# STAT 535: Chapter 13: Bayesian Logistic Regression Models

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2024

# Regression for Binary Data

- We now consider a **regression model** in which a response variable $Y$ takes on exactly two values (Pass/Fail; Survive/Die; Win/Loss, etc.), which we generally code as 0 or 1.
- The Normal or Poisson models clearly are not appropriate for modeling a response variable of this type.
- When the response variable $Y$ can only take values 0 or 1, its expected value $E(Y)$ is the same as $P(Y = 1)$.
- The model we will use will relate this $E(Y)$ to a predictor $X$ or set of predictors $X_1, X_2, \ldots, X_p$.

# Review: Odds and Probability

- Recall that if an event has probability $\pi$, then the **odds** of that event are $\pi/(1-\pi)$.
- We know the probability ranges from 0 to 1, so the odds can take on values between 0 and $\infty$.
- The odds of an event are less than 1 if and only if the event's probability $\pi < 0.5$.
- The odds of an event are equal to 1 if and only if the event's probability $\pi = 0.5$.
- The odds of an event are greater than 1 if and only if the event's probability $\pi > 0.5$.

# Real Data Example: Logistic Regression Model

- Consider a sample of senior citizens, on whom two variables, a binary $Y$ and an (approximately) continuous $X$, are measured.
- The response variable measures whether the individual is judged to be senile: Define $Y = 0$ if individual has no senility, define $Y = 1$ if individual has senility present.
- Let $X =$ the individual's score on a subset of Wechler Adult Intelligence Scale (WAIS) exam.

# Real Data Example: Logistic Regression Model

- Recall that when an individual's response $Y_i$ is binary, $E(Y_i) = P(Y_i = 1)$.
- We will model $E(Y_i) = \pi_i$ as a function of $X_i$, the WAIS score for that individual.
- So the model for the mean response **given** the predictors is

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

- So the "linear predictor" part of the model equation, $\beta_0 + \beta_1 X_{i1}$, is related to the **log-odds** that $Y_i = 1$.
- We can also write this model equation in terms of the **odds** or in terms of the **probability** that $Y_i = 1$:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{i1}} \quad \text{and} \quad \pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}$$

# General Form of the Logistic Regression Model

- In a logistic regression model with several predictors, we have

$$\log(\text{odds}) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The interpretation of, say, $\beta_1$ is as follows:
- Let $\text{odds}_x$ be odds that $Y = 1$ when $X_1 = x$ and let $\text{odds}_{x+1}$ be odds that $Y = 1$ when $X_1 = x + 1$ (an addition of one unit for $X_1$).
- Controlling for (holding constant) the other predictors $X_2, \ldots, X_p$, then $\beta_1$ is the expected change in log-odds, and $e^{\beta_1}$ is the expected **multiplicative** change in odds that $Y = 1$ when $X_1$ is increased by one unit:

$$\beta_1 = \log(\text{odds}_{x+1}) - \log(\text{odds}_x) \quad \text{and} \quad e^{\beta_1} = \frac{\text{odds}_{x+1}}{\text{odds}_x}$$

# Priors in the Logistic Regression Model

- We will need to specify priors on $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.
- Typically we choose normal priors on these model coefficients.
- If we want to be objective, we could specify prior means of 0 for all the coefficients.
- We will have to use Metropolis-Hastings (either coding it ourselves or using `stan_glm` in `rstanarm`) to sample from the posterior and estimate the parameters.
- See example with "noninformative" priors in R for the WAIS senility data.

# Specifying Subjective Priors in the Logistic Regression Model

- The book gives a process for eliciting prior information that works well especially with the `stan_glm` function.
- In the `stan_glm` function, we place a prior on the "centered" intercept. Here's an example:
- We believe that a "typical" subject might have probability 0.2 to 0.6 of senility. So the log-odds of senility for such a person should be between $\log(0.2/0.8) = -1.4$ and $\log(0.6/0.4) = 0.4$.
- So set the prior mean for the **CENTERED** $\beta_0$ (different from the $\beta_0$ in model) to be halfway between those, at -0.5.
- Set prior standard deviation to be half of the distance between -0.5 and 0.4, so prior standard deviation $= 0.45$.

# More on Specifying Subjective Priors in the Logistic Regression Model

- Here's an example of specifying the prior mean and standard deviation for $\beta_1$:
- We believe that for a one-unit increase in WAIS score, the odds of senility might be anywhere from half as large to the same, i.e., between 0.5 and 1.
- So $\beta_1$ might be between $\log(0.5) = -0.69$ and $\log(1) = 0$.
- So make prior mean on $\beta_1$ to be -0.35 and prior standard deviation around 0.175.

# Fitting the Logistic Regression Model

- We can specify the priors and use the `stan_glm` function in the `rstanarm package` to do the Metropolis-Hastings automatically, as usual.
- We would still want to do our usual MCMC diagnostics and (if necessary) remedial actions.
- The `tidy` function or `summary` function will again print summaries of the posteriors for the model coefficients.
- See R examples for the fitting of the model.

# Interpretations of Estimated Parameters

- The posterior estimate of $\beta_1$ is (around) -0.3 (it will change slightly depending on the exact type of priors chosen and even slightly based on the MCMC run).
- The odds of senility changes by a factor of $e^{-0.3} = 0.74$ (i.e., decreases by 26%) for each one-point increase in WAIS score.
- A 95% credible interval for $\beta_1$ is $(-0.498, -0.142)$, so there is high posterior probability that a higher WAIS score is associated with **lower** odds of senility.

# Using the Logistic Regression Model for Prediction

- One of the purposes of the logistic regression model is to **predict** the binary response value for a new observation.
- For example, if we have a new senior citizen with WAIS score of 10, we want to predict whether or not that person has senility.
- One approach: Plug $x = 10$ into the estimated logistic regression model, get $\widehat{E(Y|X = 10)}$, which is the estimated probability that this person is senile.
- If this estimated probability exceeds 0.5, predict $Y = 1$ for this individual; otherwise, predict $Y = 0$.
- Note that we could use a cutoff $c$ other than 0.5 if we wish.

# Defining a Classification Rule

- When predicting a binary response using our fitted logistic regression model, we are basically classifying an individual into either the $Y = 0$ or the $Y = 1$ group.
- We could use a **classification rule** as follows to do this:
- For a particular $x$ value (or set of $x_1, x_2, \ldots, x_p$ values), generate a large number of posterior predictions of $Y$.
- Let $p$ be the proportion of those posterior predictions that have $Y = 1$.
- For a chosen **classification cutoff value** $c \in [0, 1]$, we classify the individual with those specified predictor value(s) to the $Y = 1$ group if $p \geq c$; otherwise classify this individual to the $Y = 0$ group.

# Choice of Classification Cutoff Value

- The natural cutoff value to consider is $c = 0.5$, and this is what is usually used.
- But sometimes a different value makes sense, especially if the cost of one type of misclassification error is much greater than the cost of the other type of error.
- In an example in the book, predicting $Y = 1$ corresponds to predicting rain and $Y = 0$ corresponds to no rain.
- Is it worse to predict rain and carry an umbrella when no rain actually falls, or to predict no rain, forgo the umbrella, and get wet when rain actually falls?
- For this example, we might choose a smaller cutoff like $c = 0.25$, so that we will play it safe and predict rain more often.

# Assessing Model Quality

- We can again use the posterior predictive distribution to assess model quality.
- The `pp_check` function is a shortcut to generate many posterior simulated data sets. For each one, we calculate the count of $Y = 1$ values and plot these counts with a histogram.
- If the actual count of $Y = 1$ values from our observed data set falls in the middle of this distribution, it is a sign that the model fits well (see R example with WAIS data).

# Measuring Classification Accuracy

- When using the logistic regression model to classify binary observation into one class or the other, we want to be able to assess the accuracy of our classifications.
- A common way to do this is based on a **confusion matrix**.
- For a set of binary (0 or 1) observations, let $Y$ denote the actual binary value for an observation and let $\hat{Y}$ denote the predicted binary value based on whatever classification rule we've decided on.
- For each individual in our sample, $i = 1, \ldots, n$, we have the $Y_i$ value and we can calculate the $\hat{Y}_i$ value from our fitted logistic regression model.

# Confusion Matrix

- The **confusion matrix** is the $2 \times 2$ matrix with entries $a$, $b$, $c$, and $d$:

| | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
|---|---|---|
| $Y = 0$ | $a$ | $b$ |
| $Y = 1$ | $c$ | $d$ |

- The model's **overall accuracy** captures the proportion of all binary observations that are accurately classified:

$$\text{overall accuracy} = \frac{a+d}{a+b+c+d}.$$

- The **misclassification rate** is 1 minus the overall accuracy, or $\frac{b+c}{a+b+c+d}$.

# Sensitivity and Specificity

- The model's **sensitivity** (true positive rate) captures the proportion of $Y = 1$ observations that are accurately classified.
- The **specificity** (true negative rate) captures the proportion of $Y = 0$ observations that are accurately classified.

$$\text{sensitivity} = \frac{d}{c+d} \quad \text{and} \quad \text{specificity} = \frac{a}{a+b}.$$

# Aims of Sensitivity and Specificity

- Obviously we want both the sensitivity and specificity to be high, but sometimes it makes sense practically to care more about one than the other.
- If we are doing a medical test for a potentially deadly disease (e.g., breast cancer) and classifying subjects as sick or healthy, would we care more about having a high sensitivity or high specificity?
- A high sensitivity would reduce the chance of a true cancer doing undetected and thus a person with cancer going untreated.
- If our procedure has lower specificity and we misclassify some healthy people as sick, that may waste some time and money, but the consequences would not be deadly.

# Tuning the Classification Rule based on Sensitivity and Specificity

- Recall that the value of the cutoff $c$ determines our classification rule.
- We can try various values of $c$ and, for each value, check (in-sample) sensitivity and specificity based on the resulting confusion matrix.
- It may be better to use cross-validation measures of sensitivity and specificity to assess how well the model classifies **new** (i.e., out-of-sample) observations.
- There is a tradeoff: As we lower $c$, sensitivity increases, but specificity decreases. As we increase $c$, specificity increases, but sensitivity decreases.

# Bayesian Logistic Regression with Multiple Predictors

- The logistic regression model extends naturally to having several predictors, $X_1, X_2, \ldots, X_p$.
- Example from Book: Binary variable $Y$ is whether it rains tomorrow in Perth, Australia
- Predictors: $X_1$ = humidity at 9 a.m. today, $X_2$ = humidity at 3 p.m. today, $X_3$ = whether it rains today (binary).
- Model equation in terms of the mean response is:

$$\pi_i = E(Y_i|\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}{1+e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}}}$$

# Fitting of Bayesian Multiple Logistic Regression

- The (normal) priors on the $\beta$'s can be specified as usual.
- MCMC simulation from the posterior is done through direct Metropolis-Hastings or automatically via `stan_glm`.
- Estimated coefficents: $\hat{\beta}_1 = -0.007$, $\hat{\beta}_2 = 0.08$, $\hat{\beta}_3 = 1.15$ (values may change depending on priors and MCMC run).
- Inference about the $\beta$'s: The 95% credible interval for $\beta_1$ includes 0: We may not need "humidity at 9 a.m. today" as a predictor in the model, **given that** "humidity at 3 p.m. today" and "rain today" are predictors in the model.
- The predictors may be strongly associated with each other, which explains why we may not need all of them in the model.

# Model Selection in Bayesian Multiple Logistic Regression

- We can fit several models with different sets of predictors and use our usual model selection tools (CV accuracy, ELPD, BIC, etc.) to choose the "best" model.
- Rain example: Comparing the model with 3 predictors to a model with only $X_1$:
- The model with three predictors has a better CV accuracy, higher ELPD, and lower BIC, so the model with three predictors is preferred.
- However, a model that omits $X_1$ and includes $X_2$ and $X_3$ is slightly preferred over the 3-predictor model, based on these criteria.