Determinants of Access to Hysterectomy

Abstract:
This project used data from the 2017-March 2020 Pre-Pandemic data files of the National Health and Nutrition Examination Survey (NHANES) to address which predictors contribute most to the likelihood of having received a hysterectomy. After cleaning, we implemented binary logistic regression to fit our first-order full model on the remaining 11 predictors. We used stepwise regression with AIC and BIC as criterion to obtain the potential best first-order and second-order models. Each model was compared using k-fold cross validation performance metrics, and a likelihood ratio test was conducted to confirm the selection of the stepwise regression using AIC as criterion model. We found that number of pregnancies, age, and having used female hormones contributed significantly to the log odds of having received a hysterectomy. The predictors marital status, level of education and age at last live birth contributed significantly to the log odds of not having received a hysterectomy.

1. Background and Significance

The goal of this project is to find the best fitting final model to predict whether or not someone has had a hysterectomy and identify what predictors most significantly contribute to the likelihood of having received a hysterectomy. Hysterectomies are a very common procedure that involve the removal of the uterus/cervix and is typically performed as treatment for medical conditions including: cancer, endometriosis, and gender reaffirming surgery. Access to these kinds of procedures can be varied and there is a long history of obstetric racism and forced sterilization tied to it. Understanding the role of power in the data is an important feature of this analysis and factored largely into the selection of relevant demographic based predictors.

2. Data

2.1 Data Description

The dataset was sourced from the 2017 - March 2020 Pre-Pandemic data files[1] from the National Health and Nutrition Examination Survey (NHANES) with a focus on the reproductive health and demographics data files. The original dataset consists of relevant variables from the NHANES files containing 5,314 observations and 22 attributes. See the chart below for a description of the variables present in the dataset.

Variable	Description
ID	A unique ID assigned to each participant.
hysterectomy	"Had a Hysterectomy?"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
regularPeriods	"Had regular periods in the past twelve months?"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
periodReason	"Reason not having regular periods?"; Categorical variable with eight levels (Pregnancy, Breast Feeding, Hysterectomy, Menopause/Change of life, Other, Refused, Don't Know, and Missing).
doctor	"Seen a Doctor bc unable to become pregnant?"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
beenPregnant	"Ever been pregnant"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
currentlyPregnant	"Are you pregnant now"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
numPregnant	"How many times have been pregnant"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
numDeliveries	"Total number of deliveries"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
pelvicInfection	"Ever been treated for pelvic infection?"; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
ageFirstLiveBirth	"Age at first live birth?" Categorical variable with six levels (17 years or younger, ages ranging from 18-44 years , 45 years or older, Refused, Don't Know, and Missing).
ageLastLiveBirth	"Age at last live birth?" Categorical variable with six levels (17 years or younger, ages ranging from 18-44 years , 45 years or older, Refused, Don't Know, and Missing).
useFemaleHormones	"Ever use female hormones? Do not include birth control methods or use for infertility."; Categorical variable with five levels (Yes, No, Refused, Don't Know, and Missing).
age	"Age in years at screening."; Numerical variable ranging from 20 to 80.
race	"Race/ethinicity"; Categorical variable with seven levels (Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other Race - Including Multi-Racial, and Missing).
birthCountry	"County of birth"; Categorical variable with five levels (Born in the United States, Others, Refused, Don't Know, and Missing).
timeInUS	"Length of time in United States"; Categorical variable with seven levels (Less than 5 years, 5 -14 years, 15 years - 29 years, 30 years or more, Refused, Don't Know, and Missing).
education	"Education level"; Categorical variable with eight levels (Less than 9th grade, 9-11th grade/No diploma, High school graduate/GED or equivalent, Some college or AA degree, College graduate or above, Refused, Don't Know, and Missing).
maritalStatus	"Marital Status"; Categorical variable with six levels (Married/Living with Partner, Widowed/Divorced/Separated, Never Married, Refused, Don't Know, and Missing).
interviewLanguage	"Language of interview?"; Categorical variable with three levels (English, Spanish, Missing).
interpreter	"Interpreter used in interview?"; Categorical variable with three levels (Yes, No, and Missing).
incomeToPovertyRatio	"Ratio of family income to poverty"; Numerical variable ranging from 0 to 4.98.

2.2 Data Trimming

The data cleaning process first began with examining the levels of missingness present in the response and predictor variables. To conduct our analysis all missingness from our response variable was eliminated. Handling missingness in predictor variables was assessed based on the following methodology: if missingness can be explained by another predictor, a new level "Does not apply" was created. If missingness persisted at minimal percentages, observations were removed. Lastly, if the predictor presented missingness at levels too high to omit observations and ethically too concerning to impute, they were removed. After handling missingness, the following variables were removed from the dataset: ID, regularPeriod, periodReason, currentlyPregnant, and incomeToPoveryRatio. We considered the presence of multicollinearity in the data using Cramer's V to address association between two nominal categorical variables. From the matrix constructed, a threshold of 0.5 was used to identify the predictors with too significant of an association to consider for model fitting. The predictors removed were: doctor, beenPregnant, numDeliveries, birthCountry, interviewLanguage, and interpreter. The final dataset for model fitting contains 3,791 observations and 11 total attributes.

[1] "Nhanes 2017-March 2020 Pre-Pandemic Questionnaire Data." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020.

3. Methodology

3.1 Model I: First Order Binary Logistic Regression

	Accuracy	Specificity	Sensitivity	Precision
Saturated Model	0.7335795	0.7040054	0.8366864	0.4477517
Stepwise AIC	0.7354260	0.7070604	0.8343195	0.4496173
Stepwise BIC	0.7219731	0.7016293	0.7928994	0.4325371

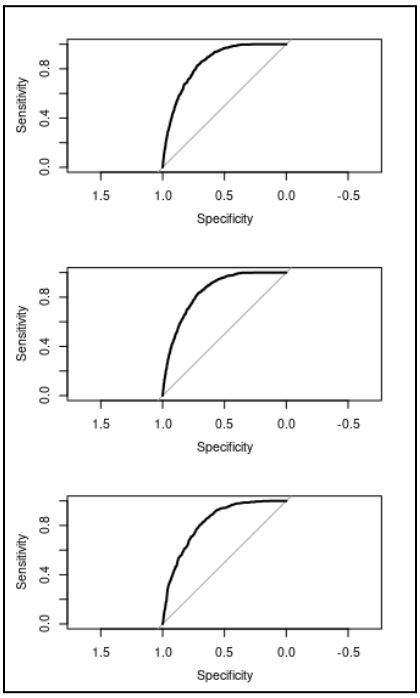


Figure 1 (Above). Performance metrics for Three Binary Logistic Regression Models.
Figure 2 (R). ROC Curves for the Three Binary Logistic Regression Models. (Top: Full First-Order Model; Middle: AIC Model; Bottom: BIC Model)

The full first-order model was fit using a binary logistic regression evaluating the relationship between our response of interest: Hysterectomy and the ten other predictors. Using stepwise regression, we created three different models: (1) the saturated model containing all predictors. (2) The refined first-order model using stepwise regression with AIC criteria, producing a model excluding three predictors: pelvicInfection, ageAtFirstFiveBirth, and timeInUS. (3) The refined first-order model using stepwise regression with BIC criteria, producing a model of three predictors: usedFemaleHormones, age, and education. We performed K-Fold cross validation with 10 folds, on our three models and computed their performance metrics with a threshold of 0.2 to account for the unequal distribution of observations with and without hysterectomies in the data (Figure 1). The performance metrics reveal all our models perform comparably well with the exception of low precision indicating these models may be over-predicting hysterectomy recipients in the data. Ultimately, we are most concerned with sensitivity and chose to select the stepwise regression model with AIC criteria as it performed best in all metrics, is only marginally less sensitive than the full model and more parsimonious. To confirm this selection we conducted a likelihood ratio test comparing the first-order model produced by stepwise regression with AIC criteria and the full first order model. The resulting p value, p=0.7981, does not provide sufficient evidence to reject the null hypothesis. We conclude the Stepwise regression model using AIC criteria sufficiently fits the data. Conclusions were confirmed by the ROC plots and AUC scores (Figure 2). The highest AUC value corresponds to the full first-order model (AUC=0.8418), with comparably high AUC scores in the AIC and BIC models (0.8399, and 0.8245 respectively). Thus, we move forward with the stepwise regression with AIC criteria as our final first-order model.

3.2 Model II: Second Order Stepwise Regression BIC Model

	Performance metric
accuracy	0.7219731
specificity	0.7016293
sensitivity	0.7928994
precision	0.4325371

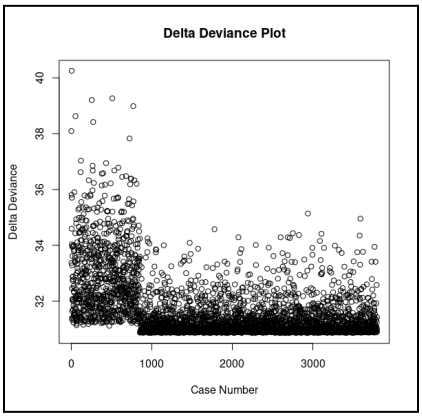
Figure 3 (L) Performance metrics of the interactive model built from the best BIC model.

Before finalizing our model we want to consider potential interaction terms. Given the relatively small number of predators in the best (BIC) model, it's reasonable to explore possible two-way interaction terms. Using stepwise regression with AIC and BIC as criterion for selecting the best second order model based on our best BIC first order model produced the same result. The resulting best second-order model included the following predictors: useFemaleHormones, age, education, and the interaction between useFemaleHormones and age. We then conducted k fold cross validation and calculated the performance metrics with a threshold of 0.2 for comparison to our first-order models. The

interaction model performed the same as the BIC model. Before moving forward with our final model selection we performed another likelihood ratio test to see if the addition of the interaction terms sufficiently fits the data better than the reduced model (the first order BIC model). While the likelihood ratio test indicates that the interaction model fits the data significantly better, with a p-value of 7.423e-07, it is likely to overfit the current data set as it is the larger model. Our first order AIC model still performs best in all performance metrics, thus we decide to go forward using the first order AIC model.

3.2 Model Diagnostics

Figure 4 (R). Delta Deviance Plot.



Lastly, we looked at model diagnostic plots for the final model to identify potential outliers. We could not find any characteristics associated with these observations to justify their outlier status. We constructed a delta deviance plot to gauge for influential outliers in our data (Figure 4). Observations falling at a delta deviance greater than 37 were removed and the final first order model was refitted on the new data.

4. Results

The individual wald tests of the final first order AIC model indicate, given the presence of all other predictors in the model, the predictors with the strongest positive relationship with hysterectomy were Num Pregnancies, Age, and Used Female Hormones. Given the presence of all other predictors in the model, the significant predictors with a negative relationship with Hysterectomy were Age at Last Live Birth (generally older ages), Marital status (never married), and Education (college graduate or above).

5. Considerations and Discussion

Some sources of bias must be taken into consideration, including the removal of observations and predictors in the data cleaning process, and the visual observation of influential outliers from the delta deviance plot. Additionally, some data was lost in the processes of re-leveling categories with too few observations at a particular level. For instance, the number of pregnancies at levels 9, 10, and 11 were low, and were thus re-coded as 8. Our analysis hoped to address the factors that contribute to someone having had a hysterectomy, as well as to define the best model for prediction of hysterectomy recipients. Our final model had a high sensitivity rate, indicating that the model can accurately predict who has had a hysterectomy. The model also indicates that Num Pregnancies, Age, Used Female Hormones, Age at Last Live Birth, Marital status, and Education are the factors that contribute most significantly to having received a hysterectomy. Age may have a positive relationship with hysterectomy given that typically a hysterectomy is a difficult procedure to procure below the age of 35. The relationship between usedFemaleHormones and hysterectomy is justified as hysterectomy recipients must also take some form of hormone replacement after the procedure. The significance of num pregnancies can be muddled in that people can choose to have a hysterectomy for a series of reasons. So it follows the logic that there are different levels of number of pregnancies that someone would still elect to have a hysterectomy.