UNIVERSITY OF TEXAS AT AUSTIN

HW Assignment 4

Logistic regression.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 4.1. (10 + 5 + 5 = 20 points) Solve Problem **3.3** from the textbook (p.122). Before you start working on the solutions to the textbook questions, explicitly write out the fit in general, the fit for high school graduates and the fit for college graduates.

Solution:

(a) The fitted values satisfy

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 \\ &= 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5\end{aligned}$$

with x_1 = GPA, x_2 = IQ, x_3 = Level (1 for College and 0 for High School), x_4 = Interaction between GPA and IQ, and x_5 = Interaction between GPA and Level. We can separate this model by level in the following way. The fit for high school graduates is

$$\hat{y} = 50 + 20x_1 + 0.07x_2 + 0.01x_1x_2.$$

The fit for college graduates is

$$\begin{aligned}\hat{y} &= 50 + 20x_1 + 0.07x_2 + 35 + 0.01x_1x_2 - 10x_1 \\ &= 85 + 10x_1 + 0.07x_2 + 0.01x_1x_2\end{aligned}$$

So, depending on the actual values of x_1 and x_2, one or the other can be higher. Thus, neither **i.** nor **ii.** are correct.

On the other hand, for high enough values of x_1, the fit for high school graduates will exceed the corresponding fit for college graduates. Hence, **iii.** is **TRUE**.

Finally, **iv.** is evidently not true (since **iii.** is true).

(b) As obtained above, the fit for college graduates is

$$\hat{y} = 85 + 10(4) + 0.07(110) + 0.01(4)(110) = 137.1.$$

(c) Since the GPA and IQ are on different scales, this is **FALSE**.

Problem 4.2. (5 points) *Source: An old SOA exam.*
You are using logistic regression to predict the probability of a particular class of driver having an accident in the next insurance period. Your predictor is a categorical random variable indicating the *Area* in which the driver does most of their driving: *Suburban, Urban, Rural*. The *Suburban* category is understood as the *baseline*. You obtain the following summary of coefficients:

Intercept	-2.358
AreaUrban	0.905
AreaRural	-1.129

What is the fitted probability that an *Urban* driver will have an accident?

Solution:

$$\hat{p} = \frac{e^{-2.358+0.905}}{1 + e^{-2.358+0.905}} = 0.1895403.$$

Problem 4.3. (7 points) *Source: An old SOA exam.*
You are using logistic regression to predict the probability of a particular class of driver having a claim in the next insurance period. Your predictors are the two categorical random variable indicating the

- *Area* in which the driver does most of their driving: *Exurban, Suburban, Urban, Rural* with *Exurban* as *baseline*.
- *Vehicle body (VB)* of the vehicle the driver drives the most: *Coupe, Sedan, Truck* with *Coupe* as the *baseline*.

You obtain the following summary of coefficients:

Intercept	-1.485
AreaSuburban	0.094
AreaUrban	0.037
AreaRural	-0.101
VBSedan	-1.175
VBTruck	-1.118

What is the fitted probability that a *Rural* driver of a *Sedan* will have an accident?

Solution:

$$\hat{p} = \frac{e^{-1.485-0.101-1.175}}{e^{-1.485-0.101-1.175}+1} = 0.05946841$$

Problem 4.4. (7 points) *Source: MAS exam, Fall 2018.*
In a study, 100 subjects were asked to choose one of three election candidates (A, B, C). The subjects were organized into four age categories $(18-30, 30-45, 45-61, 61+)$.

A logistic regression was fitted to the subjects' responses to predict their preferred candidate with age group $(18-30)$ and candidate A as reference categories.

For age group $(18-30)$ the log-odds for preference of candidate B and candidate C were -0.535 and -1.489, respectively.

Calculate the modeled probability of someone from age group $(18-30)$ preferring candidate B.

Solution: Since the log-odds are given to be -0.535, the odds are $e^{-0.535}$. Hence, the modeled probability is

$$\hat{p} = \frac{e^{-0.535}}{1+e^{-0.535}} = 0.3693515.$$

Problem 4.5. (11 points) *Source: MAS-I exam, Spring 2019.*
A statistician uses a logistic model to predict the probability of success, π, of a binomial random variable.
You are given the following information:

- There is one predictor random variable, X, and an intercept in the model.
- The estimates of π at $x = 4$ and $x = 6$ are 0.88877 and 0.96562, respectively.

Calculate the estimated intercept coefficient, b_0, and the slope coefficient, b_1, in the logistic model that produced the above probability estimates.

Solution: To obtain the fitted regression, we apply the *logit* function to the given probabilities.

$$\ln\left(\frac{0.88877}{1-0.88877}\right) = b_0 + b_1(4)$$
$$\ln\left(\frac{0.96562}{1-0.96562}\right) = b_0 + b_1(6)$$

We obtain the following system of two equations with two unknowns:

$$b_0 + 4b_1 = 2.078238$$
$$b_0 + 6b_1 = 3.335295$$

Subtracting the first from the second equation above, we get

$$2b_1 = 3.335295 - 2.078238 = 1.257057 \quad \Rightarrow \quad b_1 = 0.6285285.$$

Reusing the first equation, we obtain

$$b_0 = 2.078238 - 4(0.6285285) = -0.435876.$$