Exercises for exercise class 6a in MMS075, Feb 25, 2020

1. A data collection activity yielded the following data set:

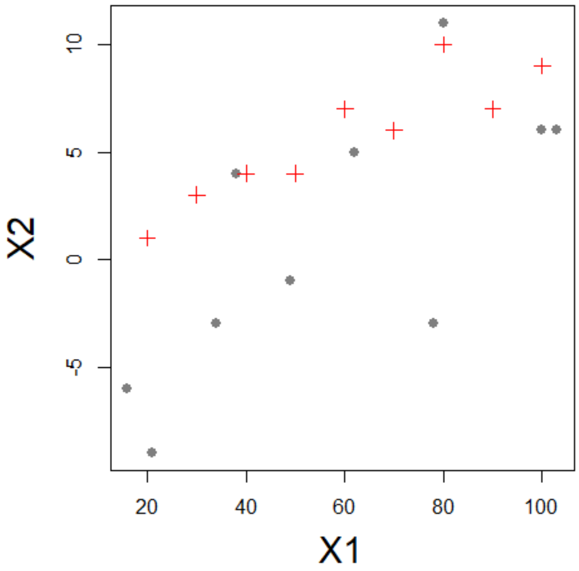
Observations when response Y = ”Yes”:

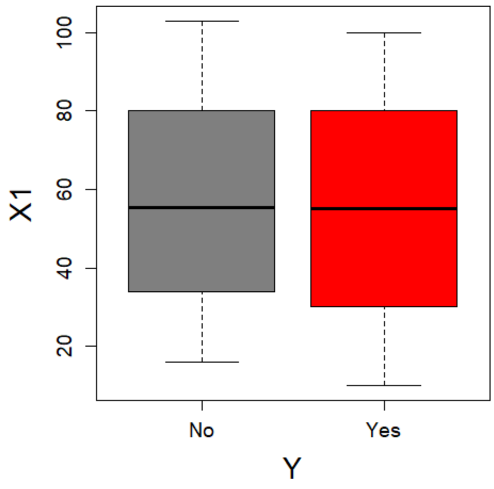
Observation number	1	2	3	4	5	6	7	8	9	10
Value of predictor X1	10	20	30	40	50	60	70	80	90	100
Value of response X2	2	1	3	4	4	7	6	10	7	9

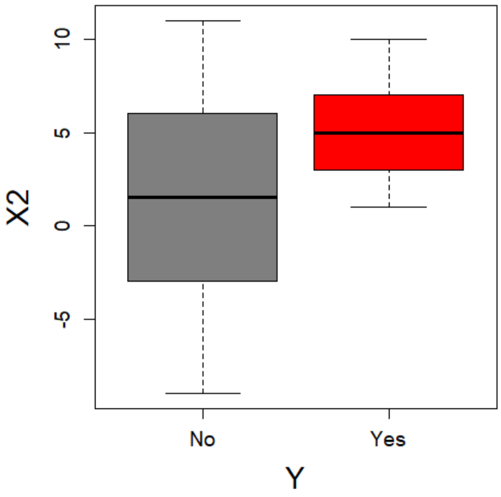
Observations when response Y = ”No”:

Observation number	1	2	3	4	5	6	7	8	9	10
Value of predictor X1	16	21	34	38	49	62	78	80	100	103
Value of response X2	-6	-9	-3	4	-1	5	-3	11	6	6

A scatter plot and box plots have been created to represent the data:

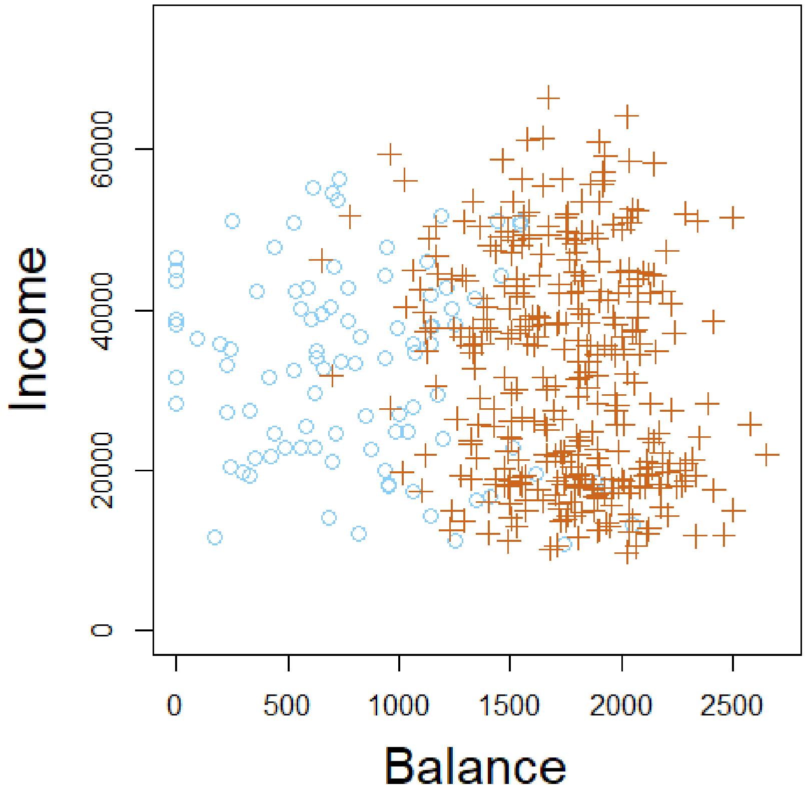






Observe these plots carefully and understand how they were created. Based on these plots, which of X1 and X2 seems better to include in a logistic regression model to predict Y?

2. The scatter plot below corresponds to the removal of data for 99% of non-defaulted individuals. The blue points represent the non-defaulted cases and the brown crosses represent the defaulted individuals. What predictions do you expect for the default probability at balance = \$500, balance = \$1000, balance = \$1500 and balance = \$2000, based on a logistic regression model that uses balance as a single predictor to predict default? Draw a function that you expect to be close to the estimated default probability curve!



3. The logistic regression model using that corresponds to the above scatter plot has the following R output:

```
Call:
glm(formula = default ~ balance, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2277   0.0489   0.1607   0.3726   2.2302

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.9439716  0.7432864  -7.997 1.28e-15 ***
balance      0.0054321  0.0005759   9.433  < 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: 456.15  on 428  degrees of freedom
Residual deviance: 196.85  on 427  degrees of freedom
AIC: 200.85

Number of Fisher Scoring iterations: 6
```

Recall that in logistic regression, the estimated probability of a “case” for given values of the predictors can be computed as follows:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}$$

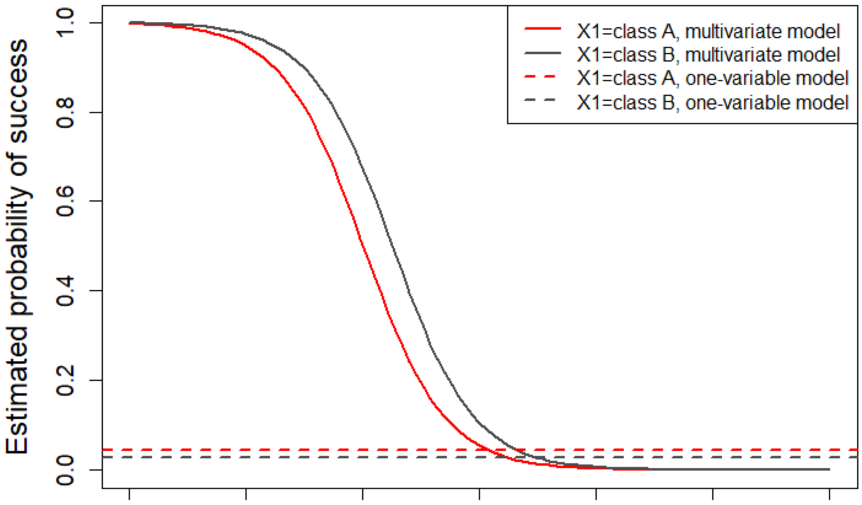
Plug the coefficient estimates from the above R output into this formula and write down the resulting equation. Using this equation, make a prediction of the probability of default at the following balance values:

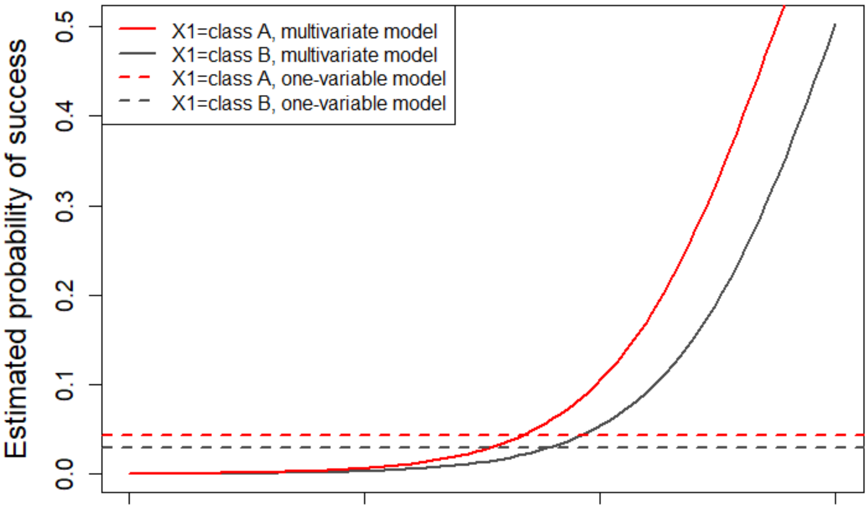
a) balance = \$1000

b) balance = \$2000.

Are the predictions close to the values that you anticipated in exercise 2?

4. Which of the figures below showing estimated probability curves from logistic regression corresponds to a model with confounding?





5. Feedback quiz (optional): Go to www.menti.com and use the code 38 50 70. Note that using this code, you can both give feedback about today's lecture and suggest topics to focus on during the classes next week.