Module 10: Logistic Regression

Rebecca C. Steorts

Agenda

We will explore a variable selecion model for Bayesian logistic regression using the data in **azdiabetes.dat**. This closesly follows the exercise 10.5 of the Hoff book.

Application to diabetes data set

Suppose we have data on health-related variables of a population of 532 women.

Our goal is to predict whether or not a patient has diabetes given the covariates below.

x_1 = number of pregnancies

x_2 = blood pressure

x_3 = body mass index

x_4 = diabetes perdigree

x_5 = age

Diabetes Data

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

```r
rm(list=ls())
azd_data = read.table("azdiabetes.dat", header = TRUE)
head(azd_data)
```

```
##   npreg glu bp skin  bmi   ped age diabetes
## 1     5  86 68   28 30.2 0.364  24       No
## 2     7 195 70   33 25.1 0.163  55      Yes
## 3     5  77 82   41 35.8 0.156  35       No
## 4     0 165 76   43 47.9 0.259  26       No
## 5     0 107 60   25 26.4 0.133  23       No
## 6     5  97 76   27 35.6 0.378  52      Yes
```

Diabetes Data (Continued)

```
diabetes <- azd_data$diabetes
data <- azd_data[-c(2,4,8)]
head(data)
```

```
##   npreg bp  bmi   ped age
## 1     5 68 30.2 0.364  24
## 2     7 70 25.1 0.163  55
## 3     5 82 35.8 0.156  35
## 4     0 76 47.9 0.259  26
## 5     0 60 26.4 0.133  23
## 6     5 76 35.6 0.378  52
```

Linear regression

Why would linear regression be inappropriate here?

```
fit.ols<-lm(diabetes~ data[,1] + data[,2] + data[,3] + data
```

```
## Warning in model.response(mf, "numeric"): using type = '
## factor response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningfu
```

```
summary(fit.ols)$coef
```

```
## Warning in Ops.factor(r, 2): '^' not meaningful for fact

##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.359189978         NA      NA       NA
## data[, 1]   0.033834780         NA      NA       NA
## data[, 2]   0.002129932         NA      NA       NA
## data[, 3]   0.017317124         NA      NA       NA
## data[, 4]   0.263748153         NA      NA       NA
```

Notation

- $X_{n \times p}$: regression features or covariates (design matrix)
- $\boldsymbol{x}_i$: ith row vector of the regression covariates
- $\boldsymbol{y}_{n \times 1}$: response variable (vector)
- $\beta_{p \times 1}$: vector of regression coefficients

Notation (continued)

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

- A column of x represents a particular covariate we might be interested in, such as age of a person.

- Denote x_i as the ith row vector of the $X_{n\times p}$ matrix.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Notation (continued)

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\boldsymbol{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

Recall that the model above is a linear model.

Logistic regression

- Logistic regression is a generalized linear model, where the response varialble is a binary value (0 or 1).
- That is the outcome Y_i takes either the value 1 or 0 depending on the application with probability p_i and $1 - p_i$.
- This is the probability that we model in relation to the covariates in our data set.

Logistic regression applied to diabetes data

The logistic regression model relates the probability that a person has diabetes (p_i) to the covariates ($x_{i1}, \ldots, x_{ip}$) through a framework much like multiple regression.

That is, we want to find a transformation such that

$$\texttt{transformation}(p_i) = X_{n \times p} \beta_{p \times 1}. \tag{1}$$

- We want to choose transformation such that this makes both mathematical and practical sense.
- For example, we want a transformation that makes the range of possibilities on the left hand side of Equation 1 equal to the range of possibilities for the right hand side.
- If there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range.

Logistic regression applied to diabetes data

One common transformation is the logit transformation:

$$\text{logit}(p_i) = \log(\frac{p_i}{1 - p_i}) \tag{2}$$

We can then re-write Equation 1 as

$$\log(\frac{p}{1 - p}) = X_{n \times p} \beta_{p \times 1} \tag{3}$$

In fact, generalized linear models are a wide class of models that are widely used in statistics and involve making a transformation like we just did. Let's see how this ties back into our original application.