Chronic Heart Failure and Risk Factors in Myocardial Infarction Dataset

Ariane, Alona, Minsu, and Jadey

Section 1

Introduction

Chronic heart failure (CHF)

According to CDC,

- More than 6 million adults in the USA have heart failure.

- About half of Americans (47%) have at least one of key risk factors.

(Figure(up): https://www.disability-benefits-help.org/resources/medical-evidence/chronic-heart-failure)

(Figure(down):https://www.verywellhealth.com/heart-failure-causes-and-risk-factors-1746181)

Topics to be covered

Question: How are the predictors of our interest associated with Chronic heart failure?

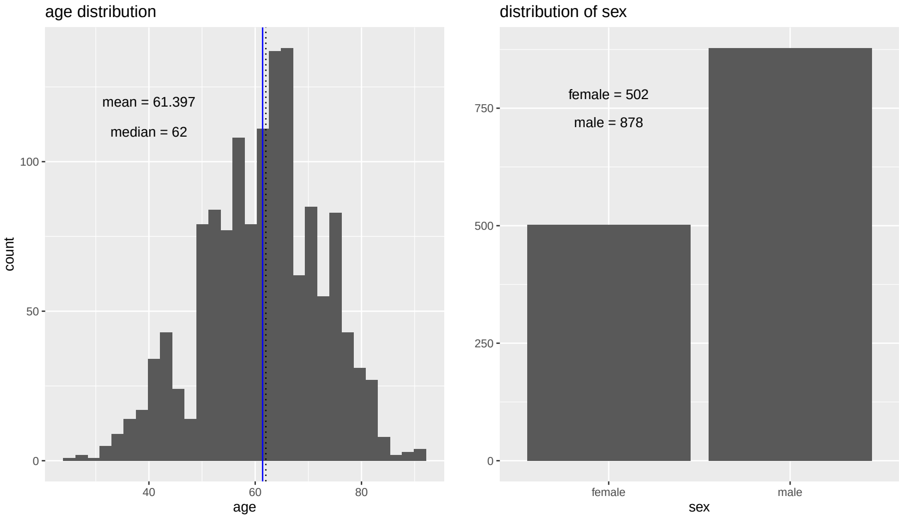
- Test independence of demographics with regards to CHF
- Association of duration of arterial hypertension and CHF
- Build a multiple logistic regression model by adding more predictors and identify the best model
- Modeling the relationship between death outcome and selected variables

Section 2

Dataset Overview

Descriptive statistics

- Demographic information



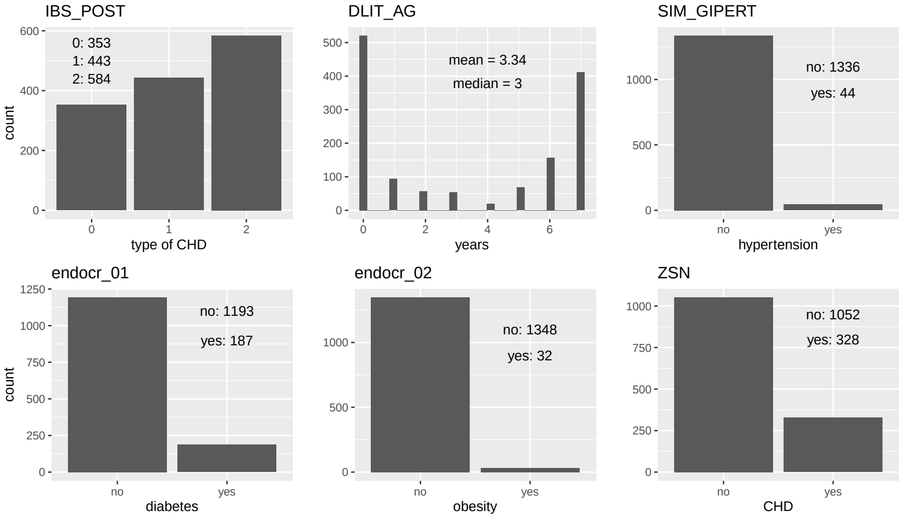
Descriptive statistics

- Patient physiological attributes
- IBS_POST: coronary heart disease in recen weeks before admission to hospital
  - 0: there was no CHD
  - 1: extertional angina pectoris
  - 2: unstable angina pectoris
- DLIT_AG: duration of arterial hypertension
  - 0: there was no arterial hypertension
  - 1: one year
  - 2: two years
  - 3: three years
  - 4: four years
  - 5: five years
  - 6: 6-10 years
  - 7: more than 10 years

Descriptive statistics

- SIM_GIPERT: systematic hypertension; 0 - no, 1 - yes
- endocr_01: diabetes mellitus in the anamnesis; 0 - no, 1 - yes
- endocr_02: obesity in the anamnesis; 0 - no, 1 - yes
- ZSN: chronic heart failure; 0 - no, 1 - yes

Descriptive statistics



Section 3

Tests for Independence of Demographics

Analysis of Sex and Chronic Heart Failure: Overview

Question: Is there an association between sex and chronic heart failure?

	Chronic Heart Failure	
Sex	No	Yes
Female	353	149
Male	699	179

Analysis of Sex and Chronic Heart Failure: Tests

Pearson χ^2 Test of Independence:

```
X-squared
 14.71773

p-value = 0.00012
```

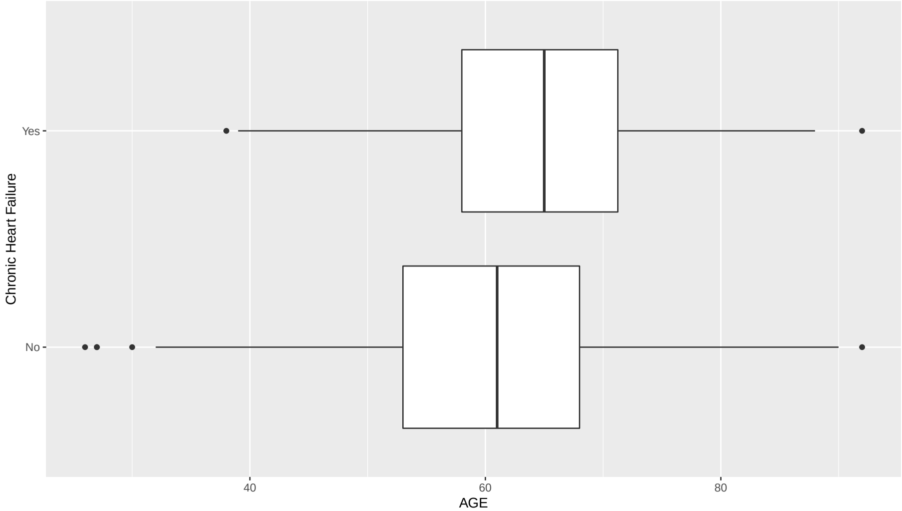
Likelihood Ratio Test of Independence:

```
       G
14.93905

p-value = 0.00011
```

Analysis of Age(Continuous) and Chronic Heart Failure: Overview

Question: Is there an association between age and chronic heart failure?



Analysis of Age(Continuous) and Chronic Heart Failure: Summary Statistics

	Chronic Heart Failure	
	No	Yes
Min.	26	38
1st Qu.	53	58
Median	61	65
Mean	60.42586	64.51220
3rd Qu.	68.00	71.25
Max	92	92

Analysis of Age(Continuous) and Chronic Heart Failure: Test

Analysis was done using a two sided Wilcoxon Rank Sum Test to test if there is a difference in Chronic Heart Failure outcome across age.

```
       W
136546.5
```

```
p-value = 1e-08
```

Analysis of Age(Categorical) and Chronic Heart Failure: Overview

Question: Is there an association between age(decade) and chronic heart failure?

```
     Chronic Heart Failure
Age     No Yes
  20s    3   0
  30s   44   2
  40s  114  24
  50s  294  67
  60s  365 126
  70s  197  86
  80s   32  22
  90s    3   1
```

Analysis of Age(Categorical) and Chronic Heart Failure: Test

Pearson χ^2 Test of Independence:

```
X-squared
 35.41942
```

```
p-value = 1e-05
```

Likelihood Ratio Test of Independence:

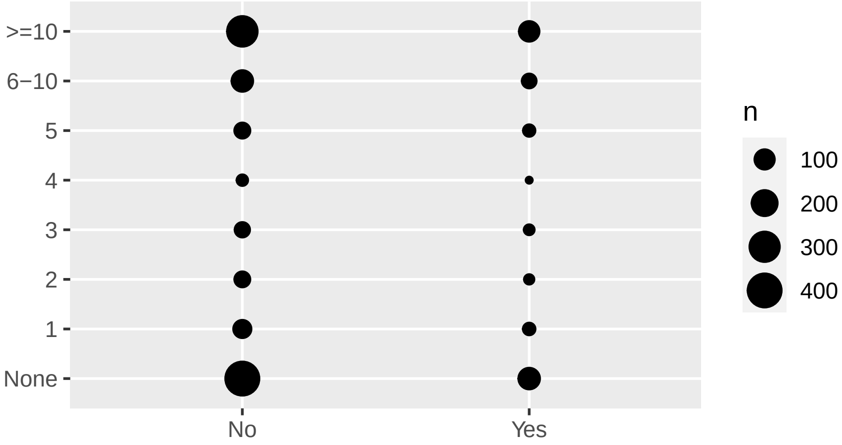
```
       G
38.86163
```

```
p-value = 2.08e-06
```

Section 4

Association of duration of arterial hypertension and CHF

Examining the relationship between Duration of Arterial Hypertension and CHF



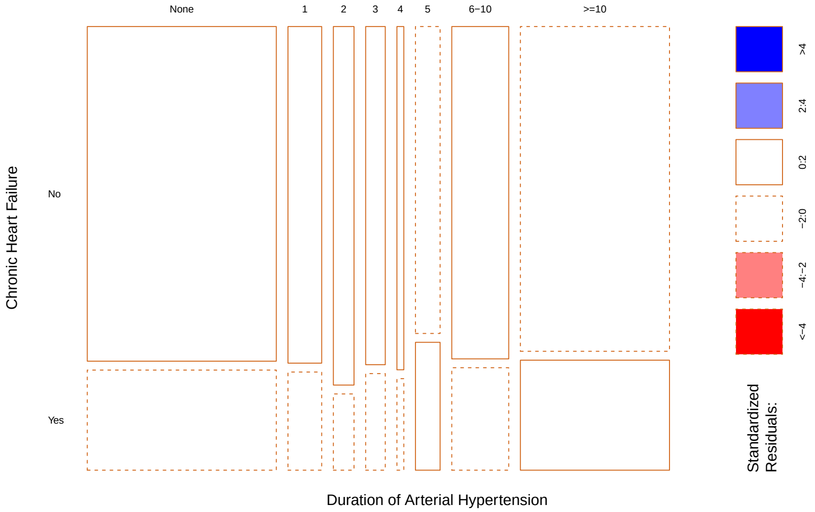
- The two classes of CHF have similar count distributions across the levels of duration of arterial hypertension.
- We will further test the hypothesis that there is an association between the two variables

Inference for contigency table.

Table 1: Duration of Arterial Hypertension by Chronic Heart Failure

	No	Yes
None	401	120
1	72	21
2	47	10
3	42	12
4	15	4
5	48	20
6-10	120	37
>=10	307	104

Examining the Standerdized residuals.



For Ix2 tables, testing for a linear trend in either response category, we use the Cochran-Armitage trend test.

```
    Cochran-Armitage test for trend

data:  dlitag
Z = -0.99455, dim = 8, p-value = 0.32
alternative hypothesis: two.sided
```

Issues to consider: Ordinal variable with unequal intervals so trend test on the original classification provides information about the direction but ignores the unequal spacing in the last two categories.

Logistic Regression model

x - Duration of Arterial Hypertension.

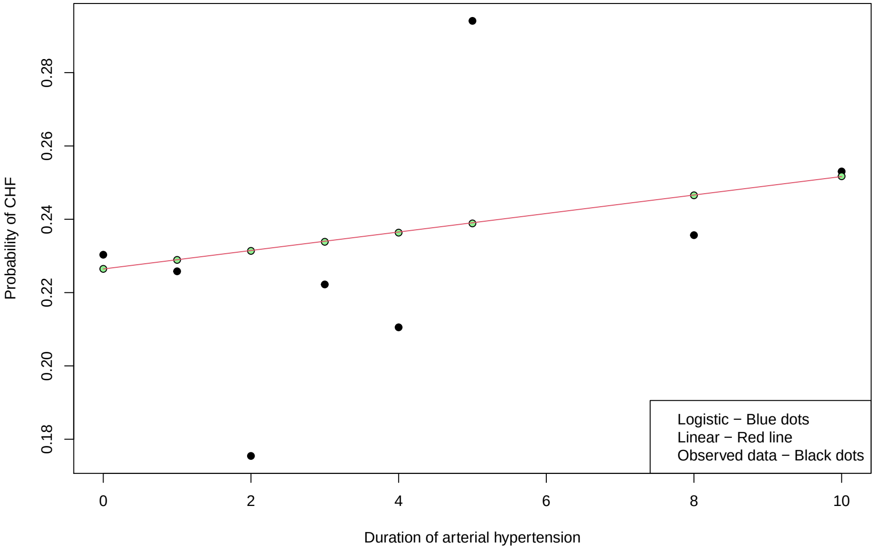
Table 2: Parameter Estimates for Logit link

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2283412	0.0915051	-13.4237468	0.0000000
x	0.0138949	0.0143812	0.9661872	0.3339505

Table 3: Parameter Estiamtes for Identity link

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2264438	0.0160982	14.0664047	0.0000000
x	0.0025212	0.0026207	0.9620338	0.3360326

Predicted probabilities for the fitted models and the observed data.



Sub-analysis

We tested the Linear probability model for the subset: Duration of arterial hypertension between 1 and 5.

Table 4: Parameter Estiamtes for subset analysis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1850895	0.0483670	3.826774	0.0001298
DLIT_AG_N	0.0167632	0.0161478	1.038107	0.2992204

The p-value for the goodness of fit went down sharply (0.16) but still didn't reach significance level to reject the null of no-fit.

Conclusions

- There is no significant association between CHF and the duration of arterial hypertension.
- By itself, duration of arterial hypertension is not predictive of CHF.

Section 5

Multiple Logistic Regression and Model Selection

Multiple Logistic Regression

- Coefficient estimates of the multiple logistic regressions of all predictors

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.02031	0.46737	-6.46239	0.00000
AGE	0.03221	0.00671	4.79990	0.00000
SEX	-0.17273	0.15006	-1.15112	0.24968
IBS_POST	-0.03283	0.08284	-0.39632	0.69187
DLIT_AG_N	-0.02809	0.01632	-1.72118	0.08522
SIM.fyes	-0.40006	0.40869	-0.97888	0.32764
endocr_01.fyes	0.75213	0.17773	4.23174	0.00002
endocr_02.fyes	0.15009	0.41093	0.36525	0.71493

- Only AGE and endocr_01 are statistically significant.
- The P-value for the overall test is much less than 0.0001, thus there is strong evidence that at least one predictor has an effect.

Multiple Logistic Regression - Goodness of Fit

Fit a multiple logistic regression model by adding AGE and endocr_01 to the logistic regression model with only DLIT_AG:

$$\text{logit}[P(ZSN = 1)] = \alpha + \beta_1\, DLIT\_AG + \beta_2\, AGE + \beta_3\, endocr\_01.f.$$

- Goodness of Fit

G.square	df	P-value
1458.899	1376	0.0591161

The model has $G^2 = 1459$ with degree of freedom $df = 1376$ (P-value$=$ $0.059 > 0.05$), which indicates a decent fit.

Multiple Logistic Regression - ANOVA test

Comparing this additive model with the initial model with DLIT_AG only,

- ANOVA Result

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1378	1512.63	NA	NA	NA
1376	1458.90	2	53.73	0

the likelihood ratios test statistic is 53.73 with degree of freedom 2, producing very tiny p-value (P<0.001). Thus, the model with AGE and endocr_01 in addition to DLIT_AG improves the goodness-of-fit.

Multiple Logistic Regression - Model selection

We perform stepwise model selection to see if there is effect of interaction between predictors.

- Backward selection

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1372	1450.891	1466.891
- DLIT_AG_N:AGE:endocr_01.f	1	0.225	1373	1451.116	1465.116
- AGE:endocr_01.f	1	0.594	1374	1451.710	1463.710
- DLIT_AG_N:AGE	1	0.421	1375	1452.131	1462.131

Multiple Logistic Regression - Model selection

- Forward selection

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1379	1513.563	1515.563
+ AGE	-1	34.509	1378	1479.054	1483.054
+ endocr_01.f	-1	17.399	1377	1461.655	1467.655
+ DLIT_AG_N	-1	2.755	1376	1458.899	1466.899
+ DLIT_AG_N:endocr_01.f	-1	6.769	1375	1452.131	1462.131

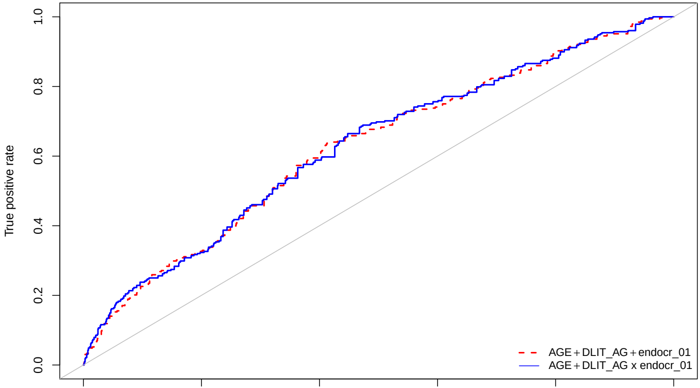
Based on the AIC, both backward elimination and forward selection choose the model of

$$\begin{aligned} \mathrm{logit}[P(ZSN = 1)] =& \alpha + \beta_1\, DLIT\_AG + \beta_2\, AGE + \beta_3\, endocr\_01.f \\ & + \beta_4\, DLIT\_AG * endocr_0 1.f\,. \end{aligned}$$

Predictive Power - ROC curves

- ROC curves of the selected model with interaction and the additive model

```
## Warning: package 'ROCR' was built under R version 4.0.4
```



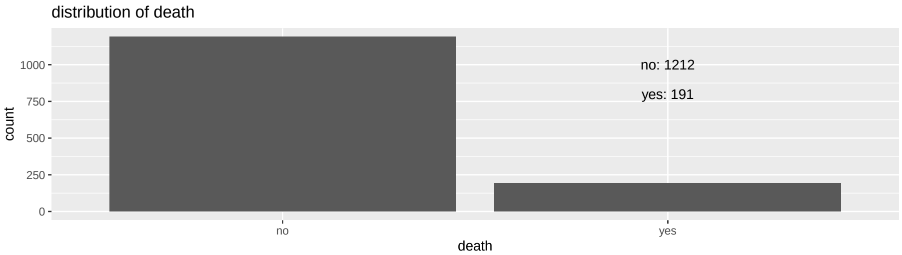
Section 6

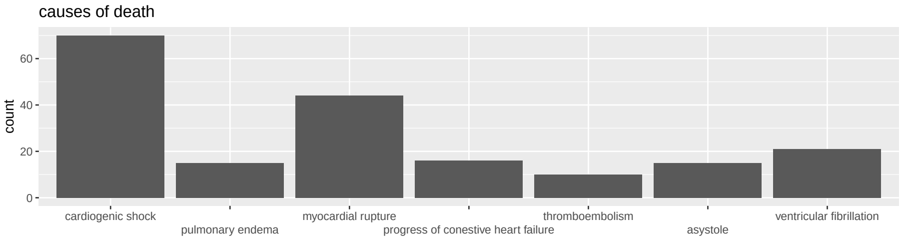
Modeling the relationship between death outcome and selected variables

Secondary analysis - modeling the relationship between death outcome and selected variables

- The dataset includes one variable indicating the causes of lethal outcome for the patients
  - LET_IS: causes of lethal outcome
    - 0: survive
    - 1: cardiogenic edema
    - 2: pulmonary edema
    - 3: myocardial rupture
    - 4: progress of congestive heart failure
    - 5: thromboembolism
    - 6: asystole
    - 7: ventricular fibrillation
- Build a logistic regression model to predict death of the patients by turning LET_IS to a binary variable "death"
- Build model with multimonial response to investigate the cause of death

Secondary analysis - modeling the relationship between death outcome and selected variables





Secondary analysis

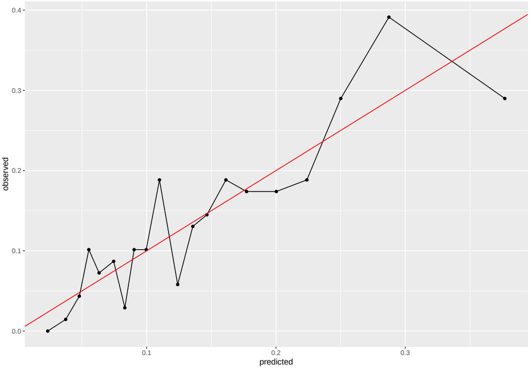
- Full model contains continuous variables AGE, DLIT_AG, categorical variables SEX, IBS_POST, SIM_GIPERT, endocr_01, endocr_02, and the interaction terms between AGE and all the other variables.
- Used stepwise step() to select the best model.
- The best model selected:

$$log \frac{\pi_i}{1-\pi_i} = -6.018 + 0.058 \times AGE + 0.073 \times I(IBS = 1) + 0.696 \times I(IBS = 2) + 0.726 \times I(SIM = 1) + 0.476 \times I(endocr01 = 1) + 1.081 \times I(endocr02 = 1)$$

- All selected predictors increase the probability of death.
  - exp(beta_age) = 1.06
  - OR(IBS_POST = 1 vs. IBS_POST = 0) = 1.08 ($p = 0.77 > 0.5$)
  - OR(IBS_POST = 2 vs. IBS_POST = 0) = 2.01
  - OR(hypertension vs. non hypertension) = 2.07 ($p = 0.065 > 0.5$)
  - OR(diabetes vs. non diabetes) = 1.61
  - OR(obese vs. not obese) = 2.95

Secondary analysis

- Goodness of fit check with Hosmer-Lemeshow test by grouping the observations into 20 groups. The test statistic is 0.4291, indicating an adequate fit of the model to the dataset.
- Plotted the predicted value against the observed value of the 20 groups. Overall the dots follow the diagonal.



Secondary analysis

- Fit baseline category logit model on cause of death. Used predictors selected in the previous analysis.
  $log\frac{\pi_j(x)}{\pi_J(x)} = \beta_{0j} + \beta_{1j} \times AGE + \beta_{2j} \times I(IBS = 1) + \beta_{3j} \times I(IBS = 2) + \beta_{4j} \times I(SIM = 1) + \beta_{5j} \times I(endocr01 = 1) + \beta_{6j} \times I(endocr02 = 1), j = 1, ..., 6$

where J = cardiogenic shock, j = 1 pulmonary edema, 2 myocardial rupture, 3 progress of congestive heart failure, 4 thromboembolism, 5 asystole, 6 ventricular fibrillation

```
multi.mod <- multinom(LET_IS ~ AGE + as.factor(IBS_POST) + as.
```

```
## # weights:  56 (42 variable)
## initial  value 371.668838
## iter  10 value 312.126386
## iter  20 value 300.807784
## iter  30 value 300.010607
## iter  40 value 299.933289
## iter  50 value 299.931699
```

Secondary analysis

```
##    intercept         AGE IBS_POST = 1 IBS_POST = 2 SIM_GIPEI
## 2 -5.208477  0.05003237    0.4172287   -0.2605801  -14.8802
## 3 -2.662446  0.04515371   -0.8725386   -1.3250667   -0.0125
## 4 -3.189649  0.02970654   -0.3969242   -0.7216065    0.0040
## 5  1.046965 -0.03766681   -0.2262002   -1.5074724  -16.1651
## 6 -2.551705  0.03088585   -2.0391936   -1.3081214  -16.8835
## 7  2.872844 -0.05676433   -0.5333366   -0.2875964    0.2570
##    endocr_02 = 1
## 2    -14.0286129
## 3      0.6681173
## 4    -15.2277093
## 5    -16.3110403
## 6      0.7648381
## 7    -15.7083009
```

Estimated $exp(\beta_{ij})$:

```
##      intercept       AGE IBS_POST = 1 IBS_POST = 2 SIM_GIPE
## 2  0.005469998 1.0513051    1.5177496    0.7706045   3.4482
## 3  0.069777345 1.0461887    0.4178893    0.2657852   9.8748
## 4  0.041186340 1.0301522    0.6723850    0.4859709   1.0040
## 5  2.848990047 0.9630338    0.7975584    0.2214691   9.5399
## 6  0.077948656 1.0313678    0.1301336    0.2703274   4.6510
## 7 17.687253718 0.9448167    0.5866443    0.7500642   1.2930
##   endocr_02 = 1
## 2  8.080734e-07
## 3  1.950562e+00
## 4  2.436071e-07
## 5  8.245276e-08
## 6  2.148646e+00
## 7  1.506509e-07
```

Section 7

Conclusion

- Age and Sex are associated with CHF
- Duration of Arterial Hypertension is predictive when included in a multivariate model
- The final multivariable model for CHF is not rejected
- Jadey