A Bayesian Model for Cervical Cancer Prediction

Marcus Choi, Aining Xu, Jingdan Zou, Ran Yu

Spring 2023

Proposal

About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide.

In this project we will be developing bayesian models in order to identify which cancer diagnostic tests will be the most accurate in predicting an individual's risk of developing cervical cancer in the future based on various risk factors such as demographics, smoking status, presence of STDs, number of sexual partners, and the use of hormonal contraceptives.

By conducting observing the effects on the posterior inference of our data, we can determine which diagnostic test is the most effective in identifying individuals at higher risk for cervical cancer. This information can be used to improve early detection and intervention efforts and contribute to public health initiatives. We will evaluate the performance of each model using a certain metric, accounting for uncertainty and variability in the data.

Overall, this project aims to identify the most accurate diagnostic test for predicting cervical cancer risk and provide valuable insights into the impact of risk factors on disease risk, benefiting public health efforts.

Attribution

Marcus Choi - PM: Create project proposal, schedule meeting times, organize final report, assist in areas of the project as needed

Aining Xu - Develop prediction models and simulations, create plots and figures, discuss results with team members, assist in areas of the project as needed

Jingdan Zou - Develop prediction models and simulations, create plots and figures, discuss results with team members, assist in areas of the project as needed

Ran Yu - Collect data and materials, make data visualization, help develop prediction models,assist in areas of the project as needed

Raw data

Data Origin

The dataset is obtained from UCI Repository, collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients.

Target Variable Selection

Hinselmanns test refers to colposcopy using acetic acid. Meanwhile, colposcopy using Lugoliodine includes Schillers test, Cytology and Biopsy. Positive exams results don't necessarily imply a diagnostic, but as multiple exams return positive, the greater the likelihood of cervical cancer).

In this project, we selected only one target variable to measure cervical cancer. We conducted separate logistic regressions on all of the original target variables (Hinselmann, Schiller, Citology, and Biopsy) and see which one would be the best for cervical cancer measurement.

```r
# use auc scores to test which model predicts the outcomes the best
set.seed(1111)

auc1 <- vector()
auc2 <- vector()
auc3 <- vector()
auc4 <- vector()
train <- data.frame()
test <- data.frame()

selected_rows <- replicate(100,sample(1:nrow(cancer),ceiling(nrow(cancer)*0.5),
                           replace=FALSE))
selected_rows <- as.data.frame(selected_rows)

for (i in 1:100) {
  train <- cancer[selected_rows[,i], ]
  test <- cancer[-selected_rows[,i], ]

  fit1 <- glm(Hinselmann ~ Age + Number.of.sexual.partners + Hormonal.Contraceptives +
              IUD + Smokes + STDs, data = train, family = "binomial")
  test$predicted.prob1 <- predict(fit1, test, type = "response")
  pred1 <- prediction(test$predicted.prob1, test$Hinselmann)
  perf1 <- performance(pred1, "auc")
  auc1[i] <- round(perf1@y.values[[1]], 3)

  fit2 <- glm(Schiller ~ Age + Number.of.sexual.partners + Hormonal.Contraceptives +
              IUD + Smokes + STDs, data = train, family = "binomial")
  test$predicted.prob2 <- predict(fit2, test, type = "response")
  pred2 <- prediction(test$predicted.prob2, test$Schiller)
  perf2 <- performance(pred2, "auc")
  auc2[i] <- round(perf2@y.values[[1]], 3)

  fit3 <- glm(Citology ~ Age + Number.of.sexual.partners + Hormonal.Contraceptives +
              IUD + Smokes + STDs, data = train, family = "binomial")
  test$predicted.prob3 <- predict(fit3, test, type = "response")
  pred3 <- prediction(test$predicted.prob3, test$Citology)
  perf3 <- performance(pred3, "auc")
  auc3[i] <- round(perf3@y.values[[1]], 3)

  fit4 <- glm(Biopsy ~ Age + Number.of.sexual.partners + Hormonal.Contraceptives +
              IUD + Smokes + STDs, data = train, family = "binomial")
  test$predicted.prob4 <- predict(fit4, test, type = "response")
  pred4 <- prediction(test$predicted.prob4, test$Biopsy)
  perf4 <- performance(pred4, "auc")
  auc4[i] <- round(perf4@y.values[[1]], 3)
}
df <- data.frame(
  Target_Variable = c("Hinselmann", "Schiller", "Citology", "Biopsy"),
  AUC = c(mean(auc1), mean(auc2), mean(auc3), mean(auc4))
)
```

```r
# Table 1: AUC
knitr::kable(df)
```

Target_Variable	AUC
Hinselmann	0.56038
Schiller	0.56557
Citology	0.48060
Biopsy	0.51989

As Table 1 shown, Schiller has the highest AUC value. In this case, we selected Schiller as the target variable. Following this, we would conduct an EDA on risk factors for cervical cancer leading to a Schiller Examination.

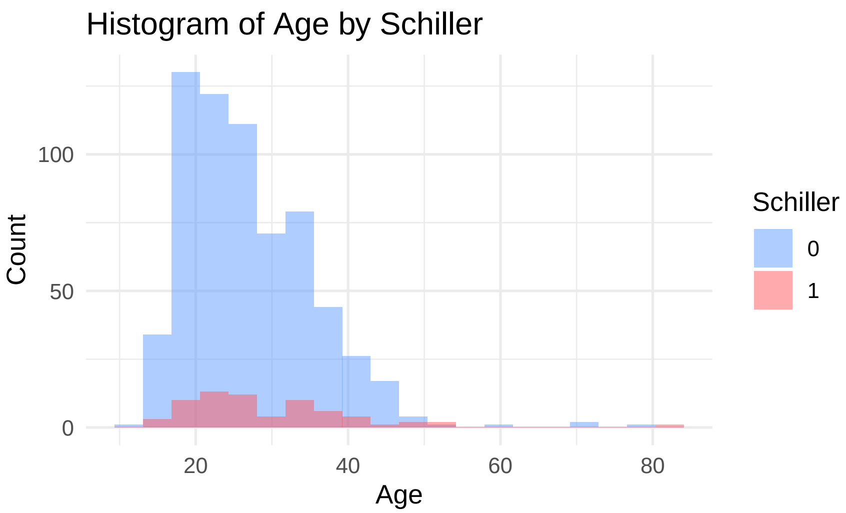
```r
# drop other target variables
cancer <- cancer %>% select(-Hinselmann, -Citology, -Biopsy)
```

Exploratory Data Analysis

Schiller Taking values between 0 (negative) or 1 (positive). 644 out of 712 observations (90.5%) were negative for abnormal cells, while the remaining 68 observations (9.5%) were positive.

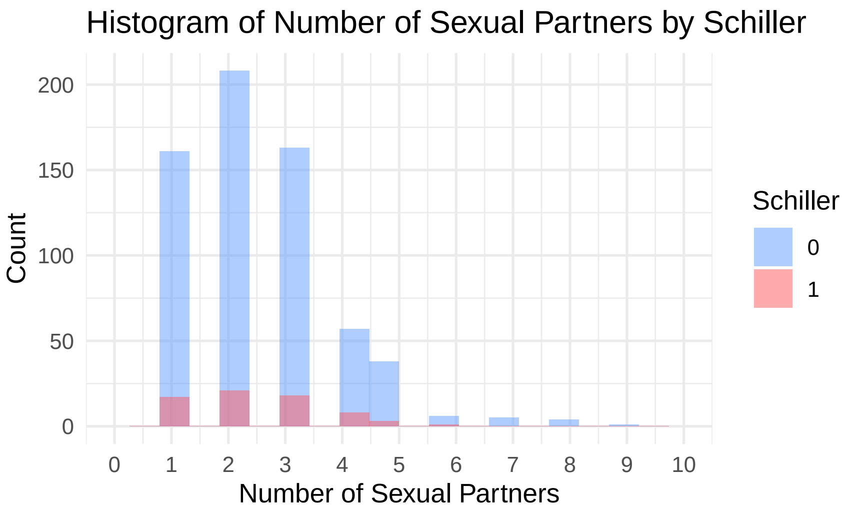
Age The age of respondents. Values ranged from 13 to 84. We computed descriptive statistics (M = 27.25; SD = 8.77; skewness = 1.40; kurtosis = 4.73).

```r
cancer %>% ggplot(aes(x = Age, fill = factor(Schiller))) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 20) +
  scale_fill_manual(values = c("#619CFF", "#FF595E"), name = "Schiller") +
  labs(x = "Age", y = "Count") +
  ggtitle("Histogram of Age by Schiller") +
  theme_minimal()
```



Number.of.sexual.partners Number of sexual partners. Values ranged from 1 to 28. We computed descriptive statistics (M = 2.51; SD = 1.64; skewness = 5.90; kurtosis = 81.68). On average, we could conclude that patients have had 2-3 sexual partners.

```r
cancer %>% ggplot(aes(x = Number.of.sexual.partners, fill = factor(Schiller))) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 20) +
  scale_fill_manual(values = c("#619CFF", "#FF595E"), name = "Schiller") +
  scale_x_continuous(limits = c(0, 10), breaks = seq(0, 10, 1)) +
  labs(x = "Number of Sexual Partners", y = "Count") +
  ggtitle("Histogram of Number of Sexual Partners by Schiller") +
  theme_minimal()
```



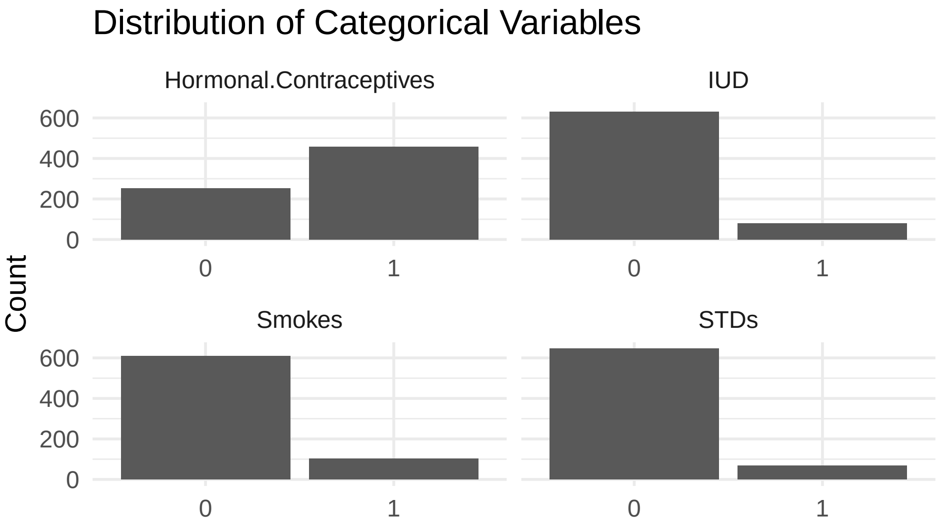
Hormonal.Contraceptives Whether use hormonal contraceptives or not. Taking values between 0 (negative) or 1 (positive). 254 out of 712 observations (35.7%) were negative for abnormal cells, while the remaining 458 observations (64.3%) were positive.

IUD Whether use the intrauterine device (hormonal control method) or not. Taking values between 0 (negative) or 1 (positive). 631 out of 712 observations (88.6%) not use intrauterine device, while the remaining 68 observations (11.4%) use intrauterine device.

Smokes Whether smokes or not. Taking values between 0 (non-smoke) or 1 (smoke). 609 out of 712 observations (85.5%) were non-smokers, while the remaining 103 observations (14.5%) were smokers.

STDs Whether have sexually transmitted diseases or not. Taking values between 0 (negative) or 1 (positive). 645 out of 712 observations (90.6%) were negative for abnormal cells, while the remaining 67 observations (9.4%) were positive.

```
#distribution of the categorical variables
cancer %>% select(Hormonal.Contraceptives, IUD, Smokes, STDs) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = factor(value))) +
  geom_bar() +
  facet_wrap(~ variable, scales = "free_x", ncol = 2) +
  labs(title = "Distribution of Categorical Variables") +
  xlab("") +
  ylab("Count") +
  theme_minimal()
```



We can observe that: - About 64% of the patients have used hormonal contraceptives. - About 11% of the patients have used IUD. - About 15% of the patients smoke. - About 9.4% of the patients have a history of sexually transmitted diseases.

Correlation We can examine the correlation among the key study variables according to following Table 2.

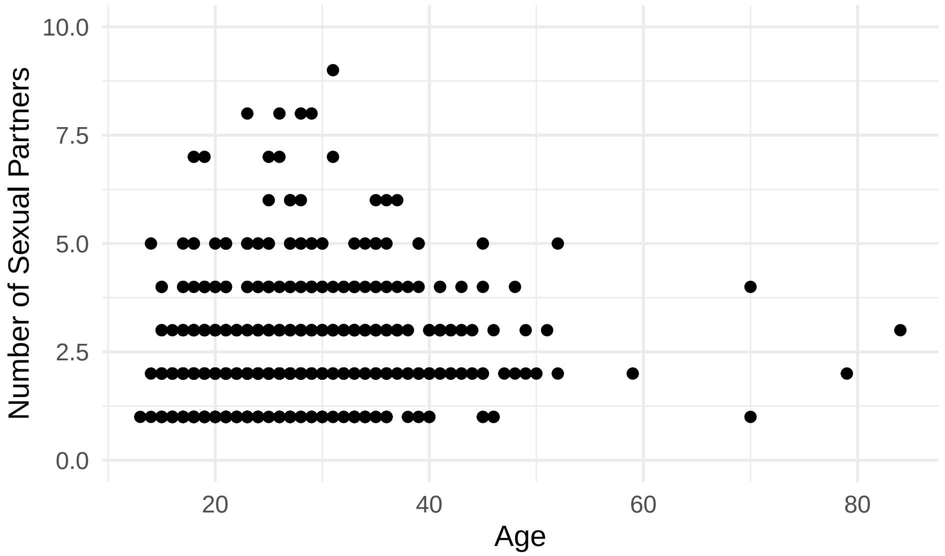
```
# Table 2: correlation matrix
cancer_table2 <- cancer %>% rename("Num.Sex.Par"=Number.of.sexual.partners) %>%
  rename("HC"=Hormonal.Contraceptives)
knitr::kable(corstarsl(cancer_table2))
```

	Age	Num.Sex.Par	HC	IUD	Smokes	STDs
Age						
Num.Sex.Par	0.09*					
HC	0.07	0.02				

	Age	Num.Sex.Par	HC	IUD	Smokes	STDs
IUD	0.28***	0.03	0.03			
Smokes	0.05	0.25***	0.01	-0.06		
STDs	0.01	0.02	-0.02	0.05	0.13***	
Schiller	0.09*	-0.01	0.00	0.08*	0.06	0.11**

All the correlation values are very low , so our model would not have the problem of multicollinearity.

```r
# removing an outlier (28)
cancer %>% ggplot(aes(Age, Number.of.sexual.partners)) +
  geom_point() +
  scale_y_continuous(limits = c(0, 10)) +
  ylab("Number of Sexual Partners") +
  theme_minimal()
```



We can observe that the number of sexual partners reach the peak in the 20-30 age group and then gradually decreases with age.

Statistical Model

Y_i is a discrete variable which can only take two values, 0 or 1. Thus, the Bernoulli probability model is the best candidate for the data. Letting π_i denote the probability of cervical cancer for a given individual i:

$$Y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$

with expected value

$$E(Y_i|\pi_i) = \pi_i. \tag{2}$$

To complete the structure of this Bernoulli data model, we must specify how the expected value of cervical cancer π_i depends upon predictor X_{ij}. To this end, the logistic regression model is part of the broader class of generalized linear models. Thus, we can identify an appropriate link function of π_i, $g(.)$, that is linearly related to $X_{i1}, ..., X_{i6}$:

$$g(\pi_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_6 X_{i6} \tag{3}$$

Let π_i and $odds_i = \pi_i/(1-\pi_i)$ denote the probability and corresponding odds of cervical cancer, $log(\pi_i/(1-\pi_i))$ is the only option that spans the entire real line. Thus, the most reasonable option is to assume that π_i depends upon predictor X_{ij} through the logit link function $g(\pi_i) = \log(\pi_i/(1-\pi_i))$:

$$Y_i|\beta_0, \beta_1, ..., \beta_6 \overset{ind}{\sim} \text{Bern}(\pi_i) \quad \text{with} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_6 X_{i6}. \tag{4}$$

That is, we assume that the log(odds of cervical cancer) is linearly related to other variables.

Prior Predictive Simulation

To complete the Bayesian logistic regression model of Y, we must put prior models on our regression parameters. As usual, since these parameters can take any value in the real line, Normal priors are appropriate for both. We'll also assume independence among the priors and express our prior understanding of the model baseline β_o through the centered intercept β_{0c}:

$$\begin{array}{lrl} \text{data:} & Y_i|\beta_0, \beta_1, ..., \beta_6 & \overset{ind}{\sim} \text{Bern}(\pi_i) \quad \text{with} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_6 X_{i6} \\ \text{priors:} & \beta_{0c} & \sim N\left(-3, 0.3^2\right) \\ & \beta_1 & \sim N\left(0.5, 0.3^2\right) \\ & ... & \\ & \beta_6 & \sim N\left(0.5, 0.3^2\right) \end{array} \tag{5}$$

Consider our prior tunings, based on the what we find on a `similar project` and `research paper`. Starting with the centered intercept β_{0c}, our prior belief is that the proportion of women in the population with cervical cancer is roughly at 0.05, i.e. $\pi \approx 0.05$. Thus, we set the prior mean for β_{0c} on the log(odds) scale to -3 in the Bayesian model:

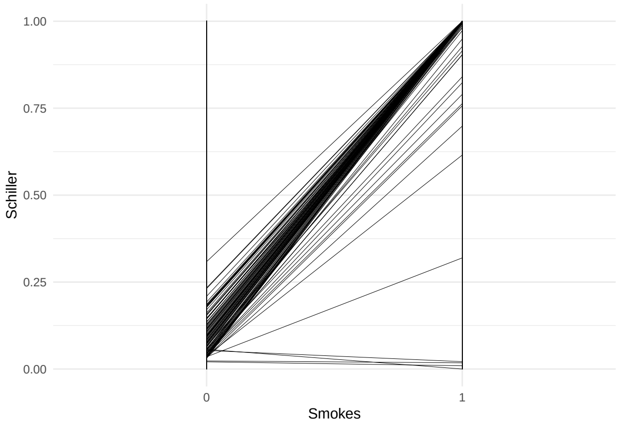
$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{0.05}{1-0.05}\right) \approx -3. \tag{6}$$

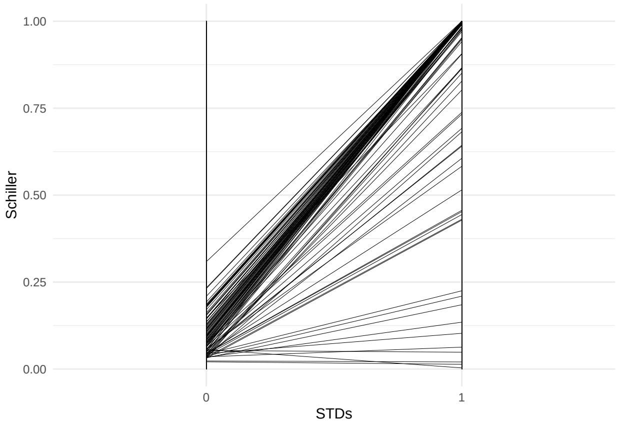
According to the research, the expected log(odds) of the coefficient ranges from -0.1 to 1. Thus we set our prior to a normal distribution with mean 0.5 and a standard error of 0.3.

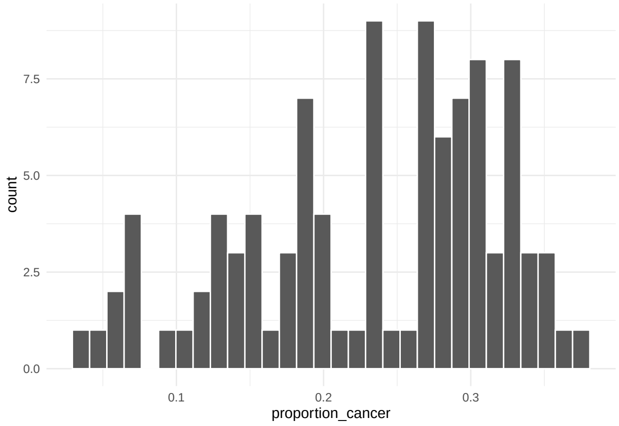
Next, we will use informative priors for the coefficients of other variables in the dataset. Then we simulated data under a variety of prior models.

```
cancer_model_prior <- stan_glm(Schiller~Age
                               +Number.of.sexual.partners
                               +Hormonal.Contraceptives
                               +IUD+Smokes+STDs,
                               data = cancer, family = binomial,
                               prior_intercept = normal(-3, 0.3),
                               prior = normal(0.5, 0.3),
                               chains = 4, iter = 5000*2, seed = 1,
                               prior_PD = TRUE)
```

```r
set.seed(11)

# plot 100 prior models, look at the prediction on the probability scale
cancer_epred <- cancer %>%
  add_epred_draws(cancer_model_prior, ndraws = 100)

cancer_epred %>%
  ggplot(aes(y = Schiller, x = Smokes)) +
  geom_line(aes(y = .epred, group = .draw), size = 0.1) +
  theme_minimal()

cancer_epred %>%
  ggplot(aes(y = Schiller, x = STDs)) +
  geom_line(aes(y = .epred, group = .draw), size = 0.1) +
  theme_minimal()

# plot the observed proportion of positive results in 100 prior datasets
cancer %>%
  add_predicted_draws(cancer_model_prior, n = 100) %>%
  group_by(.draw) %>%
  summarise(proportion_cancer = mean(.prediction == 1)) %>%
  ggplot(aes(x = proportion_cancer)) +
  geom_histogram(color = "white") +
  theme_minimal()
```







We plot 100 of these prior plausible relationships as shown above. These adequately reflect our prior understanding that the individuals who smoke or have sexually transmitted diseases have higher probability of getting positive Schiller results than the individuals who do not smoke or have diseases, as well as our prior uncertainty around the rate of this increase. The histogram of the 100 predicted positive result proportions from our 100 prior simulated datasets is displayed. We can observe that the prior predictions tend to be centered around low values. It indicates that our prior tuning is reasonable since we believe that the ratio of getting positive Schiller results should be low. Furthermore, the percent of getting positive results ranged from as low as roughly 0.03 in one dataset to as high as roughly 0.37 in another. This does adequately match our prior understanding and uncertainty about getting positive Schiller results.

Modeling Fitting and Posterior Predictive check

```
# posterior simulation
cancer_model <- stan_glm(Schiller~Age+Number.of.sexual.partners
                         +Hormonal.Contraceptives+IUD+Smokes+STDs,
                         data = cancer, family = binomial,
                         prior_intercept = normal(-3, 0.3),
                         prior = normal(0.5, 0.3),
                         chains = 4, iter = 5000*2, seed = 1)
```

```
c(auc5,mean(auc2))
```

```
## [1] 0.58100 0.56557
```

Comparing the AUC of glm model and stan_glm model, we find that stan_glm model performs better than glm model since Bayesian model is more stable and can handle model uncertainty.Thus , we use stan_glm model to continue our prediction.

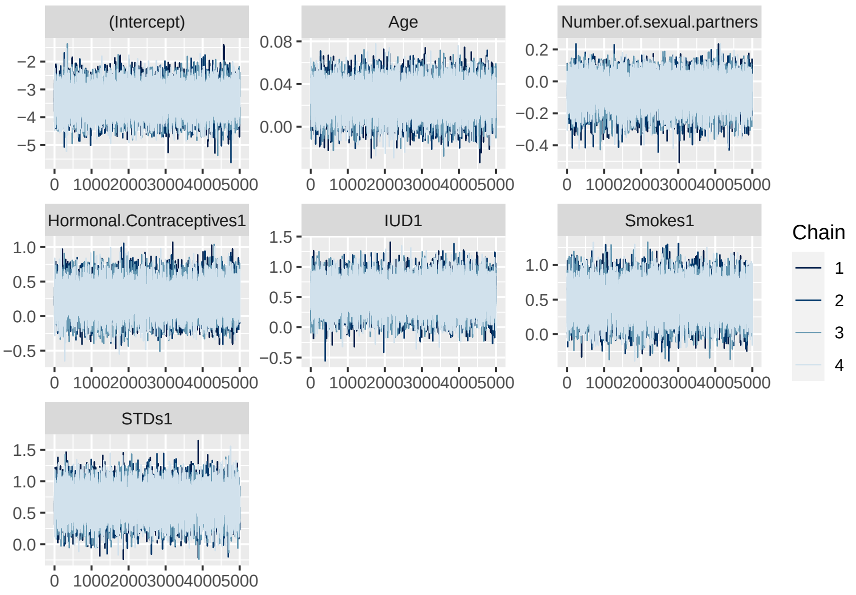
```
# posterior summaries on the log(odds) scale
posterior_interval(cancer_model, prob = 0.80)
```

```
##                                    10%         90%
## (Intercept)               -3.989118876 -2.76555814
## Age                        0.008532356  0.04370090
## Number.of.sexual.partners -0.182997063  0.04757213
## Hormonal.Contraceptives1  -0.011511939  0.52648663
## IUD1                       0.225448972  0.82164372
## Smokes1                    0.193587939  0.78940364
## STDs1                      0.368051963  0.96917364
```

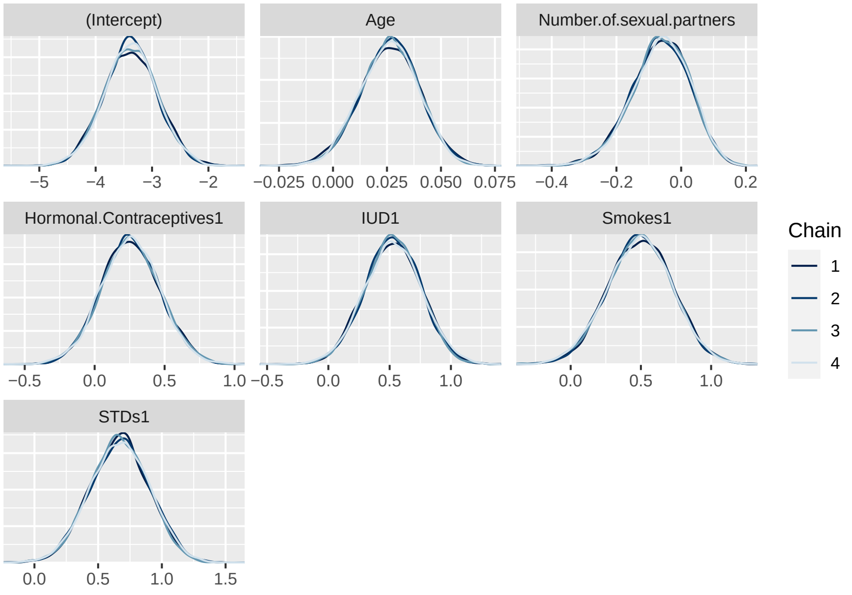
```
# posterior summaries on the odds scale
exp(posterior_interval(cancer_model, prob = 0.80))
```

```
##                                   10%        90%
## (Intercept)               0.01851602 0.06294096
## Age                       1.00856886 1.04466984
## Number.of.sexual.partners 0.83277060 1.04872184
## Hormonal.Contraceptives1  0.98855407 1.69297380
## IUD1                      1.25288510 2.27423498
## Smokes1                   1.21359610 2.20208279
## STDs1                     1.44491712 2.63576547
```

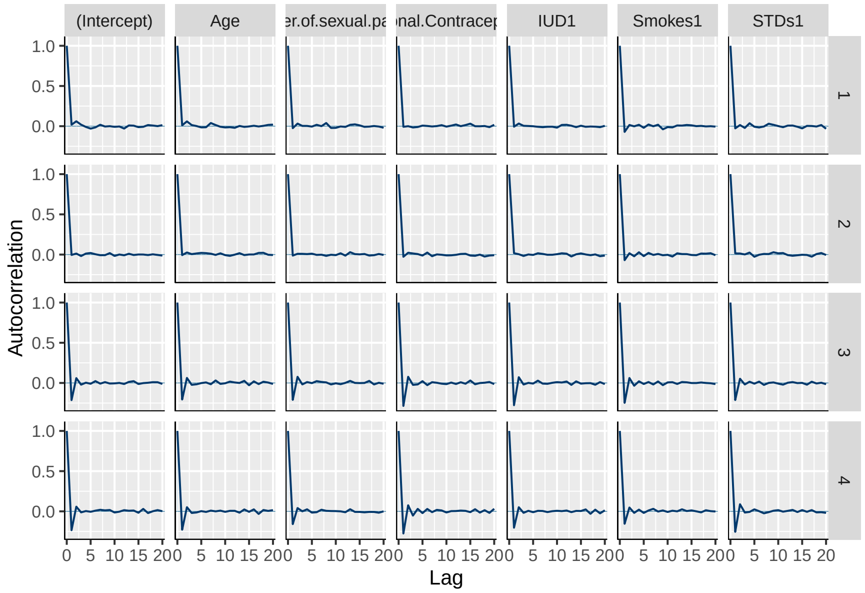
```
# check diagnostic plots for the stability of our simulation
# MCMC trace, density, & autocorrelation plots
mcmc_trace(cancer_model)
```



```
mcmc_dens_overlay(cancer_model)
```



```
mcmc_acf(cancer_model)
```



```
neff_ratio(cancer_model)
```

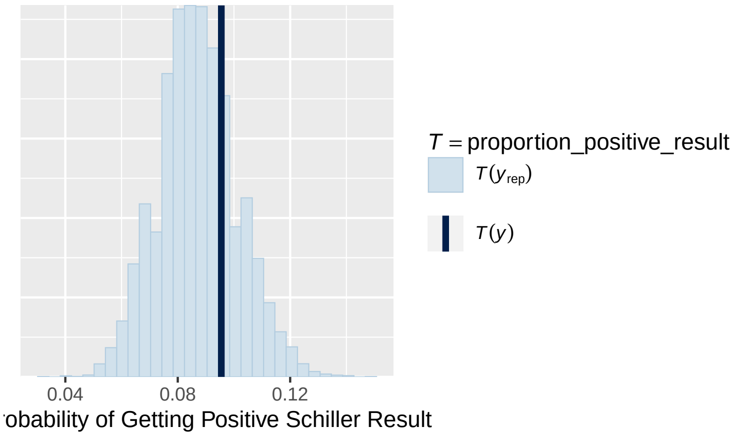
```
##                (Intercept)                         Age Number.of.sexual.partners
##                    1.15485                     1.13760                   1.09835
##   Hormonal.Contraceptives1                        IUD1                   Smokes1
##                    1.33080                     1.21915                   1.29000
##                      STDs1
##                    1.18280
```

```
rhat(cancer_model)
```

```
##                (Intercept)                         Age Number.of.sexual.partners
##                  0.9999097                   0.9999300                 0.9999499
##   Hormonal.Contraceptives1                        IUD1                   Smokes1
##                  0.9999137                   0.9998579                 0.9999558
##                      STDs1
##                  1.0001231
```

As shown above, the randomness in the trace plots, the agreement in the density plots of the four parallel chains, and Rhat value of effectively 1 suggest that our simulation is extremely stable. Further, the effective sample size ratio is satisfyingly high, indicating that our dependent chains are behaving “enough” like an independent sample so that the sampler is exploring the posterior distribution efficiently. It can also be observed from each autocorrelation plot of the four parallel chains that shows that autocorrelation is large at short lags, but then goes to zero pretty quickly.

```
# predicted probability
proportion_positive_result <- function(x){mean(x == 1)}
pp_check(cancer_model, nreps = 100,
         plotfun = "stat", stat = "proportion_positive_result") +
  xlab("Probability of Getting Positive Schiller Result")
```



```
# observed probability
sum(cancer$Schiller==1)/nrow(cancer)
```

```
## [1] 0.09550562
```

Based on the plot, we can observe that most of our posterior simulated datasets predict roughly 0.095 of positive results for Schiller test, close to the observed Schiller results in the cancer data in which the observed positive probability is 0.096, but some predict positive results as few as 0.04 or as many as 0.16. This brings us to question how accurate are our posterior classifications?

```
# check true positive rate or sensitivity of our Bayesian model
classification_summary(model = cancer_model, data = cancer, cutoff = 0.5)
```

```
## $confusion_matrix
##  y   0 1
##  0 644 0
##  1  68 0
##
## $accuracy_rates
##
## sensitivity      0.0000000
## specificity      1.0000000
## overall_accuracy 0.9044944
```

As shown, the model correctly classified 644 of the 712 (644+68) total test cases, and the overall classification accuracy rate is 90.4% (644 / 712). At face value, this seems pretty good! But look closer. Our model is much better at anticipating when it won't get positive result than when it get positive result. Among the 644 negative results, we correctly classify 644, or 100%. This figure is referred to as the true negative rate or specificity of our Bayesian model. In stark contrast, among the 68 positive results, we correctly classify 0, or 0%. This figure is referred to as the true positive rate or sensitivity of our Bayesian model. From this, we can see that our model is bad at predicating positive results. But we can do better! Sensitivity is important in our analysis. Since in the real world, it is terrible that we can not designate an individual with cervical cancer as positive. To better suit the goals of our analysis, we can increase our model's sensitivity by decreasing the cut-off from 0.5 to 0.1. That is, we'll classify a test case as positive if there's even a 10% chance of getting positive result.

```
set.seed(11)
classification_summary(model = cancer_model, data = cancer, cutoff = 0.1)
```

```
## $confusion_matrix
##  y   0   1
##  0 496 148
##  1  36  32
## 
## $accuracy_rates
## 
## sensitivity      0.4705882
## specificity      0.7701863
## overall_accuracy 0.7415730
```

As shown, after adjusting the cut-off, the sensitivity jumped from 0% to 47.1% (32 of 68). It is really an improvement in detecting potential cancer. Yet this improvement is not without consequences. In lowering the cut-off, we have a higher mistake in predicting negative result when a individual actually does not have a cervical cancer. As a result, the true negative rate dropped from 90.4% to 77.0% (496 of 644). However, if there is a trade-off, I think that sensitivity is more important in our case because identify someone who has cervical cancer as early as possible is crucial. Missing the timely and proper treatment will definitely lead to the patient's worsening conditions and even death.

Model Selection

Next, we fit more Bayesian models by using different number of predictors or having an interaction term to see which model has the best out-of-sample predictive performance.

```
cancer_model2 <- stan_glm(Schiller~Age+Number.of.sexual.partners
                         +Hormonal.Contraceptives+IUD+Smokes,
                         data = cancer, family = binomial,
                         prior_intercept = normal(-3, 0.3),
                         prior = normal(0.5, 0.3),
                         chains = 4, iter = 5000*2, seed = 1)

cancer_model3 <- stan_glm(Schiller~Age+Number.of.sexual.partners
                         +Hormonal.Contraceptives+IUD,
                         data = cancer, family = binomial,
                         prior_intercept = normal(-3, 0.3),
                         prior = normal(0.5, 0.3),
                         chains = 4, iter = 5000*2, seed = 1)

cancer_model4 <- stan_glm(Schiller~Age*Number.of.sexual.partners
                         +Hormonal.Contraceptives+IUD+Smokes,
                         data = cancer, family = binomial,
                         prior_intercept = normal(-3, 0.3),
                         prior = normal(0.5, 0.3),
                         chains = 4, iter = 5000*2, seed = 1)

cancer_model5 <- stan_glm(Schiller~Age*Smokes+
                            Number.of.sexual.partners
                          +Hormonal.Contraceptives+IUD,
                         data = cancer, family = binomial,
                         prior_intercept = normal(-3, 0.3),
                         prior = normal(0.5, 0.3),
                         chains = 4, iter = 5000*2, seed = 1)
```

```
set.seed(11)
# compare (Root) Mean Square Errors
yrep1 <- posterior_predict(cancer_model)
yrep2 <- posterior_predict(cancer_model2)
yrep3 <- posterior_predict(cancer_model3)
yrep4 <- posterior_predict(cancer_model4)
yrep5 <- posterior_predict(cancer_model5)
mse1 <- sqrt(mean((colMeans(yrep1) - (as.numeric(cancer$Schiller)-1))^2))
mse2 <- sqrt(mean((colMeans(yrep2) - (as.numeric(cancer$Schiller)-1))^2))
mse3 <- sqrt(mean((colMeans(yrep3) - (as.numeric(cancer$Schiller)-1))^2))
mse4 <- sqrt(mean((colMeans(yrep4) - (as.numeric(cancer$Schiller)-1))^2))
mse5 <- sqrt(mean((colMeans(yrep5) - (as.numeric(cancer$Schiller)-1))^2))
df2 <- data.frame(mse1, mse2, mse3, mse4, mse5)
rownames(df2) <- "Mean Square Error"
colnames(df2) <- c("cancer_model", "cancer_model2", "cancer_model3",
                   "cancer_model4", "cancer_model5")
knitr::kable(round(df2,3))
```

	cancer_model	cancer_model2	cancer_model3	cancer_model4	cancer_model5
Mean Square Error	0.29	0.291	0.292	0.292	0.291

As shown above, through checking accuracy using root mean square error for all models, we can observe that they all have similar values. Thus we need to use the other method to compare the model performances.

```
loo1 <- loo(cancer_model)
loo2 <- loo(cancer_model2)
loo3 <- loo(cancer_model3)
loo4 <- loo(cancer_model4)
loo5 <- loo(cancer_model5)
loo_compare(loo1, loo2, loo3, loo4, loo5)
```

```
##               elpd_diff se_diff
## cancer_model   0.0       0.0
## cancer_model2 -2.2       2.0
## cancer_model5 -2.3       2.0
## cancer_model3 -3.1       2.8
## cancer_model4 -3.5       2.1
```

As known, the larger the expected logged posterior predictive pdf across all possible new data points (ELPD), the more accurate the posterior predictions. From the loo_compare() which lists the models in order from the highest ELPD to the lowest, we can observe that as the number of predictors decreases, the ELPD also decreases. The cancer model with all initially selected variables and without any interaction term is the best since it has a largest ELPD among all models.

Discussion

Limitations

A small sample size and low percentage of cases with positive test results would result in a less accurate regression model with lower sensitivity, so it can not accurately identify patients from others.

Additionally, our model includes too many binary variables that can affect the regression results, as our model assumes a linear relationship between the dependent variable and the independent variables. So linear regression models usually require continuous variables as inputs.

Advantages

We use Bayesian models to identify which cancer diagnostic tests will be the most accurate in predicting an individual's risk of developing cervical cancer.

Bayesian methods perform better to handle model uncertainty. Moreover, Stan_glm is highly flexible and extensible so it can be extended and applied to other cancer predictions, as well as other disease diagnosis.

How the model can be improved

Use the data which has a higher proportion of positive test results in the sample, thus trying to get a higher sensitivity of the prediction result.

Integrate the four cancer indicators (Hinselmann, Schiller, Citology, and Biopsy) into a multilevel variable by adding them together, so that the reliability of predictions would be improved compared to the model depending on a binary outcome.

Include more numerical variables, such as the number of cigarettes smoked per day in the predictors, It is more clear to show the relationship between the probability of cancer and smoking.Also, we can select predictors by stepwise regression.