Applying STAR Models to Breast Cancer Screening Data from Central Portugal

Elisa Maria Castro Rocha Duarte

Final Project for the Master's in Statistical Techniques

Department of Statistics and Operational Research

University of Santiago de Compostela

July 2011

Abstract

Structured Additive Regression (STAR) Models is the framework chosen in this study since it provides a generalized setting for models, such as, Generalized Additive Models, Generalized Additive/Geoadditive Mixed Models, Varying Coefficient Models (VCMs), and ANOVA type interaction models. It is of the utmost importance to work with models that are flexible enough to deal with different and complex structures of data sets, taking into consideration a multitude of covariates while exploring possible spatial and temporal correlations. Using a Bayesian approach, both fixed effects and random effects are considered random variables with appropriate priors and are estimated using REML-Restricted Maximum Likelihood.

This study will comprise the analysis of a database containing approximately 260,000 data records of women that entered the Screening Program for the first time in the central region of Portugal. It is believed that the period of time between the age of menarche and the age of menopause has been increasing over time. Therefore, a new variable called Window was defined as the difference between the age of menopause and the age of menarche, which represents a woman's years of fertility. The evolution in time and space of the variables Window and the Age of Menarche will be analyzed, exploring the possible associations with other variables, such as Contraceptive Pills, Pregnancy Status, Nursing Status and Purchasing Power Index. As in the database there are records of women who have not yet reached menopause, the variable Window can be considered as censored and a Cox proportional hazards model was performed using the variable

Window as the survival time. Inference was carried out with the help of the software *BayesX* and *R*.

Our spatio-temporal analysis will be the starting point to further studies, such as the evolution of breast cancer mortality in Portugal and the possible effects of the screening program in this evolution.

Keywords: Structured Additive Regression (STAR), Breast Cancer, Screening program, Biostatistics.

1 Introduction

This study focuses on women registering for the first time in the Breast Cancer Screening Program for central Portugal. It is believed that early diagnosis of the disease can make a difference in a patient's chances of survival, resulting either from a monthly self-exam test, an annual doctor's appointment and/or the integration of women in Screening Programs.

The Portuguese Cancer League (LPCC) is a private non-profit organization dealing with a multiple of issues related to oncology. The LPCC is made up of 5 branches with head offices in Coimbra (central branch, LPCC-NRC), Lisbon (south branch, LPCC-NRS), Oporto (north branch, LPCC-NRN), Angra do Heroismo (Azores island branch) and Funchal (Madeira island branch). One of the activities of the league is The National Breast Cancer Screening Program with women 45 years of age or older, whose main goal is the diagnosis of the disease, not only in older women but also in younger women, with increasing new cases detected. As mentioned before, an early diagnosis is important to reduce the mortality rate. This study is based on 260,000 first registries of women in central Portugal, which is made up of 12 distinct sub-regions, with a total of 100 counties, representing approximately 25% of Portugal In our study data was available for a total of 83 counties.

Breast cancer risk is believed to be associated with several reproductive factors, such as early menarche and late menopause ages. Therefore, we may think that the smaller the difference between these two moments, the lower the risk of disease or, in other words, the shorter a women's years of fertility, the lower the risk of disease.

In this study the evolution in time and space of the variables Window representing a woman's years of fertility, and the Age of Menarche will be analyzed, exploring the possible associations with other variables, such as Contraceptive Pills, and Purchasing Power Index. Since we are dealing with different and complex structures of data sets, and taking into consideration a multitude of covariates while exploring possible spatial and temporal correlations, structured additive regression (STAR) models [1–3] were the choice for this work.

STAR models are very flexible, covering in a general and generic framework, such as Generalized Additive Mixed Models (GAMM) [4], Varying Coefficient Models (VCM) [5], Geoadditive Models [6]. Besides the exponential family regression, STAR models also include non-standard regression such as Cox-like hazard regression [7].

The inference performed in this work is based on empirical Bayesian estimation. Empirical Bayes [3, 8] approach can be seen as a generalized mixed model representation, where smoothing parameters are estimated based on restricted maximum likelihood (REML) or marginal likelihood estimation techniques, and regression parameters are estimated based on penalized likelihood estimation techniques. From a Bayesian perspective, this yields empirical Bayes or posterior mode estimates for the STAR models. However, estimates can also merely be interpreted as penalized likelihood estimates from a frequentist perspective. Thus, mixed model based estimation bridges the gap between a frequentist and a Bayesian approach.

The methodology is available in the open domain statistical package *BayesX* (http://www.stat.uni-muenchen.de/∼bayesx/).

2 Database description

For this study, the data were provided by the Central Regional Nucleus of the LPCC (LPCC-NRC). The data base consists of 259,652 registries of women that were registered for the first time in the Breast Cancer Screening Program in central Portugal. Due to administrative reasons, we only have information on 77 counties of the central region of Portugal, and 6 additional counties from the North of Portugal that border this region, namely in the regions of Douro and Tamega, representing only 4.4% of our data. Table 1 shows how the 83 counties are distributed along the 14 regions studied, and the total number of counties per region, according to the National Institute of Statistics (INE) of Portugal.

The use of contraceptive pills can be seen as factor that influences the

Table 1: Number of counties studied in the 14 regions that are part of this study.

REGION NUTS III	NO. COUNTIES - INE	NO. COUNTIES LPCC
Baixo Mondego	8	8
Baixo Vouga	12	7
Beira Interior Norte	9	9
Beira Interior Sul	4	4
Cova da Beira	3	3
Dão-Lafões	15	15
Médio Tejo	10	3
Oeste	12	2
Pinhal Interior Norte	14	14
Pinhal Interior Sul	5	4
Pinhal Litoral	5	5
Serra da Estrela	3	3
Douro	19	5
Tâmega	15	1

fertile period of women. This contraceptive method was introduced in Portugal around 1970. Thus, from the 259,652 registries having women born between 1900 and 1963, women born before 1920 were removed from the analysis, remaining only women that were possibly exposed to this contraceptive method.. Because of software limitations a random sample of $100,000$ registries from the database was selected. The variables considered in this study were: Menopause Age, Menarche Age, Year of Birth, Contraceptive Pills (Anovulatory), Pregnancy Status, Nursing Status, Purchasing Power Index (PPI) [9] and County Code. A new variable called Window was de-

fined as the difference between the age of menopause and the age of menarche or the age of a woman when entering the screening program, which represents a woman's years of fertility. Contraceptive Pills, Pregnancy and Nursing status are dichotomous variables, in which 1 means "Yes" and 0 means "No". Tables 2 and 3 shows some summary statistics for these variables.

Table 2: Statistics of quantitative variables.

	Year of Birth	Age of Menopause	Age of Menarche	PPI
Mean ± Std	1944 ± 9.87	47.94 ± 4.77	13.41 ± 1.79	71.21 ± 24.79
Median	1945	48	13	68.64
Min - Max	1920 - 1963	36 - 66	9 - 18	22.09 - 139.13

Table 3: Statistics of categorical variables.

	Contraceptive Pills	Pregnancy	Nursing Status
0	59%	8%	30%
1	41%	92%	70 %

The database includes in addition to the women who reached menopause at the time of screening, women who are not yet at that stage. Considering as a survival time the women's fertility period (Window), and as the event of interest reaching menopause, the problem can be viewed as a survival data study where censored data occurs for women that, at time of registration at the Breast Cancer Screening Program, did not achieved menopause. In this perspective, an analysis using a Cox model approach can be performed in order to study which covariates influence the survival time. From our 100,000 randomly selected sample, 63,009 (63%) women reached menopause (uncensored data) and the remaining 36,991 (37%) are classified as right censored data.

3 A brief description of the Analysis

Breast cancer risk is believed to be associated with early menarche and late menopause ages. So we may think, that the shorter a women's years of fertility, the lower the risk of disease. The main objective of this study is to perform a spatial-temporal analysis of the menarche and the fertility period of a woman (the Window variable) exploring possible associations with covariates, such as the purchasing power index, providing the socio-economic dimension to the study, Contraceptive Pills, Pregnancy Status and Nursing Status.

The first approach taken in this study, can be seen as two separate analysis. The first, explores the evolution of the menarche accordingly to the year of birth and the purchasing power index and in addition to the spatial effects on this variable. In this analysis all 100,000 randomly selected women are considered, including those who reached menopause and those who have not. The second one is a similar analysis but with the variable Window (the woman's years of fertility) as the response, and including other covariates, such as contraceptive pills, nursing status and pregnancy as fixed effects. In this analysis the fixed effects were all statistically significant at 5% significance level and therefore, a study was conducted to compare the continuous and spatial effects of the two categories that compose the fixed effects in this model. The next chapter will describe each model performed and the corresponding results.

In the second approach to our dataset, the analysis was focused on the variable Window including, in addition to the women who reached menopause at the time of the screening, women who are not yet at that stage. By considering the survival time as the women's fertility period (Window) and, as the event of interest, reaching menopause, the data was modeled as survival data in order to study which covariates influence the survival time. A Cox proportional hazards model was then carried out. In this scenario, 63,009 (63%) women reached menopause and 36,991 (37%) of the records were considered as right censored data.

4 Discussion and Conclusions

The main objective of this study was to perform a spatio-temporal analysis of the variables Age of Menarche and Window, the women's years of fertility, in order to understand their behavior mainly accordingly to the covariates Year of Birth and Purchasing Power Index of the counties where these women

live. The effect of reproductive factors, such as, the use of contraceptive pills, whether women were at least one time pregnant and whether or not they breastfed during birth were also considered in this study. The data analyzed resulted from women who, for the first time, were part of the Breast Cancer Screening Program for central Portugal. This program is carried out by the Portuguese Cancer League (LPPC).

Due to the complexity of the data structures present in this study, we used Structured Additive Regression (STAR) models in order to analyze the data. STAR models have the advantage to be extremely flexible, combining several types of covariates, such as non-linear, fixed effects and spatial effects. The models in this study were estimated using empirical Bayesian procedures based on mixed models methodology.

The results show early menarche seems to occur in younger women, with decreasing fertility periods for women born after 1933. Comparing the two categories of the variable contraceptive pills, we see that there is almost no change in the effect of the Year of Birth on the variable Window for women born between 1920 and 1945 and did not used contraceptive pills. This behavior contrasts with the clear decreasing effect of the Year of Birth for all women that used contraceptive pills. After 1945, there are no differences in the effect of the year of birth between the two groups, being the effect a decreasing one. Also the two groups of the variable Pregnancy Status, shows overlapping year of birth effects. The effect of the year of birth for the two groups in the variable Nursing Status is the same until 1935, with a slower decreasing effect for women who breastfed at least one time. This might suggest that breastfeeding might be associated with longer periods of fertility in women.

From an economical perspective, we conclude that counties with higher purchasing power show early menarche ages, and larger periods of women's fertility.

The maps with structured and unstructured spatial effects lead to the conclusion that interior counties have lower ages of menarche and larger periods of fertility. The Window analysis shows interior part of the central Portugal as the counties that contribute more positively to the values of the fertility period. To notice that Leiria and Mealhada are those that most negatively contribute to this variable. In the unstructured space results we have Leiria, Coimbra, Viseu and Castelo Branco with strictly negative credible intervals. A curious fact that for the moment we do not have a reason for it, but that should be addressed in future research.

Counties with the most negative structured effect are in general those with the most negative unstructured effects.

Comparing the two groups of the variable anovulatory (whether a woman

took contraceptive pills or not) there is no apparent differences in the spatial effect distribution. Nevertheless, we can distinguish the strength of the effect between the two groups, since for women who took contraceptive pills, the negative effect, i.e. contributing to a reduction in the Window variable, moves more towards the interior part of central Portugal.

The structured spatial effects have the same behavior for women who were never pregnant and who were pregnant at least one time, being the posterior mode values more positive for the later. Leiria is the only region with a strictly negative credible interval. There are no apparent differences of the unstructured spatial effects for both groups. In this case, the counties of Viseu and Coimbra present strictly negative credible intervals. For women who were never pregnant, there is one additional county, Leiria, with strictly negative credible interval.

When we look at the structured effects for both groups of the variable Nursing Status, it is clear the same decreasing effect on the variable Window when moving towards the coastline regions for both groups. Again, Leiria appears with strictly negative 95% credible interval. The unstructured spatial effects are very similar for both groups. Comparing the 95% credible intervals, Viseu, Coimbra and Leiria appear with strictly negative intervals for the group of women who never breastfed. For the other group, only Viseu and Coimbra have also strictly negative credible intervals.

The Cox model enhanced the conclusion reached for the Window spatial effects analysis. Later menopause contribute for larger woman fertility period, in the counties of southern interior central Portugal.

In conclusion, this study was able to present a clear picture of the behavior of the age of menarche and the women's years of fertility, for women born between 1920 and 1963 that were first registered in the Breast screening Program in central Portugal. It is important to notice the singular local effects that emerged from Leiria, Coimbra and Viseu in several models in our study. A fact that will deserve our utmost attention in future research studies, where other information, such as, diagnosis and data from follow up rounds in the screening program will be included.

References

1. Belitz C, Lang S (2008) Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis 53: 61-81.

2. Brezger A, Lang S (2006) Generalized structured additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis 50: 967-991.

3. Fahrmeir L, Kneib T, Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. Statistica Sinica 14: 731-761.

4. Hastie T, Tibshirani RL (1990) Generalized additive models. London: Chapman & Hall.

5. Hastie T, Tibshirani RL (1993) Varying-coefficient models. Journal of the Royal Statistical Society B 55: 757-796.

6. Kammann EE, Wand MP (2003) Geoadditive models. Journal of the Royal Statistical Society C 52: 1-18.

7. Kneib T, Fahrmeir L (2004) A mixed model approach for structured hazard regression. SFB Discussion Paper 400 .

8. Kneib T (2006) Mixed model based inference in structured additive regression. Ph.D. thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-UniversitäMünchen.

9. INE website (accessed 2011) Instituto Nacional de Estatística, Portugal http://www.ine.pt .