MAST90139: Statistical Modelling for Data Science
Assignment 1

Due time: 5 pm Friday April 1, 2022.

Submit your assignment to Gradescope on the subject Canvas LMS.

In a study of the predictors of domestic violence, a survey was conducted among women, aged 18 and over, who attended a general practitioner in Metropolitan Melbourne between November 1993 and February 1994, and who had been in a relationship during the previous 12 months. The "response" variable was whether or not the woman responded positively to questions about physical domestic violence or mental abuse experienced during the previous 12 months. The predictor variables considered included the following:

Age (in years)	0 = 18 – 29 1 = 30 – 49 2 = 50 – 64 3 = 65+
Current marital status	1 = married 2 = de Facto 3 = divorced 4 = separated 5 = widowed 6 = never married
Married/de Facto more than once	0 = yes 1 = no
Smoker	0 = yes 1 = no
Drinker (alcohol)	0 = $<$ 8 drinks/week 1 = $\geq$ 8 drinks/week
Family member's use of alcohol cause for concern when growing up	0 = no 1 = yes
Number of years in formal education	0 = $\leq$ 6 1 = 7 – 11 2 = $\geq$ 12
Region	1 = north 2 = east 3 = south 4 = west

The responses from 1316 women, for whom there were no missing values, have been stored in file (`domviolence.csv`) which can be read into `R` using the command

```
domviolence <- read.csv(file="C:/subjects/MAST90139/data/domviolence.csv")
```

provided that `domviolence.csv` is on the path `C:/subjects/MAST90139/data/`. The file can be found in the subject Canvas Module section.

The variable `dv` is the response variable: 0 = no; 1 = yes.

1. Carry out a series of data analysis using `glm()`, `anova()`, `summary()` and `step()` etc. to argue that the "best" logistic model is of the form

$$\begin{aligned}\text{logit}(\theta) \;=\; & \beta_0 + \beta_1 \times \texttt{ms.2} + \beta_2 \times \texttt{ms.3} + \beta_3 \times \texttt{ms.4} + \beta_4 \times \texttt{ms.5} + \beta_5 \times \texttt{ms.6} \\ & + \beta_6 \times \texttt{smok} + \beta_7 \times \texttt{falc} + \beta_8 \times \texttt{reg.2} + \beta_9 \times \texttt{reg.3} + \beta_{10} \times \texttt{reg.4} \\ & + \beta_{11} \times \texttt{age} + \beta_{12} \times \texttt{educ} + \beta_{13} \times \texttt{ms.2 : falc} + \beta_{14} \times \texttt{ms.3 : falc} \\ & + \beta_{15} \times \texttt{ms.4 : falc} + \beta_{16} \times \texttt{ms.5 : falc} + \beta_{17} \times \texttt{ms.6 : falc}\end{aligned}$$

   where estimates of all β_j in the above model can be obtained from the arguing process. Present your work and results in no more than 300 words (excluding `R` outputs; the `R` outputs should be presented as an appendix). [15]

   Some suggestions:

   (a) You can start the process by fitting a `model0` that includes all predictor variables as having main effects only. Then use `anova(model0, test="Chi")` alike commands repeatedly to remove any non-significant predictors. Note categorical predictors, e.g. `reg`, should be treated as factors here and be included in the model in the form of e.g. `factor(reg)`. Denote as `model1` after this process. Note `model1` may also be obtained by using commands `model0.5 = step(model0)` and `anova(model0.5, test="Chi")` to remove any non-significant predictors.

   (b) Even though `age` and `education` are observed as categorical variables, they are actually numerical variables. Thus, replace `factor(age)` and `factor(educ)` with `age` and `educ` in `model1` to get `model2`. Then fit `model2` to see whether it is significantly different from `model1` or not in terms of goodness of fit. Accept `model2` if it is not significantly different from `model1`, because `model2` is simpler than `model1` in terms of model complexity.

   (c) Expand `model2` to `model3` that contains all predictors in `model2` plus all their first-order interaction terms. Perform model comparisons by `anova(...)` to simplify `model3`. Alternatively, use `step(model3)` to select the stepwise "best" model. Denote the resultant model as `model4` and fit it. Then use `anova(model4, test="Chi")` and `summary(model4)` to see whether it can be further simplified. You can use this eventually obtained model as the "best" model.

2. Suppose the "best" model is indeed the one shown above. Interpret this "best" model in terms of odds ratios for each predictor in this mode. There will be many odds ratios to be presented here. So it should be helpful to use tables or diagrams in your description. Specifically, you can answer the question in following form: [15]

Marital status: Find the odds ratios involving `ms` by completing the following table

Odds ratios, at various levels of `falc`	`falc=0`	`falc=1`
for `ms=1` vs. `ms=1`		
for `ms=2` vs. `ms=1`		
for `ms=3` vs. `ms=1`		
for `ms=4` vs. `ms=1`		
for `ms=5` vs. `ms=1`		
for `ms=6` vs. `ms=1`		

Then provide interpretations on two representative odds ratio values of your choice in the above table.

Smoking: Find the odds ratio value in regard to `smoking` woman and interpret it.

Family alcohol: Find the odds ratios involving `falc` by completing the following table

Odds ratios, at various levels of `ms`	`ms=1`	`ms=2`	`ms=3`	`ms=4`	`ms=5`	`ms=6`
for `falc=0` vs. `falc=0`						
for `falc=1` vs. `falc=0`						

Then interpret two representative odds ratio values of your choice in the above table.

Region: Find the odds ratios involving `region` by completing the following table

	north	east	south	west
Ratio of odds (OR) for each `region` vs. `north`				

Then interpret two representative odds ratio values of your choice in the above table.

Age: Find the odds ratio associated with `age` and interpret it.

Education: Find the odds ratio associated with `education` and interpret it.

`Total marks = 30`