Modeliranje pojavnosti endemske nefropatije logističkom regresijom

Šekoranja, Maja

Master's thesis / Diplomski rad

2015

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:564721

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Maja Šekoranja

MODELIRANJE POJAVNOSTI ENDEMSKE NEFROPATIJE LOGISTIČKOM REGRESIJOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, 07, 2015

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Veliku zahvalnost, u prvom redu, dugujem prof. dr. sc. Bojanu Jelakoviću, Medicinskom fakultetu u Zagrebu, te svim sudionicima projekta "Endemska nefropatija u Hrvatskoj, epidemiologija, dijagnostika, etiopatogeneza", koji su mi omogućili korištenje svoje baze podataka o endemskoj nefropatiji.
Također, zahvaljujem se svojoj mentorici prof. dr. sc. Anamariji Jazbec koja mi je svojim savjetima pomogla pri izradi ovog diplomskog rada.
Velika hvala svima!

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Logistička regresija **3**
1.1 Regresijska analiza 3
1.2 Definicija i osnovni pojmovi 3
1.3 Logistički model 4
1.4 Testiranje adekvatnosti modela („Goodness of fit“) 5
1.5 Problem konvergencije 6
1.6 Šansa i omjer šansi 6
1.7 Statistika za računanje prediktivne snage modela 8

2 Primjer **11**
2.1 Opis varijabli 11
2.2 Deskriptivna statistika 14
2.3 Univarijatna logistička regresija 16
2.4 Multivarijatna logistička regresija 18
2.5 Zaključak 47
2.6 Prilog 48

Bibliografija **55**

Uvod

Endemska nefropatija, poznata još i pod nazivom balkanska endemska nefropatija, upalna je bolest bubrega koja se javlja na točno određenim područjima na Balkanu. Prvi slučaj takve bolesti opažen je 1920. godine, u malim zajednicama naseljenim uz rijeku Dunav i njene pritoke, na području današnjih zemalja Hrvatske, Bosne i Hercegovine, Srbije, Rumunjske te Bugarske. Područja na kojima je primijećena pojava endemske nefropatije uglavnom su ruralna, i među njima ne postoji nikakva povezanost, osim same bolesti.

Endemska nefropatija je spora i progresivna bolest, koja obično završava smrtnim ishodom. Današnja pretpostavka jest da je uzročnik neki agens iz okoliša, to jest točnije fitotoksin biljke Vučja stopa (lat. Aristolochia clematitis).

Žarište endemske nefropatije u Hrvatskoj čini 14 sela (Kaniža, Banovci, Bebrina, Zbjeg, Šumeće, Slavonski Dubočac, Slavonski Kobaš, Živike, Pričac, Malino, Slobodnica, Stupnički Kuti, Brodski Varoš i Lužani), koji sve skupa broje oko 10.800 stanovnika. Postavljanje dijagnoze endemske nefropatije temelji se na kombinaciji epidemioloških i laboratorijskih parametara.

Baza podataka korištena za modeliranje logističkom regresijom, nastala je u sklopu projekta "Endemska nefropatija u Hrvatskoj, epidemiologija, dijagnostika, etiopatogeneza" (šifra projekta: 108-0000000-0329; voditelj projekta: prof. dr. sc. Bojan Jelaković, KBC Zagreb). Originalni uzorak sadržavao je 2.378 slučajeva (observacija), iz 9 endemskih i 3 kontrola sela. Za potrebe ovog rada, napravljena je randomizacija uzorka iz postojeće baze, pri čemu je posebno randomiziran uzorak endemskih, te posebno kontrolnih sela, kako bi omjer slučajeva iz endemskih i kontrolnih sela bio kao u originalnoj bazi. Uzorak korišten u ovome radu ima 1.189 opažanja, i rezultate ovoga rada nije dopušteno koristiti kao referencu za utvrđivanje rizičnih faktora pojavnosti endemske nefropatije.

Poglavlje 1

Logistička regresija

1.1 Regresijska analiza

Općenito, regresijska analiza je metoda ispitivanja ovisnosti jedne (zavisne) varijable o jednoj ili više drugih (nezavisnih) varijabli.

Jedan od rezultata regresijske analize jest regresijski model. To je matematička jednadžba koja kvantificira povezanost između zavisne i nezavisne, odnosno zavisne i nezavisnih varijabli. Ukoliko se ispituje ovisnost zavisne varijable o jednoj nezavisnoj varijabli, rezultat regresijske analize je univarijatni regresijski model, dok je u slučaju ispitivanja ovisnosti zavisne varijable o dvije ili više nezavisnih rezultat regresijske analize multivarijatni regresijski model.

Zavisna varijabla se općenito označava sa y, i to je varijabla koja se želi opisati ili procijeniti. Nezavisna varijabla općenito se označava sa x i to je varijabla pomoću koje se želi opisati zavisna varijabla. U slučaju multivarijatne regresijske analize gdje se ispituje ovisnost zavisne varijable o n nezavisnih varijabli, pri čemu je $n \geq 2$, nezavisne varijable općenito se označavaju sa $x_1, \ldots, x_n$.

1.2 Definicija i osnovni pojmovi

Logistička regresija jest regresijska analiza u kojoj je varijabla odgovora, to jest zavisna varijabla, diskretna. To znači da varijabla odgovora može poprimiti dvije ili više vrijednosti. Za razliku od linearne regresije, nema pretpostavki za distribuciju nezavisnih varijabli.

Logistička funkcija $p : \langle -\infty, \infty \rangle \to [0, 1]$ definirana je sa $p(x) = \frac{1}{1+e^{-x}}$. To zapravo znači da je logistička funkcija ograničena u intervalu $[0, 1]$.

Logit funkcija jest funkcija inverzna logističkoj. Dakle, vrijedi $logit : [0, 1] \to \langle -\infty, \infty \rangle$, te

$$logit(x) = log\Big[\frac{x}{1-x}\Big] = log(x) - log(1-x), \tag{1.1}$$

pri čemu je $x \in [0, 1]$.

1.3 Logistički model

Najveći problem sa linearnim vjerojatnosnim modelom jest da su vjerojatnosti ograničene sa 0 i 1, ali linearne funkcije su neograničene. Iz tog razloga rješenje je transformirati vjerojatnost, kako bi maknuli donju i gornju granicu. Prvi korak je transformirati vjerojatnost u šansu, što miče gornju granicu. Drugi korak je logaritmirati šansu, što miče donju granicu.

Ako govorimo o univarijatnom logističkom modelu slijedi da je

$$log(ODD) = log\Big[\frac{p(x)}{1-p(x)}\Big] = logit(p(x)) = \beta_0 + \beta_1 x. \tag{1.2}$$

U tom slučaju koeficijent β_0 pridružen je sjecištu, a β_1 je koeficijent uz nezavisnu varijablu. Ako je $y \in [0, 1]$ varijabla odgovora, očekivanje od y uz uvjet x jednako je $\mathrm{E}[y|x] = \beta_0 + \beta_1 x$. No, istovremeno vrijedi $\mathrm{E}[y|x] = p(x)$, što povlači

$$\frac{e^{\beta_0+\beta_1 x}}{1 + e\beta_0 + \beta_1 x} = p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} \tag{1.3}$$

$$logit(p(x)) = ln\Big[\frac{p(x)}{1-p(x)}\Big] = \beta_0 + \beta_1 x \tag{1.4}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} p(x) = \frac{e\beta_0 + \beta_1 x}{1 + e^{\beta_0+\beta_1 x}} \tag{1.5}$$

Slično vrijedi i za multivarijatni logistički model. Tada je

$$logit(p(x)) = log[\frac{p(x_1, x_2, \ldots, x_k)}{1 - p(x_1, x_2, \ldots, x_k)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \tag{1.6}$$

pri čemu je k broj opisnih(nezavisnih) varijabli u modelu.

$$p(x_1, x_2, \ldots, x_k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}. \tag{1.7}$$

Koeficijenti $\beta_0, \beta_1, \ldots, \beta_k$ imaju jednostavnu interpretaciju u uvjetima omjera šansi. LOGIT model usko je vezan sa LOGLINEARNIM modelom, ima poželjne uzoračke pretpostavke, te se može lako generalizirati da dozvoli višestruke, ne poredane kategorije zavisne varijable.

1.4 Testiranje adekvatnosti modela („Goodness of fit“)

Metoda maksimalne vjerodostojnosti (u daljnjem tekstu „ML“) je često korišteni pristup za procjenu raznih statističkih modela. Upravo ML metodu koristimo za testiranje adekvatnosti logističkog modela.

Kod ML metode tražimo najmanje moguće odstupanje („DEVIANCE“) između opaženih vrijednosti (y) i prediktivnih vrijednosti ($\hat{y}$). To radimo koristeći iterativne računalne metode sve dok ne dobijemo najmanje moguće odstupanje. Dobiveno rješenje zovemo „DEVIANCE “ ili „-2LogLikelihood“ ili „Likelihood ratio“. Označimo sa $\hat{y}$ ML procjenu od y, a sa $\hat{p}(x)$ procjenu od $p(x)$. Također, neka je $\beta = (\beta_0, \beta_1)$, pri čemu su β_0 i β_1 parametri univarijatnog logističkog modela. Sa „L“ označimo „Likelihood“.

$$L(\beta) = \prod_{i=1}^{n} [p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}], \tag{1.8}$$

pri čemu su (x_i, y_i), $i = 1, \ldots, n$ promatrane vrijednosti. Princip ML metode jest da procjena parametara β maksimizira izraz $L(\beta)$.
Označimo sada sa „LL“ izraz $log(L(\beta))$. Sada slijedi

$$LL(\beta) = \sum_{i=1}^{n} [y_i ln(p(x_i)) + (1 - y_i) ln(1 - p(x_i))] \tag{1.9}$$

Maksimiziramo dobiveni izraz tako da $p(x)$, koja je definirana sa β_0 i β_1 parcijalno deriviramo po β_0 i β_1.

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0$$

$$\sum_{i=1}^{n} x_i [y_i - p(x_i)] = 0$$

Označimo sa „D“ DEVIANCE. Slijedi

$$\begin{aligned} D &= -2ln\Big[\frac{\text{Likelihood modela}}{\text{Likelihood saturiranog modela}}\Big] \\ &= -2\sum_{i=1}^{n}\Big[y_i ln\Big(\frac{\hat{p}(x_i)}{y_i}\Big) + (1-y_i)ln\Big(\frac{1-\hat{p}(x_i)}{1-y_i}\Big)\Big] \\ &= \chi^2 \end{aligned}$$

Saturirani model podrazumijeva onaj model koji sadrži onoliko parametara koliko ima podataka.

Ukoliko želimo testirati razliku modela sa i bez nezavisnih varijabli, odnosno prediktora, koristimo G statistiku. Slično kao za R^2 kod linearne regresije, kod koje dodavanjem nove varijable koja pospješuje model povećavamo R^2, u logističkoj regresiji dodavanjem nove varijable koja pospješuje model očekujemo da se „DEVIANCE“ smanji.

$$\begin{aligned} G &= D(\text{model bez varijabli}) - D(\text{model sa k varijabli}) \\ &= -2LL(0) - (-2LL(k)) \\ &= -2ln\Big[\frac{L(0)}{L(k)}\Big] \approx \chi^2(k) \end{aligned}$$

1.5 Problem konvergencije

Kao što je spomenuto, metoda maksimalne vjerodostojnosti za procjenu logističkog modela je iterativni proces koji podrazumijeva uzastopne aproksimacije. Kada je promjena koeficijenata dvije uzastopne iteracije mala, računanje prestaje i algoritam daje informaciju da je konvergencija zadovoljena. Najčešće, ovaj uvjet je zadovoljen. No ponekad iterativni proces staje i konvergencija nije zadovoljena. U većini slučajeva gdje konvergencija nije zadovoljena, nije problem u pretpostavljenom limitu broju iteracija, već procjena metodom maksimalne vjerodostojnosti ne postoji. Jedan od češćih razloga ne zadovoljavanja uvjeta konvergencije jest da postoji neka linearna kombinacija nezavisnih varijabli koja savršeno predviđa zadanu varijablu odgovora.

1.6 Šansa i omjer šansi

Šansa i omjer šansi su pojmovi usko vezani uz interpretaciju parametara β. Bazirajmo se na univarijatnu logističku regresiju. Označimo sa

$$g(x) := logit(p(x)) = ln\Big[\frac{p(x)}{1-p(x)}\Big] = \beta_0 + \beta_1 x \tag{1.10}$$

$$g(x+1) = \beta_0 + \beta_1(x+1) \tag{1.11}$$

$$g(x+1) - g(x) = \beta_1 \tag{1.12}$$

$$logit(p(x+1)) - logit(p(x)) = \beta_1 \tag{1.13}$$

$$log(odds(p(x+1))) - log(odds(p(x))) = \beta_1 \tag{1.14}$$

$$log\left(\frac{odds(p(x+1))}{odds(p(x))}\right) = \beta_1 \tag{1.15}$$

Općenito, neka je $p(x) \in [0,1]$ vjerojatnost da se događaj x dogodio. Tada je šansa (engl. „ODD“) jednaka $\frac{p(x)}{1-p(x)}$, to jest to je omjer vjerojatnosti da se događaj dogodio i vjerojatnosti da se događaj nije dogodio.

Također, vrijedi:

$$x = 1 \qquad odds(p(1)) = \frac{p(1)}{1-p(1)}$$

$$x = 0 \qquad odds(p(0)) = \frac{p(0)}{1-p(0)}$$

Sada slijedi:

$$\begin{aligned} g(1) - g(0) &= ln\left[\frac{odds(p(1))}{odds(p(0))}\right] \\ &= ln\left[\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}\right] \\ &= ln(OR(1)) \\ &= \beta_1. \end{aligned} \tag{1.16}$$

Iz toga dobivamo:

$$OR(1) = e^{\beta_1} \tag{1.17}$$

„OR“ označava omjer šansi (engl. „ODDS RATIO“).

Omjer šansi za kontinuirane nezavisne varijable

Neka je x kontinuirana nezavisna varijabla. Tada je $g(x+1) - g(x) = \beta_1$.
β_1 pokazuje promjenu u $log(odds)$ za pomak nezavisne varijable za 1. Ukoliko se želi utvrditi promjena u $log(odds)$ za pomak nezavisne varijable za $c, c > 0$, vrijedi sljedeće:

$$\begin{aligned} g(x+c) - g(x) &= c\beta_1, \\ OR(c) &= e^{c\beta_1}. \end{aligned}$$

Omjer šansi za dihotomne nezavisne varijable

U slučaju dihotomne nezavisne varijable omjer šansi računa se na sljedeći način: $OR = \frac{a \cdot d}{b \cdot c}$. Pokažimo to na sljedećem primjeru:

Tablica 1.1: Frekvencijska tablica za endemsku nefropatiju po spolu

Table of EN by SPOL			
FREKVENCIJA POSTOTAK	SPOL		
ENDEMSKA NEFROPATIJA	Ž	M	UKUPNO
NE	612 51,47	422 35,49	1034 86,96
DA	68 5,72	87 7,32	155 13,04
UKUPNO	680 57,19	509 42,81	1189 100,00

U ovome slučaju vrijedi $OR = \frac{612 \cdot 87}{422 \cdot 68} = \frac{53244}{28696} = 1,86$, što zapravo znači da prijelaz iz kategorije žena u kategoriju muškaraca povećava omjer šansi za obolijevanje od endemske nefropatije 1,86 puta, odnosno za 86%.

1.7 Statistika za računanje prediktivne snage modela

Odgovor na pitanje o prediktivnoj snazi binarnog (dihotomnog) logističkog modela maže nam dati c statistika koja je ekvivalentna površini ispod ROC („Receiver Operating Characteristic“) krivulje. Nju računamo na sljedeći način.

Definiramo da se događaj dogodio sa 1, a da se događaj nije dogodio sa 2.
Za par observacija sa različitim odgovorima kažemo da je podudaran (engl. „concordant“) ako observacija koja ima više rangirani odgovor ima nižu prediktivnu vrijednost da se događaj dogodi od observacije sa naže rangiranim odgovorom.
Za par observacija sa različitim odgovorima kažemo da je nepodudaran (engl. „discordant“) ako observacija koja ima više rangirani odgovor ima višu prediktivnu vrijednost da se događaj dogodi od observacije sa niže rangiranim odgovorom.
Ako par observacija nije niti podudaran niti nepodudaran, kažemo da je izjednačen (engl.„tie“),

to jest, radi se o paru observacija sa jednakim odgovorima.

Neka je sada *nt* broj izjednačenih parova, *nc* broj podudarnih parova, te *nd* broj nepodudarnih parova. Tada vrijedi

$$c = \frac{nc + 0,5 \cdot nt}{nc + nt + nd}. \tag{1.18}$$

Poglavlje 2

Primjer modeliranja logističkom regresijom

2.1 Opis varijabli

U bazi imamo ukupno 1.189 observacija.

Svakom ispitaniku uzeti su podatci o spolu, dobi, duljini života u endemskom selu, socio-oekonomska obilježja (edukacija, dohodak), životne navike (pušenje, alkohol, analgetici), je li po zanimanju farmer, ima li arterijsku hipertenziju, ima li dijabetes mellitus.
Od ostalih varijabli, promatrane su $\alpha 1$ mikroglobulin, albuminurija, hemoglobin, kreatinin u serumu, kreatinin, kreatinin-albumin omjer, $\alpha 1$ mikroglobulin-kreatinin omjer, $\alpha 1$ anti-tripsin, procijenjena stopa glomerularne filtracije, specifična gustoća urina, duljine lijevog i desnog bubrega te glukoza.
Također, bilježeno je ima li ispitanik pozitivnu obiteljsku anamnezu na endemsku nefropatiju, te boluje li ona sama od endemske nefropatije.

U ovom odlomku dati su kratki opisi varijabli korištenih u daljnjem modeliranju:

1. DULJINA ŽIVOTA U ENDEMSKOM SELU - Izražena je u godinama (cjelobrojno, iako ćemo je promatrati kao kontinuiranu varijablu), i značava koliko je godina života ispitanik proveo u endemskom selu. U daljnjem tekstu označavati ćemo je sa „DULJINA ENS“.

2. $\alpha 1$ MIKROGLOBULIN - Ponekad ga možemo pronaći pod nazivom Protein HC, jest mikroglobulin, to jest mali globularni protein. Možemo ga pronaći u svim kralješcima, uključujući i ljudskim, i distribuira se u krvnoj plazmi, i izvan vaskularnim

tkivima svih organa. Sintetizira se u većini stanica, ali najviše u jetri. U daljnjem tekstu označavati ćemo ga „α1M“, i izražen je u miligramima.

3. ALBUMINURIJA - Opisuje pojavu albumina u urinu. U daljnjem tekstu označavati ćemo je „ALB“, i izražena je u gramima.

4. HEMOGLOBIN - Krvni pigment, metaloprotein koji u strukturi sadrži željezo te služi za prijenos kisika. Nalazi se u crvenim krvnim stanicama. Hemoglobin prenosi kisik iz plua prema ostatku tijela, na primjer prema mišićima, gdje otpušta kisik. U daljnjem tekstu označavati ćemo ga sa „HB“, i izražen je u gramima po litri.

5. KREATININ - Organski spoj koji u ljudskom tijelu nastaje spontano razgradnjom kreatin-fosfata u mišićima, te se iz tijela izlučuje bubrezima. Također, odnosi se na razinu kreatinina u urinu. U daljnjem tekstu označavati ćemo ga sa „CR“, i izražen je u mikromolima pa litri.

6. KREATININ U SERUMU - Razinu kreatinina u krvi. U daljnjem tekstu označavati ćemo ga sa „CRS“, i izražen je u mikromolima po litri.

7. KREATININ-ALBUMIN OMJER - Rezultat se koristi za detektiranje bolesti bubrega u ranoj fazi razvoja. U daljnjem tesktu označavati ćemo ga se „CR/ALB“.

8. α1 MIKROGLOBULIN-KREATININ OMJER - Rezultat se koristi za detektiranje tubularne disfunkcije. U daljnjem tesktu označavati ćemo ga sa „α1M/CR“.

9. α1 ANTITRIPSIN- Koristi se za detektiranje α1 antitripsin manjkavosti, i mjeri razinu proteina AAT u krvi. U daljnjem tekstu označavati ćemo ga se „α1AT“, i izražen je u nanomolima po litri.

10. PROCIJENJENA STOPA GLOMERULARNE FILTRACIJE - Pomaže pri detekciji bolesti bubrega u ranoj fazi više nego test kreatinina sam. Izražena je u mililitrima po minuti i u daljnjem tekstu označavati ćemo je sa „GFR“.

11. SPECIFIČNA GUSTOĆA URINA - Testira se za provjeru koliko je dobra funkcija bubrega. U daljnjem teksu označavati ćemo ju sa „SDU“. To je omjer gustoće urina naspram gustoće vode.

12. DULJINA LIJEVOG(DESNOG) BUBREGA - Duljina bubrega izražena u milimetrima. U daljnjem tekstu označavati ćemo je sa „DLB“, odnosno sa „DDB“.

13. GLUKOZA - Najrasprostranjeniji monosaharid u prirodi. Povišena razina gukoze u krvi javlja se kod dijabetičara, a snižena kod hipoglikemije različitih uzroka. U daljnjem tekstu označavati ćemo ju sa „GL“, a izražena je u mikromolima po litri.

14. ENDEMSKA NEFROPATIJA - Prethodno je već opisana. U daljnjem tekstu označavati ćemo ju sa „EN“.

15. ARTERIJSKA HIPERTENZIJA - Kronična bolest u kojoj je povišen krvni tlak. U daljnjem tekstu označavati ćemo ju sa „AH“.

16. DIABETES MELLITUS - Poznatiji pod nazivom „dijabetes tip 1“, u čijoj se osnovi nalazi autoimuni proces. U daljnjem tekstu označavati ćemo ga sa „DM“.

17. POZITIVNA OBITELJSKA ANAMNEZA - Bilježi je li netko od obitelji ispitanika bolovao ili boluje od endemske nefropatije. U daljnjem teksu označavati ćemo ju sa „ANAMNEZA“.

U obrađenom primjeru statistički značajnom smatrana je vrijednost čija je greška tipa I(α) manja od 0.05 ($p<0.05$). Statistička obrada podataka izrađena je u statističkom softveru SAS Studio.

2.2 Deskriptivna statistika

U tablicama 2.1 i 2.2 dana je deskriptivna statistika za kategorijske varijable, dok je u tablici 2.3 dana deskriptivna statistika za kontinuirane varijable.

Tablica 2.1: Deskriptivna statistika za kategorijske varijable (1.dio)

VARIJABLA	KATEGORIJA	FREKVENCIJA	POSTOTAK
SPOL	Ž	680	57,19
	M	509	42,81
	UKUPNO	1189	100,00
EDUKACIJA	NSS	18	1,89
	SSS	601	63,20
	VŠS	290	30,49
	VSS	42	4,42
	UKUPNO	951	100,00
PRIHODI	NISKI	399	47,78
	SREDNJI	408	48,86
	VISOKI	28	3,35
	UKUPNO	835	100,00
FARMER	NE	899	93,26
	DA	65	6,74
	UKUPNO	964	100,00
PUŠENJE	NE	739	76,74
	DA	224	23,26
	UKUPNO	963	100,00
KONZUMACIJA ALKOHOLA	NIKADA	612	63,29
	PONEKAD	168	17,37
	STALNO	187	19,34
	UKUPNO	967	100,00
UZIMANJE ANALGETIKA	NIKADA	352	36,51
	PONEKAD	436	45,23
	STALNO	176	18,26
	UKUPNO	964	100,00
ARTERIJSKA HIPERTENZIJA	NE	606	52,74
	DA	543	47,26
	UKUPNO	1149	100,00
DIABETES MELLITUS	NE	1134	95,45
	DA	54	4,55
	UKUPNO	1188	100,00

Tablica 2.2: Deskriptivna statistika za kategorijske varijable (2.dio)

VARIJABLA	KATEGORIJA	FREKVENCIJA	POSTOTAK
ANAMNEZA	NE	547	46,01
	DA	642	53,99
	UKUPNO	1189	100,00
ENDEMSKA	NE	1034	86,96
	DA	155	13,04
NEFROPATIJA	UKUPNO	1189	100,00

Tablica 2.3: Deskriptivna statistika za kontinuirane varijable

VARIJABLA	N	ASR	STD	MIN	MEDIAN	MAX
DOB	1189	53,5	17,43	19	53	93
DULJINA ENS	1189	35,86	25,18	0	38	91
VISINA (cm)	1129	168,40	9,90	143	168	197
TEŽINA (kg)	1139	78,46	16,12	40	77	140
BMI	1128	27,67	5,18	16,65	27,12	50,20
ST (mmHg)	1141	142,99	24,87	80	140	235
DT (mmHg)	1140	83,95	12,86	50	83	135
α1M (mg/L)	1177	13,38	21,88	5,18	5,97	398
ALB (mg/L)	1166	21,97	69,86	2,09	6,78	1441
HB	1175	135,9	16,11	71	137	187
CRS (μmol/L)	1183	98,66	79,42	50	83	1005
CR (g/L)	895	1,1	0,59	0,23	0,98	3,38
ALB/CR (mg/g)	894	29,51	95,38	0,80	6,14	1197,55
α1M/CR (mg/L)	895	18,59	49,59	1,53	7,2	1005,43
α1AT (mg)	1166	1,51	1,55	0,01	1,16	22,58
GFR	1183	76,79	22,77	3,44	79,18	126,76
SDU	887	1,01	0,01	1	1,02	1,03
DLB (mm)	768	107,64	11,5	62	108	138
DDB (mm)	766	107,61	11,7	60	109	137
GLUKOZA	878	5,5	1,62	3,3	5,15	20,7

U tablici 2.3 sa „N“ je označen broj opažanja, sa „ASR“ označena je aritmetička sredina, a sa „STD“ standardna devijacija. Također, „MIN“ i „MAX“ redom označavaju minimum i maksimum.

2.3 Univarijatna logistička regresija

Promatramo modele sa jednom nezavisnom varijablom.

Tablica 2.4: Rezultati ML procjene parametara za univarijatne modele sa kategorijskom nezavisnom varijablom

VARIJABLA	ANALIZA ML PROCJENE				OR PROCJENA			c	N
	PROCJENA	STD. GREŠKA	WALD χ^2	p	PROCJENA	95% WALD P.I.			
SPOL_M	0,6181	0,1738	12,6501	0,0004	1,855	1,320	2,608	0,577	1189
FARMER_DA	-0,3543	0,4110	0,7432	0,3887	0,702	0,314	1,570	0,510	964
EDUKACIJA_SSS	0,6274	0,7572	0,6865	0,3796	1,873	0,425	8,260		
EDUKACIJA_VŠS	-0,5782	0,7867	0,6865	0,4073	0,561	0,120	2,621	0,619	951
EDUKACIJA_VSS	-0,9163	1,0428	0,7720	0,3796	0,400	0,052	3,088		
PRIHODI_SREDNJI	-0,8974	0,2069	18,8194	<,0001	0,408	0,272	0,611	0,610	835
PRIHODI_VISOKI	-0,4700	0,5538	0,7201	0,3961	0,625	0,211	1,851		
ALKOHOL_PONEKAD	0,4828	0,2340	4,2579	0,0391	1,621	1,025	2,564	0,556	967
ALKOHOL_STALNO	0,4282	0,2280	3,5263	0,0604	1,534	0,981	2,399		
ANALGETICI_PONEKAD	-0,4389	0,2057	4,5552	0,0328	0,645	0,431	0,965	0,553	964
ANALGETICI_STALNO	-0,0827	0,2501	0,1092	0,7410	0,921	0,564	1,503		
PUŠENJE_DA	-0,0162	0,2243	0,5221	0,4699	0,850	0,548	1,320	0,514	963
AH_DA	0,6632	0,1779	13,8993	0,0002	1,941	1,370	2,751	0,582	1149
DM_DA	0,4384	0,3614	1,4709	0,2252	1,550	0,763	3,148	0,511	1188
ANAMNEZA_DA	-0,2586	0,1724	2,2483	0,1338	0,772	0,551	1,083	0,532	1189

Iz tablice 2.4 možemo iščitati da u univarijatnim modelima logističke regresije spol, prihodi, alkohol, analgetici i arterijska hipertenzija statistički značajno utječu na vjerojatnost da osoba boluje od endemske nefropatije.

Tablica 2.5: Rezultati ML procjene parametara za univarijatne modele sa kontinuiranom nezavisnom varijablom

VARIJABLA	ANALIZA ML PROCJENE				OR PROCJENA			c	N
	PROCJENA	STD. GREŠKA	WALD χ^2	p	PROCJENA	95% WALD P.I.			
DOB	0,0630	0,00653	93,0342	<,0001	1,065	1,051	1,079	0,764	1189
DULJINA ENS	0,0579	0,00508	129,7111	<,0001	1,060	1,049	1,070	0,819	1189
VISINA (cm)	-0,0170	0,00919	3,3141	0,0646	0,983	0,966	1,001	0,548	1129
TEŽINA (kg)	-0,0158	0,00583	7,3652	0,0066	0,984	0,973	0,996	0,561	1139
BMI	-0,0320	0,0180	3,1614	0,0754	0,968	0,935	1,003	0,544	1128
ST	0,0155	0,00334	21,4918	<,0001	1,016	1,009	1,022	0,625	1141
DT	-0,00147	0,00680	0,0470	0,8284	0,999	0,985	1,012	0,515	1140
α1M (mg/L)	0,1099	0,00892	151,9864	<,0001	1,116	1,098	1,136	0,949	1177
ALB (mg/L)	0,00578	0,00142	16,4656	<,0001	1,006	1,003	1,009	0,726	1166
HB	-0,0337	0,00513	42,9599	<,0001	0,967	0,957	0,977	0,601	1175
CRS (μmol/L)	0,0200	0,00257	60,6992	<,0001	1,020	1,015	1,025	0,833	1183
CR (g/L)	-2,0817	0,3417	37,1074	<,0001	0,125	0,064	0,244	0,741	895
ALB/CR (mg/g)	0,00610	0,00111	30,1815	<,0001	1,006	1,004	1,008	0,838	894
α1M/CR (mg/L)	0,0479	0,00545	77,3217	<,0001	1,049	1,038	1,060	0,955	895
α1AT (mg)	0,4398	0,0579	57,7168	<,0001	1,552	1,386	1,739	0,672	1166
GFR	-0,0576	0,00451	162,9883	<,0001	0,944	0,936	0,952	0,803	1183
SDU	-162,6	22,8025	50,8266	<,0001	<,0001	<,0001	<,0001	0,769	887
DLB (mm)	-0,770	0,0109	50,1085	<,0001	0,926	0,906	0,946	0,735	768
DDB (mm)	-0,0781	0,0104	56,0383	<,0001	0,925	0,906	0,944	0,742	766
GLUKOZA	0,1213	0,0564	4,9284	0,0264	1,129	1,014	1,257	0,611	878

Iz tablice 2.5 možemo iščitati da u univarijatnim modelima logističke regresije, sve varijable osim visine, indeksa tjelesne mase i dijastoličkog tlaka, statistički značajno utječu na vjerojatnost da osoba boluje od endemske nefropatije.

2.4 Multivarijatna logistička regresija

U ovome poglavlju bavimo se multivarijatnim modelima te modeliramo vjerojatnost da osoba boluje od endemske nefropatije.

Napravljena su četiri različita modela i to su modeli dobiveni „FORWARD“, „BACKWARD“, „STEPWISE“ metodom selekcije, te zadani model od struke.

Pri odabiru „FORWARD“ metode selekcije, SAS prvo računa vrijednost χ^2 statistike za svaki efekt(nezavisnu varijablu) pojedinačno, te ispituje najveću od tih statistika. Ako je ona značajna na odabranoj razini značajnosti, pripadni efekt dodaje se u model. Nakon toga proces se ponavlja za preostale efekte, dokle god među njima postoji onaj čija je χ^2 statistika značajna, ili dokle god svi efekti nisu dodani u model. U ovoj metodi bitno je napomenuti da kada efekt jednom nadodamo u model, on ostaje u modelu.

Pri odabiru „BACKWARD“ metode selekcije, SAS prvo ubaci sve efekte u model pa računa vrijednost χ^2 statistike za svaki efekt(nezavisnu varijablu)u modelu, te ispituje najmanju od tih statistika. Ako ona nije značajna na odabranoj razini značajnosti, pripadni efekt izbacuje se iz model. Nakon toga proces se ponavlja za preostale efekte, dokle god među njima postoji onaj čija χ^2 statistika nije značajna, ili dokle god svi efekti nisu izbačeni iz modela. U ovoj metodi bitno je napomenuti da kada efekt jednom izbacimo iz modela, on ostaje vani.

„STEPWISE“ metoda selekcije slična je „FORWARD“ metodi, osim što efekt koji je ušao u model ne ostaje nužno u modelu. Efekti ulaze i izlaze iz modela na takav način da svaki „FORWARD“ korak selekcije može biti praćen sa jednim ili više „BACKWARD“ korakom eliminacije. „STEPWISE“ selekcija završava ako nema više elemenata koji se mogu nadodati u model, ili ako je trenutni model jednak modelu u neko prethodnom koraku.

U kreiranju naših modela „FORWARD“„BACKWARD“ te „STEPWISE“ metodama selekcije, sve dostupne varijable stavljene su u model.

Kreiranje zadanog modela je najjednostavnija metoda koja daje informaciju o točno onom modelu koji je zadan, i ne ubacuje niti izbacuje varijable iz modela. Tu metodu koristimo kada imamo neke pretpostavke što bi moglo utjecati na vjerojatnost obolijevanja, i tu nam informaciju daju struka.

2.4.1. „FORWARD“ metoda selekcije

Tablica 2.6: Dijagnostika multivarijatnog logističkog modela dobivenog „FORWARD“ metodom selekcije

Broj korištenih observacija			
425			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika za adekvatnost modela			
Kriterij	Samo sjecište	Sjecište i varijable	
-2 LOG L	251,33	36,72	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr>χ^2
Likelihood Ratio	214,61	8	<0,001
χ^2-test reziduala			
χ^2	DF	Pr >χ^2	
30,36	27	0,30	

Iz tablice 2.6 vidimo da je u modelu korišteno 425 observacija i da je kriterij konvergencije zadovoljen. Također, dobiveni model je statistički značajan ($\chi^2 = 214,61$, $st.sl. = 8$, $p < 0,001$).

Tablica 2.7: Rezultati ML procjene parametara za multivarijatni logistički model dobiven „FORWARD“ metodom selekcije

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-58,4288	16,9400	11,8967	0,0006
DULJINA ENS	1	0,4028	0,1181	11,6272	0,0006
A1M	1	0,2747	0,0766	12,8616	0,0003
ALB	1	-0,0128	0,00529	5,8197	0,0158
A1AT	1	1,4873	0,5687	6,8409	0,0089
GLUKOZA	1	0,5251	0,2290	5,2578	0,0218
PRIHODI_S	1	-4,3274	1,5201	8,1044	0,0044
AH_DA	1	3,4836	1,4182	6,0337	0,0140
ANAMNEZA_DA	1	8,0923	2,7700	8,5344	0,0035

Iz tablice 2.7 slijedi da je jednadžba modela:

$$\begin{aligned} logit(p) = &-58,4288 + 0,4028 \cdot DULJINA_ENS + 0,2747 \cdot \alpha 1M - 0,0128 \cdot ALB + \\ &+ 1,4873 \cdot \alpha 1AT + 0,5251 \cdot GLUKOZA - 4,3274 \cdot PRIHODI_SREDNJI + \\ &+ 3,4836 \cdot AH_DA + 8,0923 \cdot ANAMNEZA_DA. \end{aligned} \tag{2.1}$$

Tablica 2.8: Rezultati procjene omjera šansi efekata u multivarijatnom logističkom modelu dobivenom „FORWARD“ metodom selekcije

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
DULJINA ENS	1,496	1,187	1,886
α1M	1,316	1,133	1,529
ALB	0,987	0,977	0,998
α1AT	4,425	1,452	13,489
GLUKOZA	1,691	1,079	2,648
PRIHODI_SREDNJI	0,013	<0,001	0,260
AH_DA	32,578	2,022	524,928
ANAMNEZA_DA	>999,999	14,341	>999.999

Tablicu 2.8 tumačimo na sljedeći način:

1. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,496 puta, to jest za $49,6\%$.

2. Jedinično povećanje α1 mikroglobulina povećava omjer šansi za razvijanje bolesti 1,316 puta, to jest za $31,6\%$.

3. Jedinično povećanje albumin smanjuje omjer šansi za razvijanje bolesti za $1,3\%$.

4. Jedinično povećanje α1 antitripsina povećava omjer šansi za razvijanje bolesti 4,425 puta, to jest za $342,5\%$.

5. Jedinično povećanje glukoze povećava omjer šansi za razvijanje bolesti 1,691 puta, to jest za $69,1\%$.

6. Prijelaz iz kategorije s niskim prihodima u kategoriju sa srednjim prihodima smanjuje omjer šansi za $7474,7\%$.

7. Prijelaz iz kategorije bez arterijske hipertenzije u kategoriju sa arterijskom hipertanzijom povećava omjer šanse za razvijanje bolesti 32,578 puta, to jest za $3157,8\%$.

8. Objasnimo rezultate za ANAMNEZU sljedećom tablicom.

Tablica 2.9: Frekvencija endemske nefropatije po anamnezi za uzorak korišten u modelu izrađenom pomoću „FORWARD“ metode selekcije.

Tablica EN po varijabli ANAMNEZA			
FREKVENCIJA	ANAMNEZA		
ENDEMSKA NEFROPATIJA	NE	DA	UKUPNO
NE	207	190	397
DA	0	28	28
UKUPNO	207	218	425

U tablici 2.9 vidimo da u uzorku korištenom u ovome modelu niti jedna osoba s negativnom obiteljskom anamnezom nema bolest, što zapravo znači da je $OR = \frac{207 \cdot 28}{190 \cdot 0} = \frac{5.796}{0}$, i to je razlog rezultata da prijelaz iz kategorije s negativnom obiteljskom anamnezom u kategoriju sa pozitivnom obiteljskom anamnezom povećava omjer šansi za razvijanje bolesti više od 999,999 puta.

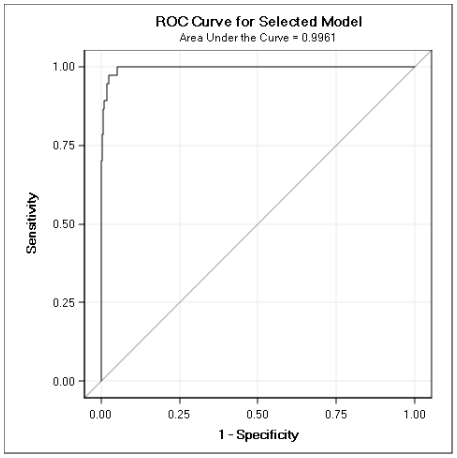
Tablica 2.10: Redoslijed ulaska varijabli u multivarijatni logistički model pomoću „FORWARD“ metode selekcije

Pregled „FORWARD“selekcije					
Korak	Ulazna varijabla	DF	Broj ulaska	χ^2	p
1	α1M	1	1	184,3957	<0,0001
2	ALB	1	2	49,2513	<0,0001
3	DULJINA ENS	1	3	31,7075	<0,0001
4	ANAMNEZA_DA	1	4	7,4236	0,0064
5	PRIHODI_SREDNJI	1	5	8,7923	0,0030
6	α1AT	1	6	7,0512	0,0079
7	AH_DA	1	7	5,9539	0,0147
8	GLUKOZA	1	8	7,1930	0,0073

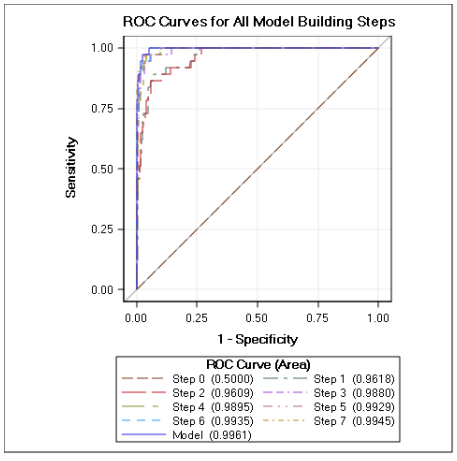
Tablica 2.11: Prediktivna snaga multivarijatnog logističkog modela dobivenog „FORWARD„ metodom selekcije

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
c	0,996

Prediktivna snaga modela jest 0,996, što znači da je $99,6\%$ pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.1 se nalazi ROC krivulja finalnog modela, dok je na slici 2.2 dana ROC krivulja svakog koraka „FORWARD“ algoritma.



Slika 2.1: ROC krivulja modela dobivenog „FORWARD“ algoritmom



Slika 2.2: ROC krivulja za svaki korak „FORWARD“ algoritma

2.4.2. „BACKWARD“ metoda selekcije

Tablica 2.12: Dijagnostika multivarijatnog logističkog modela dobivenog „BACKWARD“ metodom selekcije

Broj korištenih observacija			
425			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika adekvatnosti modela			
Kriterij	Samo sjecište	Sjecište i varijable	
-2 LOG L	251,33	43,79	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr>χ^2
Likelihood Ratio	207,54	8	<0,001
χ^2-test reziduala			
χ^2	DF	Pr >χ^2	
50,04	27	0,30	

Iz tablice 2.12 vidimo da je u modelu korišteno 425 observacija i da je kriterij konvergencije zadovoljen. Također, dobiveni model je statistički značajan ($\chi^2 = 207,54$, $st.sl. = 8$, $p < 0,001$).

Tablica 2.13: Rezultati ML procjene parametara za multivarijatni logistički model dobiven „BACKWARD“ metodom selekcije

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-52,9852	17,3025	9,3776	0,0022
DULJINA ENS	1	0,3476	0,1024	11,5321	0,0007
VISINA	1	0,1063	0,0523	4,1392	0,0419
α1M	1	0,2021	0,0528	14,6547	0,0001
α1AT	1	1,6031	0,5156	9,6663	0,0019
DDB	1	-0,1175	0,0494	5,6615	0,0173
PRIHODI_SREDNJI	1	-4,3082	1,4469	8,8654	0,0029
AH_DA	1	2,1755	1,0399	4,3771	0,0364
ANAMNEZA_DA	1	7,5814	2,4373	9,6752	0,0019

Iz tablice 2.13 slijedi da je jednadžba modela:

$$\begin{aligned} logit(p) = &- 52,9852 + 0,3476 \cdot DULJINA_ENS + 0,1063 \cdot VISINA + 0,2021 \cdot \alpha 1M + \\ &+ 1,6031 \alpha 1AT - 0,1175 \cdot DDB - 4,3082 \cdot PRIHODI_SREDNJI \\ &+ 2,1755 \cdot AH_DA + 7,5814 \cdot ANAMNEZA_DA. \end{aligned} \tag{2.2}$$

Tablica 2.14: Rezultati procjene omjera šansi efekata u multivarijatnom logističkom modelu dobivenom „BACKWARD“ metodom selekcije

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
DULJINA ENS	1,416	1,158	1,730
VISINA	1,112	1,004	1,232
α1M	1,224	1,104	1,357
α1AT	4,968	1,808	13,648
DDB	0,889	0,807	0,979
PRIHODI_SREDNJI	0,013	<0,001	0,229
AH_DA	8,807	1,147	67,601
ANAMNEZA_DA	>999,999	16.515	>999,999

Tablicu 2.14 tumačimo na sljedeći način:

1. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,416 puta, to jest za $41,6\%$.
2. Jedinično povećanje visine povećava omjer šansi za razvijanje bolesti 1,112 puta, to jest za $11,2\%$.
3. Jedinično povećanje $\alpha 1$ mikroglobulina povećava omjer šansi za razvijanje bolesti 1,224 puta, to jest za $22,4\%$.
4. Jedinično povećanje $\alpha 1$ antitripsina povećava omjer šansi za razvijanje bolesti 4,968 puta, to jest za $396,8\%$.
5. Jedinično povećanje duljine desnog bubrega smanjuje omjer šansi za razvijanje bolesti za $12,5\%$.
6. Prijelaz iz kategorije s niskim prihodima u kategoriju sa srednjim prihodima smanjuje omjer šansi za $7330,7\%$.

7. Prijelaz iz kategorije bez arterijske hipertenzije u kategoriju sa arterijskom hipertenzijom povećava omjer šanse za razvijanje bolesti 8,807 puta, to jest za 780, 7%.

8. Objašnjenje rezultata za ANAMNEZU dano je sljedećom tablicom.

Tablica 2.15: Frekvencija endemske nefropatije po anamnezi za uzorak korišten u modelu izrađenom pomoću „BACKWARD“ metode selekcije.

Tablica EN po varijabli ANAMNEZA			
FREKVENCIJA	ANAMNEZA		
ENDEMSKA NEFROPATIJA	NE	DA	UKUPNO
NE	207	190	397
DA	0	28	28
UKUPNO	207	218	425

U tablici 2.15 vidimo da u uzorku korištenom u ovome modelu niti jedna osoba s negativnom obiteljskom anamnezom nema bolest, što zapravo znači da je $OR = \frac{207 \cdot 28}{190 \cdot 0} = \frac{5.796}{0}$, i to je razlog rezultata da prijelaz iz kategorije s negativnom obiteljskom anamnezom u kategoriju sa pozitivnom obiteljskom anamnezom povećava omjer šansi za razvijanje bolesti više od 999,999 puta.

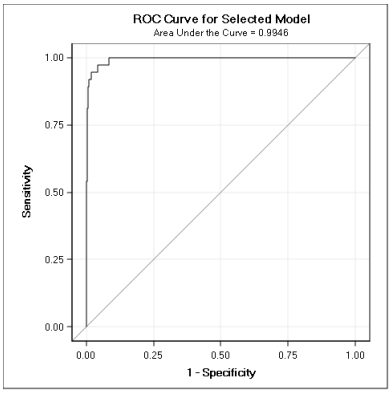
Tablica 2.16: Redoslijed izlaska varijabli iz multivarijatnog logističkog modela pomoću „BACKWARD“ metode selekcije

Pregled „BACKWARD“eliminacije					
Korak	Izlazna varijabla	DF	Broj ulaska	χ^2	p
1	EDUKACIJA_VSS	1	34	0,0000	0,9990
2	ALB	1	33	0,0005	0,9828
3	PRIHODI_V	1	32	0,0308	0,8608
4	DM_DA	1	31	0,0174	0,8952
5	α1/CR	1	30	00477	0,8271
6	SPOL_M	1	29	0,1033	0,7479
7	GFR	1	28	0,0327	0,8564
8	GLUKOZA	1	27	0,0099	0,9207
9	SDU	1	26	0,0333	0,8553
10	CRS	1	25	0,0875	07675
11	PUŠENJE_DA	1	24	0,1393	0,7089
12	CR	1	23	0,0970	0,7554
13	ALKOHOL_PONEKAD	1	22	0,0993	0,7527
14	HB	1	21	0,4201	0,5169
15	EDUKACIJA_SSS	1	20	0,2719	0,6021
16	FARMER_DA	1	19	0,4165	0,5187
17	EDUKACIJA_VŠS	1	18	0,3000	0,5839
18	ALKOHOL_STALNO	1	17	0,5568	0,4555
19	DLB	1	16	0,8515	0,3561
20	ANALGETICI_PONEKAD	1	15	1,2404	0,2654
21	DT	1	14	1,7258	0,1889
22	ALB/CR	1	13	2,9380	0,0865
23	DOB	1	12	0,0676	0,7949
24	ST	1	11	1,1212	0,2897
25	ANALGETICI_STALNO	1	10	2,3397	0,1261
26	TEŽINA	1	9	2,3914	0,1220
27	BMI	1	8	0,8561	0,3548

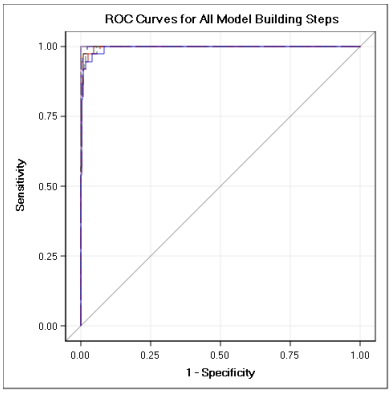
Tablica 2.17: Prediktivna snaga multivarijatnog logističkog modela dobivenog „BACKWARD“ metodom selekcije

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
C	0,995

Iz tablice 2.17 vidimo da je prediktivna snaga modela jest 0,995, što znači da je 99, 5% pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.3 se nalazi ROC krivulja finalnog modela, dok je na slici 2.4 dana ROC krivulja svakog koraka „BACKWARD“ algoritma.



Slika 2.3: ROC krivulja modela dobivenog „BACKWARD“ algoritmom



Slika 2.4: ROC krivulja za svaki korak „BACKWARD“ algoritma

2.4.3. „STEPWISE“ metoda selekcije

Tablica 2.18: Dijagnostika multivarijatnog logističkog modela dobivenog „STEPWISE“ metodom selekcije

Broj korištenih observacija			
425			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika adekvatnosti modela			
Kriterij	Samo sjecište	Sjecište i varijable	
-2 LOG L	251,33	88,12	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr>χ^2
Likelihood Ratio	163,21	3	<0,001
χ^2-test reziduala			
χ^2	DF	Pr >χ^2	
148,58	33	0,005	

Iz tablice 2.18 vidimo da je u modelu korišteno 425 observacija i da je kriterij konvergencije zadovoljen. Također, dobiveni model je statistički značajan ($\chi^2 = 163,21$, $st.sl. = 3$, $p < 0,001$).

Tablica 2.19: Rezultati ML procjene parametara za multivarijatni logistički model dobiven „STEPWISE“ metodom selekcije

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-16,0180	2,9595	29,2947	<,0001
DULJINA ENS	1	0,1343	0,0305	19,4321	<,0001
α1M	1	0,1103	0,0194	32,2145	<,0001
ANAMNEZA_DA	1	7,9604	2,7510	6,8145	0,0090

Iz tablice 2.19 slijedi da je jednadžba modela:

$$logit(p) = -16,0180 + 0,1343 \cdot DULJINA_ENS + 0,1103 \cdot \alpha 1M + 7,9604 \cdot ANAMNEZA_DA. \tag{2.3}$$

Tablica 2.20: Rezultati procjene omjera šansi efekata u multivarijatnom logističkom modelu dobivenom „STEPWISE“ metodom selekcije

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
DULJINA ENS	1,144	1,077	1,214
α1M	1,117	1,075	1,160
ANAMNEZA_DA	>999,999	1,630	>999,999

Tablicu 2.20 tumačimo na sljedeći način:

1. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,144 puta, to jest za $14,4\%$.
2. Jedinično povećanje α1 mikroglobulina povećava omjer šansi za razvijanje bolesti 1,117 puta, to jest za $11,7\%$.
3. Objašnjenje za ANAMNEZU dano je sljedećom tablicom.

Tablica 2.21: Frekvencija endemske nefropatije po anamnezi za uzorak korišten u modelu izrađenom pomoću „STEPWISE“ metode selekcije.

Tablica EN po varijabli ANAMNEZA			
FREKVENCIJA	ANAMNEZA		
ENDEMSKA NEFROPATIJA	NE	DA	UKUPNO
NE	207	190	397
DA	0	28	28
UKUPNO	207	218	425

U tablici 2.21 vidimo da u uzorku korištenom u ovome modelu niti jedna osoba s negativnom obiteljskom anamnezom nema bolest, što zapravo znači da je $OR = \frac{207 \cdot 28}{190 \cdot 0} = \frac{5.796}{0}$, i to je razlog rezultata da prijelaz iz kategorije s negativnom obiteljskom anamnezom u kategoriju sa pozitivnom obiteljskom anamnezom povećava omjer šansi za razvijanje bolesti više od 999,999 puta.

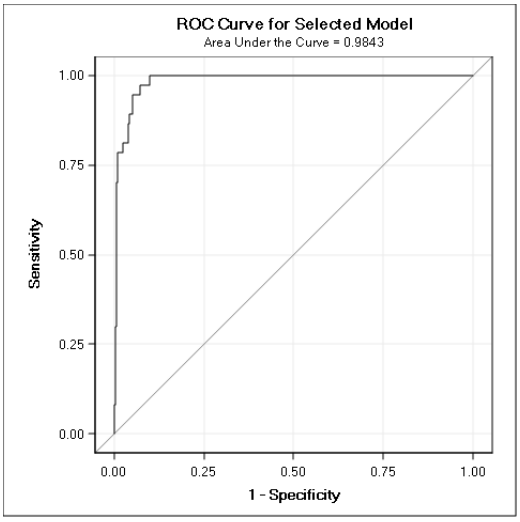
Tablica 2.22: Redoslijed ulaza i izlaza varijabli iz multivarijatnog logističkog modela pomoću „STEPWISE“ metode selekcije

Pregled koraka „STEPWISE“ selekcije							
Korak	Varijabla		DF	Broj ulaska	χ^2	Wald χ^2	p
	Ulazna	Izlazna					
1	α1M		1	1	184,3957		<,0001
2	ALB		1	2	49,2513		<,0001
3	DULJINA ENS		1	3	31,7075		<,0001
4	ANAMNEZA_DA		1	4	7,4236		0,0064
5		ALB	1	3		3,3389	0,0677
6	ALB		1	4	85,1972		<,0001
7		ALB	1	3		3,3389	

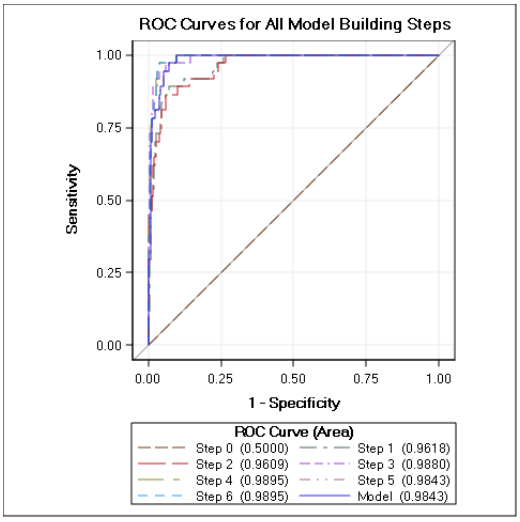
Tablica 2.23: Prediktivna snaga multivarijatnog logističkog modela dobivenog „STEPWISE“ metodom selekcije

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
C	0,984

Prediktivna snaga modela jest 0,984, što znači da je $98,4\%$ pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.5 se nalazi ROC krivulja finalnog modela, dok je na slici 2.6 dana ROC krivulja svakog koraka „STEPWISE“ algoritma.



Slika 2.5: ROC krivulja modela dobivenog „STEPWISE“ algoritmom



Slika 2.6: ROC krivulja za svaki korak „STEPWISE“ algoritma

2.4.4. Zadani modeli

Na temelju dosadašnjih istraživanja postoje određene pretpostavke što bi moglo utjecati na pojavnost ove bolesti. Zbog toge je predloženo da model za nezavisne varijable uzima spol, duljinu života u endemskom selu, podatak o zanimanju (samo je li osoba farmer ili ne), indeks tjelesne mase te sistolički tlak.

Tablica 2.24: Dijagnostika zadanog multivarijatnog logističkog modela

Broj korištenih observacija			
917			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika za adekvatnost modela			
Kriterij	Samo sjecište	Sjecište i varijable	
-2 LOG L	766,35	616,31	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr >χ^2
Likelihood Ratio	150,04	5	<0,001

Iz tablice 2.24 vidimo da je broj korištenih opažanja u ovome modelu 917 i kriterij konvergencije je zadovoljen. Dobiveni model jest statistički značajan ($\chi^2 = 150,04$, $st.sl. = 5$, $p < 0,001$).

Tablica 2.25: Rezultati ML procjene parametara za zadani multivarijatni logistički model

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-5,0684	0,8971	31,9179	<,0001
SPOL_M	1	0,6979	0,2086	11,1866	0,0008
DULJINA ENS	1	0,0480	0,00551	75,8540	<,0001
FARMER_DA	1	-0,1661	0,4452	0,1391	0,7091
BMI	1	-0,0379	0,0211	3,2278	0,0724
ST	1	0,00725	0,00420	2,9810	0,0842

Iz tablice 2.25 slijedi da je jednadžba dobivenog modela:

$$\begin{aligned} logit(p) = &- 5,0684 + 0,6979 \cdot SPOL_M + 0,0480 \cdot DULJINA_ENS \\ &- 0,1661 \cdot FARMER_DA - 0,0379 \cdot BMI + 0,00725 \cdot ST. \end{aligned} \tag{2.4}$$

Tablica 2.26: Rezultati procjene omjera šansi efekata u zadanom multivarijatnom logističkom modelu

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
SPOL_M	2,009	1,335	3,023
DULJINA ENS	1,049	1,038	1,061
FARMER	0,847	0,354	2,027
BMI	0,963	0,924	1,003
ST	1,007	0,999	1,016

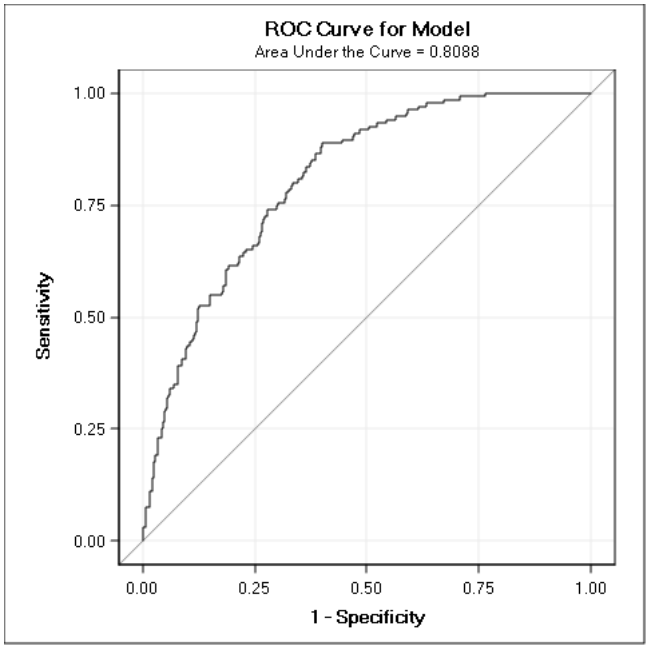
Tablicu 2.26 tumačimo na sljedeći način:

1. Prijelaz iz kategorije ženskog spola u kategoriju muškog spola povećava omjer šansi 2,009 puta, odnosno za $100,9\%$.
2. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,049 puta, to jest za $4,9\%$.
3. Prijelaz iz kategorije osoba koje po zanimanju nisu farmeri u kategoriju osoba koje su farmeri, smanjuje omjer šanse za razvijanje bolesti za $18,1\%$, i to nije statistički značajno.
4. Jedinično povećanje indeksa tjelesne mase smanjuje omjer šansi za razvijanje bolesti za $14,3\%$, i to nije statistički značajno.
5. Jedinično povećanje sistolički tlak povećava omjer šansi za razvijanje bolesti 1,007 puta, to jest za $0,7\%$, i to nije statistički značajno.

Tablica 2.27: Prediktivna snaga zadanog multivarijatnog logističkog modela

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
C	0,809

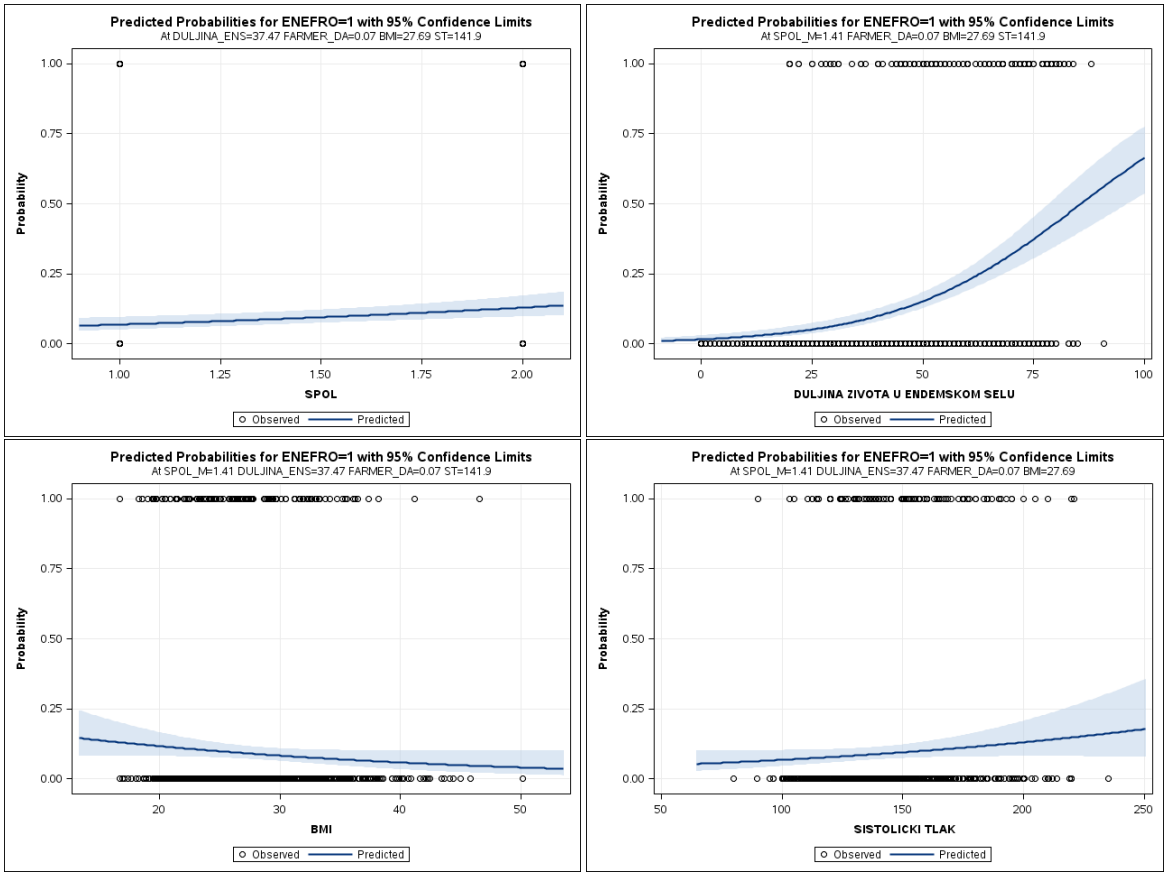
Iz tablice 2.27 iščitavamo da je prediktivna snaga modela 0,809, što znači da je $80,9\%$ pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.7 se nalazi ROC krivulja zadanog modela.



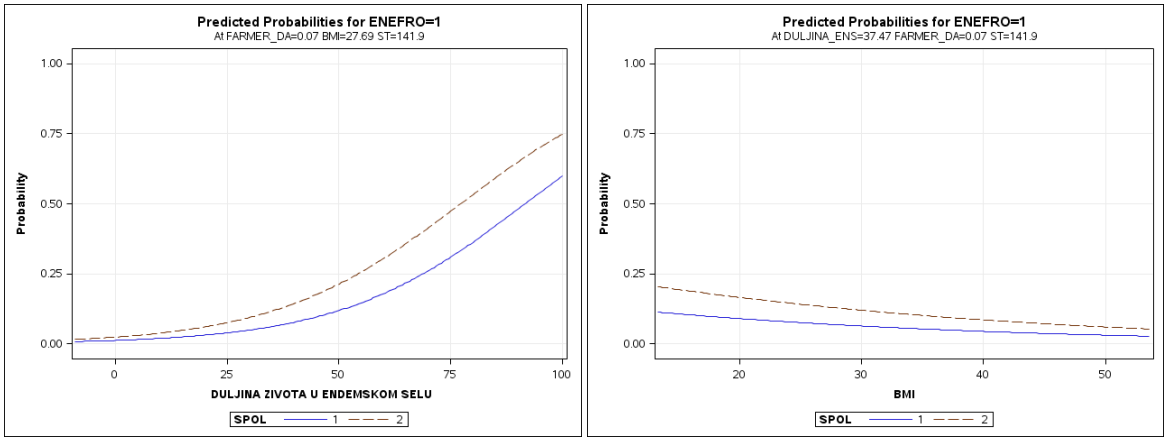
Slika 2.7: ROC krivulja prvog zadanog modela

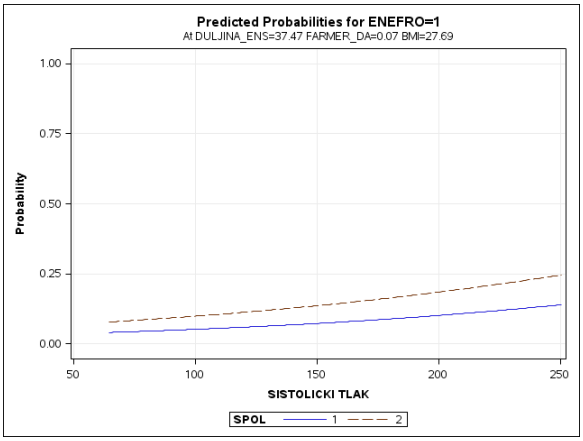
Pogledajmo još krivulje vjerojatnosti za sve varijable u zadanom modelu, osim varijable FARMER. Za svaku varijablu crtamo vjerojatnost da osoba ima bolest kada su vrijednosti preostalih nezavisnih varijabli u modelu njihova aritmetička sredina. Na slici 2.8 dana je vjerojatnost sa pripadnim pouzdanim intervalima, dok je na slici 2.9 dana vjerojatnost po spolu, pri čemu „1“ označava ženski, a „2“ muški spol.

Slika 2.8: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u zadanom modelu osim varijable FARMER, s pripadnim pouzdanim intervalima



Slika 2.9: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u zadanom modelu osim varijable FARMER, po varijabli SPOL





2.4.4.1. Prva izmjena zadanog modela

Varijabla FARMER nije statistički značajna. Napravimo model koji ne uzima za nezavisnu varijablu varijablu FARMER. Rezultati su sljedeći:

Tablica 2.28: Dijagnostika izmijenjenog zadanog multivarijatnog logističkog modela

Broj korištenih observacija			
1092			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika za adekvatnost modela			
Kriterij	Samo sjesište	Sjecište i varijable	
-2 LOG L	836,39	666,44	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr >χ^2
Likelihood Ratio	169,95	4	<0,001

Iz tablice 2.28 vidimo da je broj korištenih opažanja u ovome modelu 1092. Također, kriterij konvergencije je zadovoljen. Dobiveni model jest statistički značajan ($\chi^2 = 169,95$, $st.sl. = 4$, $p < 0,001$).

Tablica 2.29: Rezultati ML procjene parametara za izmijenjeni zadani multivarijatni logistički model

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-4,8816	0,8582	32,3541	<,0001
SPOL_M	1	0,5463	0,2011	7,3810	0,0066
DULJINA ENS	1	0,0516	0,00540	91,3345	<,0001
BMI	1	-0,0426	0,0205	4,3201	0,0377
ST	1	0,00608	0,00409	2,2157	0,1366

Iz tablice 2.29 slijedi da je jednadžba dobivenog modela:

$$\begin{aligned} logit(p) = &-4,8816 + 0,5463 \cdot SPOL_M + 0,0516 \cdot DULJINA_ENS \\ &- 0,0426 \cdot BMI - 0,00608 \cdot ST. \end{aligned} \tag{2.5}$$

Tablica 2.30: Rezultati procjene omjera šansi efekata u izmijenjenom zadanom multivarijatnom logističkom modelu

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
SPOL_M	1,727	1,164	2,561
DULJINA ENS	1,053	1,042	1,064
BMI	0,958	0,921	0,998
ST	1,006	0,998	1,014

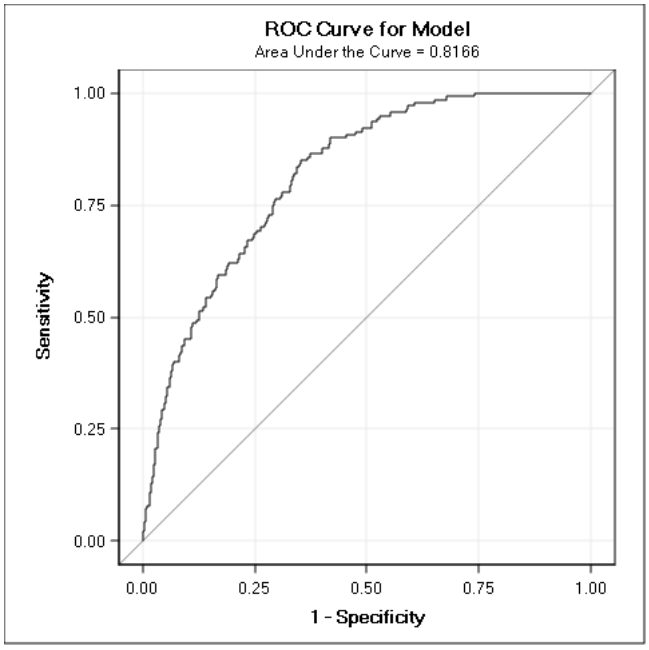
Tablicu 2.30 tumačimo na sljedeći način:

1. Prijelaz iz kategorije ženskog spola u kategoriju muškog spola povećava omjer šansi 1,727 puta, odnosno za $72,7\%$, i to je statistički značajno.
2. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,053 puta, to jest ta $5,3\%$, i to je statistički značajno.
3. Jedinično povećanje indeksa tjelesne mase smanjuje omjer šansi za razvijanje bolesti za $4,4\%$.
4. Jedinično povećanje sistoličkog tlak povećava omjer šansi za razvijanje bolesti 1,006 puta, to jest za $0,6\%$, i to nije statistički značajno.

Tablica 2.31: Prediktivna snaga izmijenjenog zadanog multivarijatnog logističkog modela

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
C	0,817

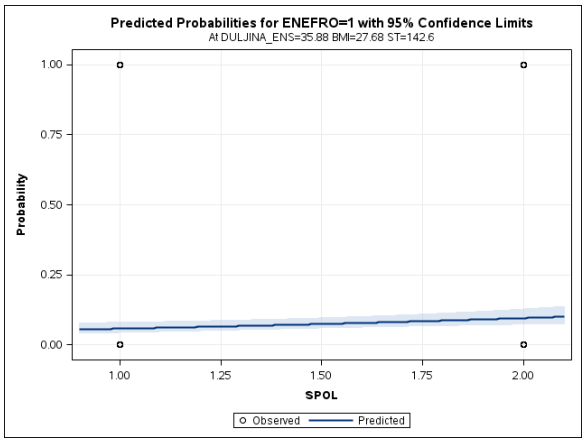
Iz tablice 2.31 iščitavamo da je prediktivna snaga modela 0,817, što znači je $81,7\%$ pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.10 se nalazi ROC krivulja zadanog modela.

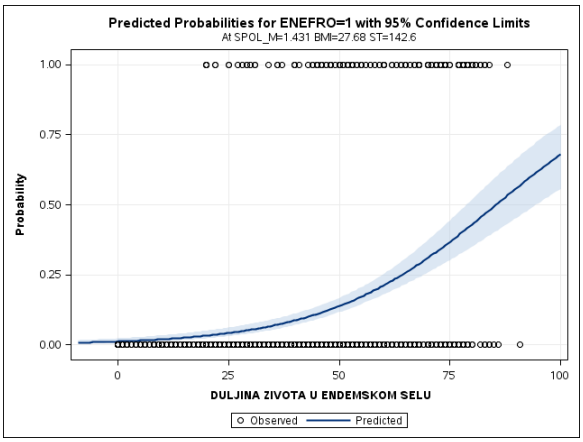


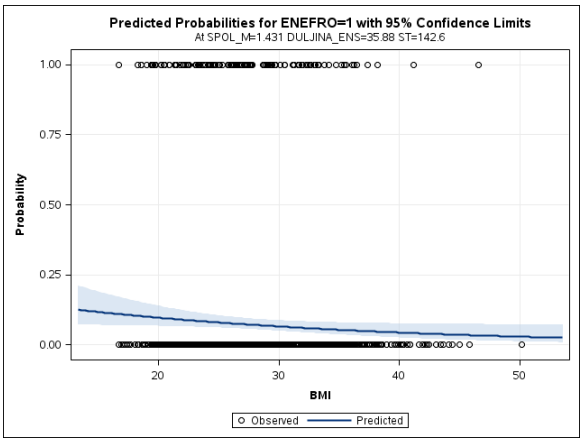
Slika 2.10: ROC krivulja prvog zadanog modela

Pogledajmo još krivulje vjerojatnosti za sve varijable u prvoj izmjeni zadanog modela, osim varijable SISTOLIČKI TLAK. Za svaku varijablu crtamo vjerojatnost da osoba ima bolest kada su vrijednosti preostalih nezavisnih varijabli u modelu njihova aritmetička sredina. Na slici 2.11 dana je vjerojatnost sa pripadnim pouzdanim intervalima, dok je na slici 2.12 dana vjerojatnost po spolu, pri čemu „1“ označava ženski, a „2“ muški spol.

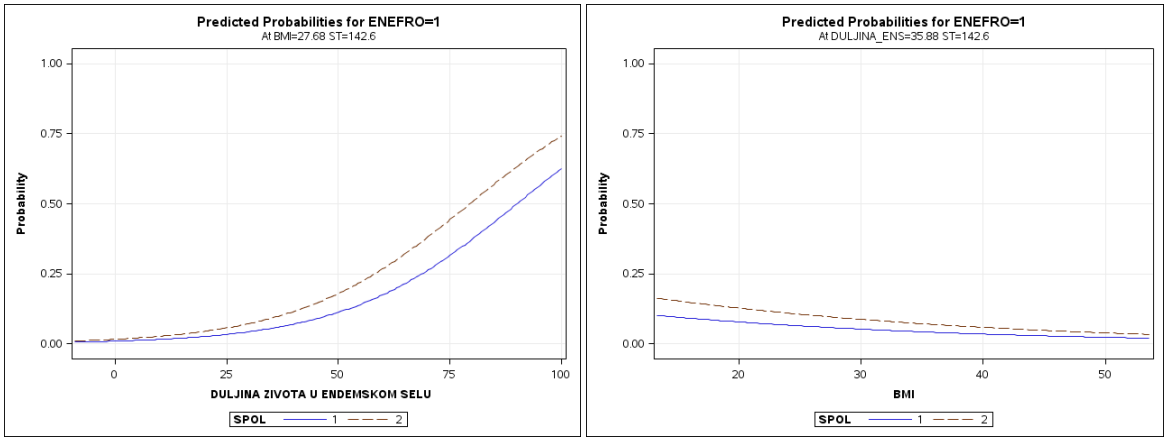
Slika 2.11: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u prvoj izmjeni zadanog modela osim varijable SISTOLIČKOG TLAKA, s pripadnim pouzdanim intervalima







Slika 2.12: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u prvoj izmjeni zadanog modela osim varijable SISTOLIČKI TLAK, po varijabli SPOL



2.4.4.2. Druga izmjena zadanog modela

Varijabla SISTOLIČKI TLAK nije statistički značajna. Napravimo model koji ne uzima za nezavisnu varijablu varijablu SISTOLIČKI TLAK. Rezultati su sljedeći:

Tablica 2.32: Dijagnostika izmijenjenog zadanog multivarijatnog logističkog modela

Broj korištenih observacija			
1092			
Status konvergencije modela			
Kriterij konvergencije (GCONV=1E-8) je zadovoljen			
Statistika za adekvatnost modela			
Kriterij	Samo sjecište	Sjecište i varijable	
-2 LOG L	857,75	681,39	
Testiranje H_0: $\beta = 0$			
Test	χ^2	DF	Pr >χ^2
Likelihood Ratio	176,36	3	<0,001

Iz tablice 2.32 vidimo da je broj korištenih opažanja u ovome modelu 1092. Također, kriterij konvergencije ja zadovoljen. Dobiveni model jest statistički značajan ($\chi^2 = 176,36$, $st.sl. = 3$, $p < 0,001$).

Tablica 2.33: Rezultati ML procjene parametara za izmijenjeni zadani multivarijatni logistički model

Analiza ML procjene					
VARIJABLA	DF	Procjena	Standardna greška	Wald χ^2	p
SJECIŠTE	1	-4,1602	0,7115	34,1863	<,0001
SPOL_M	1	0,4931	0,1984	6,1760	0,0129
DULJINA ENS	1	0,0558	0,00525	112,7779	<,0001
BMI	1	-0,0416	0,0200	4,3123	0,0377

Iz tablice 2.33 slijeda da je jednadžba dobivenog modela:

$$logit(p) = -4,1602 + 0,4931 \cdot SPOL_M + 0,0558 \cdot DULJINA_ENS - 0,0416 \cdot BMI. \quad (2.6)$$

Tablica 2.34: Rezultati procjene omjera šansi efekata u izmijenjenom zadanom multivarijatnom logističkom modelu

Odds Ratio procjena			
VARIJABLA	Procjena	95% Wald pouzdani interval	
SPOL_M	1,637	1,110	2,416
DULJINA ENS	1,057	1,047	1,068
BMI	0,959	0,922	0,998

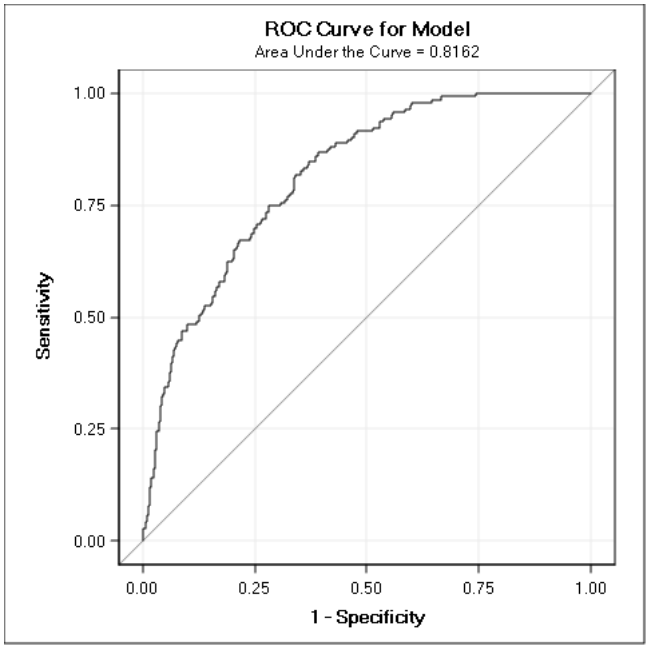
Tablicu 2.34 tumačimo na sljedeći način:

1. Prijelaz iz kategorije ženskog spola u kategoriju muškog spola povećava omjer šansi 1,637 puta, odnosno za $63,7\%$.
2. Jedna godina života dulje u endemskom selu povećava omjer šansi za razvijanje bolesti 1,057 puta, to jest ta $5,7\%$.
3. Jedinično povećanje indeksa tjelesne mase smanjuje omjer šansi za razvijanje bolesti za $4,2\%$.

Tablica 2.35: Prediktivna snaga izmijenjenog zadanog multivarijatnog logističkog modela

Povezanost predviđenih vjerojatnosti i opaženih odgovora	
C	0,816

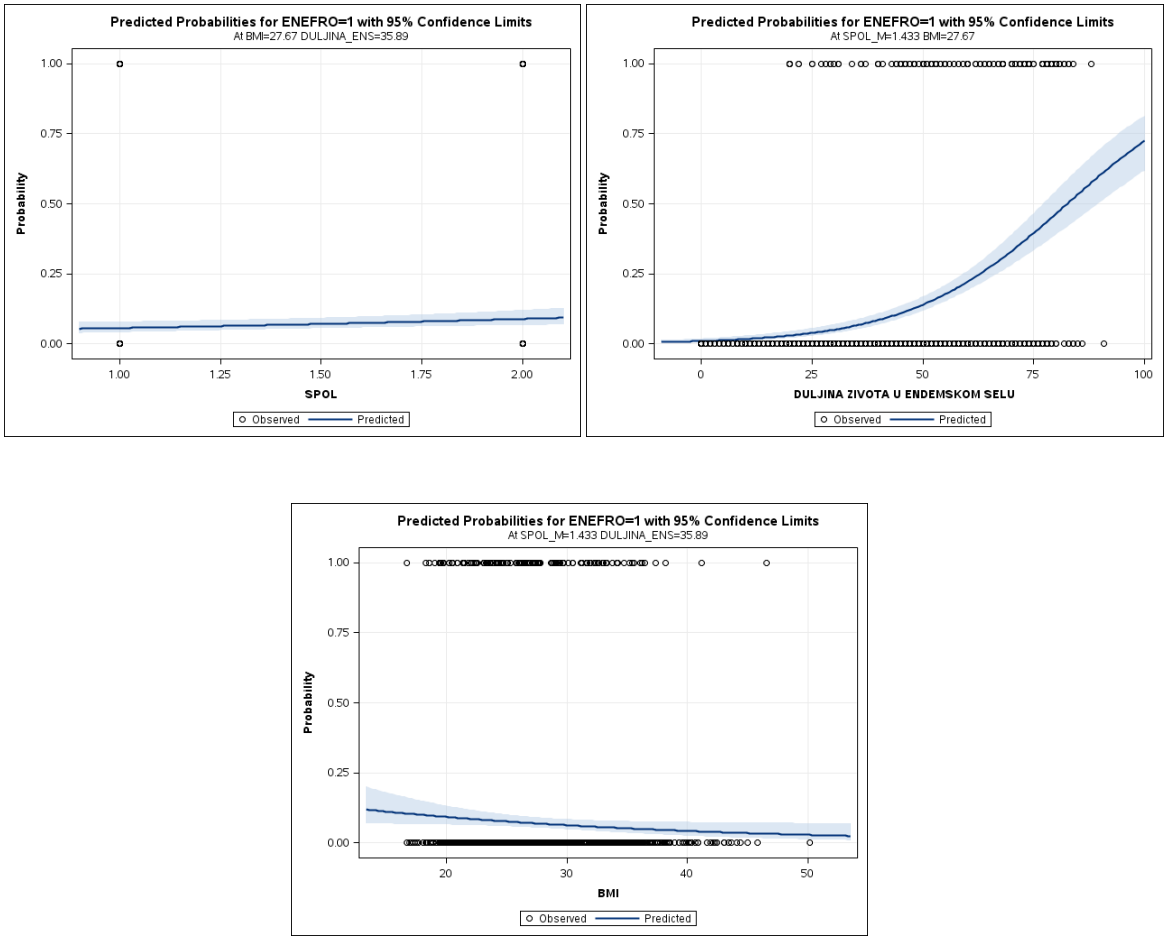
Iz tablice 2.35 isčitavamo da je prediktivna snaga modela 0,816, što znači je $81,6\%$ pojavnosti endemske nefropatije objašnjeno ovim modelom. Na slici 2.13 se nalazi ROC krivulja zadanog modela.



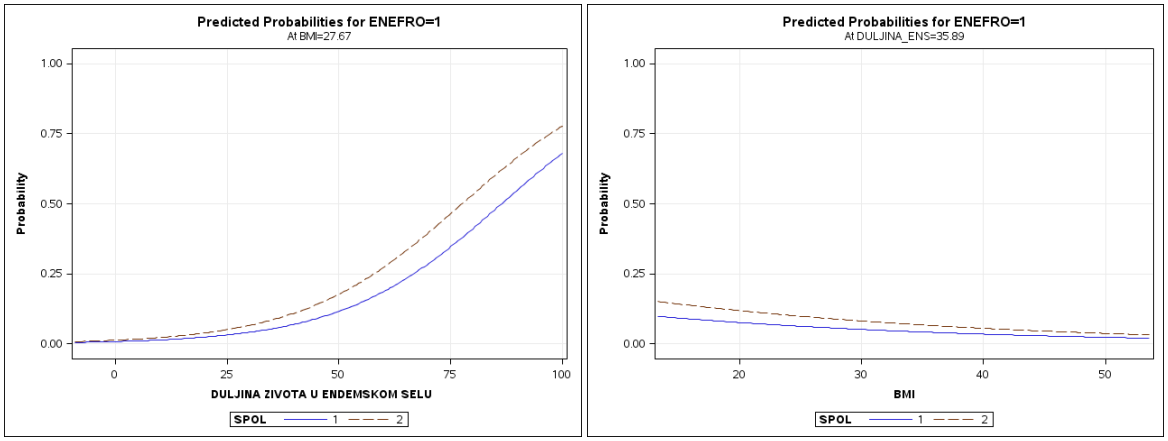
Slika 2.13: ROC krivulja prvog zadanog modela

Pogledajmo još krivulje vjerojatnosti za sve varijable u drugoj izmjeni zadanog modela. Za svaku varijablu crtamo vjerojatnost da osoba ima bolest kada su vrijednosti preostalih nezavisnih varijabli u modelu njihova aritmetička sredina. Na slici 2.14 dana je vjerojatnost sa pripadnim pouzdanim intervalima, dok je na slici 2.15 dana vjerojatnost po spolu, pri čemu „1“ označava ženski, a „2“ muški spol.

Slika 2.14: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u drugoj izmjeni zadanog modela, s pripadnim pouzdanim intervalima



Slika 2.15: Vjerojatnost da osoba ima bolest za svaku pojedinu varijablu u drugoj izmjeni zadanog modela, po varijabli SPOL



2.5 Zaključak

Vidimo da se kroz sve 3 korištene metode selekcije kao eksplanatorne varijable pojavljuju „duljina života u endemskom selu“, „razina $\alpha 1$ mikroglobulina“ te „pozitivna obiteljska anamneza na endemsku nefropatiju“. Najveću prediktivnu snagu, jednaku 0,996, ima model dobiven „FORWARD“ metodom selekcije. Usprkos tome, ja bih odabrala model dobiven „STEPWISE“ metodom selekcije, iz razloga što on sadrži samo prethodno nevedene 3 varijable, te također ima vrlo visoku prediktivnu snagu, to jest vjerojatnost za razvijanje endemske nefropatije dobivena ovim modelom objašnjava pojavu 98,43% opaženih slučajeva bolesti u korištenom uzorku.

2.6 Prilog

U nastavku je dan korišteni kod u SAS Studiu:

```
PROC IMPORT DATAFILE="/folders/myfolders/diplomski/EN.xls"
OUT=WORK.EN
DBMS=XLS
REPLACE;
run;

data dummy; set WORK.EN;
if edukacija=0 then do; edukacija1=0;edukacija2=0;edukacija3=0;end;
if edukacija=1 then do; edukacija1=1;edukacija2=0;edukacija3=0;end;
if edukacija=2 then do; edukacija1=0;edukacija2=1;edukacija3=0;end;
if edukacija=3 then do; edukacija1=0;edukacija2=0;edukacija3=1;end;
run;

data dummy; set dummy;
if income=0 then do; income1=0;income2=0;end;
if income=1 then do; PRIHODI_SREDNJI=1;income2=0;end;
if income=2 then do; income1=0;income2=1;end;
run;

data dummy; set dummy;
if alkohol=0 then do; alkohol1=0;alkohol2=0;end;
if alkohol=1 then do; alkohol1=1;alkohol2=0;end;
if alkohol=2 then do; alkohol1=0;alkohol2=1;end;
run;

data dummy; set dummy;
if analgetici=0 then do; analgetici1=0;analgetici2=0;end;
if analgetici=1 then do; analgetici1=1;analgetici2=0;end;
if analgetici=2 then do; analgetici1=0;analgetici2=1;end;
run;

proc means data=WORK.EN n nmiss mean stddev min median max maxdec=2;
var age_FINAL DULJINA_ENS visina tezina BMI ST DT alfa1_micro alb
hemoglobine creatinine_serum creatinine ACR alfa1CR alfa1 CKD_epi SDU DLB DDB
GLUCOSE;
```

```
run;

proc freq data=WORK.EN;
table SPOL_M edukacija farmer_DA income pusenje alkohol analgetici
ENEFRO AH_DA DM anamneza_DA /nocum;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=age_FINAL;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=DULJINA_ENS;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=visina;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=tezina;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=BMI;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=ST;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=DT;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=alfa1_micro;
run;
```

```
proc logistic data=WORK.EN descending;
model ENEFRO=alb;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=hemoglobine;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=creatinine_serum;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=creatinine;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=ACR;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=alfa1CR;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=alfa1;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=CKD_epi;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=SDU;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=DLB;
run;
```

```
proc logistic data=WORK.EN descending;
model ENEFRO=DDB;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=GLUCOSE;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=SPOL_M;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=farmer;
run;

proc logistic data=dummy descending;
model ENEFRO= edukacija1 edukacija2 edukacija3;
run;

proc logistic data=dummy descending;
model ENEFRO= income1 income2;
run;

proc logistic data=dummy descending;
model ENEFRO= alkohol1 alkohol2;
run;

proc logistic data=dummy descending;
model ENEFRO= analgetici1 analgetici2;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=pusenje;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=AH_DA;
```

```
run;

proc logistic data=WORK.EN descending;
model ENEFRO=DM;
run;

proc logistic data=WORK.EN descending;
model ENEFRO=anamneza_DA;
run;

proc freq data=WORK.EN;
table ENEFRO*SPOL_M;
run;

proc freq data=WORK.EN;
table ENEFRO*ANAMNEZA_DA;
run;

proc logistic data=dummy descending;
model ENEFRO=age_FINAL DULJINA_ENS visina tezina
BMI ST DT A1M ALB hemoglobine creatinine_serum
creatinine ACR ALFA1 A1AT CKD_epi SDU DLB DDB
GLUCOSE SPOL_M edukacija1 edukacija2 edukacija3
FARMER_DA PRIHODI_SREDNJI income2 pusenje alkohol1
alkohol2 analgetici1 analgetici2 AH_DA DM
ANAMNEZA_DA /selection= FORWARD outroc=prob;
run;

proc logistic data=dummy descending;
model ENEFRO=age_FINAL DULJINA_ENS visina tezina
BMI ST DT A1M ALB hemoglobine creatinine_serum
creatinine ACR ALFA1 A1AT CKD_epi SDU DLB DDB
GLUCOSE SPOL_M edukacija1 edukacija2 edukacija3
FARMER_DA PRIHODI_SREDNJI income2 pusenje alkohol1
alkohol2 analgetici1 analgetici2 AH_DA DM
ANAMNEZA_DA /selection= BACKWARD outroc=prob;
run;

proc logistic data=dummy descending;
```

```
model ENEFRO=age_FINAL DULJINA_ENS visina tezina
BMI ST DT A1M ALB hemoglobine creatinine_serum
creatinine ACR ALFA1 A1AT CKD_epi SDU DLB DDB
GLUCOSE SPOL_M edukacija1 edukacija2 edukacija3
FARMER_DA PRIHODI_SREDNJI income2 pusenje alkohol1
alkohol2 analgetici1 analgetici2 AH_DA DM
ANAMNEZA_DA /selection= STEPWISE outroc=prob;
run;

proc logistic data=dummy descending plots=effect;
model ENEFRO=SPOL_M DULJINA_ENS FARMER_DA BMI ST / outroc=prob;
run;

proc logistic data=dummy descending plots=effect;
model ENEFRO=SPOL_M DULJINA_ENS BMI ST / outroc=prob;
run;

proc logistic data=dummy descending plots=effect;
model ENEFRO=SPOL_M DULJINA_ENS BMI / outroc=prob;
run;
```

Bibliografija

[1] Allison P.D., *Logistic Regression Using SAS: Theory and Application*, SAS Institute Inc., Cary, NC, 1999.

[2] Koch G.G. Stokes M.E., Davis C.S., *Categorical Data Analysis Using SAS System*, SAS Institute Inc., Cary, NC, 2000.

Sažetak

U ovome radu modeliramo pojavnost endemske nefropatije logističkom regresijom. Promatramo modele nastale „FORWARD“, „BACKWARD“ i „STEPWISE“ metodom selekcije, te zadani model i njegove varijacije. Statistička obrada podataka rađena je u statističkom paketu SAS Studio.

U svakom modelu dobivenom SAS-ovim metodama selekcije kao rizični faktori za razvijanje endemske nefropatije pojavljuju se duljina života u endemskom selu, razina $\alpha 1$ mikroglobulina te pozitivna obiteljska anamneza na endemsku nefropatiju.

Summary

In this thesis, we are modelling prevalence of endemic nephropathy with logistic regression. We are observing models which are created with "FORWARD", "BACKWARD" and "STEPWISE" selection methods, and also demanded model and its variations. Statistical analysis was done in statistical software SAS Studio.

In each model created with SAS selection methods, living duration in endemic village, level of $\alpha 1$ microglobulin and positive family history on endemic nephropathy are appearing as risk factors for developing an endemic nephropathy.

Životopis

Rođena sam 21.04.1990. u Zagrebu. Završila sam Osnovnu glazbenu školu Zlatka Balokovića te OŠ Dobriše Cesarića u Zagrebu, a 2008. godine XV gimnaziju u Zagrebu. 2003., 2004. i 2006. godine sudjelovala sam na Državnom natjecanju iz matematike. 2013. godine završila sam Preddiplomski Sveučilišni studij matematike (inženjerski smjer), na Prirodoslovno-matematičkom fakultetu u Zagrebu, te iste godine nastavljam školovanje upisivanjem diplomskog sveučilišnog studija Matematičke statistike na Prirodoslovno - matematičkom fakultetu u Zagrebu.

2006. godine počela sam raditi kao instruktor grupnih aerobik programa u „Fitness centru JUMP“. Također, 2012. godine počinjem raditi kao vanjski suradnik u Službi za Medicinsku Informatiku i Biostatistiku Hrvatskog Zavoda za Javno Zdravstvo.

Kao vanjski suradnik na HZJZ-u, 2012. godine sudjelovala sam radu „STUDIJA PROCJENE MOGUĆEG UTJECAJA EKOLOŠIH ČIMBENIKA NA ZDRAVSTVENO STANJE STANOVNIŠTVA BRODSKO-POSAVSKE ŽUPANIJE“, objavljenom na internetskoj stranici `http://zzjzbpz.hr/images/stories/STUDIJA_PROCJENE.pdf`. Također, sudjelovala sam u projektu „ENDEMSKA NEFROPATIJA U HRVATSKOJ, EPIDEMIOLOGIJA, DIJAGNOSTIKA, ETIOPATOGENEZA“, iz kojeg je u siječnju 2015. godine u „Clinical Journal of the American Society of Nephrology“ objavljen članak „Chronic dietary exposure to aristolochic acid and kidney function in native farmers from Croatian endemic area and Bosnian immigrants“.