Procjena genetskih parametara binomne varijable koristeći GLIMMIX SAS proceduru

Matanović, Sara

SVEUČILIŠTE U ZAGREBU

AGRONOMSKI FAKULTET

Sara Matanović

PROCJENA GENETSKIH PARAMETARA BINOMNE VARIJABLE KORISTEĆI GLIMMIX SAS PROCEDURU

DIPLOMSKI RAD

Zagreb, 2016.

SVEUČILIŠTE U ZAGREBU
AGRONOMSKI FAKULTET
Genetika i oplemenjivanje životinja

SARA MATANOVIĆ

PROCJENA GENETSKIH PARAMETARA BINOMNE VARIJABLE KORISTEĆI GLIMMIX SAS PROCEDURU

DIPLOMSKI RAD

Mentor: Prof. dr. sc. Miroslav Kapš
Neposredni voditelj: Doc. dr. sc. Maja Ferenčaković

Zagreb, 2016.

Ovaj diplomski rad je ocijenjen i obranjen dana _____sa ocjenom

_____ pred Stručnim povjerenstvom u sastavu:

1. Prof. dr. sc. Miroslav Kapš

_____

2. Prof. dr. sc. Ino Čurik

_____

3. Doc. dr. sc. Vlatka Čubrić Čurik

_____

4. Doc. dr. sc. Maja Ferenčaković

_____

Zahvale

Zahvaljujem se:

- Prof. dr. sc. Miroslavu Kapšu što je pristao biti moj mentor, na svoj pomoći tijekom izrade ovoga rada, na svim savjetima te na beskrajnom strpljenju i podršci
- Doc. dr. sc. Maji Ferenčaković na pomoći pri pisanju rada i svim savjetima
- Doc. dr. sc. Vlatki Čubrić-Čurik i Prof. dr. sc. Ini Čuriku kao članovima Stručnog povjerenstva na izdvojenom vremenu
- Svojim roditeljima te bakama i djedu koji su omogućili moje školovanje i uvijek bili velika potpora

Sažetak na hrvatskom

Procijenjene su izvedbe različitih modela za analizu genetskih parametara binomne varijable. Glavni cilj je bio metodološki prikazati neku binomnu varijablu, usporediti probit i logit modele međusobno te s običnim linearnim modelom. Podatci su simulirani koristeći SAS 9.3 program. Simulirana su dva seta podataka, oba s 1000 bikova te 100 kćeri po svakome biku, a razlikovala su se u proporciji binarne varijable koja je iznosila 0,2 za jedan set podatka te 0,5 za drugi set. Podatci su zatim obrađeni procedurama mixed za obični linearni model (model očeva) i glimmix za uopćeni linearni mješoviti model. Na temelju dobivenih vrijednosti izračunati su heritabiliteti te su uspoređene vrijednosti varijance očeva i varijance ostatka za svaki od navedenih modela. Uočene su razlike između običnog linearnog te probit i logit modela. Također su uočene razlike između probit i logit modela, ali nisu bile značajno velike.

Ključne riječi: binomna varijabla, obični linearni model, probit model, logit model, heritabilitet

Sažetak na engleskom

The performances of different models for genetic analysis of binomial variable have been evaluated. The main goal was to methodologically show some binomial variable, compare the probit and logit models mutually and with a linear model. The data were simulated using SAS 9.3 software. Two sets of data were simulated, both having 1000 bulls with a 100 daughters for each bull. The data sets differed in the proportion of the binary variable that was 0,2 for one data set and 0,5 for the other. The data were processed using proc mixed for linear model (sire model) and proc glimmix for generalized linear mixed model. Based on the obtained values heritability was calculated for each model. Also, the values of sire and error variances were compared. Differences between linear model and probit and logit models were noticed. Differences between probit and logit model were also noticed but were not significantly large.

Key words: binomial variable, linear model, probit model, logit model, heritability

1. Uvod

Binomne varijable (ponavljanja binarnih varijabli) imaju široku upotrebu u istraživanju i selekciji životinja stoga je bitno kod analize takvih varijabli koristiti pravilne modele procjene genetskih parametara. Svojstva, kao što su stopa rasta ili težina, su obično izražena na kontinuiranoj ljestvici te se smatra da su normalno raspodijeljena. Svojstva kao što su stopa teljenja, lakoća teljenja ili stopa preživljavanja pripadaju kategoričkim (binomnim) svojstvima za koja se pretpostavlja da imaju u osnovi, kontinuiranu raspodjelu (Guerra, 2004). Ako se analiza binomnih varijabli provodi pod pretpostavkom da su normalne često se znaju pojaviti problemi. Drugim riječima, za modele koji provode analizu kontinuiranih odgovora često se govori da nisu prikladni za analizu kategoričkih odgovora (Thompson, 1979; Gianola, 1982; Koch i sur., 1990; Ramizez-Valverde i sur., 2001). U takvim situacijama obično se koristi transformacija kategoričkog svojstva ili se model za kontinuirane podatke primjenjuje na binomno svojstvo (Guerra, 2004). U ovome radu su uz jednostavni linearni model (model očeva) korišteni i uopćeni linearni modeli (engl., Generalized Linear Models) koji predstavljaju produžetak linearnih modela. Oni omogućavaju nelinearnost i nestalne varijance unutar podataka (Hastie i Tibshirani, 1990). Bazirani su na pretpostavljenoj vezi („link“ funkciji) između prosjeka zavisne varijable i linearne kombinacije nezavisnih varijabli te podatci mogu biti iz različitih raspodjela uključujući binomnu (Guisan, Edwards i Hastie, 2002). Unutar obitelji uobičajenih linearnih modela među najkorištenijima su probit i logit modeli koji se često koriste u procjeni genetskih parametara.

Cilj ovoga rada bio je putem simuliranih podataka metodološki prikazati postupak analize binarnih i binomnih varijabli odnosno procijeniti učinkovitost različitih modela kojima je navedena procjena provedena.

2. Binarna i binomna varijabla

Varijable su set opažanja određenog svojstva i mogu poprimiti različite vrijednosti. Dijele se na kvalitativne (kategoričke) i kvantitativne (numeričke). Kvalitativne se dalje dijele na nominalne i ordinalne, a kvantitativne na kontinuirane i diskretne. Binarna i binomna varijabla su primjeri diskretnih varijabli. Međutim, s obzirom da se svakom kategoričkom svojstvu može pridružiti određena numerička vrijednost tada će i kategorička varijabla koja može poprimiti samo jednu od dvije moguće vrijednosti biti binarna.

Binarne varijable su opažanja koja se javljaju u jednom od dva moguća stanja koja se često označavaju kao 1 ili 0 odnosno kao DA ili NE. Primjeri ovakvih varijabli u stočarstvu su: dobra/loša sperma, bolesna/zdrava životinja, muško/žensko, mrtvorođeni/živorođeni itd. Binomne varijable predstavljaju broj povoljnih pokušaja (y) u ukupno n pokušaja odnosno to su ponavljanja binarne varijable. Binomna raspodjela je raspodjela vjerojatnosti y povoljnih pokušaja (opažanja) u ukupno n pokušaja i određena je parametrom p (vjerojatnost elementarnog događaja) i brojem pokušaja n.

Dakle, karakteristike binomnog pokusa su sljedeće (Kapš i Lamberson, 2009):

1) Pokus se sastoji od n jednakovvrijednih pokušaja koji su međusobno nezavisni

2) Postoje samo dva moguća rezultata pokušaja, a označavaju se s DA ili NE (1 ili 0)

3) Vjerojatnost dobivanja rezultata DA je jednaka od pokušaja do pokušaja i označava se kao p. Vjerojatnost dobivanja rezultata NE označava se kao q. Dakle, $p + q = 1$.

4) Slučajna varijabla y označava broj povoljnih rezultata u ukupno n pokušaja.

Binomna raspodjela ima široku upotrebu u istraživanju i selekciji životinja, uključujući pitanja kao što su da li će životinja zadovoljiti neki standard, da li je krava gravidna ili nije, da li će tele preživjeti teljenje itd.

3. Model očeva

Model očeva (engl., Sire Model) je linearni mješoviti model koji sadrži fiksne i slučajne utjecaje. Kod njega se u obzir uzima samo genetski utjecaj očeva, odnosno pretpostavlja se da je svaki bik (otac) paren s prosječnom majkom te da te majke nisu srodne. To je model koji procjenjuje uzgojnu vrijednost bikova na temelju njihovih kćeri.

$$\boldsymbol{y = X_b + Z_u + e,}$$

gdje je:

y = n × 1 vektor opažanja

b = p × 1 vektor nepoznatih konstanta

u = q × 1 vektor nepoznatih utjecaja slučajnih varijabli

e = N × 1 vektor nepoznatih utjecaja ostatka

X i Z = poznate matrice N × p i N × q koje povezuju elemente b i u s elementom y

Pedesetih godina prošlog stoljeća razvijene su metode „najboljih linearnih nepristranih procjenitelja“ (engl., Best Linear Unbiased Estimates – BLUE) fiksnih utjecaja i „najboljih linearnih nepristranih previđanja“ (Best Linear Unbiased Predictions – BLUP) slučajnih utjecaja koje su omogućile da mješoviti model postane značajno područje u statističkim istraživanjima (Henderson i sur., 1959). Najranije upotrebe BLUP-a za procjenu uzgojnih vrijednosti (posebice kod mliječnih goveda) zasnovane su na modelu očeva.

Linearni mješoviti model pretpostavlja da je veza između prosjeka zavisne varijable i fiksnih i slučajnih utjecaja linearna funkcija, da varijanca nije funkcija prosjeka te da slučajni utjecaji prate normalnu raspodjelu. Zbog toga se mogu javiti problemi kada je u pitanju varijabla poput binarne. Unatoč tome, nekoliko autora je preporučilo korištenje modela očeva kod procjene genetskih parametara binomnih svojstava (Heringstad i sur., 2003; Phocas i Laloe, 2003). Oni navode kako je u slučaju kada je uzorak malen bolje koristiti linearni model.

4. Uopćeni linearni model (Generalized Linear Model - GLM)

Pojam uopćeni linearni model 1972. uveli su znanstvenici Nelder i Wedderburn. Model se bazira na pretpostavljenoj vezi (koja se naziva „link“ funkcija) između prosjeka zavisne varijable i linearne kombinacije nezavisnih varijabli. Model se može prikazati matrično:

$$\boldsymbol{\eta = g(\mu) = g[E(y)] = X\beta},$$

gdje je $\eta = g(\mu)$ funkcija prosjeka zavisne varijable y („link“ funkcija), $X\beta$ je linearna kombinacija nezavisnih varijabli i pripadajućih parametara.

Vektor prosjeka je:

$$\boldsymbol{\mu = E(y) = g^{-1}(\eta),}$$

gdje je g^{-1} inverzna „link“ funkcija odnosno funkcija koja transformira $X\beta$ na skalu prosjeka.

Primjeri nenormalnih varijabli koji se mogu analizirati uopćenim linearnim modelima su binomna, Poisson, negativna binomna i gama varijabla. U Tablici 1 prikazane su kanonske „link“ funkcije za nekoliko raspodjela vjerojatnosti. Binomna raspodjela s probit „link“ funkcijom vodi do threshold (probit) modela (Gianola and Foulley, 1983) kakav će biti korišten u ovome radu.

Tablica 1. Popis raspodjela i „link“ funkcija; izvor: Guerra J. L. L. (2004). Statistical models and genetic evaluation of binomial traits (Doctoral dissertation, Louisiana State University).

Raspodjela	“Link” funkcija
Normalna	Identitet
Binomna	Logit/Probit
Poisson	Log
Gauss	Recipročna
Negativna binomna	Log
(Ovo su najčešće i sve pripadaju eksponencijalnoj obitelji)	

4.1. Komponente uopćenih linearnih modela:

1) Zavisna varijabla s pripadajućom raspodjelom

2) Skup nezavisnih varijabli (X_1,X_2,..., X_j) s pripadajućim parametrima, npr. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. Nezavisne varijable se često zovu i prediktori

3) „Link" $g(\mu)$ funkcija koja specificira vezu između prosjeka zavisne varijable i nezavisnih varijabli

Dakle, uopćeni linearni model u skalarnom obliku glasi:

$$\boldsymbol{\eta_i = g(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j}$$

5. Uopćeni linearni mješoviti model (Generalized Linear Mixed Model – GLMM)

Uopćeni linearni mješoviti model (Generalized Linear Mixed Model – GLMM) je produžetak uopćenih linearnih modela kod kojega linearni prediktor uz fiksne utjecaje sadrži i slučajne utjecaje.

Matrično uopćeni linearni mješoviti model prikazujemo na sljedeći način

$$\boldsymbol{\eta = g[E(y|u)] = X\beta + Zu,}$$

gdje je prosjek y uvjetno definiran slučajnim utjecajem, β je skup nezavisnih varijabli s pripadajućom matricom X, a u je vektor slučajnih varijabli s pripadajućom matricom Z.

Uvjetovano očekivanje je

$$\boldsymbol{E(y|u) = g^{-1}(\eta) = g^{-1}(X\beta + Zu)},$$

što predstavlja vektor očekivanja zavisne varijable za zadani slučajni efekt u_i.

Skalarni prikaz izgleda ovako

$$\boldsymbol{\mu_i = g^{-1}(X_i\beta + u_i)},$$

gdje su μ_i i u_i prosjek i utjecaj grupe i, a X_i predstavlja redove matrice X bitne za grupu i.

Kao i kod uopćenih linearnih modela, varijanca se može sastojati od funkcije varijance i skalarnih parametara ili parametara disperzije. Dakle, uvjetovana varijanca je

$$\boldsymbol{Var}(\boldsymbol{y}|\boldsymbol{u}) = [\boldsymbol{V}(\boldsymbol{\mu})]^{\mathbf{1/2}}\boldsymbol{\Phi}^{\mathbf{2}}[\boldsymbol{V}(\boldsymbol{\mu})]^{\mathbf{1/2}},$$

gdje je $V(\mu)$ diagonalna matrica funkcija varijance $V(\mu)$, a Φ^2 je matrica skalarnih parametara. U literaturi se matrica skalarnih parametara često označava i kao R.

Varijanca slučajnih utjecaja obično se označava matricom G

$$\boldsymbol{Var}(\boldsymbol{u}) = \boldsymbol{G}$$

G je na "link" skali, a Φ^2 na opaženoj skali.

6. Modeli binarnog odgovora

Logit i probit modeli su među najkorištenijim modelima kada se govori o binarnim zavisnim varijablama (Hahn i Soyer, 2005). S obzirom da u slučaju binarnih zavisnih varijabli metoda običnih najmanjih kvadrata više ne pruža najboljeg linearnog nepristranog procjenitelja (best linear unbiased estimator (BLUE)) (Park, 2009), logit i probit modeli procjenjuju se putem metode „maximum likelihood". Ova metoda ima dobra svojstva i najpreciznija je kada se radi o velikim uzorcima (Horowitz i Savin, 2001).

6.1. Logit model

Utjecaj nezavisnih varijabli na binarnu zavisnu varijablu može biti objašnjen uporabom uopćenog linearnog modela i logit „link" funkcije. Takav model često se naziva logit model. Kada su nezavisne varijable kontinuirane odgovarajući model jest model logističke regresije koji glasi

$$\boldsymbol{\eta_i} = \boldsymbol{log}[\boldsymbol{p_i}/(\mathbf{1} - \boldsymbol{p_i})] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{1i}} + \cdots + \boldsymbol{\beta_{p-1} x_{(p-1)i}},$$

gdje je $log[p_i/(1 - p_i)]$ logit „link" funkcija, x_{1i}, x_{2i},...., $x_{(p-1)i}$ su nezavisne varijable, a β_0, β_1, β_2,...., β_{p-1} regresijski parametri.

Jednostavna logistička regresija jest logistička regresija sa samo jednom nezavisnom kontinuiranom varijablom

$$\boldsymbol{\eta_i} = \boldsymbol{log}[\boldsymbol{p_i}/(\mathbf{1} - \boldsymbol{p_i})] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_i}$$

Nezavisne varijable također mogu biti kategoričke. Primjerice, jednosmjeran logit model također može biti definiran kao

$$\boldsymbol{\eta_i = log[p_i/(1 - p_i)] = m + \tau_i},$$

gdje je *m* ukupan prosjek proporcije logaritamske skale, a τ_i je efekt grupe *i*.

Definiranje logit funkcije osigurava da su procijenjene vrijednosti zavisne varijable uvijek između 0 i 1. Greške u modelu imaju binomnu raspodjelu podijeljenu s *n*.

Funkcija varijance glasi

$$\boldsymbol{V(\mu) = V(p) = pq = p(1 - p)},$$

gdje je *q = 1 – p*.

Dakle, varijanca binomnih proporcija *y/n* je:

$$\boldsymbol{Var\left(\frac{y}{n}\right) = \frac{pq}{n} = \frac{1}{n} V(p)\phi^2}$$

Funkcija varijance *V(p)* mora biti podijeljena s *n* jer je proporcija binomna varijabla podijeljena s *n*. Iz toga slijedi da je parametar disperzije ϕ^2.

Svojstvo logističke regresije jest da je varijanca *y/n* funkcija od *p*. Model uzima u obzir heterogenost varijance definirajući funkciju varijance. Prosjek i varijanca ovise o parametru *p*. Dakle, ako nezavisne varijable utječu na parametar *p* također će utjecati na prosjek i varijancu.

6.2. Probit model

Standardna normalna varijabla se može transformirati u binarnu varijablu ako je definirano sljedeće: svim vrijednostima koje su manje od neke vrijednosti η pridružuje se vrijednost 1, a svim vrijednostima koje se veće od η pridružuje se vrijednost 0. Proporcija vrijednosti koje iznose 1 ili 0 određena je područjem ispod normalne raspodjele, tj. koristeći kumulativnu normalnu raspodjelu pomoću koje se određuje vjerojatnost.

Inverzna kumulativna normalna raspodjela se naziva probit funkcija te se stoga takvi modeli nazivaju probit modeli. Inverzna „link“ funkcija jest kumulativna normalna raspodjela, a prosjek je

$$\boldsymbol{\mu = p = F(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}}\, e^{-0.5z^2} dz},$$

gdje z označava standardnu normalnu varijablu s prosjekom 0 i varijancom 1. „Link“ funkcija se naziva probit „link“ i označava se kao

$$\boldsymbol{\eta = F^{-1}(\mu)}$$

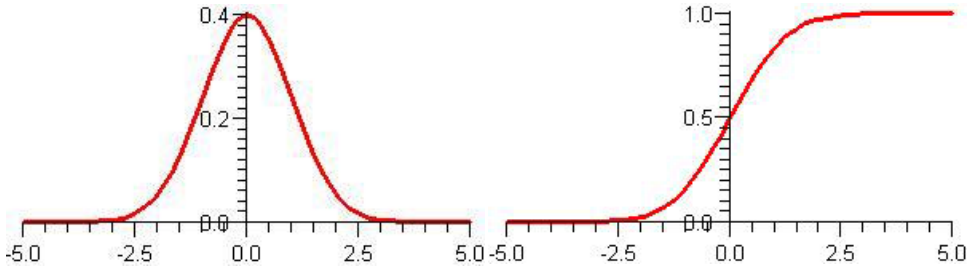
Utjecaji nezavisnih varijabli na η definirani su kao

$$\boldsymbol{\eta_i = F^{-1}(\mu_i) = x_i \beta},$$

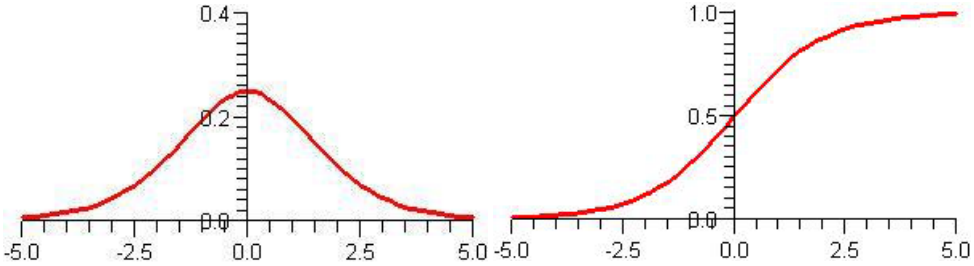
gdje je x_i nezavisna varijabla, a β je parametar regresije.

6.3. Razlike između probit i logit modela

Ključna razlika između logit i probit modela leži u raspodjeli grešaka. Kod logit modela smatra se da greške prate standardnu logističku raspodjelu gdje je prosjek 0, a varijanca $\frac{\pi^2}{3}$, $(\lambda)^{\varepsilon} = \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^{2!}}$ Kod probit modela se smatra da greške prate standardnu normalnu raspodjelu, $\emptyset(\varepsilon) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\varepsilon^2}{2}}$, a varijanca je 1 (Park, 2009.



PDF i CDF standardne normalne raspodjele



PDF i CDF standardne logističke raspodjele

Figura 1. Raspodjele standardnih normalnih i standardnih logističkih vjerojatnosti; izvor: Park H. M. (2009) Regression Models for Binary Dependent Variables Using Stata, SAS, R, LIMDEP, and SPSS. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

Funkcija gustoće vjerojatnosti (engl., probability density function (PDF)) standardne normalne varijable ima viši vrhunac i tanje repove od raspodjele standardne logističke vjerojatnosti (Figura 1). Standardna logistička raspodjela izgleda kao da je netko povinuo vrhunac standardne normalne raspodjele i zategnuo repove. Kao rezultat, funkcija kumulativne gustoće (engl., cumulative density function (CDF)) standarde normalne raspodjele je strmija u sredini od CDF standarde logističke raspodjele i brzo dostiže 0 s lijeve strane i 1 s desne strane. Dakle, može se zaključiti da logit ima nešto ravnije krajeve, tj. normalna ili probit krivulja brže dostiže osi od logističke krivulje (Park, 2009).

Ova dva modela stvaraju različite procjene parametara. Kod modela binarnog odgovora, procjene logit modela su ugrubo $\pi/\sqrt{3}$ veće nego one probit modela. Međutim, procjenitelji na kraju imaju skoro isti standardizirani utjecaj na nezavisne varijable (Long, 1997). Generalno mišljenje jest da je u većini slučajeva odabir „link“ funkcije pitanje vlastitog izbora (Maddala, 1983; Davidson i MacKinnon, 1993; Long, 1997; Powers i Xie, 2000; Fahrmeir i Tutz, 2001). Radovi vezani uz sličnosti i razlike između probit i logit modela mogu se pratiti sve do 1967. kada su Chambers i Cox uočili da su razlike vidljive samo kod velikih uzoraka i kada se pojavljuju određeni ekstremni uzorci unutar podataka. Izbor između logit i probit modela je više vezan uz procjenu i prisnost nego uz teorijske i interpretativne aspekte. Kako se razvijaju novi algoritmi tako se i važnost ovoga problema smanjuje (Park, 2009).

7. Materijali i metode

Koristeći SAS 9.3 program (SAS Institute Inc., 2011) simulirana su dva seta podataka (dvije populacije). Oba seta sastojala su se od 1000 bikova te od 100 kćeri po svakome biku, a razlikovala su se po proporciji binarne varijable p0 u populaciji. Za prvi set podataka p0 je iznosio 0,2, a za drugi set 0,5. Kod simuliranja navedenih podataka zadano je da je varijanca očeva 0,16668, a varijanca pogreške 1 te su iz njih dalje izračunate standardne devijacije te fenotipska varijanca. Uporabom funkcije RAND koja generira slučajne brojeve iz različitih kontinuiranih i diskretnih raspodjela simulirani su efekt očeva i efekt greške na standardnoj normalnoj ljestvici te binarna varijabla. Nadalje, simulirani su fiksni utjecaji 3 farme te je izračunata funkcija kumulativne raspodjele (CDF) za normalnu raspodjelu koja označava vjerojatnost da će neka varijabla poprimiti vrijednost manju ili jednaku y gdje y označava fenotip na normalnoj ljestvici. U analizi podataka korišteno je nekoliko procedura od kojih su najvažnije procedura MIXED te procedura GLIMMIX.

Procedura MIXED se koristi u analizi mješovitih linearnih modela, dakle podrazumjevajući normalnu raspodjelu podataka. Međutim, u ovome radu više je pažnje posvećeno proceduri GLIMMIX koja može analizirati i podatke s nenormalnim raspodjelama koristeći uopćene mješovite lineane modele. U radu su također korišteni logit i probit modeli.

Heritabiliteti za sva tri modela izračunati su prema formuli

h^2 = 4*var_očeva/(var_očeva + var_ostatka)

U slučaju probit i logit modela varijance ostatka su već definirane te iznose 1 za probit model i $\pi^2/3$ za logit model pa formule za heritabilitet glase

h^2 = 4*var_očeva/(var_očeva + 1)

h^2 = 4*var_očeva/(var_očeva + $\pi^2/3$)

7.1. SAS program

```
PROC UNIVARIATE DATA = bikovi;
var y_binar;
run;
```

Objašnjenje: Naredba **proc univariate** poziva proceduru dok **var** definira za koju varijablu (y_binar). Ova procedura pruža opisnu statistiku, omogućava crtanje grafova te omogućava različite statističke metode koje se kasnije mogu upotrijebiti za analizu raspodjele neke varijable.

```
proc mixed data=bikovi covtest;
class bik kcer farm;
model y_binar = farm;
random intercept / subject = bik;
LSMEANS farm /  adjust=tukey ;
run;
```

Objašnjenje: Naredba **proc mixed** poziva proceduru, **class** naređuju proceduri da varijable bik, kcer i farm uzima kao kategoričke varijable. Naredba **model** definira zavisne i nezavisne varijable. Označava y_binar kao zavisnu varijablu, a farm kao nezavisnu. **Random** definira regresijske koeficijente kao slučajne varijable, a **lsmeans** računa prosjeke najmanjih kvadrata u ovome slučaju za varijablu farm. **Covtest** omogućava dobivanje stastističkih zaključaka o parametrima kovarijance.

```
proc mixed data=bikovi covtest;
class bik kcer farm;
model y_binar = farm;
random intercept / subject = bik;
LSMEANS farm /  adjust=tukey ;
run;
```

Objašnjenje: Naredba **proc glimmix** poziva proceduru, **class** naređuju proceduri da varijablu bik uzima kao kategoričku varijablu. Naredba **model** definira zavisne i nezavisne varijable. Označava y_binar kao zavisnu varijablu, a farm kao nezavisnu. **Dist** definira raspodjelu kao binarnu, a **link** označava „link“ funkciju, tj. pretpostavljenu vezu između prosjeka zavisne varijable i linearne kombinacije nezavisnih varijabli. **Random** definira regresijske koeficijente kao slučajne varijable. Kod GLIMMIX procedure residual (ostatak) mora biti posebno naveden dok ga procedura MIXED automatski radi.

```
proc glimmix data=bikovi;
class bik;
model y_binar(event="1") = / dist=binary link=logit;
random intercept / subject=bik;
random _residual_;
run;
```

Objašnjenje: Kod radi isto kao i prijašnji samo je umjesto probit „link“ funkcije stavljena logit „link“ funkcija.

8. Rezultati i rasprava

U SAS 9.3 programu izračunati su prosjek, varijanca i standardna devijacija za 6 različitih varijabli od kojih je za ovaj rad najbitnija binarna (y_binar). Kao što je vidljivo iz Tablice 2 za p0 = 0,2 njezin prosjek je iznosio 0,27005, varijanca 0,197125, a standardna devijacija 0,4439876 dok je za p0 = 0,5 prosjek bio 0,49363, varijanca 0,249962, a standardna devijacija 0,4999619.

Tablica 2. Opisna statistika za varijable y0, y_binar, s, e i p

p0 = 0,2				
Varijabla	**Prosjek**	**Varijanca**	**Standardna devijacija**	**N**
y	-0,931948	1,1866637	1,0893409	100000
y_binar	**0,27005**	**0,197125**	**0,4439876**	**100000**
y0	-0,90906	3,08E-20	1,75E-07	100000
s	0,0011952	0,1744203	0,4176366	100000
e	0,0014739	0,9979476	0,9989733	100000
p	0,2717947	0,0623905	0,2497809	100000
p0 = 0,5				
Varijabla	**Prosjek**	**Varijanca**	**Standardna devijacija**	**N**
y	-0,022888	1,186664	1,0893409	100000
y_binar	**0,49363**	**0,249962**	**0,4999619**	**100000**
y0	-1,24E-12	1,33E-53	3,64E-24	100000
s	0,0011952	0,17442	0,4176366	100000
e	0,0014739	0,997948	0,9989733	100000
p	0,4941938	0,084085	0,2899744	100000

8.1. SAS ispis za PROC UNIVARIATE

Za varijablu y_binar korištena je procedura UNIVARIATE. Figura 2 prikazuje da njezina asimetrija (skewness) za p0 = 0,2 iznosi 1,0358605, a za p0 = 0.5 iznosi 0,0254825. To je zapravo vrijednost koja označava stupanj smjera asimetrije. Kada bi se radilo o simetričnoj raspodjeli kao što je normalna raspodjela asimetrija bi iznosila 0. Iako je asimetrija binarne varijable za p0 = 0,5 relativno blizu 0 treba uzeti u obzir i kurtosis. Kurtosis predstavlja mjeru težine repova raspodjele te je za ovu varijablu negativan i iznosi -0,9270115 za p0 = 0,2 i -1,9993906 za p0 = 0,5 što znači da su repovi „lakši“ nego što bi bili da se radi o normalnoj raspodjeli. Dakle, raspodjela bi bila normalna samo ako bi i asimetrija i kurtosis iznosili 0.

p0 = 0,2

Moments			
N	100000	Sum Weights	100000
Mean	0.27005	Sum Observations	27005
Std Deviation	0.44398758	Variance	0.19712497
Skewness	1.03586051	Kurtosis	-0.9270115
Uncorrected SS	27005	Corrected SS	19712.2998
Coeff Variation	164.409397	Std Error Mean	0.00140401

p0 = 0,5

Moments			
N	100000	Sum Weights	100000
Mean	0.49363	Sum Observations	49363
Std Deviation	0.49996192	Variance	0.24996192
Skewness	0.02548245	Kurtosis	-1.9993906
Uncorrected SS	49363	Corrected SS	24995.9423
Coeff Variation	101.282726	Std Error Mean	0.00158102

Figura 2. Statistički sažetak raspodjele

8.2. SAS ispis za PROC MIXED

Figura 3 prikazuje ukupan broj opažanja koja su očitana i korištena i kao što je vidljivo s figure sva opažanja su pročitana te iznose 100000.

Number of Observations	
Number of Observations Read	100000
Number of Observations Used	100000
Number of Observations Not Used	0

Figura 3. Broj opažanja koja su pročitana i korištena

Nadalje, izračunate su i procjene parametara kovarijance (engl., covariance parameter estimates) za p0 = 0,2 i p0 = 0,5 što prikazuje Figura 4. Kao što je vidljivo Z vrijednosti su iste u oba slučaja, a dobivene su tako što je procijenjena kovarijanca podijeljena s odgovarajućom standardnom greškom.

p0 = 0,2

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	bik	0.008804	0.000478	18.41	<.0001
Residual		0.1880	0.000845	222.48	<.0001

p0 = 0,5

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	bik	0.01233	0.000658	18.74	<.0001
Residual		0.2372	0.001066	222.48	<.0001

Figura 4. Procjene parametara kovarijance (analiza kovarijance)

Kao što je ranije spomenuto naredbom lsmeans dobiveni su prosjeci najmanjih kvadrata te razlika između istih za varijablu farm. Figura 5 prikazuje navedeno.

p0 = 0,2

Least Squares Means						
Effect	farm	Estimate	Standard Error	DF	t Value	Pr > |t|
farm	1	0.3002	0.004053	99E3	74.06	<.0001
farm	2	0.2556	0.003543	99E3	72.14	<.0001
farm	3	0.2692	0.004053	99E3	66.42	<.0001

Differences of Least Squares Means									
Effect	farm	_farm	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
farm	1	2	0.04456	0.003378	99E3	13.19	<.0001	Tukey-Kramer	<.0001
farm	1	3	0.03095	0.003909	99E3	7.92	<.0001	Tukey-Kramer	<.0001
farm	2	3	-0.01360	0.003378	99E3	-4.03	<.0001	Tukey-Kramer	0.0002

p0 = 0,5

Least Squares Means						
Effect	farm	Estimate	Standard Error	DF	t Value	Pr > |t|
farm	1	0.5253	0.004685	99E3	112.12	<.0001
farm	2	0.4759	0.004130	99E3	115.21	<.0001
farm	3	0.4980	0.004685	99E3	106.30	<.0001

Differences of Least Squares Means									
Effect	farm	_farm	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
farm	1	2	0.04943	0.003795	99E3	13.03	<.0001	Tukey-Kramer	<.0001
farm	1	3	0.02725	0.004391	99E3	6.20	<.0001	Tukey-Kramer	<.0001
farm	2	3	-0.02219	0.003795	99E3	-5.85	<.0001	Tukey-Kramer	<.0001

Figura 5. Prosjeci najmanjih kvadrata i razlike između njih

Prikazano je koliko ima farmi, koliki su njihovi procjenjeni prosjeci najmanjih kvadrata te standardne greške. Može se uočiti da postoje i razlike između farmi. Kod seta podataka gdje p0 iznosi 0,2 može se uočiti da su razlike između farme 2 i 3 manje od razlika između farmi 1 i 2 te 1 i 3. Za set podataka gdje p0 iznosi 0,5 uočljivo je da je razlika između farmi 1 i 2 veća od razlike između farmi 1 i 3 te 2 i 3.

Nadalje, izračunati su heritabiliteti za oba seta podataka prema formuli:

h^2 = 4*var_očeva/(var_očeva + var_ostatka)

Uzeti su podatci dobiveni putem procedure MIXED te je dobiveno sljedeće:

- p0 = 0, 2 h^2 = 4*0,008804/(0,008804 + 0,1880) = **0,179**
- p0 = 0,5 h^2 = 4*0,01233/(0,01233 + 0,2372) = **0,198**

Može se uočiti da razlike između heritabiliteta nisu velike.

8.3. SAS ispis za PROC GLIMMIX – „LINK“ PROBIT

Figura 6 prikazuje kolika je vjerojatnost da binarna varijabla poprimi vrijednost 1. S obzirom na zadane p vrijednosti dobiveni su odgovarajući rezultati: za p0 = 0,2 od ukupno 100000 opažanja, 27005 ih je imalo vrijednost 1 dok je za p0 = 0,5 od ukupno 100000 opažanja njih 49363 imalo vrijednost 1.

p0 = 0,2

Response Profile		
Ordered Value	y_binar	Total Frequency
1	0	72995
2	1	27005

The GLIMMIX procedure is modeling the probability that y_binar='1'.

p0 = 0,5

Response Profile		
Ordered Value	y_binar	Total Frequency
1	0	50637
2	1	49363

The GLIMMIX procedure is modeling the probability that y_binar='1'.

Figura 6. Raspodjela frekvencije zavisne varijable (ukupan broj opažanja = 100000)

p0 = 0,2

Fit Statistics	
-2 Res Log Pseudo-Likelihood	346544.6
Generalized Chi-Square	98279.04
Gener. Chi-Square / DF	0.98

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	bik	0.08504	0.004676
Residual (VC)		0.9828	0.004418

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
farm	1	98999	61.95	<.0001

p0 = 0,5

Fit Statistics	
-2 Res Log Pseudo-Likelihood	332413.3
Generalized Chi-Square	99048.29
Gener. Chi-Square / DF	0.99

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	bik	0.08330	0.004478
Residual (VC)		0.9905	0.004452

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
farm	1	98999	38.89	<.0001

Figura 7. Statistika procijenjenog modela, procjene parametara kovarijance i test fiksnih utjecaja (1)

Tip III test fiksnih utjecaja govori o značajnosti fiksnih utjecaja (farme) u modelu. Iako postoje razlike između F vrijednosti za setove podataka, u oba slučaja utjecaj farme je visoko značajan (Figura 7).

Kao i kod procedure MIXED izračunati su heritabiliteti s time da se formula razlikuje jer je za probit model varijanca ostatka već determinirana i iznosi 1. Iz podataka koji prikazuju procjene parametara kovarijance vidljivo je da rezidaul iznosi 0,9829 odnosno 0,9905.

On prikazuje dodatnu varijabilnost (uz onu teoretsku) ako je ima. S obzirom da u ovome slučaju iznosi skoro 1 nema potrebe za dodavanjem njegove vrijednosti u formulu heritabiliteta. U slučaju kada bi vrijednost reziduala bila manja od 1 tada bi nju množili s teoretskom varijabilnošću. Imajući sve navedeno u vidu dobiveni su sljedeći heritabiliteti:

- p0 = 0, 2 $h^2 = 4*0,08504/(0,08504 + 1) =$ **0,313**
- p0 = 0,5 $h^2 = 4*0,08330/(0,08330 + 1) =$ **0,333**

Ponovno je vidljivo da razlike između seta podatka nisu velike, ali su vrijednosti heritabiliteta veće od onih dobivenih procedurom MIXED.

8.4. SAS ispis za PROC GLIMMIX – „LINK“ LOGIT

p0 = 0,2

Fit Statistics	
-2 Res Log Pseudo-Likelihood	451859.2
Generalized Chi-Square	98058.45
Gener. Chi-Square / DF	0.98

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	bik	0.2406	0.01339
Residual (VC)		0.9806	0.004408

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
farm	1	98999	63.45	<.0001

p0 = 0,5

Fit Statistics	
-2 Res Log Pseudo-Likelihood	332413.3
Generalized Chi-Square	99048.29
Gener. Chi-Square / DF	0.99

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	bik	0.08330	0.004478
Residual (VC)		0.9905	0.004452

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
farm	1	98999	38.89	<.0001

Figura 8. Statistika procijenjenog modela, procjene parametara kovarijance i test fiksnih utjecaja (2)

Gledajući procjene parametara kovarijance prikazane na Figuri 8 vidljivo je da su kod ovoga modela vrijednosti znatno veće nego kod probit modela. Poznato je da se logit i probit modeli razlikuju u raspodjeli greške; kod logit modela greške imaju standardnu logističku raspodjelu, a kod probita standardnu normalnu raspodjelu. Kod usporedbe prametara koji se razlikuju u raspodjeli i u slučaju kada se zadane vjerojatnosti kreću između 0,1 i 0,9 kao što je i ovdje slučaj tada bi parametri logit modela trebali biti oko $\pi/\sqrt{3}$ veći od parametara probit modela (www.okstate.edu/sas/v8/saspdf/stat/chap54.pdf). Nadalje, F vrijednosti su slične kao kod probit modela te opet pokazuju da je utjecaj farme visoko značajan.

I kod ovog modela su izračunati heritabiliteti. Kao i za probit model varijanca greške je već definirana i iznosi $\pi^2/3$. Rezidual je ponovno blizu vrijednosti 1 te se ne dodaje u formulu. Dobiveni heritabiliteti su:

- p0 = 0,2 $h^2 = 4*0{,}2406/(0{,}2406 + (3{,}14^2/3))$ = **0,273**
- p0 = 0,5 $h^2 = 4*0{,}2143/(0{,}2143 + (3{,}14^2/3))$ = **0,245**

Kao i kod prethodnih modela razlike heritabiliteta između setova podataka su male međutim razlike između modela postoje.

Kadarmideen i sur. (2000) su u svome radu uspoređivali linearni model i probit model u analizi genetskih parametara za bolest, plodnost i proizvodnju mlijeka kod mliječnih goveda. Vrijednosti heritabiliteta su bile veće kod probit modela nego kod linearnog modela što se poklapa s vrijednostima dobivenim u ovome radu. Oni navode da se obično kod probit modela očekuju veće vrijednosti s obzirom da se heritabiliteti procjenjuju na različitim ljestvicama. Međutim, Koeck i sur. (2010) koji su također uspoređivali ove modele, došli su do zaključka da iako su vrijednosti heritabiliteta manje kod linearnog modela, kada ga se transformira na pretpostavljenu, osnovnu ljestvicu vrijednosti se približe probit i logit modelu.

9. Zaključak

- Ako je zavisna varijabla binarna za njezinu analizu bolje je koristiti probit i logit modele nego obične linearne modele
- Pokazalo se da su probit i logit modeli teoretski bolji i dobro prihvaćeni za procjenu komponenti varijance kada se radi o binarnim podatcima
- U većini istraživanja kao i u ovome radu razlike između vrijednosti dobivene probit i logit modelima postoje, ali nisu značajno velike

Popis literature

Agresti A. (2002) Categorical Data Analysis. John Wiley & Sons, Inc. New York.

Aldrich J. H., Nelson F. D. (1984) Linear probability, logit, and probit models (Vol. 45). Sage.

Austin M. P. (1987) Models for the analysis of species response to environmental gradients. Vegetatio 69: 35 – 45.

Chambers E. A., Cox D. R. (1967) Discrimination between alternative binary response models. Biometrika 54: 573–578.

Davidson R., MacKinnon J. G. (1993) Estimation and Inference in Econometrics. New York: Oxford.

Fahrmeir L., Tutz G. (2001) Multivariate Statistical Modelling Based on Generalized Linear Models (2nd ed.). New York: Springer.

Gianola D. (1982) Theory and analysis of threshold characters. J. Anim. Sci. 54:1079.

Gianola D., Foulley J. L. (1983) Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15: 201.

Guerra J. L. L. (2004) Statistical models and genetic evaluation of binomial traits (Doctoral dissertation, Louisiana State University).

Guisan A., Edwards T. C., Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling, 157(2), 89-100.

Hahn E. D., Soyer R. (2005) Probit and logit models: Differences in the multivariate realm. Submitted to The Journal of the Royal Statistical Society, Series B.

Hastie T. J., Tibshirani R. J. (1990) Generalized Additive Models. Chapman & Hall.

Henderson C. R., Kempthorne O., Searle S. R., von Krosigk C. M. (1959) "The Estimation of Environmental and Genetic Trends from Records Subject to Culling". Biometrics. International Biometric Society. **15** (2): 192–218.

Heringstad B., Rekaya R., Gianola D., Klemetsdal G., Weigel K. A. (2003) Genetic change for clinical mastitis in Norwegian cattle: a threshold model analysis. J. Dairy Sci. 86:369.

Horowitz J. L., Savin N. E. (2001) Binary response models: Logits, probits and semiparametrics. Journal of Economic Perspectives, 15(4): 43-56.

Kadarmideen H. N., Thompson R., Simm G. (2000) Linear and threshold model genetic parameters for disease, fertility and milk production in dairy cattle. *BSAS OCCASIONAL PUBLICATION*, 83-84.

Kapš M., Lamberson W. (2009) Biostatistics for animal science, 2nd edition. CABI, Wallingford, UK.

Koch C. G., Carr G. J., Strokes M. E., Uryniak T. J. (1990) Categorical Data Analysis, in Statistical Methodology in Pharmaceutical Sciences. Ed. D. A. Berry. New York: Marcel Dekkel Inc., 391

Koeck A., Heringstad B., Egger-Danner C., Fuerst C., Fuerst-Waltl B. (2010) Comparison of different models for genetic analysis of clinical mastitis in Austrian Fleckvieh dual-purpose cows. *Journal of dairy science*, *93*(9), 4351-4358.

Long J. S. (1997) Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.

Maddala G. S. (1983) Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

McCullagh P., Nelder J. A. (1989) Generalized Linear Models. Chapman and Hall, London.

Menard S. (2002) Applied logistic regression analysis (Vol. 106). Sage.

Nelder J. A., Wedderburn R. W. M. (1972) Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135 (3): 370-384.

Park H. M. (2009) Regression Models for Binary Dependent Variables Using Stata, SAS, R, LIMDEP, and SPSS. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

Phocas F., Laloe D. (2003) Evaluation models and genetic parameters for calving difficulty in beef cattle. J. Anim. Sci. 81:933.

Powers D. A., Xie Y. (2000) Statistical Methods for Categorical Data Analysis. San Diego: Academic Press.

Ramirez-Valverde R., Miztal I., Bertrand J. K. (2001) Comparison of threshold vs linear and animal vs sire models for predicting direct and maternal genetic effects on calving difficulty in beef cattle. J. Anim. Sci. 79:333.

SAS Institute Inc. 2011.SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

SAS: Business Analytics and Business Intelligence Software. http://support.sas.com/documentation/cdl/en/basess/58133/HTML/default/viewer.htm#a002645411.htm (25.8.2015.)

Söderbom M. (2009) Applied Econometrics. Lecture 10: Binary Choice Models. University of Gothenburg.

Thompson R. (1979) Sire evaluation. Biometrics. 35:111.

www.okstate.edu/sas/v8/saspdf/stat/chap54.pdf (24.9.2016.)

Popis tablica

Popis figura

Životopis

Sara Matanović rođena je 21.5.1991. u Zagrebu gdje i sada živi sa svojim roditeljima, majkom Tatjanom i ocem Željkom. Školovanje je započela 1998. godine u osnovnoj školi Većeslav Holjevac te je kroz sve razrede bila odlična učenica. Nakon završene osnovne škole upisala je IV. gimnaziju (jezičnu) gdje je pohađala dvojezičnu nastavu (hrvatski i francuski). Tijekom osnovnoškolskog i srednješkolskog obrazovanja učila je engleski i francuski jezik, francuski u sklopu nastave, a engleski u Školi stranih jezika Stara Vlaška. Nakon položene državne mature upisala je Agronomski fakultet Sveučilišta u Zagrebu, preddiplomski studij Animalne znanosti. U roku je izvršila sve obveze i položila ispite sa konačnim prosjekom veoma dobar (4,2) te stekla zvanje prvostupnice inženjerke animalnih znanosti, akademske godine 2012/2013. Nakon završenog preddiplomskog studija, akademske godine 2013/2014 upisala je diplomski studij Genetika i oplemenjivanje životinja na Agronomskom fakultetu Sveučilišta u Zagrebu. U roku je položila sve ispite te je akademske godine 2015/2016 upisala apsolventsku godinu. Sudjelovala je u istraživanju vezanom uz stručni projekt na temelju kojeg je objavljen članak „Zamjena antibiotika biološki djelatnim tvarima u hranidbi peradi“. Navedeni članak izašao je u časopisu „Krmiva, Vol.55 No.1“. Tijekom prve godine diplomskog studija počela je raditi u Školi za obuku pasa Vindor gdje radi i danas. Sudjeluje na tečajevima osnovnog odgoja i socijalizacije pasa raznih uzrasta kao pomagač u procesu te na stručnim uvodnim predavanjima za vlasnike pasa koji će polaziti tečajeve. Pomaže kod pripreme seminara o ponašanju pasa i njihovom odgoju i socijalizaciji te kod organizacije svakodnevnog rada i planiranja u školi. Također sudjeluje u ostalim aktivnostim škole (snimanja reklama i filmova, promocija škole u različitim centrima itd.).