Assignment II

Group2

2024-12-07

Introduction

Question 1:

Create table 1. Describe all available variables in the data. Show both, the original data and the imputed data.

Overview:

All available variables in the data are as follows:

- *agegroup:* Categories for age.
- *smoker:* Binary variable for smoker or not.
- *smokerf:* Smoker factor with levels No and Yes.
- *heightcm:* Convert height from inches to cm.
- *weightkg:* Convert weight from pounds to kg.
- *bmi:* Calculate BMI.
- *bmicat:* Categories for BMI.
- *cholmmol:* Convert cholesterol from mg/dl to mmol/l.
- *sbp10:* Categories of sbp (systolic blood pressure).
- *sbpcat:* Systolic blood pressure factor.
- *dibpat0f:* Dichotomous behavior pattern factor with levels A and B instead of 1 and 0. A classification system where individuals are grouped into one of two distinct categories based on their behavioral traits
- *arcus0:* Corneal arcus factor which is caused by lipid deposits in the cornea. It's presence may indicate high cholesterol levels and increased risk of heart disease.
- *chd69f:* Coronary heart disease factor.

Then, we have created tables for the original data. In the next step, we have imputed the data using Multivariate Imputation and created tables for the imputed data.

Code:

```
# Define variables
variables <- c("id", "agegroup", "age0", "cholmmol", "sbp10", "bmi", "smokerf",
               "arcus0", "dibpat0f", "chd69")
categorical <- c("smokerf", "dibpat0f", "chd69")

# Create Table 1 for the original data
table_original <- CreateTableOne(vars = variables, data = wcgs, factorVars = categorical)
```

```r
# Create Table 1 for the imputed data
table_imputed <- CreateTableOne(vars = variables, data = di, factorVars = categorical)
```

Output:

```
##
##      ### Summary of continuous variables ###
##
## strata: Overall
##             n miss p.miss  mean    sd median  p25   p75  min   max skew kurt
## id       3154    0   0.00 1e+04 6e+03  11406 3741 13115 2001 22101  0.2 -0.7
## age0     3154    0   0.00 5e+01 6e+00     45   42    50   39    59  0.5 -0.8
## cholmmol 3154   13   0.41 6e+00 1e+00      6    5     6    3    11  0.4  0.5
## sbp10    3154    0   0.00 1e+01 2e+00     13   12    14   10    23  1.2  2.8
## bmi      3154    0   0.00 2e+01 3e+00     24   23    26   11    39  0.5  2.0
## arcus0   3154    2   0.06 3e-01 5e-01      0    0     1    0     1  0.9 -1.2
##
## ================================================================================
##
##      ### Summary of categorical variables ###
##
## strata: Overall
##       var    n miss p.miss   level freq percent cum.percent
##  agegroup 3154    0    0.0 [39,45) 1448    45.9        45.9
##                            [45,55) 1384    43.9        89.8
##                            [55,60]  322    10.2       100.0
##
##   smokerf 3154    0    0.0      No 1652    52.4        52.4
##                               Yes 1502    47.6       100.0
##
##  dibpat0f 3154    0    0.0       B 1565    49.6        49.6
##                                  A 1589    50.4       100.0
##
##     chd69 3154    0    0.0       0 2897    91.9        91.9
##                                  1  257     8.1       100.0
##

##
##      ### Summary of continuous variables ###
##
## strata: Overall
##             n miss p.miss  mean    sd median  p25   p75  min   max skew kurt
## id       3154    0      0 1e+04 6e+03  11406 3741 13115 2001 22101  0.2 -0.7
## age0     3154    0      0 5e+01 6e+00     45   42    50   39    59  0.5 -0.8
## cholmmol 3154    0      0 6e+00 1e+00      6    5     6    3    11  0.4  0.5
## sbp10    3154    0      0 1e+01 2e+00     13   12    14   10    23  1.2  2.8
## bmi      3154    0      0 2e+01 3e+00     24   23    26   11    39  0.5  2.0
## arcus0   3154    0      0 3e-01 5e-01      0    0     1    0     1  0.9 -1.2
##
## ================================================================================
##
##      ### Summary of categorical variables ###
##
## strata: Overall
```

```
##       var    n miss p.miss   level freq percent cum.percent
##  agegroup 3154    0    0.0 [39,45) 1448    45.9        45.9
##                            [45,55) 1384    43.9        89.8
##                            [55,60]  322    10.2       100.0
##
##   smokerf 3154    0    0.0      No 1652    52.4        52.4
##                               Yes 1502    47.6       100.0
##
##  dibpat0f 3154    0    0.0       B 1565    49.6        49.6
##                                  A 1589    50.4       100.0
##
##     chd69 3154    0    0.0       0 2897    91.9        91.9
##                                  1  257     8.1       100.0
##
```

Conclusion:

The imputed data has been created using Multivariate Imputation where the missing data of cholmmol has been imputed. The imputed data has been created using Predictive Mean Matching (PMM) method.

Question 2:

Calculate the overall risk of CHD in the cohort.

Overview:

a. What is the outcome we are interested in?

The outcome we are interested in is Coronary Heart Disease (CHD).

b. What are the known risk factors for our outcome of interest?

The known risk factors for Coronary Heart Disease (CHD) are as follows:

- Dichotomous Behaviour type A/B (dibpat0f)
- Age (agegroup, age0)
- Cholesterol (cholmmol)
- Systolic Blood Pressure (sbp10)
- BMI (bmi)
- Smoking (smokerf)
- Corneal arcus (arcus0)

c. How many persons are included?

3154 middle-aged men, from 39 to 59 years of age, during the years 1960-1961 are included in this prospective cohort study.

d. What is the overall risk or rate and prevalence of the disease in our cohort?

The overall risk or rate and prevalence of the disease in our cohort is as follows:

```r
# Overall risk or rate
overall_rate <- table(di$chd69)

#calculate risk of CHD
overall_risk <- overall_rate / sum(overall_rate)

# extract the rate and risk into a data frame
```

```r
chd_frame <- data.frame(
  "CHD Presence" = c("No", "Yes"),
  "Overall Rate" = c(as.matrix(overall_rate)[1], as.matrix(overall_rate)[2]),
  "Overall Risk" = c(as.matrix(overall_risk)[1], as.matrix(overall_risk)[2])
)
#print overall rate and risk into a table
knitr::kable(chd_frame, col.names = c("CHD Presence", "Overall Rate", "Overall Risk"))
```

CHD Presence	Overall Rate	Overall Risk
No	2897	0.9185162
Yes	257	0.0814838

Analysis:

The overall risk of Coronary Heart Disease (CHD) in the cohort is 0.08, which indicates that prevalence of the disease is 8% in the cohort.

Question 3:

Overview:

To solve this problem, we need to build an optimal prediction model for the outcome of Coronary Heart Disease (CHD) using the available data. We will use logistic regression due to the binary nature of the outcome and select predictors that improve our predictions. Additionally, we will consider interaction terms and ensure that categorical variables are appropriately handled.

3.a. Building the Optimal Prediction Model:
Step 1: Model Selection:

Logistic regression is suitable for predicting Coronary Heart Disease (CHD) because:

- Binary outcome: CHD is a binary outcome, meaning it can be either present (1) or absent (0). Logistic regression is designed to model binary outcomes.
- Multiple predictors: There are multiple known risk factors for CHD, and logistic regression can handle multiple predictor variables.
- Quantification of risk: Logistic regression can provide estimates of the probability of developing CHD based on the values of the predictor variables, which can be useful for risk assessment and decision-making.

The `rms` package in R provides functions for regression modeling strategies, including logistic regression via the `lrm` function.

Step 2: Variable Selection:

We start by fitting a full model that includes all potential predictors:

```r
dd <- datadist(di)
options(datadist="dd")
full_model <- lrm(chd69 ~ dibpat0f + age0 + cholmmol + sbp10 + bmi + smokerf
                  + arcus0, data=di, x=TRUE, y=TRUE)
# Extract the model summary
model_summary <- as.data.frame(summary(full_model))
knitr::kable(model_summary, col.names = c("Variable", "Low", "High", "Diff", "Effect",
                                          "S.E.", "Lower 95%", "Upper 95%"),
```

```
            align = c("l", "c", "c", "c", "c", "c", "c", "c"),
            caption = "Summary of the Logistic Regression Model")
```

Table 2: Summary of the Logistic Regression Model

	Variable	Low	High	Diff	Effect	S.E.	Lower 95%	Upper 95%
age0	42.000000	50.00000	8.000000	0.4444889	0.0972089	0.2539630	0.6350148	1
X.Odds.Ratio	42.000000	50.00000	8.000000	1.5596929	NA	1.2891241	1.8870501	2
cholmmol	5.057692	6.48718	1.429487	0.5778017	0.0853722	0.4104753	0.7451282	1
X.Odds.Ratio.1	5.057692	6.48718	1.429487	1.7821166	NA	1.5075341	2.1067115	2
sbp10	12.000000	13.60000	1.600000	0.2959951	0.0654344	0.1677460	0.4242442	1
X.Odds.Ratio.2	12.000000	13.60000	1.600000	1.3444636	NA	1.1826362	1.5284347	2
bmi	22.957374	25.84272	2.885343	0.1626619	0.0761162	0.0134768	0.3118469	1
X.Odds.Ratio.3	22.957374	25.84272	2.885343	1.1766388	NA	1.0135681	1.3659455	2
arcus0	0.000000	1.00000	1.000000	0.2437805	0.1422739	-0.0350713	0.5226323	1
X.Odds.Ratio.4	0.000000	1.00000	1.000000	1.2760642	NA	0.9655366	1.6864611	2
dibpat0f...B.A	2.000000	1.00000	NA	-0.7051579	0.1442429	-0.9878687	-0.4224471	1
X.Odds.Ratio.5	2.000000	1.00000	NA	0.4940305	NA	0.3723695	0.6554409	2
smokerf...Yes.Nd	.000000	2.00000	NA	0.5773367	0.1408234	0.3013278	0.8533455	1
X.Odds.Ratio.6	1.000000	2.00000	NA	1.7812879	NA	1.3516524	2.3474872	2

```
AIC(full_model)
```

```
## [1] 1609.815
```

3.b. Including Interaction Terms

We need to check if including interaction terms between certain predictors improves the model fit. When considering interaction terms in a logistic regression model, one needs to think which variables might have a combined effect on the outcome (Coronary Heart Disease) that's different from their individual effects. Here are some potential interaction terms along with their rationale:

- **Age and Cholesterol:** As people age, their cholesterol levels may have a greater impact on their risk of Coronary Heart Disease. This interaction term can help capture the potential synergistic effect of increasing age and cholesterol levels.
- **Smoking and Age:** Smoking is a well-known risk factor for Coronary Heart Disease, and its effects may be exacerbated with increasing age. This interaction term can help account for the potential increased risk of Coronary Heart Disease among older smokers.
- **BMI and Systolic Blood Pressure:** High blood pressure is often associated with obesity, and the combination of these two factors may increase the risk of Coronary Heart Disease more than either factor alone. This interaction term can help capture the potential additive effect of high BMI and systolic blood pressure.
- **Cholesterol and Systolic Blood Pressure:** High cholesterol and high blood pressure are both risk factors for Coronary Heart Disease, and their combined effect may be greater than the sum of their individual effects. This interaction term can help account for the potential synergistic effect of these two factors.
- **Corneal arcus and Age:** Corneal arcus is a sign of lipid deposition in the cornea, which may be associated with increased risk of Coronary Heart Disease. The effect of corneal arcus may be more pronounced in older individuals, making this interaction term a potential candidate.
- **Smoking and Cholesterol:** Smoking can increase cholesterol levels, and the combination of these two factors may increase the risk of Coronary Heart Disease more than either factor alone. This interaction

term can help capture the potential additive effect of smoking and high cholesterol.

- **Age and BMI:** As people age, their BMI may have a greater impact on their risk of Coronary Heart Disease. This interaction term can help account for the potential increased risk of Coronary Heart Disease among older individuals with high BMI.
- **Cholesterol and Dichotomous Behaviour type:** The Type A behaviour type is historically linked to increased risk of heart disease. This interaction term can help capture the potential additive effect of high cholesterol and Type A behaviour type.
- **BMI and Dichotomous Behaviour type:** Type A behaviour type is associated with stress and may interact with BMI to increase the risk of Coronary Heart Disease. This interaction term can help account for the potential combined effect of high BMI and Type A behaviour type.

We first define a base formula and then consider various interaction terms to see if they improve the model fit. After fitting the models, we compare them based on their AIC values to select the best model.

Code:

```
# Define the base formula
base_formula <- as.formula("chd69 ~ dibpat0f + age0 + cholmmol + sbp10 + bmi + smokerf + arcus0")
# Define potential interaction terms
interaction_terms <- c("age0*cholmmol", "age0:smokerf", "bmi * sbp10",
                       "cholmmol*sbp10", "age0*arcus0 ", "cholmmol:dibpat0f",
                       "smokerf * cholmmol", "age0 * bmi", "sbp10:smokerf",
                       "bmi:dibpat0f")

# Initialize a list to store models and metrics
models <- list()
metrics <- data.frame(Model = character(), AIC = numeric(), stringsAsFactors =
                        FALSE)

# Loop through interaction terms
for (i in 1:length(interaction_terms)) {
  for (j in combn(interaction_terms, i, simplify = FALSE)) {
    # Create formula with interactions
    interaction_formula <- paste(base_formula, paste(j, collapse = " + "),
                                 sep = " + ")
    full_formula <- as.formula(interaction_formula)

    # Fit the model
    model <- lrm(full_formula, data = di, x = TRUE, y = TRUE)

    # Save the model and its AIC
    models[[paste(j, collapse = ", ")]] <- model
    metrics <- rbind(metrics, data.frame(Model = paste(j, collapse = ", "),
                                         AIC = AIC(model)))
  }
}

# Sort models by AIC
metrics <- metrics[order(metrics$AIC), ]

# View the first 15 top-performing models
knitr::kable(head(metrics, 15), col.names = c("Model", "AIC"))
```

	Model	AIC
29	bmi * sbp10, age0*arcus0	1606.232
126	bmi * sbp10, age0*arcus0 , cholmmol:dibpat0f	1606.383
127	bmi * sbp10, age0*arcus0 , smokerf* cholmmol	1606.748
65	age0*cholmmol, bmi* sbp10, age0*arcus0	1606.767
331	bmi * sbp10, age0*arcus0 , cholmmol:dibpat0f, smokerf* cholmmol	1606.866
128	bmi * sbp10, age0*arcus0 , age0* bmi	1607.011
210	age0*cholmmol, bmi* sbp10, age0*arcus0 , cholmmol:dibpat0f	1607.176
332	bmi * sbp10, age0*arcus0 , cholmmol:dibpat0f, age0* bmi	1607.206
335	bmi * sbp10, age0*arcus0 , smokerf* cholmmol, age0 * bmi	1607.380
212	age0*cholmmol, bmi* sbp10, age0*arcus0 , age0* bmi	1607.424
120	bmi * sbp10, cholmmol*sbp10, age0*arcus0	1607.470
211	age0*cholmmol, bmi* sbp10, age0*arcus0 , smokerf* cholmmol	1607.523
41	age0*arcus0 , cholmmol:dibpat0f	1607.531
602	bmi * sbp10, age0*arcus0 , cholmmol:dibpat0f, smokerf* cholmmol, age0 * bmi	1607.551
5	age0*arcus0	1607.579

Explanation

After exploring various models with various combinations of interaction terms along with the full model, we went through a model selection process using AIC to compare the goodness-of-fit. We ultimately chose the model including the interactions between `bmi` and `sbp10`, and between `age0` and `arcus0` as it gave the lowest AIC on comparing with every other model combination.

- **Interaction Terms**: Interaction terms allow us to assess whether the effect of one predictor on the outcome depends on the level of another predictor. For example, the effect of BMI on CHD might vary depending on systolic blood pressure.
- **Model Comparison**: The likelihood ratio test helps determine if the addition of interaction terms significantly improves the model fit. Since the p-value is significant ($p < 0.05$), we include the interaction terms; otherwise, we would have retained the model without interactions. Coupled with the fact that the AIC was slightly better than the model without interactions though it adds a certain level of complexity given the additional number of parameters

```
# Compare using LR test
final_model <- models[[29]]
lrtest(full_model, final_model)
```

```
##
## Model 1: chd69 ~ dibpat0f + age0 + cholmmol + sbp10 + bmi + smokerf +
##     arcus0
## Model 2: chd69 ~ dibpat0f + age0 + cholmmol + sbp10 + bmi + smokerf +
##     arcus0 + bmi * sbp10 + age0 * arcus0
##
## L.R. Chisq       d.f.          P
## 7.58273819 2.00000000 0.02256469
```

3.c. Calculating Predicted Risks Once the final model is selected, we calculate the predicted probabilities of CHD for each individual in the dataset and add these predictions to the dataset.

Predicted Risks: These probabilities provide an estimate of each individual's risk of developing CHD based on the predictor values in the model. This information can be crucial for further analysis, such as assessing model calibration or making risk-based decisions.

```
di$predicted_risk <- predict(final_model, di, type="fitted")
head(di$predicted_risk)
```

```
## [1] 0.050265671 0.118150744 0.008609505 0.010966470 0.149331048 0.026458752
```

These values can be compared to the previously calculated overall risk of 0.0814838 in the cohort to see how individual risks vary based on the predictor variables.

Conclusion:

1. **Model Selection and Variable Selection**:
   - Started with a full logistic regression model.
2. **Interaction Terms**:
   - Assessed interaction effects between various predictors based on domain knowledge. Considered multiple interaction terms to improve model fit.
   - Used likelihood ratio test to compare models with and without interactions.
3. **Predicted Risks**:
   - Calculated predicted probabilities of CHD for each individual and added them to the dataset.

This approach ensures that the final model is both statistically sound and practically useful for predicting CHD risk.

Question 4

Overview :

Code :

Output :

Analysis :

Question 5

Overview :

Code :

Output :

Analysis :

Question 6

Overview :

Code :

Output :

Analysis :

Question 7

Overview :

Code :

Output :

Analysis :

Question 8

Overview :

Code :

Output :

Analysis :

Question 9

Overview :

Code :

Output :

Analysis :

Question 10

Overview :

Code :