MATH 8050 Homework 5

Parampreet Singh

Due April 26 at 11:59pm

Instructions: For the problems that follow, please show all of your work. For problems that you solve using software, please provide the code you used either with the problem or in an appendix at the end of your homework document. To submit your answers for homework, please upload a **single** PDF to Canvas.

Problem 1

Load the diabetes data set from Canvas into R with `read.csv("~/diabetes.csv")`, where `~/` is the location where you have saved the file. Complete each part of this question based on the `diabetes` data. This data contains 5 variables:

- `glucose`: blood glucose concentration in mg/dL (milligrams per deciliter)
- `pressure`: Diastolic blood pressure in millimeters of mercury (mm Hg)
- `bmi`: body mass index (weight in kilograms by height in metres squared)
- `age`: age in years
- `diabetes`: yes/no indicator of whether the patient has diabetes. This will be our response variable.

(a) Identify the 3 components of a generalized linear model for this data, where `diabetes` is the response and all other variables are treated as predictors.

(b) Fit a generalized linear model (GLM) with `diabetes` as the response variable and all other variables as predictors, using the logit link for your model. Print the model summary and interpret each of the estimated coefficients for the predictors: `glucose`, `pressure`, `bmi`, and `age`.

(c) At the $\alpha = 0.05$ significance level, test whether the regression coefficient for `age` is equal to 0.05 or not, i.e.:

$$H_0 : \beta_4 = 0.05$$
$$H_A : \beta_4 \neq 0.05$$

(d) Calculate 90% confidence intervals for each of the coefficients for `glucose`, `pressure`, `bmi`, and `age`.

(e) What is the estimated probability of being diagnosed with diabetes when `glucose`= 150, `pressure`=100, `bmi`=20, and `age`=45? Keeping `glucose`, `bmi`, and `age` fixed at those values, try a few different values for `pressure` (within the range of the data). Do you think that `pressure` has much impact on the probability of diabetes for this particular data set?

Problem 2

Suppose we collect data about the number of patients that different primary care doctors in the Clemson area have. We collect the following variables:

- y_i: number of patients for doctor i, where $i = 1, \ldots, n$
- x_{i1}: distance that doctor i works from the university
- x_{i2}: average cost of the first office visit to doctor i

(a) Identify the correct distribution and link function to describe the response variable and then write down the log-likelihood for the corresponding GLM.

(b) Take the first and second derivatives of the log-likelihood with respect to the vector of regression coefficients, $\boldsymbol{\beta}$.

(c) What will the entries of the Fisher information matrix look like for this model?

Homework 5 (Data Analysis)

Parampreet SIngh

2024-04-26

PROBLEM 1

Part A: Random Component - diabetes, likely follow a binomial distribution given its yes/no nature. Specifies the distribution of the response variable. Systematic Component - predictors or independent variables (glucose, pressure, bmi, age). Link function - logit link function -> g(E(Y)) = log(E(Y) / 1 - E(Y)), which connects the mean of the random component to the systematic component.

```r
#Part B: Fit a generalized linear model (GLM) with diabetes as the response variable:
diabetes_data <- read.csv("diabetes.csv")

glm_model <- glm(diabetes ~ glucose + pressure + bmi + age,
                 data = diabetes_data, family = binomial(link = "logit"))

print(summary(glm_model))
```

```
##
## Call:
## glm(formula = diabetes ~ glucose + pressure + bmi + age, family = binomial(link = "logit"),
##     data = diabetes_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.582963   1.136690  -8.431  < 2e-16 ***
## glucose      0.036352   0.004931   7.373 1.67e-13 ***
## pressure    -0.002339   0.011422  -0.205 0.837752
## bmi          0.078953   0.020789   3.798 0.000146 ***
## age          0.054867   0.013810   3.973 7.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 354.32  on 387  degrees of freedom
## AIC: 364.32
##
## Number of Fisher Scoring iterations: 5
```

Interpretation - Intercept: The model intercept is approximately -9.582963. The intercept represents the log odds of the outcome (having diabetes) when all the predictors are held at zero, which is not practically interpretable for these variables.

Glucose: The estimated coefficient for glucose is approximately 0.036352. Since it's positive and highly significant ($p < 0.001$), higher glucose levels are associated with an increased likelihood of diabetes.

Pressure: The estimated coefficient for pressure is approximately -0.002339. The negative sign suggests that higher pressure is associated with a slightly decreased likelihood of diabetes, although the effect is very small and not statistically significant ($p > 0.05$).

BMI: The estimated coefficient for BMI is approximately 0.078953. The positive coefficient, which is statistically significant ($p < 0.001$), suggests that higher BMI is associated with an increased likelihood of diabetes.

Age: The estimated coefficient for age is approximately 0.054867. The positive coefficient indicates that the likelihood of diabetes increases with age, and this effect is statistically significant ($p < 0.001$).

```r
# Part C: Hypothesis Test
# H0: beta = 0.05
# Ha: beta != 0.05
# alpha = 0.05

# Hypothesized value for age coefficient
hypothesized_value <- 0.05

# Estimated coefficient and standard error for age from the GLM summary
estimated_coefficient <- 0.054867
standard_error <- 0.013810

# Compute the z-statistic
z_statistic <- (estimated_coefficient - hypothesized_value) / standard_error
p_value <- 2 * (1 - pnorm(abs(z_statistic)))

print(paste("z_statistic:", z_statistic))
```

```
## [1] "z_statistic: 0.352425778421433"
```

```r
print(paste("p_value:", p_value))
```

```
## [1] "p_value: 0.724518972164205"
```

The computed z-statistic is approximately 0.352 and the corresponding p-value is approximately 0.724. Since the p-value is greater than the significance level $= 0.05$, we fail to reject the null hypothesis. This means there isn't enough statistical evidence to say that the regression coefficient for age is different from 0.05 at the 5% significance level.

```r
#Part D: calculate the 90% confidence intervals
confint(glm_model, level = 0.90)
```

```
## Waiting for profiling to be done...

##                      5 %        95 %
## (Intercept) -11.53251585 -7.78567802
## glucose       0.02848386  0.04473905
## pressure     -0.02111615  0.01658231
## bmi           0.04532746  0.11389356
## age           0.03256096  0.07810079
```

```r
# Part E: Predicting the probability of being diagnosed with diabetes

predict_data <- data.frame(glucose = 150, pressure = 100, bmi = 20, age = 45)
logit_prob <- predict(glm_model, newdata = predict_data, type = "response")

print(paste("Diabetic Probability at pressure (",predict_data$pressure,"):",logit_prob))
```

```
## [1] "Diabetic Probability at pressure ( 100 ): 0.42166494863745"
```

```r
# Trying with different pressure values
pressure_values <- c(70, 100, 120, 150)
probabilities <- sapply(pressure_values, function(p) {
  predict_data$pressure <- p
  predict(glm_model, newdata = predict_data, type = "response")
})

names(probabilities) <- pressure_values
print(probabilities)
```

```
##        70       100       120       150
## 0.4388636 0.4216649 0.4103012 0.3934370
```

This trend suggests that as blood pressure increases, the model estimates a lower probability of being diagnosed with diabetes, given the specific values for the other variables. This outcome might seem counter-intuitive, given that one might expect higher blood pressure to be associated with a higher risk of diabetes. However, this is what the model has estimated based on the data it was trained on.

~~Pb~~ Problem 2

Given,

$y_i \rightarrow$ number of patients for doctor i
$x_{i1} \rightarrow$ distance that doctor i works from University
$x_{i2} \rightarrow$ average cost of 1st office visit

Response $\rightarrow$ y_i
Predictors $\rightarrow$ x_{i1}, x_{i2}

As our response variable deals with the count data, Poisson Distribution would be appropriate here.

(a) Response variable or random component : $y_i \sim \text{Poisson}(\lambda_i)$

where λ_i is the expected number of events

Poisson distribution density :-

$$f(y_i | \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} , \quad y_i = 0, 1, 2, \ldots \quad \lambda_i \in [0, \infty)$$

$E(y_i) = \lambda_i$
$Var(y_i) = \lambda_i$

GLM components

① Random : $y_i \sim \text{Poisson}(\lambda_i)$

② systematic : $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

③ Link : $g(E(y_i)) = g(\lambda_i) = \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

Scanned with CamScanner

Joint likelihood :

~~$L(\beta_0, \beta_1 | X, y)$~~

$$L(\beta_0, \beta_1, \beta_2 | X, y) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}$$

$$\lambda_i = \exp\{\beta_0 + \beta_1 x_i + \beta_2 x_{i_2}\}$$

Substituting λ_i,

$$L(\beta_0 \beta_1 \beta_2 | X, y) = \prod_{i=1}^{n} \frac{e^{-\exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\}} \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\}^{y_i}}{y_i!}$$

Taking the log,

$$\ell(\beta_0 \beta_1 \beta_2 | X, y) = \sum_{i=1}^{n} -\exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} + y_i(\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}) - \log(y_i!)$$

(b) Taking 1st & 2nd order derivatives w.r.t $\beta_0, \beta_1, \beta_2$:-

β_0

$$\frac{\partial \ell(\beta_0 \beta_1 \beta_2 | X, y)}{\partial \beta_0} = \sum_{i=1}^{n} -\exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} + y_i$$

$$\frac{\partial^2 \ell(\beta_0 \beta_1 \beta_2 | X, y)}{\partial \beta_0 \partial \beta_0^T} = \sum_{i=1}^{n} -\exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\}$$

β_1

$$\frac{\partial \ell(\beta_0 \beta_1 \beta_2 | X, y)}{\partial \beta_1} = \sum_{i=1}^{n} -x_{i_1} \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} + x_{i_1} y_i$$

$$\beta_1 \left\{ \frac{\partial^2 l(\beta_0 \beta_1 \beta_2 | X, Y)}{\partial \beta_1 \partial \beta_1^T} = \sum_{i=1}^{n} -x_{i_1}^2 \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} \right.$$

$$\beta_2 \left\{ \begin{array}{l} \frac{\partial l(\beta_0 \beta_1 \beta_2 | X, Y)}{\partial \beta_2} = \sum_{i=1}^{n} -x_{i_2} \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} + y_i x_{i_2} \\ \frac{\partial^2 l(\beta_0 \beta_1 \beta_2 | X, Y)}{\partial \beta_2 \partial \beta_2^T} = \sum_{i=1}^{n} -x_{i_2}^2 \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} \end{array} \right.$$

(C) our model $\rightarrow$

$$g(E(y_i)) = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}$$

As we have 2 predictors, our fisher Information matrix -

$$I\beta = \begin{pmatrix} I(\beta_0) & I(\beta_0 \beta_1) & I(\beta_0 \beta_2) \\ I(\beta_0 \beta_1) & I(\beta_1) & I(\beta_1 \beta_2) \\ I(\beta_0 \beta_2) & I(\beta_1 \beta_2) & I(\beta_2) \end{pmatrix}$$

where

$$I(\beta_0) = E\left(-\frac{\partial^2 l(\beta_0 \beta_1 \beta_2 | X, Y)}{\partial^2 \beta_0}\right) = \sum_{i=1}^{n} \exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\}$$

let, $\exp\{\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2}\} \rightarrow v_i$

$$I(\beta_0) = \sum_{i=1}^{n} v_i$$

Similarly,

$$I(\beta_0 \beta_1) = E\left(-\frac{\partial^2 l(\beta_0 \beta_1 \beta_2 | X, Y)}{\partial \beta_0 \partial \beta_1}\right) = \sum_{i=1}^{n} x_{i_1} v_i$$

Scanned with CamScanner

$$I(\beta_0 \beta_2) = \sum_{i=1}^{n} x_{i2} v_i \qquad I(\beta_1) = \sum_{i=1}^{n} x_{i1}^2 v_i$$

$$I(\beta_1 \beta_2) = \sum_{i=1}^{n} x_{i1} x_{i2} v_i \qquad I(\beta_2) = \sum_{i=1}^{n} x_{i2}^2 v_i$$

Fisher Information Matrix →

$$I(\beta) = \begin{pmatrix} \sum_{i=1}^{n} v_i & \sum_{i=1}^{n} x_{i1} v_i & \sum_{i=1}^{n} x_{i2} v_i \\ \sum_{i=1}^{n} x_{i1} v_i & \sum_{i=1}^{n} x_{i1}^2 v_i & \sum_{i=1}^{n} x_{i1} x_{i2} v_i \\ \sum_{i=1}^{n} x_{i2} v_i & \sum_{i=1}^{n} x_{i1} x_{i2} v_i & \sum_{i=1}^{n} x_{i2}^2 v_i \end{pmatrix}$$