

Comparative characteristics of logistic regression implementation in SAS and R on the example of cardiovascular diseases

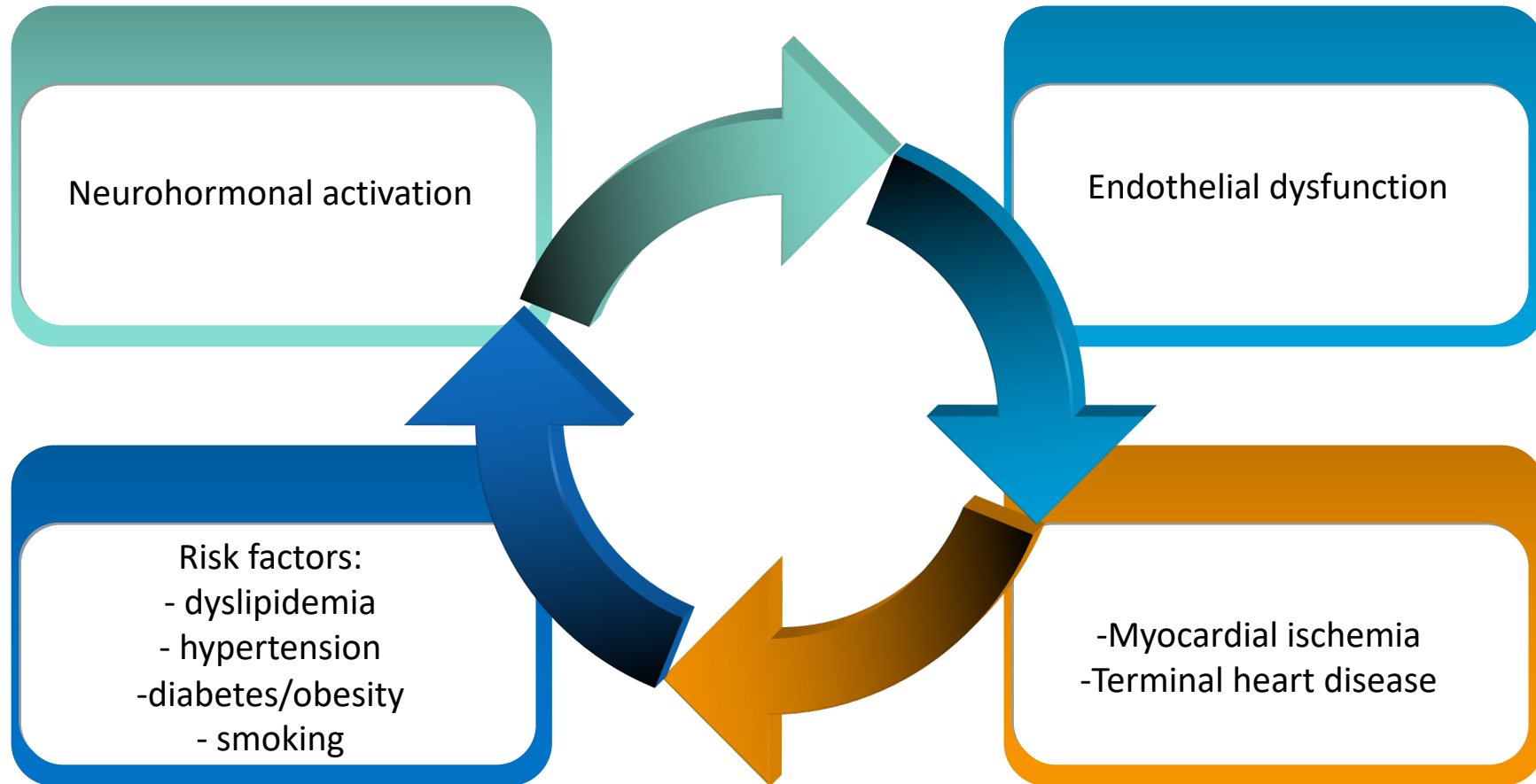
Kashyn Serhii

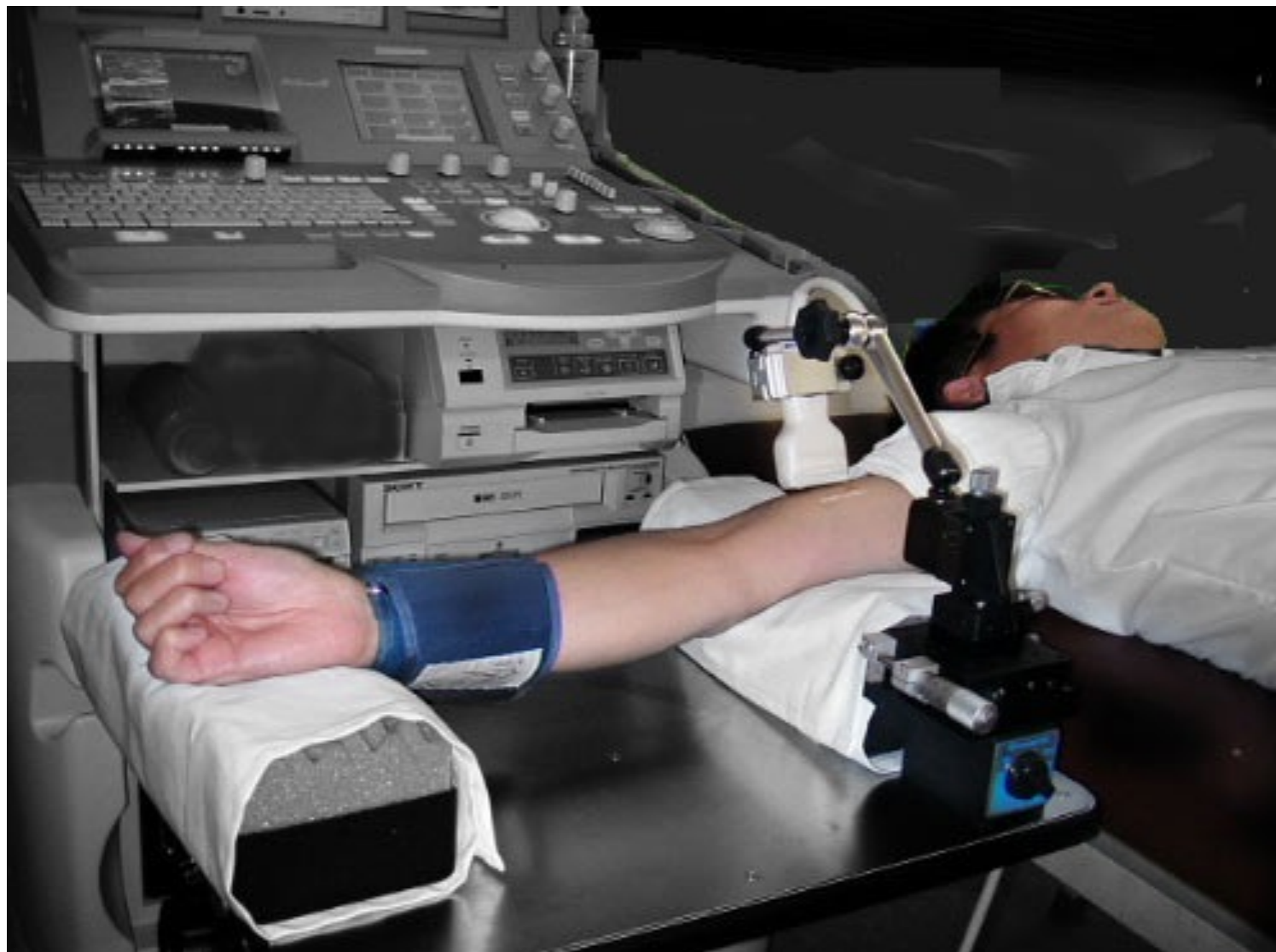
PHUSE 2021 US Connect
14th–18th June 2021

Agenda

- 1. Description of the medical problem
- 2. Logistic regression
- 3. Realization in SAS
- 4. Realization in R
- 5. Comparative characteristics
- 6. Conclusion

Cardiovascular continuum





Indicators

Anthropometry

Height
Body weight
Body mass index (BMI)
Weight-growth index
Shoulder width
Arm length
Hand span

Cardio

Left atrium
Left atrium / body area
Myocardial thickness
Right ventricle
Right ventricle / body area
TIS (thickness of the interventricular septum)
DLVD (diameter of the left ventricle in diastole)
LVEF (left ventricular ejection fraction)
SV (stroke volume of the left ventricle in systole)

Linear classification model

$$a(x, w) = \text{sign}\langle x, w \rangle$$

Training sample

$$X^l = (x_i, y_i)_{i=1}^l, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$$

Continuous approximation of a binary loss function

$$Q(w) = \sum_{i=1}^l [a(x_i, w)y_i < 0] \leq \sum_{i=1}^l f(\langle x_i, w \rangle y_i) \rightarrow \min_w$$

Maximum likelihood principle

$$L(w) = \log \prod_{i=1}^l p(x_i, y_i; w) = \sum_{i=1}^l \log P(y_i | x_i; w) p(x_i) \rightarrow \max_w$$

$$P(y_i | x_i; w) = \frac{1}{1 + \exp(-(x_i, w)y_i)} = \sigma(\langle x_i, w \rangle y_i)$$

where $\sigma(M) = \frac{1}{1 + e^{-M}}$ - sigmoid function

Estimation of the probability coefficient

$$\text{logit}(X) \equiv \log \left(\frac{\Pr(\text{event}|X)}{\Pr(\text{noevent}|X)} \right) = \alpha + X\beta$$

Logarithm of the odds ratio

$$\log(\phi) \equiv \log(\phi(X = 1, X = 0)) = \text{logit}(X = 1) - \text{logit}(X = 0) = (\alpha + 1 \times \beta) - (\alpha + 0 \times \beta) = \beta$$

$$\phi = \exp[(b - a)\beta] = [\exp(\beta)]^{b-a} \equiv [\exp(\beta)]^c$$

Effect coding

Hemodynamic type	Design Variables	
	X1	X2
2 type	1	0
3 type	0	1
1 type	-1	-1

Dummy coding

Hemodynamic type	Design Variables	
	X1	X2
2 type	1	0
3 type	0	1
1 type	0	0

Coefficient logarithms

1

$$\text{logit}(\text{secondtype}) = \alpha + (X_1 = 1)\beta_1 + (X_2 = 0)\beta_2 = \alpha + \beta_1$$

$$\text{logit}(\text{thirdtype}) = \alpha + (X_1 = -1)\beta_1 + (X_2 = -1)\beta_2 = \alpha - \beta_1 - \beta_2 - \beta_3$$

$$\begin{aligned}\log(\phi(\text{secondtype}, \text{thirdtype})) &= \text{logit}(\text{secondtype}) - \text{logit}(\text{thirdtype}) \\ &= 2\beta_1 + \beta_2\end{aligned}$$

2

$$\begin{aligned}\log(\phi(\text{secondtype}, \text{thirdtype})) &= \text{logit}(\text{secondtype}) - \text{logit}(\text{thirdtype}) \\ &= (\alpha + (X_1 = 1)\beta_1 + (X_2 = 0)\beta_2) - (\alpha + (X_1 = 0)\beta_1 + (X_2 = 0)\beta_2) = \beta_1\end{aligned}$$

Analysis results for the maximum likelihood method for both parameterizations

$$\begin{cases} \beta_1^d = 2\beta_1^e + \beta_2^e \\ \beta_2^d = 2\beta_2^e + \beta_1^e \end{cases}$$

Effect coding

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-1.5706	0.3455	20.6679	<.0001
Intercept	2	1	0.7241	0.2821	6.5896	0.0103
tog	2	1	0.3168	0.3330	0.9050	0.3414
tog	3	1	-0.3143	0.3805	0.6821	0.4089

Dummy coding

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-1.5731	0.4966	10.0349	0.0015
Intercept	2	1	0.7216	0.4529	2.5384	0.1111
tog	2	1	0.3194	0.5741	0.3095	0.5780
tog	3	1	-0.3117	0.6567	0.2253	0.6350

Analysis results for the maximum likelihood method for both parameterizations - Continued

<i>Param=</i>	<i>Parameter Estimates</i>		<i>Odds Ratio Estimates</i>
	β^1	β^2	
<i>Effect</i>	0.3168	-0.3143	$\exp(2*0.3168-0.3143)=$ $\exp(0.3193)=1.376$
<i>Ref</i>	0.3194	-0.3117	$\exp(0.3194)=1.376$

$$\begin{cases} \beta_1^d = 2\beta_1^e + \beta_2^e \\ \beta_2^d = 2\beta_2^e + \beta_1^e \end{cases} \rightarrow \begin{cases} 0.3194 \approx 2 * 0.3168 - 0.3143 \\ -0.3117 \approx 2 * (-0.3143) + 0.3168 \end{cases}$$

Construction of logistic regression models with one category predictor

```
proc logistic data = cardio_disease descending;  
  class sex(ref='1') / param = effect;  
  model range = sex / clodds=pl;  
run;
```

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	3.4695	0.0625

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-2.0242	0.4512	20.1305	<.0001
Intercept	2	1	0.3393	0.3714	0.8343	0.3610
sex	2	1	0.6767	0.3633	3.4695	0.0625

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	20.6	Somers' D	0.156
Percent Discordant	5.0	Gamma	0.607
Percent Tied	74.3	Tau-a	0.098
Pairs	1109	c	0.578

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
sex 2 vs 1	1.0000	3.871	1.003	16.984

Model evaluation

1

Concordant

If the entity with the higher predicted probability corresponds to the subject that fell, and the entity with the lower predicted probability corresponds to the entity that did not fall

2

Discordant

If the entity with the higher predicted probability correspond to the subject that did not fall, and the entity with the lower predicted probability corresponds to the entity that fell

3

Tied

If the calculated probability is the same

Construction of a logistic regression model with one continuous predictor

```
proc logistic data = cardio_disease;
  model range (event='1') = sv;
  units sv = 10;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sv	0.982	0.954	1.011

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	58.3	Somers' D	0.178
Percent Discordant	40.5	Gamma	0.180
Percent Tied	1.3	Tau-a	0.111
Pairs	1109	c	0.589

Comparison of AIC and SC values for two models

	Model without interaction	Model with interaction
AIC	131.055	124.75
SC	149.904	131.034

Comparison of pairs

	Model without interaction	Model with interaction
Concordant Pairs	62.9	20.6
Discordant Pairs	36.5	5.0
Ties	0.5	74.3
c Statistic	0.632	0.854

Realization in R

Dummy coding

```
mylogit <- glm(sex ~ tog,  
               data = adis  
               ,family = binomial(link = 'logit')  
             )  
  
summary(mylogit)  
exp(coef(mylogit))
```

Effect coding

```
mylogit <- glm(sex ~ tog,  
               data = adis  
               ,family = binomial(link = 'logit')  
               ,contrasts = list(tog = contr.sum)  
             )  
  
summary(mylogit)  
exp(coef(mylogit))
```

Comparative characteristics

SAS

Dummy coding

Parameter		DF	Estimate
Intercept		1	1.6740
tog	2	1	0.8109
tog	3	1	-0.6624

Effect coding

Parameter		DF	Estimate
Intercept		1	1.7235
tog	2	1	0.7614
tog	3	1	-0.7119

R

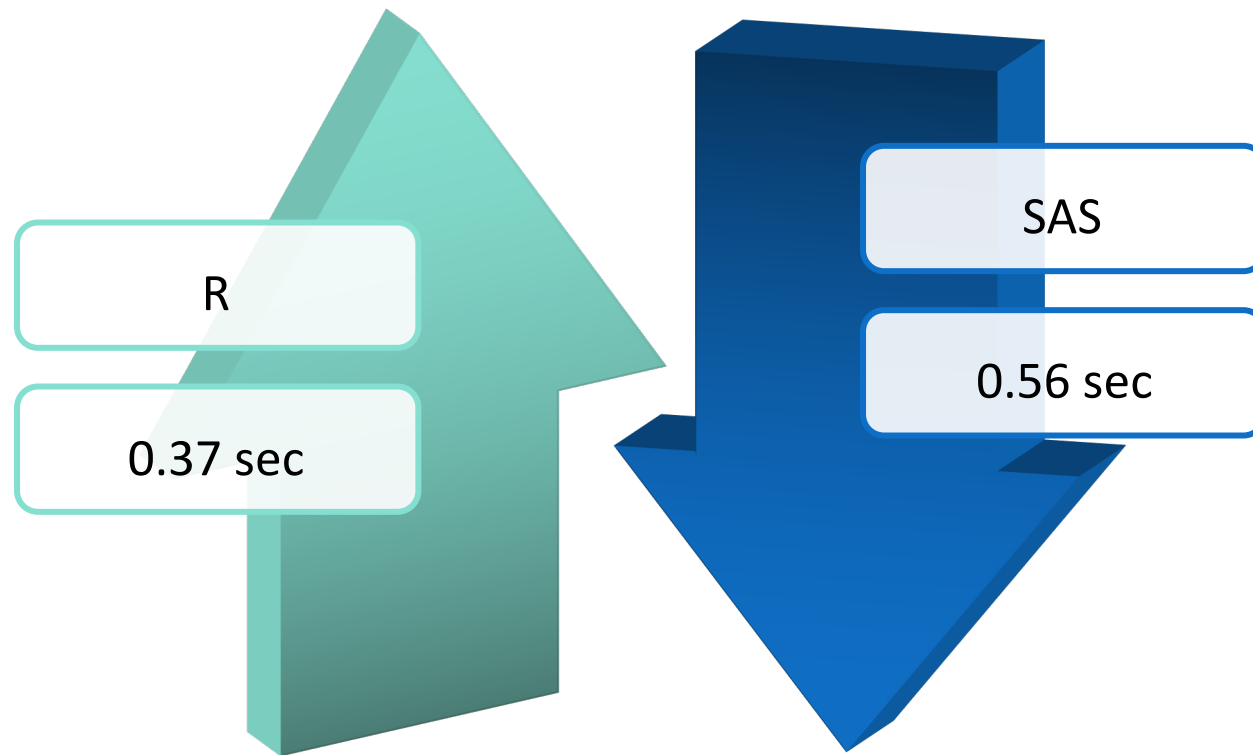
Dummy coding

```
Coefficients:
              Estimate
(Intercept)  1.6740
tog2         0.8109
tog3        -0.6624
```

Effect coding

```
Coefficients:
              Estimate
(Intercept)  1.7235
tog1         0.7614
tog2        -0.7119
```

Launch speed



Conclusion

- 1. Solution of the main problem. The profitability of logistic regression.
- 2. The value of the data obtained.
- 3. Comparison of programming languages.

You can find the codes here

https://github.com/sergeykashyn/logistic_regression



Thank you for your attention!