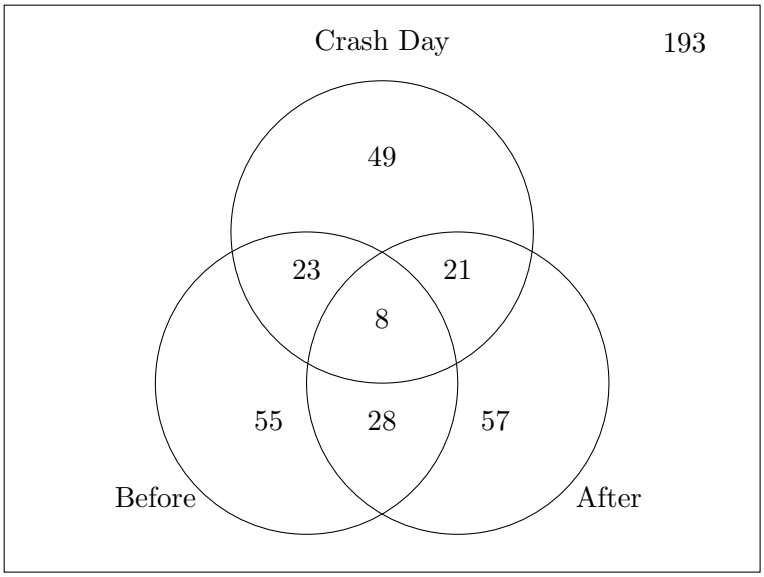
STA303H1S/STA1002HS Assignment 2
Due on 15th August, 2020 11:59 PM in Quercus
All relevant work must be shown for credit.

Note: In any question, if you are using `R`, all `R` codes and `R` outputs must be included in your answers. You should assume that the reader is not familiar with R outputs and so explain all your findings, quoting necessary values form your outputs. Please note that academic integrity is fundamental to learning and scholarship. You may discuss questions with other students. However, the work you submit should be your own. If I feel suspicious of any assignment (e.g. if your work doesn't appear to be consistent with what we have discussed in class), I will not mark the assignment. Instead, I will ask you to present your work in my office and your grade will be assigned based on your presentation. Assignments can be hand written but the `R` codes and outputs should be printed.



1. The numbers in the Figure above indicate the weather (overcast or not) of 434 location-matched triplets of days, one day on which a traffic accident took place, and two control days without an accident (the day before the accident and the day after the accident). This dataset could be analyzed as a 1:2-matched case-control. The Venn diagram presentation of the data is rather unconventional. In matched case-control studies the data could alternatively be presented in 2×2-tables

	exposed	unexposed
case	a_i	b_i
control	c_i	d_i

for each matched set $i = 1, \ldots, 434$. We denote $n_i = a_i + b_i + c_i + d_i$.

   (a) **[5 Marks]** We note that there are six types of location-specific 2×2-tables with the same exposure-case configuration. List these tables (i.e. different combinations of the numbers a_i, b_i, c_i and d_i) and their counts.

(b) **[3 Marks]** The null hypothesis assumes that there is no relationship between being a case and being exposed. Under the null hypothesis the distribution of the cell count a_i conditional on the row and column marginals is hypergeometric. Find $E(a_i \mid a_i + c_i)$ and $\text{Var}(a_i \mid a_i + c_i)$ under the null.

(c) **[3 Marks]** Test the null hypothesis of no association between weather and accidents using the Cochrane-Mantel-Haenszel (CMH) test statistic, given by

$$\frac{\left(\sum_{i=1}^{434} a_i - \sum_{i=1}^{434} E(a_i|a_i + c_i)\right)^2}{\sum_{i=1}^{434} \text{Var}(a_i \mid a_i + c_i)},$$

which is asymptotically distributed as χ^2 with one degree of freedom.
Note: $\chi^2_{0.95}(1) = 3.84$.

(d) **[4 Marks]** Let's assume we have the following table is a triplet specific contingency table from a 1:2 matched case control study.

	exposed	unexposed	Total
case	a	b	1
control	c	d	2
Total	$a+c$	$b+d$	3

The odds of being exposed in the case groups is θ times of the odds of being exposed in the control group. Also, let's assume that $P(a = 1) = \dfrac{\theta\Omega}{1+\theta\Omega}$ and $P(c = 1) = \dfrac{\Omega}{1+\Omega}$
Show that, $P(a = 1 \mid a + c = 1) = \dfrac{\theta}{2+\theta}$
(Hint: The 2 in the denominator comes from 2 controls).

2. For this question you have to simulate a dataset.

(a) **[2 Marks]** Perform the following simulations.

- Generate 500 random values from $X_1 \sim \text{Uniform}[0,1]$, $X_2 \sim \text{Uniform}[0,1]$, $X_3 \sim \text{Uniform}[0,1]$, $X_4 \sim \text{Uniform}[0,1]$, $X_5 \sim \text{Uniform}[0,1]$
- Generate, $f(\boldsymbol{X}) = 4[\sin(\pi x_1 x_2) + 8(x_3 - 0.5)^3 + 1.5x_4 - x_5 - 0.77]$. Here, $\pi = 3.14.....$
- Generate $Y \sim \text{Bernoulli}\left(p(\boldsymbol{X}) = \dfrac{\exp(f(\boldsymbol{X}))}{1+\exp(f(\boldsymbol{X}))}\right)$

(b) **[5 Marks]** Fit a logistic regression where Y is the outcome and $X_1, X_2, ..., X_5$ are the predictors. Show the coefficients table. Produce the ROC curve. State the AUC value and interpret.

(c) **[5 Marks]** Now instead of using the original $X_1, X_2, ..., X_5$ as predictors, transform the variables in such a way that they resembels the individual terms in $f(\boldsymbol{X})$. That, is create new variables from $X_1, X_2, ..., X_5$ in such a way that $f(\boldsymbol{X})$ is transformed to a linear predictor. Now run a logistic regression using the new variables. Show the coefficients table. Produce the ROC curve. State the AUC value.
(**Hint:** You have to create 4 new variables from $X_1, X_2, ..., X_5$)

(d) **[3 Marks]** Compare your results in (b) and (c): how did your coefficients and AUC change from (b) to (c)? Explain why you think this happened.

3. For this problem you need to load the NHANES dataset using the following command

```
## If the package is not installed then use ##
install.packages('NHANES') ## And install.packages('tidyverse')
library(tidyverse)
library(NHANES)
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12"
& NHANES$Age > 17,c(1,3,4,8:11,13,25,61)])
small.nhanes <- small.nhanes %>%
group_by(ID) %>% filter(row_number()==1)
```

This is data collected by US National Center for Health Statistics (NCHS). The preceeding codes creates a small dataset of the original NHANES dataset. With this dataset answer the following questions,

(a) **[5 Marks]** Randomly select 500 observations from the data. For this selection use your student ID as seed. Fit a logistic regression to predict smoking status (variable `SmokeNow`), using all the other variables (excluding `ID`). Explain your results in few sentences.

(b) **[5 Marks]** Perform a model selection procedure based on stepwisemethods (both AIC and BIC) and also using elastic-net. Do they select the same model.

(c) **[5 Marks]** Perform an internal validation using cross-validation. Explain your results.

(d) **[5 Marks]** Construct the Receiver operating characteristic (ROC) curve. Calculate the area under the curve (AUC). How would you interpret the AUC.

(e) **[5 Marks]** Predict the probabilities for the remaining 310 observations. Calculate the deciles for the predicted probabilities. Does the observed and the predicted probabilities differ for the deciles?

(f) **[5 Marks]** For this problem you need to load the NHANES dataset but keeping all the rows of the data. You can use the following commands

```
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12"
& NHANES$Age > 17,c(1,3,4,8:11,13,25,61)])
```

Fit a mixed effects logistic regression. Only consider random intercept for subject ID. Use all the available predictors.