Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 7, No. 3, 2025

Research on Personalized Recommendations for Optimal Detection Timing Based on K-Means Clustering and Logistic Regression

Jing Ding*, Huilin Sun, Ziquan Feng

Lanzhou University, Lanzhou, China

* **Corresponding author**: Jing Ding (Email: 3042710682@qq.com)

Abstract: This paper focuses on selecting the optimal timing for specific detection scenarios. It integrates multiple modeling approaches—including multiple linear regression, K-Means clustering, and logistic regression—to construct a quantitative decision-making system that optimizes detection efficiency and reliability. The study first analyzes the linear correlation between target indicator concentrations and multidimensional features using a Pearson correlation coefficient matrix to identify core influencing variables. Subsequently, a multiple linear regression model is constructed using the statsmodels toolkit for solution and significance verification, precisely quantifying the influence strength of each feature on the target indicator. During data preprocessing, IQR and Z-score methods identify outliers, with effective information retained through trimming. K-Means clustering combined with the elbow rule then determines the optimal number of clusters, dividing key features into four significantly distinct groups. Separate logistic regression models were established for each group to fit the temporal variation patterns of detection success rates. Optimal detection timepoints were determined for each group using a 90% success rate threshold, alongside risk level assessments. The resulting technical framework provides data-driven support for personalized detection timing selection and establishes a quantitative analytical paradigm for optimizing similar detection scenarios.

Keywords: Multiple Linear Regression Model, K-Means Clustering, Logistic Regression.

1. Introduction

In specific detection scenarios, the precise selection of detection timing directly impacts detection success rates and result reliability. Since target indicator concentrations are influenced by the cross-interaction of multidimensional features, traditional empirical timing selection methods struggle to adapt to detection demands under varying feature combinations. There is an urgent need to establish a data-driven quantitative analysis framework. Current research predominantly focuses on the influence of single variables on detection outcomes, lacking systematic decomposition of multi-feature synergistic effects. Moreover, it fails to provide personalized timing solutions for groups with differing characteristics, resulting in compromised detection efficiency and risk control [1].

To address this, this study integrates target indicator concentration and associated feature data, combining Pearson correlation analysis [2], multiple linear regression [3], K-Means clustering [4], and logistic regression [5]. First, correlation analysis and regression modeling identify core influencing factors and quantify their impact strength. Then, clustering algorithms enable precise segmentation of feature groups. Finally, success rate prediction models are constructed for different groups to determine optimal detection timepoints. This research approach not only fills the gap in applying multi-feature collaborative analysis to detection timing optimization but also provides a reusable technical paradigm for process optimization in similar detection scenarios, thereby enhancing the scientific rigor and precision of detection decisions [6].

2. Correlation analysis of indicators

To investigate the various factors influencing Y chromosome concentration, this study analyzed the linear relationships among variables using a Pearson correlation coefficient matrix on the processed dataset. Subsequently, a multiple linear regression model was established and solved using Python's statsmodels library, thereby quantifying the impact of each feature on Y chromosome concentration.

2.1. Pearson correlation analysis

Correlation analysis aims to measure the linear relationship among multiple continuous variables. This study employs Pearson's correlation coefficient to quantitatively assess the strength and direction of linear relationships between variables. Its formula is:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

Where n represents the sample size, and (X, Y) denotes any pair of variables requiring correlation analysis. These can be any two of the following: Y chromosome concentration (Y_concentration), maternal BMI (BMI), gestational week (Gestational_Week), maternal age (Age), number of pregnancies (Gravidity), or number of births (Parity). The resulting correlation coefficient matrix is visualized as a heatmap in Figure 1 below.

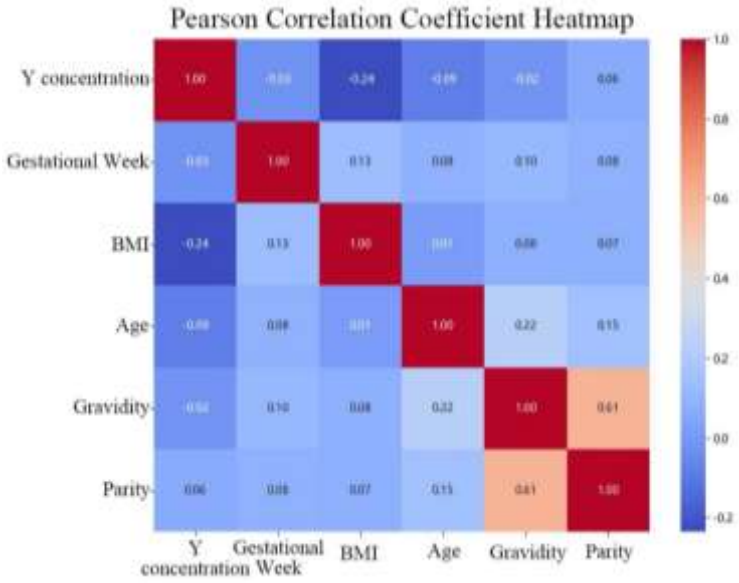

Figure 1. Pearson Correlation Coefficient Heatmap

This heatmap illustrates the Pearson correlation coefficients between all variables. Red indicates positive correlation, blue denotes negative correlation, with darker shades signifying stronger associations. The numerical values within each cell represent the magnitude of the correlation coefficient.

From the data in the first row or first column, this study yields the following conclusions: The correlation coefficient between Y chromosome concentration and maternal BMI is -0.24, the highest absolute value among all independent variables. This clearly indicates a moderately strong negative correlation. The correlation coefficient between Y chromosome concentration and gestational week is -0.03, suggesting virtually no linear relationship between the two in the current data. The absolute values of the correlation coefficients between Y chromosome concentration and age, number of pregnancies, and number of births are all less than 0.1, indicating that there is almost no linear relationship between them and Y chromosome concentration.

2.2. Establishing the multiple linear regression model

This study selected a multiple linear regression model to quantify the influence of each characteristic on Y chromosome concentration. The model expression is as follows:

$$Y_{\text{concentration}} = \beta_0 + \beta_1 X_{\text{Gestational\_Week}} + \beta_2 X_{\text{BMI}} + \beta_3 X_{\text{Age}} + \beta_4 X_{\text{Gravidity}} + \beta_5 X_{\text{Parity}} + \epsilon \quad (2)$$

Where β_0 is the model's intercept term, β_1 to β_5 are the regression coefficients for each feature, reflecting the influence of the corresponding independent variable on Y chromosome concentration under controlled conditions, and ϵ is the random error term.

2.3. Model solution and analysis

Statsmodels is a Python library for statistical modeling and econometrics, emphasizing the analytical interpretation of relationships between variables. Here, this study employs the statsmodels library to fit the model, yielding the regression results presented in Tables 1 and Table 2 below.

Table 1. Multivariate Linear Regression Model Coefficient Details

Variable	Coefficient	SE	t-value	P-value	[0．025	0．975]
Intercept	0.1622	0.028	5.835	0.000	0.107	0.217
Gestational Week	0.0002	0.002	0.127	0.899	−0．003	0.003
BMI	−0．0023	0.001	−4．004	0.000	−0．003	−0．001
Age	−0．0008	0.000	−1．546	0.123	−0．002	0.000
Gravidity	−0．0022	0.003	−0．854	0.394	−0．007	0.003
Parity	0.0060	0.003	1.794	0.074	−0．001	0.013

Table 2. Overall Results of the Multiple Linear Regression Model

Indicator Name	R-squared	Adj. R-squared	F−statistic	Prob（F−statistic）
Value	0.075	0.057	4.231	0.00103

This section analyzes the multiple linear regression model presented in the table, focusing on the independent effects of each explanatory variable within the model. The aim is to rigorously evaluate the model and the relationships between variables it reveals.

2.4. Significance analysis of regression coefficients

The regression coefficient for maternal BMI is -0.0023, with a p-value of 0.000. The p-value is far below the significance level of 0.05, indicating that the effect of maternal BMI on Y chromosome concentration is highly significant. The negative coefficient signifies a significant negative linear relationship between the two variables. This implies that, controlling for other variables, a one-unit increase in maternal BMI is associated with a 0.0023-unit decrease in Y chromosome concentration.

Regarding other regression coefficients, this study's analysis is as follows:

1. Non-significant variables: For parity, gestational week at testing, age, and number of pregnancies, their p-values were 0.074, 0.899, 0.123, and 0.394 respectively, all exceeding the 0.05 significance level. Therefore, this study cannot conclude that they significantly affect Y chromosome concentration. Among these, the p-value of 0.074 for number of births is borderline significant, with a regression coefficient of 0.0060. This may indicate a weak positive relationship, but the statistical evidence is insufficient given the current sample size. The p-values for the other three variables are large, indicating their coefficients are not significantly different from zero.

2. The regression coefficient for the intercept term is 0.1622 with a p-value of 0.000, indicating a significant effect of the intercept on Y chromosome concentration. However, the theoretical significance of the intercept lies in representing the value of the dependent variable when all independent variables are zero. In practice, it is impossible for all independent variables to be zero simultaneously, thus limiting the practical significance of this intercept term.

In summary, this multiple linear regression model holds. Among the five characteristics, the mother's pre-pregnancy BMI is the only factor exhibiting highly statistically significant influence, showing a significant negative correlation with Y chromosome concentration. Number of previous births, gestational week at testing, age, and number of pregnancies did not demonstrate independent significant effects.

3. Exploring optimal grouping and testing timing

This study aims to establish a quantitative model for non-invasive prenatal genetic testing (NIPT) that selects the optimal testing timing while minimizing risks.

This section addresses a binary classification scenario: NIPT testing outcomes are either "successful" or "unsuccessful." Therefore, by grouping pregnant women based on BMI and establishing logistic regression models for each group, this section aims to recommend optimal testing time points that minimize potential risks for different BMI cohorts.

3.1. Data preprocessing and outlier analysis

To enhance model accuracy, the raw data undergoes preprocessing. The dataset contains 1081 records with one missing value in the "Test Gestational Age" field, which is imputed using the median value of 16.00. Next, outlier detection was performed on three key variables: gestational week at detection, maternal BMI, and Y chromosome concentration. The methods employed were the IQR method and the Z-score method.

The IQR method is a distribution-based outlier detection technique. It first calculates the first quartile (Q_1) and third quartile (Q_3) of the dataset, with their difference forming the interquartile range:

$$\mathrm{IQR} = Q_3 - Q_1 \tag{3}$$

Defining the normal range:

$$[Q_1 - 1.5 \times \mathrm{IQR}, Q_3 + 1.5 \times \mathrm{IQR}] \tag{4}$$

Any data point falling outside this range is considered an outlier.

In contrast, the Z-score method assumes of a normal distribution. It measures deviation by calculating the distance between each data point X and its mean μ, where σ is the standard deviation. The formula is:

$$Z = \frac{X-\mu}{\sigma} \tag{5}$$

If the absolute value of a data point's Z-score exceeds a preset threshold, it is flagged as an outlier.

After processing the raw data using both methods, no outliers were found in the gestational age data, while a small number of outliers existed in maternal BMI and Y chromosome concentration.

Using the IQR method, 26 outliers were detected in maternal BMI, accounting for 2.41% of the data. The Z-score method identified 14 outliers in maternal BMI, representing 1.30% of the data.

Using the IQR method, 10 outliers were detected in Y chromosome concentration, accounting for 0.93% of the data. The Z-score method identified 8 outliers in Y chromosome concentration, representing 0.74% of the data.

Considering that direct deletion of outliers may result in loss of clinical information, this study employed trimmed outlier processing. Values exceeding the normal range were replaced with values corresponding to specified percentiles, thereby preserving data records while reducing the impact of extreme values on model stability.

3.2. BMI Clustering Based on the K-Means Algorithm

The K-Means algorithm is a classic unsupervised clustering method whose objective is to partition a dataset into K distinct clusters such that each data point belongs to the nearest cluster center.

Since the number of clusters was not explicitly specified, this study employed the "elbow rule" to determine the optimal number of clusters—a core heuristic in cluster analysis for identifying the optimal K value. The core principle involves observing the trend of the within-cluster sum of squares (WCSS) as the number of clusters K increases. The inflection points where the curve transitions from a steep decline to a plateau corresponds to the optimal K value. WCSS, also known as the within-cluster sum of squares, calculates the sum of squares of the Euclidean distances between each sample and its cluster center. A smaller WCSS

indicates greater concentration of samples within the same cluster. Its formula is:

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (6)$$

Where K denotes the number of clusters; C_k represents the set of all samples in the kth cluster; x_i is the ith sample in the kth cluster; and μ_k is the center of the kth cluster.

After thorough analysis and comprehensive consideration of medical practice and application requirements, k=4 was determined to be the optimal choice. Subsequently, the process begins with randomly selecting 4 BMI values as initial cluster centers, followed by an iterative procedure:

1. Assignment step: Calculate the distance from each pregnant woman's BMI value to the 4 cluster centers and assign her to the nearest cluster.

2. Update step: Recalculate the average BMI of all pregnant women within each cluster and use this as the new cluster center.

This assignment and update process repeats until cluster assignments no longer change or cluster centers stabilize.

Based on the algorithm's objective, write the objective function expression J:

$$J = \sum_{k=1}^{K=4} \sum_{x_i \in \text{ Cluster } k} \|x_i - \mu_k\|^2 \quad (7)$$

Where x_i is the BMI value of the i-th pregnant woman, and μ_k is the centroid of the k-th cluster.

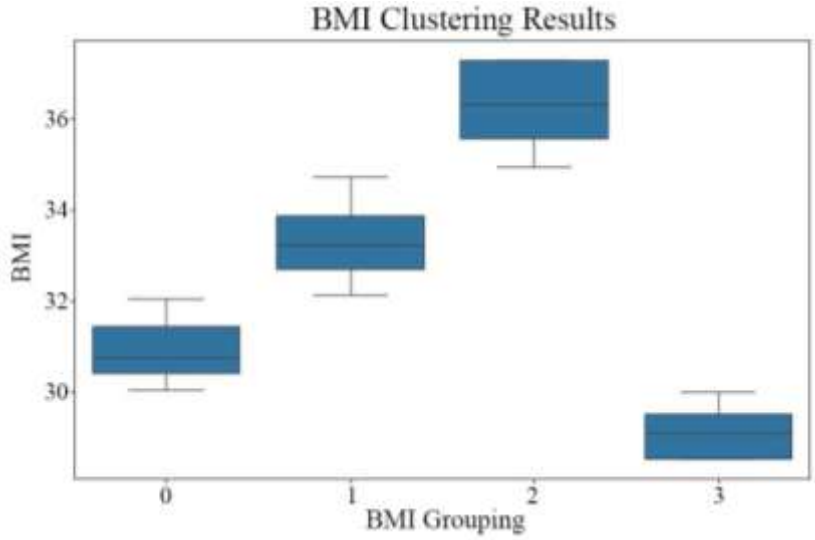

Figure 2. BMI Clustering Results

Figure 2 is a box plot illustrating the distribution of BMI values within each group after pregnant women were divided into 4 clusters using the K-Means clustering algorithm. The box represents the middle 50% of the data, and the horizontal line inside the box indicates the median. Thus, the K-Means algorithm successfully divided the BMI data into four groups with significant differences. The specific BMI ranges for each group are as follows:

1. BMI Group 0: 30.0–32.1
2. BMI Group 1: 32.1–34.7
3. BMI Group 2: 34.9–37.3
4. BMI Group 3: 28.5–30.0

3.3. Determining the optimal detection timepoint using logistic regression

For each BMI group, this study established a model to predict the variation in NIPT detection success rate across gestational weeks.

The model expression is as follows:

$$P(S \mid t) = \frac{1}{1+e^{-(\beta_0+\beta_1 t)}} \quad (8)$$

Where $P(S \mid t)$ is the probability of successful detection at gestational week t, and β_0 and β_1 are model parameters estimated separately for each BMI group. This study defines the "optimal timing" t^* as the gestational week where the success rate first reaches θ=90.0%, i.e.:

$$t^* = \min\{t \mid P(S \mid t) \geq \theta\} \quad (9)$$

By fitting and solving the data for each group, this study obtained the following recommended timing and risk assessments (see Table 3).

Table 3. Recommended Optimal Testing Timing for Each BMI Group (Default 90% Success Rate)

BMI Group	BMI Range	Recommended Optimal Timing (t^*)	Corresponding Risk Level
0	30.0 − 32.1	20.8 weeks	High risk
1	32.1 − 34.7	11.1 weeks	Lower risk
2	34.9 − 37.3	Not reached	Close monitoring recommended
3	28.5 − 30.0	17.4 weeks	High risk

Visualization of NIPT success rate trends by gestational week for each BMI group, with recommended optimal testing time points. Partial results (Group 1) are shown in Figure 3 below. In the figure:

Blue scatter points represent actual test results (vertical axis: 1 = success, 0 = failure); The red curve shows the model-predicted success probability trend over gestational weeks; the green dashed line represents the 90% target success rate

threshold; the purple dashed line marks the intersection point between the success rate curve and the 90% threshold line, indicating the "optimal timing."

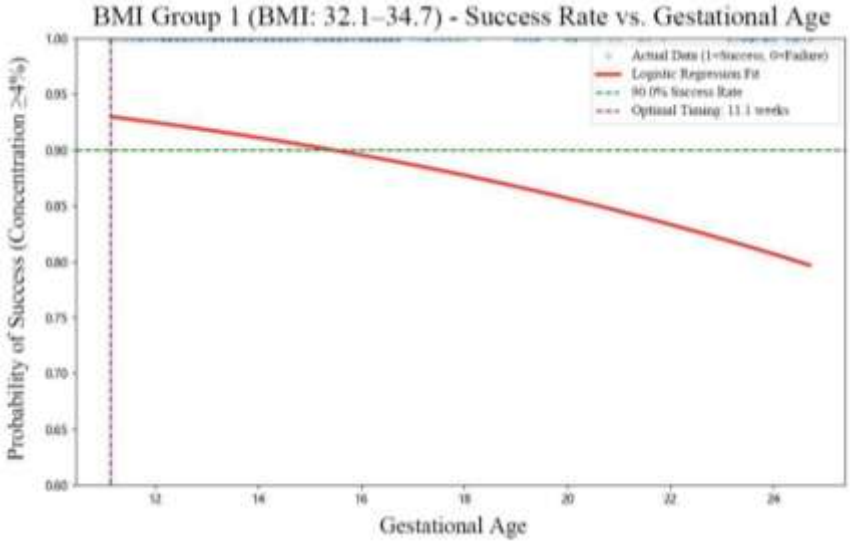

Figure 3. 32.1–34.7 Group Achievement Rate vs. Gestational Week

Final analysis results: For Group 3, the recommended optimal timing is 17.4 weeks; for Group 0, it is 20.8 weeks. Comparing these two "normally functioning" groups reveal that as BMI increases, the gestational week required to achieve the same success rate (90%) significantly delays. This aligns with the theory that Y chromosome concentration becomes "diluted" in high-BMI pregnancies. However, an anomaly emerged in Group 1, where the logistic regression curve exhibited a downward trend. The recommended optimal timing was 11.1 weeks, as success rates were already high early in observation (around 11 weeks) and subsequently declined slightly. This may stem from data bias within this group, or other confounding factors not captured by the model.

For Group 2, the 90% success rate threshold was consistently unattainable. This indicates that for women with extremely high BMI, achieving a 90% detection success rate within the standard testing window may remain unfeasible, classifying them as high-risk individuals for NIPT failure. Based on these findings, the study assesses potential risk levels for each group: Group 1 (11.1 weeks) is classified as "low risk." Groups 3 (17.4 weeks) and 0 (20.8 weeks) are categorized as "high risk." Group 2, unable to consistently achieve the target success rate, carries the highest risk of test failure and delayed diagnosis.

4. Conclusions

This study addresses the optimization of optimal testing time points for specific detection scenarios. Through multi-model collaborative analysis and validation, it establishes a comprehensive technical solution spanning from identifying influencing factors to personalized time point recommendations. Its core conclusions and value are reflected in three aspects. First, at the factor level, combined validation through Pearson correlation analysis and multiple linear regression models confirmed that the key feature is the sole variable exerting a highly significant negative impact on target indicator concentration. Other features showed no independent significant effect, anchoring the core research dimension for subsequent analysis. Second, regarding group segmentation and timing adaptation, after dividing key features into four distinct clusters using K-Means clustering, the success rate prediction system built with logistic regression revealed significant differences in optimal detection timing across groups: some groups could achieve 90% success rates as early as 11.1 weeks, indicating lower risk; while others required delaying until 17.4–20.8 weeks to meet success rate requirements, indicating higher risk. The group with extremely high feature values consistently failed to achieve the target success rate within the conventional detection window, classifying it as a high-risk cohort. Third, regarding technological application value, The analytical framework established in this study—"quantification of influencing factors, precise group segmentation, and dual assessment of timing and risk"—not only overcomes the limitations of traditional empirical timing selection but also ensures the reliability and stability of results through data processing techniques like truncation and model optimization methods such as the elbow rule. This provides actionable quantitative decision-making support for process optimization in similar detection scenarios and offers practical reference for the collaborative application of multiple models in detection optimization.

References

[1] Li Jia'an. Research on Collaborative Saliency Detection Algorithms Based on Multi-Scale Feature Fusion [D]. Nanjing University of Information Science and Technology, 2024. DOI: 10. 27248/ d. cnki. gnjqc.2024.001709.

[2] Zhang Yuhang. Research on Time-Delay Pearson Correlation Analysis and Key Variable Prediction Methods for Air Separation Equipment [D]. Hangzhou Dianzi University, 2023. DOI: 10.27075/d.cnki.ghzdc.2023.000752.

[3] Yan Jueyuan, Shen Xiangyu, Wu Yingkang, et al. From Linear Correlation to Univariate Linear Regression Models: Exploring Relationships Between Two Continuous Variables [J]. Mathematics Teaching, 2025, (10): 27-35+50.

[4] Wang Bingcan., Wang Guochang, & Wei Yanhua. (2025). An improved method for determining the number of clusters based

on the k-means algorithm [J]. Statistics and Decision Making, 41(07), 59-64. DOI: 10.13546/j.cnki.tjyjc.2025.07.010.

[5] Kong Liru, Chen Yuming, Fu Xingyu, et al. A Logistic Regression Algorithm Based on Rotational Granularity [J]. Research on Computer Applications, 2024, 41(08): 2398-2403. DOI: 10.19734/j. issn. 1001-3695.2023.11.0578.

[6] Ge Yan. Research on Time Series Anomaly Detection Method Based on Dual Feature Collaborative Analysis [D]. Hangzhou Dianzi University, 2025. DOI: 10.27075/ d. cnki. ghzdc. 2025. 001938.