

# Analysis of Risk Factors for Fibrocystic Breast Disease

Bin Fang

Department of Mathematics & Statistics, San Diego State University, San Diego, USA

## Abstract

Fibrocystic Breast Disease, which is characterized by noncancerous breast lumps which can sometimes cause discomfort periodically related to the menstrual cycle, affects an estimated 30-60% of women all over the world. A case-control study involving 255 women with Fibrocystic Breast Disease and 790 controls was conducted at two hospitals in New Haven, Connecticut from 1977 to 1979. In order to study the risk factors of this disease, conditional logistic regression model is applied. The final model is selected with individual Wald test and information criteria and the goodness-of-fit are assessed by diagnostic statistics. From the model it is found that weight, age at menarche and marital status are significant risk factors for Fibrocystic Breast Disease.

## Introduction

Fibrocystic Breast Disease are marked by appearance of fibrous tissue and a lumpy, cobblestone texture in the breasts. These lumps are smooth with defined edges, and are usually free-moving in regard to adjacent structures. The exact mechanism of the condition is not fully understood, although it is known to be relevant to hormone levels and the menstrual cycle. Fibrocystic breast, with elevated risk of developing breast cancer, is common among females, with estimated prevalence ranging from 30% to 60% of all women. Imaging and biopsy are the most widely used diagnostic method for this disease and personal medical history is also helpful for the diagnostics.

A case-control study was conducted at two hospitals in New Haven, Connecticut between November 1977 and May 1979. The case group includes 255 women aged 20-74 years with a biopsy-confirmed diagnosis of fibrocystic breast disease, while the control group involves 790 women admitted to surgical services in the same hospitals over the same time period for disorders not involving the breast. Matching was based on the age of the subject at the time of interview. The data we have are a subset of this study, involving 50 cases and 150 matched controls, with three controls per case.

Conditional logistic regression model is used to fit the data. The stratum-specific logistic regression model can be written as:  $\Pi_k(\mathbf{x}) = \exp(\alpha_k + \beta' \mathbf{x}) / [1 + \exp(\alpha_k + \beta' \mathbf{x})]$ . In this model,  $\Pi_k(\mathbf{x})$  denotes the conditional probability of Fibrocystic Breast Disease on the  $k$ th stratum,  $\alpha_k$  denotes the contribution to the log odds of matched variables (in this case this is the age of each subject) within the  $k$ th stratum,  $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$  denotes the parameter vector,  $\mathbf{x}$  denotes the design matrix. On each stratum the summation of  $\Pi_k(\mathbf{x})$  equal 1.

The full model is fit first and some of the predictors are eliminated from the model under model selection criteria. The final model shows that the association between three predictors (weight, age at menarche and marital status) and the occurrence of Fibrocystic Breast Disease are significant, indicating they are risk factors for Fibrocystic Breast Disease. No significant interaction is found among them. Model diagnostics shows strata 4, 12, 18 have large influences on the estimated coefficients in the model.

R studio and SAS are used to analyze the data.

## Summary Information

There are 8 variables included in our study. Final Diagnosis of Fibrocystic Breast Disease (FNDX) is the response variable, and as stated before, there are 50 cases and 150 cases which comprises 25% and 75% of all the subjects respectively. 7 Predictors includes Regular Medical Check-ups (CHK), Marital Status (MST), Weight of the Subject (WT), Age at Menarche (AGMN), Age at Last Menstrual Period (AGLP), Highest Grade in School (HIGD) and Degree (DG). The means and standard deviations of continuous variables and numbers and percentages of categorical variables are summarized in Table 1. (HIGD and DEG are ordinal variables and will be treated as continuous variables in the final conditional logistic regression model. Further discussion is presented in statistical analysis section.)

Table 1: Descriptive Statistics and Summary Information for the Case Group and Control Group

| Variable | Description   | Codes/Values      | Cases     |          | Controls  |          |
|----------|---|-------------------|-----------|----------|-----------|----------|
|          |   |                   | Mean(No.) | SD(Pct.) | Mean(No.) | SD(Pct.) |
| FNDX     | Final Diagnosis                                       | Outcome Variable  | 50        | 25%      | 150       | 75%      |
| CHK      | Regular Medical Check-ups                             | 1 = Yes           | 41        | 82%      | 78        | 52%      |
|          |   | 2 = No            | 9         | 18%      | 72        | 48%      |
| MST      | Marital Status  | 1 = Married       | 36        | 72%      | 109       | 73%      |
|          |   | 2 = Divorced      | 4         | 8%       | 16        | 11%      |
|          |   | 3 = Separated     | 1         | 2%       | 5         | 3%       |
|          |   | 4 = Widowed       | 2         | 4%       | 15        | 10%      |
|          |   | 5 = Never Married | 7         | 14%      | 5         | 3%       |
| WT       | Weight of the Subject                                 | Pounds            | 126.06    | 18.27    | 149.6     | 33.35    |
| AGMN     | Age at Menarche                                       | Years             | 13.92     | 1.75     | 12.63     | 1.62     |
| AGLP     | Age at Last Menstrual Period                          | Years             | 42.36     | 7.72     | 40.91     | 7.19     |
| HIGD     | Highest Grade in School<br>(as a continuous variable) | 5-20              | 13.2      | 3.02     | 12.46     | 2.39     |
| DEG      | Degree<br>(as a continuous variable)                  | 0 = None          | 1.6       | 1.43     | 1.16      | 0.99     |
|          |   | 1 = High School   |           |          |           |          |
|          |   | 2 = Jr. College   |           |          |           |          |
|          |   | 3 = College       |           |          |           |          |
|          |   | 4 = Masters       |           |          |           |          |
|          |   | 5 = Doctoral      |           |          |           |          |

From the summary table, we can intuitively feel there are some differences of predictors, with regard to the mean (for continuous variables) and the percentage (for categorical variables) between the case group and control group, although we are not sure whether these differences are significant right here. This means these predictors could be potential candidates for the risk factors of Fibrocystic Breast Disease. Since our response variable is a binary variable, conditional logistical regression model are supposed to be helpful for analysis of the relations between the disease and its risk factors.

## Statistical Analysis

### 1. Ordinal Variables

HIGD and DEG are both ordinal variables. Before fitting the full model, we need to assess whether to treat them as continuous variables or as categorical variables. After fitting the model with continuous and categorical HIGD respectively, Likelihood Ratio Test (LRT) suggests the model with continuous HIGD is better. Similarly, LRT between the model with continuous and categorical DEG infers the model with continuous DEG is better. Thus, HIGD and DEG are treated as continuous variables in the following statistical analysis. The results of LRT are attached in Appendix A.

### 2. Full Model

Next, we fit the conditional logistical regression model with all the predictors. The summary of the full model is attached in Appendix B, from where we can see some of the coefficient estimates are not significant at  $\alpha = 0.05$  level, suggesting the multicollinearity between some of those variables. Thus, model selection is required to select the “best” models to interpret the relationship between the occurrence of Fibrocystic Breast Disease and risk factors. The likelihood Ratio between the full model and the null model is 56.6 and the p-value for LRT is very small, suggesting all those variables together have good predicting power on the occurrence of Fibrocystic Breast Disease.

### 3. Model Selection

We use backward elimination strategy to kick out the most insignificant variable at a time in each step and the criteria for insignificance are individual Wald test, AIC and AICC values. The details of each step of model selection are summarized in Appendix C. Notice that during each step, p-value of the Wald test for each eliminated variables are greater than  $\alpha = 0.05$ ; Besides, AIC and AICC values decreases as the elimination goes on. The likelihood Ratio between the full model and the final reduced model is 8.84 and the p-value for LRT is greater than  $\alpha = 0.05$ , suggesting the full reduced model is preferred and sufficient for prediction of Fibrocystic Breast Disease.

### 4. Final Reduced Model

There are 4 predicting variables remaining in the final reduced model: Medical check up, Weight, Age at Menarche and Marital Status (Never Married vs. Others). The summary of the final model is shown in Table 2.

Table 2: Summary of the Final Reduced Model

| Variables             | Coefficient | Standard Error | Odds Ratio | 95% CI for OR   | P-value |
|-----------------------|-------------|----------------|------------|-----------------|---------|
| Medical check up      | -1.161      | 0.447          | 0.313      | (0.130, 0.752)  | 0.009   |
| Never Married         | 1.593       | 0.736          | 4.920      | (1.162, 20.821) | 0.030   |
| Weight of the Subject | -0.028      | 0.010          | 0.972      | (0.953, 0.991)  | 0.005   |
| Age at Menarche       | 0.359       | 0.128          | 1.423      | (1.115, 1.840)  | 0.005   |

The association between medical check up and the occurrence of Fibrocystic Breast Disease is probably due to the selection bias of the case-control study design. Since cases and controls are samples from two hospitals in New Haven, patients with the disease will probably have more medical check up than those who are not in hospitals. Thus, the negative value of the coefficient does not imply a negative association between medical check up and the disease; actually in common sense the disease

itself has no relation with regard to the medical check up.

The other three variables are supposed to be the risk factors for the disease: the decrease of weight and increase of age at menarche and never-married status would elevate the probability of the occurrence of Fibrocystic Breast Disease. The Odds Ratio shows the strength of association between one risk factor and the disease adjusting for all the other risk factors, respectively. Never-married status have the strongest association with the disease: a never-married female will be 4.92 times likely to develop the Fibrocystic Breast Disease than the other female adjusting for all other risk factors. The Odds Ratio associated with 1-year increase of age at menarche and 1-pound increase in weight is 1.423 and 0.972 respectively adjusting for all other risk factors.

To check if there is interaction between those predictors in the final reduced model, each interaction term are added to the model respectively and individual Wald test shows that no significant interaction between those predictors exist. The tests for each interaction term are attached in Appendix D.

## 5. Model Diagnosis

Since Overall goodness fit of test for conditional logistic regression is not available right now (Hosmer 2013), we directly study the goodness of fit for single covariate patterns. Plots of Leverages, Pearson's Chi-square and Cook's Distance versus Estimated Probabilities are presented from Appendix E to G, respectively. Strata with large values of leverage,  $\hat{\alpha}^2$ ,  $\hat{\beta}$  are summarized in Appendix H.

To study whether those poorly fit or influential strata have great influence on the coefficient estimates of the model, each stratum is deleted at one time, and new coefficient estimates and percent of change compared to the original one is calculated. Results are presented in Appendix I. Some coefficient estimates are dramatically changed in three Strata (Stratum 4, 12, 18). However, without further information we suggest keeping these data right now and meanwhile checking for data errors for those subjects and consulting with a specialist on the clinical plausibility of the data.

## Conclusion

The conditional logistic regression model suggests three risk factors for Fibrocystic Breast Disease: weight, age at menarche and marital status. Since it is known that this disease is tied to hormone levels and the menstrual cycle, it is not surprising that age at menarche and marital status are significantly correlated with the disease. The decrease of weight among fibrocystic breast patients might be due to the comfort and uneasiness which go hand in hand with the disease.

Although stratum 4, 12, 18 have large influences on the estimated coefficients of predictors, we suggest retain them in the model for the time being before strong evidence of data errors have been found. Besides, due to unavoidable selection bias in the case-control study, a relevant prospective study may help verify the risk factors.

## Bibliography

1. Pastides, H., Kelsey, J.L., Holford, T.R., and LiVolsi, V.A., (1985). The Epidemiology of Fibrocystic Breast Disease. *American Journal of Epidemiology*, 121, 440-447.
2. Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression Analysis*.
3. Therneau T. (2015). *A Package for Survival Analysis in S* (version 2.38).
4. R 3.2.0 Vienna, Austria
5. SAS 9.4 Cary, NC

6. <http://en.wikipedia.org/wiki/>

## Appendix

### A. Likelihood Ratio Test to Decide the Treatment of Ordinal Variables

| Variables | Model                       | Loglikelihood | Likelihood Ratio | Df. | P-value |
|-----------|-----------------------------|---------------|------------------|-----|---------|
| DEG       | model with categorical DEG  | -66.69612     | 4.31             | 4   | 0.365   |
|           | model with continuous DEG   | -64.53987     |                  |     |         |
| HIGD      | model with categorical HIGD | -67.81183     | 17.09            | 14  | 0.251   |
|           | model with continuous HIGD  | -59.26214     |                  |     |         |

### B. Summary of the Full Model

| Variables                    | Coefficient | Standard Error | Odds Ratio | P-value |
|------------------------------|-------------|----------------|------------|---------|
| Medical check up             | -0.987      | 0.485          | 0.373      | 0.042   |
| Divorced                     | -0.200      | 0.717          | 0.819      | 0.780   |
| Separated                    | -0.521      | 1.549          | 0.594      | 0.740   |
| Widowed                      | -0.848      | 0.872          | 0.428      | 0.330   |
| Never Married                | 1.937       | 0.855          | 6.397      | 0.024   |
| Weight of the Subject        | -0.032      | 0.011          | 0.968      | 0.003   |
| Age at Menarche              | 0.405       | 0.141          | 1.499      | 0.004   |
| Age at Last Menstrual Period | 0.107       | 0.055          | 4.113      | 0.050   |
| Highest Grade in School      | -0.343      | 0.218          | 0.710      | 0.120   |
| Degree                       | 0.926       | 0.494          | 2.524      | 0.061   |

### C. Model Selection Details

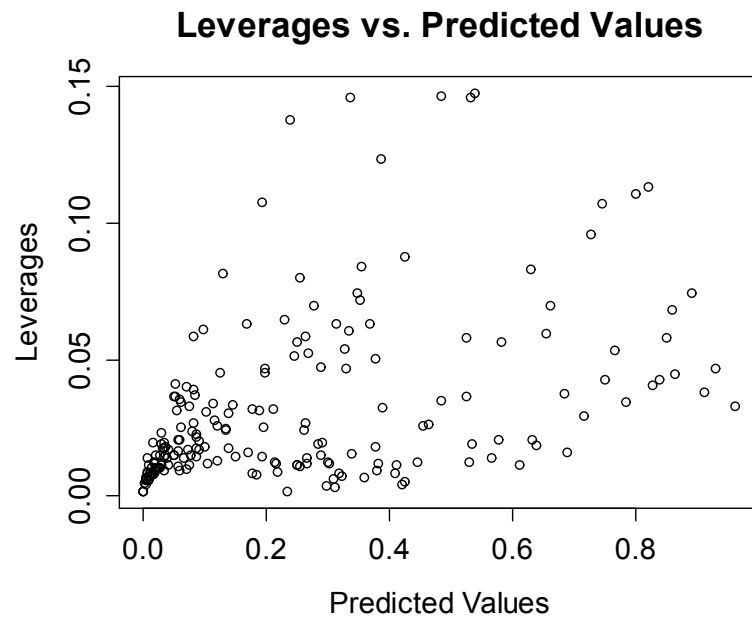
|        | Most Insignificant Variable  | P-value of Wald Test | AIC after elimination | AICC after elimination |
|--------|------------------------------|----------------------|-----------------------|------------------------|
| step 1 | Divorced                     | 0.78                 | 100.04                | 100.51                 |
| step 2 | Separated                    | 0.77                 | 98.87                 | 98.98                  |
| step 3 | Widowed                      | 0.34                 | 98.13                 | 98.50                  |
| step 4 | Highest Grade in School      | 0.12                 | 97.41                 | 97.63                  |
| step 5 | Degree                       | 0.18                 | 97.23                 | 97.41                  |
| step 6 | Age at Last Menstrual Period | 0.08                 | 97.12                 | 97.38                  |

### D. Test for Interaction

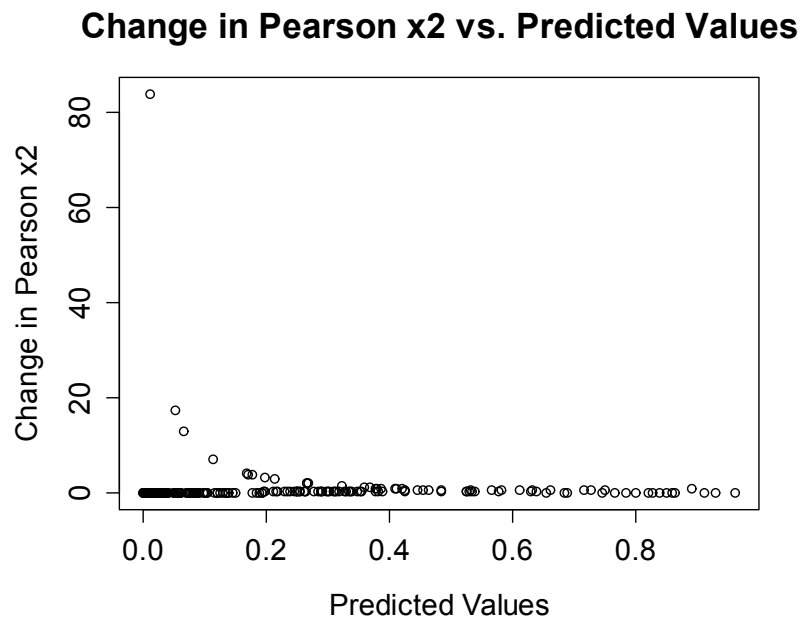
| Interaction Term                      | P-value |
|---------------------------------------|---------|
| Medical Check Up with Age at Menarche | 0.53    |
| Medical Check Up with Weight          | 0.17    |
| Medical Check Up with Never Married   | 0.54    |

|                                    |      |
|------------------------------------|------|
| Age at Menarche with Never Married | 0.93 |
| Age at Menarche with Weight        | 0.85 |
| Weight with Never Married          | 0.42 |

#### E. Plots of Leverages Versus the Estimated Probability

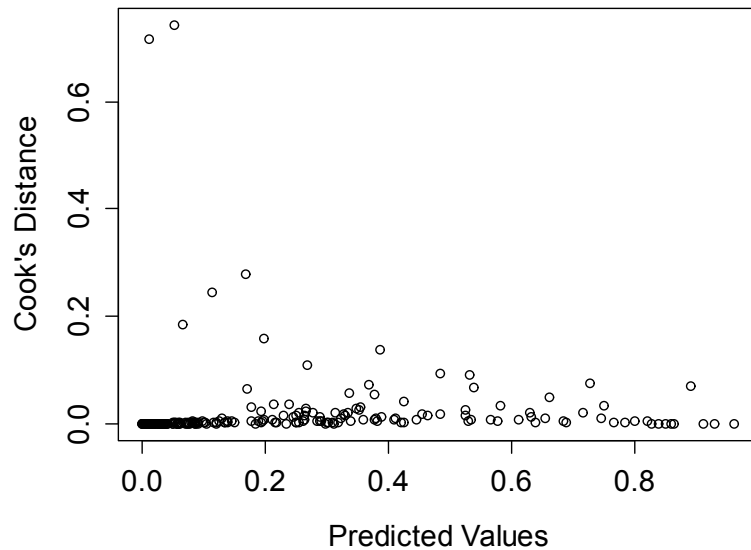


#### F. Plots of Change in Pearson's Chi-square Versus the Estimated Probability





### G. Plots of Cook's Distance Versus the Estimated Probability

### Cook's Distance vs. Predicted Values



## H. Stratum with Large Values of Diagnostic Statistics

| STR | OBS | h              |  2 |  hat |
|-----|-----|----------------|---|---|
| 4   | 1   | <b>0.14657</b> | 0.551021  | 0.0946328   |
| 4   | 2   | 0.025642       | 0.120734  | 0.0031773   |
| 4   | 3   | 0.01697        | 0.04082   | 0.0007047   |
| 4   | 4   | 0.084196       | 0.354722  | 0.0326119   |
| 10  | 1   | 0.034026       | <b>6.9891</b>   | <b>0.246189</b>   |
| 10  | 2   | 0.032118       | 0.3886  | 0.0128954   |
| 10  | 3   | 0.017254       | 0.07376   | 0.001295  |
| 10  | 4   | 0.087562       | 0.424982  | 0.0407833   |
| 12  | 1   | 0.00847        | <b>83.8379</b>  | <b>0.716207</b>   |
| 12  | 2   | 0.079964       | 0.255071  | 0.0221691   |
| 12  | 3   | 0.013713       | 0.007385  | 0.0001027   |
| 12  | 4   | 0.095932       | 0.725892  | 0.0770255   |
| 14  | 1   | 0.050412       | 1.029405  | 0.054649  |
| 14  | 2   | <b>0.14615</b> | 0.336615  | 0.057616  |
| 14  | 3   | 0.010152       | 0.02313   | 0.0002372   |
| 14  | 4   | 0.026943       | 0.263235  | 0.0072887   |
| 17  | 1   | 0.046664       | 3.257341  | 0.1594412   |
| 17  | 2   | 0.06464        | 0.230496  | 0.0159288   |
| 17  | 3   | <b>0.14624</b> | 0.531352  | 0.0910176   |



|    |   |                |                |                 |
|----|---|----------------|----------------|-----------------|
| 17 | 4 | 0.01147        | 0.040513       | 0.0004701       |
| 18 | 1 | 0.041259       | <b>17.2595</b> | <b>0.742763</b> |
| 18 | 2 | 0.074417       | 0.890561       | 0.0716008       |
| 18 | 3 | 0.008518       | 0.007842       | 6.738E-05       |
| 18 | 4 | 0.036663       | 0.049534       | 0.0018852       |
| 24 | 1 | 0.013902       | <b>13.044</b>  | <b>0.18389</b>  |
| 24 | 2 | 0.013129       | 0.119726       | 0.0015928       |
| 24 | 3 | 0.046497       | 0.329859       | 0.0160855       |
| 24 | 4 | 0.034758       | 0.483647       | 0.0174157       |
| 26 | 1 | 0.012486       | 0.69131        | 0.008741        |
| 26 | 2 | <b>0.13765</b> | 0.237629       | 0.0379322       |
| 26 | 3 | 0.006425       | 0.003863       | 2.498E-05       |
| 26 | 4 | 0.063304       | 0.313291       | 0.0211728       |
| 31 | 1 | 0.063238       | 4.121987       | <b>0.278262</b> |
| 31 | 2 | 0.009154       | 0.034658       | 0.0003202       |
| 31 | 3 | 0.012001       | 0.216299       | 0.0026273       |
| 31 | 4 | 0.056198       | 0.58109        | 0.0346005       |
| 39 | 1 | <b>0.14773</b> | 0.396022       | 0.0686456       |
| 39 | 2 | 0.007706       | 0.014036       | 0.000109        |
| 39 | 3 | 0.05622        | 0.24901        | 0.0148334       |
| 39 | 4 | 0.045168       | 0.198663       | 0.0093976       |

OBS 1: Case      OBS 2-4: Control

High leverage,  $\hat{\beta}_1^2$ ,  $\hat{\beta}_2^2$  values are highlighted. ( $h > 0.13$ ,  $\hat{\beta}_1^2 > 5$ ,  $\hat{\beta}_2^2 > 0.18$ )

### I. Estimated Coefficients and Percent of Change when Strata Are Deleted

| Data             | Medical check up |               | Never Married |                | Weight of the Subject |               | Age at Menarche |               |
|------------------|------------------|---------------|---------------|----------------|-----------------------|---------------|-----------------|---------------|
| All Kept         | -1.161           |               | 1.593         |                | -0.028                |               | 0.359           |               |
| <b>Delete 4</b>  | -1.108           | -4.57%        | 1.270         | <b>-20.28%</b> | -0.029                | 3.57%         | 0.366           | 1.95%         |
| Delete 10        | -1.342           | 15.59%        | 1.685         | 5.78%          | -0.026                | -7.14%        | 0.404           | 12.53%        |
| <b>Delete 12</b> | -1.241           | 6.89%         | 1.679         | 5.40%          | -0.035                | <b>25.00%</b> | 0.452           | <b>25.91%</b> |
| Delete 14        | -1.187           | 2.24%         | 1.798         | 12.87%         | -0.027                | -3.57%        | 0.347           | -3.34%        |
| Delete 17        | -1.086           | -6.46%        | 1.888         | 18.52%         | -0.028                | 0.00%         | 0.380           | 5.85%         |
| <b>Delete 18</b> | -1.479           | <b>27.39%</b> | 2.247         | <b>41.05%</b>  | -0.029                | 3.57%         | 0.368           | 2.51%         |
| Delete 24        | -1.366           | 17.66%        | 1.687         | 5.90%          | -0.030                | 7.14%         | 0.357           | -0.56%        |
| Delete 26        | -1.109           | -4.48%        | 1.574         | -1.19%         | -0.028                | 0.00%         | 0.367           | 2.23%         |
| Delete 31        | -1.350           | 16.28%        | 1.665         | 4.52%          | -0.032                | 14.29%        | 0.312           | -13.09%       |
| Delete 39        | -1.220           | 5.08%         | 1.345         | -15.57%        | -0.027                | -3.57%        | 0.358           | -0.28%        |

Stratum 4, 12, 18 are highlighted due to large change of some coefficients after deletion.