Master of Applied Mathematics Program
Math 574 — Bayesian Computational Mathematics
Department of Applied Mathematics
Illinois Institute of Technology

Bayesian Inferential Regression on Cervical Cancer Risk Factors

Submitted by

Robert L. Mead II

Master of Applied Mathematics Program

05-01-2022

1 Introduction

Cervical cancer occurs in the lining of the cervix, or the lower part of the uterus. The cervix connects the uterus, where a fetus grows, to the vagina, the birth canal. The cervix is comprised of three regions, the endocervix, transformation zone, and the ectocervix. The endocervix is the canal from the uterus to the transformation zone that is primarily glandular cells. The ectocervix is the portion of the cervical canal that connect the transformation zone to the vagina that contain squamous cells. The transformation zone is the canal that is between the endocervix and the ectocervix where the cells change from glandular cells to squamous cells. The transformation zone is the most frequent location cell abnormalities and precancerous cells are to develop [3]. Screenings for the precancerous changes in the transformation zone assess the degree of abnormalities in the cervical tissue. The cervical screenings are used to monitor any abnormalities in the cervical tissue in the transformation zone. The precancerous cells, on average, do not need any treatment and will go away. In some cases, the precancers turn into cancer cells on the cervix.

The American Cancer Society, also referred to as the ACS, identifies the three common cervical cancers are squamous cell carcinomas, adenocarcinomas and adenosqaumous carcinomas [2]. Squamous cell carcinomas is the most common of cervical cancer, as it develops in the ectocervix with proliferation of cancer from the squamous cells in the transformation zone. The second common cervical cancer, adenocarcinomas, develops in the endocervix proliferating from the glandular cells. The least common of the cervical cancers has proliferation of cancer from the squamous cells of the ectocervix and the glandular cells of the endocervix. Women between the ages of thirty five and forty four are most frequently diagnosed with cervical cancer, and has been the most common cause of cancer deaths for American women. The development of the Papanicolaou Test, commonly referred to as The Pap Test, is a procedure that monitors the cells from the cervical canal to monitor for abnormal and precancerous cell growth. The development of the Pap Test has provided an accurate detection of cervical cancer and has helped the decline of cervical cancer cases, it still remains the second most prominent type of cancer for women globally.

2 Background

2.1 Dataset Description

The dataset used focuses on the indicators and diagnosis of cervical cancer by featuring demographic information, habits, and medical records of 858 patients from the Hospital Universitario de Caracas, in Caracas, Venezuela. The dataset was obtained from the UCI Machine Learning Repository[4]. In the dataset, four test methods are used to indicate cervical cancer. The dataset represents each of the test methods as an indicator variable, where ”1” represents a positive test result and ”0” represents a negative test result for cervical cancer. The test methods for cervical cancer used in the dataset are the Hinselmann Test, Schiller Test, Cytology Test, and a Biopsy. The Hinselmann Test examines cervical cells on an instrument called the colpopscope, which enables a healthcare provider to examine the cells of the cervix more closely. The Schiller Test colors the cervical cells in iodine. The iodine colors of the cervix allow the healthcare provider to determine if there are cancerous cells forming in the cervix. Healthy cervical cells will turn brown from the iodine, while the abnormal cells will remain unchanged by the application of the iodine. A Cytology Test, similar to the Schiller Test, samples cells from the cervix and are examined under a microscope. Lastly, the Biopsy surgically removes a piece of cerivcal tissue and is tested for adnormal, precancerous or cancerous cells [6]. There are thirty four predictor variables that describe the patients age, sexual tendencies detailed through history of types of contraception (IUD or Hormonal Contraceptives) sexual transmitted diseases (Condylomatosis, Vaginal Condylomatosis (VC), Cervical Condylomatosis (CC), Vulvo Perineal Condylomatosis (VPC), Syphilis, Pelvic Inflammatory Disease (PID), Genital. Herpes (GH), Molluscum Contagiosum (MC), Hepatitis B, HPV, HIV, AIDS), pregnancy, number of sexual partners, and age of first intercourse.

Fifteen binary explanatroy variables are represented in the dataset where a response of $x_i = 1$ indicates a response of ”yes”, and $x_i = 0$ indicates a patient response of ”no” for the fifteen explanatory variables. For the i^{th} patient $Smokes_i$ denotes if the patient is a smoker, IUD_i indicates if the patient has an IUD, $STDs_i$ indicates if the patient has a general STD, $STDs.Condylomatosis_i$, $STDs.Cervical.Condylomatosis_i$,$STDs.Vaginal.Condylomatosis_i$, $STDs.Vulvo.Perineal.Condylomatosis_i$, $STDs.Syphilis_i$,$STDs.Pelvic.Inflammatory.Disease_i$, $STDs.Genital.Herpes_i$, $STDs.Molluscum.Contagiosum_i$,$STDs.AIDS_i$,$STDs.HIV_i$,$STDs.HPV_i$ and $STDS.Hepatitis.B_i$ denotes if the patient has the specific STD with a response of either 0 or 1.

2.2 Dataset Cleaning

The dataset needed to be cleaned. The removal of two predictor variables, ” STD’s Time Since First Diagnosis” and ”STD’s Time Since Last Diagnosis” are removed due to the lack of responses from the patients, thus, showing the futility of those predictor variables. The predictor variable ”STDs” was removed due to the similarity the data in the column shared with the corresponding predictor variable ”STDs Number”. Similarly, there are missing observations in almost all the predictor variables. Shown in the table below. To account for those missing values, median imputation is used to account for the skewed data in each of those columns.

Table 1: Frequency Table of Missing Values from Each Explanatory Variable

Explanatory Variable	Count
Age	0
Number.of.sexual.partners	26
First.sexual.intercourse	7
Num.of.pregnancies	56
Smokes	13
Smokes..years.	13
Smokes..packs.year.	13
Hormonal.Contraceptives	108
Hormonal.Contraceptives..years.	108
IUD	117
IUD..years.	117
STDs	105
STDs..number.	105
STDs.condylomatosis	105
STDs.cervical.condylomatosis	105
STDs.vaginal.condylomatosis	105
STDs.vulvo.perineal.condylomatosis	105
STDs.syphilis	105
STDs.pelvic.inflammatory.disease	105
STDs.genital.herpes	105
STDs.molluscum.contagiosum	105
STDs.AIDS	105
STDs.HIV	105
STDs.Hepatitis.B	105
STDs.HPV	105
STDs..Number.of.diagnosis	0
Dx.Cancer	0
Dx.CIN	0
Dx.HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0

3 Bayesian Inference

3.1 The Prior Distribution

An essential part of Bayesian Inference starts with a prior distribution. The prior distribution represents the description of knowledge about the true value of a set of parameters, θ. There are several priors that can be used in a Bayesian Inference such as, Noninformative Prior Distributions, Conjugate Prior Distributions, and Improper Prior Distributions. The selection of the prior distribution used indicates the knowledge about the parameters in the distribution. A noninformative prior, that will be used, assumes that each β follows the Multivariate Normal Distribution and has a covariance matrix $\sigma\varepsilon_0$.

$$\beta_0, \beta_1, ..., \beta_{33}|\sigma^2 \sim N((b_0, b_1, ..., b_{33})^T, \sigma\varepsilon_0) \tag{3.1}$$

$$\frac{1}{\sigma^2} \sim Gamma(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}) \tag{3.2}$$

The values of all the hyperparameters $\theta = (b_0, b_1, ...b_{33}, \varepsilon_0, v_0, \sigma^2)$ have a multivaraite Normal-Gamma conjugate family. In starting the Bayesian Inference, the values of the hyperparameters need to be specified. Since there is not relevant prior information about the variances, covariances and coefficients and the hyperparameters adopting a noninformative prior distribution is used to specify the values of the hyperparameters. The prior distribution of the coefficients and the prior distribution of the variance are represented by the equations (3.3) and (3.4) respectively. .

$$p(\beta_0, \beta_1, ..., \beta_{33}|\sigma^2) \propto 1 \tag{3.3}$$

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{3.4}$$

By choosing an noninformative prior distribution, this will have minimal effect on the computing the posterior distribution of the coefficients and the covariance matrix.

3.2 The Regression Model

The objective of the Bayesian inference is to fit a model that describes the relationship between a Cervical Cancer Diagnosis and its thirty three predictor variables. The Cervical Cancer Diagnosis offers a binary response, where y_i is an indicator of cervical cancer. When $y_i = 1$ indicates the patient has a positive diagnosis for cervical cancer and $y_i = 0$ indicates a negative cervical cancer diagnosis. The health measurements are related to thirty three covariates through a regression model. The regression model

$$E(Dx_i|x_i, \theta) = \beta_0 + \beta_1 Age_i + ... + \beta_{33} Biopsy_i \tag{3.5}$$

The analysis of the explanatory variables that have a strong effect, and other explanatory variables with a weaker effect on the Cervical Cancer Diagnosis can be observed in Table 2. The explanatory variables Smokes (Packs/Year), STDs (Number), STDs Vaginal Condylomatosis, STDs Vulvo Perineal Condylomatosis, STDs Syphilis, STDs Pelvic Inflammatory Disease, STDs Genital Herpes, STDs Molluscum Contagiosum, STDs HIV, STDs Hepatitis B, STDs (Number of Diagnosis), Dx Cancer,Dx CIN all have a strong effect on the Diagnosis of Cervical Cancer. Where explanatory variables, Smokes, Hormonal Contraceptives Per Year , STDs Condylomatosis and STDs HPV have a weaker effect on the Diagnosis of Cervical Cancer.

3.3 The Posterior Distribution

The posterior analysis for the regression model is similar to the posterior analysis of the mean and the variance of a sampling model. The joint density of the Ordinary Least Sqaures Regression is characterized by the product of the posterior distribution of the regression vector, β, conditional on the variance, and the marginal posterior distribution of the variance. In order to simulate the posterior distribution of the regression coefficients conditioned on the variance it is necessary to simulate each β from the conditional posterior distribution $f(\beta|\sigma^2, y)$. The posterior distribution of the variance is simulated from the marignal posterior distribution $f(\sigma^2|y)$.

$$f(\beta, \sigma^2|y) = f(\beta|y, \sigma^2)f(\sigma^2|y) \tag{3.6}$$

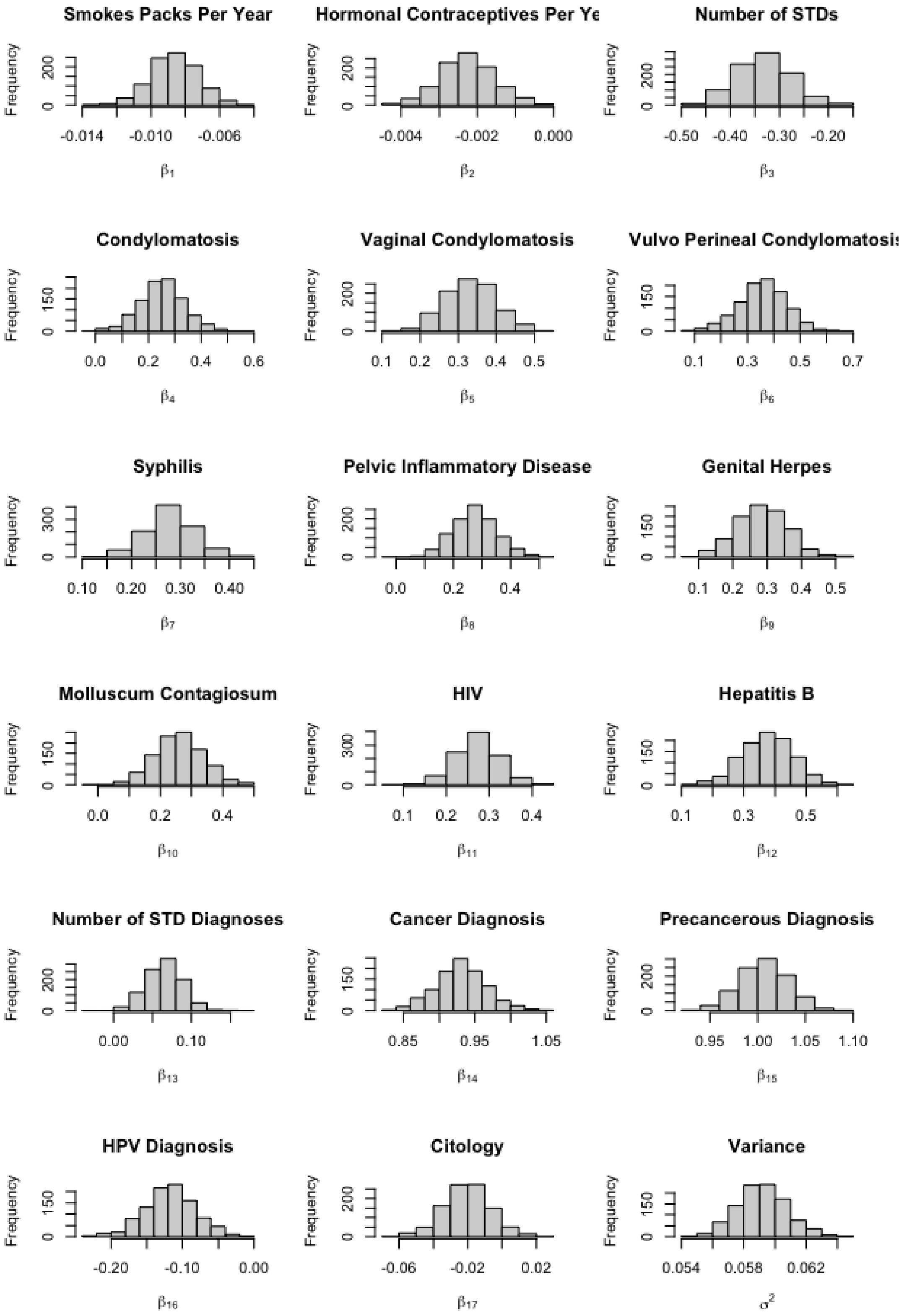
The regression vector, β is simulated from the multivariate normal density (MVND) with a mean of $\hat{\beta}$ and a variance matrix $V_\beta \sigma^2$. The simulated draws of the regression coefficients from the posterior distribution of β conditional on the variance , and the simulated draws of the marginal posterior distribution of the variance are represented through Figure 1.

Generally, the posterior means of the regression parameters are similar to the estimated coefficient values in the ordinary least squares regression. This result can be anticipated due to the prior for β was non-informative. There may exist deviations from the posterior means due to the errors that were in the 1000 simulated draws. The absence of strong deviation of the posterior distribution means of the regression coefficients and the estimates in the ordinary least squares regression show that the significant predictor variables help indicate the diagnosis of cervical cancer from the patients. Next, the posterior predictive distribution will be analyzed in the context of Cervical Cancer Diagnosis.

Table 2: Regression for Cervical Cancer Diagnosis

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0202	0.0154	-1.31	0.1898
Age	-0.0006	0.0004	-1.59	0.1126
Number.of.sexual.partners	0.0016	0.0014	1.17	0.2438
First.sexual.intercourse	0.0016	0.0009	1.71	0.0880
Num.of.pregnancies	0.0033	0.0020	1.66	0.0972
Smokes	0.0216	0.0090	2.40	0.0166
Smokes..years.	-0.0005	0.0010	-0.48	0.6315
Smokes..packs.year.	-0.0086	0.0014	-6.13	0.0000
Hormonal.Contraceptives	0.0023	0.0053	0.43	0.6653
Hormonal.Contraceptives..years.	-0.0023	0.0007	-3.12	0.0019
IUD	0.0215	0.0109	1.98	0.0482
IUD..years.	-0.0003	0.0017	-0.16	0.8768
STDs	-0.0067	0.0263	-0.26	0.7977
STDs..number.	-0.3348	0.0531	-6.30	0.0000
STDs.condylomatosis	0.2564	0.0847	3.03	0.0026
STDs.vaginal.condylomatosis	0.3339	0.0626	5.33	0.0000
STDs.vulvo.perineal.condylomatosis	0.3609	0.0872	4.14	0.0000
STDs.syphilis	0.2821	0.0486	5.80	0.0000
STDs.pelvic.inflammatory.disease	0.2823	0.0754	3.74	0.0002
STDs.genital.herpes	0.2921	0.0769	3.80	0.0002
STDs.molluscum.contagiosum	0.2677	0.0756	3.54	0.0004
STDs.HIV	0.2752	0.0504	5.46	0.0000
STDs.Hepatitis.B	0.3781	0.0792	4.78	0.0000
STDs..Number.of.diagnosis	0.0649	0.0225	2.89	0.0040
Dx.Cancer	0.9319	0.0353	26.36	0.0000
Dx.CIN	1.0075	0.0249	40.45	0.0000
Dx.HPV	-0.1186	0.0364	-3.26	0.0012
Hinselmann	0.0114	0.0141	0.81	0.4198
Schiller	-0.0114	0.0127	-0.90	0.3693
Citology	0.0339	0.0105	3.22	0.0013
Biopsy	-0.0204	0.0135	-1.51	0.1319

Figure 1: Histogram of simulated draws from the conditional posterior distributions of β_1, β_2, ..., β_{17}, and simulated draws from the marginal posterior distribution of σ^2



3.4 The Posterior Predictive Distribution

The Posterior Predictive Distribution is used to predict future observations, $\tilde{y}$,that correspond to the a covariate vector, x^*. The posterior predictive distribution is the product of two distributions, the sampling distribution, $p(\tilde{y}|\beta, \sigma^2)$, the posterior distribution, $f(\beta, \sigma^2|y)$ both averaged over the parameters β and σ^2. The posterior predictive distribution is represented by:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\beta, \sigma^2) f(\beta, \sigma^2|y) d\beta d\sigma^2 \tag{3.7}$$

To obtain future responses that correspond to a specific covariate in the covariate vector x^* simulations of β and σ^2 from the joint posterior distribution are essential. Also, simulations from the the sampling distribtuion, $p(\tilde{y}|\beta, \sigma^2)$ will produce simulated $\tilde{y}$. Through both of these simulations the posterior predictive distribution can give an indication of what the future data might look like given the data and the regression model. The posterior predictive distribution can be used to examine the descriptive adaquecy of the model that is being considered to identify the significant coefficients that are indicative of diagnosing cervical cancer. Draws of the posterior predictive distribution are simulated for all $y_1^* ... y_{853}^*$.Each of the 853 distributions are represented in the plot below to show the relationship between the posterior predictive distributions of each of the 853 observations.On the plot, the observed response, y is represented as a point for each observation. The presence of both the posterior predictive distributions and each response variable for all 853 observations determines the significance of the posterior predictive distribution. If any of the observations is not within the 95 percent interval based region for each observation, the corresponding observations could be classified as an outlier, however, that requires more analysis.

The Posterior Predictive Distribution for each patient $y_1 ... y_{853}$ shows that the posterior predictive distribution was relatively accurate in comparison to the original response variable. There are fourteen observable response variables that are exceeding the 95 percent confidence interval of the posterior predictive distribution. Those observations are that are exceeding the band of the posterior predictive distribution can be classified as outliers. Further analysis will be needed to determine the validity of each outlier.

4 Conclusion

In this Bayesian Analysis, a non-informative prior, which provides little information about the dataset and the explanatory variables, is utilized to allow the dataset to have greater influence on the posterior distribution. First, a ordinary least squares regression was performed under the assumptions of normality. From the thirty three predictor variables that were used in the dataset,

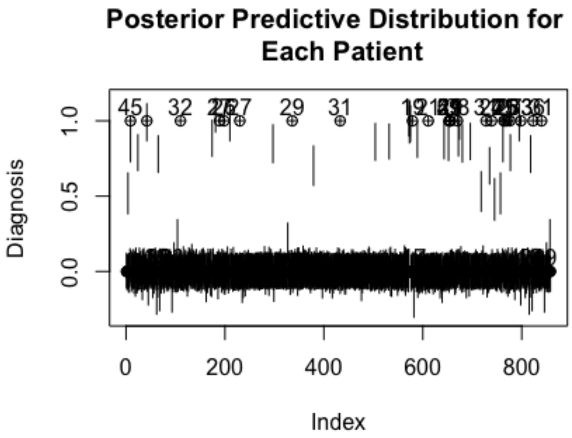


Figure 2: Posterior Predictive Distribution

seventeen of the predictor variables were classified as having a significance in predicting the diagnosis of cervical cancer. Following the Ordinary Least Squares Regression, simulating draws of the regression coefficients conditioned on the variance , the conditional posterior distribution, and simulating draws of the variance conditioned on the data, the marginal posterior distribution, enabled the opportunity to calculate the posterior distribution. The posterior distribution showed that the regression coefficients and the mean values for the posterior distribution of each of the regression coefficients were nearly identical. Lastly, the Posterior Predictive Distribution was calculated through simulating draws of the sampling distribution and the posterior distribution. The posterior predictive distribution showed that some of the future observations would not be similar to the already observed outcomes in the response variable. There were fourteen observations that were classified as outliers in a 95 confidence interval. Further analysis of the posterior distribution can be completed as an extension to determine the validity of the outliers.

References

[1] Albert, Jim. “Regression Models.” Bayesian Computation with R, 2nd ed., Springer New York, New York, NY, 2009, pp. 205–229.

[2] American Cancer Society Medical Content, et al. “Cervical Cancer - Symptoms and Causes - Mayo Clinic.” Mayo Clinic, 17 June 2021, https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501.

[3] “Cervical Cancer Overview.” Cervical Cancer Overview — What Is Cervical Cancer? , National Cervical Cancer Coalition, https://www.nccc-online.org/hpvcervical-cancer/cervical-cancer-overview/.

[4] Fernandes, Kelwin, et al. “Cervical Cancer (Risk Factors) Dataset.” UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+/28Risk+Factors/29. Accessed 26 Mar. 2022.

[5] Gelman, Andrew, et al. Bayesian Data Analysis. 3rd ed., CRC Press, 2013.

[6] “NCI Dictionary of Cancer Terms.” National Cancer Institute, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/.