PSTAT127 Homework 2

Billy Dang

2024-10-16

```
library(ggplot2)
```

Problem 1: A simple simulation for Bernoulli R.V. and Logistic Regression

Suppose we have a single covariate x. Generate n = 1000 independent Bernoulli(π_i) random variables $\{Y_i$: i = 1, . . . , n$\}$, according to the model:

$$logit(\mu_i) = \beta_0 + \beta_1 x_i, \quad i = 1, ..., n$$

where $\mu_i = E(Y_i)$, with $\beta_0 = 1$, $\beta_1 = 0.5$, and 1000 design points x_i regularly spaced on the interval (-10, 10)

1.(a-c):

- You can use command seq to generate regularly spaced design points
- You can do all steps using vectors, you don't need any loops here
- After assigning values for n, $/beta_0$, $/beta_1$, and $\{x_i : i = 1, ..., 1000\}$, calculate the corresponding true π_i values for $i = 1, ..., n$.

```
set.seed(333)

#Num of design points & parameters
n = 1000
beta_0 <- 1
beta_1 <- 0.5

#Generate xi's
x <- seq(-10, 10, length.out = n)
```

```
#Create vector of R.V.'s as well as vector of mu values

#Logit (linear predictor) that transforms probabilities into log-odds
logit_mu <- beta_0 + beta_1*x
#Inverse logit (sigmoid function) to recover probability mu_i from the logit function and restrains it
mu <- 1 / (1 + exp(-logit_mu))
```

In a logistic regression model for **Bernoulli** random variables, π_i represents the probability of success (i.e. the probability that $y_i = 1$). This means:

$$\pi_i = P(y_i = 1 | x_i)$$

Where μ_i represents the same probability in that:

$$\mu_i = E(Y_i) = P(Y_i = 1|x_i)$$

Additionally, in a Generalized Linear Model (GLM), with 3 components:

- Random Component: The distribution of the response variable (in this case Bernoulli for binary outcomes)
- Systematic Component: The linear predictor, which is a linear combination of the covariates
- Link function: The function that connects the expected value μ_i (which is π_i here) to the linear predictor.

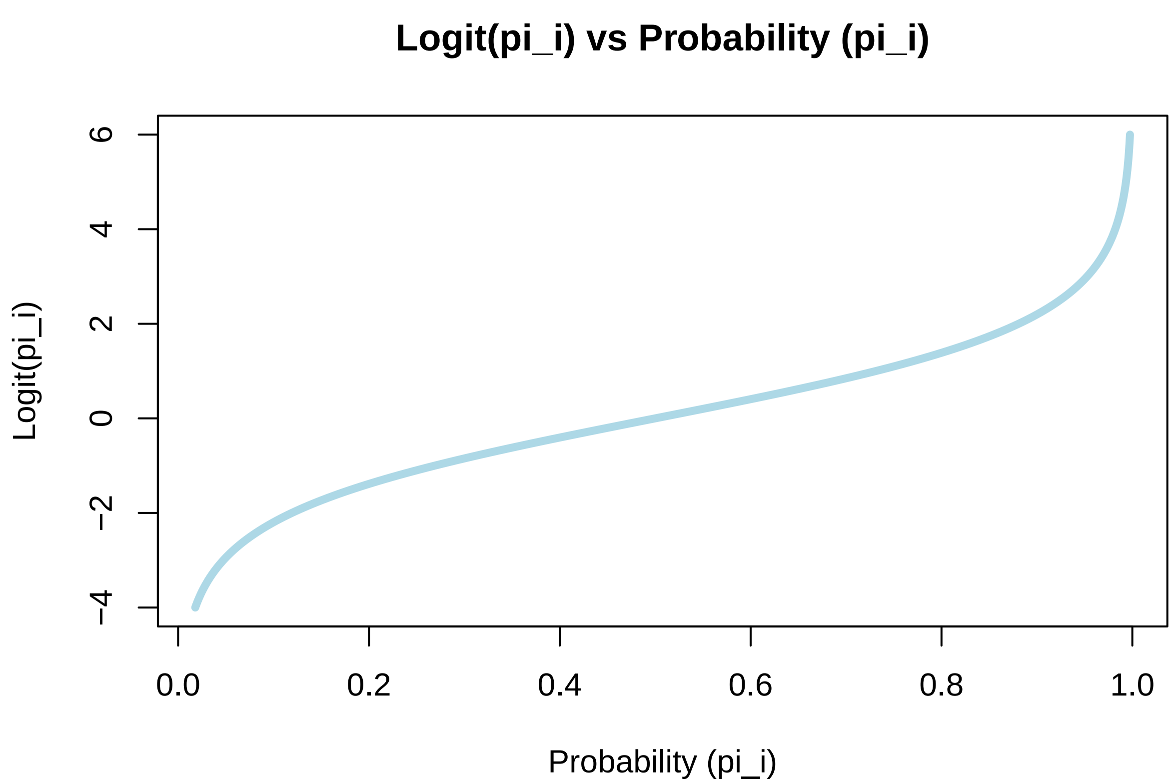
For logistic regression, the link function is the logit link function:

$$logit(\mu) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

1.d:

Plot $logit(\pi_i)$ (vertical axis) versus π_i (horizontal axis) for these n = 1000 points.

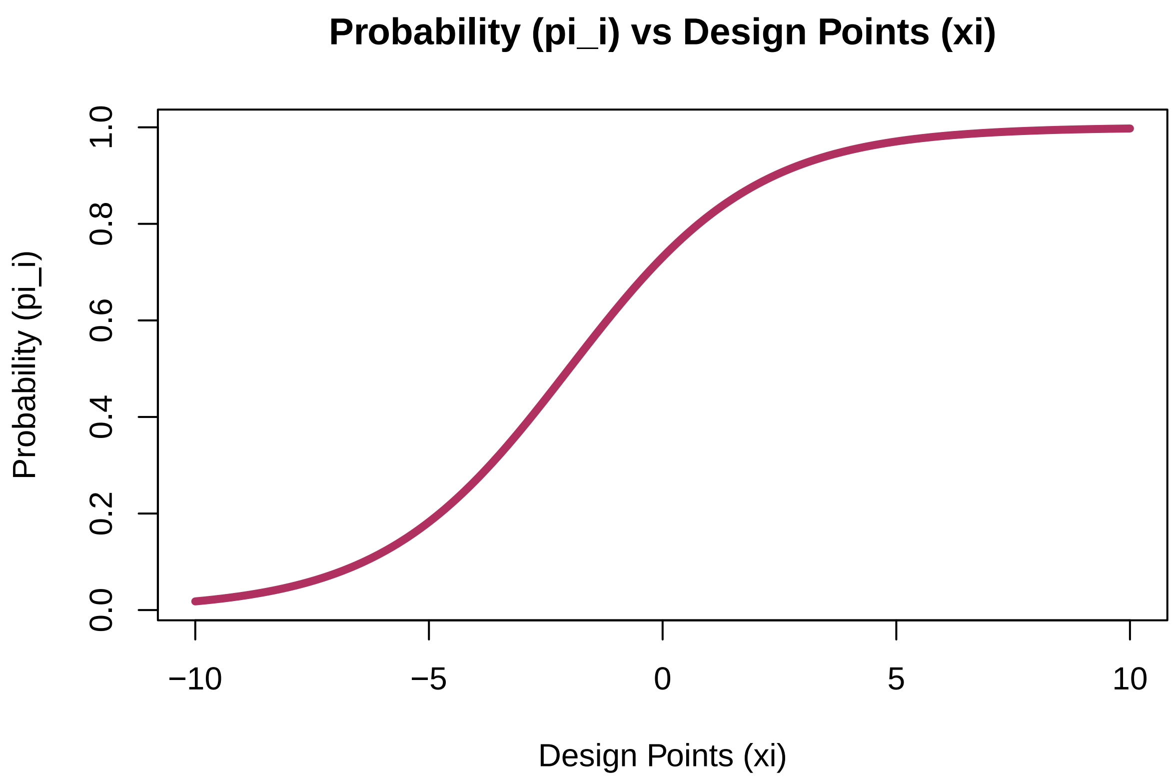
```r
plot(mu, logit_mu, type = "l", col = "lightblue", lwd = 4,
     main = "Logit(pi_i) vs Probability (pi_i)",
     xlab = "Probability (pi_i)", ylab = "Logit(pi_i)")
```



1.e:

Plot π_i (vertical axis) versus x_i (horizontal axis) for these n = 1000 points.

```
plot(x, mu, type = "l", col = "maroon", lwd = 4,
     main = "Probability (pi_i) vs Design Points (xi)",
     xlab = "Design Points (xi)", ylab = "Probability (pi_i)")
```



1.f

Now use R command "rbinom", with the appropriate parameter vector and constans, to simulate $Y_i \overset{\text{indep}}{\sim}$ Bernoulli(π_i). What are the only possible values for y_i? Save your results as a vector y.

** The only possible values for y_i are 0 or 1 due to the log transformation done to the linear predictor values.

```
set.seed(333)
#Generate n Bernoulli R.V.'s with mu vector
y <- rbinom(n, size = 1, prob = mu)
y[50:55] #Looking at 5 random results in the sim
```

```
## [1] 0 0 0 0 0 0
```

Now fit a logistic regression model to your simulated data vector using the command (see pdf). What link function is being used (i.e., what is the default link function in R for binomial family)?

R uses the Logit Link function denoted as:

$$logit(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

where μ represents the probability of success (π) in this case and the log function is the natural logarithm of the odds.

```r
glm1 <- glm(y ~ x, family = binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.06665    0.11823   9.022   <2e-16 ***
## x            0.50450    0.03088  16.339   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1342.71  on 999  degrees of freedom
## Residual deviance:  626.02  on 998  degrees of freedom
## AIC: 630.02
##
## Number of Fisher Scoring iterations: 6
```

The values of my parameter estimates are $\hat{\beta}_0 = .9011$ and $\hat{\beta}_1 = .4860$.