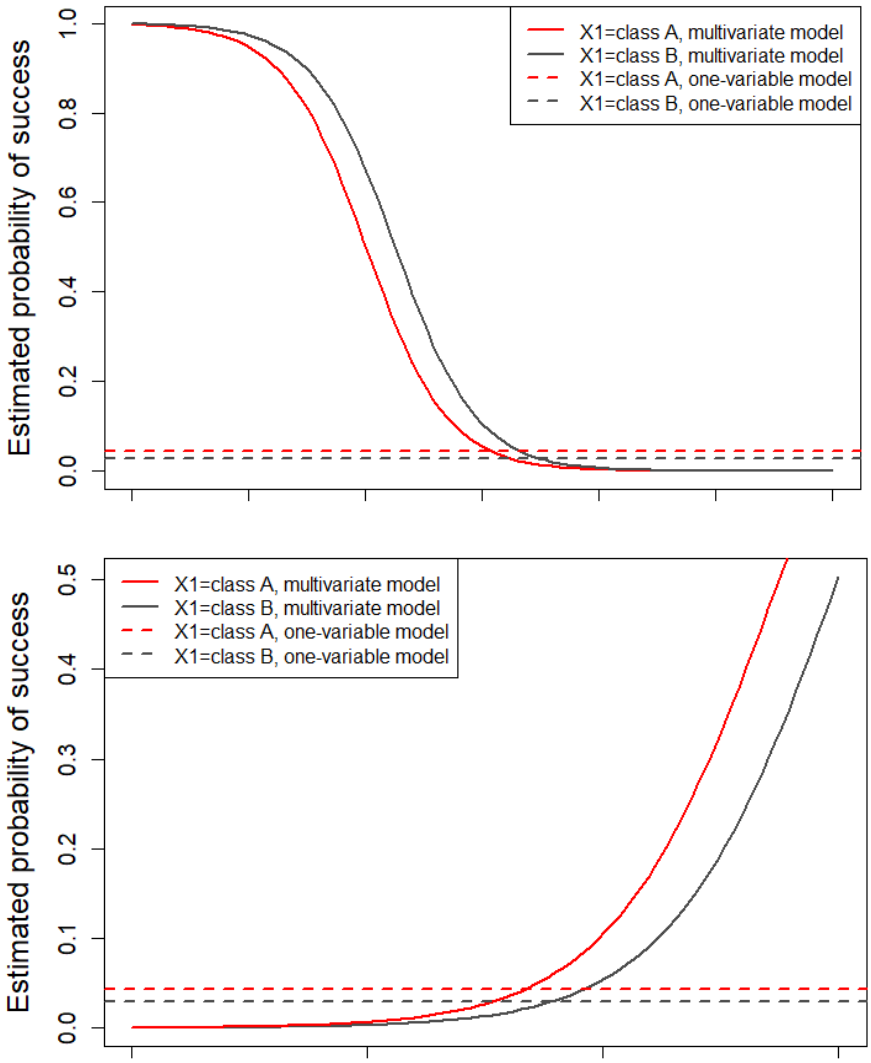
**Exercises for exercise class 6b in MMS075, Feb 26, 2020**

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




How would you describe the effect of changing the value of X1 from class A to class B in the different models?

2. An air conditioning company collects information about factors that affect the probability of installing central air conditioning. As a first step, they analyze the **Housing** dataset in R and consider the following variables as predictors: **price**, denoting the sale price of a house ($), **gashw** indicating whether the house uses gas for hot water heating (yes/no), and **stories** indicating the number of stories excluding basement (numerical value). Their analysis in R returns the following summary output:

```
Call:
glm(formula = airco ~ gashw + stories + price, family = "binomial",
    data = Housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7904  -0.7341  -0.5198   0.6765   3.1124

Coefficients:
              Estimate   Std. Error z value Pr(>|z|)
(Intercept) -4.275690040  0.387730993 -11.027  < 2e-16 ***
gashwyes    -3.689570472  1.086346957  -3.396 0.000683 ***
stories      0.294471594  0.130692028   2.253 0.024248 *
price        0.000043195  0.000005342   8.085  6.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 681.92  on 545  degrees of freedom
Residual deviance: 532.87  on 542  degrees of freedom
AIC: 540.87

Number of Fisher Scoring iterations: 6
```

**Using this output, do the following tasks:**

a) Specify the equations predicting the probability that a house with 2 stories has air conditioning, depending on its price and whether it uses gas for hot water heating!

b) Estimate the probability that a house with a sale price of $90000 and 2 stories that does not use gas for hot water heating has central air conditioning!

In addressing parts a) and b), remember that the equation for estimating the probability of a "case" in logistic regression is as follows:

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

3. The air conditioner company considered in exercise 2 investigates the different factors individually, to understand how much we know if there is only limited information available. The coefficients from the three single variable models are given in the R outputs below. Comparing these with the output for the multivariate model, is there a sign of confounding?

**Output 1:**

```
            Estimate Std. Error z value           Pr(>|z|)
(Intercept) -0.70758    0.09316  -7.595 0.0000000000000308 ***
gashwyes    -2.47048    1.02460  -2.411             0.0159 *
```

**Output 2:**

```
            Estimate Std. Error z value        Pr(>|z|)
(Intercept)  -2.1371     0.2350  -9.094         < 2e-16 ***
stories       0.7276     0.1114   6.533 0.0000000000643 ***
```

**Output 3:**

```
              Estimate  Std. Error z value Pr(>|z|)
(Intercept) -3.76328283  0.34599029  -10.88   <2e-16 ***
price        0.00004223  0.00000459    9.20   <2e-16 ***
```

4. Show that in case of binary response, the alternative formulation of the multinomial logistic regression model (see equation below) is indeed the same as the usual (i.e. binomial) logistic regression model

$$\log\left(\frac{\Pr(Y=k|X)}{\Pr(Y=s|X)}\right) = \beta_{0,k,s}+\beta_{1,k,s}X_1+\beta_{2,k,s}X_2+\cdots+\beta_{p,k,s}X_p$$

5. The **Mode** dataset in the **Ecdat** library in R contains data about travel modes: the estimated cost and time of car, carpool, bus or rail for different trips and the decision of which travel mode is chosen for the trip. We want to understand how the decision depends on the various parameters, in particular, what influences people to choose the different alternatives instead of driving a car. Therefore, we fit a multinomial logistic regression model with "car" as reference level, see the R commands and the output below.

```
> Mode$TravelMode=relevel(Mode$choice,ref="car")
> multinom(TravelMode~.-choice,data=Mode)
# weights:  40 (27 variable)
initial  value 627.991346
iter  10 value 394.826983
iter  20 value 360.160469
iter  30 value 342.077240
final  value 342.073501
converged
Call:
multinom(formula = TravelMode ~ . - choice, data = Mode)

Coefficients:
        (Intercept)  cost.car cost.carpool   cost.bus  cost.rail
carpool   -4.106305 0.6360771   -0.4472983 0.04505779 -0.5501033
bus       -4.788587 0.8461170    0.2162670 0.01013198 -0.5276687
rail      -4.299980 0.8900743    0.2058304 0.56600590 -1.2752642

  time.car time.carpool     time.bus    time.rail
0.12366772  -0.06922734  0.007442305 -0.027732676
0.02285042   0.09461637 -0.107750616 -0.006393055
0.03639195   0.07297203 -0.018132194 -0.075282632
```

Based on this output, how are the following changes expected to affect the probability of choosing various alternatives:

a) Increasing the cost of car trips;
b) Decreasing the time of car trips;
c) Increasing the price of train tickets;
d) Building new, faster train connections;
e) Increasing fuel price?

6. Consider a simple model A and a very flexible model B. Evaluate whether the following inequalities always hold:
   a. Training MSE for model B ≤ Training MSE for model A;
   b. Test MSE for model B ≤ Test MSE for model A.

7. 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.