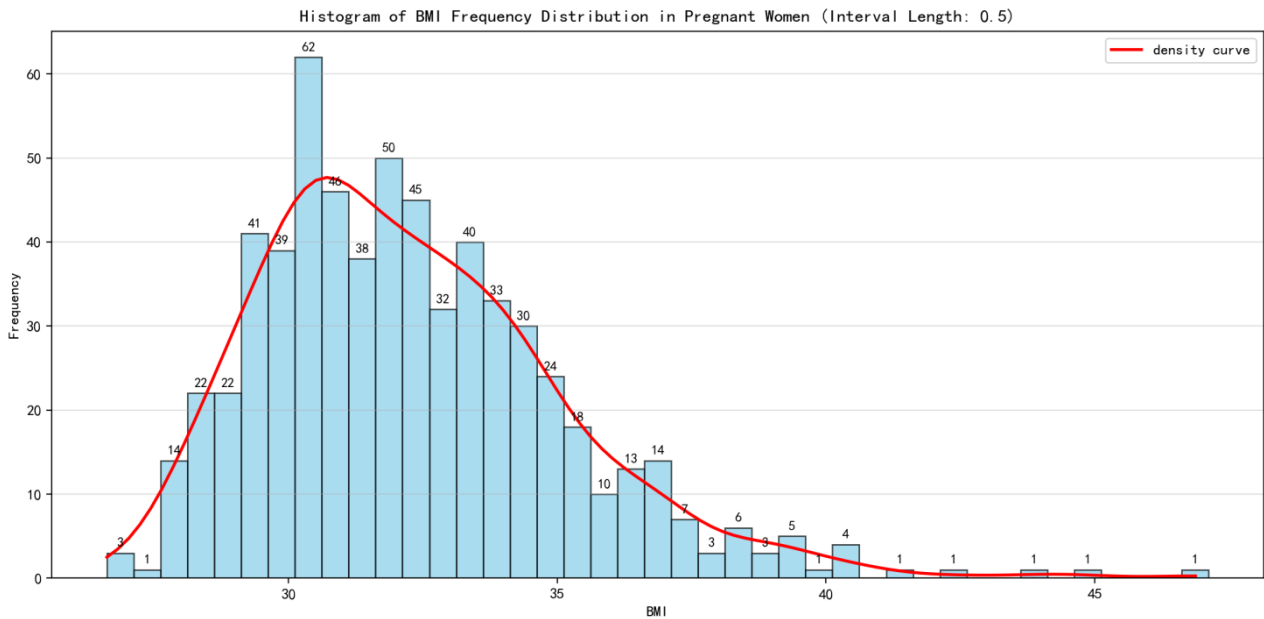


Figure 2. BMI distribution histogram

To study the correlation between Y chromosome concentration and gestational weeks as well as BMI, a scatter plot (Figure 3) was first used to roughly observe the relationship between each covariate and the dependent variable. The distribution of sample points in the scatter plot was scattered, making it impossible to intuitively determine their relationships.

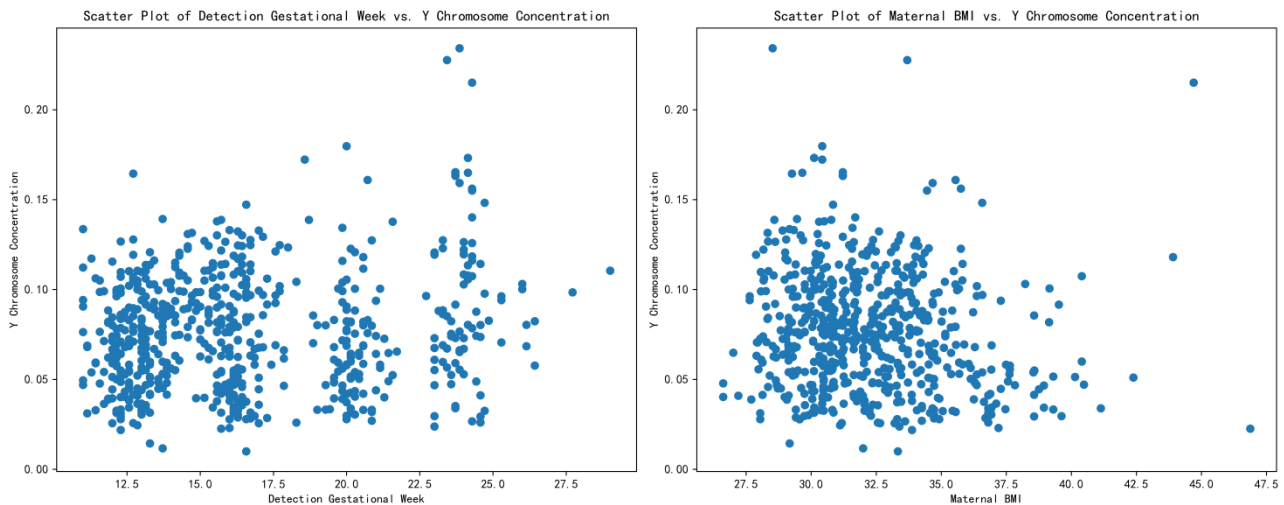


Figure 3. Scatter plot of Y chromosome concentration versus gestational age and BMI

Based on the research in medicine on the relationship between the proportion of fetal free DNA and its influencing factors, this paper separates the two covariates of gestational weeks and BMI, and establishes regression models for them and Y chromosome concentration respectively. Inspired by scholars such as Xue Ying, the covariates (BMI, gestational weeks) were grouped, the mean values were taken, and then regression analysis was conducted.

3.2. Establishment of the regression model

3.2.1. Regression equation of Y chromosome concentration and BMI

Step1: Group BMI according to the principle of equal frequency grouping

The average BMI and the average Y chromosome concentration in each group were calculated, and the results are shown in Table 1.

Table 1. BMI grouping situation

BMI grouping	Frequency	Average BMI	Mean concentration of the Y chromosome
(26.618, 29.297]	83	28.56157	0.080856
(29.297, 30.301]	76	29.84526	0.083167
(30.301, 30.851]	78	30.59245	0.077362
(30.851, 31.793]	79	31.39178	0.081684
(31.793, 32.812]	79	32.329	0.073294
(32.812, 33.887]	78	33.35933	0.077525
(33.887, 35.496]	79	34.54995	0.075346
(35.496, 46.875]	79	37.59595	0.065552

Step2: Establish a univariate linear regression model
 The correlation between the two variables is expressed as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

Among them, β_0, β_1 is the regression coefficient, y_i is the average Y chromosome concentration of group i , x_i is the mean BMI of group i , ε_i is the random error, and $\varepsilon_i \sim N(0, \sigma^2)$.

Step3: Least squares estimation of regression coefficients:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

Among them, n_i represents the sample size of group n_i , and $\hat{\beta}_0, \hat{\beta}_1$ should satisfy

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) \tag{3}$$

Step4: Solve and obtain the regression equation

$$y = 0.1316 - 0.0017x \tag{4}$$

3.2.2. Regression equation of Y chromosome concentration and gestational weeks

Step1: First, determine the optimal number of groups

According to the principle of Jenks' natural breakpoint method [8-10]: minimizing intra-group variance and maximizing inter-group variance to find the best classification points of the data, the optimal number of groups and grouping boundaries are determined based on three indicators: variance goodness of fit (GVF), pseudo-F statistic, and Bayesian information criterion (BIC). First, introduce the total variance, intra-group variance and inter-group variance.

Total variance (SST) :

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu})^2 \tag{5}$$

Sum of within-group variances (SSW) :

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu}_i)^2 \tag{6}$$

Sum of inter-group variances (SSB):

$$SSB = \sum_{i=1}^k n_i (\bar{\mu}_i - \bar{\mu})^2 \tag{7}$$

Where, x_{ij} is the j th specific sample value ($j = 1, 2, \dots, n_i$) of group i ; n is the total sample number; n_i is the sample number of group i ; $\bar{\mu}_i$ is the intra-group mean of group i (satisfying $\bar{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$); $\bar{\mu}$ is the overall mean value (satisfying $\bar{\mu} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$). The following is a formula for the three indicators.

Variance Goodness of Fit:

$$GVF = \frac{SST - SSW}{SST} \tag{8}$$

Pseudo F-statistic:

$$F = \frac{SSB / (k - 1)}{SSW / (n - k)} \tag{9}$$

Bayesian Information Criteria (BIC):

$$BIC = n \ln(SSW / n) + k \ln(n) \tag{10}$$

Use Python to calculate the above three indicators of gestational weeks, and the results are shown in Figure 4.

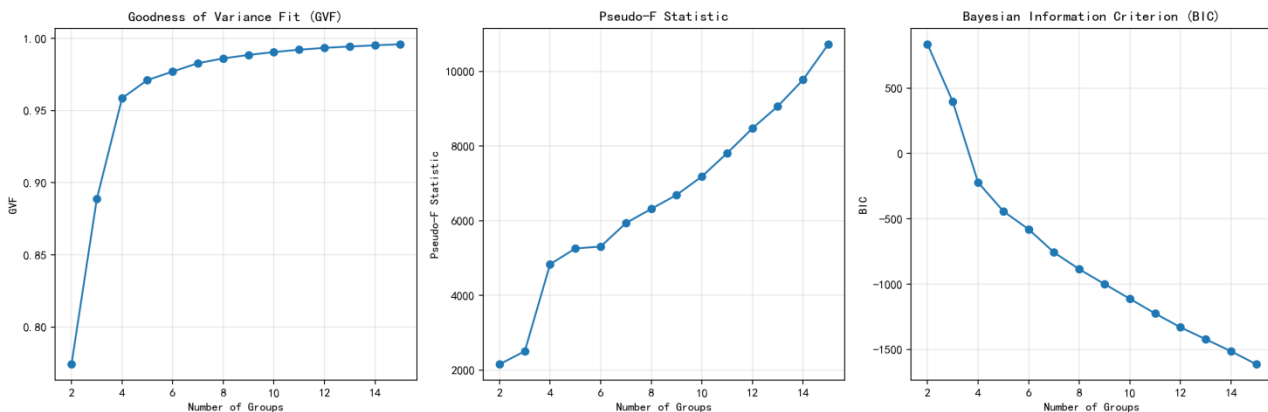


Figure 4. Changes in three indicators with gestational weeks and number of groups

For GVF, as the number of groups increases, GVF increases rapidly from around 0.77. After the number of groups reaches 4, the growth rate slows down and gradually approaches 1. This shows that increasing the number of groups can improve the model's interpretation of the total variance of the data. However, when the number of groups exceeds a certain range, the marginal improvement in explanatory ability will weaken; The pseudo-F statistic continues to increase overall with the increase of the number of groups, especially when the number of groups increases rapidly from 2 to 4, and continues to increase after that, indicating that increasing the number of groups will significantly enhance the difference between groups relative to the difference within groups., that is, the ability of groups to distinguish data structures increases with the increase of the number of groups; The BIC continues to decrease with the increase of the number of groups, rapidly decreasing from a higher value when the number of groups is 2, and then gradually approaching a lower value, indicating that in this analysis scenario, more groups are more favored under the BIC criterion.

The number of recommended groups is calculated based on the optimal values of the three indicators, and the respective weights of the three indicators are calculated using the entropy weight method. Table 2 below presents the results of the number of recommended groups and their weights for the three indicators.

Table 2. Number of recommended groups for three indicators and their weights

Indicators	Number of recommended groups	Weight
GVF	4	0.2048
BIC	15	0.3225
pseudo F-statistic	15	0.4727

Finally, the optimal recommended number of groups based on the entropy weight method-TOPSIS method was 15 groups. The final gestational week grouping results obtained according to Jenks natural breakpoint method are shown in Table 3.

Table 3. Group of gestational weeks

Gestational weeks	Frequency	Mean Y chromosome concentration	Mean number of gestational weeks tested
(11.0,11.7]	24	0.074531	11.327381
(11.7,12.6]	77	0.064929	12.254174
(12.6,13.4]	82	0.073746	13.054007
(13.4,14.3]	56	0.078941	13.887755
(14.3,15.1]	31	0.091971	14.797235
(15.1,16.0]	68	0.076684	15.691176
(16.0,16.9]	75	0.072711	16.409524
(16.9,18.0]	27	0.081462	17.365079
(18.0,19.4]	12	0.075330	18.904762
(19.4,20.4]	59	0.068343	20.009685
(20.4,21.7]	31	0.072746	20.944700
(21.7,23.7]	34	0.084954	23.348739
(23.7,24.9]	44	0.101605	24.240260
(24.9,26.4]	9	0.083686	25.888889
(26.4,29.0]	2	0.104440	28.357143

Step 2: Repeat the steps of the first regression model

A model of the correlation between Y chromosome concentration and gestational age was established, and recorded as the mean value of gestational age in the group, and a regression equation was given.

$$y = 0.055034 + 0.001376t \tag{11}$$

3.3. Significance test

F-test was performed on two one-variable linear regression equations respectively, and their correlation coefficient r and determination coefficient R^2 were calculated.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{12}$$

Python is used to calculate and solve the problem, and the significance test results are shown in Table 4. Y chromosome concentration was negatively correlated with BMI, with a strong and

statistically significant correlation ($p = 0.0054 < 0.05$); Y chromosome concentration was positively correlated with gestational age, with a moderate and statistically significant correlation ($p = 0.0112 < 0.05$).

Table 4. Significance test

Correlation	correlation coefficient r	coefficient of determination R^2	P value
Y chromosome concentration and BMI	-0.8658	0.7496	0.0054
Y chromosome concentration and gestational age	0.6338	0.4017	0.0112

In order to intuitively reflect the change trend of Y chromosome concentration with BMI and gestational age, we drew fitting regression plots based on the distribution and mean values of each variable in each group.

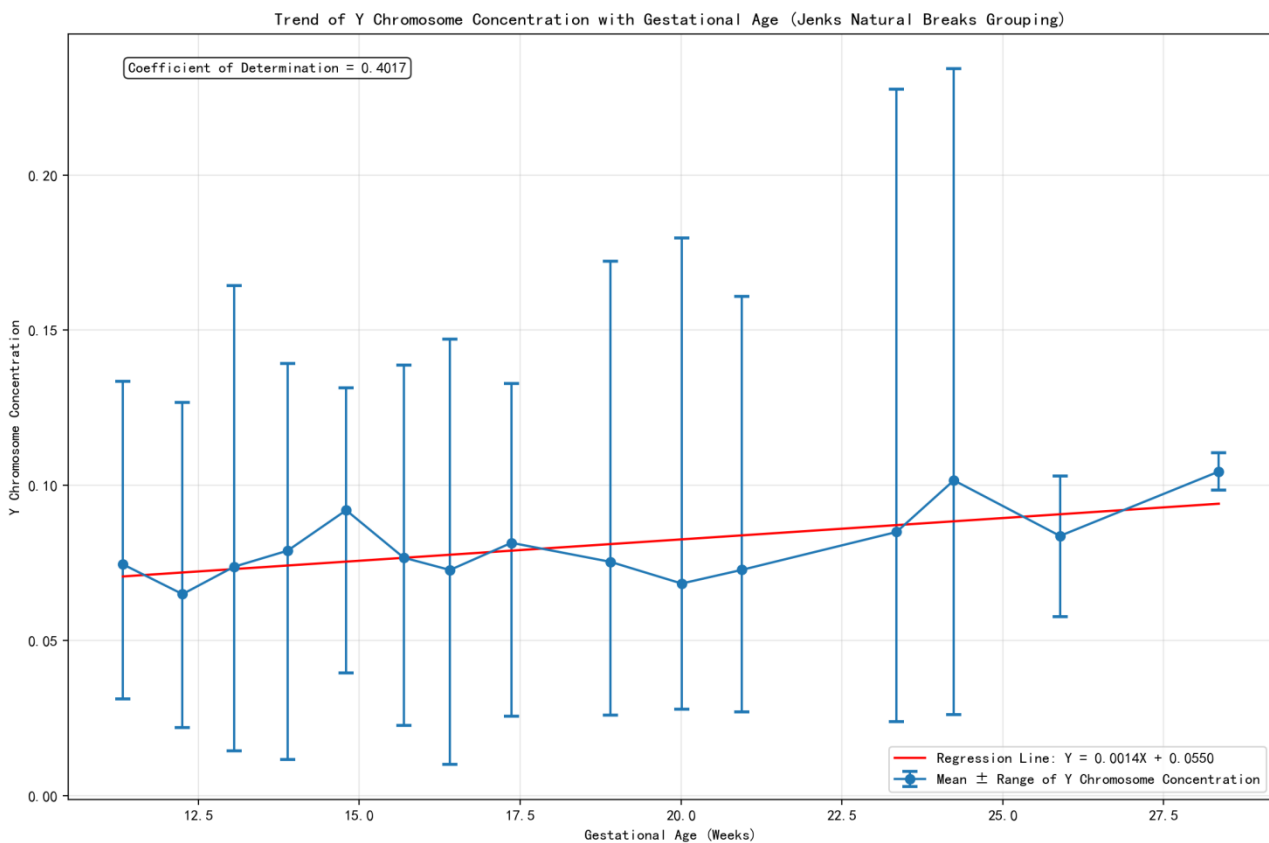


Figure 5. Relationship between Y chromosome concentration and BMI

It can be intuitively observed from Figure 5 that the Y chromosome concentration decreases linearly with the increase of the mean BMI, and there is a negative linear correlation between the two; although the error bars of the Y chromosome concentration under different BMI groups are long, that is, the variation range is large, the actual observed mean points are generally relatively close to the red regression line, and the change trend of the mean points is highly consistent with the trend of the regression line, and there is no significant deviation from the regression line for the mean points. This intuitively shows that the linear regression line can better capture the overall trend of Y chromosome concentration changing with the mean BMI, and the fitting effect is ideal.

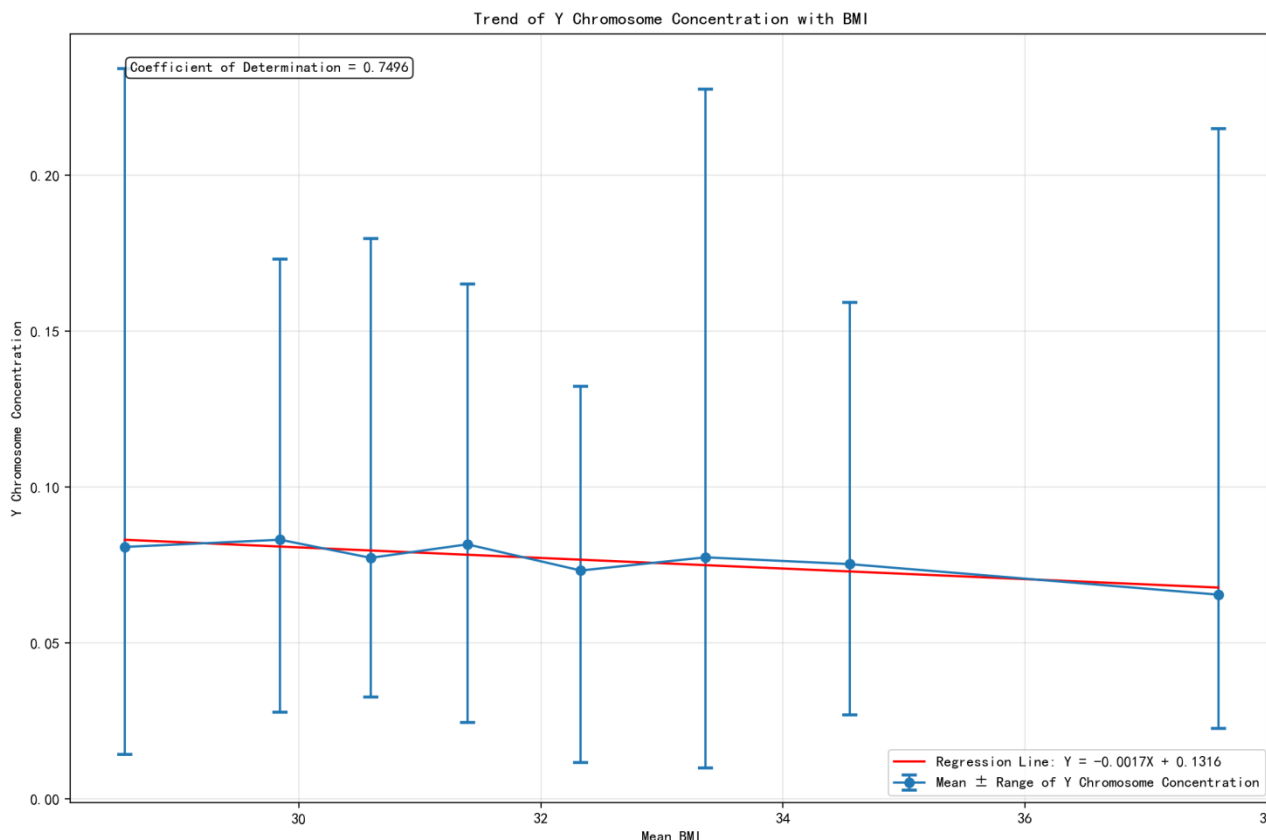


Figure 6. Relationship between Y chromosome and gestational age

It can be intuitively seen from Figure 6 that the actual observed mean points generally follow the upward trend of the red regression line, and there is a positive linear correlation between the Y chromosome concentration and the gestational age; although the variation range of the Y chromosome concentration under different gestational age groups is large, and there are small fluctuations in individual mean points, the overall change direction of the mean point is highly consistent with the trend of the regression line, and there is no significant deviation from the regression line of the mean point. This shows that the linear regression line can capture the overall increasing trend of Y chromosome concentration with gestational age and has a certain fitting effect.

4. Conclusions

Through systematic data processing and modeling analysis, this study clarified the quantitative correlation between pregnant women's gestational weeks, BMI and fetal Y chromosome concentration: after data pretreatment to ensure sample quality, a group regression strategy was used to avoid the limitations of direct regression, among which the BMI is divided into equal frequency groups and the optimal 15 groups of gestational weeks effectively captured the linear relationship between variables; The final verification showed that the Y chromosome concentration showed a moderately strong positive correlation with the increase of gestational weeks and a strong negative correlation with the increase of BMI. Both types of correlation relationships were statistically significant, providing a key influencing factor quantification basis for clinical NIPT testing. Especially for pregnant women with high BMI, detection strategies can be adjusted based on the negative correlation characteristics of BMI to improve accuracy. In the future, the research dimension can be further expanded to include the age of the pregnant woman and the gender of the fetus.(Female fetal X chromosome concentration) and other variables, a multi-factor regression model is constructed to more comprehensively explain Y chromosome concentration variations; at the same time, quantile regression and machine learning algorithms can be combined to optimize grouping methods and model fitting effects to explore different risk levels (such as critical scenarios with Y chromosome concentration < 4%) The detection

time threshold under the critical scenario provides more refined technical support for the formulation of NIPT personalized testing plans.

References

- [1] Linthorst J ,Sisternans A E . Noninvasive Prenatal Testing: Mosaic Ratio Score as a Predictor for Confined Placental Mosaicism. [J]. Clinical chemistry,2025.
- [2] Ren Y ,Hao N ,Chang J , et al. Clinical Implications of Noninvasive Prenatal Testing Failures Due to Low Fetal Fraction: Associations With Adverse Maternal and Fetal Outcomes. [J]. Prenatal diagnosis,2025.
- [3] Masouleh M A A ,Yazdi E P ,Sadrabadi E A , et al. Embryo metabolism as a novel non-invasive preimplantation test: nutrients turn over and metabolomic analysis of human spent embryo culture media (SECM). [J].Human reproduction update,2025.
- [4] Zhong G ,Wu J ,Zhong Z , et al. Case Report: A prenatal case with sex discordance between non-invasive prenatal testing and fetal genetic testings due to maternal rare chromosome karyotype [J].Frontiers in Genetics,2025,161546579-1546579.
- [5] Peng H ,Wang D ,Guo F , et al. Prenatal diagnosis of imprinted associated chromosome abnormalities identified by noninvasive prenatal testing (NIPT) [J].Scientific Reports,2025,15 (1):12830-12830.
- [6] Warton C ,Vears F D . Healthcare professionals' perspectives on and experiences with non-invasive prenatal testing: a systematic review [J]. Human Genetics,2025,144 (4):1-32.
- [7] Song J ,Zheng Y ,Huang X , et al. Enhancing Thermodynamic and Kinetic Performance of Microfluidic Interface-Based Circulating Fetal Cell Isolation for Noninvasive Prenatal Testing. [J]. Analytical chemistry,2025,97 (13).
- [8] Saputri W A ,Ari H P ,Kaila G K , et al. Clustering the Depression Prevalence in Indonesia Provinces through Natural Breaks Jenks Method [J].Clinical Practice & Epidemiology in Mental Health,2025,21e17450179375982.
- [9] Gui R ,Song W ,Lv J , et al. Digital Elevation Model-Driven River Channel Boundary Monitoring Using the Natural Breaks (Jenks) Method [J].Remote Sensing,2025,17 (6):1092-1092.
- [10] Chaoying K ,Shu H ,Yigen Q . Comparison of natural breaks method and frequency ratio dividing attribute intervals for landslide susceptibility mapping [J]. Bulletin of Engineering Geology and the Environment,2023,82 (10).