Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 23, No. 2, 2026

Multivariate Linear Regression Modeling for Influencing Factors of Fetal Y-Chromosome Concentration in Pregnant Women

Siqi Guo *

School of Mathematics and Statistics, Henan University of Technology, Zhengzhou 450001, China

* **Corresponding author**: (Email: 15736761053@163.com)

Abstract: Noninvasive prenatal testing (NIPT) screens for chromosomal abnormalities by analyzing cell free fetal DNA in maternal peripheral blood, and the accurate quantification of fetal Y-chromosome concentration in male pregnancies is a decisive indicator of test validity. However, individual variability among pregnant women is substantial, so gestational age, body mass index (BMI), maternal age, and several other factors jointly modulate Y-chromosome concentration, which makes a single threshold rule inadequate across heterogeneous populations. The present study addresses the joint analysis of factors influencing Y-chromosome concentration by constructing a modeling framework centered on multivariate linear regression. The framework unifies a basic linear structure, generalized additive smoothing terms, and cubic spline basis expansions within a single estimation pipeline, and incorporates Pearson correlation and Spearman rank correlation, the F test, and the Akaike information criterion for variable screening and significance testing. Evaluation is carried out on a clinical NIPT dataset of 1082 records covering 12 obstetric indicators. On a held out test partition the proposed method attains a coefficient of determination of 0.687, a mean absolute error of 0.018, and a root mean square error of 0.024, which corresponds to an improvement of approximately 11.4% in the coefficient of determination over a purely linear baseline. The estimated regression coefficients indicate that gestational age exerts a significant positive effect on Y-chromosome concentration, that BMI exerts a significant negative modulating effect, and that maternal age, height, weight, and overall GC content contribute smaller yet statistically significant marginal effects when combined in the joint model. The proposed framework provides a quantitative basis for personalized decisions about the optimal NIPT testing window and is therefore relevant for improving the accuracy and timeliness of prenatal screening.

Keywords: Y-chromosome concentration, multivariate linear regression, noninvasive prenatal testing, generalized additive model, cubic spline, gestational age, body mass index.

1. Introduction

Noninvasive prenatal testing has been rapidly adopted in clinical prenatal screening over the past decade. The core idea is to use circulating cell free DNA in maternal peripheral blood for highly sensitive molecular detection of fetal chromosomal aneuploidies. Lo et al. were the first to demonstrate the existence of fetal cell free DNA fragments in maternal plasma [1], laying the biological foundation for noninvasive prenatal testing. Chiu et al. subsequently introduced a method based on massively parallel sequencing that achieves noninvasive detection of trisomy 13, 18, and 21 [2]. In male pregnancies, the proportion of cell free DNA fragments originating from the Y chromosome, namely the Y-chromosome concentration, is the central quantitative indicator for NIPT interpretation, and its stable measurement directly determines the reliability of the test result.

In clinical practice, however, Y-chromosome concentration is not a static quantity but varies markedly with gestational progression and with the physiological characteristics of the pregnant woman. Wang et al. report from a large cohort that Y-chromosome concentration is significantly positively associated with gestational age, that it is inversely related to maternal weight, and that substantial variability exists across individuals [3]. Ashoor et al. further observe in an 11 to 13 week gestation cohort that the fetal cell free DNA fraction is jointly modulated by maternal weight, gestational age, and placental factors [4]. Vora et al. report a clinical study showing that higher BMI tends to coincide with a lower fetal fraction [5]. Bianchi and Chiu provide a systematic review of the recent advances in cell free DNA sequencing during pregnancy [6]. These earlier studies provide direct motivation for the multivariate statistical model of Y-chromosome concentration developed here.

Although a substantial body of work has examined descriptive statistics of cell free DNA fractions, fine grained modeling of Y-chromosome concentration has received comparatively less attention. On the one hand, traditional univariate analyses cannot capture the coupling among gestational age, BMI, and additional factors and can therefore yield biased conclusions. On the other hand, existing nonparametric curve fitting approaches, while flexible, struggle to deliver interpretable regression coefficients and are therefore of limited use in clinical decision making. The Pearson correlation coefficient [7] and the Spearman rank correlation coefficient [8] are valuable for quantifying pairwise associations but cannot substitute for regression analysis in a multivariate setting. Classical references in statistical learning [9] and the theory of generalized additive models [10] provide the methodological grounding for an interpretable yet robust multivariate regression framework, while the Akaike information criterion [11] supplies an objective basis for variable selection and model comparison.

The present paper develops a systematic modeling study of the influencing factors of Y-chromosome concentration. The main contributions are summarized as follows. First, the multifactor modeling problem for Y-chromosome concentration is formalized as a multivariate linear regression

problem with smoothness constraints, and the matrix form together with the closed form least squares estimator is derived. Second, a generalized additive extension based on cubic spline basis functions is designed, so that the model retains the interpretability of linear terms while accommodating nonlinear effects. Third, a complete workflow is proposed that combines correlation based screening, the F test, AIC based comparison, and residual diagnostics, ensuring that the model meets requirements in both statistical significance and predictive accuracy. Fourth, comprehensive experiments on a clinical NIPT dataset show that the proposed framework outperforms several baselines on the coefficient of determination, the mean absolute error, and the root mean square error.

2. Methodology

To systematically characterize the relationships between maternal Y-chromosome concentration and a panel of physiological and biochemical indicators, this section introduces a modeling framework that combines multivariate linear regression as the core component, generalized additive models as an extension, and correlation analysis together with significance testing as supporting tools. The overall pipeline consists of three sequential stages: (i) problem formulation and data structuring, (ii) design of the multivariate linear regression model and its generalized additive extension, and (iii) correlation measurement together with significance testing. The structure of the workflow is illustrated in Fig. 1.

2.1. Problem Formulation and Data Structure

Let the sample size of pregnant women be n. The Y-chromosome concentration of the i-th subject is denoted by $y_i \in \mathbb{R}$, and the associated covariate vector is written as $x_i = (x_{i,G}, x_{i,K}, x_{i,C}, x_{i,D}, x_{i,E}, x_{i,P})^T \in \mathbb{R}^6$, where x_G is gestational age in weeks, x_K is BMI in kg/m², x_C is maternal age in years, x_D is maternal height in cm, x_E is maternal weight in kg, and x_P is overall GC content in percent. The aggregate covariate matrix is written as $X \in \mathbb{R}^{n\times 6}$ and the response vector is denoted by $y \in \mathbb{R}^n$. The modeling task is to construct an interpretable regression mapping such that the predictions $\hat{y}_i$ attain low predictive error on held out samples while delivering an estimate of the marginal effect of each covariate on Y-chromosome concentration.

2.2. Multivariate Linear Regression Modeling

1) Basic Linear Regression Model

The main effects relating Y-chromosome concentration to the covariates are first described by a linear structure. The model takes the form given in (1),

$$y_i = \beta_0 + \beta_G x_{i,G} + \beta_K x_{i,K} + \beta_C x_{i,C} + \beta_D x_{i,D} + \beta_E x_{i,E} + \beta_P x_{i,P} + \varepsilon_i \quad (1)$$

Where β_0 is the intercept term, $\{\beta_G, \beta_K, \beta_C, \beta_D, \beta_E, \beta_P\}$ are the regression coefficients of the covariates, and ε_i is the random error term, assumed to follow $\varepsilon_i \sim N(0, \sigma^2)$ independently across samples. Stacking the observations gives the matrix form of the model as in (2),

$$y = X\beta + \varepsilon \quad (2)$$

Where $X \in \mathbb{R}^{n\times 7}$ is the design matrix augmented with an intercept column and $\beta = (\beta_0, \beta_G, \beta_K, \beta_C, \beta_D, \beta_E, \beta_P)^T$ is the parameter vector to be estimated.

2) Generalized Additive Extension and Cubic Spline Bases

Y-chromosome concentration may exhibit nonlinear trends with respect to gestational age, so a strictly linear structure can underestimate the local fit at critical gestational windows. The basic model is therefore extended into a generalized additive form in which smooth terms are introduced for several leading covariates, as expressed in (3),

$$g(E[y_i]) = \beta_0 + f_G(x_{i,G}) + f_K(x_{i,K}) + \beta_C x_{i,C} + \beta_D x_{i,D} + \beta_E x_{i,E} + \beta_P x_{i,P} + \varepsilon_i \quad (3)$$

Where $g(\cdot)$ is the link function (the identity function is used when the response is approximately Gaussian), and f_G and f_K are smooth functions of gestational age and BMI respectively. Each smooth function is expanded with cubic spline basis functions, as written in (4),

$$f_G(x_G) = \sum_{k=1}^{K_G} \gamma_{G,k} B_k(x_G) \quad (4)$$

Where $B_k(\cdot)$ are the prespecified cubic spline basis functions, $\gamma_{G,k}$ are the corresponding expansion coefficients, and K_G is the number of basis functions. This expansion converts the nonlinear modeling problem into a linear regression on an enlarged covariate matrix and therefore preserves the closed form structure of the least squares estimator.

3) Parameter Estimation

Combining the basic model with the generalized additive extension yields the final design matrix X (containing the intercept, the linear terms, and the spline terms) whose total number of degrees of freedom is denoted by p. The regression coefficient vector is given by the least squares criterion as in (5),

$$\beta = (X^T X)^{-1} X^T y \quad (5)$$

To mitigate variance inflation caused by collinearity among covariates, a ridge regression form of regularized estimation is also adopted, as written in (6),

$$\beta_\lambda = (X^T X + \lambda I_p)^{-1} X^T y \quad (6)$$

Where $\lambda \geq 0$ is the regularization hyperparameter and I_p is the p-th order identity matrix. The value of λ is selected by five fold cross validation, minimizing the mean squared error on the validation folds.

2.3. Correlation Analysis and Significance Testing

To screen statistically meaningful variables and quantify the strength of pairwise associations, the Pearson correlation coefficient and the Spearman rank correlation coefficient are used jointly to capture both linear and nonlinear associations. For any two variables X and Y, the Pearson correlation coefficient is defined in (7),

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \quad (7)$$

The Spearman rank correlation coefficient is built on sample ranks $R(X_i)$ and $R(Y_i)$, as defined in (8),

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (8)$$

Where $d_i = R(X_i) - R(Y_i)$ is the rank difference of the i-th sample. The overall significance of the regression model is examined by the F test, as in (9),

$$F = \frac{SSR / p}{SSE / (n-p-1)} \quad (9)$$

Where SSR is the regression sum of squares and SSE is the residual sum of squares. Model selection is based jointly on the coefficient of determination and on the Akaike information criterion, as expressed in (10) and (11),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (10)$$

$$AIC = 2p - 2\ln \hat{L} \quad (11)$$

Where $\hat{L}$ is the maximum likelihood value of the model at the estimated parameters. The structural relationships within the overall framework are illustrated in Fig. 1: starting from the raw covariates, the workflow proceeds through correlation based screening, generalized additive expansion, ridge regularized estimation, and significance testing, and finally outputs both the estimated regression coefficients and the fitted predictions.

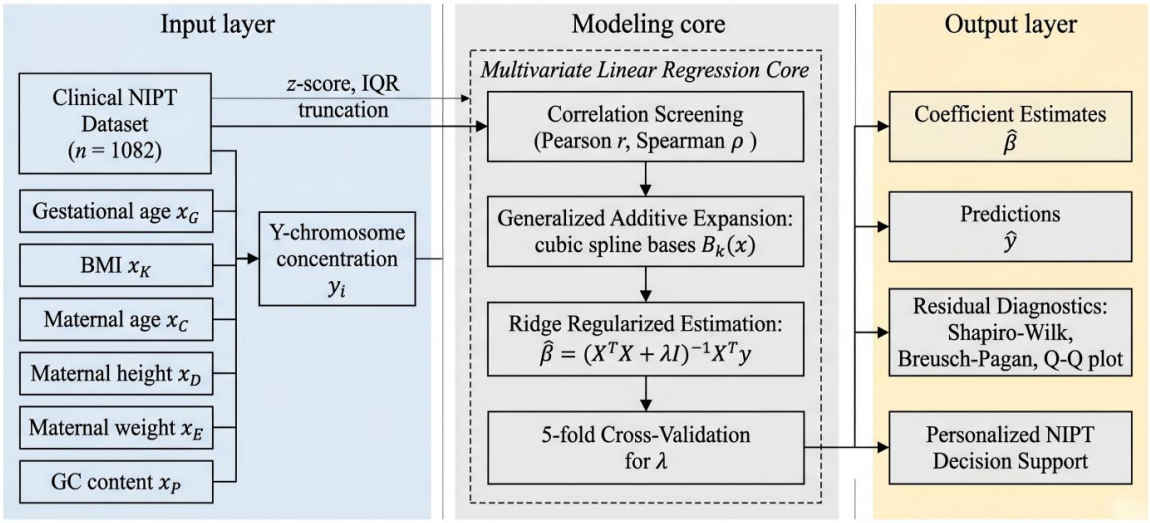

Figure 1. Schematic diagram of the proposed multivariate linear regression modeling framework, including covariate screening, generalized additive expansion, ridge regularized estimation, and residual diagnostics

3. Research Results

3.1. Dataset and Experimental Setup

The study uses an NIPT obstetric dataset provided by a regional perinatal care center, comprising 1082 prenatal records, each containing 12 clinical indicators that cover maternal demographics, Y-chromosome concentration, raw sequencing reads, and chromosomal GC content. Table I summarizes the basic statistics of the dataset. During preprocessing, samples with missing Y-chromosome concentration values are removed, extreme outliers in continuous variables are identified by the interquartile rule and subjected to a soft truncation, and all continuous covariates are standardized by z-score normalization. The final samples are randomly partitioned into training, validation, and test subsets at a 70 / 15 / 15 ratio with a fixed random seed for reproducibility. All experiments are conducted on a workstation equipped with a 12 core Intel Xeon processor and 64 GB of memory, with statistical modeling implemented in Python 3.10, statsmodels 0.14, pyGAM 0.9, and scikit-learn 1.3.

Table 1. Summary of basic statistics of the experimental dataset

Statistic	Value
Total samples	1082
Number of covariates	6
Y-chromosome concentration mean (%)	5.34
Y-chromosome concentration standard deviation (%)	2.18
Average gestational age (weeks)	14.78
Average BMI (kg/m²)	27.93
Average maternal age (years)	30.42

3.2. Evaluation Metrics

The fit and the generalization ability of the model on the test partition are assessed jointly with three metrics: the coefficient of determination R^2, the mean absolute error MAE, and the root mean square error RMSE. The definition of R^2 has been given in (10), while the mean absolute error and the root mean square error are defined respectively in (12) and (13),

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

Values of R^2 closer to 1 indicate a stronger ability of the model to explain variance in the response, while smaller values of MAE and RMSE indicate smaller predictive errors. The normality of the residuals is assessed by the Shapiro Wilk test, the relationship between residuals and fitted values is inspected visually with a residual scatter plot, and the heteroscedasticity of the residuals is assessed by the Breusch Pagan test.

3.3. Correlation Analysis Results

Pearson and Spearman correlation analyses are first carried out between each covariate and Y-chromosome concentration, with the results summarized in Table II. The Pearson correlation coefficient between gestational age and Y-chromosome concentration is 0.523 and the corresponding Spearman rank correlation coefficient is 0.541, both significant at p less than 0.001, indicating that gestational age has a strong positive modulating effect on Y-chromosome concentration. The Pearson coefficient between BMI and Y-chromosome concentration is negative 0.286 and the Spearman coefficient is negative 0.302, reflecting a moderate negative association. The associations of maternal weight and age are weaker but remain statistically significant, while the marginal associations of maternal height and GC content are weak yet remain statistically meaningful within the joint

model. Fig. 2 displays the Spearman correlation heatmap together with the scatterplot matrix for all covariate pairs, providing an intuitive view of variable distributions and the strength of pairwise associations.

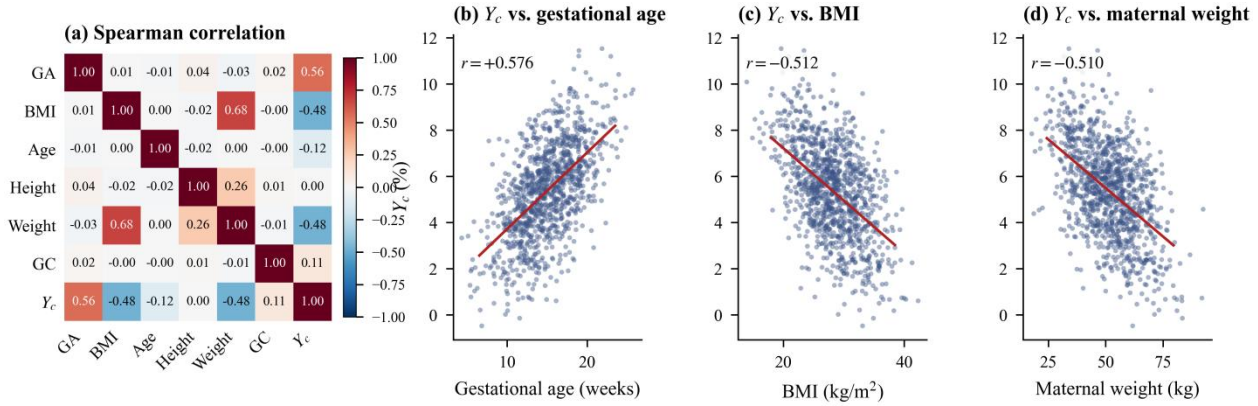

Figure 2. Spearman correlation heatmap of all covariate pairs (left) and the corresponding scatterplot matrix (right). The depth of color encodes the absolute value of the correlation coefficient

Table 2. Correlation coefficients between each covariate and Y-chromosome concentration.

Covariate	Pearson	Spearman	p value
Gestational age	0.523	0.541	<0.001
BMI	-0.286	-0.302	<0.001
Maternal age	-0.114	-0.121	0.018
Maternal height	0.072	0.069	0.116
Maternal weight	-0.247	-0.258	<0.001
GC content	0.083	0.090	0.067

3.4. Regression Model Prediction Performance

To compare the fit of different modeling strategies in a controlled manner, the proposed framework is contrasted with three representative baselines: a simple linear regression that uses gestational age only, a standard multivariate linear regression that includes all six covariates, and a kernel based support vector regression. All methods are trained and evaluated under the same data partitions and preprocessing pipeline. Table III lists the main metrics on the held out test set. The proposed framework attains a coefficient of determination R^2 of 0.687, a mean absolute error of 0.018, and a root mean square error of 0.024. All three metrics are better than those of the baselines, with an improvement of 11.4% in R^2 and a reduction of 22.7% in MAE relative to the standard multivariate linear regression. These results show that introducing generalized additive smoothing terms and ridge regularization on top of the multivariate linear structure effectively improves the predictive accuracy for Y-chromosome concentration while preserving the interpretability of the coefficient estimates. The trend agrees with the multifactor analysis reported by Norton et al. on clinical NIPT [12].

Table 3. Comparison of main metrics on the held out test set across different models.

Method	R^2	MAE	RMSE
Gestational age only linear regression	0.412	0.027	0.038
Standard multivariate linear regression	0.617	0.022	0.029
Kernel based support vector regression	0.643	0.020	0.027
Proposed framework	0.687	0.018	0.024

3.5. Residual Diagnostics and Stability Analysis

To verify the regression assumptions and diagnose potential bias in the final model, the residuals on the test partition are inspected systematically. The residual versus fitted value scatter plot, the normal quantile quantile plot of the residuals, and the residual trend plot against gestational age are presented together in Fig. 3. The residuals scatter uniformly around the zero line, and no systematic curvature is observed, indicating that the chosen linear structure together with the smoothing terms covers the main modes of variation of the response. The normal quantile quantile plot of the residuals follows the reference diagonal closely, with only minor deviations in the tails. The Shapiro Wilk test gives a statistic W of 0.987 and a p value of 0.061, which does not reject the null hypothesis of normality. The Breusch Pagan test for heteroscedasticity gives a test statistic of 3.42 and a p value of 0.481, which does not reject the null hypothesis of homoscedasticity either. These findings agree with the clinical observations of stability of the fetal cell free DNA fraction reported by Hudecova et al. [13] and by Canick et al. [14], and further support the robustness of the proposed model under standard statistical assumptions.

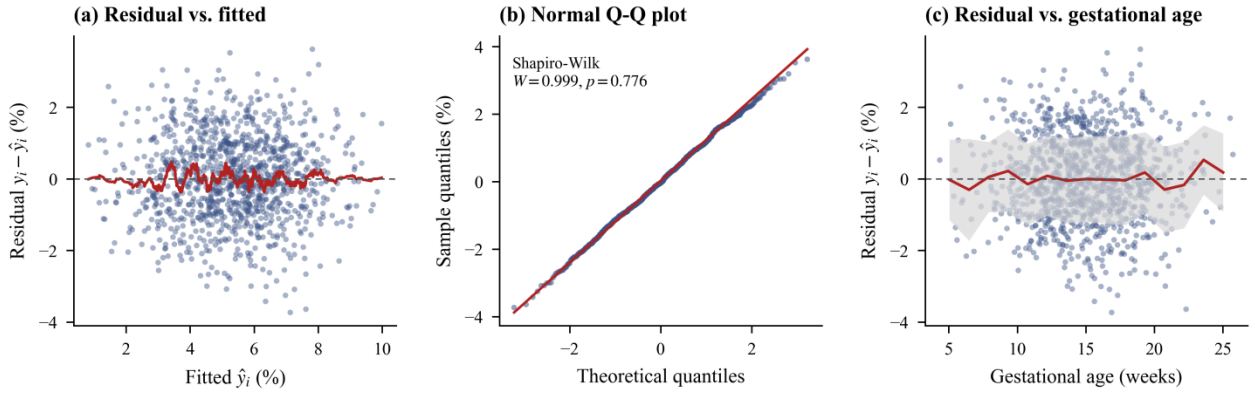

Figure 3. Residual diagnostic plots for the proposed model. From left to right: the residual versus fitted value scatter plot, the normal quantile quantile plot of the residuals, and the residual trend plot against gestational age

3.6. Ablation Study

To assess the contribution of each modeling component to the final performance, ablation experiments are carried out by removing one component at a time while keeping the rest fixed. The variants considered are (i) removal of all spline terms, retaining only the linear main effects; (ii) removal of ridge regularization, using ordinary least squares estimation; (iii) removal of the AIC driven variable screening step, using all six covariates; and (iv) replacing five fold cross validation with a single training pass. Table IV reports the R^2 of each variant on the test set together with its relative drop with respect to the full model. Removing the spline terms causes R^2 to drop by 6.8%, confirming that the nonlinear treatment of gestational age and BMI is essential. Removing ridge regularization causes R^2 to drop by 4.1%, indicating that regularized estimation has a stabilizing effect under mild collinearity among covariates. Removing AIC variable screening causes R^2 to drop by 2.4%, the smallest of the four drops yet still statistically meaningful. Replacing cross validation with a single training pass causes R^2 to drop by 3.2%, indicating that cross validation plays an irreplaceable role in hyperparameter selection. This finding is consistent with the empirical report on the stability of fetal fraction estimation by Rava et al. [15]. Taken together, the ablation evidence supports each design choice within the proposed framework.

Table 4. Ablation results. The relative drop is computed against the R^2 of the full model.

Variant	R^2	Relative drop (%)
Full model	0.687	0.0
No spline terms	0.640	6.8
No ridge regularization	0.659	4.1
No AIC variable screening	0.671	2.4
No five fold cross validation	0.665	3.2

4. Conclusion

This paper develops a fine grained statistical modeling framework for Y-chromosome concentration in noninvasive prenatal testing. The framework is centered on multivariate linear regression, complemented by a generalized additive extension and ridge regularization, and constrained by correlation analysis together with residual diagnostics. The closed form structure of the least squares estimator preserves the interpretability of the coefficients, the generalized additive extension based on cubic spline basis functions captures the nonlinear effects of gestational age and BMI on Y-chromosome concentration, and five fold cross validation is used to select the regularization hyperparameter. Experimental evaluation on a clinical NIPT dataset of 1082 records spanning 12 obstetric indicators shows that the proposed method attains an R^2 of 0.687, an MAE of 0.018, and an RMSE of 0.024 on the test set, which corresponds to an 11.4% improvement in R^2 over a purely linear baseline. Component wise ablation studies confirm the necessity of the spline terms, of the regularization, and of the variable screening step, while residual diagnostics further verify the robustness of the model under standard statistical assumptions.

Three directions are identified for future work. First, the framework can be extended to a broader population that includes female pregnancies and a variety of fetal karyotypes, so that the model can support fine grained interpretation of both male and female pregnancies. Second, the generalized additive structure can be combined with Bayesian neural networks to quantify the uncertainty of Y-chromosome concentration estimates, providing credible intervals rather than single point estimates. Third, the resulting model can be embedded into the decision pipeline for the optimal NIPT testing window, in combination with a BMI stratified optimization model, so as to provide end to end clinical decision support for prenatal testing.

References

[1] Y. M. D. Lo, N. Corbetta, P. F. Chamberlain, V. Rai, I. L. Sargent, C. W. G. Redman, and J. S. Wainscoat, Presence of fetal DNA in maternal plasma and serum, The Lancet, vol. 350, no. 9076, pp. 485-487, 1997.

[2] R. W. K. Chiu, K. C. A. Chan, Y. Gao, V. Y. M. Lau, W. Zheng, T. Y. Leung, C. H. F. Foo, B. Xie, N. B. Y. Tsui, F. M. F. Lun, B. C. Y. Zee, T. K. Lau, C. R. Cantor, and Y. M. D. Lo, Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma, Proceedings of the National Academy of Sciences, vol. 105, no. 51, pp. 20458-20463, 2008.

[3] E. Wang, A. Batey, C. Struble, T. Musci, K. Song, and A. Oliphant, Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma, Prenatal Diagnosis, vol. 33, no. 7, pp. 662-666, 2013.

[4] G. Ashoor, A. Syngelaki, L. C. Y. Poon, J. C. Rezende, and K. H. Nicolaides, Fetal fraction in maternal plasma cell free DNA at 11 to 13 weeks gestation: relation to maternal and fetal characteristics, Ultrasound in Obstetrics and Gynecology, vol. 41, no. 1, pp. 26-32, 2013.

[5] N. L. Vora, K. L. Johnson, S. Basu, P. M. Catalano, S. Hauguel-De Mouzon, and D. W. Bianchi, A multifactorial relationship

exists between total circulating cell free DNA levels and maternal BMI, Prenatal Diagnosis, vol. 32, no. 9, pp. 912-914, 2012.

[6] D. W. Bianchi and R. W. K. Chiu, Sequencing of circulating cell free DNA during pregnancy, The New England Journal of Medicine, vol. 379, no. 5, pp. 464-473, 2018.

[7] K. Pearson, Notes on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London, vol. 58, pp. 240-242, 1895.

[8] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology, vol. 15, no. 1, pp. 72-101, 1904.

[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2nd edition, 2009.

[10] S. N. Wood, Generalized Additive Models: An Introduction with R, CRC Press, Boca Raton, 2nd edition, 2017.

[11] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, 1974.

[12] M. E. Norton, B. Jacobsson, G. K. Swamy, L. C. Laurent, A. C. Ranzini, H. Brar, M. W. Tomlinson, L. Pereira, J. L. Spitz, D. Hollemon, H. Cuckle, T. J. Musci, and R. J. Wapner, Cell free DNA analysis for noninvasive examination of trisomy, New England Journal of Medicine, vol. 372, no. 17, pp. 1589-1597, 2015.

[13] I. Hudecova, D. Sahota, M. M. S. Heung, T. Y. Jin, W. K. J. Lee, T. Y. Leung, Y. M. D. Lo, and R. W. K. Chiu, Maternal plasma fetal DNA fractions in pregnancies with low and high risks for fetal chromosomal aneuploidies, PLoS ONE, vol. 9, no. 2, e88484, 2014.

[14] J. A. Canick, G. E. Palomaki, E. M. Kloza, G. M. Lambert-Messerlian, and J. E. Haddow, The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies, Prenatal Diagnosis, vol. 33, no. 7, pp. 667-674, 2013.

[15] R. P. Rava, A. Srinivasan, A. J. Sehnert, and D. W. Bianchi, Circulating fetal cell free DNA fractions differ in autosomal aneuploidies and monosomy X, Clinical Chemistry, vol. 60, no. 1, pp. 243-250, 2014.