Solutions to exercises for exercise class 6a in MMS075, Feb 25, 2020

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

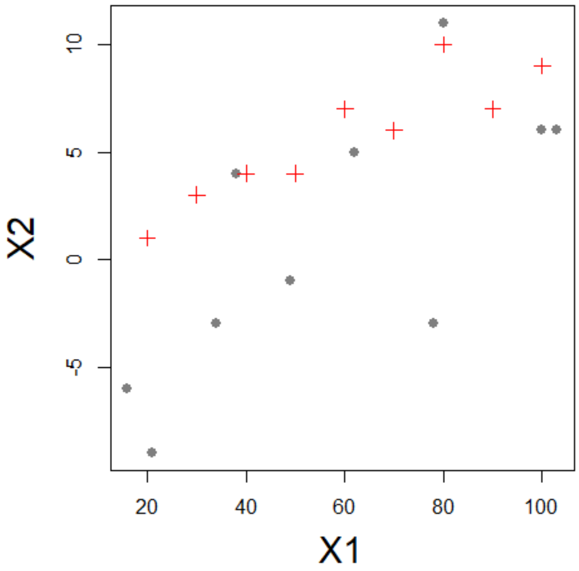
Observations when response Y = ”Yes”:

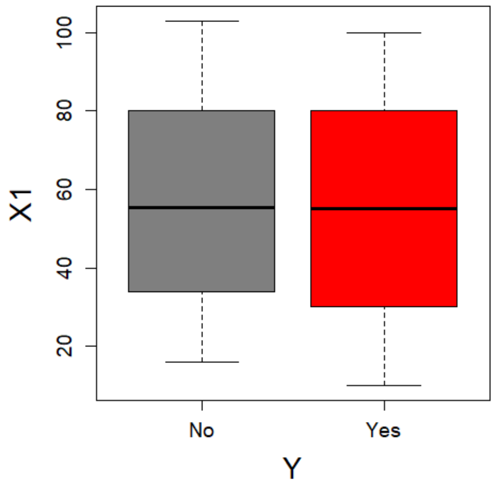
Observation number	1	2	3	4	5	6	7	8	9	10
Value of predictor X1	10	20	30	40	50	60	70	80	90	100
Value of response X2	2	1	3	4	4	7	6	10	7	9

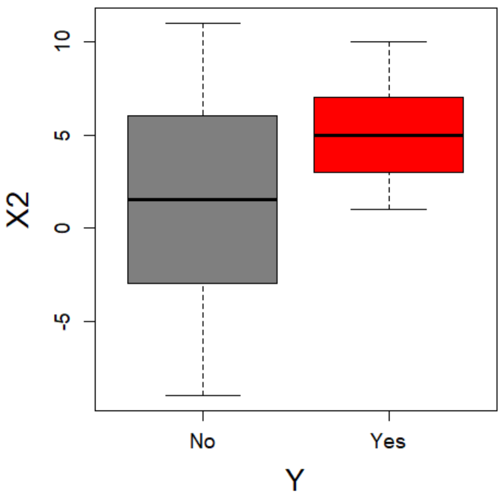
Observations when response Y = ”No”:

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:



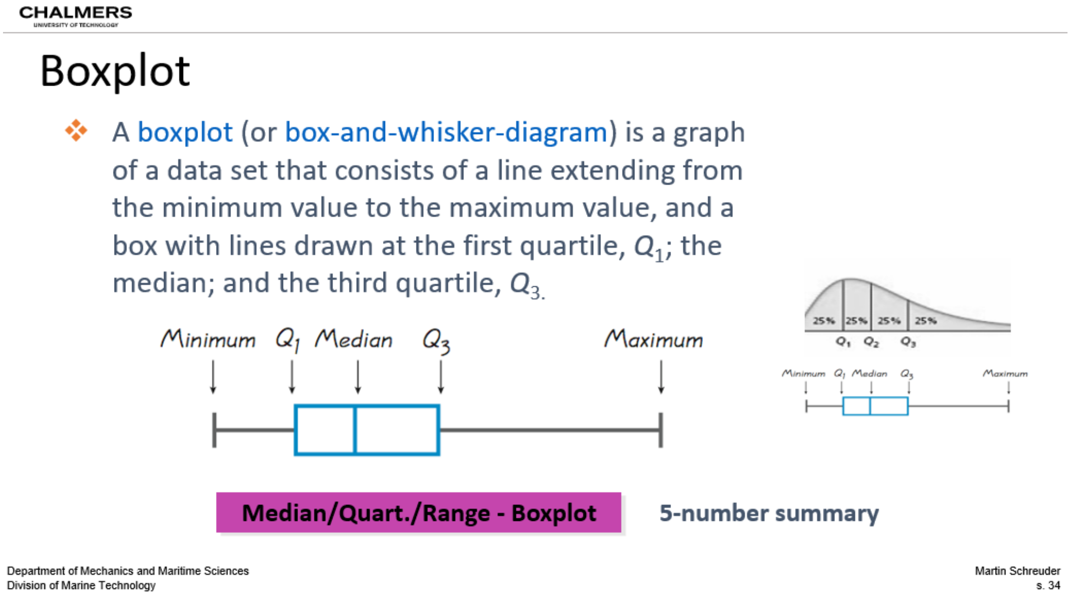




Observe these plots carefully and understand how they were created. Based on these plots, which of X1 and X2 seems better to include in a logistic regression model to predict Y?

The scatter plot is plotting the X1 and X2 values of all 20 observations listed in the two tables above. The (X1,X2) pairs given in the first table are marked by red crosses while the (X1,X2) pairs in the second table are marked by grey points. The reason for this is that the predictor pairs in the first table are those for which a "Yes" response was observed and the predictor pairs in the lower table are those for which a "No" response was observed (as specified in the text of the exercise).

There are four box plots in the figure above, each one contains the 5-number summary (i.e. minimum, Q1, median, Q3, maximum) for one of the rows of the above tables. The red ones show the graphs corresponding to the upper table and the grey ones show the graphs for the lower table. The construction of box plots was also explained in SJO915:



CHALMERS
UNIVERSITY OF TECHNOLOGY

Boxplot

- A boxplot (or box-and-whisker-diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

Minimum
Q_1
Median
Q_3
Maximum
25%
25%
25%
25%
Q_1
Q_2
Q_3
Minimum
Q_1
Median
Q_3
Maximum

Median/Quart./Range - Boxplot

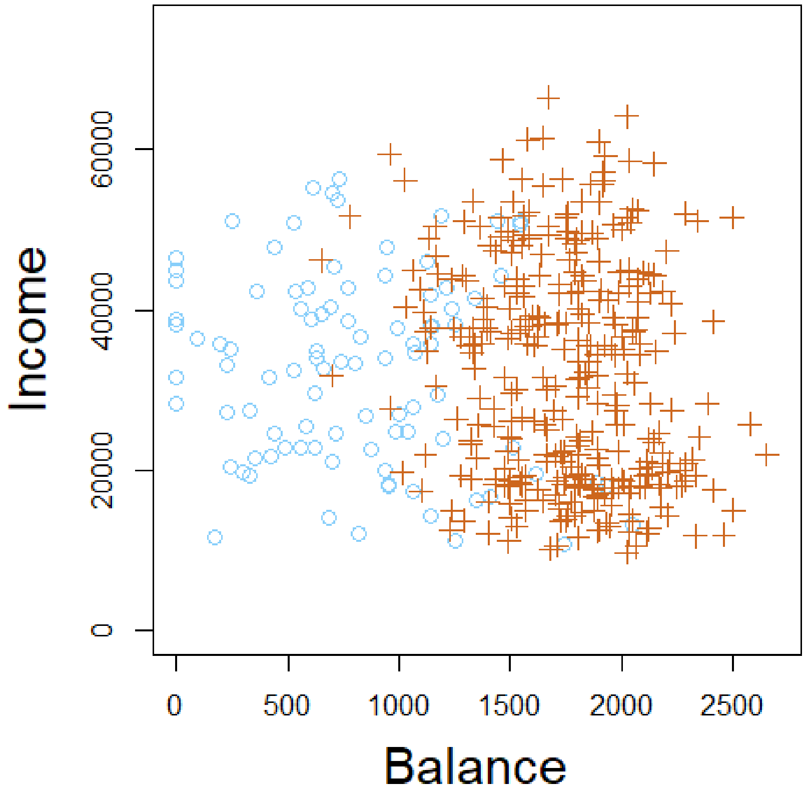
5-number summary

Department of Mechanics and Maritime Sciences
Division of Marine Technology

Martin Schreuder
s. 34

Based on all these plots, X2 seems to be a better predictor, because it seems to provide a clearer separation between "Yes" and "No" values of Y. In the scatter plot, one can observe a tendency of red crosses (representing "Yes" values for Y) being higher up than grey points (representing "No" values for Y), while it is much less clear whether there is a similar separation horizontally, i.e. along the values of the X1 variable.

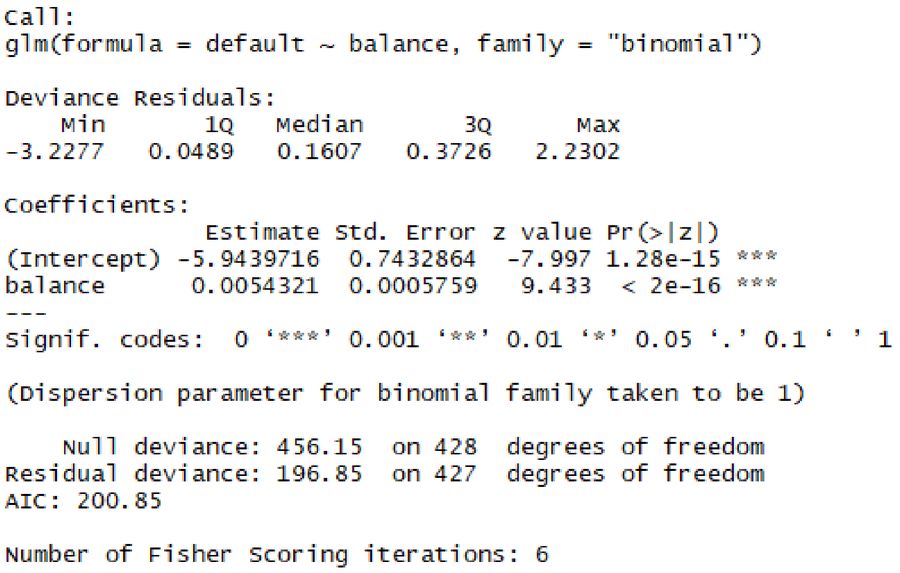
2. The scatter plot below corresponds to the removal of data for 99% of non-defaulted individuals. The blue points represent the non-defaulted cases and the brown crosses represent the defaulted individuals. What predictions do you expect for the default probability at balance = \$500, balance = \$1000, 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!



For each value along the x-axis, i.e. each given balance value, we would expect an estimated probability of default that would correspond to the relative frequency of defaulted cases among people with approximately the given balance value. For example, we see around 15 points around Balance=$500 and 1-2 of them is a brown cross, i.e. corresponds to a defaulted individual – therefore, we expect about 7-13% estimated probability of default for this balance value. For Balance=$1000, we see about 15 points of which 5-6 are brown → we expect about 33-40% estimated probability. For Balance=$1500, the brown crosses are clearly dominating among all such points, maybe 80-90% of points with this balance value are brown. Finally, for Balance=$2000, essentially all points are brown, so we expect the probability to be very close to 1.

Therefore, we expect that the estimated probability curve will be an S-shaped curve (as usual for logistic regression) that passes through the identified probability values or intervals.

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.

In this case, there is a single variable, balance, which can be taken to be X_1 in the above formula. Additionally, we know the value of the intercept. Therefore, the formula for predicting probabilities is as follows:

$$\hat{p}(X) = \frac{e^{-5.94+0.0054\times \mathrm{balance}}}{1+e^{-5.94+0.0054\times \mathrm{balance}}}.$$

Using this equation, make a prediction of the probability of default at the following balance values:

a) balance = $1000

b) balance = $2000.

The two predicted probabilities are:

$$\frac{e^{-5.94+0.0054\times 1000}}{1+e^{-5.94+0.0054\times 1000}} = 0.37 = 37\%$$

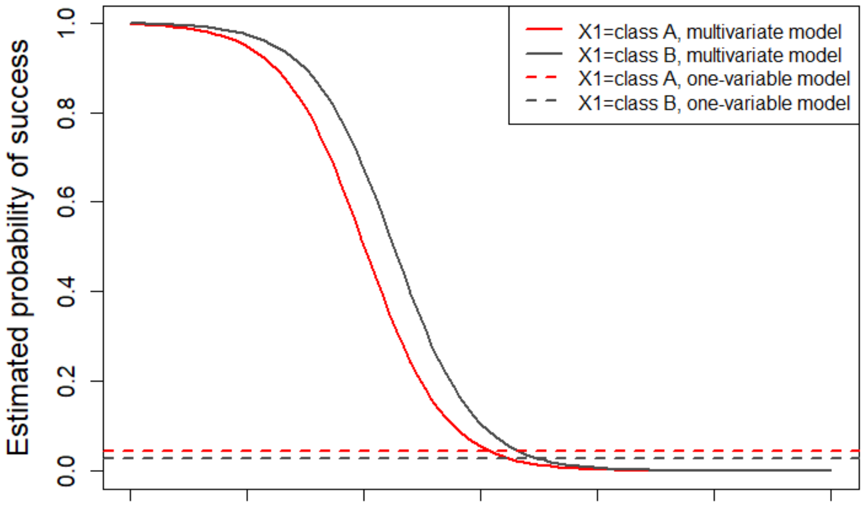
and

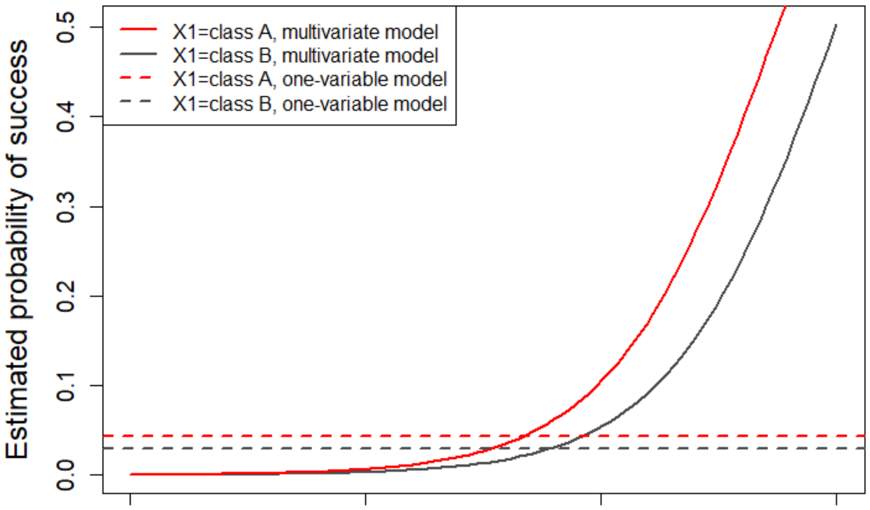
$$\frac{e^{-5.94+0.0054\times 2000}}{1+e^{-5.94+0.0054\times 2000}} = 0.99 = 99\%$$

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

For balance=$1000, we expected 33-40% and for balance=$2000, we expected a probability very close to 1, so the predictions in parts a) and b) are very close to the anticipated values.

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





The solution is described in Exercise class 6b – exercises with solutions.pdf, see Canvas.

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.