REGRESIJSKI MODELI V BIATLONU

ANJA ŽAVBI KUNAVER

Fakulteta za matematiko in fiziko
Univerza v Ljubljani

Članek je raziskovalne narave. S pomočjo testiranj slovenske reprezentance biatloncev in tekačev na smučeh so predstavljeni različni modeli, izdelani v programskem jeziku *R*. Izdelava modelov temelji na teoriji mešanih linearnih modelov in uporabi treh metod - metode najmanjših kvadratov, navadne in restringirane metode največjega verjetja. Za preverjanje ustreznosti modelov so uporabljene različne mere, kot najpomembnejši pa je upoštevan determinacijski koeficient multiple regresije.

Poudarek je na napovedi maksimalne porabe kisika tekmovalcev, kjer se kot statistično značilne spremenljivke izkažejo mišičje, višina in starost testirancev. Model tekmovalne uspešnosti je predstavljen zgolj kot zanimivost, saj bi bilo za zanesljive sklepe potrebno pridobiti veliko več podatkov. Na podlagi podatkov, pridobljenih iz 24-ih minut testiranja na tekoči preprogi se da pojasniti več kot 80% variabilnosti v doseženem času testiranja, maksimalni vrednosti ventilacije, porabe kisika in srčnega utripa. Vse dobljene modele bi se dalo izboljšati z večjo skupino testirancev.

REGRESSION MODELS IN BIATHLON

This article is based on real-life data that was gathered from tests of members of Slovenian national teams of biathletes and cross country skiers. We presented different kinds of models which were implemented in programming language R. Most of the time we employ Gaussian linear mixed models. Estimations were made with three different types of methods - least squares method, maximum likelihood method, and restricted maximum likelihood method. For testing goodness of fit of the investigated models we use different measures, most importantly the coefficient of determination. The emphasis is on prediction of the maximum rate of oxygen consumption. We find that muscular mass, height, and age of a competitor are statistically important covariates for this prediction. The model of competitive success is presented only as an interesting fact. For accurate findings much more data should be available. From results gathered from 24 minutes of measuring on the treadmill we can explain over 80% of variability of maximum testing time, maximum ventilation, maximum rate of oxygen consumption, and maximum heart rate. All models could be improved with more tests and participants.

1. Uvod

Biatlon je vse bolj popularen olimpijski šport z dolgo zgodovino. Slovenci so z največjih tekmovanj prinesli že mnoga odličja. Ker sem bila tudi sama biatlonka, sem za zaključek študija na 1. stopnji želela povezati matematiko in šport. Cilj dela je bil generirati modele, vezane na biatlon. Programiranja sem se lotila v programskem jeziku *R*, za delo pa sem potrebovala tako programersko, kot tudi ekonometrično in statistično teorijo.

V 2. poglavju je predstavljena teorija, potrebna za izdelavo modelov. Teorija zajema poglavja iz statistike in ekonometrije. Najprej so predstavljene ekonometrične definicije, nato pa linearni mešani modeli in vsebine, povezane z njimi. 3. poglavje je namenjeno predstavitvi podatkov in opisu spremenljivk. V 4. poglavju je predstavljen začetek dela in pregled podatkov, potrebnih za nadaljnje delo. V 5. poglavju so predstavljeni modeli za napoved maksimalne porabe kisika in je najbolj obširno izmed poglavij o izdelavi modelov. Tu so uporabljene različne metode in oblike modelov. 6. poglavje je namenjeno razlagi modela tekmovalne uspešnosti. V zadnjem poglavju je opis napovedi maksimalnih vrednosti na podlagi 24 minut testiranja.

2. Teorija

2.1 Osamelci in škatla z ročaji

Osamelci so vrednosti spremenljivke, ki so zunaj intervala $(Q_1 - \frac{3}{2}Q_r, Q_3 + \frac{3}{2}Q_r)$, kjer s Q označujemo kvartile in s $Q_r = Q_3 - Q_1$ kvartilni razmik. Škatla z ročaji (tudi škatla z brki ali okvir z ročaji, angleško box plot) je vrsta grafa. Okvir določata kvartila Q_1 in Q_3, njegovo prečko pa mediana Q_2. Spodnji in zgornji ročaj določata pogojni minimum in pogojni maksimum. Le ta dobimo tako, da poiščemo najmanjšo oz. največjo vrednost spremenljivke, ki ni osamelec. Škatla z brki zelo nazorno prikaže obliko porazdelitve spremenljivke, njene kvartile, variacijski razmik in kvartilni razmik ([9] Kraner Šumenjak).

2.2 Bimodalna porazdelitev

Grafi gostot imajo lahko različno število vrhov. Najbolj poznana je unimodalna porazdelitev, katere graf gostote ima en vrh. Klasičen primer je normalna porazdelitev. Pri bimodalni porazdelitvi ima graf gostote dva vrhova. Ta dva vrhova ponazarjata lokalna maksimuma, torej lokalno najpogostejše vrednosti. Ponavadi dva vrhova nakazujeta na to, da imamo dve različni skupini. Poznamo še večmodalne porazdelitve, ki imajo več kot dva vrhova ([10 Statistics how to]).

2.3 Domneva, testna statistika, funkcija moči

Domneva je neka trditev o populaciji. Testna statistika je funkcija vzorca, na podlagi katere se odločamo o resničnosti domneve. Naj bo Z zavrnitveno območje za testno statistiko T. Potem je funkcija moči za testno statistiko T funkcija, ki jo zapišemo kot $\beta(\theta) = P_\theta(T \in Z)$ (Pohar Perme, 2019).

2.4 Determinacijski koeficient in p-vrednost

Determinacijski koeficient multiple regresije (R^2) je najpogosteje uporabljena mera primernosti oziroma zanesljivosti regresijskega modela. Pove, kolikšen delež celotne variance odvisne spremenljivke y je pojasnjen z linearnim regresijskim modelom, na podlagi katerega je izračunan. Lahko ga zapišemo kot kvadrat korelacijskega koeficienta med dejanskimi in z obravnavanim regresijskim modelom izračunanimi vrednostmi odvisne spremenljivke. Torej velja $0 \leq \ R^2 = r^2_{y\hat{y}} \ \leq 1$. Vrednosti R^2 niso primerljive med modeli z različno definirano odvisno spremenljivko y. Velja, da se z izločanjem pojasnjevalnih spremenljivk v modelu R^2 lahko zmanjša ali ostane enak, z dodajanjem pojasnjevalnih spremenljivk pa se vrednost ne more zmanjšati. Potrebno se je zavedati, da visoka vrednost determinacijskega koeficienta še ne pomeni, da smo v model vključili prave pojasnjevalne spremenljivke. Nizke vrednosti pa še ne pomenijo, da model ne vključuje pomembnih pojasnjevalnih spremenljivk (Pfajfar, 2018).

Naj bo T testna statistika in ničelna domneva $H_0 : \theta \in \Theta_0$. Denimo, da zavračamo za velike vrednosti T in naj bo t vrednost testne statistike na vzorcu. Potem $p - vrednost$ definiramo kot

$$p = \sup_{\theta \in \Theta_0} P(T \geq t).$$

Če za $p - vrednost$ velja $p \leq \alpha$, zavrnemo ničelno domnevo in sklepamo, da je spremenljivka statistično značilna. Običajno za α vzamemo vrednost 0,05. Če torej velja $p \leq 0,05$, lahko s 95% gotovostjo posplošimo rezultate iz vzorca na populacijo. Če to ne velja, se moramo vzdržati vsakega sklepanja iz vzorca na populacijo (Pohar Perme, 2019).

Program R sam izračuna vrednosti R^2 in $p-vrednost$, zato se izračunom v diplomskem delu ne bom posebej posvečala. Več o možnih postopkih izračuna je v svojem delu zapisal Pfajfar (Pfajfar, 2018).

2.5 Linearni mešani modeli

Posplošeni linearni mešani model izrazimo kot

$$Y = X\beta + Z\alpha + \epsilon,$$

kjer je Y opazovani slučajni vektor, X matrika znanih vrednosti pojasnjevalnih spremenljivk, β neznan vektor regresijskih koeficientov (fiksni učinki), Z znana matrika, α vektor naključnih učinkov in ϵ vektor napak. α in ϵ sta neopazovana. Predpostavimo, da sta nekorelirana.

V matrični obliki model izgleda takole:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & 0 & \dots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,q} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,q} \\ \vdots & \vdots & & \vdots \\ Z_{n,1} & Z_{n,2} & \dots & Z_{n,q} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_q \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Linearni mešani modeli se delijo na Gaussove ali normalne in ne-Gaussove. Pomembna predpostavka pri normalnih linearnih mešanih modelih je normalna porazdeljenost vektorja slučajnih učinkov $\alpha \sim N(0, \sigma^2 I_q)$ in vektorja slučajnih odstopanj $\epsilon \sim N(0, \tau^2 I_n)$, ki nista nujno enakih razsežnosti. Druga pomembna predpostavka je neodvisnost slučajnih vektorjev α in ϵ. Prednost uporabe nenormalnih linearnih mešanih modelov pred normalnimi je v tem, da so bolj fleksibilni za modeliranje. Uporabni so na mnogo različnih področjih (npr. v medicini, financah, izobraževanju,...). Za modeliranje pridejo prav, kadar so meritve na testirancih opravljene večkrat skozi neko časovno obdobje (Maver, 2018, str. 6).

2.5.1 Osnovni model enostavne linearne regresije

Za razumevanje linearnih mešanih modelov se je dobro spomniti osnovnega modela enostavne linearne regresije. Zapišemo ga kot:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

kjer so ϵ_i med seboj neodvisne slučajne spremenljivke, x_i pa dane vrednosti. Velja $\epsilon_i \sim N(0, \sigma^2)$ za vsak i in tako $Y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$. Model lahko razširimo na več linearnih parametrov:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

kjer so ϵ_i neodvisne enako porazdeljene slučajne spremenljivke, za $1 \leq i \leq n$.

Lahko ga zapišemo tudi v matrični obliki:

$$Y = X\beta + \epsilon.$$

V tem modelu je predpostavljeno, da so regresijski koeficienti fiksni, obstajajo pa primeri, v katerih je smiselno predvidevati, da so nekateri izmed koeficientov naključni (Jiang, 2007, str. 1).

2.5.2 Longitudinalni/vzdolžni model

Spada med normalne linearne mešane modele. V splošnem ga zapišemo kot

$$Y_i = X_i\beta + Z_i\alpha_i + \epsilon_i,$$

kjer je Y_i opazovani vektor i-tega posameznika, X_i in Z_i sta znani matriki, $1 \leq i \leq n$. Kot običajno je β neznan vektor regresijskih koeficientov, α_i vektor naključnih učinkov, ϵ_i pa vektor napak. Število stolpcev matrike Z_i predstavlja razsežnost slučajnega vektorja α_i, število vrstic pa število meritev na i-tem posamezniku. Predpostavimo, da sta α_i in ϵ_i neodvisna in porazdeljena $\alpha_i \sim N(0, G_i)$ in $\epsilon_i \sim N(0, R_i)$, kjer sta $G_i = Var(\alpha_i)$ in $R_i = Var(\epsilon_i)$. Če razlik med meritvami znotraj ene enote ni, je variančno-kovariančna matrika R_i diagonalna. Vzdolžne modele se uporablja, kadar imamo skupino posameznikov, kjer je na vsakem posamezniku izvedenih več meritev v različnih časovnih trenutkih. Najpogosteje so uporabljeni v analizah longitudinalnih podatkov (Jiang, 2007, str. 6).

2.5.3 Multikolinearnost in avtokorelacija

Med temeljne predpostavke regresijskega modela spada predpostavka, da med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multikolinearnosti. Najbolj tipičen vzrok za kršenje te predpostavke je, da smo v model kot neodvisni vključili dve spremenljivki, med katerima obstaja močna linearna povezanost. Do multikolinearnosti pride tudi, če v model vključimo več spremenljivk kot je velikost vzorca. Na multikolinearnost posumimo, če se v modelu determinacijski koeficient izkaže za statistično značilnega, od regresijskih koeficientov pa nobeden.

Medsebojna neodvisnost vrednosti slučajne spremenljivke ϵ prav tako spada med temeljne predpostavke linearnega regresijskega modela. Pravimo, da spremeneljivke niso avtokorelirane oz. da v modelu ni avtokorelacije. To predpostavko zapišemo $Cov(\epsilon_i, \epsilon_j) = 0$ za vsak $i \neq j$ (Pfajfar, 2018).

2.6 Metode ocenjevanja

V tem razdelku bom na kratko opisala tri metode, ki sem jih uporabila pri modeliranju v programskem jeziku *R*. Prva metoda je metoda najmanjših kvadratov (MNK), ki je najbolj enostavna. Nato bom opisala metodo največjega verjetja (MNV) in restringirano metodo največjega verjetja (RMNV), ki je bila razvita zaradi pomanjkljivosti prejšnje. Pri obeh standardnih metodah (MNV in RMNV) ločimo dva načina ocenjevanja in sicer točkovno ocenjevanje (za majhen vzorec) in ocenjevanje z asimptotično kovariančno matriko (za velike vzorce). V delu bom metodi predstavila le za točkovno ocenjevanje. V primeru osnovnega modela linearne regresije sta oceni koeficientov po metodah MNK in MNV ekvivalentni, če predpostavimo normalno porazdelitev ϵ in α.

2.6.1 Metoda najmanjših kvadratov (MNK)

Pri 16 letih jo je odkril nemški matematik Carl F. Gauss. Zaradi svojih lastnosti je najbolj razširjena metoda ocenjevanja regresijskih koeficientov (Pfajfar, 2018, str.53).

Pri MNK na primeru osnovnega regresijskega modela velikosti $p = 1$ iščemo β_0 in $\beta 1$ tako, da bo vsota kvadratov ostankov najmanjša možna. Pri danih (x_i, y_i) torej iščemo

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

2.6.2 Metoda največjega verjetja (MNV)

Naj bodo $X_1, ..., X_n$ n.e.p. s.s., porazdeljene z gostoto $f_Y(x;\ \theta_1, \ldots, \theta_k)$. Funkcijo

$$L(\theta;\ x) = L(\theta_1, \ldots, \theta_k;\ x_1, \ldots, x_n) = \prod_{i=1}^{n} f_Y(x_i;\ \theta_1, \ldots, \theta_k)$$

imenujemo funkcija verjetja za vzorec velikosti n.

Za vsak vzorec x naj bo $\hat{\theta}(x)$ vrednost parametra, v katerem funkcija L doseže maksimum. Cenilka po MNV za parameter θ na osnovi vzorca X je $\hat{\theta}(X)$. Ta cenilka je pod ustreznimi predpostavkami dosledna, vendar ne nujno nepristranska.

MNV ima asimptotsko najmanjšo varianco (je učinkovita), kar sledi kot posledica neenakosti Cramér-Raa. V primeru diskretne porazdelitve je verjetje kar produkt verjetnosti:

$$L(x;\ \theta) = \prod_{i=1}^{n} P(X_i = x_i;\ \theta).$$

Gaussov linearni mešani model lahko predstavimo z njegovo marginalno porazdelitvijo. Tako je Y porazdeljen kot

$$Y \sim N(X\beta, V),$$

kjer je $V = R + ZGZ^T$ ter $R = \mathrm{diag}(R_1, ..., R_n)$, $G = \mathrm{diag}(G_1, ..., G_n)$ in $Z = \mathrm{diag}(Z_1, ..., Z_n)$. Vektor $\theta = (R_1, ..., R_n, G_1, ..., G_n)$ predstavlja vektor vseh variančnih komponent, vključenih v V. V nadaljevanju je predstavljena izpeljava cenilke po MNV, kjer y in Y označujejo vektorje.

Porazdelitev slučajnega vektorja Y je dana z gostoto

$$f(y) = \frac{1}{(2\pi)^{\frac{n}{2}}|V|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta)},$$

kjer je n dimenzija Y. Torej je logaritmirana funkcija verjetja enaka

$$l(\beta, \theta) = c - \frac{1}{2}\log(|V|) - \frac{1}{2}(y - X\beta)^T V^{-1}(Y - X\beta),$$

kjer je c neka konstanta. Ko odvajamo po parametru β, dobimo

$$\frac{\partial l}{\partial \beta} = X^T V^{-1} Y - X^T V^{-1} X\beta.$$

Odvod po komponenti vektorja θ pa je enak

$$\frac{\partial l}{\partial \theta_r} = \frac{1}{2}\{(Y - X\beta)^T V^{-1} \frac{\partial V}{\partial \theta_r} V^{-1}(Y - X\beta) - sled(V^{-1}\frac{\partial V}{\partial \theta_r})\}, r = 1, ..., q,$$

kjer je θ_r r-ta komponenta vektorja $\theta = (R_1, ..., R_n, G_1, ..., G_n)$. Le ta je dimenzije q.

Cenilka bo rešitev vektorskih enačb

$$\frac{\partial l}{\partial \beta} = 0, \frac{\partial l}{\partial \theta} = 0.$$

Dobljena cenilka za vektor koeficientov po MNV je

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} Y.$$

Podrobnejšo izpeljavo z vsemi potrebnimi predpostavkami je v svojem delu predstavil Jiang (Jiang, 2007).

Da še malo poenostavimo, poglejmo primer osnovnega modela linearne regresije v matrični obliki $Y = X\beta + \epsilon$, kjer so

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & 0 & \dots & X_{n,p} \end{bmatrix}, \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Tu je matrika X dimenzije $n \times (p+1)$ in p število spremenljivk. Predpostavimo, da je matrika X polnega ranga in velja

$$\epsilon \sim N(0, \sigma^2 I_{n\times n}).$$

Cenilka, dobljena po MNV, je

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Ta cenilka je nepristranska, saj velja

$$E(\hat{\beta}) = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X \beta = \beta.$$

Izpeljava variančno-kovariančne matrike:

$$var(\hat{\beta}) = var((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T var(Y) X (X^T X)^{-1} =$$
$$= (X^T X)^{-1} X^T \sigma^2 I_{n\times n} X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Velja

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 (X^T X)^{-1}_{ii}}} \sim N(0,1).$$

V primeru, ko je $p = 1$, je ocena regresijskega koeficienta β_1 po metodah MVN in MNK enaka

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

2.6.3 Restringirana metoda največjega verjetja (RMNV)

Zaradi pomanjkljivosti metode največjega verjetja se je razvila restringirana metoda največjega verjetja. Ta ima več prednosti pred MNV, med drugim omogoča pridobivanje cenilk zgolj za parametre, ki so predmet zanimanja in se ne ozira na nezanimive parametre. Cenilke, dobljene po RMNV, so invariantne za fiksne učinke modela, metoda pa le-teh niti ne ocenjuje (Maver, 2018, str.17).

Poglejmo še podrobnejši opis RMNV, povzet po obsežnejši literaturi (Jiang, 2007). Naj bo $Y \sim N(X\beta, V)$. Brez škode za splošnost naj bo izpolnjena predpostavka, da ima matrika $X \in R^{m\times p}$ poln rang, torej $\text{rang}(X) = p$. Naj bo A taka matrika dimenzije $m \times (m-p)$, za katero velja $\text{rang}(A) = m - p$ in $A^T X = 0$. Za potrebe izpeljave naj bo $Z = A^T Y$. Velja: $E(A^T Y) = A^T X \beta = 0$, $Var(A^T Y) = A^T V A \Rightarrow Z \sim N(0, A^T V A)$. Sledi, da je zvezna porazdelitvena gostota slučajne spremenljivke Z dana kot

$$f_{Z,R}(z) = \frac{1}{(2\pi)^{(m-p)/2} |A^T V A|^{1/2}} e^{-\frac{1}{2} z^T (A^T V A)^{-1} z}.$$

Indeks R tu označuje, da gre za restringirano metodo. Restringirana logaritmirana funkcija verjetja je

$$l_R(\theta) = c - \frac{1}{2}\log(|A^T V A|) - \frac{1}{2} Z^T (A^T V A)^{-1} Z, \tag{1}$$

kjer je c konstanta. Z odvajanjem pa dobimo

$$\frac{\partial l_R}{\partial \theta_i} = \frac{1}{2}\{Y^T P \frac{\partial V}{\partial \theta_i} P Y - sled(P \frac{\partial V}{\partial \theta_i})\}, i = 1, ..., q.$$

Tu je $P = A(A^T V A)^{-1} A^T$. Cenilka po restringirani metodi največjega verjetja je definirana kot točka, v kateri funkcija (1) zavzame maksimum. Dobi se jo kot rešitev enačbe $\frac{\partial l_R}{\partial \theta} = 0$ (Jiang, 2007, str. 13). RMNV je torej metoda, s katero se direktno ne pridobi cenilke parametra β, pač pa le cenilko parametra θ. Razlog je v tem, da se β izloči že pred ocenjevanjem. Cenilko za β se dobi iz enačbe $V = V(\hat{\theta})$, kjer je $\hat{\theta}$ cenilka za θ, dobljena z RMNV. Izpeljave in podrobnejšo razlago metode lahko pogledamo v (Jiang, 2007).

2.7 Biatlon in VO2max

Biatlon je zimski šport, sestavljen iz teka na smučeh v drsalni tehniki in streljanja z malokalibrsko puško. Biatlonci tekmujejo v več različnih individualnih disciplinah, ki se razlikujejo v dolžini proge, štartu in številu strelskih prihodov. Štart je lahko individualen ali skupinski. Poleg individualnih disciplin tekmujejo tudi v štafetnih preizkušnjah. Ne glede na disciplino streljajo v dveh položajih - leže in stoje. Vsak zgrešeni strel se kaznuje s časovnim pribitkom ali kazenskim krogom. Cilj vsakega tekmovalca je preteči progo v čim hitrejšem času, poleg tega pa je za uspeh potrebno hitro in učinkovito streljanje.

VO2max ali maksimalna poraba kisika ali maksimalna aerobna kapaciteta je največja količina kisika, ki jo organizem lahko porabi v 1 minuti. Ta volumen izražamo v litrih na minuto ali mililitrih na kilogram telesne mase na minuto. Odraža se v delovanju pljuč (prenos kisika v kri), v vezavi kisika na hemoglobin, črpalni sposobnosti srca in cirkulaciji krvi v mišice. Pri merjenju porabe kisika posredno merimo tudi posameznikovo maksimalno aerobno zmogljivost (Podgornik, 2013). Zaželeno je, da je maksimalna vrednost VO2max čim večja. Ker je VO2max eden izmed najpomembnejših dejavnikov pri uspešnosti v smučarskem teku, sem temu namenila več pozornosti.

3. Podatki

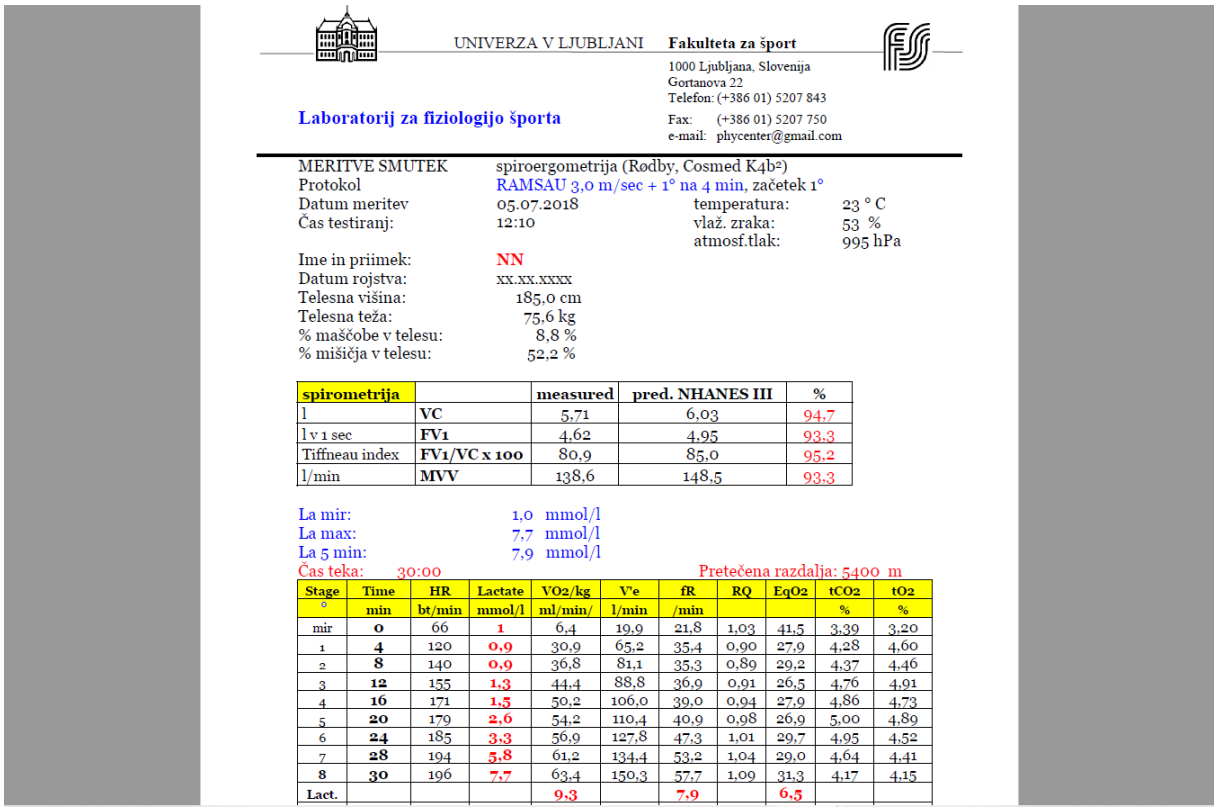
V tem razdelku so predstavljeni podatki, katere analiziramo s pomočjo različnih modelov.

Zbrana so testiranja biatlonske moške članske in mladinske reprezentance ter tekaške moške članske reprezentance za sezoni 2016/17 in 2017/18. Sodelovali so 4 tekači in 9 biatloncev, od tega 4 mladinci. Običajno tekmovalci opravijo po 2 testiranji letno, eno v poletnem delu pripravljalnega obdobja in eno v jesenskem delu. Tako imamo za vsakega tekmovalca zbranih več meritev (od vsakega vsaj 2 in največ 4).

Najpomembnejša testiranja za spremljanje telesne pripravljenosti biatlonci in tekači opravljajo v Športno diagnostičnem centru na Fakulteti za šport (FŠ). Na testiranju vsak testiranec s tekaškimi rolkami teče na tekoči preprogi. Hitrost teka je konstantna, 3m/s, naklon se začne pri 2 stopinjah, na vsake 3 ali 4 minute (odvisno od protokola) pa se dvigne za 1 stopinjo. S pomočjo dihalne maske in drugih naprav se merijo različni parametri, ki pričajo o telesni pripravljenosti testiranca. Po koncu testiranja vsak tekmovalec dobi svoje izvide. Primer takega testiranja je na Sliki 1. Zaradi anonimnosti so tekmovalci označeni s črkami namesto imen. Po izločanju nepopolnih testiranj je ostalo 40 uporabnih testov.

V začetku izvida so podatki o telesni višini, teži in telesni sestavi tekmovalca. V prvi tabeli so podatki o njegovi pljučni kapaciteti (VC), količini izdihanega zraka v 1 sekundi (FV1), »tifenopinelli«-jev indeks in količini predihanega zraka v 1 minuti (MVV). Ti podatki so izmerjeni že pred začetkom testiranja. V drugem stolpcu so pričakovane vrednosti testa glede na starost, višino in telesno težo tekmovalca (le te izračunajo na fakulteti za šport); v zadnjem pa kvocient izmerjene in pričakovane vrednosti. V glavni tabeli so podatki testiranja na preprogi. Kot je že bilo omenjeno, se na vsake 4 minute poveča naklon preproge. Takrat se tekmovalec ustavi, da se mu izmeri laktat. To naredijo tako, da mu vzamejo kapljico krvi iz ušesa. Hkrati zabeležijo tudi srčni utrip. Laktat je stranski produkt mlečne kisline in daje informacije o zakisanosti telesa. VO2 je poraba kisika, kar je eden izmed najpomembnejših dejavnikov uspešnosti v smučarskem teku. Naslednji stolpec predstavlja ventilacijo, to je količino predihanega zraka v litrih na minuto. Zadnjih 5 stolpcev predstavljajo različni indeksi, ki pa nimajo tako velikega pomena kot prejšnji. Testiranje poteka, dokler tekmovalec zdrži.

Slika 1. Primer rezultatov testiranja na FŠ



UNIVERZA V LJUBLJANI **Fakulteta za šport**

1000 Ljubljana, Slovenija
Gortanova 22
Telefon: (+386 01) 5207 843
Fax: (+386 01) 5207 750
e-mail: phycenter@gmail.com

Laboratorij za fiziologijo športa

MERITVE SMUTEK — spiroergometrija (Rødby, Cosmed K4b²)
Protokol — RAMSAU 3,0 m/sec + 1° na 4 min, začetek 1°
Datum meritev — 05.07.2018 — temperatura: 23 ° C
Čas testiranj: — 12:10 — vlaž. zraka: 53 %
atmosf.tlak: 995 hPa

Ime in priimek: NN
Datum rojstva: xx.xx.xxxx
Telesna višina: 185,0 cm
Telesna teža: 75,6 kg
% maščobe v telesu: 8,8 %
% mišičja v telesu: 52,2 %

spirometrija		measured	pred. NHANES III	%
l	VC	5,71	6,03	94,7
l v 1 sec	FV1	4,62	4,95	93,3
Tiffneau index	FV1/VC x 100	80,9	85,0	95,2
l/min	MVV	138,6	148,5	93,3

La mir: 1,0 mmol/l
La max: 7,7 mmol/l
La 5 min: 7,9 mmol/l
Čas teka: 30:00 — Pretečena razdalja: 5400 m

Stage °	Time min	HR bt/min	Lactate mmol/l	VO2/kg ml/min/	V'e l/min	fR /min	RQ	EqO2	tCO2 %	tO2 %
mir	0	66	1	6,4	19,9	21,8	1,03	41,5	3,39	3,20
1	4	120	0,9	30,9	65,2	35,4	0,90	27,9	4,28	4,60
2	8	140	0,9	36,8	81,1	35,3	0,89	29,2	4,37	4,46
3	12	155	1,3	44,4	88,8	36,9	0,91	26,5	4,76	4,91
4	16	171	1,5	50,2	106,0	39,0	0,94	27,9	4,86	4,73
5	20	179	2,6	54,2	110,4	40,9	0,98	26,9	5,00	4,89
6	24	185	3,3	56,9	127,8	47,3	1,01	29,7	4,95	4,52
7	28	194	5,8	61,2	134,4	53,2	1,04	29,0	4,64	4,41
8	30	196	7,7	63,4	150,3	57,7	1,09	31,3	4,17	4,15
Lact.				9,3		7,9		6,5		

4. Začetek dela

Dela sem se lotila z izpisom podatkov v Excelovo tabelo in uvozom le te v programski jezik R. Najprej sem pregledala podatke. Izkazalo se je, da sta spremenljivki MVV in FV1 popolnoma korelirani, zato sem morala eno izločiti. Odločila sem se za MVV. Močna korelacija je prisotna še med spremenljivkama FV1 in VC (0,81) ter med mišičjem in maščobo (0,86). Najverjetneje bo potrebno izločiti tudi po eno od teh dveh spremenljivk, vendar jih zaenkrat še pustim v tabeli.

Z vgrajeno funkcijo *boxplot()* sem vsako spremenljivko posebej pregledala, če ima kaj osamelcev. Izkazalo se je, da ima spremenljivka VC dva osamelca, srčni utrip v miru in ventilacija pa vsaka po enega. Ker se osamelci pojavijo v različnih vrsticah (5., 7., 16., 20.), jih ne izločim iz tabele, saj imam že tako malo podatkov.

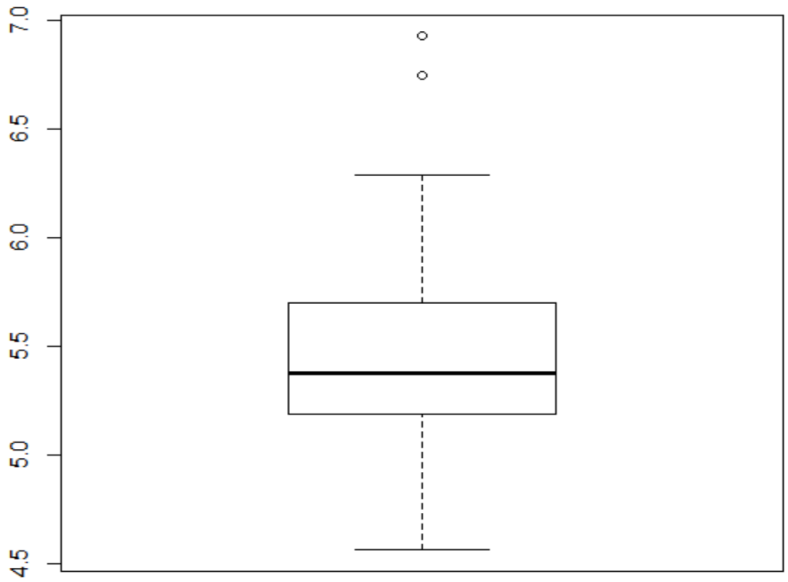
Slika 2 prikazuje škatlo z brki za spremenljivko VC. Osamelca pripadata istemu tekmovalcu.

Z vgrajenima funkcijama *density()* in *plot()* sem opazovala grafe vsake spremenljivke posebej. Zanimalo me je, če so grafi podobni normalni ali kateri drugi porazdelitvi. Graf, najbolj podoben normalni, ima spremenljivka FV1. Pri spremenljivki VC je najbolj izrazita bimodalna porazdelitev. Ugotovila sem, da drugi vrh pri vrednosti približno 6,8 l/sec povzročita osamelca iz 5. in 7. vrstice. Gre za rezultata istega tekmovalca. Če ju izločim iz tabele podatkov, se drugi vrh izravna in imam unimodalno porazdelitev. Ostali grafi nimajo večjih posebnosti. Predvidevam, da bi grafi pri večjem številu testiranj bili bolj podobni normalni porazdelitvi.

S pomočjo funkcije *ggplot()* sem vsako pojasnjevalno spremenljivko posebej primerjala s spremenljivko VO2max. Ugotovila sem, da je najmočnejša linearna povezanost prisotna med ventilacijo in VO2max. To je bilo pričakovati že iz samih definicij teh spremenljivk. Iz ostalih grafov ne morem sklepati o posebni povezanosti pojasnjevalnih spremenljivk z VO2max.

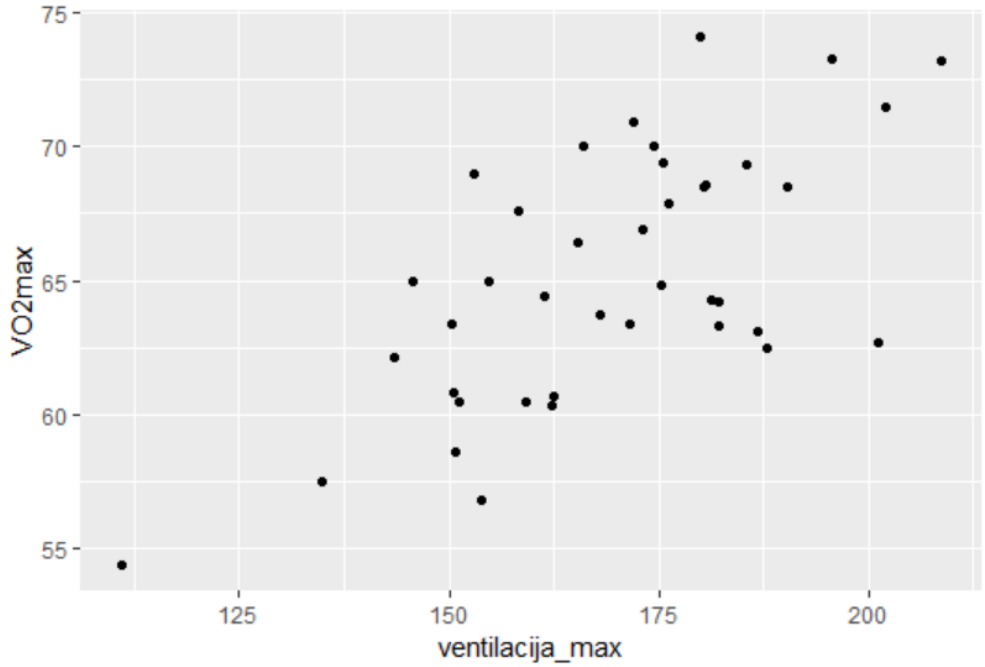
S spodnje slike je razvidna pozitivna linearna povezanost med maksimalno ventilacijo in maksi-

Slika 2. Pregled osamelcev spremenljivke VC



malno porabo kisika.

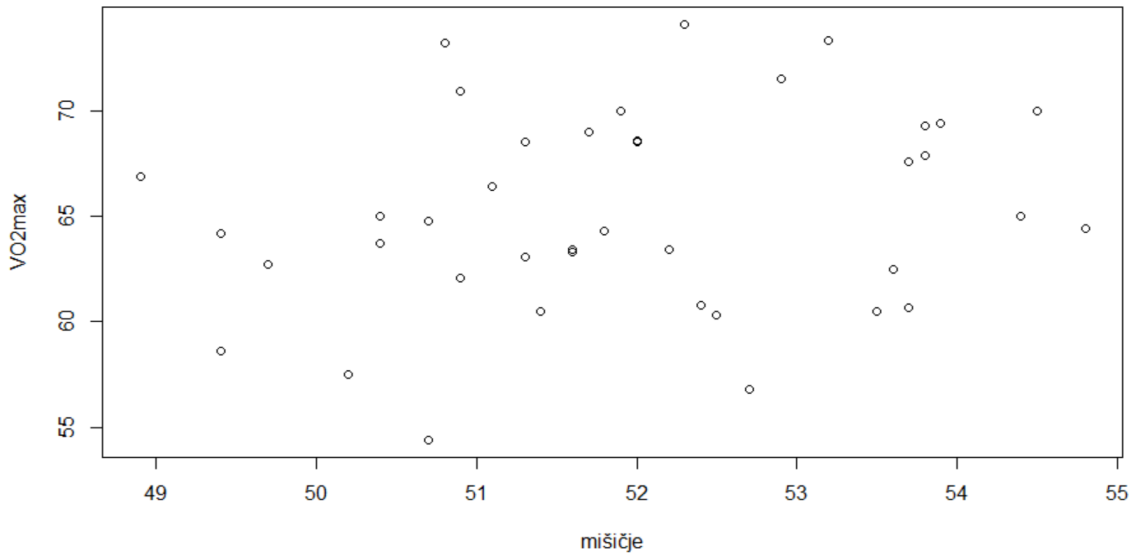
Slika 3. Povezava med ventilacijo in VO2



Z naslednje slike je razvidno, da med mišično maso in maksimalno porabo kisika ni mogoče razbrati posebne povezanosti. Predvidevam, da bi bila v primeru večjega števila podatkov razvidna pozitivna korelacija.

V literaturi ([5] Wikipedia) sem zasledila, da se da maksimalno porabo kisika dobro napovedati tudi iz kvocienta med maksimalnim srčnim utripom in utripom v mirovanju. Žal se je izkazalo, da v primeru testiranj na preprogi to ne drži. Razlog vidim v tem, da imajo tekmovalci na preprogi

Slika 4. Povezava med mišičjem in VO2



pred testiranjem tremo, zato srčni utrip v mirovanju ni povsem relevanten. Potrebno bi bilo izmeriti srčni utrip v mirovanju brez motečih dejavnikov.

V naslednjih treh razdelkih so predstavljeni postopki in rezultati generiranja modelov za napoved maksimalne porabe kisika, tekmovalne uspešnosti in maksimalnih vrednosti na podlagi 24-ih minut testiranja na tekoči preprogi.

5. Model VO2max

Najprej sem se lotila modela za napoved maksimalne porabe kisika. Za boljši pregled sem naredila novo tabelo, v kateri so le podatki, izmerjeni pred samim testiranjem na preprogi. To so FV1, VC, teža, višina, mišičje, maščoba in starost. Z ustreznim modelom bi lahko ocenili VO2max tekmovalca brez testiranja na preprogi. S tem bi zmanjšali stroške testiranj in dobili pomembne rezultate brez večjih naporov. Taka testiranja bi lahko opravljali pogosteje in s tem dosledno spremljali pripravljenost tekmovalca.

Na temo napovedi maksimalne porabe kisika športnika je bilo narejenih že več raziskav in najti je možno različne formule za izračun ocenjene vrednosti. Glede na to, da biatlonci in tekači na smučeh pogosto dosegajo najvišje vrednosti VO2, bi bilo smiselno narediti model, prilagojen značilnostim športnikov iz teh panog.

5.1 Model z MNK

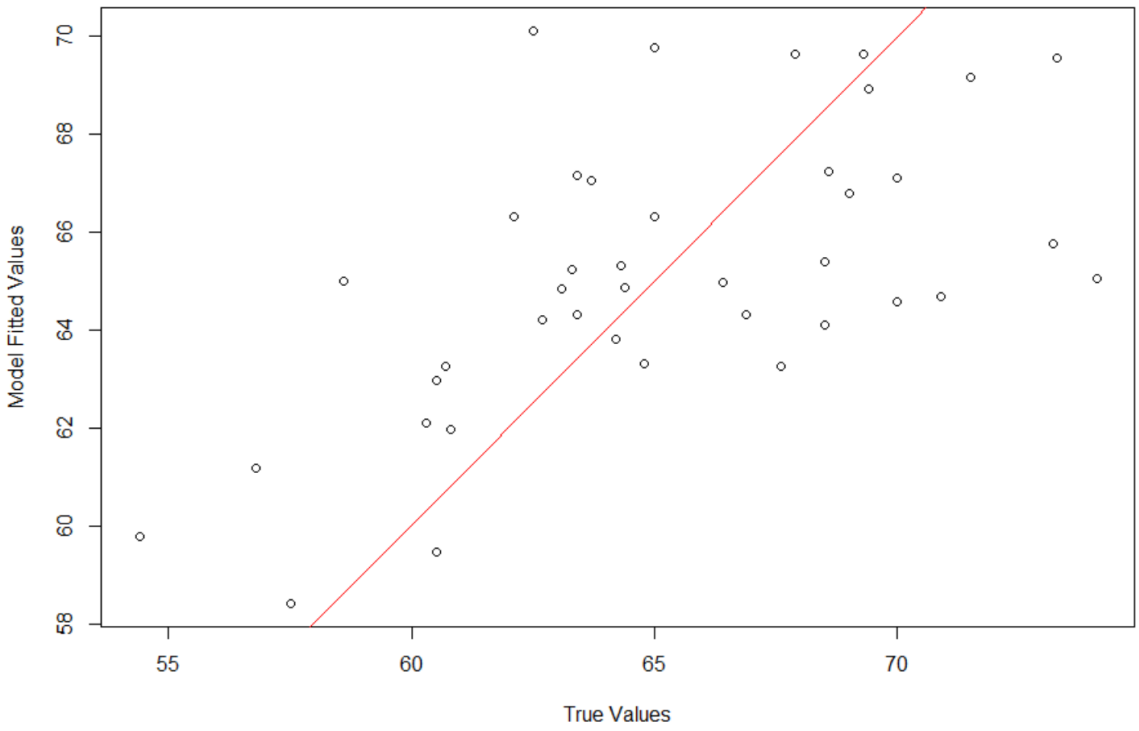
Začela sem z linearnim modelom, in sicer z ocenjevanjem parametrov modela po metodi najmanjših kvadratov. Za generiranje linearnega modela ima R že vgrajeno funkcijo *lm()*. Najprej sem uporabila vseh 7 pojasnjevalnih spremenljivk. Narisala sem graf ostankov, kjer le ti izgledajo naključni, kar je dobro. Ostanki ne smejo tvoriti nobenega vzorca, saj bi to kazalo na problem avtokorelacije. V primeru nenaključnih ostankov bi bilo smiselno linearni model zamenjati s kakšnim drugim. Model ima $R^2 = 0,3984$, kar je zadovoljivo. Višina in starost imata p-vrednost nižjo od 0,05, mišičje blizu 0,05, ostale spremenljivke pa višjo.

Model sem poskusila izboljšati s postopnim izločanjem spremenljivk. Izločala sem jih glede na stolpec p-vrednosti. Najprej sem izločila FV1 in dobila model z isto vrednostjo R^2, p-vrednosti pa

so se izboljšale. Na podlagi p-vrednosti lahko sklepam o statistični značilnosti višine, mišičja in starosti. Nato sem izločila maščobo in ponovno dobila boljši model. Nadaljevala sem z izločitvijo teže. Tu se je R^2 v primerjavi s prvim modelom zmanjšal za 0,01, p-vrednosti pa so se ponovno izboljšale. Ko sem izločila še VC, so mi ostale le še spremenljivke z dovolj nizkimi p-vrednostmi, da sem lahko sklepala, da so statistično značilne. Te spremenljivke so višina, mišičje in starost. R^2 v tem modelu je 0,3698. Začetni model je imel sicer boljšo vrednost R^2 (0,3984), vendar pa so bile $p - vrednosti$ previsoke za sklepanje o statistični značilnosti regresijskih koeficientov.

Slika 5 prikazuje ujemanje pravih in ocenjenih vrednosti maksimalne porabe kisika v modelu, kjer so parametri ocenjeni po MNK. Ta model vključuje le statistično značilne pojasnjevalne spremenljivke. Zaželjeno je, da so vrednosti porazdeljene čimbližje rdeči premici, torej simetrali lihih kvadrantov.

Slika 5. Ujemanje ocenjenih in pravih vrednosti VO2max



5.2 Model z logaritmiranjem

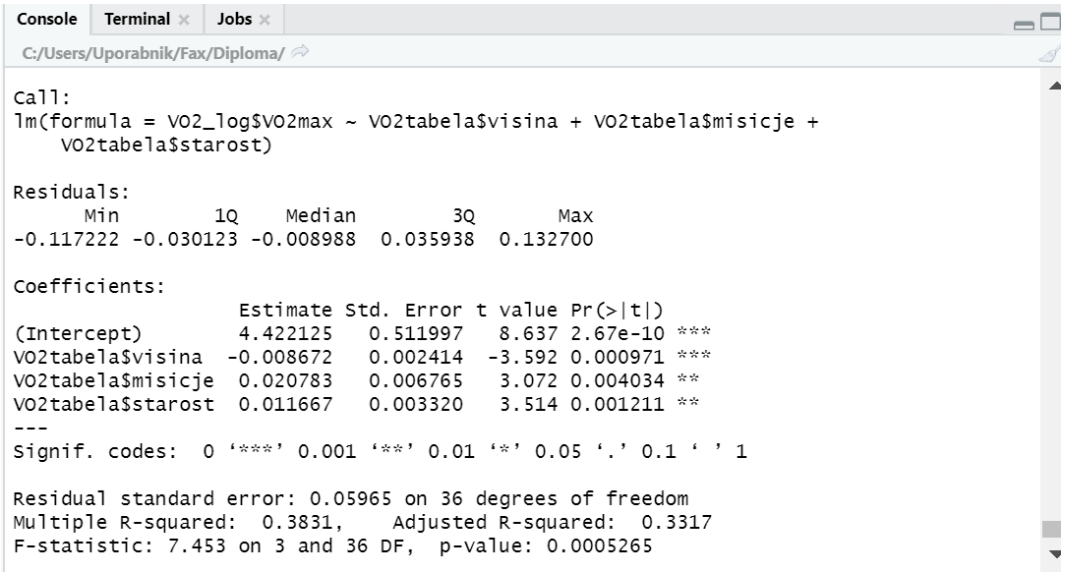
Poskusila sem z logaritmiranjem spremenljivke VO2max. To sem naredila tako, da sem v tabeli v stolpcu z vrednostmi VO2max logaritmirala vsako vrednost posebej. Dobljeni model ima $R^2 = 0,413$, vendar pa ta vrednost ni primerljiva s prejšnjim modelom iz razdelka 5.1, ker je odvisna spremenljivka drugače definirana. Že takoj sem lahko sklepala o statistični značilnosti pojasnjevalnih spremenljivk višina, mišičje in starost, saj imajo vrednosti nižje od 0,05. Ponovno sem poskusila z izločanjem spremenljivk priti do boljšega modela. Zopet sem uspela priti do modela z nižjimi p-vrednostmi pojasnjevalnih spremenljivk in posledično sem lahko sklepala le o statistični značilnosti višine, mišičja in starosti. Ostale spremenljivke imajo previsoke p-vrednosti.

Ugotovila sem še, da proste konstante ne smem izločiti iz modela, saj tako dobljen model ni pravilno specificiran. Prosto konstanto sem poskusila izločit, ker je imela visoko p-vrednost, vendar pa je R^2 takoj narastel na 0,99, kar je kazalo na napako. Kot sem navedla v teoriji, do tega tipično

pride, če je število pojasnjevalnih spremenljivk večje od števila testiranj, vendar pa v tem primeru temu ni tako. Če iz modela odstranimo prosto konstanto, moramo R^2 izračunati po drugi formuli. Determinacijski koeficient za model brez regresijske konstante označimo z R^2_*. Še vedno pa velja $0 \leq R^2_* \leq 1$. Programski jezik R avtomatsko izračuna R^2 po formuli, ki velja le za regresijske modele, ki vključujejo konstantni člen. Primerjava vrednosti R^2_* in R^2 ne more biti podlaga za izbiro med dvema regresijskima modeloma, saj je prvi izračunan na podlagi kvadratov odklonov od vrednosti 0, v v drugem primeru pa na podlagi kvadratov odklonov od aritmetične sredine odvisne spremenljivke.

Na spodnji sliki je izpis iz programskega jezika R. Predstavljen je povzetek zadnjega modela z logaritmiranjem spremenljivke $VO2max$. Razvidno je, da so vse $p - vrednosti$ dovolj nizke, da lahko sklepamo o statistični značilnosti.

Slika 6. Povzetek modela z logaritmiranjem



5.3 Posplošeni linearni modeli

V linearnem modelu $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$ velja $E(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$. Slednjo enačbo lahko z uporabo primerno definirane funkcije g posplošimo do

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}.$$

Tu z indeksom i označujemo i-tega posameznika. Prejšnja enačba je poseben primer, kjer je g identiteta.

Z uporabo funkcije *glm()* in znotraj primerno definirane funkcije g sem generirala več posplošenih linearnih modelov. Najprej sem za g uporabila logaritemsko funkcijo, za porazdelitev ostankov pa normalno porazdelitev. Z vključitvijo vseh 7 pojasnjevalnih spremenljivk sem dobila model, kjer lahko sklepam o statistični značilnosti spremenljivk starost in višina. Ponovno ima mišičje p-vrednost blizu 0,05, ostale spremenljivke pa več. Ponovila sem postopek s postopnim izločanjem spremenljivk in na koncu dobila model s prosto konstanto in pojasnjevalnimi spremenljivkami starost, višina in mišičje. Vse p-vrednosti so nižje od 0,01. Povzetek slednjega modela si lahko ogledamo na Sliki 7. Podobne rezultate sem dobila, ko sem za funkcijo g uporabila inverzno funkcijo. Ocene regresijskih koeficientov se sicer razlikujejo, ostanejo pa iste pojasnjevalne spremenljivke s p-vrednostmi, nižjimi od 0,01.

Slika 7. Povzetek posplošenega linearnega modela z logaritemsko funkcijo

```
Call:
glm(formula = Tabela_test$VO2max ~ VO2tabela$starost + VO2tabela$visina +
    VO2tabela$misicje, family = gaussian(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5894  -2.0309  -0.6419   2.3939   9.0936

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.436801   0.513530   8.640 2.65e-10 ***
VO2tabela$starost  0.011068   0.003357   3.297  0.00220 **
VO2tabela$visina  -0.008363   0.002449  -3.415  0.00159 **
VO2tabela$misicje  0.019694   0.006766   2.911  0.00615 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 15.3848)

    Null deviance: 867.11  on 39  degrees of freedom
Residual deviance: 553.85  on 36  degrees of freedom
AIC: 228.64

Number of Fisher Scoring iterations: 4

> |
```

5.4 Linearni mešani modeli: Vzdolžni model

Glede na to, da je vsak posameznik opravil več testov v različnih časovnih obdobjih, so ti testi med seboj korelirani. Iz tega razloga bi moral biti vzdolžni model najbolj primeren za napoved. Za generiranje vzdolžnega modela sem v tabelo dodala nov stolpec *oseba*, kjer sem vsakemu tekmovalcu priredila število med 1 in 8. Najprej sem uporabila knjižnico *nlme* in pripadajočo funkcijo *lme*(). Ta avtomatsko uporablja restringirano metodo največjega verjetja (RMNV). Z dobljenimi p-vrednostmi nisem bila zadovoljna, saj nisem mogla sklepati o statistični značilnosti nobene od spremenljivk. Zato sem ponovno izločala spremenljivke, dokler nisem prišla do modela, kjer sem lahko sklepala o statistični značilnosti vseh preostalih pojasnjevalnih spremenljivk. Ponovno so se za take spremenljivke izkazale višina, mišičje in starost. Poskusila sem še z osnovno metodo največjega verjetja. Prvotne p-vrednosti so bile boljše kot pri prejšnji metodi. Že v prvem poskusu se je kot statistično značilna izkazala višina. Po izločanju spremenljivk so ponovno ostale višina, mišičje in starost, ki pa imajo nižje p-vrednosti kot pri RMNV. Nato sem poskusila še z uporabo knjižnice *lme*4, ki je novejša od *nlme*. Z vgrajeno funkcijo *glmer*() sem dobila model z zelo nizkimi p-vrednostmi spremenljivk višina, mišičje in starost. Tudi po izločanju spremenljivk mi ostanejo iste 3 pojasnjevalne spremenljivke, p-vrednosti pa so najnižje do zdaj.

5.5 Modeli po podskupinah

Zaradi suma vpliva razvoja mladincev na model sem naredila še modele po podskupinah. Najprej sem posebej naredila tabeli za člane in mladince in ponovno uporabila modele po MNK. Presenetljivo je bil za model mladincev $R^2 = 0,7543$, za model članov pa 0,2268. Vseeno pa so bile p-vrednosti v obeh modelih previsoke, da bi lahko sklepala o statistični značilnosti katerekoli od spremenljivk. Po izločanju spremenljivk v modelu mladincev kot statistično značilni ostaneta mišičje in FV1, $R^2 = 0,6748$, v modelu članov pa se kot statistično značilna spremenljivka izkaže samo mišičje. Rezultati so me presenetili, saj sem pričakovala, da bodo težave v modelu povzročali mladinci, ker se še razvijajo in se njihovi rezultati pri vsakem testiranju zelo spremenijo.

Izkazalo se je, da model članov slabše napoveduje, zato sem želela ugotoviti, v čem je problem. Pogledala sem razsevne grafikone in poiskala osamelce. Ko sem iz modela izločila vse 3 najdene

osamelce, se je R^2 izboljšal na 0,3073. To še vedno ni najboljši model. Rezultate pripisujem premajhnemu številu testirancev. Testiranja posameznega člana so si med seboj zelo podobna, tako da bi bolj kot več testiranj potrebovala več testirancev. Podobni testi lahko vodijo do problema multikolinearnosti. Razliko med modeloma pripisujem tudi večji razpršenosti podatkov mladincev. Večja kot je variabilnost pojasnjevalne spremenljivke, bolj zanesljive so ocene regresijskih koeficientov (Fajfar, 2018, str.90).

5.6 Zadnja testiranja

Na koncu sem naredila še tabelo, v kateri so spravljeni le zadnji testi vsakega leta. To so načeloma najboljši testi vsakega tekmovalca za vsako leto, saj naj bi bili biatlonci in tekači najbolje pripravljeni jeseni, ko se približuje tekmovalno obdobje. S funkcijo *ggplot*() sem pogledala razsevne grafikone, da bi preverila, kakšne povezave obstajajo med pojasnjevalnimi spremenljivkami in VO2max. Grafi kažejo na pozitivno linearno povezanost VO2max s starostjo in mišičjem. Pri spremenljivki FV1 izgleda, kot da ima negativno povezavo z VO2max. Pri ostalih spremenljivkah se iz grafa ne da razbrati, kakšne bi bile povezave. Ponovno sem naredila vzdolžni model, tokrat le iz zadnjih testov vsakega leta. Dobljeni model je imel vse p-vrednosti višje od 0,05. Po izločanju se je kot statistično značilna spremenljivka izkazala samo FV1.

5.7 Ugotovitve

Kot statistično značilne spremenljivke se v skoraj vseh modelih izkažejo višina, mišičje in starost. V večini primerov imajo ocene regresijskih koeficientov pri mišičju in starosti pozitiven, pri višini pa negativen predznak. Pozitiven vpliv mišičja je bil pričakovan, saj delež mišičja posredno kaže na telesno pripravljenost tekmovalca. Zaželjen je visok delež mišičja, vseeno pa delež maščobe ne sme biti prenizek, saj bi to lahko ogrožalo zdravje.

V teoriji ([5] Wikipedia) sem zasledila, da se VO2max s starostjo zmanjšuje, zato sem pri oceni regresijskega koeficienta starosti pričakovala negativen predznak. Tu je treba poudariti, da so vsi tekmovalci v vzorcu mlajši od 30 let, VO2max pa naj bi začel upadati šele po 30. letu starosti. Pozitiven predznak si tako razlagam z napredkom tekmovalcev skozi čas. Starejši kot so, bolje so pripravljeni in posledično dosegajo višje vrednosti VO2max. Večina tekačev na smučeh in biatloncev svoj vrhunec doseže okoli 30. leta starosti.

Negativen predznak ocene regresijskega koeficienta višine je težje pojasniti. Možno je, da je do tega prišlo zaradi majhnega vzorca (imam namreč le 13 testirancev). Druga možna razlaga izhaja iz osnovne formule VO2max, ki se izraža v $\frac{ml}{kg \cdot min}$. Višji kot je tekmovalec, večja je njegova telesna masa, le ta pa negativno vpliva na posameznikov VO2max. V teku na smučeh na dolge proge boljše rezultate dosegajo tekmovalci z manjšo telesno maso.

Model mladincev se je izkazal za obetavnega. Sklepam, da zato, ker imajo bolj razpršene vrednosti pojasnjevalnih spremenljivk. Zanimivo bi bilo spremljati njihov razvoj, za kar bi bilo potrebno opazovanje skozi daljše časovno obdobje.

Za izboljšavo modelov bi bilo potrebno pridobiti bistveno več podatkov. Raziskava bi lahko zajemala več let in več testirancev.

6. Model tekmovalne uspešnosti

Model tekmovalne uspešnosti sem lahko naredila le za 3 biatlonce. Tekači na smučeh ne tekmujejo na istih tekmovanjih, zato jih ne morem primerjati. Eden izmed biatloncev v sezoni 2016/2017 ni opravljal testiranj z dihalno masko, v sezoni 2017/18 pa je bil poškodovan, zato ga prav tako ni bilo

mogoče uporabiti kot podatek. Mladinci tekmujejo na drugem nivoju, zato njihovi rezultati prav tako niso primerljivi.

Za izdelavo modela sem najprej uvozila 2 novi tabeli. V prvi tabeli so rezultati strelskih testov pod obremenitvijo, in sicer uspešnost streljanja leže, stoje in skupaj, izražena v procentih zadetih strelov izmed vseh strelov. Druga tabela vsebuje podatke o povprečnih zaostankih na tekmovanjih (izraženi v procentih zaostanka za prvouvrščenim), povprečnem mestu na tekmovanjih in strelski uspešnosti na tekmovanjih v sezoni (izraženi v procentih zadetih strelov).

Najprej sem poskusila iz strelskih testov napovedati povprečno strelsko uspešnost na tekmovanjih. Žal je model neuporaben, saj z njim lahko pojasnim le 3% variabilnosti pri uspehu v streljanju na tekmovanjih. Podgornik je v svoji diplomski nalogi ugotovil, da na osnovi psiholoških spremenljivk predtekmovalnih stanj in stresa lahko pojasnimo 39% variabilnosti pri uspehu v skupnem biatlonskem streljanju. Stoje lahko pojasnimo 56% variabilnosti, leže pa le 27%, kar je statistično nepomembno (Podgornik, 2013).

Po mojem mnenju na strelsko uspešnost vpliva tudi fizična pripravljenost. Tekmovalec mora biti stabiliziran, da lahko drži puško, ki je težka vsaj $3,5kg$. Poleg tega fizično slabše pripravljen tekmovalec ponavadi pride do strelišča bolj utrujen in zadihan. Za preverjanje te domneve pa bi bilo potrebnih veliko več podatkov, kot sem jih uspela zbrati.

Glede na to, da se tekmovalci od spomladi do jeseni bolj razvijejo in da jesenski testi bolje odražajo pripravljenost pred zimsko sezono, sem vzela le zadnje teste vseh tekmovalcev iz vsakega leta. Skupno tabelo za model tekmovalne uspešnosti sem naredila tako, da sem združila tabele tekmovalne strelske uspešnosti, strelskih testov in zadnjih testov s Fakultete za šport. Uporabila sem lahko le podatke za 3 tekmovalce in za vsakega po 2 leti, torej skupaj 6 testiranj. Ker mora biti število opazovanih enot večje od števila pojasnjevalnih spremenljivk, sem za model lahko uporabila največ 5 pojasnjevalnih spremenljivk. Napovedati sem želela stolpec *tek.str.*, ki predstavlja povprečno mesto tekmovalca v sezoni. Stolpec *skupaj* predstavlja skupno strelsko uspešnost na treningih v sezoni, *cas_teka* dosežene minute na testiranju na FŠ, stolpci $VO2max, ventilacija, visina, misicje, teza, FV1, VC, ...$ pa so iz tabele, ki sem jo uporabila že pri modelu za napoved VO2max.

Najprej sem preverila, kako močna je povezanost med spremenljivkami. Koreliranost med spremenljivkami s testiranj s FŠ sem preverila že pri modelu VO2max, tako da sem morala preveriti le korelacijo z ostalimi spremenljivkami. Nikjer ni bilo tako močne koreliranosti, da bi bilo potrebno kakšno spremenljivko izločit iz modela.

Nato sem poskusila narediti model, kjer sem kot pojasnjevalne spremenljivke vzela VO2max, čas teka, ventilacijo in skupno strelsko uspešnost na treningih v sezoni. Napovedati sem želela povprečno mesto tekmovalca v sezoni. Model po MNK je imel $R^2 = 0,9057$, vendar pa se nobena pojasnjevalna spremenljivka ni izkazala za statistično značilno. Tudi izločanje spremenljivk ni pomagalo.

Za model, kakršnega sem želela narediti, bi morala imeti bistveno več podatkov, zato je neuporaben. Na tekmovalno uspešnost vpliva bistveno več spremenljivk kot sem jih uspela pridobiti s testiranj. Po mojem mnenju bi se tudi iz teh spremenljivk dalo napovedati precejšen delež variabilnosti v biatlonski tekmovalni uspešnosti, vendar pa bi bilo potrebno vzorec testirancev bistveno povečati. V Sloveniji to žal ni mogoče, ker imamo premalo tekmovalcev. Taka raziskava bi bila mogoča le na nivoju svetovnega pokala. Poleg tega bi bilo smiselno opraviti tudi psihološko oceno tekmovalcev, saj ima psihološko stanje velik vpliv tako na strelski kot tudi tekaški del tekmovanja.

7. Napoved na podlagi 24 minut

Zanimalo me je, ali se da na podlagi rezultatov testiranja do nekega časa napovedati maksimalne vrednosti testiranja. Uporabila sem rezultate s 24 minut, saj so do teh minut prišli vsi testiranci.

Nekaterim testirancem so vrednosti izmerili na vsake 3, drugim na vsake 4 minute, na 24-ih pa se torej vsi ustavijo. Lahko bi uporabila še rezultate z 12-ih minut, vendar so se mi zdeli rezultati bližje maksimalnim vrednostim bolj ustrezni.

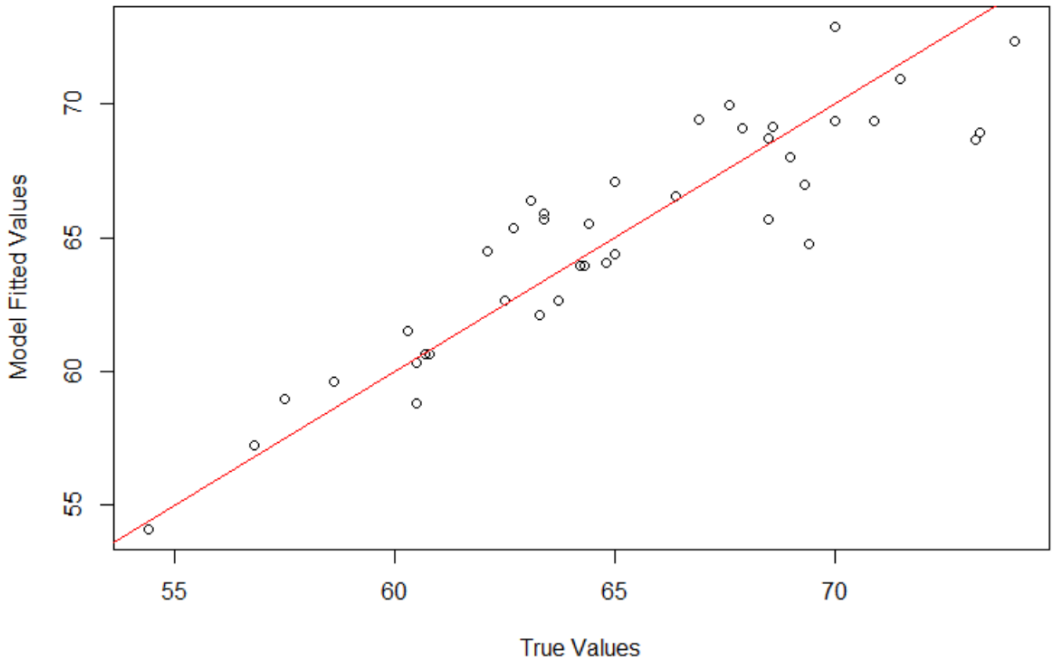
Najprej sem uvozila tabelo z rezultati na 24-ih minutah. Tabelo z maksimalnimi vrednostmi testiranj sem uvozila že za model VO2max, zaradi boljše preglednosti sem spremenila imena stolpcev. Ti dve tabeli sem nato združila v skupno tabelo. Ponovno sem preverila, da med spremenljivkami ni previsoke koreliranosti.

Nato sem s podatki iz 24-ih minut poskusila napovedati vsako spremenljivko maksimalnih vrednosti posebej. Za pojasnjevalne spremenljivke sem torej vzela VO2, ventilacijo, laktat in srčni utrip, vse izmerjene na 24-ih minutah. Parametre modelov sem ocenjevala po MNK.

Ugotovitve: Z laktatom in VO2 lahko pojasnim $82,05\%$ variabilnosti v maksimalni vrednosti VO2. Do tega rezultata sem prišla z izločitvijo ventilacije in srčnega utripa.

Slika 8 prikazuje ujemanje pravih vrednosti $VO2max$ z ocenjenimi na podlagi laktata in VO2 pri 24-ih minutah testiranja. Vrednosti se dobro prilegajo simetrali lihih kvadrantov.

Slika 8. Ujemanje pravih in ocenjenih vrednosti VO2max



Z laktatom, VO2, ventilacijo in srčnim utripom lahko pojasnim $80,68\%$ variabilnosti v maksimalni vrednosti ventilacije. Z laktatom in ventilacijo lahko pojasnim $80,1\%$ variabilnosti v doseženem času testiranja. Z laktatom in srčnim utripom lahko pojasnim $87,74\%$ variabilnosti v maksimalni vrednosti srčnega utripa.

Pri napovedi maksimalne vrednosti laktata sem naletela na težave, saj je imel prvotni model $R^2 = 0,0525$, zato sem poskusila z dodajanjem spremenljivk, katerih vrednosti so bile izmerjene že pred samim testiranjem na preprogi. S tem sem R^2 uspela nekoliko izboljšati, vendar pa so bile vrednosti še vedno prenizke. Prav tako se nobena spremenljivka ni izkazala za statistično značilno, zato sem sklenila, da na podlagi vrednosti, dobljenih na 24-ih minutah ne morem sklepati o doseženi maksimalni vrednosti laktata. Razlog vidim v tem, da se vsakemu posamezniku telo drugače odziva, poleg tega pa je zakisanost telesa odvisna še od mnogih drugih dejavnikov.

8. Zaključek

Linearne modele za napovedovanje v tem delu ocenjujem kot uspešne. Zelo dobro lahko na podlagi 24-ih minut testiranja napovedujemo maksimalne vrednosti testiranja tekmovalcev, kar pomeni, da bi lahko tekmovalce na preprogi ustavili že pred največjim naporom in na podlagi dotedanjih vrednosti z veliko verjetnostjo napovedali maksimalne. Modeli za napoved maksimalne porabe kisika so se izkazali za zadovoljive. Verjamem, da bi se jih dalo izboljšati s povečanjem vzorca, mogoča pa je tudi njihova razširitev ob uvedbi dodatnih spremenljivk, kot je na primer število opravljenih ur treninga do samega testiranja. V tem delu smo na podlagi spremenljivk, merjenih v mirovanju, uspeli pojasniti okoli 40% variabilnosti slučajne spremenljivke VO2max, kar je še vedno premalo, da bi lahko povsem opustili testiranja pod obremenitvijo. Na izide testiranja vedno vplivajo še slučajni dejavniki, kot je vreme, dnevno počutje, vlaga zraka,... Le te bi se prav tako dalo upoštevati v modelu, potrebovali pa bi več podatkov o testirancih in pogojih. Zagotovo je z uporabo linearnih modelov mogoče napovedovati mnoge parametre v športu, zavedati pa se je potrebno, da je vsako tekmovanje zgodba zase in vsak tekmovalec odreagira drugače. Zanimivo bi bilo v modele vključiti še psihološko komponento, ki pa je težje merljiva.

LITERATURA

[1] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics, Springer Science + Business Media, LLC, New York, 2007.

[2] D. Bates et al., *Package 'lme4'*, v: Linear Mixed-Effects Models using 'Eigen' and S4, [ogled 4.6.2019], dostopno na `https://cran.r-project.org/web/packages/lme4/lme4.pdf`.

[3] *lme*, v: RDocumentation, [ogled 15.5.2019], dostopno na `https://www.rdocumentation.org/packages/nlme/versions/3.1-137/topics/lme`.

[4] *glm*, v: RDocumentation, [ogled 15.5.2019], dostopno na `https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/glm`.

[5] *VO2 max*, v: Wikipedia: The Free Encyclopedia, [ogled 20.5.2019], dostopno na `https://en.wikipedia.org/wiki/VO2_max`.

[6] *Box plot*, v: Wikipedia: The Free Encyclopedia, [ogled 20.5.2019], dostopno na `https://en.wikipedia.org/wiki/Box_plot`.

[7] V. Maver, *Normalni linearni mešani modeli*, diplomsko delo, Fakulteta za matematiko in fiziko, Univerza v Ljubljani, 2018.

[8] M. Podgornik, *Značilnosti reagiranja v stresu pri streljanju v biatlonu*, diplomsko delo, Fakulteta za šport, Univerza v Ljubljani, 2013.

[9] T. Kraner Šumenjak, *Statistika 2.predavanje*, [ogled 14.6.2019], dostopno na `http://www.fk.uni-mb.si/images/stories/matematika/2pred-stat1.pdf`

[10] *Bimodal Distribution: What is it?*, v: Statistics How To, [ogled 20.5.2019], dostopno na `https://www.statisticshowto.datasciencecentral.com/what-is-a-bimodal-distribution/`.

[11] M. Pohar Perme, *Statistika 1*, [ogled 31.7.2019], dostopno na `https://ucilnica.fmf.uni-lj.si/pluginfile.php/58675/mod_resource/content/18/predavanja_FMF_nova19s.pdf`.

[12] L. Pfajfar, *Osnovna ekonometrija*, učbeniki Ekonomske fakultete, Ljubljana, 2018.