2. Logistic Regression

(a)

5/5 points (graded)

In this problem, we will model the likelihood of a particular client of a financial firm defaulting on his or her loans based on previous transactions. There are only two outcomes, “Yes" or “No", depending on whether the client eventually defaults or not. It is believed that the client's current balance is a good predictor for this outcome, so that the more money is spent without paying, the more likely it is for that person to default.

For each x, we will write Y_x as the 0-1 outcome of defaulting/not defaulting, given a particular current balance x. In other words, we will model the distribution of Y_x as a Bernoulli distribution with with a parameter x, which is reasonable given that there are only two possible outcomes.

First, recall the likelihood of a Bernoulli RV Y in terms of the parameter p:

$$\mathbf{P}(Y=y) = p^y (1-p)^{1-y}.$$

Rewrite this in terms of an exponential family

$$\mathbf{P}(Y=y) = h(y) \exp\left[\eta(p) T(y) - B(p)\right].$$

Since this representation is only unique up to re-scaling by constants, take the convention that $T(y) = y$.

$\eta(p) =$ ln(p/(1-p)) ✔ **Answer:** ln(p/(1-p))

$B(p) =$ -ln(1-p) ✔ **Answer:** -ln(1-p)

$h(y) =$ 1 ✔ **Answer:** 1

We can write this in canonical form, e.g. as

$$\mathbf{P}(Y=y) = h(y) \exp\left[y\eta - b(\eta)\right].$$

What is $b(\eta)$?

$b(\eta) =$ ln(1+exp(eta)) ✔ **Answer:** ln(1+exp(eta))

Recall that the mean of a Bernoulli(p) distribution is p. What is the canonical link function $g(\mu)$ associated with this exponential family, where $\mu = \mathbb{E}[Y]$? Write your answer in terms of p.

$g(\mu) =$ ln(p/(1-p)) ✔ **Answer:** ln(p/(1-p))

STANDARD NOTATION

Solution:

We can rewrite the likelihood as

$$\mathbf{P}(Y=y) = \exp\left(y \ln p + (1-y) \ln(1-p)\right)$$

$$= \exp\left(y \ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right)$$

Hence, given the convention $T(y) = y$ for this specific case, we set

$$\begin{aligned} h(y) &= 1 \\ B(p) &= -\ln(1-p) \\ \eta(p) &= \ln\left(\frac{p}{1-p}\right). \end{aligned}$$

In order to rewrite this in canonical form, solve

$$\ln\left(\frac{p}{1-p}\right) = \eta \iff p = \frac{e^\eta}{1+e^\eta}.$$

so

$$b(\eta) = \ln(1+e^\eta).$$

The canonical link function is b'^{-1}, which is

$$b'^{-1}(p) = \ln\left(\frac{p}{1-p}\right).$$

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

(b)

2/2 points (graded)
What range will the values in Y belong to?

- (selected, correct) $\{0,1\}$, the set of two values 0 and 1.
- $(0,1)$, the open interval between 0 and 1.
- $[0,1]$, the closed interval enclosed by 0 and 1.
- Other

According to the canonical Generalized Linear Model (your answer from (a)), what is the range of possible predictions for p?

- $\{0,1\}$, the set of two values 0 and 1.
- (selected, correct) $(0,1)$, the open interval between 0 and 1.
- $[0,1]$, the closed interval enclosed by 0 and 1.
- Other

Solution:

Y is Bernoulli, so it lives in $\{0, 1, 2, \ldots\}$.

Since the canonical model states $\lambda = e^{\eta}$, the range of λ_t is the full range of parameters for a Poisson distribution: $\mathbb{R}_{>0} = \{\lambda \in \mathbb{R} : \lambda > 0\}$.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

(c)

3/3 points (graded)
Return to the original model. We now introduce a bias parameter p_x for every possible amount of money that measures the balance in a person's account.

Denote the parameter (η) that gives the canonical exponential family representation as above by θ_x. We choose to employ a linear model connecting the balance x with the canonical parameter θ_x of the Bernoulli distribution above, i.e.,

$$\theta_x = a + bx.$$

In other words, we choose a generalized linear model with the Bernoulli distribution and its canonical link function. That also means that conditioned on x, we assume the Y_x to be independent.

Imagine we observe the following data:

$x_1 = -100$ 0 (Never defaulted)
$x_2 = 2000$ 0 (Never defaulted)
$x_3 = 2000$ 1 (Defaulted)
$x_4 = 5000$ 1 (Defaulted)

We want to produce a maximum likelihood estimator for (a, b). To this end, write down the log likelihood $\ell(a, b)$ of the model for the provided four observations at x_1, x_2, x_3 and x_4 (plug in their values).

$\ell(a, b) =$

-ln(1+exp(a-100*b))-2*ln(1+exp(a+2000*b))+(a+2000*b)+(a+5000*b)-ln(1+exp(a+5000*b))

Answer: 2*a+7000*b-ln(1+exp(a-100*b))-2*ln(1+exp(a+2000*b))-ln(1+exp(a+5000*b))

$$-\ln(1+\exp(a-100\cdot b)) - 2\cdot\ln(1+\exp(a+2000\cdot b)) + (a+2000\cdot b) + (a+5000\cdot b) - \ln(1+\exp(a+5000\cdot b))$$

What is its gradient? Enter your answer as a pair of derivatives.

$\partial_a \ell(a, b) =$

- exp(a - 100*b)/(1 + exp(a - 100*b)) - (2*exp(a + 2000*b))/(1 + exp(a + 2000*b)) - exp(a + 5000*b)/(1

Answer: 2-(1/(1+exp(-a+100*b)))-(2/(1+exp(-a-2000*b)))-(1/(1+exp(-a-5000*b)))

$$-\frac{\exp(a-100\cdot b)}{1+\exp(a-100\cdot b)} - \frac{2\cdot\exp(a+2000\cdot b)}{1+\exp(a+2000\cdot b)} - \frac{\exp(a+5000\cdot b)}{1+\exp(a+5000\cdot b)} + 2$$

$\partial_b \ell(a, b) =$

(100*exp(a - 100*b))/(1 + exp(a - 100*b)) - (4000*exp(a + 2000*b))/(1 + exp(a + 2000*b)) - (5000*exp(

Answer: 7000+(100/(1+exp(-a+100*b)))-(4000/(1+exp(-a-2000*b)))-(5000/(1+exp(-a-5000*b)))

$$\frac{100\cdot\exp(a-100\cdot b)}{1+\exp(a-100\cdot b)} - \frac{4000\cdot\exp(a+2000\cdot b)}{1+\exp(a+2000\cdot b)} - \frac{5000\cdot\exp(a+5000\cdot b)}{1+\exp(a+5000\cdot b)} + 7000$$

Solution:

The likelihood for one observation is given by

$$\mathbf{P}(Y_x = y) = \exp\left(y(a+bx) - \ln\left(1+e^{a+bx}\right)\right)$$

That means the log likelihood for the model for n observations is

$$\ell(a,b) = \sum_{i=1}^{n}\left[y_i(a+bx_i) - \ln\left(1+e^{a+bx_i}\right)\right].$$

Plugging in the provided values, we get

$$\begin{aligned}\ell(a,b) = \quad & -\ln\left(1+e^{a-100b}\right) - 2\ln\left(1+e^{a+2000b}\right) - \ln\left(1+e^{a+5000b}\right)\\ & +2a+7000b.\end{aligned}$$

Its derivative with respect to a is

$$\partial_a \ell(a,b) = \quad 2 - \frac{exp(a-100b)}{(1+exp(a-100b))} - \frac{2exp(a+2000b)}{(1+exp(a+2000b))} - \frac{exp(a+5000b)}{(1+exp(a+5000b))}.$$

Its derivative with respect to b is

$$\partial_b \ell(a,b) = \quad 7000 + \frac{100exp(a-100b)}{(1+exp(a-100b))} - \frac{4000exp(a+2000b)}{(1+exp(a+2000b))} - \frac{5000exp(a+5000b)}{(1+exp(a+5000b))}.$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

(d)

1/1 point (graded)

Assume that we can reasonably estimate the likelihood estimator by using numerical methods to solve $\nabla_{(a,b)}\ell = 0$. Consider the scenario where, using many more samples, we obtain the estimates

$$\widehat{a} \approx 0.0012, \quad \hat{b} \approx 0.00035.$$

Given these results, what would be the predicted expected outcome $\mathbb{E}[Y_x]$ for $x = 4000$? Round your answer to the nearest 0.001.

0.80237 ✔ **Answer:** 0.802374

Solution:

We obtain the expected outcome as

$$\mathbb{E}[Y_x|x] = p_x = \frac{e^{a+bx}}{1+e^{a+bx}}.$$

Using the estimates for a and b and $x = 4000$, we obtain the prediction

$$p_{4000} = \frac{e^{0.0012+0.00035\cdot 4000}}{1+e^{0.0012+0.00035\cdot 4000}} \approx 0.802374.$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Discussion

Topic: Unit 7 Generalized Linear Models:Homework 11 / 2. Logistic Regression

Show Discussion