Univerza v Mariboru
Fakulteta za logistiko
Mariborska cesta 7
3000 Celje, Slovenija

UNIVERZA V MARIBORU
FAKULTETA ZA LOGISTIKO

Statistična analiza dokumentov Incoterm s pomočjo statističnega modeliranja tipa regresijski LOGIT in ugotavljanje ključnih preferenc

Delno poročilo študije

izr. prof. dr. Dejan Dragan, doc. dr. Marjan Strnad, prof. dr. Bojan Rosi (dekan FL UM)

Fakulteta za logistiko UM

Celje, 06.02.2019

Dokument: **Statistična analiza dokumentov Incoterm s pomočjo statističnega modeliranja tipa regresijski LOGIT in ugotavljanje ključnih preferenc**

Avtorji: Dejan Dragan, Marjan Strnad, Bojan Rosi

Izdelano: Leto 2019

KAZALO

1. UVOD IN OZADJE RAZISKAVE

Raziskava obravnava Klavzule lncoterms, ki natančno definirajo odgovornosti in dolžnosti kupcev in prodajalcev v mednarodni trgovini pri prevozu blaga na podlagi prodajne pogodbe. Prvič so se pojavile že leta 1936, njihovo tolmačenje pa je uvedla Mednarodna trgovinska zbornica v Parizu.

Klavzule vsebujejo več neodvisnih spremenljivk (regresorjev), kot npr. 'razdalja', 'Zabojnik', 'Šifra_lege_kraja', 'Vrsta posla', 'promet na meji', 'promet v notranjosti' , 'Carinski postopek', 'število EUL', 'število postavk', 'Statistična vrednost blaga', 'Neto masa', ter 'vrednost na kg'. Ključni podatek Incoterma pa je njegova koda (odvisna spremenljivka y).

Glavni namen raziskave je bil statistično ugotoviti, katere neodvisne spremenljivke signifikantno vplivajo na vrednost binomsko porazdeljene naključne spremenljivke y. Analizirani so bili podatki **uvoza in izvoza za Slovenijo za leto 2016**. Kot statistični model je bil izbran **logistični regresijski model logit**.

Ker je bil v prvotni obliki y porazdeljen po Bernoullijevi porazdelitvi (0 ali 1 – pripadnost enemu ali drugemu razredu), smo morali pri izbrani kombinaciji neodvisnih spremenljivk prešteti število »ugodnih« in »neugodnih« izidov spremenljivke y. Pri tem smo tudi neodvisne spremenljivke 'Statistična vrednost blaga', 'Neto masa', ter 'vrednost na kg' razvrstili v razrede, da smo sploh lahko prešteli število »ugodnih« in »neugodnih« izidov spremenljivke y.

Najzahtevnejši del raziskave je predstavljalo iskanje optimalne kombinacije regresorjev, ki statistično signifikantno vpliva na vrednost y, pri čemer morajo biti hkrati izpolnjeni vsi statistični testi in pogoji glede kvalitete prileganja izhoda modela realnim podatkom, kot npr. Waldova signifikantnost regresorjev, Devianca manjša od kritične vrednosti, McFaddenov R2 in ostali pseudo-R2 morajo biti dovolj veliki, ocenjeni logit se mora dobro prilegati dejanskemu (iz podatkov), ocenjene verjetnosti, da y zavzame določeno vrednost, pa dejanskim (na osnovi podatkov).

Za potrebe iskanja optimalne kombinacije regresorjev je bila uvedena Monte Carlo (MC) simulacija, ki je za vsako izbrano število regresorjev (med tri in šest – pri večjem številu so se pojavile težave s slabo pogojenostjo matrik za ocenjevanje regresorjev), preigrala 100.000 možnih scenarijev, tako za podatke uvoza kot tudi izvoza. V MC postopku so se poleg omenjenih statističnih testov izračunavali tudi številni drugi testi, kot npr. vrednost

normaliziranega korelacijskega koeficienta med ocenjenimi in dejanskimi vrednostmi logita, oz. ocenjenimi in dejanskimi verjetnostmi, mere povezane s Pearsonovim residualom, itn.

Po koncu postopka in celotne raziskave se je izkazalo, da **pri uvozu** na vrednosti y najbolj (signifikantno!) vplivajo naslednji regresorji: **'konstanta', 'Zabojnik', 'Šifra_lege_kraja', 'Statistična vrednost blaga (v razredih)' , ter 'Neto masa (v razredih)'**. **Pri izvozu** pa najbolj vplivajo: **'konstanta', 'Zabojnik', 'Šifra_lege_kraja', ter 'Statistična vrednost blaga (v razredih)'**.

Pomembno je še omeniti, da smo zaradi preprečevanja nepotrebne redundantnosti izločili dogodke, ki so se zgodili le enkrat. Kot se izkaže, se tako pri uvozu kot tudi izvozu, poleg vseh izpolnjenih statističnih testov, model dobro prilega danim podatkom (logit, verjetnosti, dovolj velik normaliziran korelacijski koeficient, »število zadetkov« modela).

Glavna uporabna vrednost te raziskave je, da bi na osnovi statističnega modeliranja bolje razumeli subjektivne preference, ki so v letu 2016 vodile odločevalce k izbiri določenega Incoterma, ter preučili, kateri regresorji pri tovrstnih odločitvah sploh igrajo pomembno vlogo. Gre za eno prvih tovrstnih raziskav v svetu, kar je gotovo doprinos raziskave.

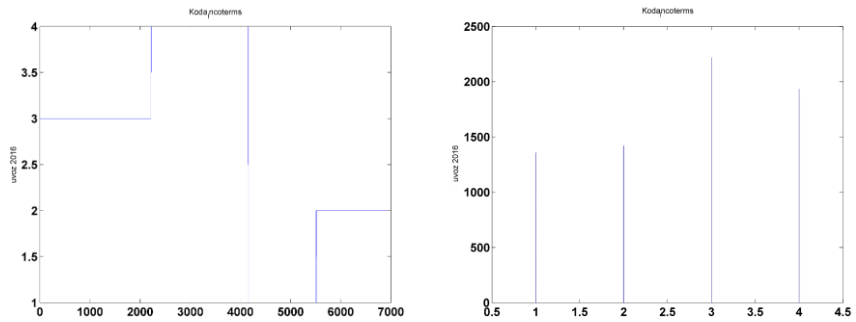
Proces modeliranja je bil implementiran v programskem orodju Matlab, pri čemer so bili uporabljeni tudi Statistics and Machine Learning toolbox, ter Econometrics Toolbox. Koncept in rezultati modeliranja, ki so bili pridobljeni s pomočjo razvite programske opreme, bodo predstavljeni tudi v okviru končnega poročila študije. Poleg bodo rezultati, pridobljeni s pomočjo razvite programske opreme, predstavljeni tudi v izvirnem znanstvenem JCR ali SNIP članku.

2. PREDSTAVITEV PODATKOV IN DEFINICIJA SPREMENLJIVK

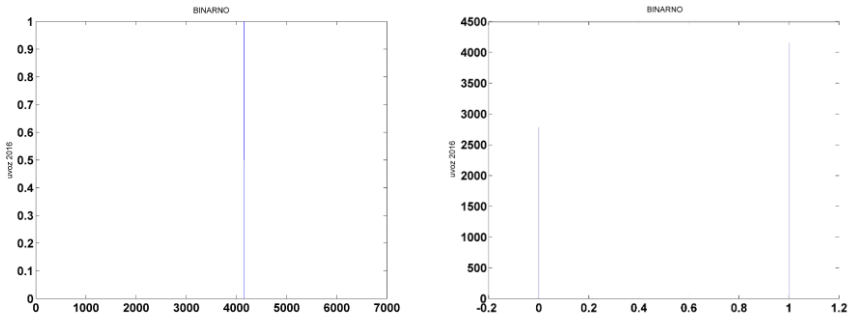
V tem poglavju bomo na kratko predstavili grafični prikaz danih podatkov Incotermov, njihove verjetnostne porazdelitve, ter ustrezno priredili simbole posameznim spremenljivkam.

2.1 Predstavitev podatkov za uvoz 2016

Izhodna spremenljivka y v originalni obliki (delitev za 4 razrede) je prikazana na sliki 1, njena binarna inačica pa na sliki 2. Na desni strani obeh slik je prikazana tudi ustrezna porazdelitev.

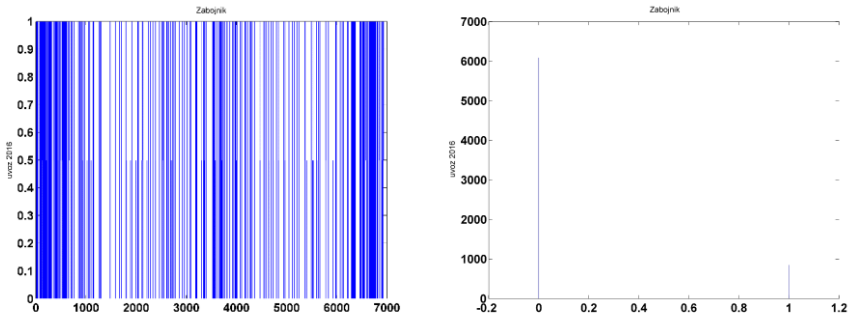


Slika 1: Izhodna spremenljivka y v originalni obliki (delitev za 4 razrede) in njena porazdelitev.



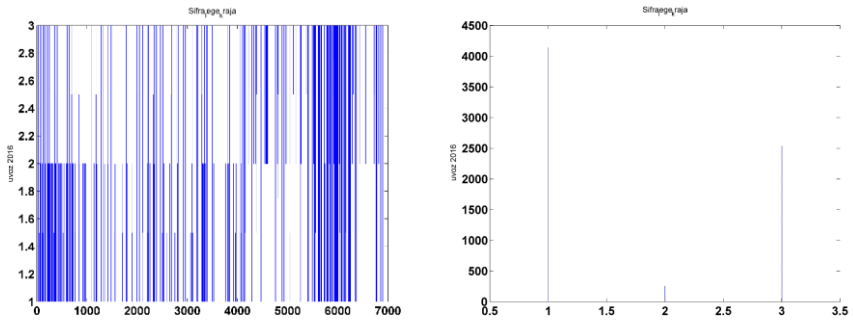
Slika 2: Izhodna spremenljivka y v binarni obliki in njena porazdelitev.

V nadaljevanju predstavimo vhodne spremenljivke. Zaradi večje preglednosti je zgornja meja oordinatne osi povsod omejena (na prikazih slik). Vhodna spremenljivka $X_1 = X_{ZABOJNIK}$ in njena porazdelitev je predstavljena na sliki 3.



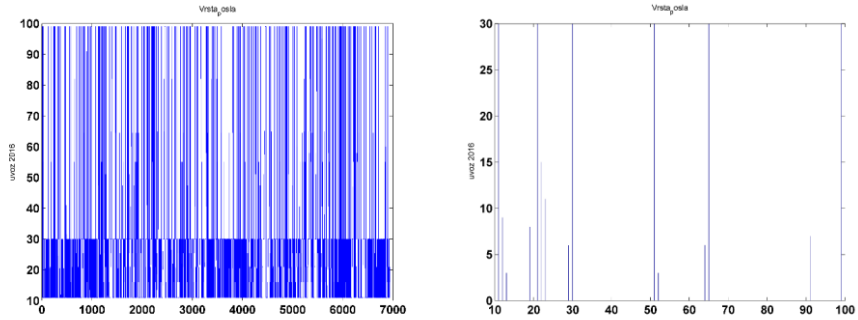
Slika 3: Vhodna spremenljivka $X_1 = X_{ZABOJNIK}$ in njena porazdelitev.

Vhodna spremenljivka $X_2 = X_{SIFRA\_LEGE\_KRAJA}$ in njena porazdelitev je predstavljena na sliki 4.



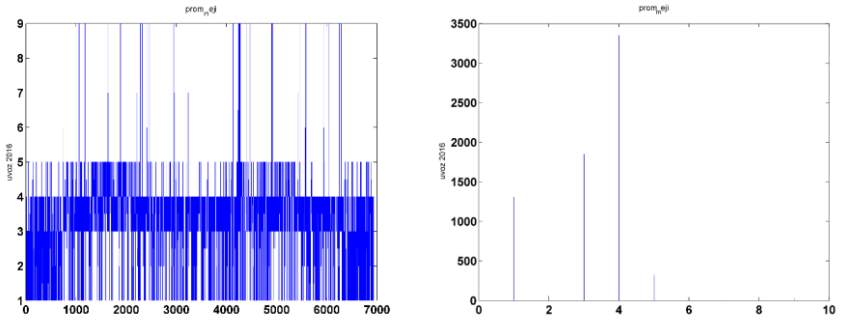
Slika 4: Vhodna spremenljivka $X_2 = X_{SIFRA\_LEGE\_KRAJA}$ in njena porazdelitev.

Vhodna spremenljivka $X_3 = X_{VRSTA\_POSLA}$ in njena porazdelitev je predstavljena na sliki 5.



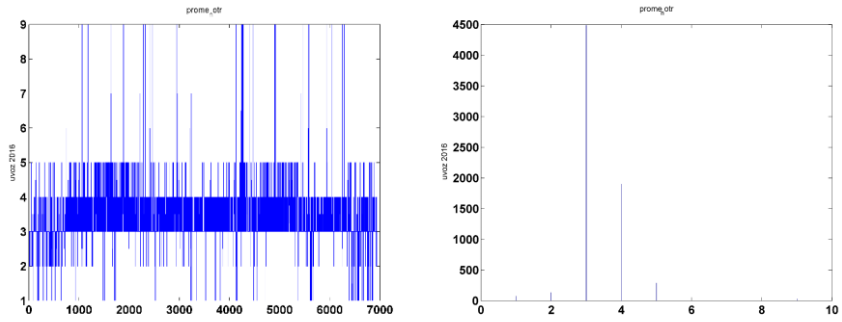
Slika 5: Vhodna spremenljivka $X_3 = X_{VRSTA\_POSLA}$ in njena porazdelitev.

Vhodna spremenljivka $X_4 = X_{PROM\_NA\_MEJI}$ in njena porazdelitev je predstavljena na sliki 6.



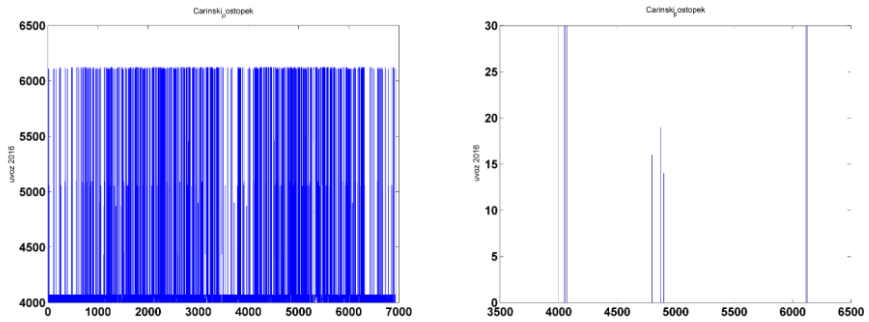
Slika 6: Vhodna spremenljivka $X_4 = X_{PROM\_NA\_MEJI}$ in njena porazdelitev.

Vhodna spremenljivka $X_5 = X_{PROM\_V\_NOTR}$ in njena porazdelitev je predstavljena na sliki 7.



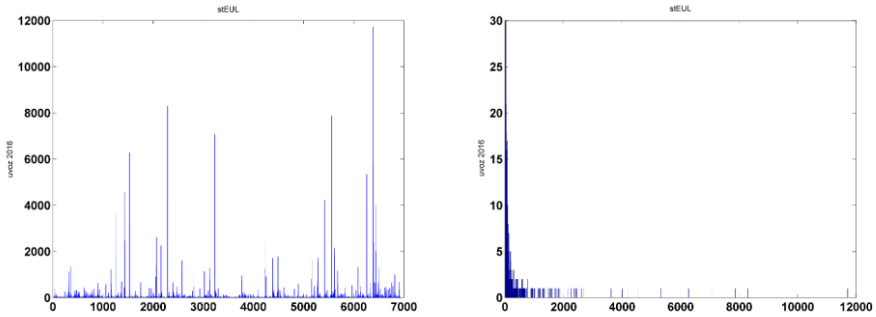
Slika 7: Vhodna spremenljivka $X_5 = X_{PROM\_V\_NOTR}$ in njena porazdelitev.

Vhodna spremenljivka $X_6 = X_{CARIN\_POST}$ in njena porazdelitev je predstavljena na sliki 8.



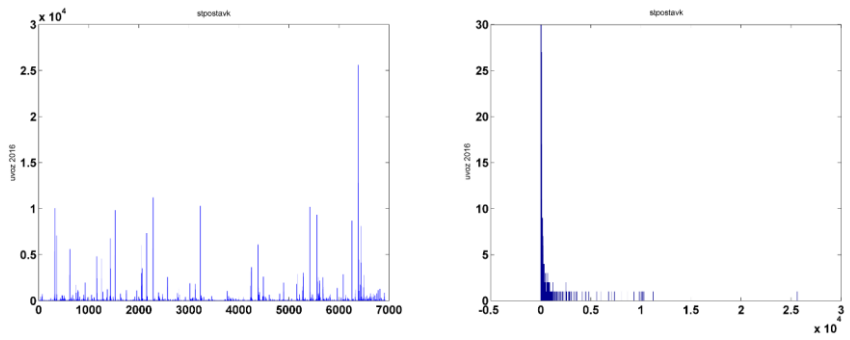
Slika 8: Vhodna spremenljivka $X_6 = X_{CARIN\_POST}$ in njena porazdelitev.

Vhodna spremenljivka $X_7 = X_{ST\_EUL}$ in njena porazdelitev je predstavljena na sliki 9.



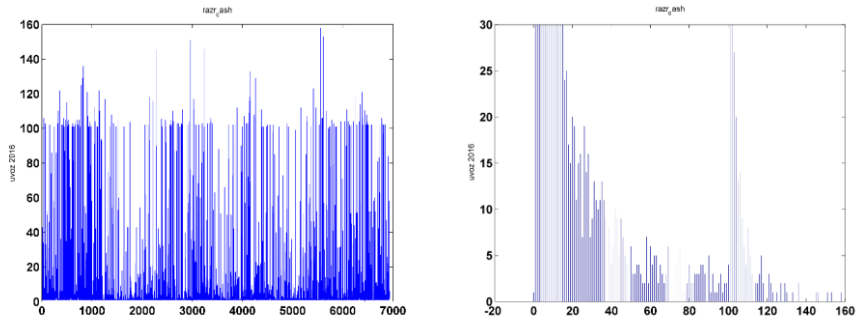
Slika 9: Vhodna spremenljivka $X_7 = X_{ST\_EUL}$ in njena porazdelitev.

Vhodna spremenljivka $X_8 = X_{ST\_POSTAVK}$ in njena porazdelitev je predstavljena na sliki 10.



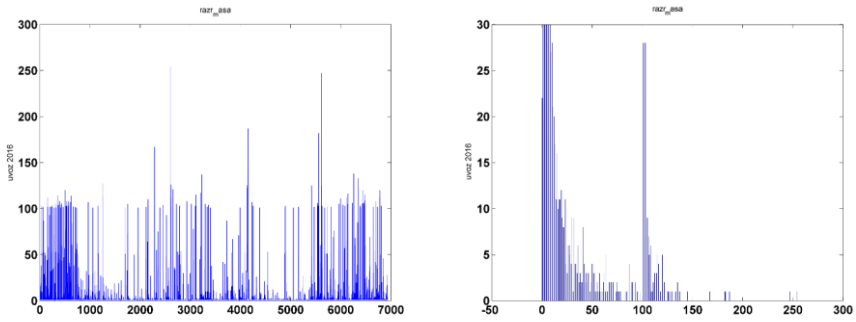
Slika 10: Vhodna spremenljivka $X_8 = X_{ST\_POSTAVK}$ in njena porazdelitev.

Vhodna spremenljivka $X_9 = X_{CASH\_RAZR}$ - Statistična vrednost blaga (v razredih) in njena porazdelitev je predstavljena na sliki 11. Na x osi leve slike 11 je število vzorcev, na y osi pa številka razreda denarne vrednosti. Višje denarne vrednosti imajo seveda višjo številko razreda. Interpretacija desne slike 11 pa je: na x osi je velikostni razred denarja, na oordinati pa frekvenca (število nastopov) vsakega posameznega velikostnega razreda.



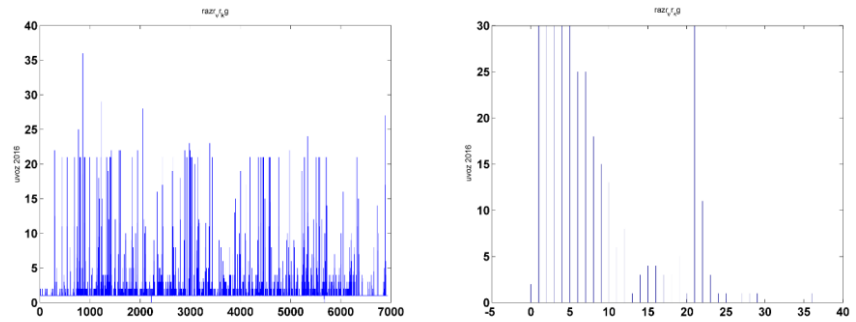
Slika 11: Vhodna spremenljivka $X_9 = X_{CASH\_RAZR}$ in njena porazdelitev.

Vhodna spremenljivka $X_{10} = X_{MASA\_RAZR}$ - Masa blaga (v razredih) in njena porazdelitev je predstavljena na sliki 12.



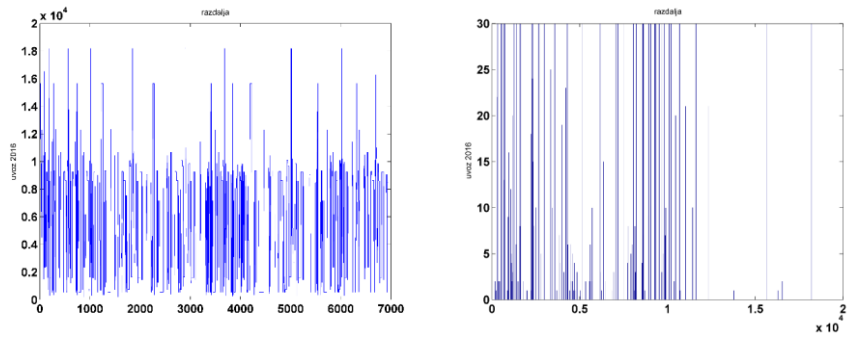
Slika 12: Vhodna spremenljivka $X_{10} = X_{MASA\_RAZR}$ in njena porazdelitev.

Vhodna spremenljivka $X_{11} = X_{VRED\_NA\_KG\_RAZR}$ - Vrednost/kg blaga (v razredih) in njena porazdelitev je predstavljena na sliki 13.



Slika 13: Vhodna spremenljivka $X_{11} = X_{VRED\_NA\_KG\_RAZR}$ in njena porazdelitev.

Kako so vrednosti spremenljivk X_9, X_{10}, X_{11} točno porazdeljene na razrede, je razvidno iz poročila o programski opremi za dotično raziskavo, ter seveda samega programa. Slika 14 pa prikazuje še neodvisno spremenljivko razdalja in njeno porazdelitev, za katero pa se je izkazalo, da ni relevantna pri izgradnji modela.



Slika 14: Vhodna spremenljivka razdalja in njena porazdelitev.

2.2 Predstavitev podatkov za izvoz 2016

Podatkov za izvoz 2016 ne bomo predstavljali, saj so podobne oblike in značilnosti. Seveda so določene razlike, tako v velikosti vzorca, kot tudi amplitudah in frekvencah spremenljivk. Opravka pa imamo seveda s popolnoma enakimi tipi spremenljivk, kot so bile predstavljene v prejšnjem poglavju.

3. MONTE CARLO SIMULACIJA IN KRITERIJI

V splošnem lahko zapišemo naslednjo relacijo, ki povezuje binomski izhod (tip incoterma – 0 ali 1) z vhodnimi neodvisnimi spremenljivkami X_p, $p = 1,2,...,P = 11$:

$$Y_i \in B\left(m_i, p_i\left(X_p\right)\right) \text{ za } i = 1,...,N; p = 1,...,P = 11 \quad (1)$$

torej se njegova vrednost porazdeljuje po binomski porazdelitvi B, kjer je N velikost vzorca. Pri tem lahko tvorimo polno matriko vseh vhodov:

$$\mathbf{X_{polni}} = \left[\mathbf{X_1}, ..., \mathbf{X_{11}}\right] = \begin{bmatrix} X_1(1) & X_2(1) & ... & X_{11}(1) \\ X_1(2) & X_2(2) & ... & X_{11}(2) \\ ... & ... & ... & ... \\ X_1(N) & X_2(N) & ... & X_{11}(N) \end{bmatrix} \quad (2)$$

Seveda se v praksi izkaže, da izhod ni statistično signifikantno odvisen od vseh vhodov, pač pa le nekaterih. Torej je potrebno poiskati takšno kombinacijo vhodov $X_l, l < 6 < p = 11$, pri kateri se bo uporabljeni statistični model, ki povezuje vhode z izhodom, najbolje obnesel. Kot se izkaže, ima lahko model največ 6 regresorjev (vhodov), sicer pride so slabe pogojenosti sistema, postopek iskanja parametrov modela pa divergira.

Za potrebe iskanja optimalne kombinacije regresorjev je bila uvedena Monte Carlo (MC) simulacija, ki je za vsako izbrano število regresorjev (med tri in šest), preigrala 100.000 možnih scenarijev, tako za podatke uvoza kot tudi izvoza. Z drugimi besedami to pomeni, da je bila v vsakem scenariju naključno izbrana drugačna sekvenca regresorjev, v paketih po tri, štiri, pet, ali šest regresorjev. Pri tem se je pri vsakem scenariju testiralo, če so hkrati izpolnjeni vsi statistični testi in pogoji glede kvalitete prileganja izhoda modela realnim podatkom, kot npr.:

- Waldova signifikantnost regresorjev,

- Devianca manjša od kritične vrednosti,
- McFaddenov R2 in ostali pseudo-R2 morajo biti dovolj veliki,
- Ocenjeni logit se mora dobro prilegati dejanskemu (iz podatkov),
- Ocenjene verjetnosti, da y zavzame določeno vrednost, se morajo dobro prilegati dejanskim (na osnovi podatkov),
- Vrednost normaliziranega korelacijskega koeficienta med ocenjenimi in dejanskimi vrednostmi logita mora biti dovolj velika,
- Vrednost normaliziranega korelacijskega koeficienta med ocenjenimi in dejanskimi verjetnostmi mora biti dovolj velika,
- Mere povezane s Pearsonovim residualom morajo biti ustrezne,
- Ostale residualne metrike morajo biti ustrezne v smislu čim manjšega residuala (pogreška modela).

Ko je bil postopek preigravanja scenarijev $j = 1, 2, 3, ..., 100000$ končan, se je tako za uvoz kot tudi izvoz poiskala tista optimalna reducirana matrika vhodov $\mathbf{X} = \mathbf{X}^* = \mathbf{X}\left(j^*\right)_{N \times l}$, $j^* = j_{OPTIM}$, ki je vsebovala takšno kombinacijo *l* regresorjev, pri kateri je statistični model imel najboljše lastnosti.

4. LOGISTIČNA REGRESIJA IN LOGIT MODEL

V tem poglavju si bomo pogledali nekaj kratkih teoretičnih osnov za logit model. Več podrobnosti pa lahko bralec zasledi v viru: Dejan Dragan, 2016: *Logistična regresija s programskim orodjem Matlab.*

4.1 Logit model

Za matriko v izrazu (2) lahko za vzorec velikosti N podamo vrednosti statističnih enot:

$$\mathbf{x}_i^T = \begin{bmatrix} x_1(i) \\ \vdots \\ x_p(i) \end{bmatrix}^T, \quad i = 1, ..., N, \quad \mathbf{x}_i^T = \left[x_1(i), ..., x_p(i) \right] \tag{3}$$

oziroma:

$$\begin{aligned}\mathbf{x}_1^T &= \left[x_1(1),...,x_p(1)\right]\\ \mathbf{x}_2^T &= \left[x_1(2),...,x_p(2)\right]\\ &\vdots\\ \mathbf{x}_N^T &= \left[x_1(N),...,x_p(N)\right]\end{aligned} \tag{4}$$

Za binarno odvisno spremenljivko Y v splošnem velja:

$$\begin{aligned}&Y_1 = y_1, Y_2 = y_2, ..., Y_N = y_N\\ &y_i = \begin{cases}1\\0\end{cases} \text{ za vsak } i = 1,...,N\end{aligned} \tag{5}$$

Ker pa smo pri izbranem vektorju neodvisnih spremenljivk $\mathbf{x}_i^T$ naredili m_i realizacij, pa y_i pomeni število nastopov vrednosti 1. Če predpostavimo, da so naključne spremenljivke Y_i med seboj neodvisne binomske naključne spremenljivke, lahko zapišemo:

$$Y_i \in B(m_i, p_i) \text{ za } i = 1,...,N \tag{6}$$

kjer za prave verjetnosti na osnovi podatkov velja:

$$p_i = p(x_i)_{i=1,...,N} = \frac{e^{b_0+b_1x_1(i)+b_2x_2(i)+...+b_px_p(i)}}{1+e^{b_0+b_1x_1(i)+b_2x_2(i)+...+b_px_p(i)}} = \frac{e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}{1+e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}; \quad \mathbf{B} = \begin{bmatrix}b_0\\b_1\\\vdots\\b_p\end{bmatrix} \tag{7}$$

kjer:

$$\mathbf{x}_{ia}^T = \begin{bmatrix}1 & \mathbf{x}_i^T\end{bmatrix}$$

Za logit funkcijo na osnovi podatkov velja izraz:

(8)

$$\log it\left(p_i\right)=\ln\frac{p_i}{1-p_i}=\ln\frac{\dfrac{e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}{1+e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}}{1-\dfrac{e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}{1+e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}}=\ln\frac{e^{\mathbf{x}_{ia}^T\cdot\mathbf{B}}}{1}=\mathbf{x}_{ia}^T\cdot\mathbf{B}=\underset{i=1,...,N}{b_0+b_1}\,x_1\left(i\right)+...b_p x_p\left(i\right)$$

Vektor regresijskih koeficientov, ki ga je treba oceniti, je:

$$\mathbf{B}=\begin{bmatrix}b_0\\b_1\\\vdots\\b_p\end{bmatrix}\quad\rightarrow\text{ Vektor regresijskih koeficientov}\tag{9}$$

Za ocenjene verjetnosti velja:

$$\hat{p}_i=\frac{e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}}}{1+e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}}};\quad i=1,2,...,N\tag{10}$$

Če tvorimo funkcijo največjega verjetja, po izpeljavi dobimo:

$$\ln L\left(\hat{\mathbf{B}}\right)=C+\sum_{i=1}^{N}\left[y_i\cdot\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}+m_i\ln\left(1-\hat{p}_i\right)\right]=\tilde{L}\left(\hat{\mathbf{B}}\right)\tag{11}$$

Funkcijo največjega verjetja moramo odvajati po parametrih $\hat{b}_0,\hat{b}_1,...,\hat{b}_p$. Izvesti moramo torej naslednjo operacijo:

$$\mathbf{S}\left(\hat{\mathbf{B}}\right)=\begin{bmatrix}S_0\left(\hat{\mathbf{B}}\right)\\\vdots\\S_p\left(\hat{\mathbf{B}}\right)\end{bmatrix}=\begin{bmatrix}\dfrac{\partial\tilde{L}\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_0}\\\dfrac{\partial\tilde{L}\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_1}\\\vdots\\\dfrac{\partial\tilde{L}\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_p}\end{bmatrix}=\frac{\partial}{\partial\hat{\mathbf{B}}}\tilde{L}\left(\hat{\mathbf{B}}\right)=\mathbf{0}\tag{12}$$

Po daljši izpeljavi dobimo za vektor **S**:

$$\mathbf{S}\left(\hat{\mathbf{B}}\right)=\frac{\partial}{\partial\hat{\mathbf{B}}}\tilde{L}\left(\hat{\mathbf{B}}\right)=\sum_{i=1}^{N}\left(y_i\cdot\mathbf{x}_{ia}-m_i\cdot\mathbf{x}_{ia}\cdot\hat{p}_i\right)=\sum_{i=1}^{N}\mathbf{x}_{ia}\cdot\left(y_i-m_i\cdot\hat{p}_i\right) \tag{13}$$

Sistem enačb $\mathbf{S}\left(\hat{\mathbf{B}}\right)=\mathbf{0}$ ponavadi ni rešljiv z analitičnimi postopki. Zato moramo uporabiti postopke numerične matematike za rešitev sistema enačb (13). V ta namen bomo uporabili informacijsko matriko razsežnosti $\left(p+1\right)\cdot\left(p+1\right)$, ki ima obliko:

$$\mathbf{I}\left(\hat{\mathbf{B}}\right)=-\frac{\partial}{\partial\hat{\mathbf{B}}^T}\mathbf{S}\left(\hat{\mathbf{B}}\right)=-\begin{bmatrix}\frac{\partial S_0\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_0} & \cdots & \frac{\partial S_0\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_p}\\ \vdots & \cdots & \vdots\\ \frac{\partial S_p\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_0} & \cdots & \frac{\partial S_p\left(\hat{\mathbf{B}}\right)}{\partial\hat{b}_p}\end{bmatrix} \tag{14}$$

Dokazati se da, da je njena vrednost enaka:

$$\mathbf{I}\left(\hat{\mathbf{B}}\right)=\underbrace{\begin{bmatrix}1 & \ldots & 1\\ x_1(1) & \ldots & x_1(N)\\ \vdots & \ldots & \vdots\\ x_p(1) & & x_p(N)\end{bmatrix}}_{\mathbf{X}^T}\cdot\underbrace{\begin{bmatrix}m_1\cdot\hat{p}_1\cdot(1-\hat{p}_1) & \emptyset\\ \emptyset & \ldots\\ & m_N\cdot\hat{p}_N\cdot(1-\hat{p}_N)\end{bmatrix}}_{\mathbf{W}}\cdot\underbrace{\begin{bmatrix}1 & x_1(1) & \cdots & x_p(1)\\ \ldots & \vdots & \ddots & \vdots\\ 1 & x_1(N) & \cdots & x_p(N)\end{bmatrix}}_{\mathbf{X}} \tag{15}$$

$$\mathbf{I}\left(\hat{\mathbf{B}}\right)=\mathbf{X}^T\cdot\mathbf{W}\cdot\mathbf{X}$$

kjer velja relacija za matriko **X**:

$$\mathbf{X}=\begin{bmatrix}\mathbf{x_{1a}^T}\\ \ldots\\ \mathbf{x_{Na}^T}\end{bmatrix} \tag{16}$$

Tudi vektor **S** se da še nekoliko preoblikovati, pri čemer po krajši izpeljavi dobimo:

$$\mathbf{S}\left(\hat{\mathbf{B}}\right)=\mathbf{X^T}\cdot\left(\mathbf{Y}-\mathbf{E}\right) \tag{17}$$

kjer je **Y** vektor izmerjenih vrednosti odvisne spremenljivke (izmerjenih frekvenc), **E** je pa vektor ocenjenih vrednosti odvisne spremenljivke (teoretičnih frekvenc) [Jesenko].

Največjo vrednost funkcije največjega verjetja bomo iskali z Newton-Raphsonovo iteracijsko metodo, pri čemer bomo poskusili numerično rešiti sistem enačb $\mathbf{S}\left(\hat{\mathbf{B}}\right)=\mathbf{X}^{\mathbf{T}}\cdot\left(\mathbf{Y}-\mathbf{E}\right)=\mathbf{0}$. Za začetne vrednosti ponavadi izberemo vektor koeficientov 0, to je:

$$\hat{\mathbf{B}}(1)=\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix} \tag{90}$$

Iteracijski postopek pa temelji na naslednjem zaporedju približkov rešitev za vektor ocenjenih koeficientov [Jesenko]:

$$\hat{\mathbf{B}}(k+1)=\hat{\mathbf{B}}(k)+\left(\mathbf{I}\left[\hat{\mathbf{B}}(k)\right]\right)^{-1}\cdot\mathbf{S}\left(\hat{\mathbf{B}}(k)\right),\qquad k=1,2,... \tag{18}$$

Iteracijski postopek izvajamo toliko časa, dokler ni dosežena prepisana toleranca ε in se ocenjeni parametri $\hat{b}_p$ v dveh zaporednih iteracijah skorajda ne razlikujejo več med seboj:

$$\left\|\hat{\mathbf{B}}\left(k_{končni}\right)-\hat{\mathbf{B}}\left(k_{končni-1}\right)\right\|\le\varepsilon \tag{19}$$

Končni rezultat je enak:

$$\hat{\mathbf{B}}_{OPT}=\hat{\mathbf{B}}^{*}=\hat{\mathbf{B}}\left(k_{končni}\right) \tag{20}$$

Ocena kovariančne matrike optimalnih ocenjenih koeficientov $\hat{\mathbf{b}}_0^{*},...,\hat{\mathbf{b}}_P^{*}$ je enaka:

$$\hat{\mathbf{\Sigma}}^{*}=\mathbf{I}^{-1}\left(\hat{\mathbf{B}}^{*}\right)=\begin{bmatrix}S_{11}^{*} & \cdots & S_{1,p+1}^{*}\\ \vdots & \ddots & \vdots\\ S_{p+1,1}^{*} & \cdots & S_{p+1,p+1}^{*}\end{bmatrix} \tag{21}$$

Kvadratni koreni diagonalnih elementov te matrike pa so standardne napake optimalnih ocen koeficientov:

$$\sigma\left(\hat{b}_0^*\right)=\sqrt{S_{11}^*},\quad \sigma\left(\hat{b}_1^*\right)=\sqrt{S_{22}^*},\ ...,\quad \sigma\left(\hat{b}_P^*\right)=\sqrt{S_{p+1,\,p+1}^*} \tag{22}$$

Ocenjene verjetnosti $\hat{p}_i^*$ (točkaste ocene) in ocenjene logit vrednosti so enake:

$$\begin{aligned}\hat{p}_i^* &= \frac{e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}{1+e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}=\frac{e^{\hat{b}_0^*+\hat{b}_1^*x_1(i)+...+\hat{b}_P^*x_P(i)}}{1+e^{\hat{b}_0^*+\hat{b}_1^*x_1(i)+...+\hat{b}_P^*x_P(i)}},\qquad i=1,...,N\\ \log it\left(\hat{p}_i^*\right) &= \ln\frac{\hat{p}_i^*}{1-\hat{p}_i^*}=\ln\left(\frac{\dfrac{e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}{1+e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}}{1-\dfrac{e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}{1+e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}}}\right)=\ln e^{\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*}=\mathbf{x}_{ia}^T\cdot\hat{\mathbf{B}}^*\end{aligned} \tag{23}$$

Ocenjene vrednosti odvisne spremenljivke pri njih pa so enake:

$$\hat{y}_i^*=\hat{p}_i^*\cdot m_i,\qquad i=1,...,N \tag{24}$$

Pri logit modelu so pomembna tudi tki. ODDS razmerja za ocenjene parametre, ki izražajo elastičnost odvisne spremenljivke glede na spremembo neodvisne spremenljivke. Razmerje ODDS ratio potemtakem pomeni, za koliko se procentualno poveča logit funkcije, če katerega izmed regresorjev povečamo za eno enoto.

Denimo neko vhodno spremenljivko $x_p(i)$ povečamo na $x_p(i)+1$. Potem bi za parameter $\hat{b}_p$ veljal naslednji izraz:

$$OR_p^*=e^{\hat{b}_p^*} \tag{25}$$

Sklep bi torej bil: Če torej regresor x_p povečamo za eno enoto, se ODDS RATIO poveča za $e^{\hat{b}_p^*}$ enot, pri čemer gre $p=0,...,P$. Na primer, če bi prišel $e^{\hat{b}_p^*}=1{,}175$, bi to pomenilo, da se je OR_p^* povečal za 17,5%

4.2. Validacija modela

4.2.1. Devianca in Likelihood ratio R2L

Devianca polnega modela logistične regresije je enaka:

$$D = -2\left[\sum_{i=1}^{N}\left[y_i \cdot \ln\frac{\hat{y}_i^{*}}{y_i} + \left(m_i - y_i\right)\ln\frac{\left(m_i - \hat{y}_i^{*}\right)}{m_i - y_i}\right]\right] =$$
$$= 2\left[\sum_{i=1}^{N}\left\{ y_i \cdot \ln\frac{y_i}{\hat{y}_i^{*}} + \left(m_i - y_i\right)\ln\frac{m_i - y_i}{m_i - \hat{y}_i^{*}}\right\}\right] = D_{P_{\text{POLNI}}} \tag{26}$$

kjer je $\hat{y}_i^{*} = m_i \cdot \hat{p}_i^{*}$,

$\hat{p}_i^{*} = \dfrac{e^{\mathbf{x}_{ia}^{T}\cdot\hat{\mathbf{B}}^{*}}}{1+e^{\mathbf{x}_{ia}^{T}\cdot\hat{\mathbf{B}}^{*}}}$,

$i = 1,\ldots,N$

Devianca je vedno nenegativna in je zato lahko tudi neka mera za prilagajanje modela logistične regresije podatkom. Za velike m_i in pri dobrem prilagajanju modela podatkom (ko je majhna devianca) je D približno χ^2 naključna spremenljivka z $N - p - 1$ prostostnimi stopnjami. Iz tega izhaja, da se za velike vrednosti deviance model slabo prilagodi podatkom. Devianco pa pogostokrat uporabljamo tudi za primerjavo različnih modelov logistične regresije z različnim številom parametrov.

Kot se izkaže, velja za velike m_i pri dobrem prilagajanju modela podatkom:

$$D \approx \chi^2\left(N - p - 1\right) \tag{27}$$

Celovit test kakovosti polnega modela pa izvedemo z naslednjo primerjavo:

H_0 : Ocenjeni model se dobro prilagodi podatkom (28)

H_1 : Ocenjeni model se slabo prilagodi podatkom

$$D = D_P \overset{?}{>} \chi_{krit}^{\,2}\left(\alpha, N - p - 1\right)$$

Če je $D_P < D_{Pkrit} = \chi_{krit}^{\ 2}(\alpha, N-p-1)$, ne zavrnemo ničelne hipoteze, torej se ocenjeni model dobro prilagodi podatkom. Likelihood ratio R2L dobimo, če izračunamo tudi devianco reduciranega (ničelnega) sistema D_{null}:

$$R_L^2 = 1 - \frac{D_P}{D_{null}} \tag{29}$$

Na podoben način dobimo tudi ostale pseudo-R2 metrike, kot npr. McFaddenov R2, itn.

4.2.2. Waldov test za testiranje posameznih regresorjev

Za test enega samega parametra oz. testiranje parametrov posameznih regresorjev običajno uporabimo takoimenovano **Waldovo statistiko**. Za velik N in pri pravilni ničelni hipotezi $H_0 : b_p = 0$ je Waldova statistika približno standardizirana normalna naključna spremenljivka. Pri izbrani stopnji pomembnosti α in nasprotni hipotezi $H_1 : b_p \neq 0$ sprejmemo ničelno hipotezo, če velja $w_p^* \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}} \right)$. Test hipoteze lahko zgradimo tudi s statistiko w_p^{*2}, ki je χ^2 naključna spremenljivka s stopnjo prostosti 1 [Jesenko].

Dokazati se da, da pri dovolj velikem N približno velja:

$$\frac{\hat{b}^*(p) - b_p}{\sigma\left(\hat{b}^*(p)\right)} = \frac{\hat{b}^*(p) - b_p}{\sqrt{S^*_{p+1,p+1}}} \in Z_p^* = \underset{p=0,\ldots,P}{N(0,1)}, \tag{30}$$

kjer so $\sigma\left(\hat{b}^*(p)\right)$ standardne napake optimalnih ocen koeficientov iz izraza (22). Postavimo naslednji hipotezi:

$$\begin{aligned} &H_0 : b_p = 0 \\ &H_1 : b_p \neq 0, \\ &p = 0,\ldots,P \end{aligned} \tag{31}$$

Če ničelna hipoteza velja, sledi:

$$\frac{\hat{b}^*(p)-0}{\sqrt{S^*_{p+1,p+1}}}=\frac{\hat{b}^*(p)}{\sqrt{S^*_{p+1,p+1}}}=Z^*_p=w^*_p \tag{32}$$

kjer je w^*_p takoimenovana Waldova statistika. Za ta test torej velja:

$$\begin{aligned} &H_0 \text{ sprejmemo, če velja: } w^*_p \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right) \Rightarrow \text{ Potem } \hat{b}^*_p \text{ ni SIGNIFIKANTEN} \\ &H_0 \text{ zavrnemo, če velja: } w^*_p \notin \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right) \Rightarrow \ \hat{b}^*_p \text{ je SIGNIFIKANTEN} \end{aligned} \tag{33}$$

V praksi to pomeni, da je nek regresor statistično signifikanten, če velja:

$$\left|\frac{\hat{b}^*(p)}{\sqrt{S^*_{p+1,p+1}}}\right|=\left|Z^*_p\right|=\left|w^*_p\right|>w^*_{p\ krit}=z_{\frac{\alpha}{2}} \tag{34}$$

4.2.3. Residualni testi

Za testiranje kvalitete modela so pomembni tudi različni tki. residualni testi, povezani z napako (pogreškom) prileganja izpeljanega modela danim podatkom. Analiziramo lahko različne tipe pogreškov (residualov) modela in odtod izpeljemo razne metrike za merjenje kvalitete modela. V izrazih (35) so prikazane nekatere najbolj tipične residualne metrike, ki smo jih dobili za najboljši model (pri najboljšem scenariju j^* po končanem Monte Carlo postopku), kjer znak '*' pomeni optimalne vrednosti, $n\left(j^*\right)<<N$ je velikost binomskega vzorca pri združevanju ugodnih Bernoulijevih binarnih izidov in izključitvi enkratnih dogodkov, $N_{SIGN\_REGR}\left(j^*\right)$ pa je število signifikantnih regresorjev. Poleg tega se v (35) pojavijo tudi normaliziran korelacijski koeficient CORR, kovarianca COV, ter varianca VAR. V izrazu (35) se pojavi tudi distančna metrika za simetrično Kullback-Leibler mero, ki se običajno uporablja za merjenje oddaljenosti dveh signalov v postopku Dynamic Time Warping pri signalnem procesiranju (Matlab, Signal Processing Toolbox, 2018).

$$e\left(i,j^*\right)=\begin{cases} p(i)-\hat{p}\left(i,j^*\right)=e_1 \\ \log it(i)-\overset{\wedge}{\log it}\left(i,j^*\right)=e_2 \\ e_{pearson}\left(i,j^*\right)=e_3=r_{pi}=\dfrac{\tilde{p}_i-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}}=\dfrac{\dfrac{y_i}{m_i}-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}}, \quad i=1,\ldots,N \end{cases}$$

$$MSE\left(j^*\right)=\frac{1}{n}\cdot\sum_{i=1}^{n}e^2\left(i,j^*\right), \quad RMSE\left(j^*\right)=\sqrt{MSE\left(j^*\right)}, \quad MAE\left(j^*\right)=\frac{1}{n}\cdot\sum_{i=1}^{n}\left|e\left(i,j^*\right)\right|$$

$$MAPE_g\left(j^*\right)=\frac{1}{n}\cdot\sum_{i=1}^{n}\left|\frac{e\left(i,j^*\right)}{g(i)}\right|\cdot 100;\ g(i)=\begin{cases} p(i)|e_1 \\ \log it(i)|e_2 \end{cases}, \quad \max\_err\left(j^*\right)=\max\left|e\left(i,j^*\right)\right|$$

$$KL_{\log it}\left(j^*\right)=\sum_{i=1}^{n}\left[\log it(i)-\overset{\wedge}{\log it}\left(i,j^*\right)\right]\cdot\left[\log\left[\log it(i)\right]-\log\left[\overset{\wedge}{\log it}\left(i,j^*\right)\right]\right]$$

$$KL_p\left(j^*\right)=\sum_{i=1}^{n}\left[p(i)-\hat{p}\left(i,j^*\right)\right]\cdot\left[\log\left[p(i)\right]-\log\left[\hat{p}\left(i,j^*\right)\right]\right]\Rightarrow Kullback-Leibler$$

$$S_1\left(j^*\right)=\sum_{i=1}^{n}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|>0.1\right];\quad S_2\left(j^*\right)=\sum_{i=1}^{n}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|<0.1\right]$$

$$RAZ_{DOBRI/SLABI}\left(j^*\right)=R_L^2\left(j^*\right)\cdot N_{SIGN\_REGR}\left(j^*\right)\cdot n\left(j^*\right)\cdot\frac{S_2\left(j^*\right)}{S_1\left(j^*\right)}$$

$$CORR\left(p(i),\hat{p}\left(i,j^*\right)\right)=\frac{COV\left(p(i),\hat{p}\left(i,j^*\right)\right)}{\sqrt{VAR\left(p(i)\right)}\cdot\sqrt{VAR\left(\hat{p}\left(i,j^*\right)\right)}} \tag{35}$$

$$CORR\left(\log it(i),\overset{\wedge}{\log it}\left(i,j^*\right)\right)=\frac{COV\left(\log it(i),\overset{\wedge}{\log it}\left(i,j^*\right)\right)}{\sqrt{VAR\left(\log it(i)\right)}\cdot\sqrt{VAR\left(\overset{\wedge}{\log it}\left(i,j^*\right)\right)}}$$

Poleg vseh metrik, podanih v izrazih (35), je izvedeno tudi štetje ugodnih zadetkov. Matlabova koda (K1) je naslednja:

```
% Matlab koda K1:
Ctrue=0;
Cfalse=0;
Cuncert=0;
for i = 1:length(pi)
   if (pi(i) < 0.5)&&(p_oc(i) < 0.5)
```

```
      Ctrue=Ctrue+1;  % zadetki modela
   elseif (pi(i) > 0.5)&&(p_oc(i) > 0.5)
      Ctrue=Ctrue+1;   % zadetki modela
   elseif (pi(i) == 0.5)
      Cuncert = Cuncert + 1;   % uncertain
   else
      Cfalse = Cfalse + 1;  % zgrešitve modela
   end
end
Chits = 1 – Cfalse/Ctrue;   % procent uspešnosti zadetkov modela
```

5. IZRAČUNAN REGRESIJSKI LOGIT MODEL

5.1. Logit model za uvoz 2016

V raziskavi smo izvedli tudi **Belsleyev test multikolinearnosti**, ki je za dane regresorje bil ustrezen. Dekompozicija razmerij varianc je pokazala, da je maksimalni pogojni indeks (Conditional index) le za malenkost večji od od 30 ($CI_{\max} = 36 > 30$), vendar so razmerja varianc za vse pogojne indekse $CI_1,...,CI_{\max}$ bila ustrezna. Tudi faktorji inflacije variance VIF so bili ustrezni ($VIF\left(p\right) \le 7.8 < 10,\ za\ \forall p$). Po koncu Monte Carlo postopka in celotne raziskave se je izkazalo, da **pri uvozu** na vrednosti y najbolj (signifikantno!) vplivajo naslednji regresorji: **'konstanta', 'Zabojnik', 'Šifra_lege_kraja', 'Statistična vrednost blaga (v razredih)' , ter 'Neto masa (v razredih)'**. To pomeni, da imajo signifikanten vpliv na tip incoterma, to je binarne spremenljivke, gledane skozi prizmo seštetih zaporedij ugodnih Bernoulijevih poskusov kot izhodne binomske spremenljivke, naslednje vhodne spremenljivke:

- Vhodna spremenljivka $X_1 = X_{ZABOJNIK}$ (glej sliko 3),
- Vhodna spremenljivka $X_2 = X_{SIFRA\_LEGE\_KRAJA}$ (glej sliko 4),
- Vhodna spremenljivka $X_9 = X_{CASH\_RAZR}$ - Statistična vrednost blaga (v razredih) (glej sliko 11), ter

- Vhodna spremenljivka $X_{10} = X_{MASA\_RAZR}$ - Masa blaga (v razredih) (glej sliko 12).

Simbolično bi to lahko označili na naslednji način (glej tudi izraza (1), (2)):

$$y = f\left(X_1, X_2, X_9, X_{10}, link = logit\right)$$

$$\mathbf{X} = \left[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_9}, \mathbf{X_{10}}\right] = \begin{bmatrix} X_1(1) & X_2(1) & X_9(1) & X_{10}(1) \\ X_1(2) & X_2(2) & X_9(2) & X_{10}(2) \\ \dots & \dots & \dots & \dots \\ X_1\left(n\left(j^*\right)\right) & X_2\left(n\left(j^*\right)\right) & X_9\left(n\left(j^*\right)\right) & X_{10}\left(n\left(j^*\right)\right) \end{bmatrix} \quad (36)$$

Na osnovi izraza (23) lahko za ocenjene verjetnosti in ocenjeni logit zapišemo:

$$\hat{p}_i^{\,*} = \frac{e^{\mathbf{x}_{ia}^T \cdot \hat{\mathbf{B}}^*}}{1 + e^{\mathbf{x}_{ia}^T \cdot \hat{\mathbf{B}}^*}} = \frac{e^{\hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i) + \hat{b}_4^* x_{10}(i)}}{1 + e^{\hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i) + \hat{b}_4^* x_{10}(i)}}, \qquad i = 1, \dots, n\left(j^*\right) = \mathbf{134}$$
$$\log it\left(\hat{p}_i^*\right) = \hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i) + \hat{b}_4^* x_{10}(i) \quad (37)$$

Ocenjeni parametri najboljšega Monte Carlo scenarija pridejo:

$$\tilde{b}_0^* = \hat{b}_0^* = \mathbf{5.9497}, \ \tilde{b}_1^* = \hat{b}_1^* = \mathbf{-1.1012}, \tilde{b}_2^* = \hat{b}_2^* = \mathbf{-2.0852},$$
$$\tilde{b}_9^* = \hat{b}_3^* = \mathbf{-0.25753}, \ \tilde{b}_{10}^* = \hat{b}_4^* = \mathbf{-1.0053} \quad (38)$$

ODDS razmerja za ocenjene parametre pridejo (glej izraz (25)):

$$OR_1^* = e^{\hat{b}_1^*} = \mathbf{0.33246}$$
$$OR_2^* = e^{\hat{b}_2^*} = \mathbf{0.12428}$$
$$OR_3^* = e^{\hat{b}_3^*} = \mathbf{0.77296} \quad (39)$$
$$OR_4^* = e^{\hat{b}_4^*} = \mathbf{0.36594}$$

Devianca polnega modela logistične regresije pride enaka (glej (26)):

$$D_P = 2\left[\sum_{i=1}^{134}\left\{y_i \cdot \ln\frac{y_i}{\hat{y}_i^*} + \left(m_i - y_i\right)\ln\frac{m_i - y_i}{m_i - \hat{y}_i^*}\right\}\right] = \mathbf{125.47}$$
kjer je $\hat{y}_i^* = m_i \cdot \hat{p}_i^*$,
$$\hat{p}_i^* = \frac{e^{b_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_9^* x_9(i) + \hat{b}_{10}^* x_{10}(i)}}{1 + e^{b_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_9^* x_9(i) + \hat{b}_{10}^* x_{10}(i)}}, \quad (40)$$
$i = 1, \dots, 134$

Ker velja: $D_P < D_{Pkrit} = \chi_{krit}{}^2(\alpha, N-p-1) = \mathbf{171.8}$ (glej izraz (28)), se očitno dobljeni logit model dobro prilagodi podatkom ($\alpha = 0.05$). Likelihood ratio R2L dobimo, če izračunamo tudi devianco reduciranega (ničelnega) sistema D_{null} (glej izraz (29)) :

$$R_L^2 = 1 - \frac{D_P}{D_{null}} = 1 - \frac{125.47}{1265.7} = 0.90087 \tag{41}$$

Torej tudi dokaj visoka vrednost Likelihood ratio R2L potrdi, da se dobljeni logit model dobro prilagodi podatkom. Za Waldovo signifikantnost regresorjev velja (glej (34)):

$$w^*_{p\ krit} = z_{\frac{\alpha}{2}} = \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(0)}{\sqrt{S^*_{1,1}}}\right| = \left|\frac{\hat{b}^*_0}{\sqrt{S^*_{1,1}}}\right| = \left|Z^*_0\right| = \left|w^*_0\right| = \mathbf{7.4251} > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(1)}{\sqrt{S^*_{2,2}}}\right| = \left|\frac{\hat{b}^*_1}{\sqrt{S^*_{2,2}}}\right| = \left|Z^*_1\right| = \left|w^*_1\right| = \left|\mathbf{-2.7177}\right| > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(2)}{\sqrt{S^*_{3,3}}}\right| = \left|\frac{\hat{b}^*_2}{\sqrt{S^*_{3,3}}}\right| = \left|Z^*_2\right| = \left|w^*_2\right| = \left|\mathbf{-26.596}\right| > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(3)}{\sqrt{S^*_{4,4}}}\right| = \left|\frac{\hat{b}^*_3}{\sqrt{S^*_{4,4}}}\right| = \left|Z^*_3\right| = \left|w^*_3\right| = \left|\mathbf{-2.6969}\right| > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(4)}{\sqrt{S^*_{5,5}}}\right| = \left|\frac{\hat{b}^*_4}{\sqrt{S^*_{5,5}}}\right| = \left|Z^*_4\right| = \left|w^*_4\right| = \left|\mathbf{-2.1711}\right| > \mathbf{1.96} \tag{42}$$

Tudi Waldovo testiranje signifikantnosti regresorjev potrdi ustrezno veljavnost dobljenega logit modela.

Nazadnje si poglejmo še, kako so se izračunali nekateri najbolj tipični residualni testi (glej izraz (35)):

$$e\left(i,j^*\right)=\begin{cases} p(i)-\hat{p}\left(i,j^*\right)=e_1 \\ \log it(i)-\overset{\wedge}{\log it}\left(i,j^*\right)=e_2 \\ e_{pearson}\left(i,j^*\right)=e_3=r_{pi}=\dfrac{\tilde{p}_i-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}}=\dfrac{\dfrac{y_i}{m_i}-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}},\quad i=1,...,134 \end{cases}$$

$$KL_p\left(j^*\right)=\sum_{i=1}^{134}\left[p(i)-\hat{p}\left(i,j^*\right)\right]\cdot\left[\log\left[p(i)\right]-\log\left[\hat{p}\left(i,j^*\right)\right]\right]=\mathbf{0.24017}$$

$$S_1\left(j^*\right)=\sum_{i=1}^{134}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|>0.1\right]=\mathbf{0.4403};\quad S_2\left(j^*\right)=\sum_{i=1}^{n}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|<0.1\right]=\mathbf{0.5597}$$

$$RAZ_{DOBRI/SLABI}\left(j^*\right)=R_L^2\left(j^*\right)\cdot N_{SIGN\_REGR}\left(j^*\right)\cdot n\left(j^*\right)\cdot\frac{S_2\left(j^*\right)}{S_1\left(j^*\right)}=\mathbf{0.90\times 5\times 134\times\frac{0.5597}{0.4403}=766.5}$$

$$CORR\left(p(i),\hat{p}\left(i,j^*\right)\right)=\frac{COV\left(p(i),\hat{p}\left(i,j^*\right)\right)}{\sqrt{VAR\left(p(i)\right)}\cdot\sqrt{VAR\left(\hat{p}\left(i,j^*\right)\right)}}=\mathbf{0.86306}$$

$$CORR\left(\log it(i),\overset{\wedge}{\log it}\left(i,j^*\right)\right)=\frac{COV\left(\log it(i),\overset{\wedge}{\log it}\left(i,j^*\right)\right)}{\sqrt{VAR\left(\log it(i)\right)}\cdot\sqrt{VAR\left(\overset{\wedge}{\log it}\left(i,j^*\right)\right)}}=\mathbf{0.81326} \qquad (43)$$

$$C_{true}=\mathbf{103},\ C_{false}=\mathbf{4},\ C_{uncert}=27,\ C_{hits}=\mathbf{0.96117}$$

Tudi prikazani residualni testi potrdijo dobro kvaliteto modela, kar se tiče residualov oz. pogreškov prileganja realnim podatkom. Enako velja za tiste retultate tesotv, ki niso prikazani. KL metrika je zelo majhna (0.24017), normirano razmerje med dobrimi in slabimi izidi (pravilen ali napačen razred – 0 ali 1) napovedi modela je visoko (766.5), oba korelacijska koeficienta zadovoljivo visoka (0.86306 oz. 0.81326), model pa pravilno zadane napoved kar 103-krat, štirikrat zgreši, 27 je pa nedoločenih situacij (ko so realne verjetnosti točno 0.5). Tako procentualno zagotovi kar 96.117% zadetkov izida razreda (0 ali 1).

5.2. Logit model za izvoz 2016

V raziskavi smo ponovno izvedli tudi **Belsleyev test multikolinearnosti**, ki je za dane regresorje tudi tu bil ustrezen. Dekompozicija razmerij varianc je pokazala, da je maksimalni

pogojni indeks (Conditional index) manjši od 30 ($CI_{\max} = 24 < 30$), razmerja varianc za vse pogojne indekse $CI_1, ..., CI_{\max}$ pa so bila ustrezna. Tudi faktorji inflacije variance VIF so bili ustrezni ($VIF\left(p\right) \le 8.7 < 10,\ za\ \forall p$). Po koncu Monte Carlo postopka in celotne raziskave se je izkazalo, da **pri izvozu** na vrednosti y najbolj (signifikantno!) najbolj vplivajo naslednji regresorji: **'konstanta', 'Zabojnik', 'Šifra_lege_kraja', ter 'Statistična vrednost blaga (v razredih)'**. To pomeni, da imajo signifikanten vpliv na tip incoterma, to je binarne spremenljivke, gledane skozi prizmo seštetih zaporedij ugodnih Bernoulijevih poskusov kot izhodne binomske spremenljivke, naslednje vhodne spremenljivke:

- Vhodna spremenljivka $X_1 = X_{ZABOJNIK}$ (glej sliko 3),
- Vhodna spremenljivka $X_2 = X_{SIFRA\_LEGE\_KRAJA}$ (glej sliko 4),
- Vhodna spremenljivka $X_9 = X_{CASH\_RAZR}$ - Statistična vrednost blaga (v razredih) (glej sliko 11).

Kot vidimo, dobimo dokaj podobno situacijo kot pri uvozu, le da smo tam imeli še regresor za maso.

Simbolično bi funkcijsko odvisnost izhoda od vhodov lahko označili na naslednji način (glej tudi izraza (1), (2)):

$$y = f\left(X_1, X_2, X_9, link = logit\right)$$

$$\mathbf{X} = \left[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_9}\right] = \begin{bmatrix} X_1(1) & X_2(1) & X_9(1) \\ X_1(2) & X_2(2) & X_9(2) \\ \dots & \dots & \dots \\ X_1\left(n\left(j^*\right)\right) & X_2\left(n\left(j^*\right)\right) & X_9\left(n\left(j^*\right)\right) \end{bmatrix} \quad (44)$$

Na osnovi izraza (23) lahko za ocenjene verjetnosti in ocenjeni logit zapišemo:

$$\hat{p}_i^{\,*} = \frac{e^{\mathbf{x}_{ia}^T \cdot \hat{\mathbf{B}}^*}}{1 + e^{\mathbf{x}_{ia}^T \cdot \hat{\mathbf{B}}^*}} = \frac{e^{\hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i)}}{1 + e^{\hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i)}}, \qquad i = 1, ..., n\left(j^*\right) = \mathbf{237}$$
$$\log it\left(\hat{p}_i^*\right) = \hat{b}_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_3^* x_9(i) \quad (45)$$

Ocenjeni parametri najboljšega Monte Carlo scenarija pridejo:

$$\tilde{b}_0^* = \hat{b}_0^* = \textbf{-2.3366},\ \tilde{b}_1^* = \hat{b}_1^* = \textbf{-0.57533}, \tilde{b}_2^* = \hat{b}_2^* = \textbf{1.3162},$$
$$\tilde{b}_9^* = \hat{b}_3^* = \textbf{-0.10589} \qquad (46)$$

ODDS razmerja za ocenjene parametre pridejo (glej izraz (25)):

$$OR_1^* = e^{\hat{b}_1^*} = \textbf{0.56252}$$
$$OR_2^* = e^{\hat{b}_2^*} = \textbf{3.7294} \qquad (47)$$
$$OR_3^* = e^{\hat{b}_3^*} = \textbf{0.89953}$$

Devianca polnega modela logistične regresije pride enaka (glej (26)):

$$D_P = 2\left[\sum_{i=1}^{237}\left\{ y_i \cdot \ln\frac{y_i}{\hat{y}_i^*} + (m_i - y_i)\ln\frac{m_i - y_i}{m_i - \hat{y}_i^*}\right\}\right] = \textbf{167.54}$$
kjer je $\hat{y}_i^* = m_i \cdot \hat{p}_i^*$, (48)
$$\hat{p}_i^* = \frac{e^{b_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_9^* x_9(i)}}{1 + e^{b_0^* + \hat{b}_1^* x_1(i) + \hat{b}_2^* x_2(i) + \hat{b}_9^* x_9(i)}},$$
$$i = 1, \ldots, 237$$

Ker velja: $D_P < D_{Pkrit} = {\chi_{krit}}^2(\alpha, N - p - 1) = \textbf{288.99}$ (glej izraz (28)), se očitno dobljeni logit model tudi pri izvozu dobro prilagodi podatkom ($\alpha = 0.05$). Likelihood ratio R2L dobimo, če izračunamo tudi devianco reduciranega (ničelnega) sistema D_{null} (glej izraz (29)) :

$$R_L^2 = 1 - \frac{D_P}{D_{null}} = 1 - \frac{167.54}{746.06} = \textbf{0.7754} \qquad (49)$$

Tu je sicer vrednost manjša kot pri uvozu (glej izraz (41)), a še vedno zadovoljivo visoka. Torej tudi tukaj relativno visoka vrednost Likelihood ratio R2L potrdi, da se dobljeni logit model dobro prilagodi podatkom. Za Waldovo signifikantnost regresorjev velja (glej (34)):

$$w^*_{p\ krit} = z_{\frac{\alpha}{2}} = \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(0)}{\sqrt{S^*_{1,1}}}\right| = \left|\frac{\hat{b}^*_0}{\sqrt{S^*_{1,1}}}\right| = \left|Z^*_0\right| = \left|w^*_0\right| = \left|\mathbf{-4.8394}\right| > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(1)}{\sqrt{S^*_{2,2}}}\right| = \left|\frac{\hat{b}^*_1}{\sqrt{S^*_{2,2}}}\right| = \left|Z^*_1\right| = \left|w^*_1\right| = \left|\mathbf{-2.3871}\right| > \mathbf{1.96} \qquad (50)$$

$$\left|\frac{\hat{b}^*(2)}{\sqrt{S^*_{3,3}}}\right| = \left|\frac{\hat{b}^*_2}{\sqrt{S^*_{3,3}}}\right| = \left|Z^*_2\right| = \left|w^*_2\right| = \left|\mathbf{21.648}\right| > \mathbf{1.96}$$

$$\left|\frac{\hat{b}^*(3)}{\sqrt{S^*_{4,4}}}\right| = \left|\frac{\hat{b}^*_3}{\sqrt{S^*_{4,4}}}\right| = \left|Z^*_3\right| = \left|w^*_3\right| = \left|\mathbf{-2.1731}\right| > \mathbf{1.96}$$

Tudi Waldovo testiranje signifikantnosti regresorjev potrdi ustrezno veljavnost dobljenega logit modela.

Nazadnje si poglejmo še, kako so se izračunali nekateri najbolj tipični residualni testi (glej izraz (35)):

$$e\left(i,j^*\right)=\begin{cases} p(i)-\hat{p}\left(i,j^*\right)=e_1 \\ \log it(i)-\hat{\log it}\left(i,j^*\right)=e_2 \\ e_{pearson}\left(i,j^*\right)=e_3=r_{pi}=\dfrac{\tilde{p}_i-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}}=\dfrac{\frac{y_i}{m_i}-\hat{p}_i^*}{\sqrt{\hat{p}_i^*\cdot\left(1-\hat{p}_i^*\right)}},\quad i=1,\ldots,237 \end{cases}$$

$$KL_p\left(j^*\right)=\sum_{i=1}^{237}\left[p(i)-\hat{p}\left(i,j^*\right)\right]\cdot\left[\log\left[p(i)\right]-\log\left[\hat{p}\left(i,j^*\right)\right]\right]=\mathbf{0.0829}$$

$$S_1\left(j^*\right)=\sum_{i=1}^{237}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|>0.1\right]=\mathbf{0.48945};\quad S_2\left(j^*\right)=\sum_{i=1}^{237}\left[\left|p(i)-\hat{p}\left(i,j^*\right)\right|<0.1\right]=\mathbf{0.51055}$$

$$RAZ_{DOBRI/SLABI}\left(j^*\right)=R_L^2\left(j^*\right)\cdot N_{SIGN\_REGR}\left(j^*\right)\cdot n\left(j^*\right)\cdot\frac{S_2\left(j^*\right)}{S_1\left(j^*\right)}=\mathbf{0.775\times 4\times 237\times\frac{0.51055}{0.48945}=766.3}$$

$$CORR\left(p(i),\hat{p}\left(i,j^*\right)\right)=\frac{COV\left(p(i),\hat{p}\left(i,j^*\right)\right)}{\sqrt{VAR\left(p(i)\right)}\cdot\sqrt{VAR\left(\hat{p}\left(i,j^*\right)\right)}}=\mathbf{0.77196}$$

$$CORR\left(\log it(i),\hat{\log it}\left(i,j^*\right)\right)=\frac{COV\left(\log it(i),\hat{\log it}\left(i,j^*\right)\right)}{\sqrt{VAR\left(\log it(i)\right)}\cdot\sqrt{VAR\left(\hat{\log it}\left(i,j^*\right)\right)}}=\mathbf{0.75955} \qquad (51)$$

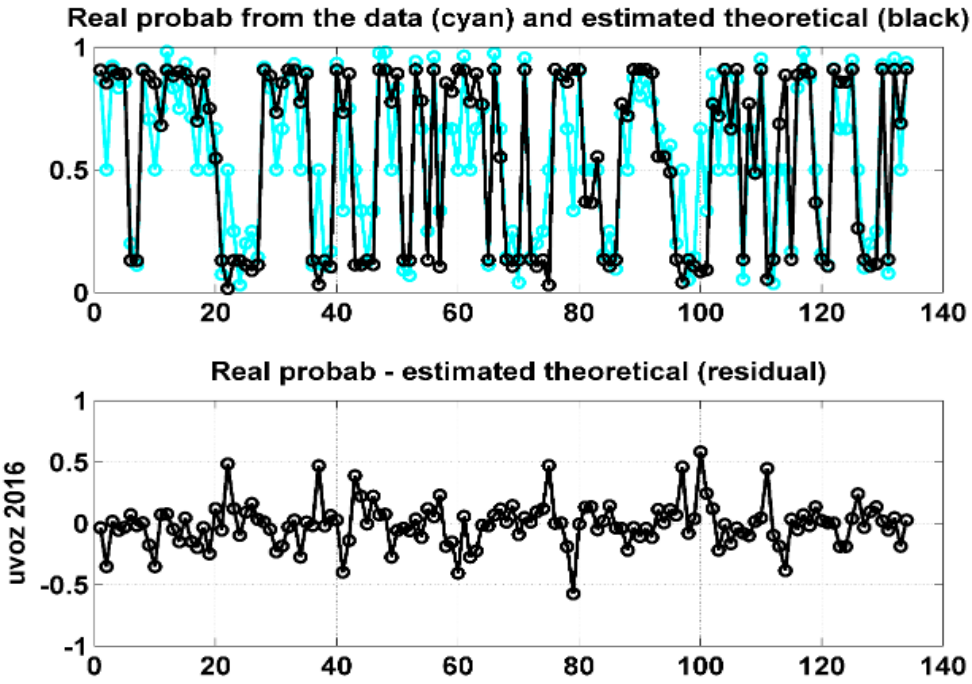
$$C_{true}=\mathbf{153},\ C_{false}=\mathbf{16},\ C_{uncert}=\mathbf{68},\ C_{hits}=\mathbf{0.8954}$$

Tudi prikazani residualni testi potrdijo dobro kvaliteto modela, kar se tiče residualov oz. pogreškov prileganja realnim podatkom. Enako velja za tiste retultate tesotv, ki niso prikazani. KL metrika je zelo majhna (0.0829), normirano razmerje med dobrimi in slabimi izidi (pravilen ali napačen razred – 0 ali 1) napovedi modela je visoko (766.3), oba korelacijska koeficienta zadovoljivo visoka (0.77196 oz. 0.75955), model pa pravilno zadane napoved kar 153-krat, 16 krat zgreši, 68 je pa nedoločenih situacij (ko so realne verjetnosti točno 0.5). Tako procentualno zagotovi kar 89.54% zadetkov izida razreda (0 ali 1).

6. KVALITETETA PRILEGANJA MODELA REALNIM PODATKOM

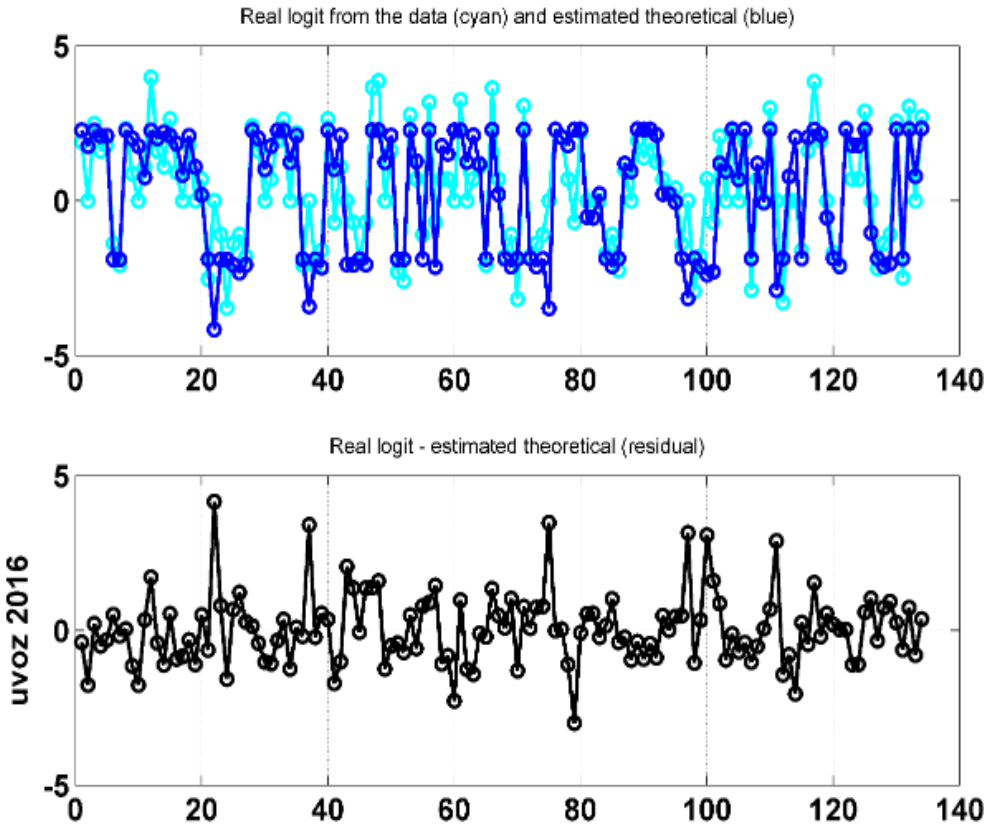
6.1. Kvaliteteta Prileganja za Logit model za uvoz 2016

Slika 15 prikazuje dejansko in ocenjeno verjetnost glede na indeks danih 134 podatkov. Kot lahko vidimo, se ocenjene verjetnosti $\hat{p}\left(i, j^*\right)$ za status Incoterma v razredu 0 oz. 1, dokaj dobro prilegajo dejanskim verjetnostim $\tilde{p}_i = \frac{y_i}{m_i}$, kjer je y_i število nahajanj (število enk) v razredu 1 pri dani kombinaciji signifikantnih vhodnih spremenljivk $X(i) = \left[X_1(i), X_2(i), X_9(i), X_{10}(i)\right], i = 1,...,134$, in spreminjajočim se kombinacijam vrednosti ostalih spremenljivk $X_{ost}(i) = \left[X_3(i), X_4(i),..., X_8(i), X_{11}(i)\right], i = 1,...,134$, ki se m_i krat ponovijo. Kot je razvidno na sliki 15, je spodaj prikazan tudi residual, to je pogrešek modela $p(i) - \hat{p}\left(i, j^*\right) = e_1$. Slednji je v 103 primerih pod mejo 0.1 in da model pravo oceno razreda (glej $C_{true} = \mathbf{103}$ v izrazu (35)), štirikrat je napačna ocena razreda ($C_{false} = \mathbf{4}$), 27 je pa nedoločenih situacij (ko so realne verjetnosti točno 0.5).



Slika 15: Dejanska ($\tilde{p}_i = \frac{y_i}{m_i}$) in ocenjena verjetnost $\hat{p}\left(i, j^*\right)$ glede na indeks danih 134 podatkov, ter residual $p(i) - \hat{p}\left(i, j^*\right) = e_1$.

Slika 16 prikazuje dejanske in ocenjene vrednosti logit funkcije glede na indeks danih 134 podatkov. Kot lahko vidimo, se tudi ocenjene vrednosti logit funkcije $\log it\left(\hat{p}_i^*\right)$ dokaj dobro prilegajo dejanskim vrednostim logit funkcije $\log it\left(\tilde{p}_i = \frac{y_i}{m_i}\right)$.



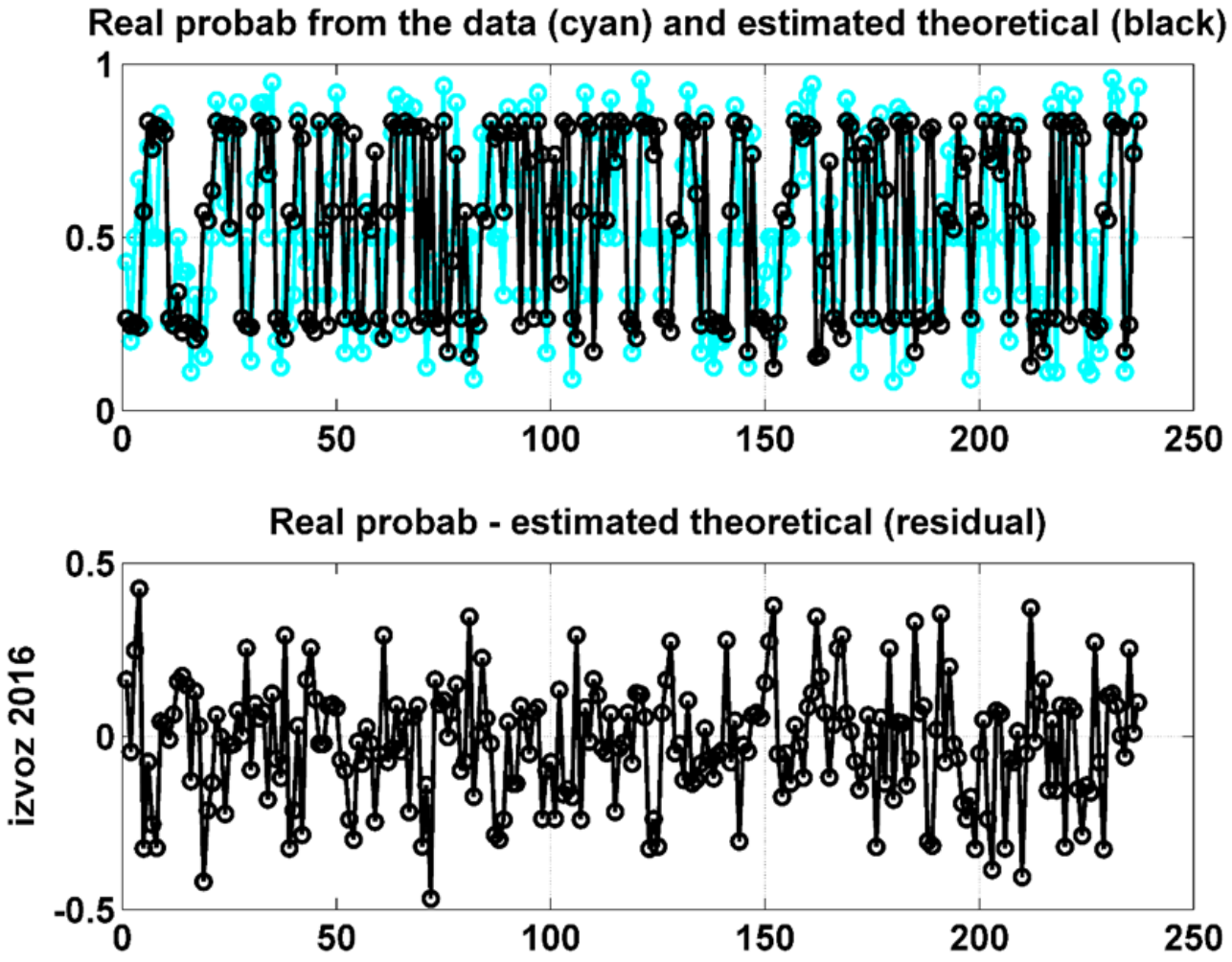
Slika 16: Ocenjene vrednosti logit funkcije $\log it\left(\hat{p}_i^*\right)$ in dejanske vrednosti logit funkcije $\log it\left(\tilde{p}_i = \frac{y_i}{m_i}\right)$ glede na indeks danih 134 podatkov, ter residual $\log it(i) - \hat{\log it}\left(i, j^*\right) = e_2$

(pri tem veljata enakosti: $\log it\left(\tilde{p}_i = \frac{y_i}{m_i}\right) = \log it(i)$; $\log it\left(\hat{p}_i^*\right) = \hat{\log it}\left(i, j^*\right)$).

6.2. Kvaliteteta Prileganja za Logit model za izvoz 2016

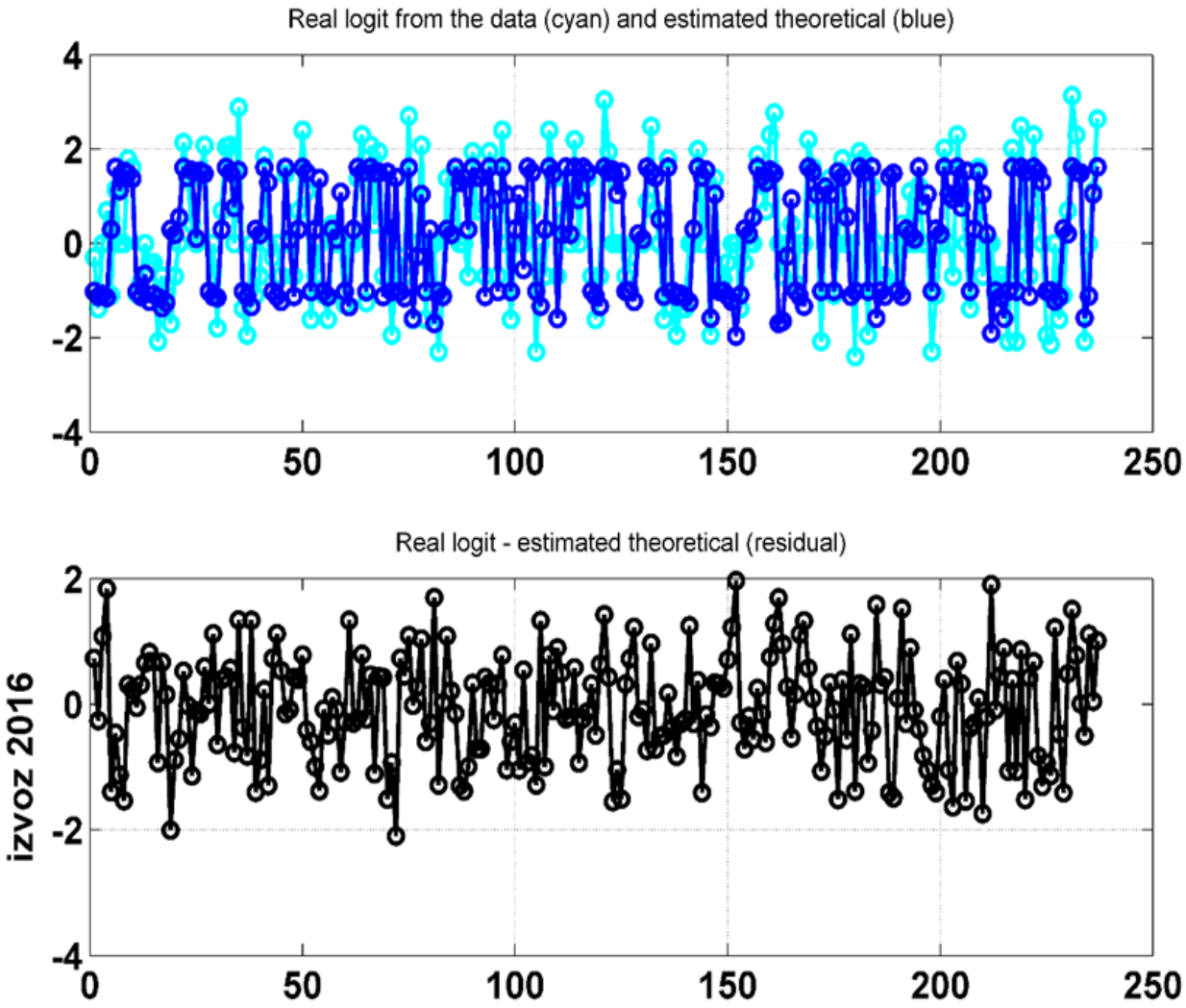
Slika 17 prikazuje dejansko in ocenjeno verjetnost glede na indeks danih 237 podatkov. Kot lahko vidimo, se ocenjene verjetnosti $\hat{p}\left(i, j^*\right)$ za status Incoterma v razredu 0 oz. 1, tudi v

primeru izvoza solidno prilegajo dejanskim verjetnostim $\tilde{p}_i = \frac{y_i}{m_i}$, kjer je y_i število nahajanj (število enk) v razredu 1 pri dani kombinaciji signifikantnih vhodnih spremenljivk $X(i) = \left[X_1(i), X_2(i), X_9(i)\right], i = 1,...,237$, in spreminjajočim se kombinacijam vrednosti ostalih spremenljivk $X_{ost}(i) = \left[X_3(i), X_4(i),..., X_8(i), X_{10}(i), X_{11}(i)\right], i = 1,...,237$, ki se m_i krat ponovijo. Kot je razvidno na sliki 17, je spodaj prikazan tudi residual, to je pogrešek modela $p(i) - \hat{p}(i, j^*) = e_1$. Slednji je v 153 primerih pod mejo 0.1 in da model pravo oceno razreda (glej $C_{true} = \mathbf{153}$ v izrazu (51)), 16 krat je napačna ocena razreda ($C_{false} = \mathbf{16}$), 68 je pa nedoločenih situacij (ko so realne verjetnosti točno 0.5).



Slika 17: Dejanska ($\tilde{p}_i = \frac{y_i}{m_i}$) in ocenjena verjetnost $\hat{p}(i, j^*)$ glede na indeks danih 237 podatkov, ter residual $p(i) - \hat{p}(i, j^*) = e_1$.

Slika 18 prikazuje dejanske in ocenjene vrednosti logit funkcije glede na indeks danih 237 podatkov. Kot lahko vidimo, se tudi ocenjene vrednosti logit funkcije $\log it\left(\hat{p}_i^*\right)$ dokaj dobro prilegajo dejanskim vrednostim logit funkcije $\log it\left(\tilde{p}_i = \frac{y_i}{m_i}\right)$.



Slika 18: Ocenjene vrednosti logit funkcije $\log it\left(\hat{p}_i^*\right)$ in dejanske vrednosti logit funkcije $\log it\left(\tilde{p}_i = \frac{y_i}{m_i}\right)$ glede na indeks danih 237 podatkov, ter residual $\log it(i) - \hat{\log it}\left(i, j^*\right) = e_2$.

7. VIRI IN LITERATURA

Dejan Dragan, 2016: *Logistična regresija s programskim orodjem Matlab*, Fakulteta za logistiko UM.

V Celju, 06.02.2019

Poročilo izdelal: izr. prof. dr. Dejan Dragan