DSCI 562 Lab 1

Introduction to Generalized Linear Models

Contents

Lab Mechanics

rubric={mechanics:5}

- Paste the URL to your GitHub repo here: **INSERT YOUR GITHUB REPO URL HERE**
- Once you finish the assignment, you must **knit** this `R` markdown to create a `.pdf` file and push everything to your GitHub repo using `git push`. You are responsible for ensuring all the figures, texts, and equations in the `.pdf` file are appropriately rendered.
- You must submit this `.Rmd` **and** the rendered `.pdf` files to Gradescope.

  Heads-up: You need to have a minimum of 3 commits.

Code Quality

rubric={quality:3}

The code that you write for this assignment will be given one overall grade for code quality. Check our **code quality rubric** as a guide to what we are looking for. Also, for this course (and other MDS courses that use `R`), we are trying to follow the `tidyverse` code style. There is a guide you can refer too: http://style.tidyverse.org/

Each code question will also be assessed for code accuracy (i.e., does it do what it is supposed to do?).

Writing

rubric={writing:3}

To get the marks for this writing component, you should:

- Use proper English, spelling, and grammar throughout your submission (the non-coding parts).
- Be succinct. **This means being specific about what you want to communicate, without being superfluous.**

Check our **writing rubric** as a guide to what we are looking for.

A Note on Challenging Questions

Each lab will have a few challenging questions. These are usually low-risk questions and will contribute to maximum 5% of the lab grade. The main purpose here is to challenge yourself or dig deeper in a particular area. When you start working on labs, attempt all other questions before moving to these questions. If you are running out of time, please skip these questions.

We will be more strict with the marking of these questions. If you want to get full points in these questions, your answers need to

- be thorough, thoughtful, and well-written,
- provide convincing justification and appropriate evidence for the claims you make, and
- impress the reader of your lab with your understanding of the material, your analytical and critical reasoning skills, and your ability to think on your own.

Setup

If you fail to load any packages, you can install them and try loading the library again.

```r
library(AER, quietly = TRUE)
library(MASS, quietly = TRUE)
library(tidyverse, quietly = TRUE)
library(broom, quietly = TRUE)
library(performance, quietly = TRUE)
library(qqplotr, quietly = TRUE)
library(cowplot, quietly = TRUE)
library(digest, quietly = TRUE)
library(testthat, quietly = TRUE)
library(ISLR2, quietly = TRUE)
library(faraway, quietly = TRUE)
library(see, quietly = TRUE)
```

This lab focuses on exploring generalized linear models (GLMs) with Binary Logistic and count regression models.

(Challenging) Exercise 1: Maximum Likelihood Estimation in Binary Logistic Regression

In **DSCI 552's `lecture8`**, we explored analytical univariate maximum likelihood estimation (MLE) via the Exponential distribution. Moreover, in **DSCI 552's `lab4`**, we explored another analytical MLE approach for the Poisson distribution.

Possibly to some of us, these univariate cases might seem trivial. Nevertheless, let us suppose we aim to estimate the **regression coefficients** (also called **weights** in Machine Learning) in a **Binary Logistic regression model**. Then, under this framework, MLE becomes crucial to estimate these coefficients and is the basis of the so-called `glm()` function.

Therefore, this challenging exercise will introduce you to the foundations of multivariate MLE for regression coefficients in a specific GLM (such as the Binary Logistic regression) using the analytical steps we saw in DSCI 552 along with the principle of the link function.

Q1.1.

rubric={reasoning:1}

Suppose you have a **training dataset** of size n with the ith response Y_i subject to p regressors $X_{i,1}, \dots, X_{i,p}$ $(i = 1, \dots, n)$. **Under this framework, the regressors can take a continuous or discrete nature.**

> **Heads-up:** Note the uppercase notation in all the variables (response and regressors). Theoretically speaking, they are assumed as random variables.

Now, let us begin with **step 1**: *choosing the right response distribution.* In Binary Logistic regression, the response takes on the following values:

$$Y_i = \begin{cases} 1 & \text{if there is a success,} \\ 0 & \text{otherwise.} \end{cases}$$

The **most basic** Binary Logistic regression model assumes that the n Y_is are **independent only** (not identically distributed!). Given this assumption, what theoretical distribution can we assume for Y_i to perform MLE? Why are you choosing this distribution along with its parameter(s)? **Answer in two or three sentences.**

Type your answer here, replacing this text.

Q1.2.

rubric={reasoning:1}

Using your answer for **Q1.1**, execute **step 2**: *obtaining the joint probability mass function (PMF).* You will need the corresponding PMFs of the n discrete Y_is. Recall that the notation in the likelihood function for random variables changes to lower cases since we depict n **observed values**.

> **Heads-up:** Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.3.

rubric={reasoning:1}

Using your answer for **Q1.2**, execute **step 3**: *obtaining the joint likelihood function*

$$l(\pi_1, \dots, \pi_n \mid y_1, \dots, y_n).$$

Briefly, justify your answer.

> **Heads-up:** Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.4.

rubric={reasoning:1}

Let β_0 be the regression intercept along with the p regression coefficients $\beta_1, \dots, \beta_p$ corresponding to the p regressors (also known as features) $X_{i,1}, \dots, X_{i,p}$. Unlike Ordinary Least-Squares (OLS), we cannot directly relate the ith response Y_i with $\beta_0, \beta_1, \dots, \beta_p$ and the p regressors.

Let π_i be the probability of success in Y_i. Then, from `lecture1`, we know that Binary Logistic regression will need to use a **link function** $h(\pi_i)$ with the above regression terms. Specifically, we do it via the natural logarithm (i.e., on the base e) of the odds $\frac{\pi_i}{1-\pi_i}$:

$$h(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}.$$

Hence, using this equation

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}, \tag{1}$$

show that $h(\pi_i)$ is mononotic; i.e.,

$$\pi_i = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^{p} \beta_j X_{i,j}\right)}. \tag{2}$$

> **Heads-up:** Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.5.

rubric={reasoning:1}

What is Equation (2)'s name? What is one of the most important modelling characteristics of this equation?

Answer in one or two sentences.

Type your answer here, replacing this text.

Q1.6.

rubric={reasoning:1}

We can easily show via Equation (1) the following:

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}\right). \tag{3}$$

Thus, using Equation (3), show that:

$$1 - \pi_i = \frac{1}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}\right)}. \tag{4}$$

Heads-up: Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.7.

rubric={reasoning:1}

Using the likelihood function from **Q1.3.**, along with Equations (3) and (4), show that:

$$l(\beta_0, \ldots, \beta_p \mid y_1, \ldots, y_n, x_{1,1}, \ldots, x_{1,p}, \ldots, x_{n,1}, \ldots, x_{n,p}) = \prod_{i=1}^{n} \left[\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}\right)\right]^{y_i} \left[\frac{1}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}\right)}\right].$$

Heads-up: Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.8.

rubric={reasoning:1}

Using the function from **Q1.7**, execute **step 4**: *obtaining the joint log-likelihood function*

$$\log\left[l(\beta_0, \ldots, \beta_p \mid y_1, \ldots, y_n, x_{1,1}, \ldots, x_{1,p}, \ldots, x_{n,1}, \ldots, x_{n,p})\right].$$

Heads-up: Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.9.

rubric={reasoning:1}

Using the log-likelihood function from **Q1.8**, execute **step 5**: *obtaining the first partial derivatives with respect to $\beta_0, \beta_1, \ldots, \beta_p$.*

Heads-up: Use LaTeX to show your work.

Type your answer here, replacing this text.

Q1.10.

rubric={reasoning:1}

In three or four sentences, state why we cannot continue with **step 6** *to obtain the analytical MLEs $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.* Moreover, explain an alternative to solve this issue.

Type your answer here, replacing this text.

Exercise 2: The Orange Juice Problem

Let us dig into a marketing-related problem to practice what we have learned about Binary Logistic regression. We will use the data set `OJ` from package `ISLR2` (James et al., 2021):

> *"The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded."*

```
str(OJ)
```

```
## 'data.frame':    1070 obs. of  18 variables:
##  $ Purchase      : Factor w/ 2 levels "CH","MM": 1 1 1 2 1 1 1 1 1 1 ...
##  $ WeekofPurchase: num  237 239 245 227 228 230 232 234 235 238 ...
##  $ StoreID       : num  1 1 1 1 7 7 7 7 7 7 ...
##  $ PriceCH       : num  1.75 1.75 1.86 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceMM       : num  1.99 1.99 2.09 1.69 1.69 1.99 1.99 1.99 1.99 1.99 ...
##  $ DiscCH        : num  0 0 0.17 0 0 0 0 0 0 0 ...
##  $ DiscMM        : num  0 0.3 0 0 0 0 0.4 0.4 0.4 0.4 ...
##  $ SpecialCH     : num  0 0 0 0 0 0 1 1 0 0 ...
##  $ SpecialMM     : num  0 1 0 0 0 1 1 0 0 0 ...
##  $ LoyalCH       : num  0.5 0.6 0.68 0.4 0.957 ...
##  $ SalePriceMM   : num  1.99 1.69 2.09 1.69 1.69 1.99 1.59 1.59 1.59 1.59 ...
##  $ SalePriceCH   : num  1.75 1.75 1.69 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceDiff     : num  0.24 -0.06 0.4 0 0 0.3 -0.1 -0.16 -0.16 -0.16 ...
##  $ Store7        : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 2 2 2 ...
##  $ PctDiscMM     : num  0 0.151 0 0 0 ...
##  $ PctDiscCH     : num  0 0 0.0914 0 0 ...
##  $ ListPriceDiff : num  0.24 0.24 0.23 0 0 0.3 0.3 0.24 0.24 0.24 ...
##  $ STORE         : num  1 1 1 1 0 0 0 0 0 0 ...
```

Our variables of interest will be the following:

- `Purchase`: The response of interest indicating whether the customer purchased Minute Maid (`MM`) or Citrus Hill (`CH`).
- `PriceDiff`: A continuous explanatory variable, the difference in juice price between both brands at the moment of purchase (`SalePriceMM - SalePriceCH`).
- `STORE`: A label indicating at what possible store a given juice was purchased. It is originally a numeric-type column (`0`, `1`, `2`, `3`, or `4`), but we will convert it to a nominal factor.

Run the below code before proceeding.

```
OJ <- OJ %>%
  select(Purchase, PriceDiff, STORE) %>%
  mutate(STORE = as.factor(STORE))

OJ$STORE <- recode_factor(OJ$STORE,
    '0' = 'Store 0', '1' = 'Store 1',
    '2' = 'Store 2', '3' = 'Store 3',
    '4' = 'Store 4'
  )
```

Suppose you are part of the Data Science team from Citrus Hill, and you want to assess which explanatory variables (from the ones above) are **statistically associated** to the following response (and by how much!):

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th customer purchased Citrus Hill,} \\ 0 & \text{otherwise.} \end{cases}$$

Q2.1.

rubric={autograde:2}

We will start with some data wrangling. Convert the labels in `OJ$Purchase` as `1` for `CH` and `0` for `MM`. **The column should be numeric.**

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/Q2.1.R")
```

```
## Test Q2.1 - 1 failed:
## Purchase be assigned numeric
## is.numeric(OJ$Purchase) is not TRUE
##
##   `actual`:   FALSE
##   `expected`: TRUE
##
## Test Q2.1 - 2 failed:
## labelling is not correct
## 'sum' not meaningful for factors
```

Q2.2.

rubric={viz:2,reasoning:3}

What if we treat the categories of `Purchase` as probabilities?

Use `geom_point()` to plot these *"purchase probabilities"* from `Purchase` on the y-axis versus the `PriceDiff` on the x-axis, along with the OLS regression model estimated line on top with `geom_smooth()`. Include proper axis labels and title.

Assign your plot to `OJ_scatterplot`.

```
OJ_scatterplot <- NULL

# YOUR CODE HERE

OJ_scatterplot
```

```
## NULL
```

Now, answer the following:

1. What probabilistic quantity does this OLS linear regression aim to model and with what explanatory variable? **Explain in one or two sentences.**

Type your answer here, replacing this text.

2. What behaviour do you see in the estimated regression line? **Explain in one or two sentences.**

Type your answer here, replacing this text.

3. Based on the OLS regression framework, would this be a good model to use? Why or why not? **Explain in one or two sentences.**

Type your answer here, replacing this text.

Q2.3.

rubric={reasoning:3}

Because the OLS linear regression of `Purchase` versus `PriceDiff` might be questionable, let us try a Binary Logistic regression model. The **logit** link function of this GLM **will not** assume a linear relationship between the estimated purchase probability and `PriceDiff`.

Let π_i be purchase probability of the ith customer and $\texttt{PriceDiff}_i$ the continuous explanatory variable. Answer the following:

1. Write the sample's model equation of the *logit* link function.

   Heads-up: Use LaTeX to show your work.

Type your answer here, replacing this text.

2. What is the distributional assumption we are making on each customer Y_i given a `Purchase`? **Explain briefly.**

Type your answer here, replacing this text.

3. What is the distributional parameter in this framework? **Explain briefly.**

Type your answer here, replacing this text.

Q2.4.

rubric={autograde:2}

Estimate a Binary Logistic regression of `Purchase` versus `PriceDiff` and call it `bin_log_model`.

```r
bin_log_model <- NULL

# YOUR CODE HERE

bin_log_model
```

```
## NULL
```

```r
. = ottr::check("tests/Q2.4.R")
```

```
## Test Q2.4 - 1 passed
##
##
## Test Q2.4 - 2 failed:
## the correct fitting function is not being used
## "glm" %in% class(bin_log_model) is not TRUE
##
##   `actual`:   FALSE
##   `expected`: TRUE
##
## Test Q2.4 - 3 failed:
## check the formula and data arguments in the fitting function
## digest(round(sum(bin_log_model$coefficients), 2)) not equal to "2d5a88a9304983ce583b433b7974c08c".
##   1/1 mismatches
##   x[1]: "908d1fd10b357ed0ceaaec823abf81bc"
##   y[1]: "2d5a88a9304983ce583b433b7974c08c"
```

Q2.5.

rubric={viz:1,reasoning:1}

Plot the `bin_log_model` estimated regression equation on the `OJ_scatterplot`.

```r
# YOUR CODE HERE
```

What do you notice between both estimated regression equations? **Explain in one or two sentences.**

Type your answer here, replacing this text.

Q2.6.

rubric={viz:2,reasoning:1}

Make suitable plots comparing `PriceDiff` on the y-axis by each level of `Purchase` on the x-axis, which has to be faceted by `STORE` (check `facet_wrap()`). Include proper axis labels and title.

```r
# YOUR CODE HERE
```

In one or two sentences,, comment on what you observe about the relationship of `PriceDiff` and `Store` on `Purchase`.

Type your answer here, replacing this text.

Q2.7.

rubric={autograde:2}

Now, estimate a second Binary Logistic regression of `Purchase` versus `PriceDiff` and `STORE`. Call it `bin_log_model_2`.

```r
bin_log_model_2 <- NULL

# YOUR CODE HERE

bin_log_model_2
```

```
## NULL
```

```r
. = ottr::check("tests/Q2.7.R")
```

```
## Test Q2.7 - 1 passed
##
##
## Test Q2.7 - 2 failed:
## the correct fitting function is not being used
## "glm" %in% class(bin_log_model_2) is not TRUE
##
##   `actual`:   FALSE
##   `expected`: TRUE
##
## Test Q2.7 - 3 failed:
## check the formula and data arguments in the fitting function
## digest(round(sum(bin_log_model_2$coefficients), 2)) not equal to "6e76d62738b6d511f99f4c042626a94d".
##   1/1 mismatches
##   x[1]: "908d1fd10b357ed0ceaaec823abf81bc"
##   y[1]: "6e76d62738b6d511f99f4c042626a94d"
```

Q2.8.

rubric={accuracy:1,reasoning:2}

Compare both Binary Logistic regression models, `bin_log_model` and `bin_log_model_2`, which one fits the data better? You can use **either one** of the methods **introduced in `lecture2`**. State your hypotheses and use a significance level $\alpha = 0.05$, **if necessary**. Provide the necessary code to support your conclusion.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

Q2.9.

rubric={accuracy:1,reasoning:2}

Using `bin_log_model_2`, are the explanatory variables `PriceDiff` and `STORE` statistically associated to the response `Purchase`? State your conclusions with a significance level $\alpha = 0.05$. Provide the necessary code to support these conclusions.

> **Heads-up:** Recall that `STORE` is a nominal factor. Thus, the statistical conclusions will be in function of certain levels when compared to the baseline.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

Q2.10.

rubric={accuracy:1,reasoning:4}

Using `bin_log_model_2`, interpret those statistically significant estimated coefficients (with $\alpha = 0.05$) in terms of the odds coming from the response `Purchase`. Provide the necessary code to support these interpretations.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

Q2.11.

rubric={autograde:1}

Let us consider a customer at `Store 2` who finds Citrus Hill $0.50 less expensive than Minute Maid. Predict their purchase probability of Citrus Hill using `bin_log_model_2`. Then, bind your results to the **vector** `pred_CH_purchase`.

```
pred_CH_purchase <- NULL

# YOUR CODE HERE

pred_CH_purchase
```

```
## NULL
```

```
. = ottr::check("tests/Q2.11.R")
```

```
## Test Q2.11 - 1 failed:
## pred_CH_purchase should be a vector
## "numeric" %in% class(pred_CH_purchase) is not TRUE
##
##   `actual`:   FALSE
##   `expected`: TRUE
##
## Test Q2.11 - 2 failed:
```

```
## wrong predicted probability
## non-numeric argument to mathematical function
```

(Challenging) Q2.12.

rubric={viz:1,reasoning:2}

Obtain the corresponding binned residual plot with `bin_log_model_2`.

```
# YOUR CODE HERE
```

Then, answer the following **in or two sentences**:

- Can we say that it is a proper model fitting?
- Why or why not?

Type your answer here, replacing this text.

Exercise 3: Doctor Visits in Australia

The `dvisits` dataset (from package `faraway`) contains data from the **Australian Health Survey of 1977-78** and consists of 5190 single elderly adults. We will model the relationship between the count of consultations with a doctor or specialist in the past two weeks (`doctorco`) to the regressor `age` in years along with the factor-type `sex` (**encoded in this dataset** as binary with levels `1` for female and `0` for male as baseline) and the individual's `income` in thousands of Australian dollars.

Run the below code before proceeding.

```
dvisits <- dvisits %>%
  select(doctorco, age, sex, income) %>%
  mutate(
    sex = as.factor(sex),
    age = age * 100
  )
```

Q3.1.

rubric={viz:2,reasoning:1}

Create a proper visualization to summarize the relationship between the count of consultations with a doctor or specialist in the past two weeks to `age` and `income` (**separately!**).

```
# YOUR CODE HERE
```

Comment on these relationships **in one or two sentences**.

Type your answer here, replacing this text.

Q3.2.

rubric={reasoning:1}

In what way(s) would OLS regression be inappropriate for this data, if at all? **Answer in one or two sentences.**

Type your answer here, replacing this text.

Q3.3.

rubric={accuracy:2}

Using `doctorco` as response along with `age`, `income`, and `sex` as regressors; fit a Poisson regression model with a log-link function and call it `poisson_model`.

```
poisson_model <- NULL

# YOUR CODE HERE

summary(poisson_model)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

Q3.4

rubric={accuracy:1,reasoning:1}

Using `poisson_model`, are all regression coefficients statistically associated to `doctorco`? State your conclusions with a significance level $\alpha = 0.05$. Provide the necessary code to support your conclusions.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

Q3.5.

rubric={accuracy:1,reasoning:3}

Using `poisson_model`, interpret those statistically significant estimated coefficients in terms of the **original scale** of the response `doctorco`. Provide the necessary code to support these interpretations.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

(Challenging) Q3.6.

rubric={accuracy:1,viz:1,reasoning:1}

We can also compute the deviance residuals for a count regression model. These residuals can be used in a diagnostic plot of in-sample fitted values on the x-axis versus deviance residuals on the y-axis.

> **Heads-up:** The scatterplot of **in-sample fitted values versus deviance residuals** is used to evaluate the linearity in the GLM graphically. Under a linearity assumption, we would expect a **flat trend** in the data points.

Using `augment()`, extract the in-sample predictions and deviance residuals from `poisson_model`. Then, make a scatterplot using `geom_point()`. Add a horizontal dashed line on 0 in the y-axis along with a locally estimated scatterplot smoothing (LOESS) line using `geom_smooth()`.

```
# YOUR CODE HERE
```

Do you see any pattern in the plot? If so, does this indicate a non-linearity matter? **Answer in one or two sentences.**

Type your answer here, replacing this text.

Q3.7.

rubric={accuracy:1,reasoning:1}

Test for overdispersion in your `poisson_model` using a significance level of $\alpha = 0.05$. Provide the necessary code to support your statistical conclusions **in one or two sentences**.

Type your answer here, replacing this text.

```
# YOUR CODE HERE
```

Q3.8.

rubric={accuracy:2}

Because there is overdispersion, we might want to improve our distributional assumption as a Negative Binomial whose variance is freed up from the mean (recall the Poisson variance is equal to its mean). Note that the Poisson distribution is a particular case of Negative Binomial.

Fit a Negative Binomial regression model on the data with the same regressors and response and call it `nb_model`.

```
nb_model <- NULL

# YOUR CODE HERE

summary(nb_model)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

Q3.9.

rubric={accuracy:1,reasoning:2}

Compare the estimates and their standard errors between the `poisson_model` and the `nb_model`. What is the impact on these standard errors when we free up the variance with the `nb_model`? Describe it **in plain words in three or four sentences**. Provide the necessary code to support your conclusions.

Type your answer here, replacing this text.

```
# YOUR CODE HERE
```

Submission

Congratulations! You are done the lab. Do not forget to:

- Knit the assignment to generate the `.pdf` file and push everything to your Github repo.
- Double check all the figures, texts, equations are rendered properly in the `.pdf` file
- Submit the `.Rmd` and the `.pdf` files to Gradescope.

Reference

- Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani (2021). ISLR2: Introduction to Statistical Learning, Second Edition. R package version 1.3. https://CRAN.R-project.org/package=ISLR2