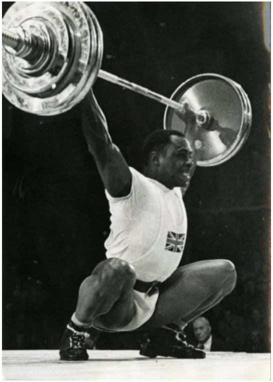
Lifting Big Numbers: Logistic Regression Insights in Olympic Weightlifting

Jean Peyen

Department of Statistics – University of Leeds

August 2023

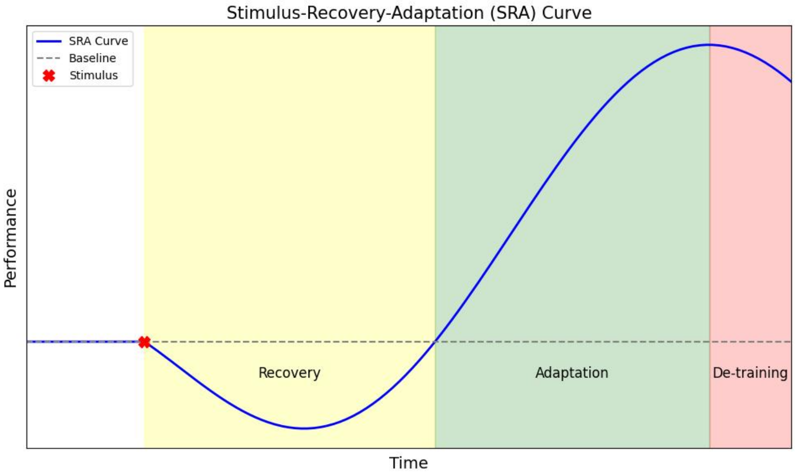


Louis Martin snatch

Isaac Berger clean

Tommy Kono jerk

The Stimulus Recovery Adaptation paradigm



Activity. *Discuss how an injury would qualitatively affect the profile of the SRA curve.*

Dataset

Covariates

`sex`	1.0
`OA`	1.0
`age_dec`	4.2
`age_start_dec`	4.1
`nutrition`	1.0
`train_total`	1.6
`pown`	0.0
`pa`	1.0
`sport0`	1.0

Responses

`shoulder`	0
`knees`	1
`back`	0
`wrist`	0
`hips`	0

Two approaches

Predictive Modelling
Goal : predict outcomes

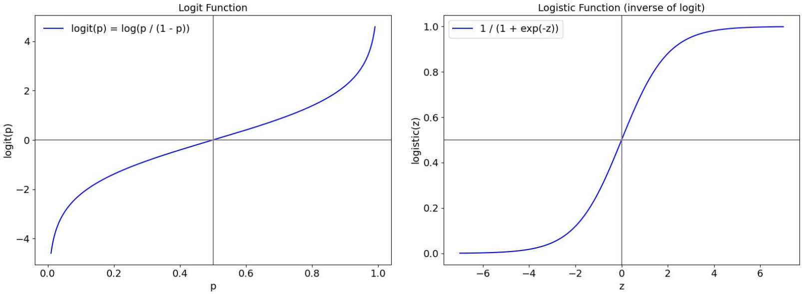
Establishing Influence
Goal: understand relationships

Activity. *Download the dataset and the code from the links provided in the references.*

Logistic regression : model

Each subject $i \in [\![1, n]\!]$ is characterised by a set of covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and a binary response y_i. In the model, the probability of response is given by:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$



Logistic regression: interpretation and challenges

Interpretation of Coefficients (Log-Odds)

For a one-unit increase in the covariate j, the odds of the event occurring are multiplied by e^{β_j}.

- If $\beta_j < 0$: the covariate j reduces the odds of the response.
- If $\beta_j > 0$: the covariate j increases the odds of the response.

Likelihood

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} \cdot (1 - p_i)^{1-y_i}.$$

Multicollinearity

The model is harder to interpret when one variable can be linearly predicted from the others, leading to unstable estimates and inflated standard errors. Regularisation techniques or variable selection may be employed to address multicollinearity

Data analysis : multicollinearity

Variance Inflation Factors

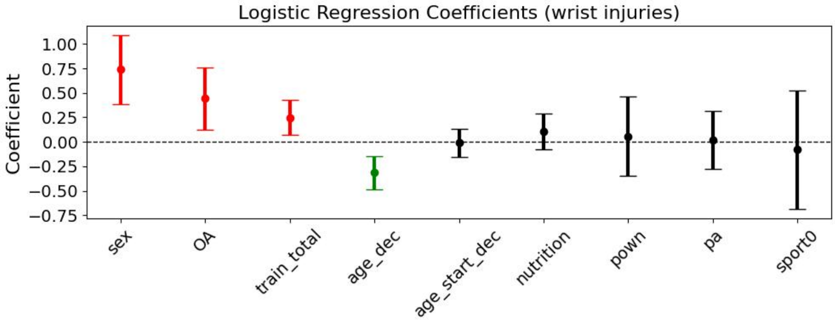
VIF Factor	features
57.237730	const
1.315278	sex
1.030166	OA
1.333104	age_dec
1.518194	age_start_dec
1.102661	nutrition
1.119851	train_total
1.505378	pown
1.033219	pa
1.025428	sport0

Activity. *Which covariates may cause problems of collinearity in the original dataset?*

Data analysis : fit summary

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  wrist   No. Observations:                  976
Model:                          Logit   Df Residuals:                      966
Method:                           MLE   Df Model:                            9
Date:                                   Pseudo R-squ.:                 0.04895
Time:                                   Log-Likelihood:                -484.57
converged:                       True   LL-Null:                       -509.51
Covariance Type:            nonrobust   LLR p-value:                 1.133e-07
==============================================================================
                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0106      0.614     -1.646      0.100      -2.214       0.193
sex            0.7403      0.181      4.090      0.000       0.386       1.095
OA             0.4459      0.162      2.745      0.006       0.127       0.764
age_dec       -0.3128      0.087     -3.602      0.000      -0.483      -0.143
age_start_dec -0.0087      0.074     -0.118      0.906      -0.154       0.136
nutrition      0.1069      0.093      1.150      0.250      -0.075       0.289
train_total    0.2488      0.091      2.719      0.007       0.069       0.428
pown           0.0558      0.207      0.270      0.787      -0.349       0.461
pa             0.0166      0.151      0.110      0.912      -0.279       0.312
sport0        -0.0811      0.311     -0.261      0.794      -0.691       0.528
==============================================================================
```

Data analysis : regression coefficients



Activity. *Based on this analysis, what are your conclusions? Use the downloaded code to analyse other types of injuries. Identify key predictors and their impacts. Did you find any unexpected results?*

Bonus. *How might a coach or an athlete benefit from this analysis? Suggest additional covariates and expected effects. You may refer to the supplementary resources available on the GitHub repository.*

Sources

- David G. Kleinbaum , Mitchel Klein, *Logistic Regression A Self-Learning Text* 3rd edition, Springer New York
- Seabold Skipper, and Josef Perktold, *statsmodels: Econometric and statistical modeling with python*, Proceedings of the 9th Python in Science Conference, 2010, *https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf*
- Marianne Huebner, *Weightlifting during the COVID-19 pandemic. A transnational study regarding motivation, barriers, and coping of master athletes.*, *https://datadryad.org/stash/dataset/doi:10.5061/dryad.qfttdz0hto*
- GitHub repository of the presentation, *https://github.com/Etamunu/weightliftingLogitReg*
- Bob Takano, *Weightlifting programming : a winning coach's guide* Catalyst Athletics