

# Machine learning in sports science

Damian Horna



# About me



Group of  
Horribly  
Optimistic  
STatisticians



# The project

How to use machine learning to predict the **perceived exertion** of young soccer players using **external measures of training load**?



# Why it is important

- Distance and number of squats can be measured, but we don't know how they will influence the player because it is subjective
- Each player can react differently to the same external load.
- But it is the **perceived exertion that has the biggest influence** on the results of the players

# Details of the study

- 2018-2019 in-season competition period (18 weeks)
- 18 youth soccer players(age  $17.81 \pm 0.96$ )
- 804 observations ( $43 \pm 17$  per player)
- Training duration:  $68 \pm 15$  minutes
- Only field-based soccer sessions with warm-ups performed on the field were included for the purpose of the study.
- All of the analyzed training sessions took place in the same part of day, on the same outdoor grass training pitch with a break of 24 hours between consecutive training sessions.

# GPS devices for measuring the external load

Portable, non-differential 10 Hz GPS integrated with a 400 Hz triaxial accelerometer and a 10 Hz triaxial magnetometer (PLAYERTEK, Dundalk, Ireland).



<https://www.playertek.com/us/playertek/>

DASHBOARD

SEASON

SQUADS

SESSIONS

ZONAL

PRO MAPS



game x Set the session's tags...

Split: All

VS

### Session Summary

Start Time: 18:00:20 (03h 03m)

Pitch: portadown FC

Pod: 10

78



11.16 km

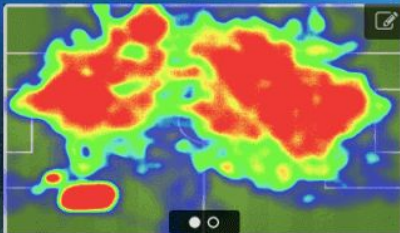


6.51 m/s



43

### Activity Chart



### Volume

Distance : 11.16 km

112%

Sprint Dist : 448 m

45%

Power Plays : 43

54%

Energy : 1303.6 kcal

93%

### Intensity

Top Speed : 6.51 m/s

72%

Distance/min : 60.9 m/min

55%

Power Score : 4.74 w/kg

47%

Player Load : 759.4

152%

# What was measured

- Distance
- Distance > 19.8 km/h
- Distance in Acceleration/Deceleration zones
- Accelerations/Decelerations zone count
- Impacts >3g
- Player load
- Time of the session and all of the above per minute



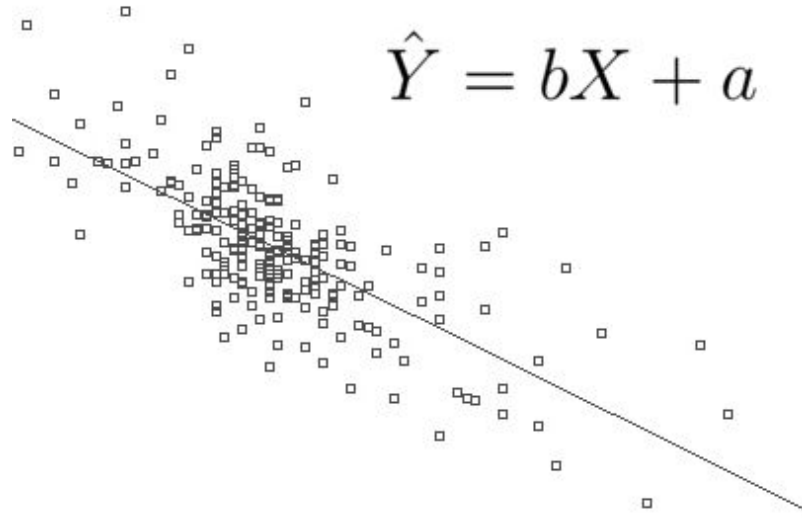
# Sample data

A	B	C	D	E	F	G	H	I	J	K
Player	Distance p	Player Load p	Distance > 1	Impact > 3	Distance in	Distance in	Accelerations	Deceleration	sRPE	
Player 1	69.93	4.35	0.28	3.18	2.89	2.69	3.29	2.76	4	
Player 1	95.57	4.56	5.10	5.79	3.07	2.68	2.66	2.37	8	
Player 1	72.96	4.26	1.21	3.54	3.11	2.64	2.95	2.97	6	
Player 1	84.71	4.88	1.26	5.73	3.68	2.81	3.17	2.99	2	
Player 1	99.65	5.00	2.75	6.43	3.80	3.00	3.10	3.03	9	
Player 1	87.48	4.76	6.88	5.39	3.37	2.77	2.75	2.56	7	
Player 1	62.72	3.87	0.06	3.57	3.76	3.17	3.34	3.38	5	
Player 1	101.25	5.09	5.21	4.77	3.35	2.59	2.87	2.62	7	
Player 1	68.00	3.87	0.44	1.94	2.40	2.33	2.91	2.19	3	
Player 1	80.77	4.58	1.72	2.61	2.66	2.59	2.70	2.25	3	
Player 1	82.55	4.56	5.42	4.39	3.30	2.82	2.91	2.70	5	
Player 1	65.78	3.70	1.63	4.09	4.10	3.50	3.18	2.96	5	
Player 1	89.39	4.65	0.92	4.20	2.91	2.06	2.85	2.69	5	
Player 1	97.37	5.09	6.03	6.66	4.24	2.87	3.32	3.11	5	
Player 1	46.56	2.50	2.08	1.55	2.78	1.97	1.55	1.72	2	
Player 1	80.39	4.40	1.65	2.97	3.10	2.80	2.91	2.39	3	



# sRPE: CR-10 Borg scale modified by Foster

Rating	Descriptor
0	Rest
1	Very easy
2	Easy
3	Moderate
4	Somewhat hard
5	Hard
6	
7	Very hard
8	
9	
10	Maximal

# Why not use OLS linear regression?



# Quick recap: linear regression assumptions

- Linearity (but no multicollinearity)
- Homoscedasticity
- Normally distributed errors
- **Independence of observations** 
- **Continuous, normally distributed response variable** 

# Generalized linear models (GLM)

- Unification of various linear statistical models, such as linear regression, logistic regression, Poisson regression.
- Response variable  $\mathbf{Y}$  does not need to be normally distributed, neither continuous
- In general  $\mathbf{Y}$  is assumed to follow an exponential family distribution (e.g. normal, binomial, Poisson, multinomial, ...)

# Three components of any GLM


- Random component - probability distribution of the response variable  $Y_i$
- Systematic component - explanatory variables  $(X_1, X_2, \dots, X_n)$ , which form the linear predictor:  $\eta_i = \sum_{j=1}^k x_{ij} \beta_j$
- Link function  $g(\mu_i)$  - link between random and systematic components, where  $\mu_i = E(Y_i)$  and  $g(\mu_i) = \eta_i$

# Summary of GLMs

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

<https://online.stat.psu.edu/stat504/node/216/>

# GLM assumptions

- Response variable  $Y$  does NOT need to follow normal distribution
- GLM does NOT assume linear relationship
- The homogeneity of variance does NOT need to be satisfied
- Errors do NOT need to be normally distributed
- MLE rather than OLS
- **But still Independent observations** 



GEE to the rescue!

# GEE overview

- First introduced by Liang and Zeger (1986)
- Generalized Estimating Equation (GEE) is an extension of GLMs
- Semiparametric: uses only mean and variance but not any particular distribution to model  $Y$
- Useful for modelling the **average response** over the population, especially in the longitudinal studies
- **Allow for the correlation** between observations without explaining its origin



# Increasing popularity of GEE

GEE is becoming increasingly popular, because it overcomes the classical assumptions of statistics, i.e. independence and normality.

Let's see how GEE  
works

# GEE: objective

Fit a model to repeated **categorical** responses, that is **correlated** and clustered responses, by GEE methodology.

# GEE: variables

- $n$  - number of subjects
- $Y_i$  - vector of responses of  $i$ -th subject
- $X_i$  - matrix of covariates corresponding to  $i$ -th subject
- $x_{ij}$  - vector of covariates corresponding to  $i$ -th subject at  $j$ -th timepoint
- $\beta$  - regression parameters
- $m_i$  - measurements for  $i$ -th subject

$$Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{im_i})^T$$

$$X_i = (x_{i1}, x_{i2}, \dots, x_{im_i})^T$$

$$x_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp})$$

$$\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$$

for  $i = 1, \dots, n$ ;  $j = 1, \dots, m_i$ ; and  $k = 1, \dots, p$

# GEE: model

Its form is like GLM, but full specification of the joint distribution not required, and thus no likelihood function:

$$g(\mu_i) = X_i \beta$$

# GEE: random component

Any distribution of the response that we can use for GLM, e.g., binomial, multinomial, normal, etc.



# GEE: systematic component

A linear predictor of any combination of **continuous** and discrete variables.

# GEE: link function

Can be any  $g(\mu_i)$  e.g., **identity**, log, logit, etc.

# GEE: covariance structure

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as in the case of independent responses. The random component is described by the same variance functions as in the independence case, but the **covariance structure** of the correlated responses **must also be specified** and modeled now!

Therefore we need to **assume some form of variance** that depends on  $\mu$  and a **model of longitudinal correlation**.

# GEE: variance of the response variable

Continuous outcome:  $V_i = \sigma^2$

Count outcome:  $V_i = \mu_i$

Binary outcome:  $V_i = \mu_i(1 - \mu_i)$

$$S_i = \text{diag}(V_i)$$

## GEE: specifying correlation structure $R_i$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Independence

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Exchangeable

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Autoregressive AR1

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Unstructured

$\rho_{ij} = \text{corr}(Y_{ij}, Y_{ik})$  for the  $i^{\text{th}}$  subject at times  $j$  and  $k$ .

# How to choose the right correlation structure?

- Analyse within-subject correlation structure of the observed data and find the best approximation
- Basically the best choice is the simplest one that fits the data well.

Finally, the ‘working’ variance-covariance matrix

$$Cov(Y_i) = V_i(\beta, \rho) = S_i^{\frac{1}{2}} R_i S_i^{\frac{1}{2}}$$

# The Estimating Equation in GEE

In a GEE, the  $\beta$ -vector estimator can be obtained by iteratively solving an equation:

$$\sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \hat{\mu}_i) = 0$$

Where:

$$\hat{\mu}_i = g^{-1}(X_i \hat{\beta})$$

$$D_i = \frac{\partial \mu_i}{\partial \beta}$$

$V_i$  - our 'working' variance-covariance matrix



# GEE summary

1. Specify distribution
2. Specify link function
3. Specify covariance structure
4. Estimate model parameters using quasi-likelihood

# GEE: advantages

- Computationally more simple than MLE for categorical data
- Does not require multivariate distribution
- Consistent estimation even with mis-specified correlation structure (sandwich estimator)

# GEE: limitations

- There is no likelihood function since the GEE does not specify completely the joint distribution; thus some do not consider it a model but just a method of estimation.
- Likelihood-based methods are NOT available for testing fit, comparing models, and conducting inferences about parameters.
- Empirical based standard errors underestimate the true ones, unless very large sample size.

# Results of the project

Two articles: one in review in Journal of Strength and Conditioning Research, another one in progress...

When it comes to the study - **distance and distance above 19.8 km/h** turned out to be the strongest predictors of perceived exertion.

# References

- Liang, K.Y. and Zeger, S.L.(1986) "Longitudinal data analysis using generalized linear models". Biometrika, 73:1322.
- Zeger, S.L. and Liang, K.Y.(1986) "Longitudinal data analysis for discrete and continuous outcomes". Biometrics, 42:121130.
- <https://online.stat.psu.edu/stat504/node/180/>

Thank you for your  
attention