Statitical Models Homework 1

Exploratory Data Analysis

After setting up a virtual environment suitable for carrying out the scientific analysis, we load some useful packages to improve the capabilities of the analysis software. In particular, as suggested we load *ISLR*, *ISLR2*, *ROCR* and *tidyverse*. For the purposes of this research, we will also use *e1071*, *mice*, *caret*, *corrplot*, *patchwork*

The first step in our analysis is to load the data contained in *chd.csv*. With the function *glimpse()* obtain an overview of the data.

```
Rows: 4,238
Columns: 13
$ sex       <chr> "Male", "Female", "Male", "Female", "Female", "Female", "Fem~
$ age       <dbl> 39, 46, 48, 61, 46, 43, 63, 45, 52, 43, 50, 43, 46, 41, 39, ~
$ education <dbl> 4, 2, 1, 3, 3, 2, 1, 2, 1, 1, 1, 2, 1, 3, 2, 2, 3, 2, 2, 2, ~
$ smoker    <dbl> 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, ~
$ cpd       <dbl> 0, 0, 20, 30, 23, 0, 0, 20, 0, 30, 0, 0, 15, 0, 9, 20, 10, 2~
$ stroke    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ HTN       <dbl> 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, ~
$ diabetes  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ chol      <dbl> 195, 250, 245, 225, 285, 228, 205, 313, 260, 225, 254, 247, ~
$ DBP       <dbl> 70.0, 81.0, 80.0, 95.0, 84.0, 110.0, 71.0, 71.0, 89.0, 107.0~
$ BMI       <dbl> 26.97, 28.73, 25.34, 28.58, 23.10, 30.30, 33.11, 21.68, 26.3~
$ HR        <dbl> 80, 95, 75, 65, 85, 77, 60, 79, 76, 93, 75, 72, 98, 65, 85, ~
$ CHD       <chr> "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No", "No"~
```

From the output we observe that the variables sex and CHD are represented as character variables, while the remaining predictors are stored as double-precision numerical values (real numbers). This suggests that most of the variables are already in a format suitable for analysis, although the character variables may have to be converted to factors depending on the modelling approach.

First of all, let us proceed with some basic checks on the data. The wise thing to do first is always to look for missing values.

```
      sex       age education    smoker       cpd    stroke       HTN  diabetes
        0         0       105         0        29         0         0         0
     chol       DBP       BMI        HR       CHD
       50         0        19         1         0
```

Since we have a small data set to use, we think it is best to use an imputator to replace the missing values with the actual expected values. We will now make a copy of our data set, as we will apply the imputer function.

We then create *data_simple*, where we impute missing values with basic metrics such as mean, median (if the data distribution is skewed) for continuous data and mode for the categorical ordinal *education*.

Before we apply the *simple_imputer()* we first want to evaluate the skewness of continuous numeric variables.

```
      cpd      chol       BMI        HR
1.2470206 0.8707979 0.9812762 0.6440255
```

We can asses moderate skewness in all variables, except for *cpd*, which proves to be significantly skewed. Therefore, the median will be a preferable metric for imputing values.

We will use mode to complete missing values for *education*. Mode imputation is commonly used for the categorical variables but has some limitations. It is generally reasonable when the distribution is highly imbalanced and the number of missing values is relatively small, making it acceptable to substitute with the most frequent category. However, mode imputation ignores the ordinal nature of the variable and does not take into account potential correlations with other variables. In our case, the number of missing values is small, so these limitations are unlikely to significantly affect the results. The output below confirms that the missing values were correctly replaced by the imputer function.

```
      sex       age education    smoker       cpd    stroke       HTN  diabetes
        0         0         0         0         0         0         0         0
     chol       DBP       BMI        HR       CHD
        0         0         0         0         0
```

Now that we have handled the missing values, we will search for possible significant relationships between variables. For data exploration we'll plot only information from the data_simple() copy of the data.

The first step will be to draw a heat map of the Pearson correlation matrix between the variables. This will help us identify potential multicollinearity issues between predictors and assess the strength of linear relationships between each predictor and the response variable. Since correlation requires numerical inputs, the categorical variables *CHD* and *SEX* must first be converted into binary numerical variables (e.g., 1 = Yes, 0 = No). Once transformed, they can be included in the correlation matrix. From Figure 1, it can be seen that there are no strong linear relationships between the predictors and the response variable. The correlation with CHD is weak for all characteristics, with *AGE* (r = 0.23), *HTN* (r = 0.18) and *DBP* (r = 0.15) showing the highest associations. These weak positive correlations suggest limited linear discriminative power, although they could still contribute useful information in a multivariate model. As expected, positive correlations were also found between *SMOKER* and *CPD* (r = 0.77), *HNT* and *DBP* (r = 0.61), *DBP* and *BMI* (r = 0.38). This indicates potential multicollinearity, which should be monitored during model fitting, particularly in the context of regression analysis, where correlated predictors may inflate the variance in coefficient estimates and reduce interpretability.

To further investigate the discriminatory power of continuous and categorical predictors, we examined the distributions of those variables that showed some correlation with CHD. Specifically, we used boxplots for continuous predictors and bar plots with proportions for categorical ones. As shown in Figures 2 and 3, the distribution of CHD cases varies across different levels of the predictors *AGE*, *DBP*, *HTN* and *SEX*, providing additional visual evidence of a relationship between these variables and the response. These differences support the hypothesis that certain predictors may contribute useful information to CHD risk modeling. We care to

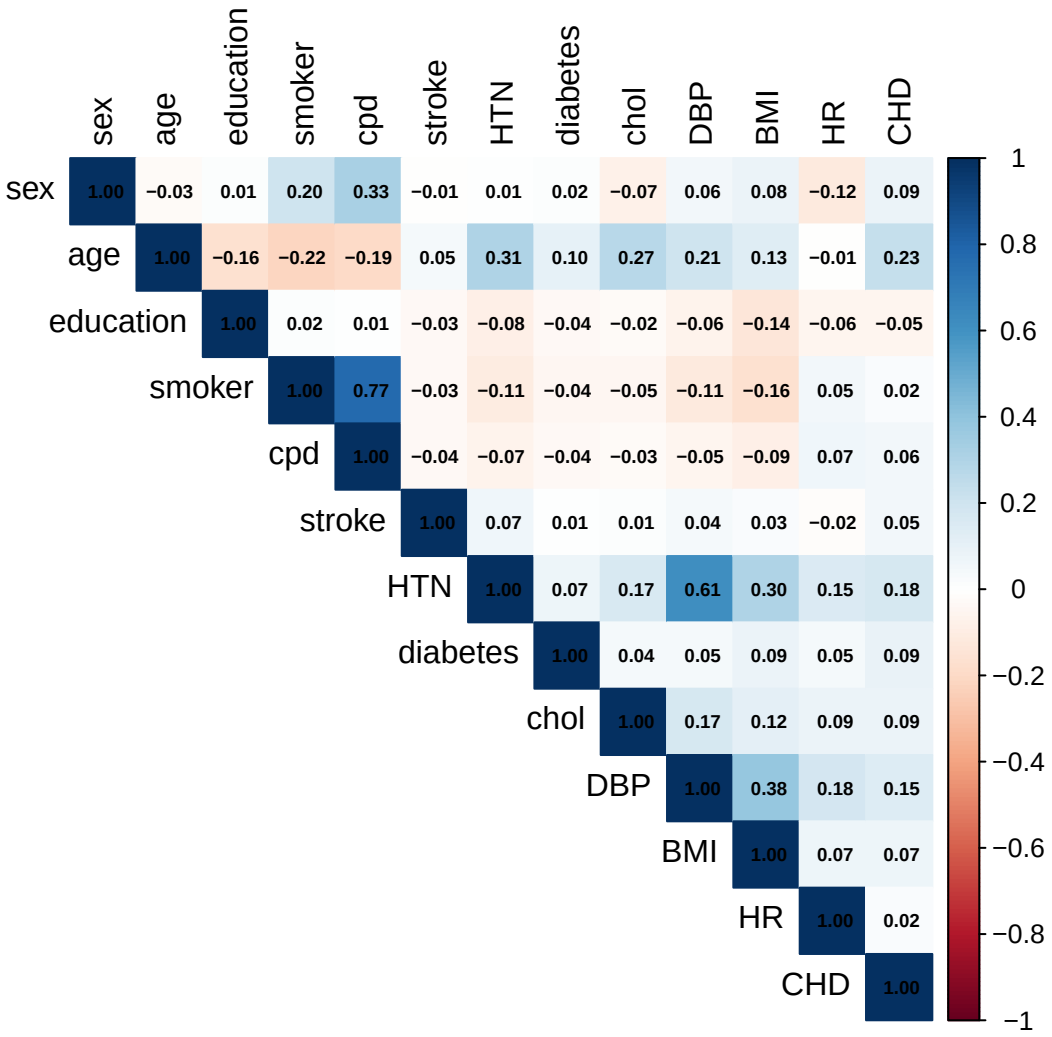

Figure 1: Correlation heatmap (including CHD)

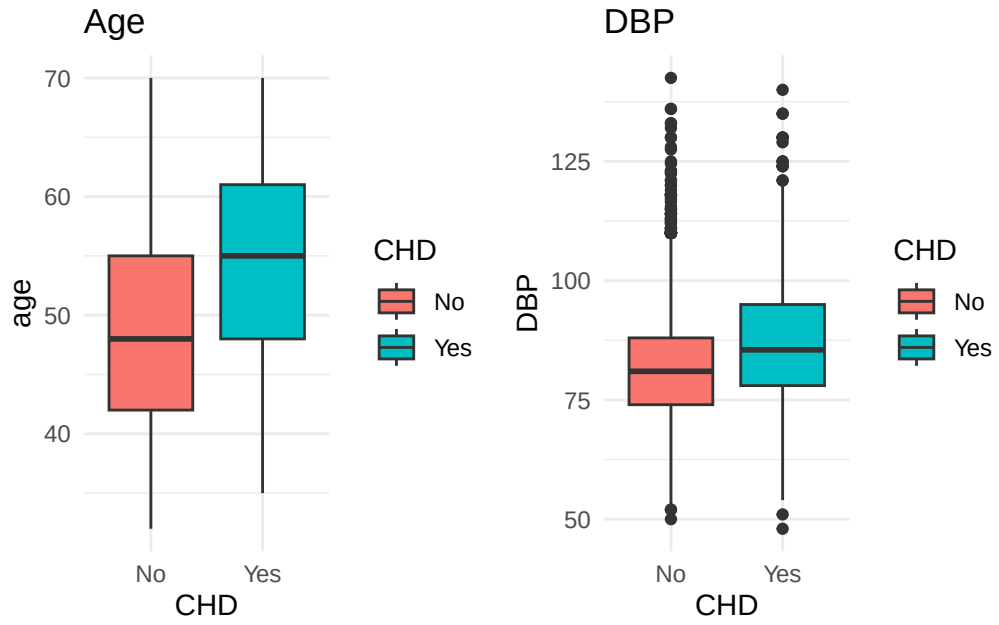

Figure 2: Distribution of continuous predictors against CHD

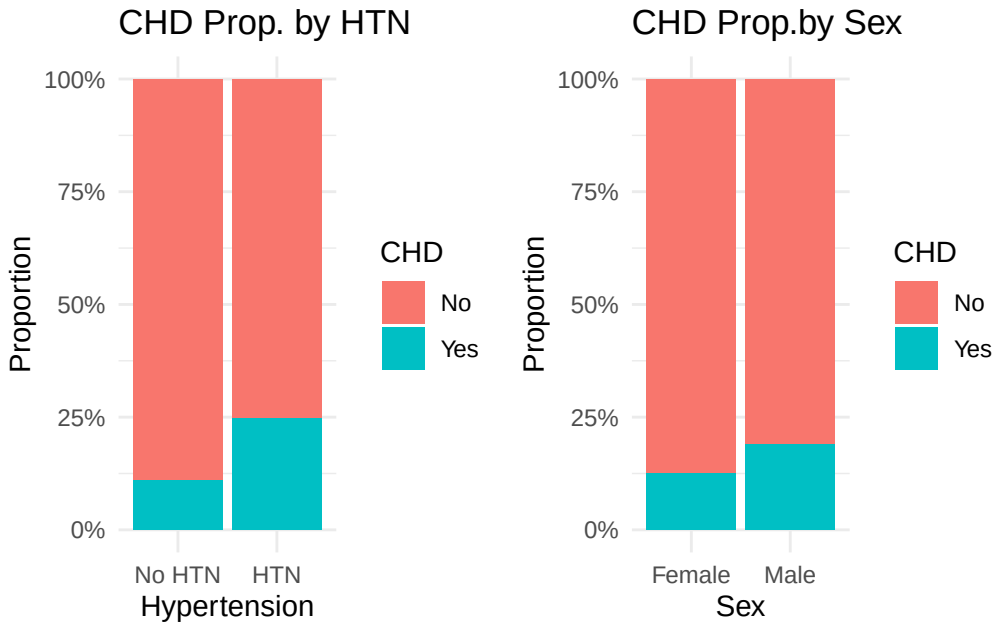

Figure 3: CHD proportions by hypertension status

We proceed with the analysis of the nature of the response variable. Since the response variable is categorical and binary we want to asses the level of balance inside the two categories by computing the proportion of each one of them on the total number of observations.

```
       No       Yes
0.8480415 0.1519585
```

The output clearly shows that the response variable is strongly skewed toward the "No" class. This means that most observations do not have the event of interest (e.g., no coronary artery disease within 10 years). This has some implications for modeling, such as the fact that standard classifiers, e.g., logistic regression, will likely favor the most representative class, and perform poorly in making predictions about the minority class. In this scenario, the Accurancy is misleading, since the prediction of the majority class alone would still provide high accuracy (e.g., 84%) even if all cases of the minority class were missing.

In addition, we must prevent the random partitioning of data into a training and test set from leading to unbalanced sets where the true proportions of the sample are not actually represented correctly. To do this we use a technique called stratified sampling, which is aimed precisely at maintaining the proportions of the categories of a variable in both subsets of the sample. In practice, we will divide the sample into two groups filtered by the two categories of the response variable (in this case "yes" and "no"). These two groups are called *strata*. We then branch out some instances from the two groups and place them in our desired training or test set in a certain proportion (we chose to put 70% of the instances in the training set and the other 30% in the test). This maintains the original proportion in both sets. In R, we use *caret::createDataPartition()* which is the function provided to do this. After that, we verify the new proportions of both the sets.

```
       No       Yes
0.8479946 0.1520054
```

```
       No       Yes
0.8481511 0.1518489
```

We fit the following GLM model:

$$\text{logit}(\text{E}(\text{CHD})) = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + \beta_3 \text{education} + ... + \beta_{12} \text{HR}$$

```
Call:
glm(formula = CHD ~ sex + age + education + smoker + cpd + stroke +
    HTN + diabetes + chol + DBP + BMI + HR, family = binomial,
    data = data_glm)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.597e+00  6.466e-01 -11.750  < 2e-16 ***
sexMale      4.398e-01  9.837e-02   4.470 7.81e-06 ***
age          7.032e-02  5.949e-03  11.822  < 2e-16 ***
education   -2.475e-02  4.536e-02  -0.546 0.585404
smoker       1.658e-02  1.423e-01   0.117 0.907209
cpd          2.129e-02  5.626e-03   3.784 0.000154 ***
stroke       1.023e+00  4.382e-01   2.334 0.019592 *
HTN          4.397e-01  1.174e-01   3.744 0.000181 ***
diabetes     8.603e-01  2.151e-01   3.999 6.37e-05 ***
chol         2.034e-03  1.019e-03   1.996 0.045911 *
DBP          1.286e-02  4.638e-03   2.772 0.005566 **
BMI          3.369e-03  1.171e-02   0.288 0.773455
HR           5.551e-05  3.857e-03   0.014 0.988516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3611.5  on 4237  degrees of freedom
Residual deviance: 3238.7  on 4225  degrees of freedom
AIC: 3264.7

Number of Fisher Scoring iterations: 5
```

The output represent the fitted logistic regression model, in which we estimate the log-odds of developing *CHD* as a function of an intercept and 12 predictor variables. The intercept (-7.597) represents the log-odds of developing *CHD* for an individual belonging to all the reference categories—such as being male, having the lowest education level, not smoking, and having no history of stroke, hypertension, or diabetes. Its strongly negative value reflects a very low baseline likelihood of developing *CHD* in this group. The estimated coefficient

for *SEX* is positive and highly significant, indicating that, all else being equal, males have a higher log-odds of developing CHD compared to females. *AGE* is also highly significant, with each additional year increasing the log-odds of developing CHD by 0.07032. In terms of the odds ratio, exp(0.07032) is approximately 1.072, indicating that each extra year of *AGE* raises the odds of developing CHD by about 7%. The variable *CPD* (cigarettes per day) is highly significant, whereas smoking status itself is not. This may suggest that the intensity of smoking provides more meaningful information about the risk of developing CHD than simply whether someone smokes or not.