Model predikcije malignosti tumora dojke logističkom regresijom

Krajina, Jelena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Jelena Krajina

MODEL PREDIKCIJE MALIGNOSTI TUMORA DOJKE LOGISTIČKOM REGRESIJOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, rujan, 2017

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem roditeljima, Ani, Ivanu i prijateljima na svoj ljubavi i podršci koju su mi dali. Posebno zahvaljujem svojoj mentorici, prof. dr. sc. Anamariji Jazbec koja mi je svojim strpljenjem i savjetima pomogla pri izradi ovog rada.

Sadržaj

Popis engleskog nazivlja iz SAS tablica

-2LogL (-2LogLikelihood) - -2Log(vjerodostojnost)
Distribution-distribucija
Estimated Probability - procijenjena vjerojatnost
Intercept - ili β_0, očekivana vrijednost zavisne varijable kada su vrijednosti nezavisnih varijabli jednake 0
Intercept and Covariates - misli se na puni, satuirani model
Maximum - najveća vrijednost
Median-medijan
Mean-očekivanje
Minimum - najmanja vrijednost
Likelihood Ratio (χ^2) - omjer vjerodostojnosti χ^2 testa da je barem jedan od parametara u regresijskom modelu različiti od 0
OR (odds ratio) - omjer šansi
ROC curve - ROC krivulja
Sensitivity-osjetljivost
Specificity-specifičnost
Std Dev - standardna devijacija
Tolerance - tolerancija
Variance Inflation - inflacija varijance
Wald χ^2 - vrijednost testne statistike za hipotezu da je vrijednost procijenjenog parametra 0 ako su ostale varijable u modelu

Uvod

Rak dojke je najčešća zloćudna bolest žena u razvijenom svijetu, iako, vrlo rijetko, od raka dojke mogu oboljeti i muškarci. Manifestira se pojavom nove tvorbe u području dojke. Skoro polovica žena koje obole od raka dojke razvije metastatsku bolest. Oko trećinu svih malignih tumora kod žena čini upravo rak dojke. U Hrvatskoj je stopa incidencije (broj novooboljelih na 100 000 stanovnika) viša nego u Europi i ima trend rasta. Također, rak dojke je visoko zastupljen među uzrocima smrti kod žena. To je postao javnozdravstveni problem i veliki se napori ulažu kako u ranom otkrivanju te bolesti, tako i u njenom liječenju. Rak dojke je najčešća zloćudna (maligna) bolest u žena. U oko 80% novih slučajeva bolest bude otkrivena u fazi ranog raka, a oko 20% u trenutku postavljanja dijagnoze bolest već bude u uznapredovaloj fazi. Prema podacima Registra za rak, Zavoda za javno zdravstvo Republike Hrvatske godišnje od raka dojke oboli preko 2500 žena. Prema posljednim dostupnim epidemiološkim podacima rak dojke je treći uzrok smrti u ženskoj populaciji u 2012. (iza ishemijske bolesti srca i cerebrovaskularne bolesti). Stopa smrtnosti od raka dojke u Hrvatskoj je među najvišima u Europi. 2012 godine preko 1000 žena umrlo je od raka dojke.[9]

U ovom radu ćemo koristeći model logističke regresije napraviti prediktivni model malignosti tumora dojke koristeći podatke „Breast Cancer Wisconsin (Diagnostic) Data Set“. Podatke je prikupio dr. William H. Wolber (University of Wisconsin Hospitals, Madison). Podatci sadrže 699 opservacija i 11 varijabli, odnosno atributa, a to su: Identifikacijski broj uzorka, Debljina grumena, Uniformnost veličine stanice, Uniformnost staničnog oblika, Marginalna adhezija, Veličina epitelne stanice, Gole jezgre, Normalan kromatin, Normalne jezgrice, Mitoza i Razred. Atributi su dobiveni iz digitalizirane slike tkiva uzetog pri punkcijsko-aspiracijskoj biopsiji. [7]

Podatke ćemo obrađivati u statističkom programu SAS.

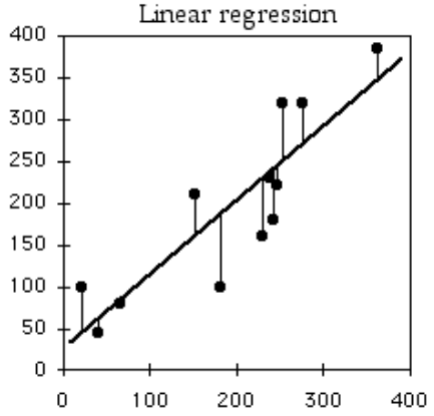
Poglavlje 1

Logistička regresija

1.1 Regresijska analiza

Regresijska analiza je metoda ispitivanja ovisnosti jedne (zavisne) varijable o jednoj ili više drugih (nezavisnih) varijabli. Jedan od rezultata regresijske analize jest regresijski model. Regresijski model je matematička jednadžba koja kvantificira povezanost između zavisne i nezavisne, odnosno zavisne i nezavisnih varijabli. Varijablu odaziva ili zavisnu varijablu označavamo s y, dok varijablu poticaja ili nezavisnu varijablu označavamo s x.

Povezanost između zavisne i nezavisne varijable može biti linearna i u tom slučaju govorimo o linearnoj regresiji.[5]



Slika 1.1: Primjer linearne regresije
izvor:http://www.biostathandbook.com/linearregression.html,2017

1.2 Linearna regresija

Kod linearne regresije povezanost između zavisne i nezavisne varijable opisana je jednadžbom pravca. Pravac koji najbolje opisuje povezanost tih dviju varijabli odredimo tako da iz skupa svih pravaca odaberemo pravac čija je suma odstupanja svake točke od pravca najmanja (metoda najmanjih kvadrata).

Promotrimo jednodimenzionalni linearni model

$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \varepsilon$$

gdje su:

- $x_1, x_2, ..., x_p$- varijable poticaja ili nezavisne varijable
- ε - slučajna greška
- y - varijabla odaziva ili zavisna varijabla
- $\beta_0, \beta_1, ..., \beta_p$- parametri modela

Linearan model u kojemu je p=1 zovemo univarijatna linearna regresija, dok se model u kojemu je p >1 naziva multivarijatna linearna regresija.

U primjeni imamo više opažanja pa to zapisujemo:

$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, ..., n$$

gdje pretpostavljamo da su greške $\varepsilon_1, ..., \varepsilon_n$ nezavisne s distribucijom N(0,σ^2).
Kraće to zapisujemo u matričnom obliku

$$\mathbf{Y} = \mathbf{X}b + \varepsilon$$

Pritom je:

- $\mathbf{Y} = (y_1, ..., y_n)^T$

- $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T \sim N(0, \sigma^2 II)$
- $b = (\beta_0, ..., \beta_p)^T$
- $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & . & . & . & x_{1p} \\ 1 & x_{21} & . & . & . & x_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & . & . & . & x_{np} \end{bmatrix}$

Ako želimo minimizirati

$$\|\varepsilon\|_2 = \|\mathbf{Y} - \mathbf{X}b\|_2$$

po b dobivamo da je najbolja ocjena za b

$$\hat{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

(uz uvjet da je $\mathbf{X}^T\mathbf{X}$ regularna)

Procijenjene vrijednosti tada su jednake

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

gdje je $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
a ostaci

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{Y}}(II - \mathbf{H})\mathbf{Y}$$

Glavne pretpostavke koje opravdavaju korištenje linearnog regresijskog modela u svrhu analize podataka i predviđanja su:

- linearni odnos između varijabli poticaja i odaziva
- nekoreliranost varijable poticaja i greške
- nezavisnost grešaka (nekoreliranost povlači nezavisnost)
- homogenost grešaka
- normalna distribuiranost grešaka

Ukoliko neka od pretpostavki nije opravdana, naša previđanja mogu biti nevaljana.[5],[6]

1.3 Linearna regresija i dihotomna zavisna varijabla

Kao što je već navedeno, pretpostavke linearne regresije su:

$$\begin{cases} 1. & y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ 2. & \mathbf{E}(\varepsilon_i) = 0 \\ 3. & var(\varepsilon_i) = \sigma^2 \\ 4. & cov(\varepsilon_i, \varepsilon_j) = 0 \\ 5. & \varepsilon_i \sim N(0, \sigma^2) \end{cases}$$

Radi jednostavnosti pokazati ćemo za jednu nezavisnu varijablu x za koju pretpostavljamo da je fiksna. Indeksi $\imath$ predstavljaju različite članove uzorka. Pretpostavke 2, 3 i 4 poznate su i kao Gauss-Markovljevi uvjeti.
Pretpostavka 1 kaže da je y linearna funkcija od x plus slučajna greška (šum) ε za sve članove uzorka. Sve ostale pretpostavke govore nešto o distribuciji slučajne greške ε .
Pretpostavka 2 kaže da $\mathbf{E}(\varepsilon_i)$ ne ovisi o x_i što implicira da su x_i i ε_i nekorelirane za svaki $\imath$.
Pretpostavka 3, znana i kao homoskedastičnost (postojanost varijance), kaže da je varijanca od ε jednaka za sve obzervacije.
Pretpostavka 4 kaže da su slučajne greške različitih obzervacija međusobno nekorelirane.
Konačno, pretpostavka 5 kaže da su slučajne greške normalno distribuirane.
Ako su pretpostavke zadovoljene procijenitelji od β_0 i β_1 su nepristrani i imaju minimalnu varijabilnost kroz ponavljanje uzorka.
Sada pretpostavimo da je y dihotomna varijabla s mogućim vrijednostima 0 i 1. I dalje je razumno tvrditi da pretpostavke 1, 2 i 4 vrijede. No ako su 1 i 2 vrijede, tada 3 i 5 nužno ne vrijede.
Prvo promotrimo pretpostavku 5.
Pretpostavimo da je $y_i = 1$. Tada pretpostavka 1 implicira da je

$$\varepsilon_i = 1 - \beta_0 - \beta_1 x_1$$

Obrnuto, ako je $y_i = 0$ tada je

$$\varepsilon_i = -\beta_0 - \beta_1 x_1$$

Budući da ε_i poprima samo 2 vrijednosti nemoguće je da ima normalnu razdiobu (normalna razdioba ima kontinuirane vrijednosti bez donje i gornje granice). Pa slijedi da pretpostavku 5 moramo odbaciti.
Promotrimo sada pretpostavku 3. Očekivanje od y_i je po definiciji

$$\mathbf{E}(y_i) = 1 \times \mathbf{P}(y_i = 1) + 0 \times \mathbf{P}(y_i = 0)$$

Ako definiramo $p_i = \mathbf{P}(y_i = 1)$ slijedi da je $\mathbf{E}(y_i) = p_i$

(Generalno, za sve pomoćne (*engl. dummy*) varijable, njihova očekivana vrijednost jednaka je vjerojatnosti da su jednake 1)

Pretpostavke 1 i 2 također impliciraju

$$\begin{aligned}\mathbf{E}(y_i) &= \mathbf{E}\left[\beta_0 + \beta_1 x\right] \\ &= \mathbf{E}\left[\beta_0\right] + \mathbf{E}\left[\beta_1 x_i\right] + \mathbf{E}\left[\varepsilon_i\right] \\ &= \beta_0 + \beta_1 x_i\end{aligned}$$

Pa slijedi da je $p_i\beta_0 + \beta_1 x_i + \varepsilon_i$.
Ovaj izraz se ponekad naziva linearni vjerojatnosni model, odnosno vjerojatnost da je $y = 1$ je linearna funkcija od x.
Promotrimo sada varijancu od ε_i. Budući da x tretiramo kao fiksnu varijablu, varijanca od ε_i je jednaka varijanci od y_i. (Generalno varijanca pomoćne varijable je $p_i(1 - p_i)$. Slijedi:

$$var(\varepsilon_i) = p_i(1 - p_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)$$

Očito, varijanca od ε_i mora biti različita za različite obzervacije i uglavnom varira kao funkcija od x.

Pokazali smo da dihotomna zavisna varijabla u linearnoj regresiji krši pretpostavke homoskedastičnosti i normalnosti greške. Koje su posljedice toga? I nisu tako ozbiljne.

1. Nisu nam potrebne sve pretpostavke kako bismo dobili nepristrane procjenitelje. Ako su zadovoljene samo pretpostavke 1 i 2 metodom najmanjih kvadrata ćemo dobiti nepristrane procjenitelje za β_0 i β_1

2. Uvjet normalnosti nije potreban ako je uzorak relativno velik. Centralni granični teorem osigurava da procijenjeni parametri imaju distribuciju koja je aproksimativno normalna ikao slučajna greška nije normalno distribuirana. To znači da još uvijek možemo koristiti tablicu normalne razdiobe za računanje p-vrijednosti i pouzdanih intervala. No ako je uzorak malen dobivene vrijednosti mogu biti loše.

Kršenje homoskedastičnosti ima 2 neželjene posljedice. Prvo, procijenjeni parametri nisu više dovoljno dobri, odnosno postoje alternativne metode procjene paramaetara s manjom standardnom devijacijom. Drugo i ozbiljnije, procjene standardne devijacije nisu više konzistentni procjenitelji prave standardne devijacije. To znači da procjena standardne de-

vijacije može biti pristrana, a budući da standardnu devijaciju koristimo u računanju testne statistike, testna statistika također može biti pristrana.

Kao dodatak ovim tehničkim problemima postoji jedan fundamentalniji problem s pretpostavkama linearnog modela. Pokazali smo da za dihotomnu zavisnu varijablu pretpostavke 1 i 2 impliciraju linearan vjerojatnosni model

$$p_i = \beta_0 + \beta_1 x_i$$

Iako nema ništa krivo u ovom modelu, on je nerealan, pogotovo ako je x kontinuirana varijabla. Ako x nema gornju ili donju granicu tada za bilo koju vrijednost od β postoje vrijednosti od x za koje je p_i ili veće od 1 ili manje od 0 što je nemoguće budući da je p_i vjerojatnost.

Zbog ovakvih problema s linearnom regresijom statističari su razvili alternativne pristupe koji konceptualno imaju više smisla i imaju bolja statistička svojstva. Najpopularnij pristup je logistička regresija u kojoj se procjene rade metodom maksimalne vjerodostojnosti (*engl. maximum likelihood*).[3]

1.4 Logistička regresija - osnovni pojmovi

Logistička regresija je u osnovi regrsijski model čija zavisna varijabla je kategorijska. U njegovoj najraširenijoj primjeni, zavisna varijabla je jednostavna dihotomna, dok nezavisne varijable mogu biti ili kvantitativne ili kategorijske. Logistička regresija se može generalizirati i primjenjivati za zavisne varijable koje imaju više od dvije kategorije, bile one poredane ili neporedane.

Logistička funkcija definirana je sa

$$p(x) = \frac{1}{1 + e^{-x}} \tag{1.1}$$

gdje je $x \in \langle -\infty, \infty \rangle$, a $p(x) \in \langle 0, 1 \rangle$.

Logit funkcija je funkcija inverzna logističkoj. Odnosno:

$$logit(p(x)) = log\left[\frac{p(x)}{1 - p(x)}\right] = log(p(x)) - log(1 - p(x)) \tag{1.2}$$

gdje je $p \in \langle 0, 1 \rangle$, a $logit(p) \in \langle -\infty, \infty \rangle$. [5]

1.5 Šansa i logit transformacija

Šansa (*engl. odds*) nekog događaja je omjer očekivanog broja puta kada će se događaj dogoditi naspram očekivanog broja puta kada se događaj neće dogoditi. Postoji jednostavna veza između vjerojatnosti i šanse. Ako je p(x) vjerojatnost nekog događaja i odds šansa događaja tada je

$$odds = \frac{p(x)}{1 - p(x)} = \frac{\text{vjerojatnost da se događaj dogodio}}{\text{vjerojatnost da se događaj nije dogodio}}$$

$$p(x) = \frac{odds}{1 + oods}$$

Tablica 1.1: Odnos između vjerojatnosti i šanse

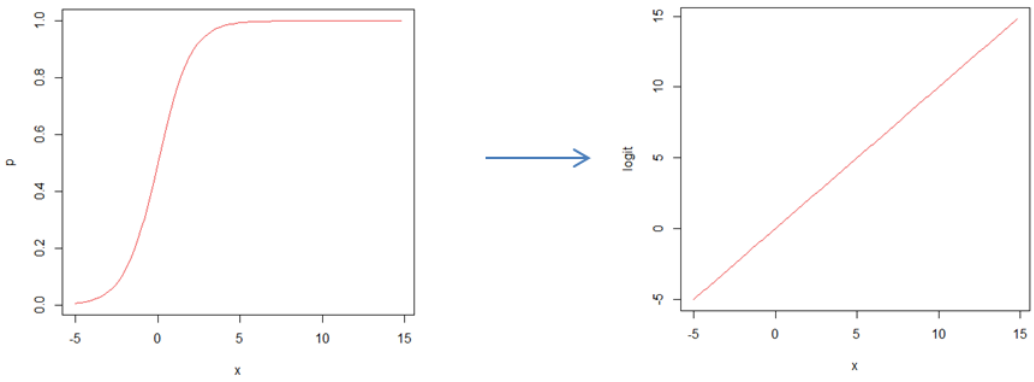
Vjerojatnost	Šansa
0,1	0,11
0,2	0,25
0,3	0,43
0,4	0,67
0,5	1,00
0,6	1,50
0,7	2,33
0,8	4,00
0,9	9,00

Primijetimo da šansa manja od 1 odgovara vjerojatnosti manjoj od 0.5, dok šansa veća od 1 odgovara vjerojatnosti većoj od 0.5. Kao i vjerojatnost, šansa ima donju granicu 0 ali nema gornju granicu.

Kod linearnog vjerojatnosnog modela vjerojatnost je ograničena s 0 i 1. Zato ju transformiramo kako bismo maknuli granice. Transformiranjem vjerojatnosti u šansu mičemo gornju granicu, a logaritmiranjem donju.

Ako govorimo o univarijatnom logističkom modelu slijedi da je

$$log(odds) = log\left[\frac{p(x)}{1-p(x)}\right] = logit(p(x)) = \beta_0 + \beta_1 x \tag{1.3}$$



Slika 1.2: Logit transformacija

Sada imamo da je očekivanje od y uz uvjet x jednako $\mathbf{E}[\, y \mid x] \; = \beta_0 + \beta_1 x$. No, budući da je y dihotomna varijabla vrijedi $\mathbf{E}[\, y \mid x] \; = p(x)$, što povlači

$$\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{1.4}$$

$$logit(p(x)) = ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x \tag{1.5}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} \tag{1.6}$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1.7}$$

Analogno vrijedi i za multivarijatni logistički model. Tada je

$$logit(p(x)) = log\left[\frac{p(x_1, x_2, ..., x_k)}{1-p(x_1, x_2, ..., x_k)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{1.8}$$

gdje je k broj nezavisnih varijabli u modelu.

$$p(x_1, x_2, ..., x_k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}} \tag{1.9}$$

[5]

1.6 Testiranje adekvatnosti modela (*engl. Goodness of fit*)

Za razliku od linearne regresije gdje metodom najmanjih kvadrata minimiziramo kvadrirane reziduale, kod logističke regresije koristimo metodu maksimalne vjerodostojnosti (oznaka ML). Kod ML metode tražimo najmanje moguće odstupanje (*engl. Deviance*) (oznaka D) između opaženih vrijednosti y i prediktivnih vrijednosti $\hat{y}$ koristeći iterativne računalne metode sve dok ne dobijemo najmanje moguće odstupanje. Jednom kada se nađe najbolje rješenje to odstupanje zovemo Deviance ili -2LogLikelihood ili Likelihood ratio

Označimo sa $\hat{y}$ ML procjenu od y, a sa $\hat{p}(x)$ procjenu od p(x). Takoder, neka je $\beta = (\beta_0, \beta_1)$, pri čemu su β_0 i β_1 parametri univarijatnog logističkog modela. Sa L označimo vjerodostojnost (*engl. Likelihood*).

$$L(\beta) = \prod_{i=1}^{n} \left[p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \right] \tag{1.10}$$

pri čemu su (x_i, y_i), $\imath = 1, \ldots$,n promatrane vrijednosti. Princip ML metode jest da procjena parametara β maksimizira izraz L(β). Označimo sa LL Log-likelihood, tj log(L(β)).

$$LL(\beta) = \sum_{i=1}^{n} [y_i ln(p(x_i)) + (1 - y_i) ln(1 - p(x_i))] \tag{1.11}$$

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

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

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

Slijedi

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

Saturirani model je model koji sadrži onoliko parametara koliko ima podataka.

Kada želimo testirati razliku modela sa i bez varijabli (prediktora) koristimo G statistiku. To je sličo kao i R^2 kod linearne regresije samo što tamo dodavanjem nove varijable koja pospješuje model povećavamo R^2, dok ovdje očekujemo da se Deviance smanji.

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

1.7 Interpretacija parametara

$$g(x) = logit(p(x)) = ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x \tag{1.12}$$

$$g(x+1) = \beta_0 + \beta_1(x+1) \tag{1.13}$$

$$g(x+1) - g(x) = \beta_1 \tag{1.14}$$

$$logit(p(x+1)) - logit(p(x)) = \beta_1 \tag{1.15}$$

$$log(odds(p(x+1))) - log(odds(p(x))) = \beta_1 \tag{1.16}$$

$$log\Big(\frac{odds(p(x+1))}{odds(p(x))}\Big) = \beta_1 \tag{1.17}$$

Uzmimo za primjer logističku regresiju s dihotomnom zavisnom i nezavisnom varijablom. Sada je

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

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

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

Pritom oznakom OR označavamo omjer šansi (*engl. odds ratio*).

Za kontinuirane nezavisne varijable

$$g(x + 1) - g(x) = \beta_1$$

β_1 pokazuje promjenu u log oddsu za pomak nezavisne varijable x za 1. Analogno tome, pomak u log oddsu za pomak nezavisne varijable x za konstantu c jednak je

$$g(x + c) - g(c) = c\beta_1$$

$$OR(c) = OR(x + c, x) = e^{c\beta_1}$$

β_0 ili intercept je očekivana vrijednost zavisne varijable y kada je x=0. [5]

1.8 Testiranje značajnosti parametara

Kod univarijatnog logističkog modela testiramo hipoteze:

$$\mathbf{H}_0 : \beta_1 = 0$$

$$\mathbf{H}_1 : \beta_1 \neq 0$$

Za parametar β_1 kažemo da je statistički značajan ako se on statistički značajno razlikuje od 0. Ako je on približno jednak nuli, tada nezavisna varijabla nije statistički značajna i njezin utjecaj na zavisnu varijablu je zanemariv.

Kod multivarijatnog logističkog modela testiramo hipoteze:

$$\mathbf{H}_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$\mathbf{H}_1 : \text{bar jedan } \beta_i \neq 0, \quad \imath = 1, 2, ...k$$

Slično kao i kod univarijatnog logističkog modela, za parametare $\beta_1, \beta_2, ..., \beta_k$ multivarijatnog logističkog modela kažemo da su statistički značajni ako se statistički značajno razlikuju od 0.

Najčešće korištena razina značajnosti za određivanje statističke značajnosti parametara je $\alpha = 5\%$. I to je upravo razina značajnosti koju ćemo koristiti prilikom obrade podataka. Ako je p-vrijednost $< \alpha$, tada odbacujemo $\mathbf{H}_0$ u korist alternative $\mathbf{H}_1$ te kažemo da je nezavisna varijabla statistički značajna na razini značajnosti α. Ako je p-vrijednost $> \alpha$, tada ne možemo odbaciti $\mathbf{H}_0$ u korist alternative $\mathbf{H}_1$ te kažemo da nezavisna varijabla nije statistički značajna na razini značajnosti α. [5]

1.9 Konvergencija i separabilnost

Kao što je već rečeno procjena ML metodom je iterativni proces uspješnih aproksimacija. Kada je promjena u procjenama parametara između 2 iteracije dovoljno malena iteracije prestaju i kažemo da je algoritam iskonvergirao. Uglavnom ovaj proces prolazi relativno glatko, no ponekad procedura ne konvergira. Razlog tome mogu biti potpuna separabilnost i kvazi-potpuna separabilnost. [3] [8]

Potpuna separabilnost

Potpuna separacija događa se kada zavisna varijabla savršeno odvaja nezavisnu varijablu ili kombinaciju nezavisnih varijabli. Albert i Anderson (1984) definirali su to kao "postoji vektor koji točno alocira sve opservacije njihovoj grupi". Pogledajmo na primjeru

y	x_1	x_2
0	1	3
0	2	2
0	3	-1
0	3	-1
1	5	2
1	6	4
1	10	1
1	11	0

U gornjoj tablici y je zavisna varijabla, a x_1i x_2 nezavisne. Vidimo da opservacije kod kojih je y=0 sve imaju vrijednosti $x_1 \leq 3$, dok opservacije kod kojih je $y = 1$ sve imaju vrijednosti $x_1 > 3$. Drugim riječima y savršeno razdvaja x_1. Kada pokušamo koristiti logističku regresiju na takvim podatcima, ML za β_1 (koeficijent uz x_1) ne postoji. Preciznije, što je β_1 veći, veći je i likelihood, odnosno β_1 će biti ∞. [8]

Kvazi-potpuna separabilnost

Kvazi-potpuna separabilnost događa se kada zavisna varijabla razdvaja nezavisnu varijablu ili kombinaciju nezavisnih varijabli do određene razine.

y	x_1	x_2
0	1	3
0	2	0
0	3	-1
0	3	4
1	3	1
1	4	0
1	5	2
1	6	7
1	10	3
1	11	4

Zavisna varijabla y odvaja nezavisnu varijablu x_1 poprilično dobro, osim kada je $x_1 = 3$. Kao i kod potpune separabilnosti, kada koristimo logističku regresiju na takvim podacima ML od β_1 ne postoji. [8]

1.10 SAS procedure

Backward, Forwar i Stepwise

Postupno uključivanje (*engl. Forward*)
Metoda počinje samo sa interceptom, i zatim u svakom koraku dodaje varijablu koja najbolje maksimizira fit modela. Proces završava kada se više ne može postići značajno poboljšanje dodavanjem varijable.
Postupno isključivanje (*engl. Backward*)
Metoda počinje s punim modelom koji uključuje sve nezavisne varijable. Zatim eliminira iz modela varijablu po varijablu koja ima najmanji doprinos adekvatnosti modela.
Postupno uključivanje i isključivanje (*engl. Stepwise*)
Ova metoda je kombinacija postupnog uključivanja i postupnog isključivanja. Počinje kao

metoda postupnog uključivanja, samo što varijabla koja je bila u modelu ne mora tamo i ostati.[5]

ROC krivulja

Definirajmo da se dogadaj dogodio s 1, a da se nije dogodio s 0. Za par opservacija s različitim odgovorima, kažemo da su usklađene (*engl. concordant*) ako opservacija koja ima više rangirani odgovor (npr. 2 "događaj se ne dogodi"), ima nižu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom (npr. 1 "događaj se dogodi"). Za par opservacija s različtim odgovorima, kažemo da su neusklađene (*engl. discordant*) ako opservacija koja ima više rangirani odgovor, ima višu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom. Ako par opservacija nije ni usklađen ni neusklađen, kažemo da imaju jednak odgovor (*engl. tie*).

$$c = \frac{nc + 0.5(t - nc - nd)}{t}$$

gdje je t broj parova s različitim vrijednostima odgovora, nc broj concordant parova, a nd broj discordant parova. c zapravo predstavlja površinu ispod ROC krivulje (površinu ispod ROC krivulje ponekad nazivamo i c-statistika). ROC krivulja (*engl. receiver operating charasteristic curve*) je često korištena metoda za ispitivanje valjanosti dijagnostičkog testa. Valjanost dijagnostičkog testa je složeni pokazatelj i ima dvije komponente: osjetljivost i specifičnost. Osjetljivost testa je proporcija dobro detektiranih bolesnih osoba od sveukupnog broja bolesnih, a specifičnost testa je proporcija zdravih osoba koje su dobro detektirane kao zdrave, od ukupnog broja zdravih osoba. ROC krivulja prikazuje odnos proporcija lažno pozitivnih (1-specifičnost) i stvarno pozitivnih (osjetljivost). Površina ispod ROC krivulje (odnosno c) je mjera točnosti testa.

Mjere točnosti testa	
Izvrstan test	0,9 - 1
Dobar test	0,8 - 0,9
Osrednji test	0,7 - 0,8
Slabiji test	0,6 - 0,7
Test bez uspjeha	0,5 - 0,6

Npr, ako u proceduri dobijemo da je c=0.897, to tumačimo da je 89.7% bolesti objašnjeno modelom.[5], [4]

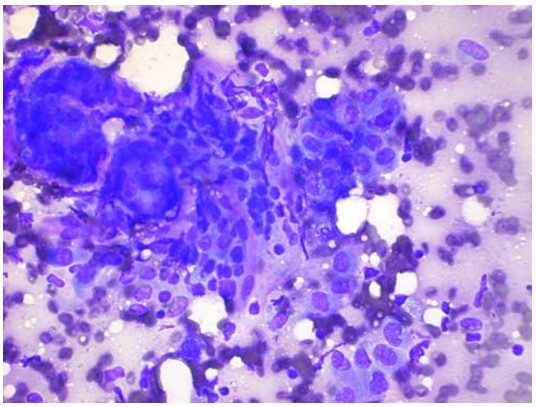
Poglavlje 2

Modeliranje malignosti tumora dojke logističkom regresijom

2.1 Opis podataka

Termin tumor odnosi se na grupu bolesti kod kojih stanice abnormalno rastu. Također tumor možemo definirati kao ”maligna neoplazma”. Neoplazma znači novi rast. Masa tkiva koja se formira kao rezultat abnormalnog, pretjeranog, nekordiniranog, autonomnog i besmislenog povećanja broja stanica naziva se tumor. Grana znanosti koja se bavi proučavanjem tumora naziva se onkologija (*lat.* *oncos*=tumor, *logos*=učenje). Tumori mogu biti benigni ili maligni. Benigni tumori ne napadaju i ne šire se na druge dijelove tijela, niti uništavaju tkivo unutar kojeg se stvaraju. Maligni tumori su suprotno od benignih. Oni se šire na druge dijelove tijela, te napadaju i uništavaju tkivo koje ih okružuje. Podatci koje ćemo obraditi sastoje se od 699 biopsijskih uzoraka tumora dojke. Podatci su preuzeti sa stranice https ://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin + (Diagnostic), 2017. godine, a prikupio ih je Dr. William H. Wolberg (University of Wisconsin Hospitals, Madison, USA). Dobiveni su iz digitalizirane slike tkiva uzetog pri punkcijsko-aspiracijskoj biopsiji dojke. Prikupljani su u razdoblju od 1989-1991. godine u nekoliko faza.

Faze prikupljanja podataka	
Grupa 1	367 uzoraka (siječanj 1989)
Grupa 2	70 uzoraka (listopad 1989)
Grupa 3	31 uzorak (veljača 1990)
Grupa 4	17 uzoraka (travanj 1990)
Grupa 5	48 uzoraka (kolovoz 1990)
Grupa 6	49 uzoraka (siječanj 1991)
Grupa 7	31 uzorak (lipanj 1991)
Grupa 8	86 uzoraka(studeni 1991)



Slika 2.1: Prikaz tkiva uzetog pri punkcijsko-aspiracijskoj biopsiji

Svaki uzorak sastoji se od 11 varijabli. A to su : Identifikacijski broj uzorka (*engl. Sample code number*), Debljina grumena (*engl. Clump Thickness*), Uniformnost veličine stanice (*engl. Uniformity of Cell Size*), Uniformnost staničnog oblika (*engl. Uniformity of Cell Shape*), Marginalna adhezija (*engl. Marginal Adhesion*), Veličina epitelne stanice (*engl. Single Epithelial Cell Size*), Gole jezgre (*engl. Bare Nuclei*), Normalan kromatin (*engl. Bland Chromatin*), Normalne jezgrice (*engl. Normal Nucleoli*), Mitoza (*engl. Mitoses*) i Razred (*engl. Class*).

- Identifikacijski broj uzorka: oznaka uzorka pomoću koje raspoznajemo o kojem se uzorku radi
- Debljina grumena: Benigne stanice se uglavnom grupiraju u jednom sloju, dok se maligne stanice često grupiraju u više slojeva.

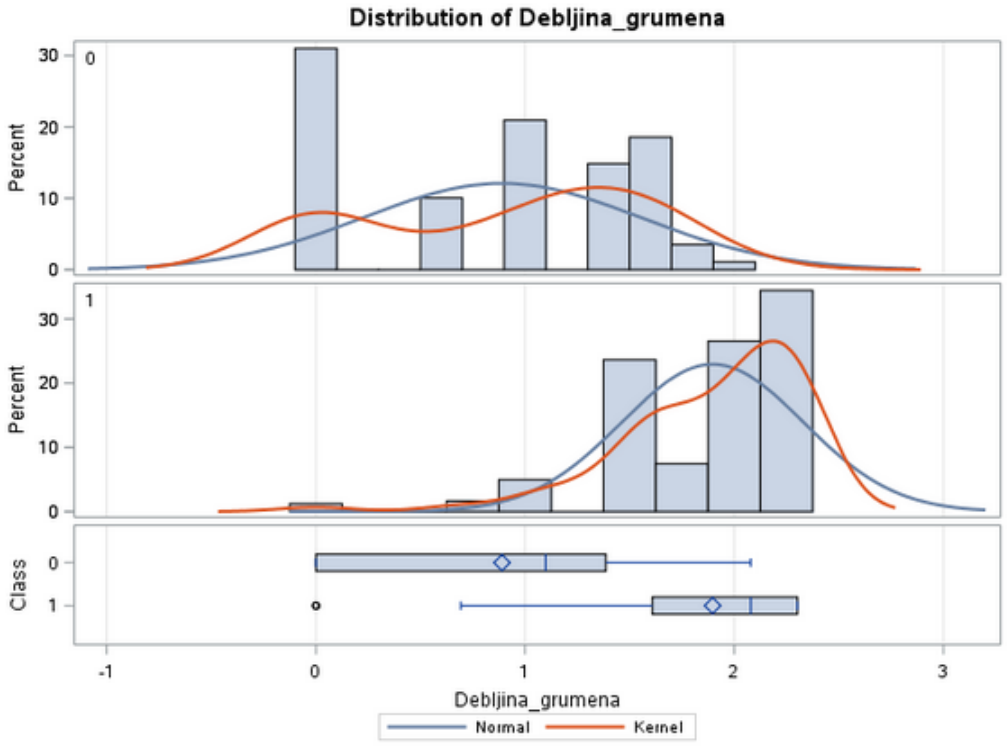
- Uniformnost veličine stanice/Uniformnost staničnog oblika: Stanice raka imaju tendenciju da variraju u veličini i obliku. Iz tog razloga su svojstva veličine i oblika stanice važna u determiniranju malignosti.
- Marginalna adhezija: Normalne stanice se drže skupa, dok stanice raka gube tu sposobnost. Iz tog razloga je gubitak adhezije znak malignosti.
- Veličina epitelne stanice: Povećane epitelne stanice mogu biti znak malignosti.
- Gole jezgre: Termin koji se koristi za jezgre koje nisu okružene citoplazmom (ostatak stanice). Uglavnom ih viđamo kod benignih tumora.
- Normalan kromatin: Opisuje uniformnu teksturu jezgre koju često viđamo kod benignih stanica. Kromatin je uglavnom povećan kod malignih stanica.
- Normalne jezgrice: Jezgrice su malene strukture koje se nalaze u jezgri. Kod normalnih stanica jezgrice su uglavnom vrlo malene, ako su uopće vidljive. Kod malignih stanica jezgrice su istaknutije.
- Mitoza- dijeljenje jezgre: Proces se naziva podjela stanice i sastoji se od 4 faze. To su: profaza, prometafaza, metafaza, anafaza i telofaza. Abnormalno velika podjela stanica je često znak malignosti.
- Varijabla Razred ima 2 vrijednosti i govori nam je li uzorak maligni (1) ili benigni (0).

Ukupno imamo 458 (65.5%) benignih i 241 (34.5%) malignih uzoraka. Na temelju mišljenja citologa svakoj od varijabli dodijeljena je vrijednost od 1-10. Da bi bilo jasnije prikazati ćemo to u tablici.

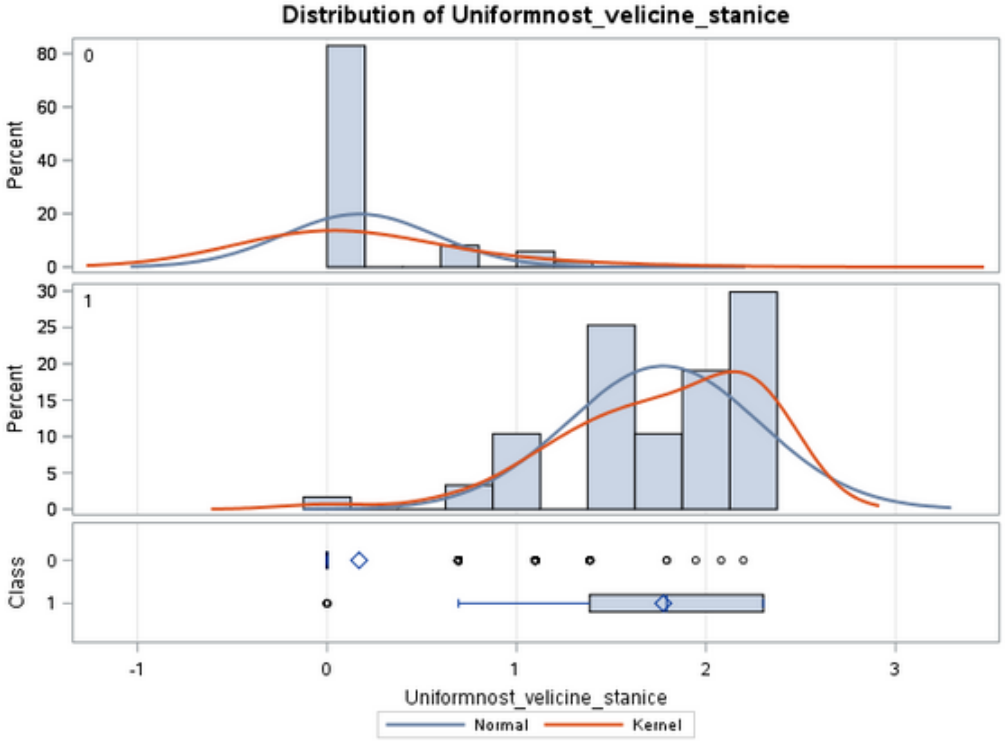
Identifikacijski broj uzorka	identifikacijski broj
Debljina grumena	1-10
Uniformnost veličine stanice	1-10
Uniformnost staničnog oblika	1-10
Marginalna adhezija	1-10
Veličina epitelne stanice	1-10
Gole jezgre	1-10
Normalan kromatin	1-10
Normalne jezgrive	1-10
Mitoza	1-10
Razred	0-benigni, 1-maligni

Tablica 2.1: Deskriptivna statistika za analizu varijabli (ispis iz SASa)

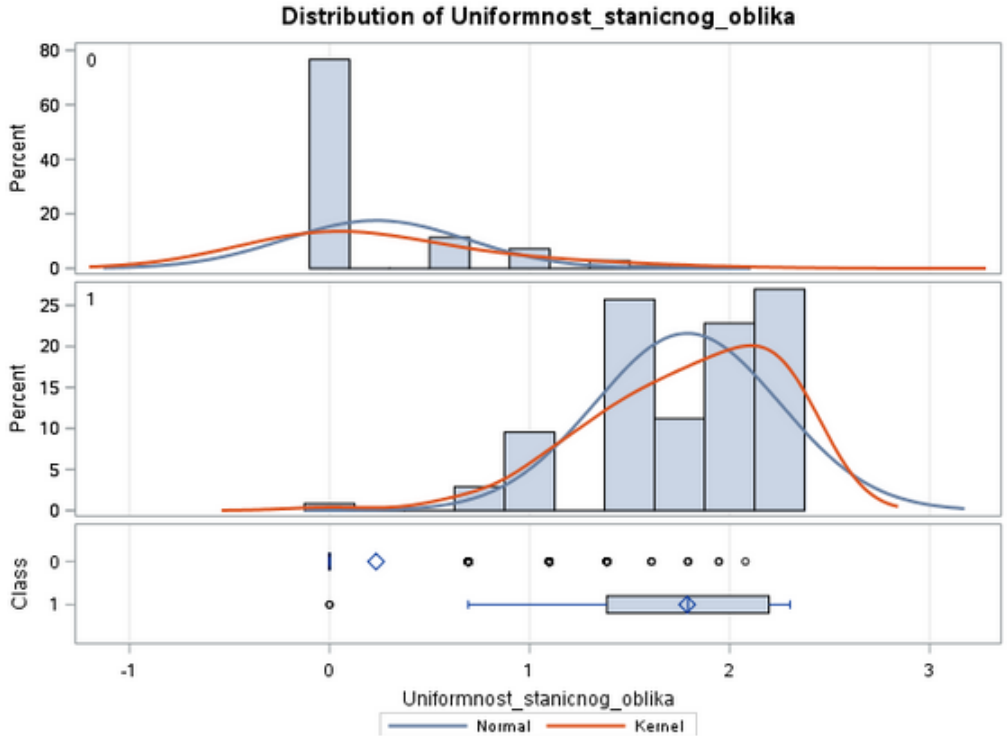
Varijabla	N	Mean	Median	Std Dev	Minimum	Maximum
Debljina grumena	699	4,41774	4	2,815741	1	10
Uniformnost veličine stanice	699	3,134478	1	3,051459	1	10
Uniformnost staničnog oblika	699	3,207439	1	2,971913	1	10
Marginalna adhezija	699	2,806867	1	2,855379	1	10
Veličina epitelne stanice	699	3,216023	2	2,2143	1	10
Gole jezgre	683	3,544656	1	3,643857	1	10
Normalan kromatin	699	3,437768	3	2,438364	1	10
Normalne jezgrice	699	2,866953	1	3,053634	1	10
Mitoza	699	1,589413	1	1,715078	1	10



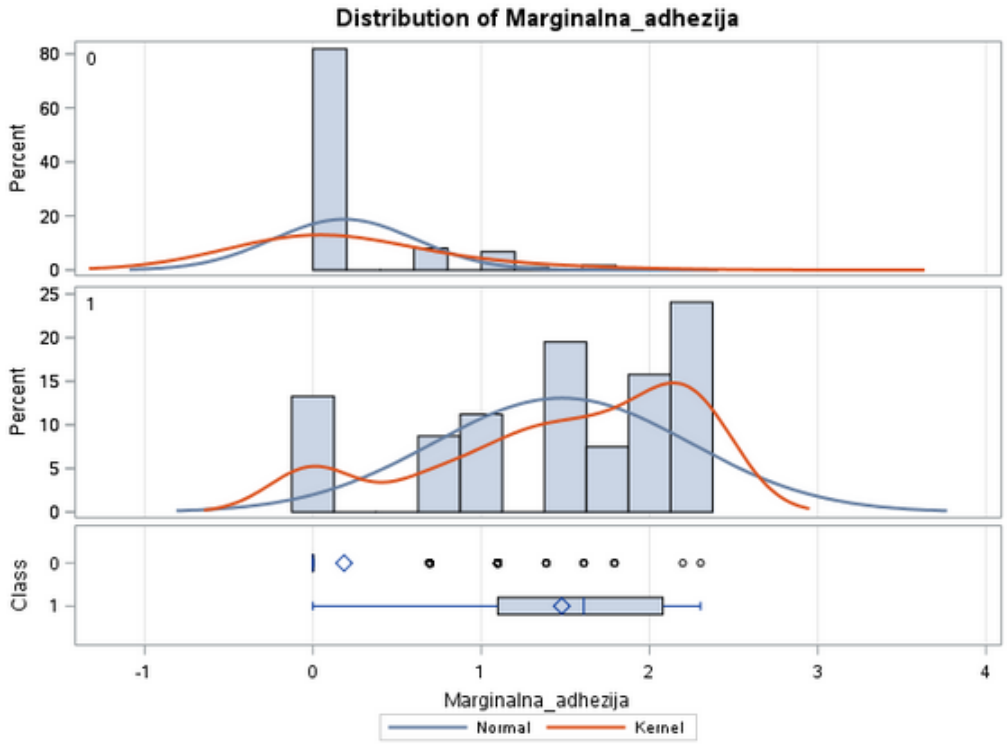
Slika 2.2: Distribucija logaritmiranih podataka za Debljinu grumena (ispis iz SASa)



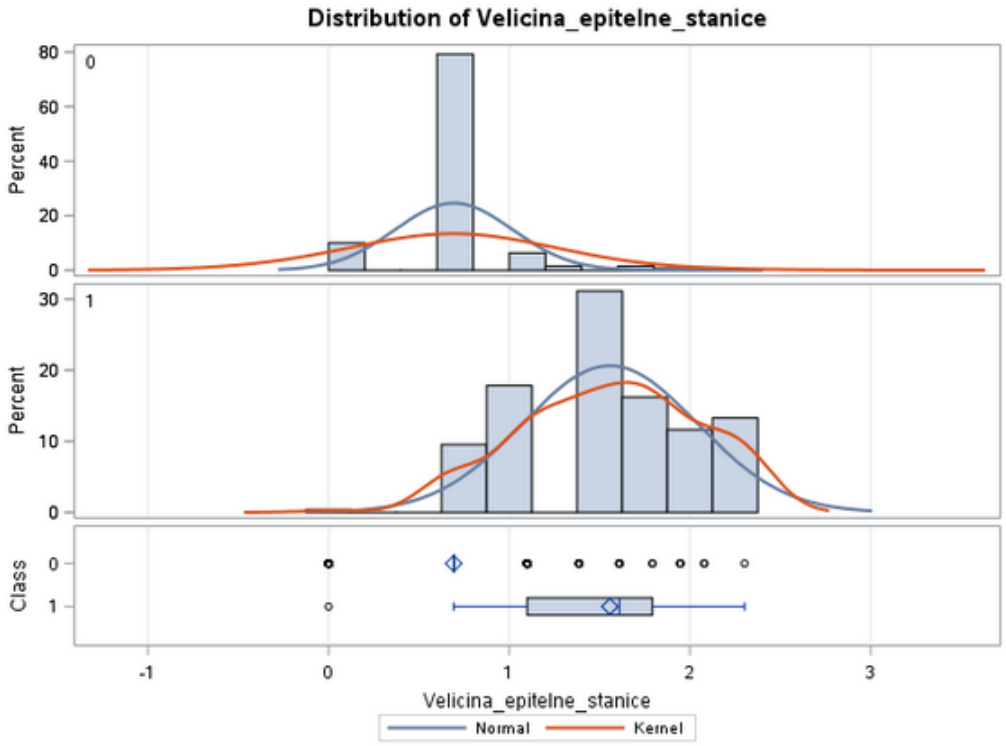
Slika 2.3: Distribucija logaritmiranih podataka za Uniformnost veličine stanice (ispis iz SASa)



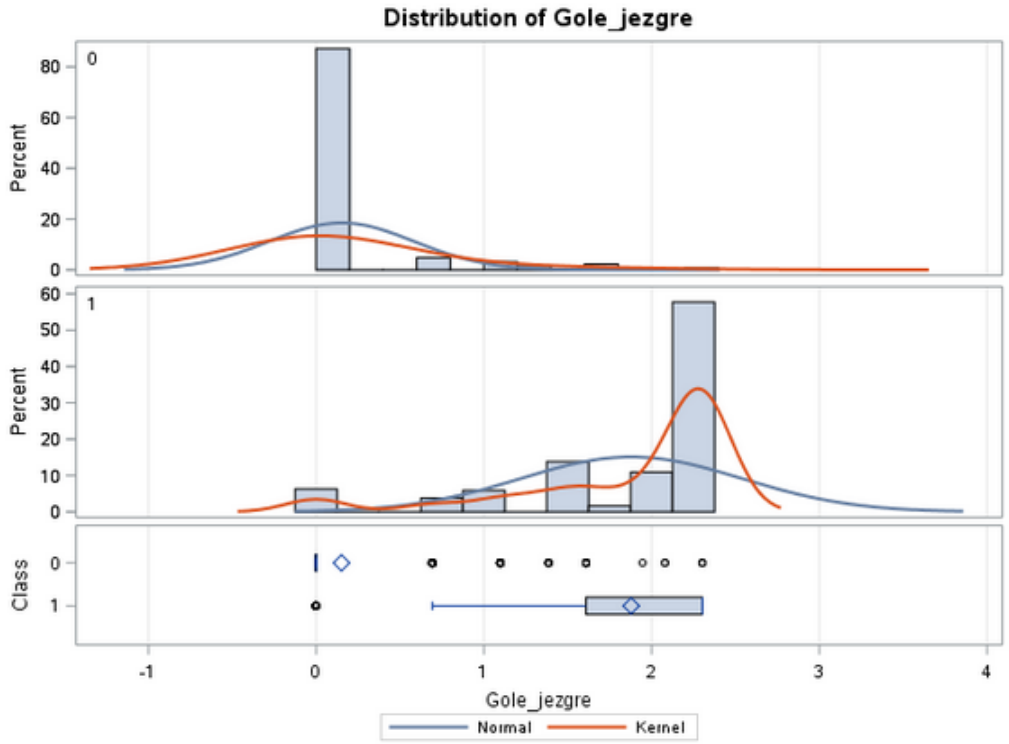
Slika 2.4: Distribucija logaritmiranih podataka za Uniformnost staničnog oblika (ispis iz SASa)



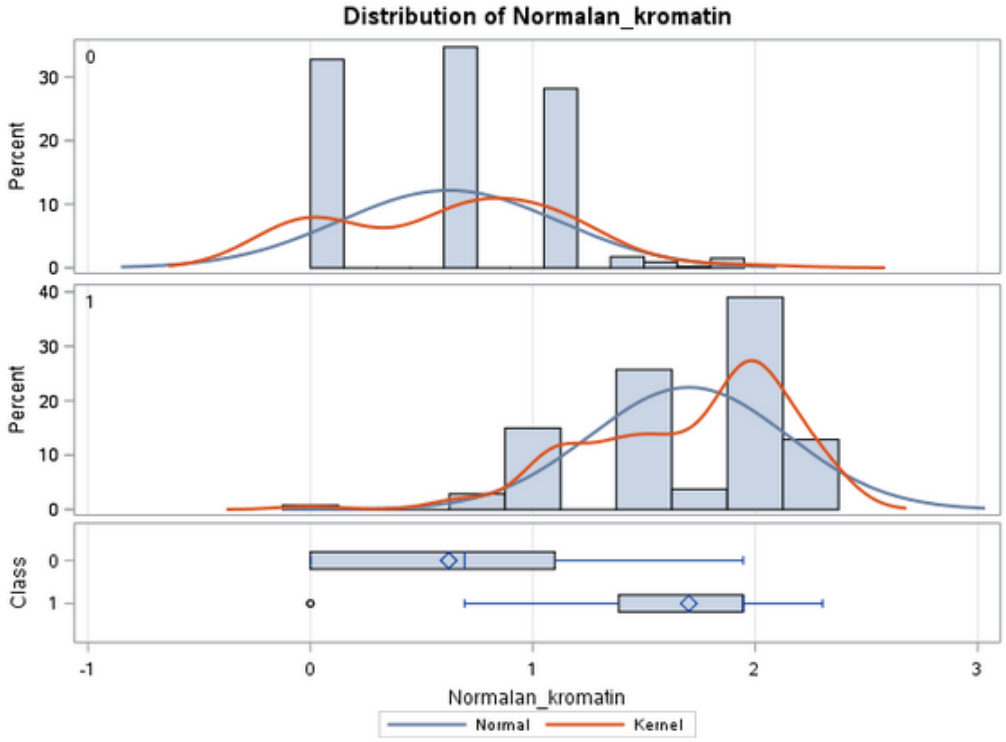
Slika 2.5: Distribucija logaritmiranih podataka za Marginalnu adheziju (ispis iz SASa)



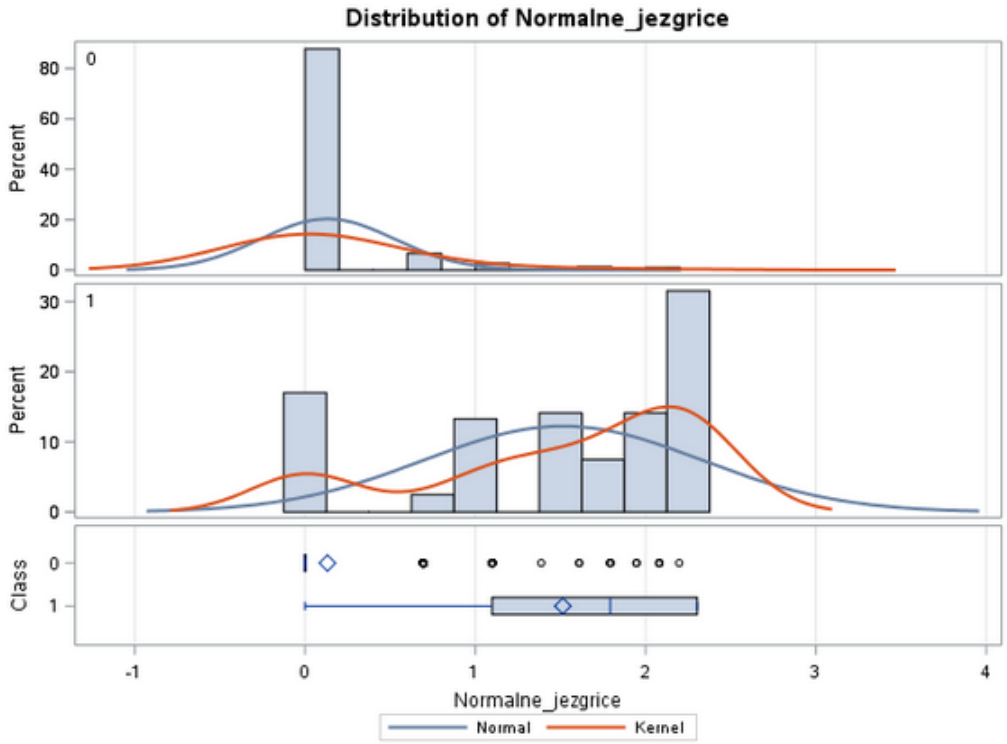
Slika 2.6: Distribucija logaritmiranih podataka za Veličinu epitelne stanice (ispis iz SASa)



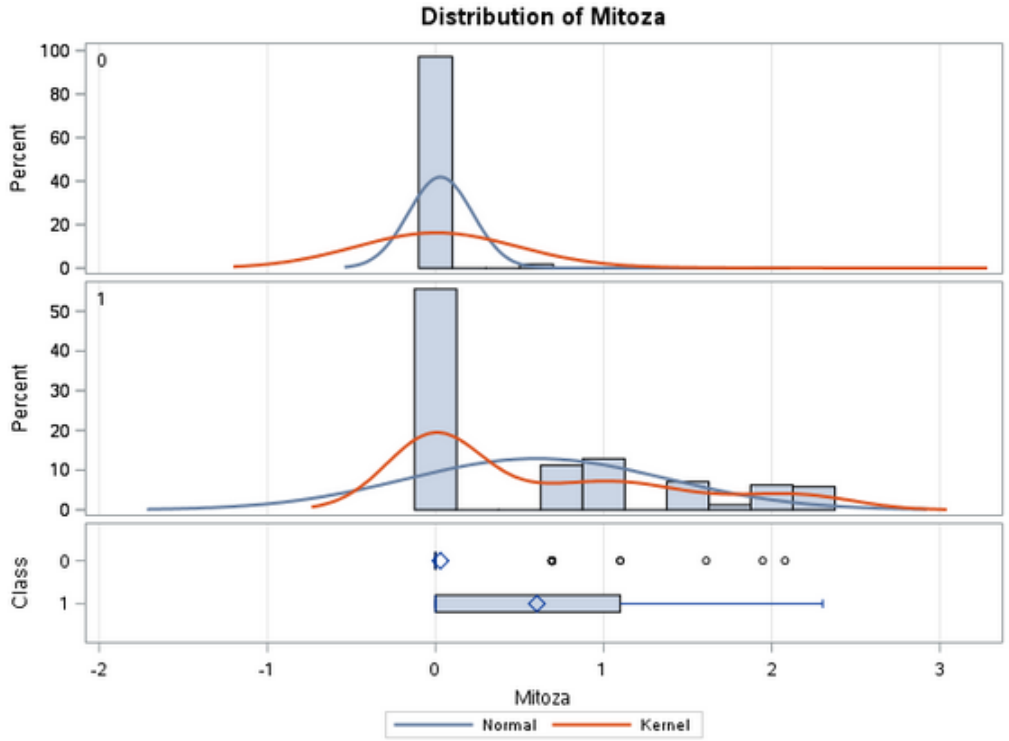
Slika 2.7: Distribucija logaritmiranih podataka za Gole jezgre (ispis iz SASa)



Slika 2.8: Distribucija logaritmiranih podataka za Normalan kromatin (ispis iz SASa)



Slika 2.9: Distribucija logaritmiranih podataka za Normalne jezgrice (ispis iz SASa)



Slika 2.10: Distribucija logaritmiranih podataka za Mitozu (ispis iz SASa)

Prema slikama 2.2 do 2.10 vidimo da podatci nisu normalno distribuirani. Također vidimo da se distribucija malignih i benignih tumora značajno razlikuje za sve varijable.[7],[1],[2]

2.2 Univarijatna logistička regresija

Za svaku od nezavisnih varijabli ćemo provesti univarijatnu logističku regresiju. Na taj način ćemo odrediti statističku značajnost svake varijable zasebno.

Procedura je iskonvergirala za sve varijable. Osnovne dobivene vrijednosti prikazane su na donjim tablicama.

Tablica 2.2: Rezultati analize univarijatnih logističkih modela (ispis iz SASa)

Varijable	DF	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vrijednost
Debljina grumena	1	900,527	464,054	436,4735	<.0001
Uniformnost veličine stanice	1	900,527	275,554	624,9731	<.0001
Uniformnost staničnog oblika	1	900,527	284,246	616,281	<.0001
Marginalna adhezija	1	900,527	492,554	407,9732	<.0001
Veličina epitelne stanice	1	900,527	481,706	418,8217	<.0001
Gole jezgre	1	884,35	340,628	543,7224	<.0001
Normalan kromatin	1	900,527	401,345	499,1824	<.0001
Normalne jezgrice	1	900,527	488,509	412,0182	<.0001
Mitoza	1	902,527	735,081	169,446	<.0001

Kriterij konvergencije je zadovoljen za sve varijable. Iz tablice 2.2 vidimo da su dobiveni modeli statistički značajni, budući da je $\chi^2 \geq 169,446$ za sve varijable, i p-vrijednost <0,0001 za sve varijable.

Tablica 2.3: Rezultati ML procjene parametara za univarijatne logističke modele (ispis iz SASa)

Varijable	DF	Procjena	Standardna greška	Wald χ^2	p-vrijednost
Intercept	1	-5,1602	0,3779	186,4049	<.0001
Debljina grumena	1	0,9355	0,0738	160,8118	<.0001
Intercept	1	-4,9602	0,36	189,8772	<.0001
Uniformnost veličine stanice	1	1,4887	0,121	151,2705	<.0001
Intercept	1	-5,0618	0,3703	186,867	<.0001
Uniformnost staničnog oblika	1	1,4068	0,1125	156,3005	<.0001
Intercept	1	-3,125	0,2079	225,9375	<.0001
Marginalna adhezija	1	0,9658	0,0813	140,9726	<.0001
Intercept	1	-4,7725	0,3258	214,5706	<.0001
Veličina epitelne stanice	1	1,3594	0,1107	150,8968	<.0001
Intercept	1	-3,5221	0,232	230,3911	<.0001
Gole jezgre	1	0,8593	0,0709	146,8161	<.0001
Intercept	1	-5,1932	0,3753	191,4302	<.0001
Normalan kromatin	1	1,3228	0,111	142,0265	<.0001
Intercept	1	-2,8853	0,1879	235,8525	<.0001
Normalne jezgrice	1	0,8592	0,0745	133,0785	<.0001
Intercept	1	-2,4656	0,2277	117,2134	<.0001
Mitoza	1	1,349	0,1792	56,6721	<.0001

Iz tablice 2.3 vidimo da je p-vrijednost <0,0001 za sve varijable, pa možemo zaključiti da su sve varijable statistički značajne na razini značajnosti 5%.

Također, iz tablice 2.3 vidimo da jednadžbe dobivenih modela glase:
Za Debljinu grumena:

$$logit(p) = -5,1602 + 0,9355 \cdot \text{Debljina grumena}$$

Za Uniformnost veličine stanice:

$$logit(p) = -4,9602 + 1,4887 \cdot \text{Uniformnost veličine stanice}$$

Za Uniformnost staničnog oblika:

$$logit(p) = -5,0618 + 1,4068 \cdot \text{Uniformnost staničnog oblika}$$

Za Veličinu epitelne stanice:

$$logit(p) = -4,7725 + 1,3594 \cdot \text{Veličina epitelne stanice}$$

Za Gole jezgre:

$$logit(p) = -3,5221 + 0,8593 \cdot \text{Gole jezgre}$$

Za Normalan kromatin:

$$logit(p) = -5,1932 + 1,3228 \cdot \text{Normalan kromatin}$$

Za Normalne jezgrice:

$$logit(p) = -2,8853 + 0,8592 \cdot \text{Normalne jezgrice}$$

Za Mitozu:

$$logit(p) = -2,4656 + 1,349 \cdot \text{Mitoza}$$

Tablica 2.4: Procjene omjera šansi za univarijatne logističke modele (ispis iz SASa)

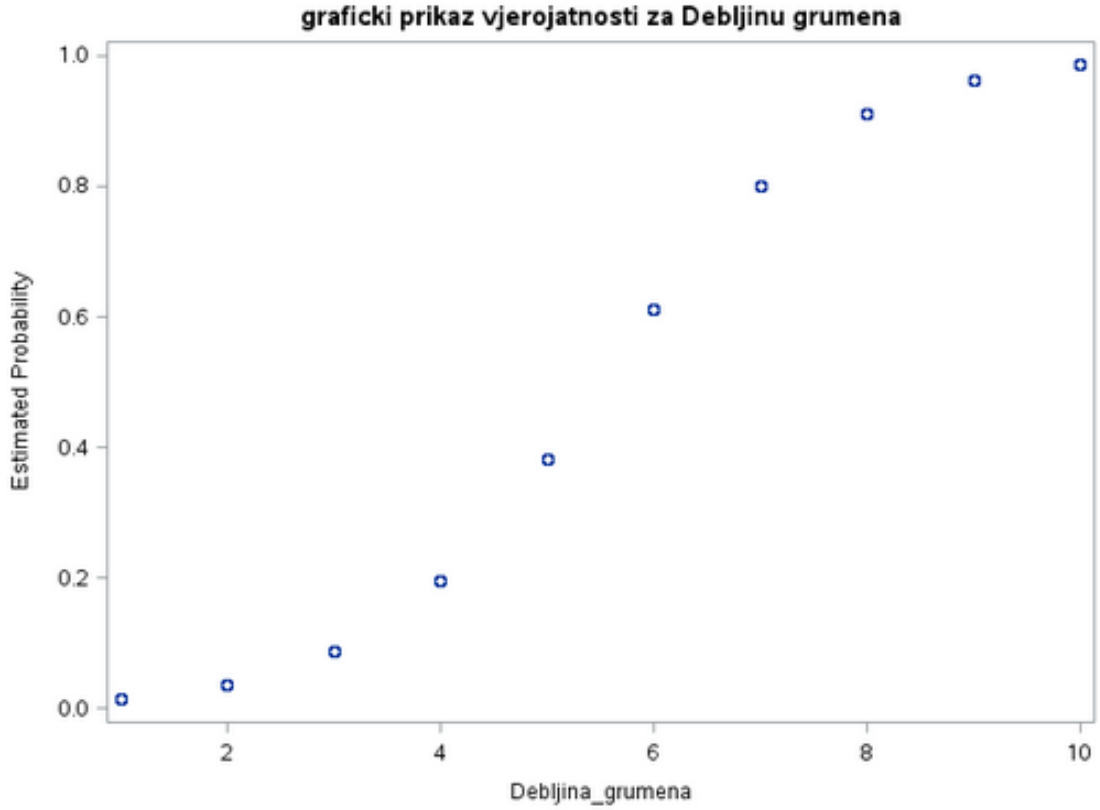
Varijable	Procjena OR	95% pouzdani interval
Debljina grumena	2,548	2,205 - 2,945
Uniformnost veličine stanice	4,431	3,495 - 5,618
Uniformnost staničnog oblika	4,083	3,275 - 5,090
Marginalna adhezija	2,627	2,240 - 3,081
Veličina epitelne stanice	3,894	3,134 - 4,837
Gole jezgre	2,362	2,055 - 2,714
Normalan kromatin	3,754	3,020 - 4,666
Normalne jezgrice	2,361	2,041 - 2,732
Mitoza	3,854	2,712 - 5,475

Još jedan način na koji možemo vidjeti koje varijable su statistički značajne je putem 95% pouzdanih intervala. Ako interval sadrži jedinicu, varijabla nije statistički značajna. Iz tablice 2.4 vidimo da niti jedan interval ne sadrži jedinicu, pa zaključujemo da su sve varijable statistički značajne.

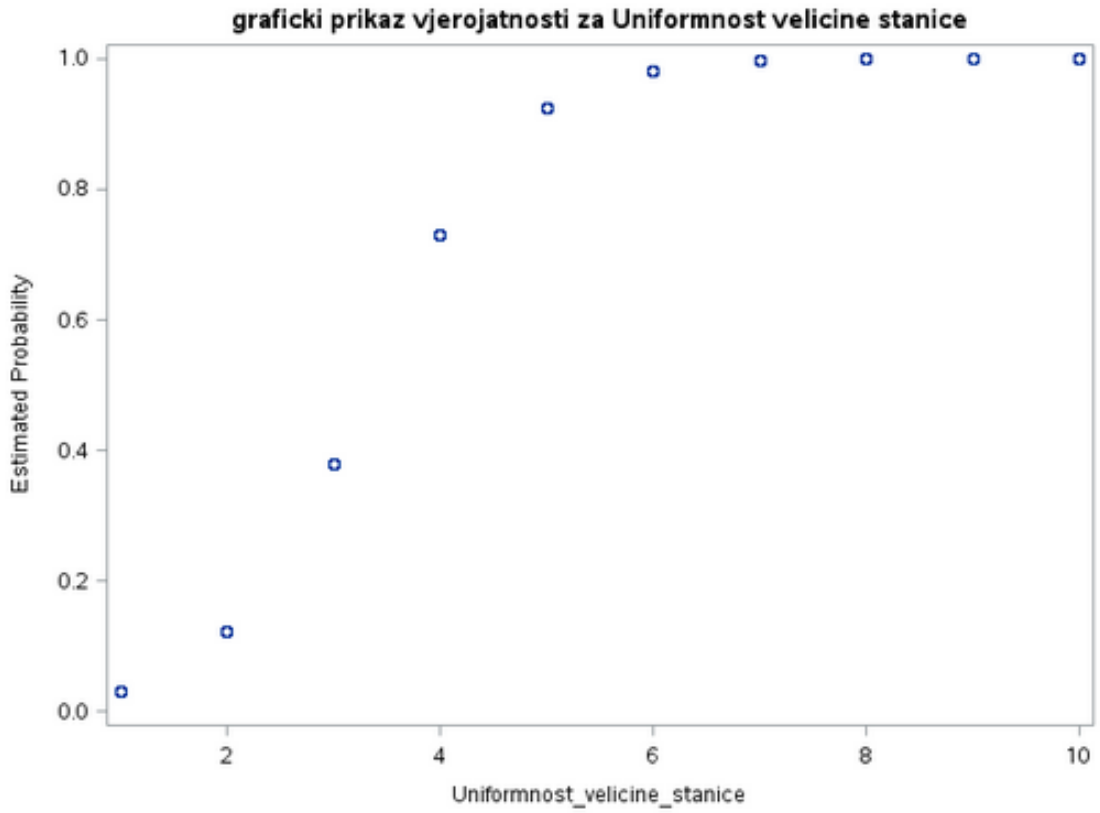
Vrijednosti u tablici 2.4 pod "Procjena OR" predstavljaju omjer šanse prelaska tumora iz benignog u maligni, uz prelazak nezavisne varijable iz niže u višu kategoriju. Te vrijednosti dobivene su relacijom $OR(x+1, x) = e^{\beta}$, gdje je β parametar modela. Stupac "Procjena OR" u tablici 2.4 tumačimo na sljedeći način:

- Povećanje Debljine grumena za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 2,548 puta.
- Povećanje Uniformnosti veličine stanice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 4,431 puta.
- Povećanje Uniformnosti oblika stanice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 4,083 puta.
- Povećanje Marginalne adhezije za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 2,627 puta.
- Povećanje Veličine epitelne stanice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 3,894 puta.
- Povećanje Gole jezgre za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 2,362 puta.
- Povećanje Normalnog kromatina za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 3,754 puta.
- Povećanje Normalne jezgrice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 2,361 puta.
- Povećanje Mitoze za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 3,854 puta.

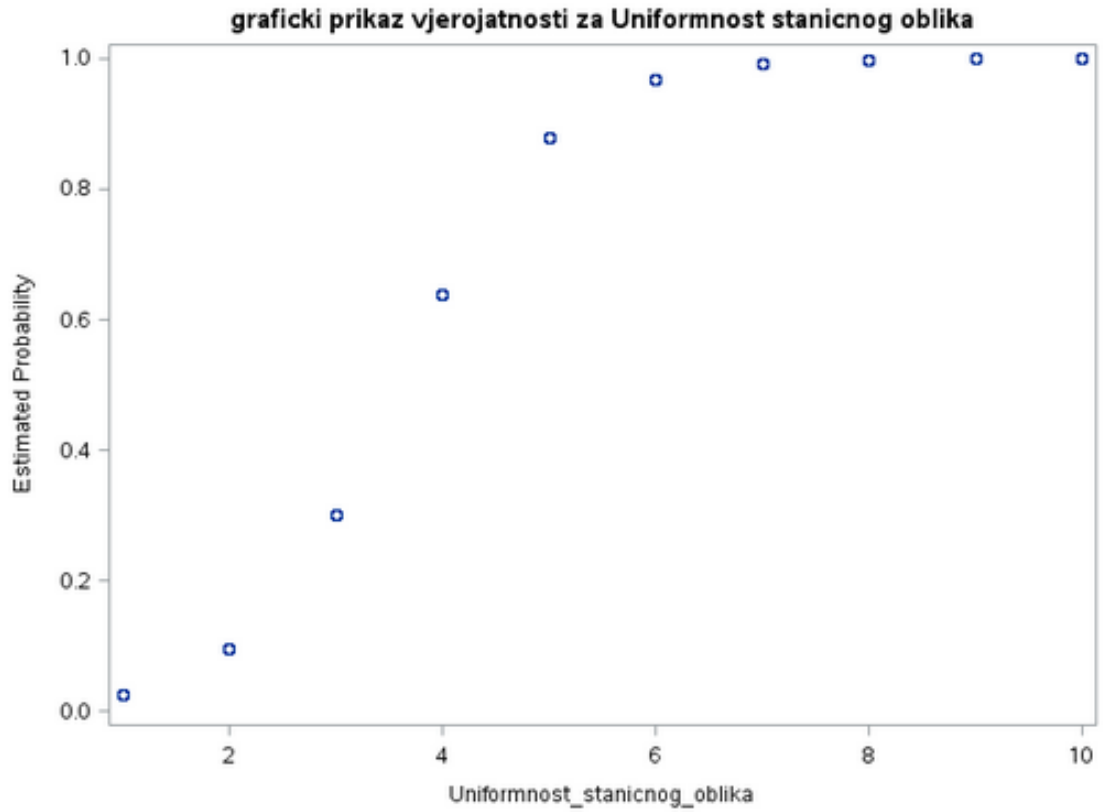
Također, značajnost varijable možemo promotriti i crtanjem vjerojatnosti. To nam omogućava da jasnije vidimo koliko promjena u varijabli utječe na malignost.



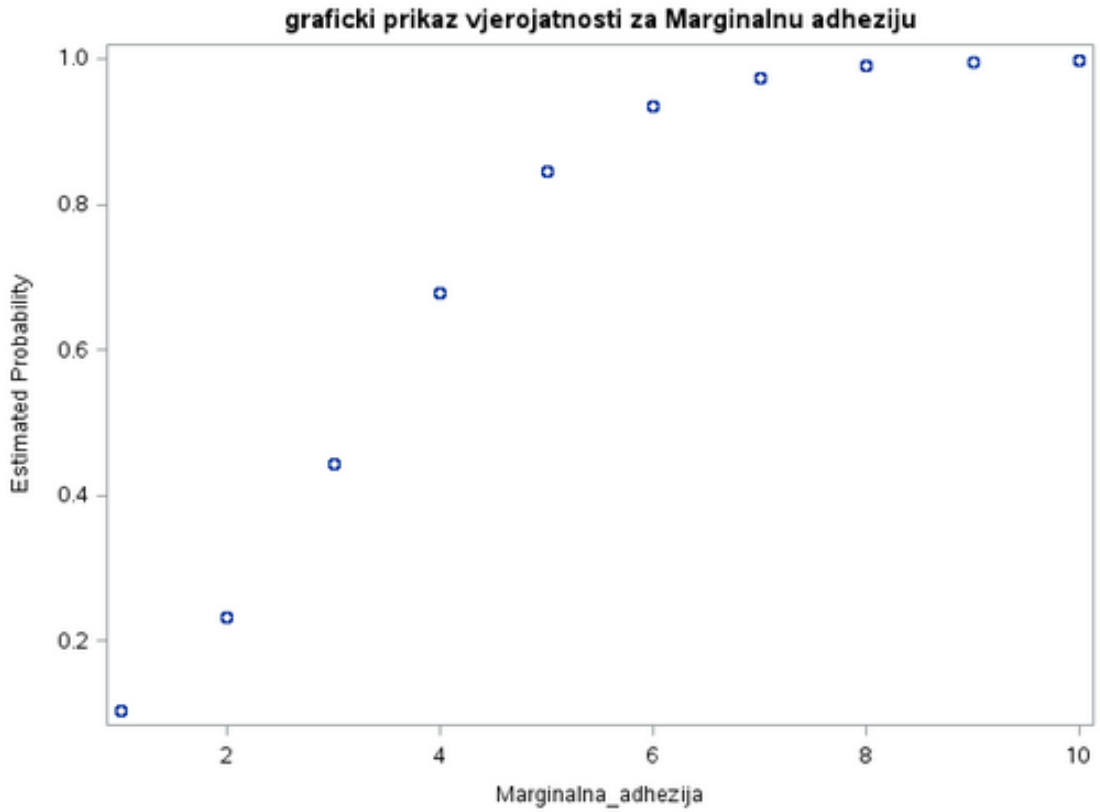
Slika 2.11: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Debljinu grumena (ispis iz SASa)



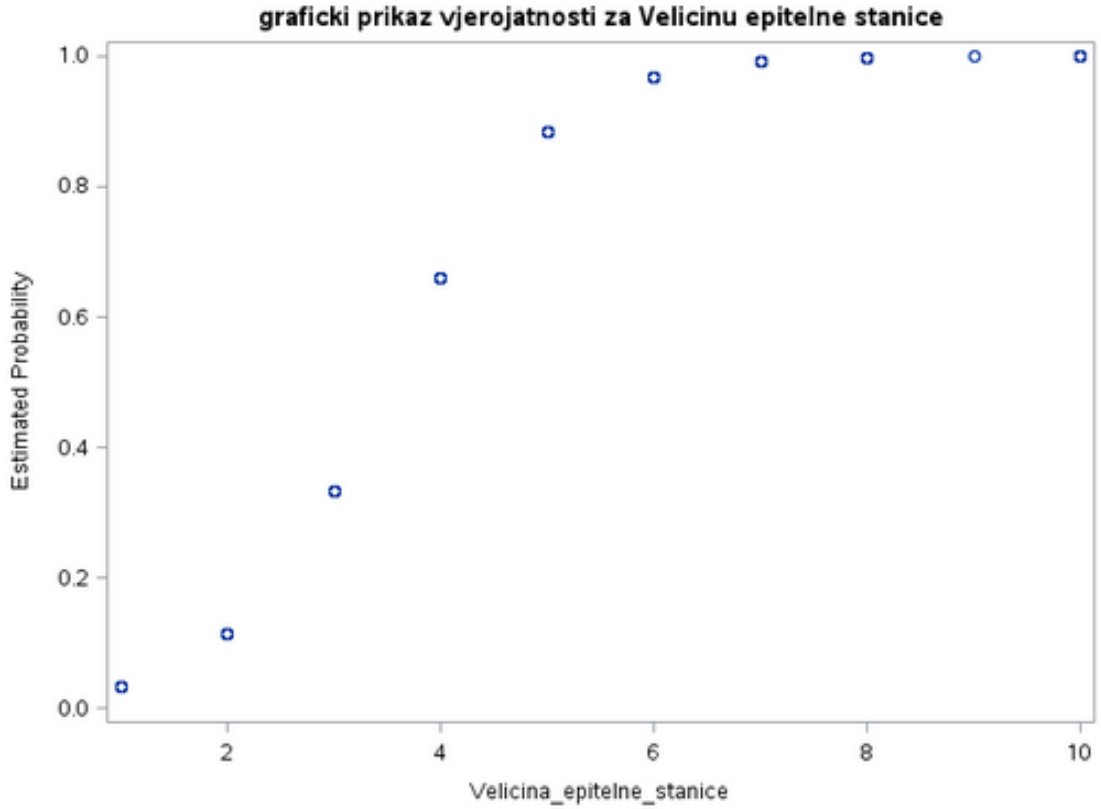
Slika 2.12: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Uniformnost veličine stanice (ispis iz SASa)



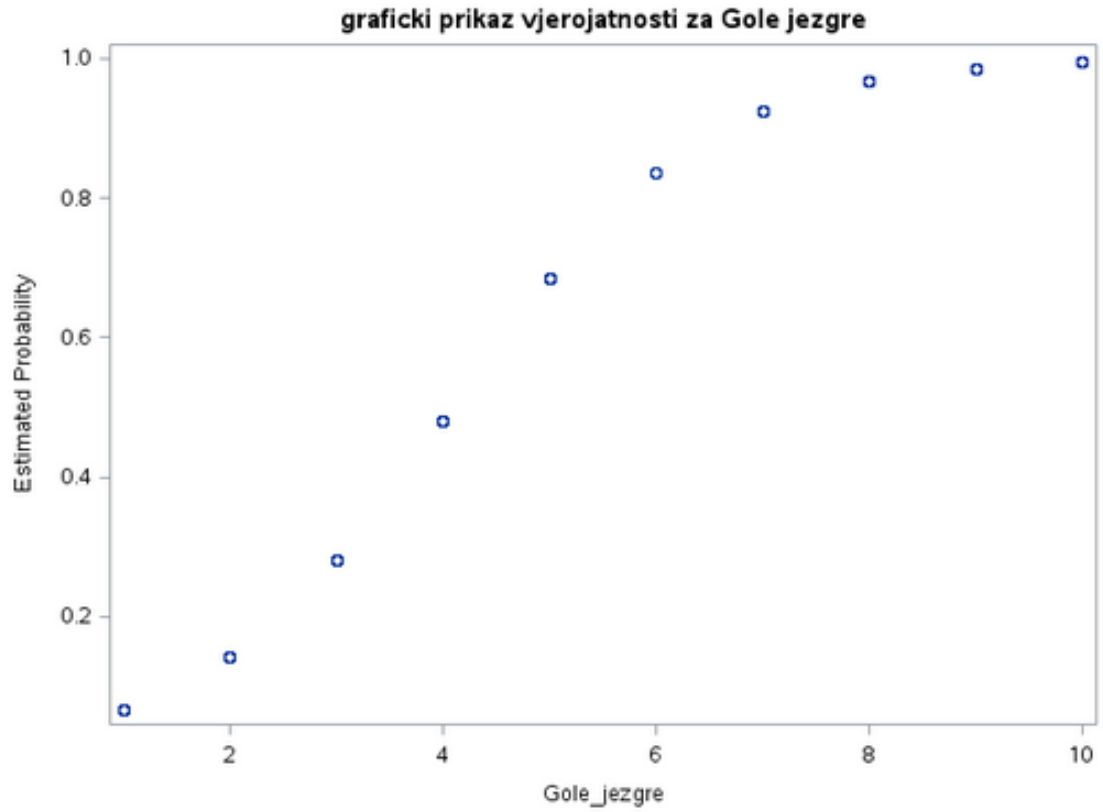
Slika 2.13: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Uniformnost staničnog oblika (ispis iz SASa)



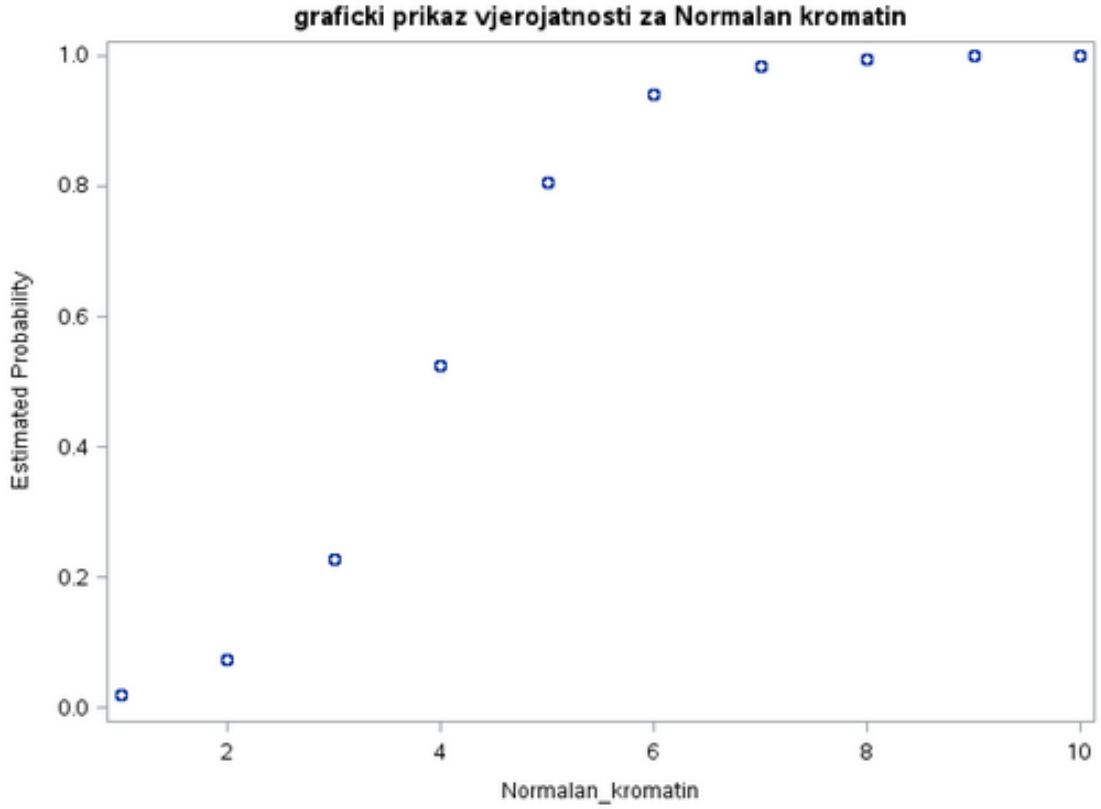
Slika 2.14: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Marginalnu adheziju (ispis iz SASa)



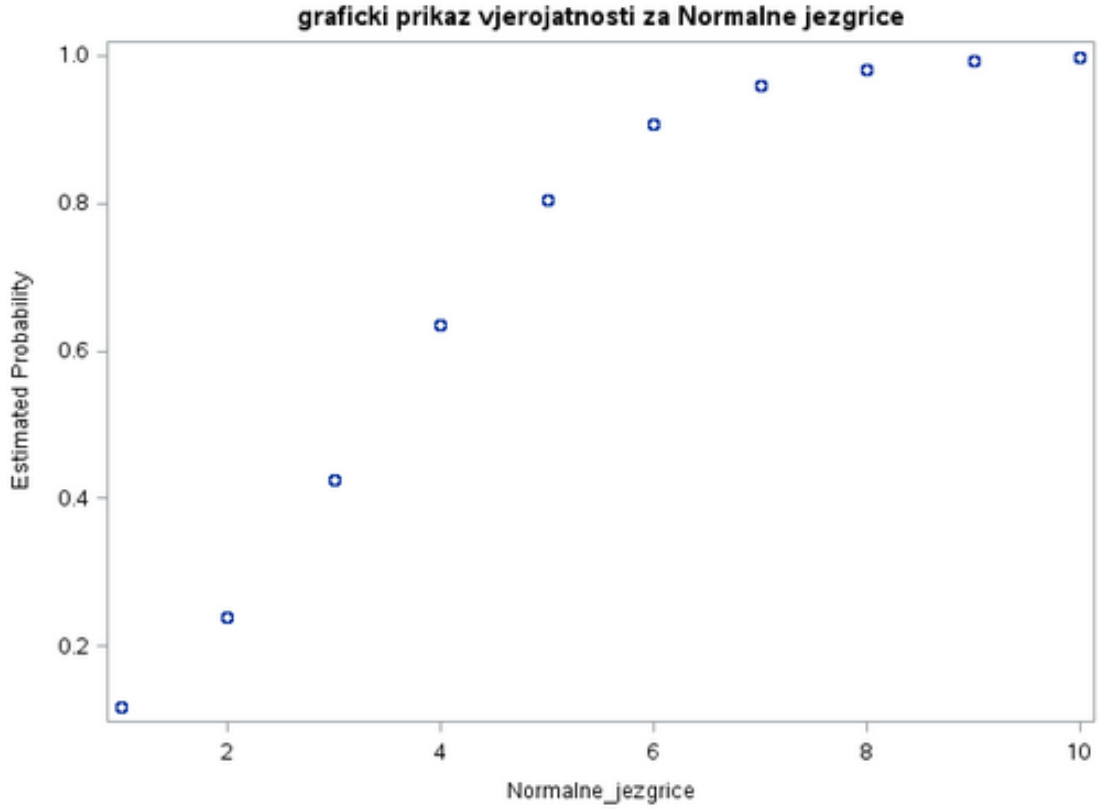
Slika 2.15: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Veličinu epitelne stanice (ispis iz SASa)



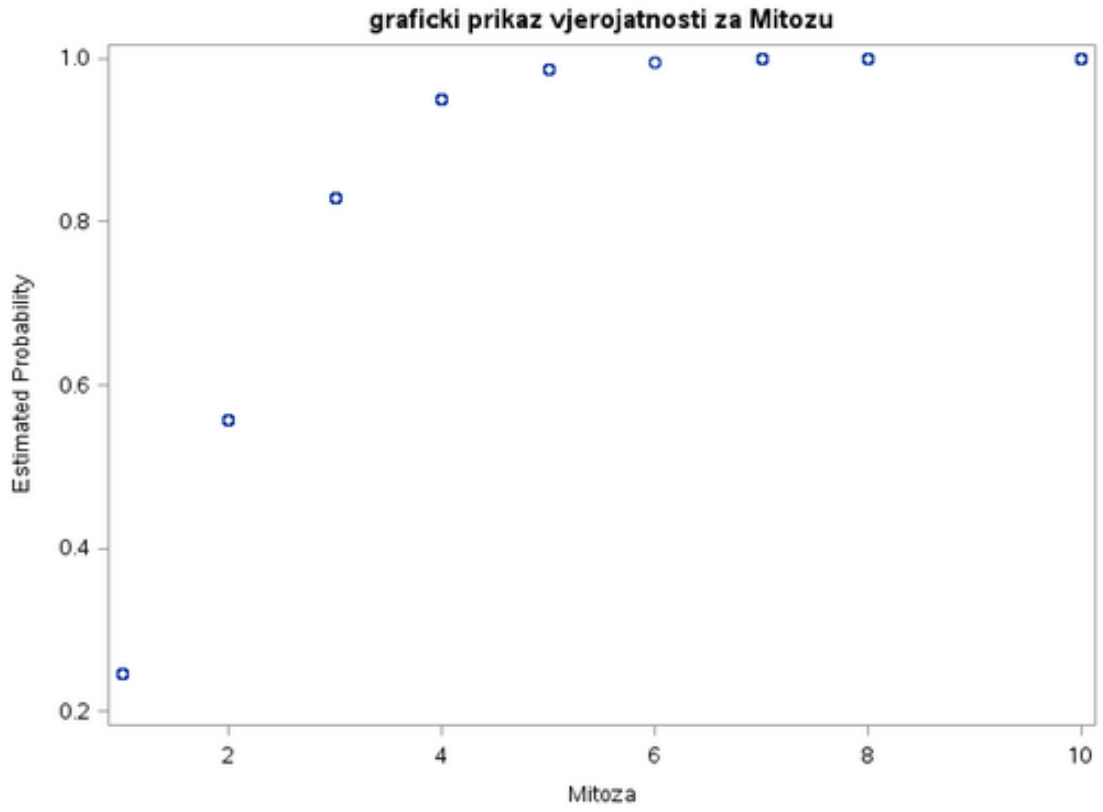
Slika 2.16: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Gole jezgre (ispis iz SASa)



Slika 2.17: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Normalan kromatin (ispis iz SASa)



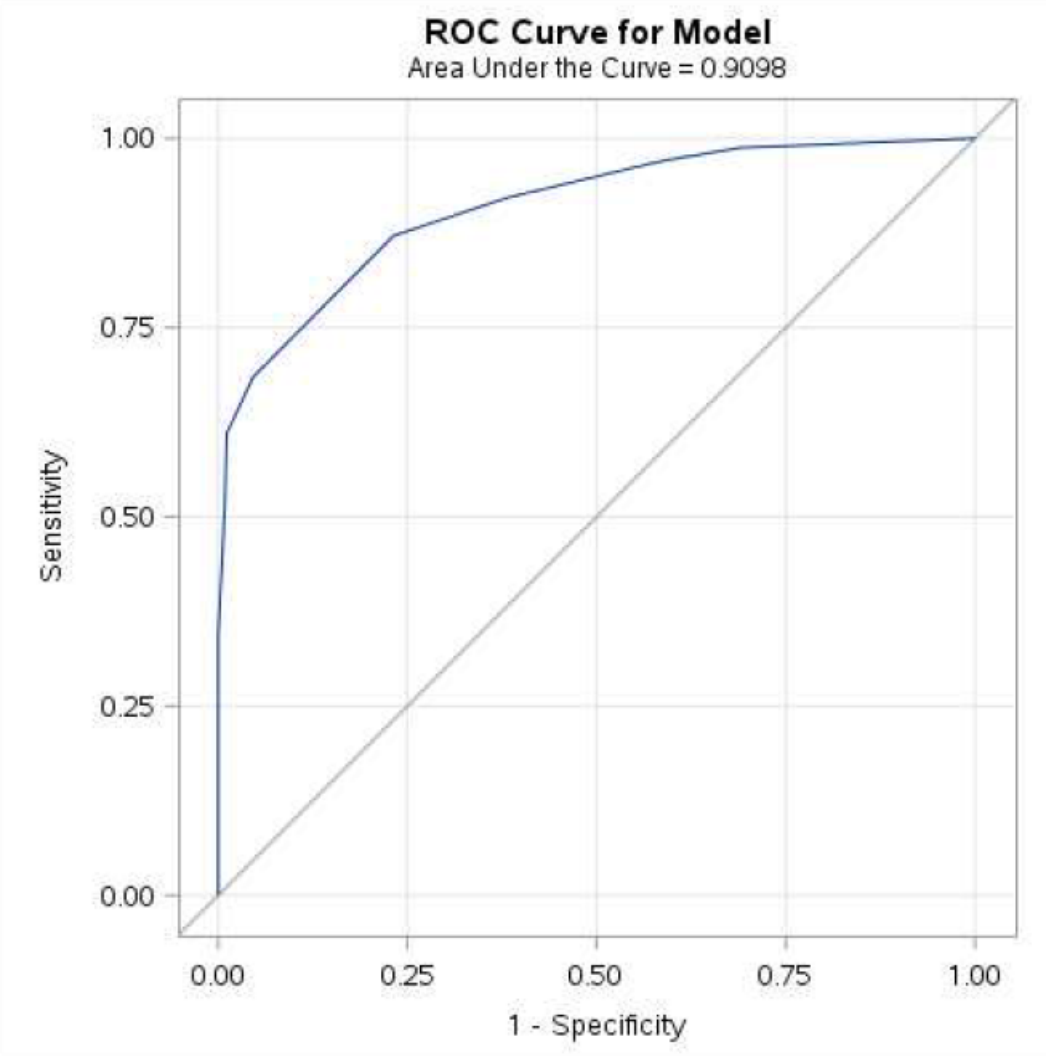
Slika 2.18: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Normalne jezgrice (ispis iz SASa)



Slika 2.19: Grafički prikaz vjerojatnosti pojavnosti malignog tumora za Mitozu (ispis iz SASa)

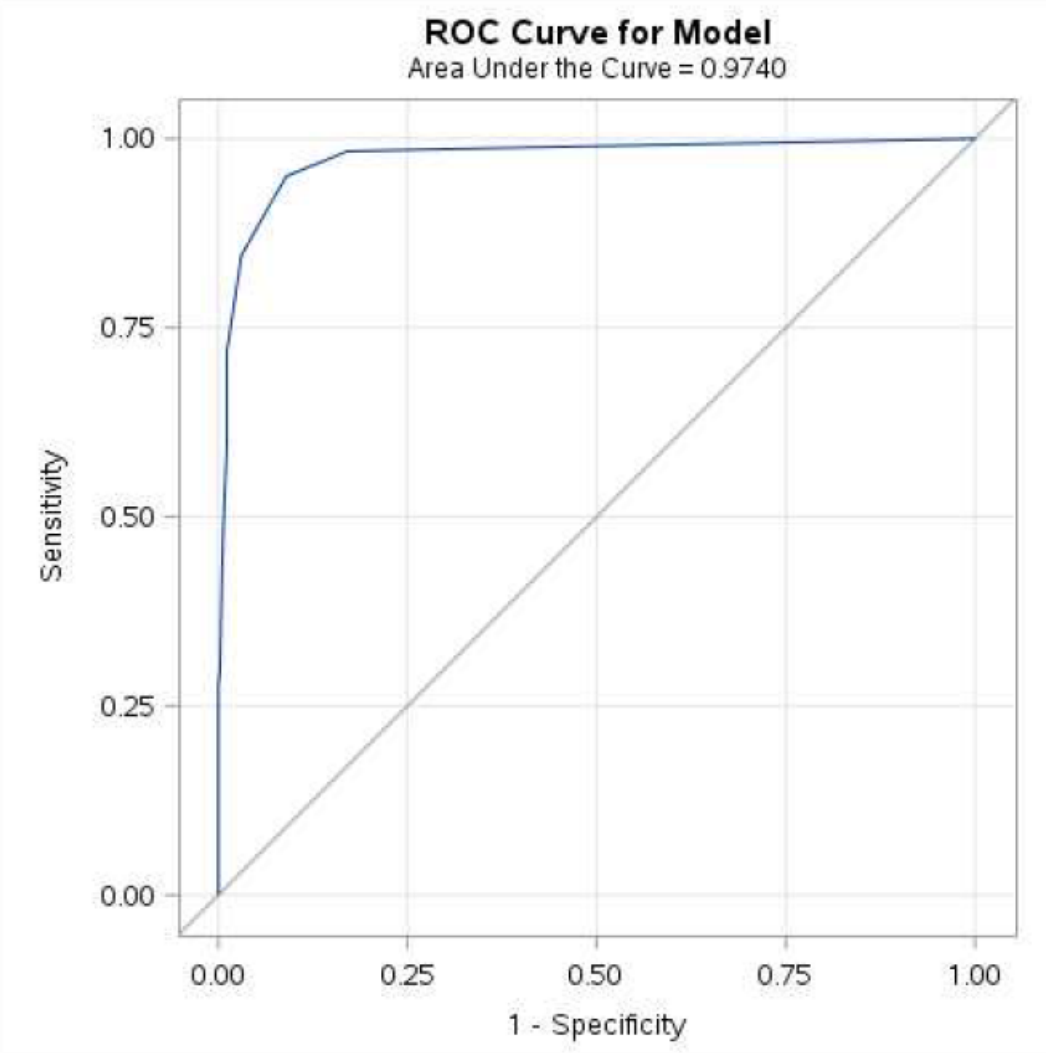
Iz gornjih prikaza (slika 2.11 - slika 2.19) vidimo kako se povećanjem nezavisne varijable povećava vjerojatnost malignog tumora. Prema slici 2.19 vidimo da je najveći rast za varijablu Mitoza.

Pogledajmo sada ROC krivulje za univarijatne logističke modele.



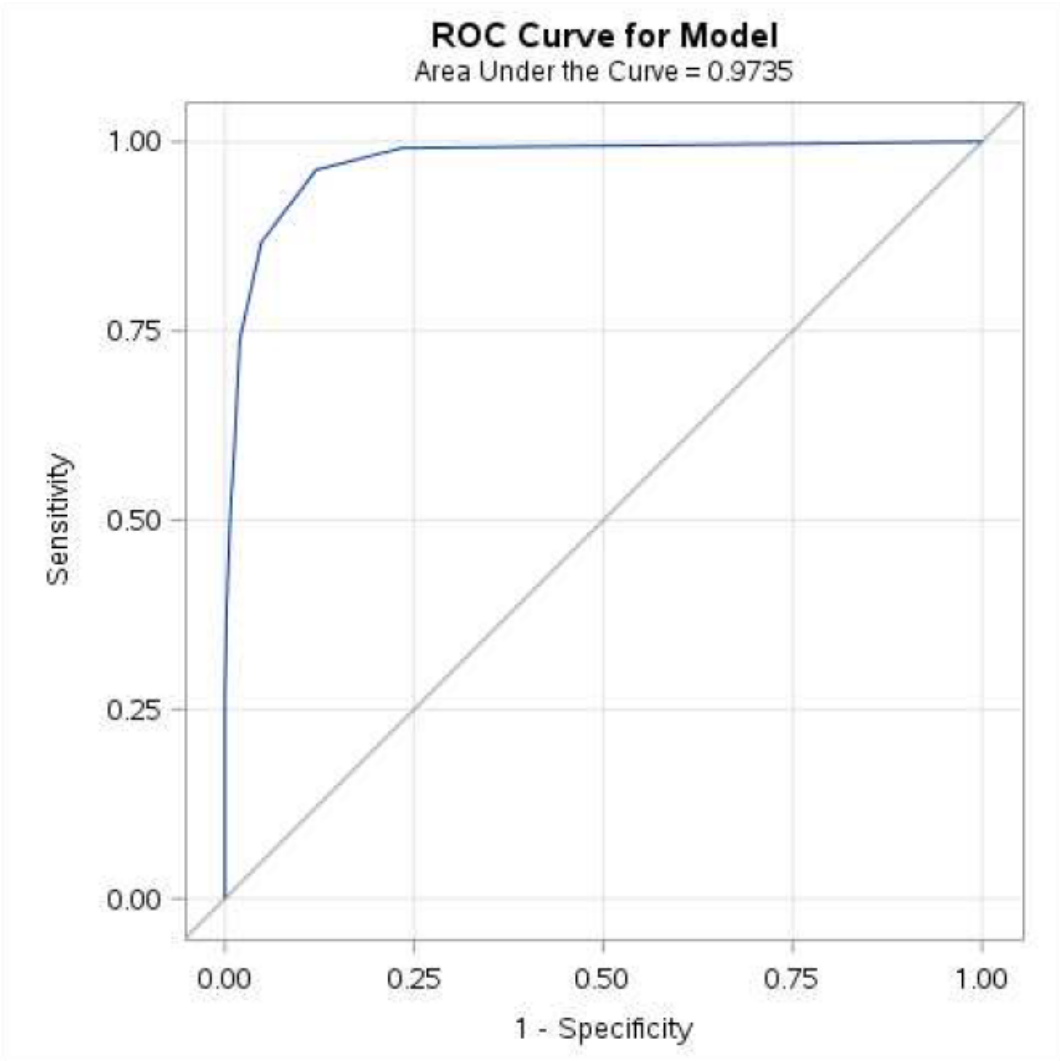
Slika 2.20: ROC krivulja za Debljinu grumena (ispis iz SASa)

c=0,9098, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 90,98%.



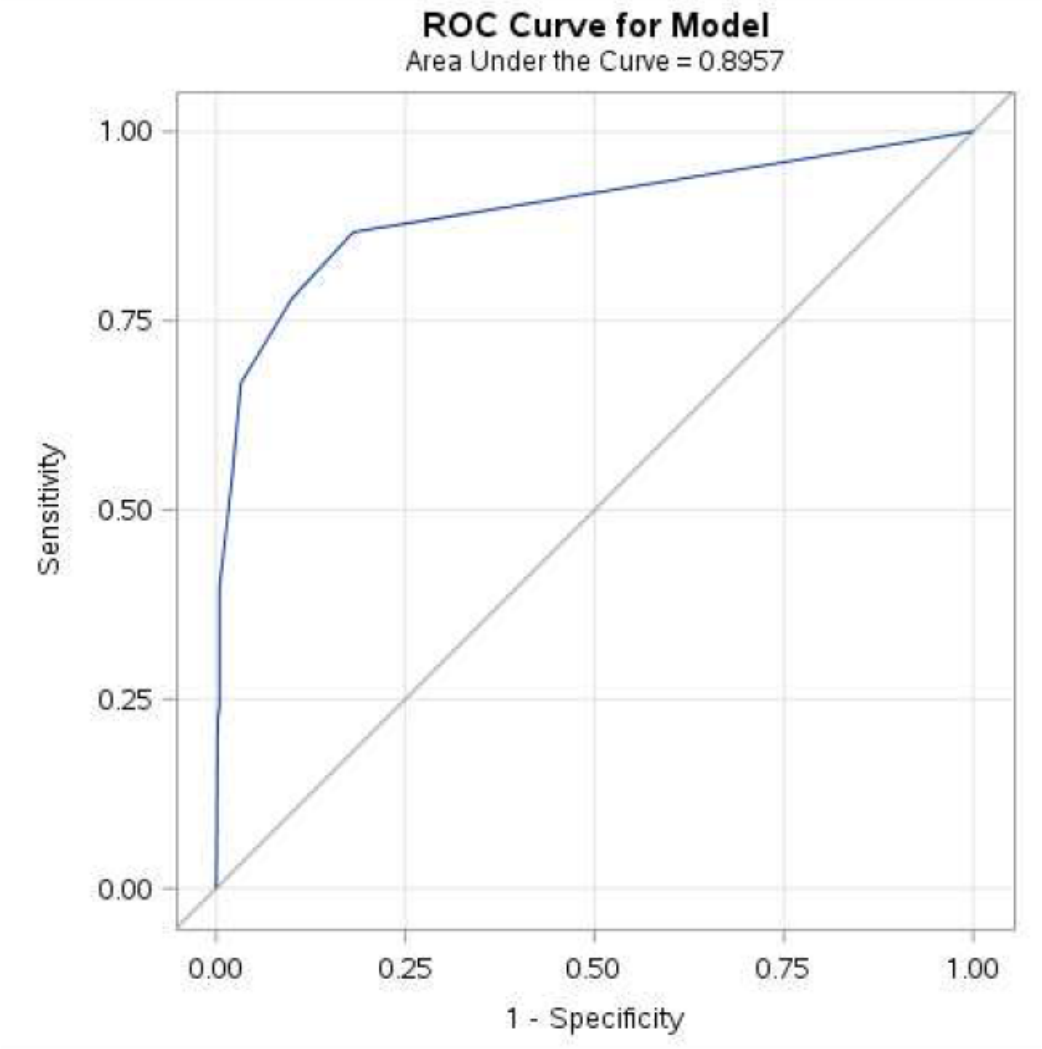
Slika 2.21: ROC krivulja za Uniformnost veličine stanice (ispis iz SASa)

c=0,974, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 97,4%.



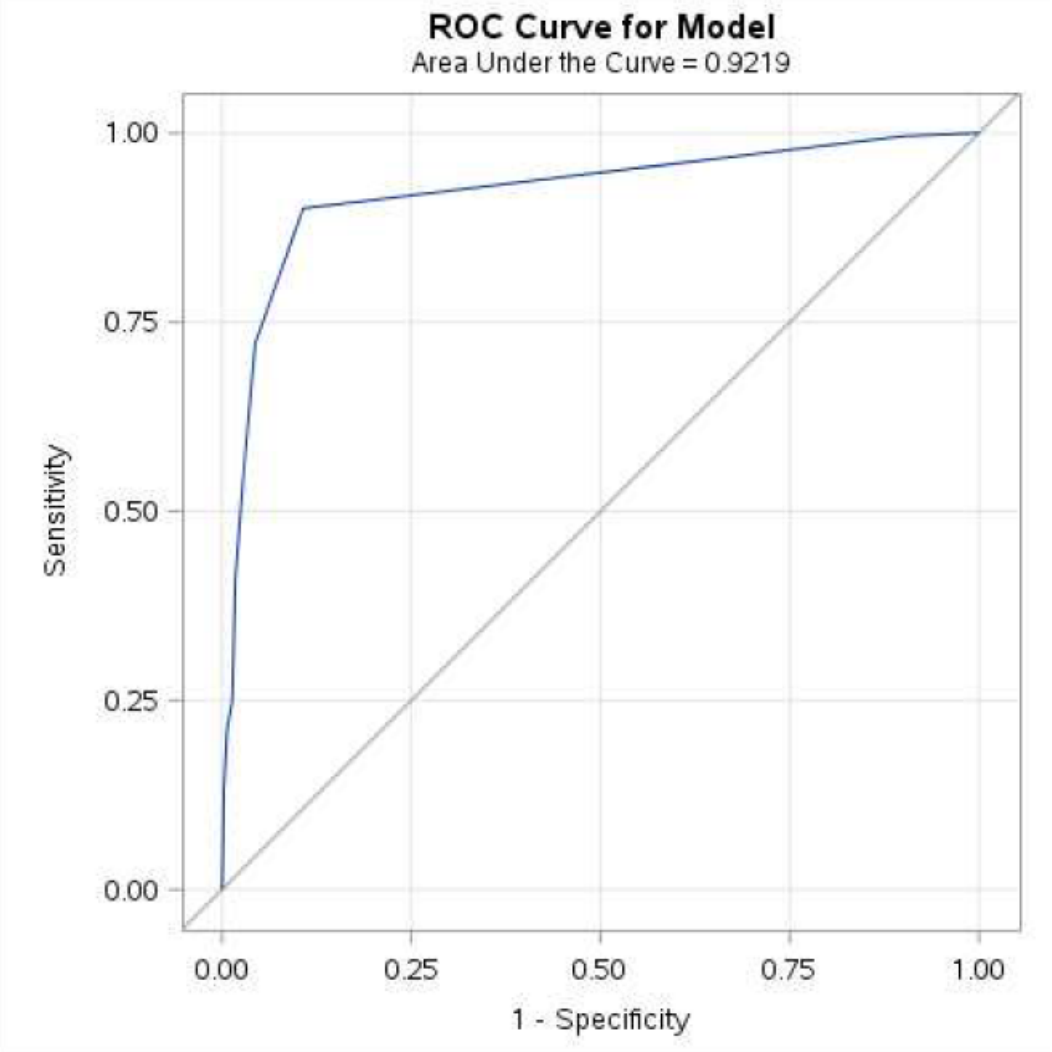
Slika 2.22: ROC krivulja za Uniformnost staničnog oblika (ispis iz SASa)

c=0,9735, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 97,35%.



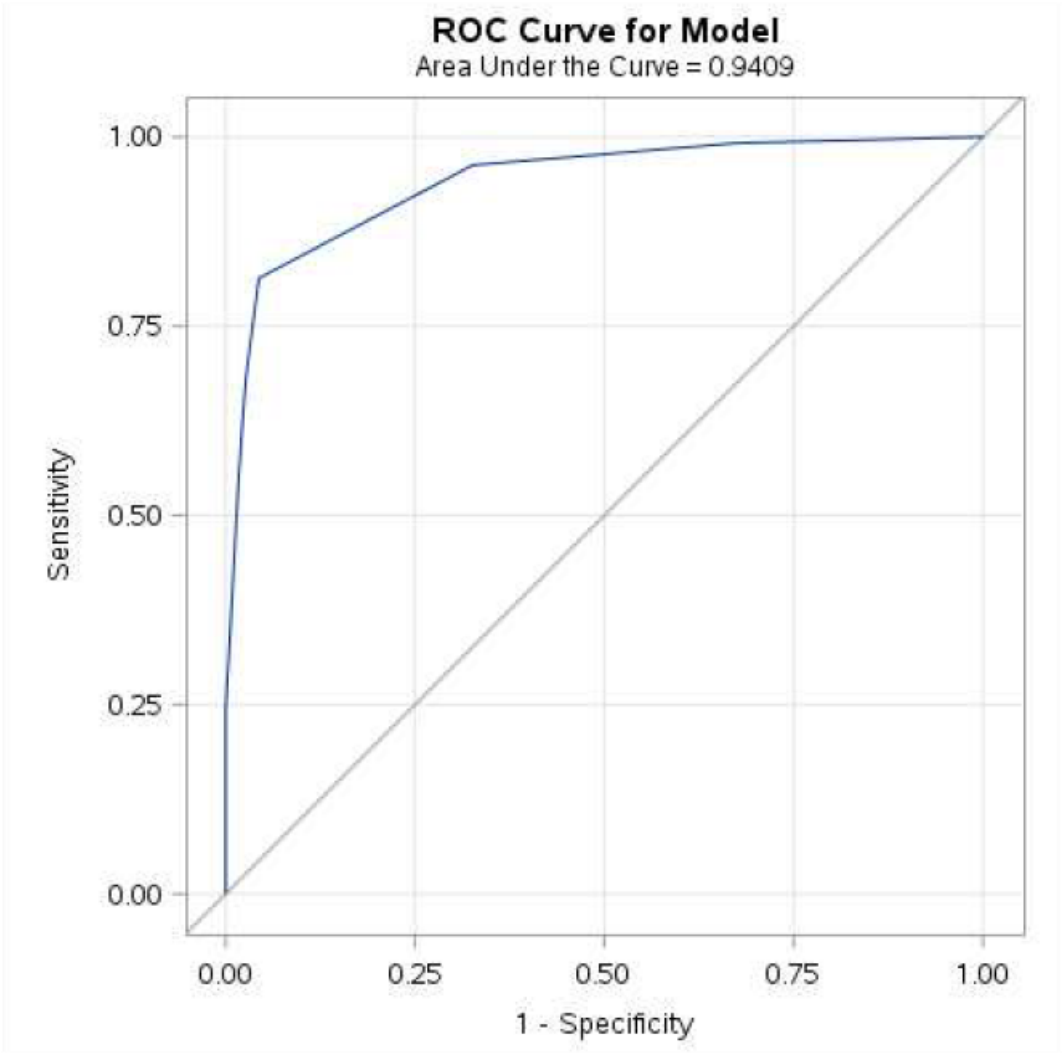
Slika 2.23: ROC krivulja za Marginalnu adheziju (ispis iz SASa)

c=0,8957, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 89,57%.



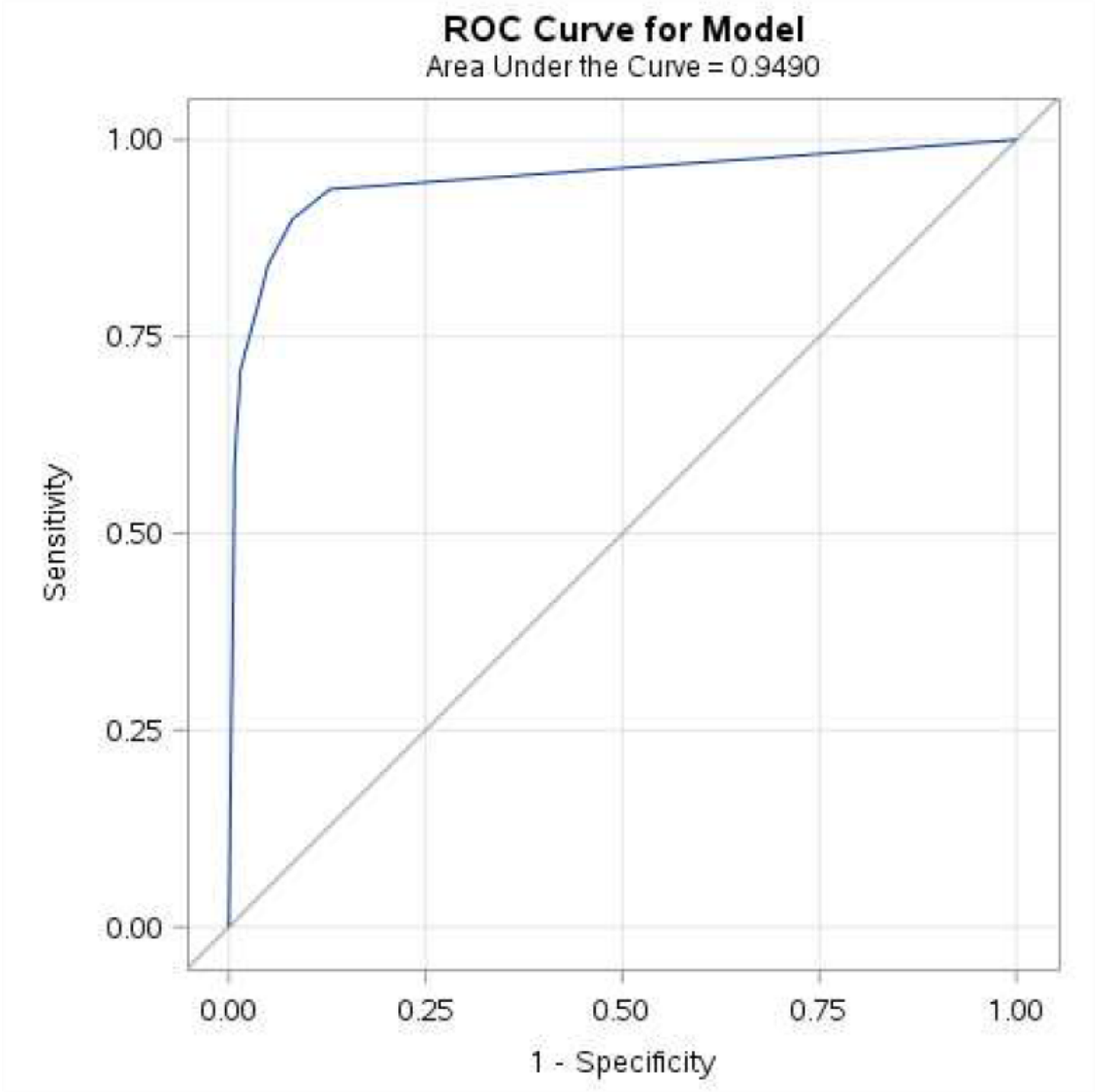
Slika 2.24: ROC krivulja za Veličinu epitelne stanice (ispis iz SASa)

c=0,9219, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 92,19%.



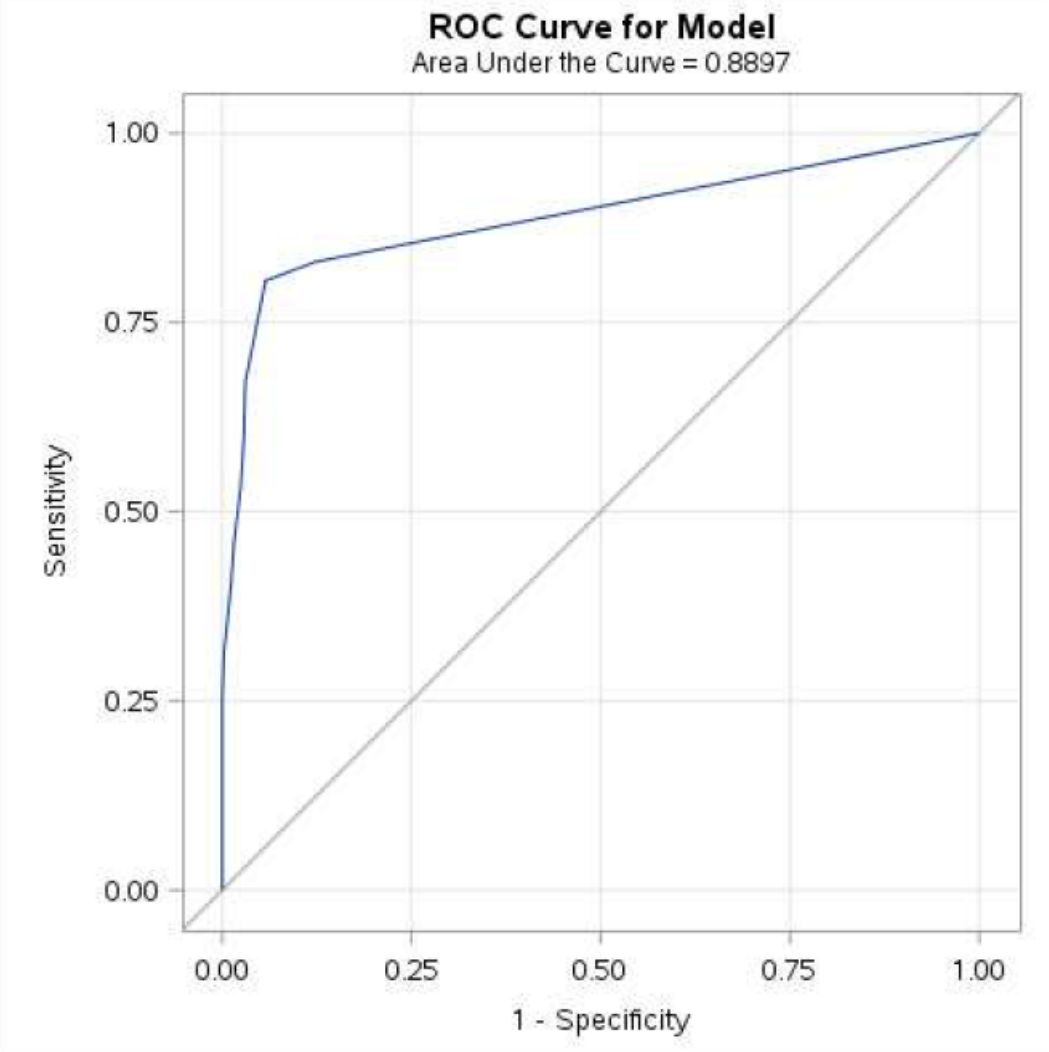
Slika 2.25: ROC krivulja za Normalan kromatin (ispis iz SASa)

c=0,9409, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 94,09%.



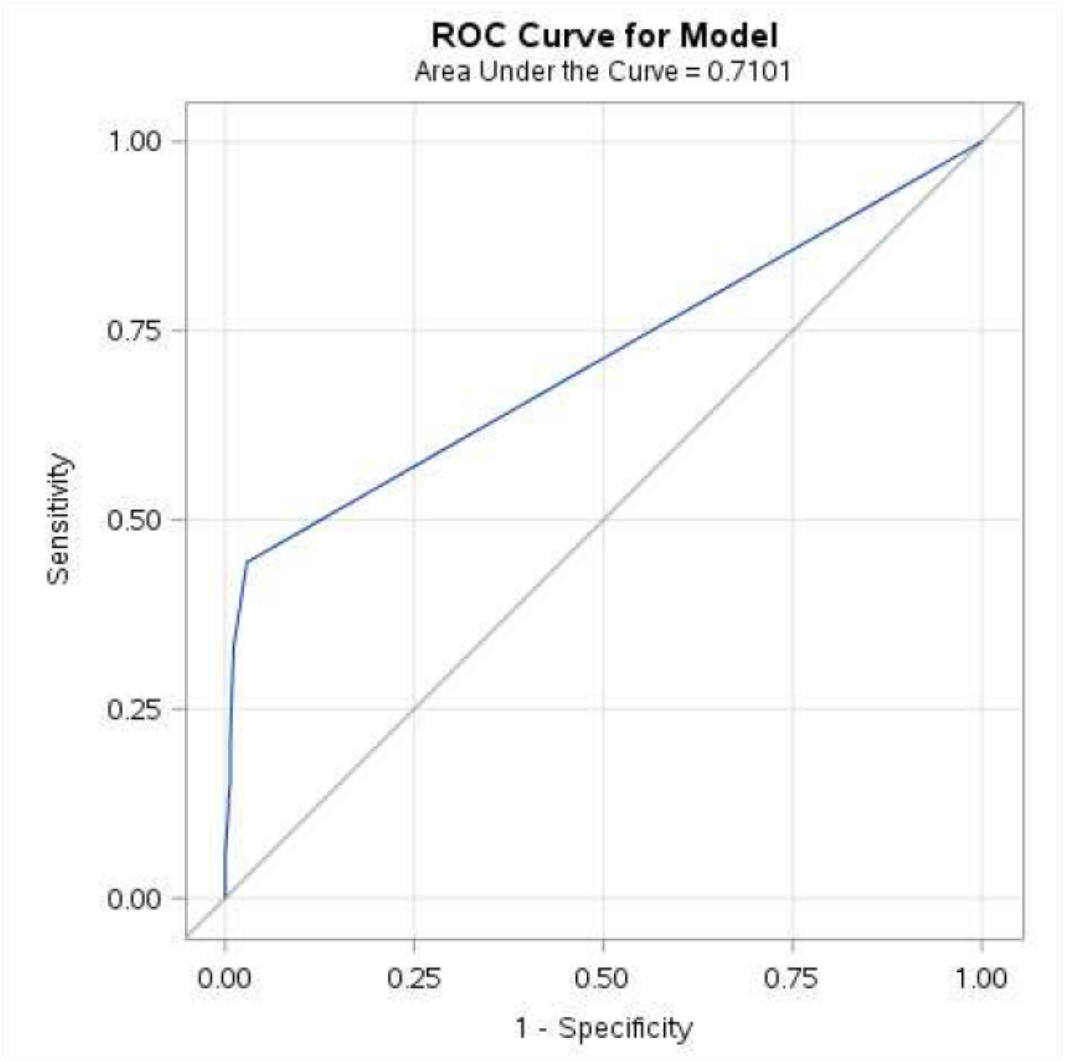
Slika 2.26: ROC krivulja za Golu jezgru (ispis iz SASa)

c=0,949, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 94,9%.



Slika 2.27: ROC krivulja za Normalne jezgrice (ispis iz SASa)

c=0,8897, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 88,97%.



Slika 2.28: ROC krivulja za Mitozu (ispis iz SASa)

c=0,710, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 71%.

Prema slikama 2.20 do 2.28 vidimo da sve varijable imaju jako dobru prediktivnu snagu za određivanje malignosti tumora dojke. Najveću ima Uniformnost veličine stanice (0,974), dok najmanju ima Mitoza (0,71).

2.3 Multivarijatna logistička regresija

U ovom poglavlju ćemo za sve nezavisne varijable iz baze provesti multivarijatnu logističku regresiju. Na taj način ćemo odrediti statističku značajnost varijabli, tj. statističku značajnost pripadnih (procijenjenih) parametara. Kod multivarijatnog modela koristiti ćemo svih 9 varijabli. Podatci koji nas zanimaju prikazani su u tablicama ispod.

Tablica 2.5: Rezultati analize multivarijatnog logističkog modela (ispis iz SASa)

DF	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vrijednost
9	884,35	102,888	781,462	<.0001

Kriterij konvergencije je zadovoljen. Iz tablice 2.5 vidimo da je dobiveni model statistički značajan, budući da je $\chi^2 = 781,4620)$ i p-vrijednost <0,0001.

Tablica 2.6: Rezultati ML procjene parametara za multivarijatni logistički model (ispis iz SASa)

Varijable	DF	Procjena	Standardna greška	Wald χ^2	p-vrijednost
Intercept	1	-10,1037	1,1748	73,9609	<.0001
Debljina grumena	1	0,535	0,142	14,1916	0,0002
Uniformnost veličine stanice	1	-0,00628	0,2091	0,0009	0,9761
Uniformnost staničnog oblika	1	0,3227	0,2306	1,9584	0,1617
Marginalna adhezija	1	0,3306	0,1234	7,173	0,0074
Veličina epitelne stanice	1	0,0966	0,1566	0,3808	0,5372
Gole jezgre	1	0,383	0,0938	16,6591	<.0001
Normalan kromatin	1	0,4472	0,1714	6,8082	0,0091
Normalne jezgrice	1	0,213	0,1129	3,562	0,0591
Mitoza	1	0,5348	0,3288	2,6459	0,1038

Iz tablice 2.6 vidimo da je jednadžba dobivenog modela:

$$\begin{aligned} logit(p) = &-10,1037 + 0,535 \cdot \text{Debljina grumena} - 0,00628 \cdot \text{Uniformnost veličine stanice} \\ &+ 0,3227 \cdot \text{Uniformnost staničnog oblika} + 0,3306 \cdot \text{Marginalna adhezija} \\ &+ 0,0966 \cdot \text{Veličina epitelne stanice} + 0,383 \cdot \text{Gole jezgre} \\ &+ 0,4472 \cdot \text{Normalan kromatin} + 0,213 \cdot \text{Normale jezgrice} \\ &+ 0,5348 \cdot \text{Mitoza} \end{aligned}$$

Varijable koje su statistički značajne na razini značajnosti od 5% su varijable kojima je p-vrijednost manja od 0.05. Iz tablice 2.6 vidimo da su to Debljina grumena, Marginalna adhezija, Gole jezgre i Normalan kromatin.

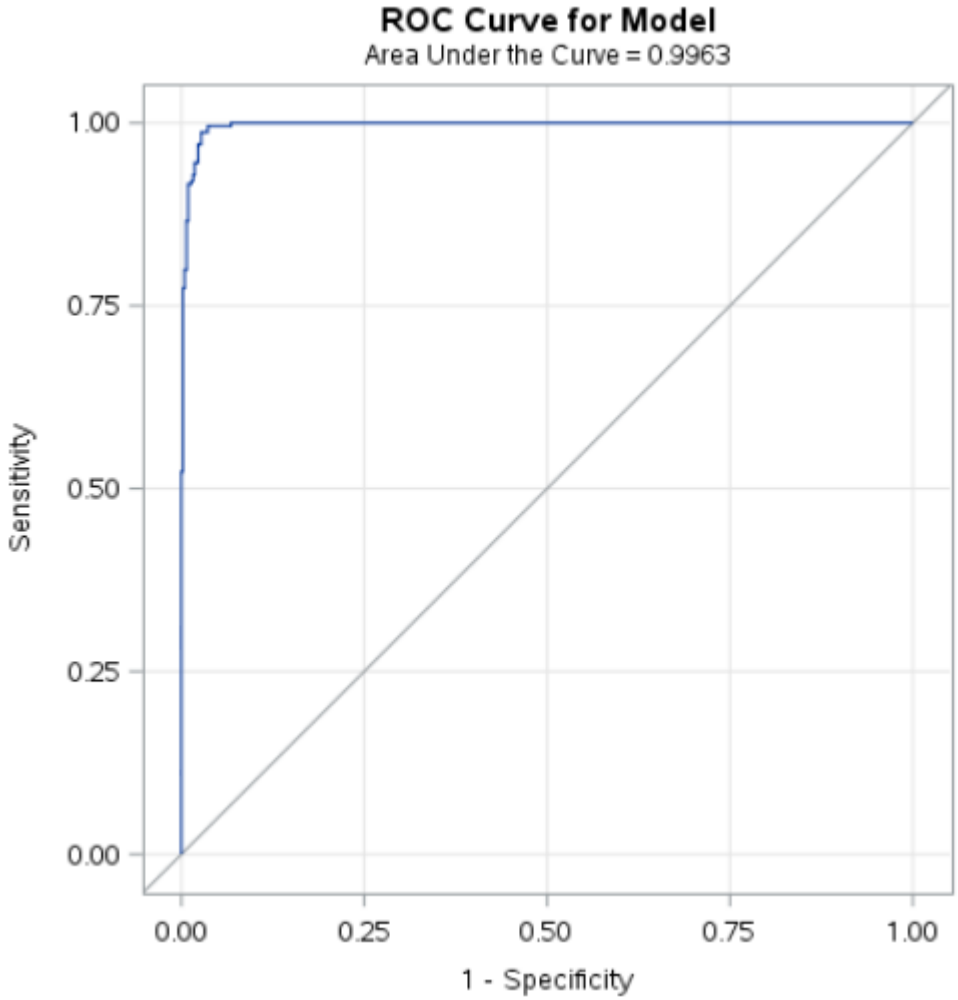
Tablica 2.7: Procjene omjera šansi za multivarijatni logistički model (ispis iz SASa)

Varijable	Procjena OR	95% pouzdani interval
Debljina grumena	1,707	1,293 - 2,255
Uniformnost veličine stanice	0,994	0,660 - 1,497
Uniformnost staničnog oblika	1,381	0,879 - 2,170
Marginalna adhezija	1,392	1,093 - 1,773
Veličina epitelne stanice	1,101	0,810 - 1,497
Gole jezgre	1,467	1,220 - 1,763
Normalan kromatin	1,564	1,118 - 2,188
Normalne jezgrice	1,237	0,992 - 1,544
Mitoza	1,707	0,896 - 3,252

Iz tablice 2.7 vidimo da su 95% pouzdani intervali koji ne sadrže jedinicu intervali vezani uz varijable Debljina grumena, Marginalna adhezija, Gole jezgre i Normalan kromatin. Pa zaključujemo da su navedene varijable statistički značajne.

”Procjenu OR” za statistički značajne varijable iz tablice 2.7 tumačimo na sljedeći način:

- Povećanje Debljine grumena za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,707 puta.
- Povećanje Marginalne adhezija za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,392 puta.
- Povećanje Gole jezgre za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,467 puta.
- Povećanje Normalnog kromatina za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,564 puta.



Slika 2.29: ROC krivulja za multivarijatni logistički model

c=0,9963, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 99,63%.

2.4 Stepwise procedura

Podatke ćemo prikazati onim redoslijedom kojim su ulazili u model.

Tablica 2.8: Rezultati analize stepwise procedure (ispis iz SASa)

Varijable	DF	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vrijednost
Gole jezgre	1	884,35	340,628	543,7224	<.0001
Uniformnost staničnog oblika	2	884,35	171,385	712,9649	<.0001
Debljina grumena	3	884,35	140,246	744,104	<.0001
Normalan kromatin	4	884,35	122,743	761,6071	<.0001
Marginalna adhezija	5	884,35	112,566	771,7841	<.0001
Normale jezgrice	6	884,35	107,144	777,2065	<.0001

Kriterij konvergencije je zadovoljen. Iz tablice 2.8 vidimo da je dobiveni model statistički značajan, budući da je $\chi^2 > 107,144$) za sve varijable i p-vrijednost <0.0001 za sve varijable.

Tablica 2.9: Rezultati ML procjene parametara za stepwise proceduru (ispis iz SASa)

Varijable	DF	Procjena	Standardna greška	Wald χ^2	p-vrijednost
Intercept	1	-9,767	1,0851	81,0263	<.0001
Debljina grumena	1	0,6225	0,1371	20,6122	<.0001
Uniformnost staničnog oblika	1	0,3495	0,165	4,4854	0,0342
Marginalna adhezija	1	0,3375	0,1156	8,524	0,0035
Gole jezgre	1	0,3785	0,0938	16,2841	<.0001
Normalan kromatin	1	0,4713	0,1661	8,0508	0,0045
Normalne jezgrice	1	0,2432	0,1086	5,0179	0,0251

Iz tablice 2.10 vidimo da jednadžba dobivenog modela glasi:

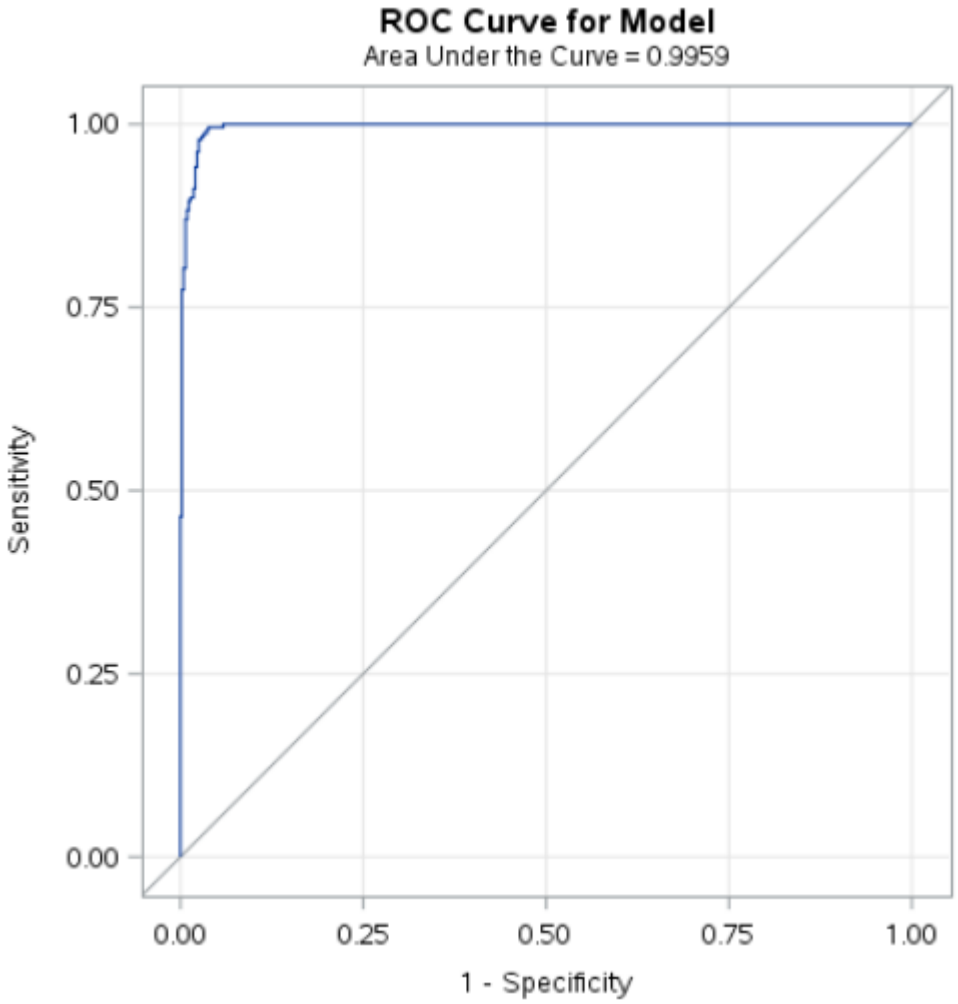
$$\begin{aligned} logit(p) = &-9,767 + 0,6225 \cdot \text{Debljina grumena} + 0,3495 \cdot \text{Uniformnost staničnog oblika} \\ &+ 0,3375 \cdot \text{Marginalna adhezija} + 0,3785 \cdot \text{Gole jezgre} \\ &+ 0,4713 \cdot \text{Normalan kromatin} + 0,2432 \cdot \text{Normalne jezgrice} \end{aligned}$$

Tablica 2.10: Procjene omjera šansi za stepwise proceduru (ispis iz SASa)

Varijable	Procjena OR	95% pouzdani interval
Debljina grumena	1,864	1,424 - 2,438
Uniformnost staničnog oblika	1,418	1,026 - 1,960
Marginalnacadhezija	1,401	1,117 - 1,758
Gole jezgre	1,46	1,215 - 1,755
Normalan kromatin	1,602	1,157 - 2,219
Normalne jezgrice	1,275	1,031 - 1,578

Iz tablice 2.10 (stupac ”Procjena OR”) slijedi:

- Povećanje Debljine grumena za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,864 puta.
- Povećanje Uniformnosti staničnog oblika za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,418 puta.
- Povećanje Marginalne adhezije za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,401 puta.
- Povećanje Gole jezgre za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,46 puta.
- Povećanje Normalnog kromatina za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,602 puta.
- Povećanje Normalne jezgrice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,275 puta.



Slika 2.30: ROC krivulja za stepwise proceduru

c=0,9959, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 99,59%.

2.5 Multikolinearnost

Budući da se multivarijatni logistički model i stepwise procedura razlikuju bilo bi zgodno pogledati ima li koreliranosti među varijablama. Ako su dvije ili više varijabli međusobno jako korelirane teško je dobiti dobru procjenu njihovog utjecaja na zavisnu varijablu.

Tablica 2.11: Multikolinearnost (ispis iz SASa)

Varijable	DF	Tolerance	Variance Inflation
Debljina grumena	1	0,52474	1,9057
Uniformnost veličine stanice	1	0,13899	7,19472
Uniformnost staničnog oblika	1	0,15268	6,54981
Marginalna adhezija	1	0,40547	2,46628
Veličina epitelne stanice	1	0,39119	2,55633
Gole jezgre	1	0,38504	2,59716
Normalan kromatin	1	0,34734	2,87902
Normalne jezgrice	1	0,41129	2,4314
Mitoza	1	0,71604	1,39658

Tolerancija se računa tako da se radi regresijski model svake nezavisne varijable na sve ostale nezavisne varijable, pritom računajući R^2 i zatim se R^2 oduzima od 1. Prema tome, niska tolerancija odgovara visokoj multikolinearnosti. Iako ne postoji neko striktno pravilo za toleranciju, sve značajno manje od 0,3 trebalo bi biti zabrinjavajuće. Isto tako visoka inflacija varijance (*engl. variance inflation*) odgovara viskoj multikolinearnosti. Inflacija varijance je zapravo recipročna vrijednost tolerancije. Govori nam koliko će se varijanca koeficijenta "napuhati" u odnosu na to kakva bi bila kada varijabla ne bi bila korelirana s drugim varijablama. Slično kao kod tolerancije, vrijednosti inflacije varijance veće od 4 trebale bi biti zabrinjavajuće. [3]

Prema tablici 2.11 vidimo da varijable Uniformnost veličine stanice i Uniformnost staničnog oblika imaju jako malenu toleranciju, pa možemo zaključiti da su te varijable korelirane s nekim drugim varijablama.

Iako je uočiti multikolinearnost relativno lako, mnogo veći problem je šta učiniti nakon toga. Jedno od mogućih rješenja je isključiti iz modela varijable koje najviše doprinose multikolinearnosti. No to ne znači nužno da će tada model biti dobar. No pogledajmo što ćemo dobiti u tom postupku.

Tablica 2.12: Multikolinearnost nakon što smo izbacili Uniformnost veličine stanice i Uniformnost staničnog oblika iz modela (ispis iz SASa)

Varijable	DF	Tolerance	Variance Inflation
Debljina grumena	1	0,56647	1,76532
Marginalna adhezija	1	0,42337	2,36198
Veličina epitelne stanice	1	0,45587	2,19361
Gole jezgre	1	0,4012	2,49253
Normalan kromatin	1	0,37706	2,65206
Normalne jezgrice	1	0,44002	2,27261
Mitoza	1	0,71877	1,39127

Prema tablici 2.12 vidimo da nakon izbacivanja Uniformnosti veličine stanice i Uniformnosti staničnog oblika iz modela više nemamo multikolinearnosti.

Pogledajmo sada kako će izgledati multivarijatni logistički model i stepwise procedura nakon što Uniformnost veličine stanice i Uniformnost staničnog oblika izbacimo iz modela.

Multivarijatni model

Tablica 2.13: Rezultati analize multivarijatnog logističkog modela (ispis iz SASa)

DF	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vrijednost
7	884,35	106,663	777,687	<.0001

Kriterij konvergencije je zadovoljen. Iz tablice 2.13 vidimo da je dobiveni model statistički značajan, budući da je $\chi^2 = 777,687$ i p-vrijednost <0,0001

Tablica 2.14: Rezultati ML procjene parametara za multivarijatni logistički model (ispis iz SASa)

Varijable	DF	Procjena	Standardna greška	Wald χ^2	p-vrijednost
Intercept	1	-10,5815	1,1704	81,7339	<.0001
Debljina grumena	1	0,6396	0,1342	22,7015	<.0001
Marginalna adhezija	1	0,3743	0,1214	9,5064	0,002
Veličina epitelne stanice	1	0,1684	0,15	1,2607	0,2615
Gole jezgre	1	0,4342	0,0894	23,6009	<.0001
Normalan kromatin	1	0,5237	0,1542	11,5406	0,0007
Normalne jezgrice	1	0,2726	0,107	6,4886	0,0109
Mitoza	1	0,574	0,3201	3,2149	0,073

Prema tablici 2.14 jednadžba modela glasi:

$$\begin{aligned} logit(p) = &-10,5815 + 0,6396 \cdot \text{Debljina grumena} + 0,3743 \cdot \text{Marginalna adhezija} \\ &+ 0,1684 \cdot \text{Veličina epitelne stanice} + 0,4342 \cdot \text{Gole jezgre} \\ &+ 0,5237 \cdot \text{Normalan kromatin} + 0,2726 \cdot \text{Normale jezgrice} \\ &+ 0,574 \cdot \text{Mitoza} \end{aligned}$$

Također iz tablice 2.14 vidimo da su varijable koje su statistički značajne za naš model: Debljina grumena, Marginalna adhezija, Gole jezgre, Normalan kromatin i Normalne jezgrice.

Tablica 2.15: Procjena omjera šansi za multivarijatni logistički model (ispis iz SASa)

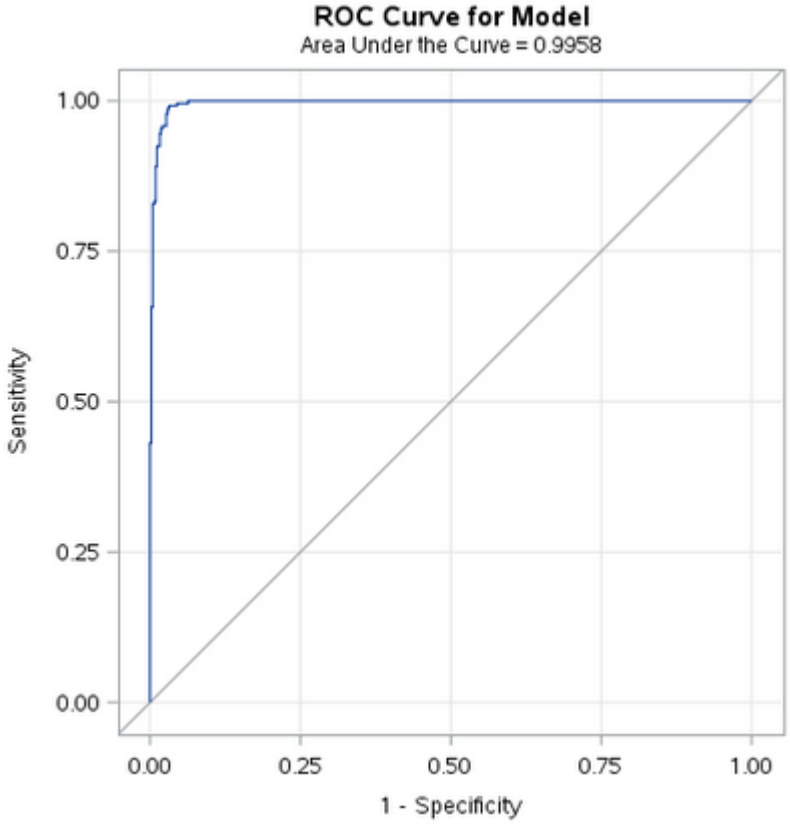
Varijable	Procjena OR	95% pouzdani interval
Debljina grumena	1,896	1,457 - 2,466
Marginalna adhezija	1,454	1,146 - 1,845
Veličina epitelne stanice	1,183	0,882 - 1,588
Gole jezgre	1,544	1,296 - 1,839
Normalan kromatin	1,688	1,248 - 2,284
Normalne jezgrice	1,313	1,065 - 1,620
Mitoza	1,775	0,948 - 3,325

95% pouzdani intervali koji ne sadrže jedinicu, prema tablici 2.15 su intervali vezani uz varijable: Debljina grumena, Marginalna adhezija, Gole jezgre, Normalan kromatin

i Normalne jezgrice, pa opet zaključujemo da su te varijable statistički značajne za naš model.

Iz tablice 2.15 (stupac "Procjena OR") za statistički značajne varijable slijedi:

- Povećanje Debljine grumena za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,896 puta.
- Povećanje Marginalne adhezije za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,454 puta.
- Povećanje Gole jezgre za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,544 puta.
- Povećanje Normalnog kromatina za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,688 puta.
- Povećanje Normalne jezgrice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,313 puta.



Slika 2.31: ROC krivulja za multivarijatni logistički model

c=0,9958, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 99,58%.

Stepwise model

Tablica 2.16: Rezultati analize stepwise procedure (ispis iz SASa)

Varijable	DF	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vrijednost
Gole jezgre	1	884,35	340,628	543,7224	<.0001
Normalne jezgrice	2	884,35	213,615	670,7355	<.0001
Debljina grumena	3	884,35	147,34	737,0102	<.0001
Normalan kromatin	4	884,35	125,96	758,3905	<.0001
Marginalna adhezija	5	884,35	112,264	772,0867	<.0001

U tablici 2.16 varijable smo prikazali onim redoslijedom kojim su ulazile u model. Dobiveni model je statistički značajan, budući da je $\chi^2 \geq 543,7224$ za sve varijable i p-vrijednost <0,0001 za sve varijable

Tablica 2.17: Rezultati ML procjene parametara za stepwise proceduru (ispis iz SASa)

Varijable	DF	Procjena	Standardna greška	Wald χ^2	p-vrijednost
Intercept	1	-10,1306	1,0945	85,6657	<.0001
Debljina grumena	1	0,7413	0,1319	31,5924	<.0001
Marginalna adhezija	1	0,3952	0,1159	11,6199	0,0007
Gole jezgre	1	0,4473	0,088	25,8562	<.0001
Normalan kromatin	1	0,5529	0,1502	13,5514	0,0002
Normalne jezgrice	1	0,3342	0,0978	11,673	0,0006

Prema tablici 2.17 jednadžba modela glasi:

$$\begin{aligned} logit(p) = &-10,13065 + 0,7413 \cdot \text{Debljina grumena} + 0,3952 \cdot \text{Marginalna adhezija} \\ &+ 0,4473 \cdot \text{Gole jezgre} + 0,5529 \cdot \text{Normalan kromatin} \\ &+ 0,3342 \cdot \text{Normale jezgrice} \end{aligned}$$

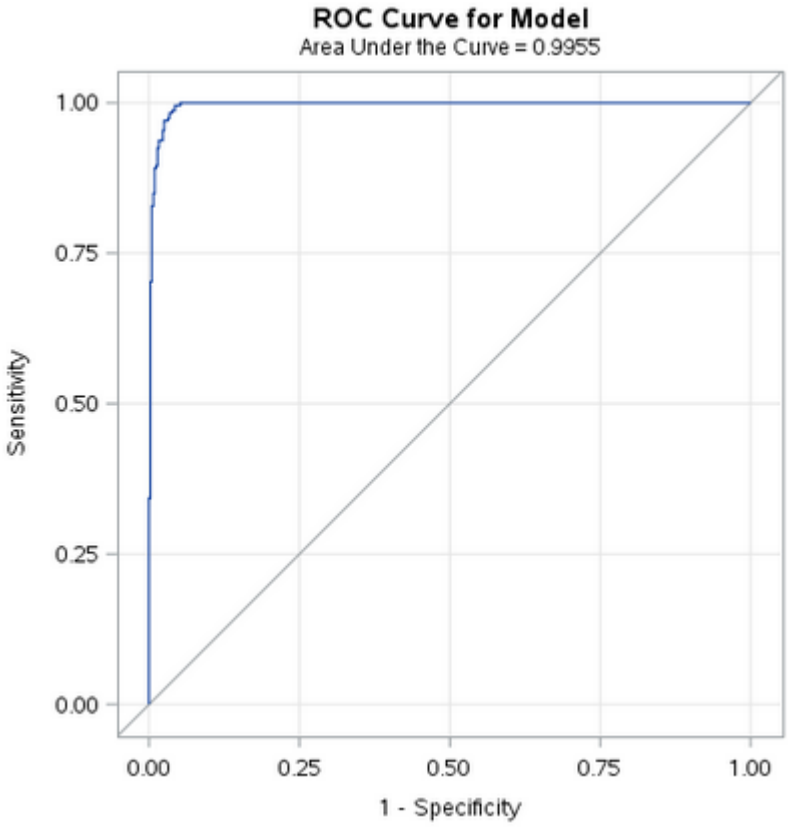
Tablica 2.18: Procjena omjera šansi za stepwise proceduru (ispis iz SASa)

Varijable	Procjena OR	95% pouzdani interval
Debljina grumena	2,099	1,621 - 2,718
Marginalna adhezija	1,485	1,183 - 1,863
Gole jezgre	1,564	1,316 - 1,858
Normalan kromatin	1,738	1,295 - 2,333
Normalne jezgrice	1,397	1,153 - 1,692

Iz tablice 2.18 (stupac "Procjena OR") slijedi:

- Povećanje Debljine grumena za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 2,099 puta.
- Povećanje Marginalne adhezije za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,485 puta.

- Povećanje Gole jezgre za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,564 puta.
- Povećanje Normalnog kromatina za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,738 puta.
- Povećanje Normalne jezgrice za 1 povećava omjer šanse za prelazak tumora iz benignog u maligni za 1,397 puta.



Slika 2.32: ROC krivulja za stepwise proceduru

c=0,9955, dakle prediktivna snaga malignosti tumora ovim modelom iznosi 99,55%.

2.6 Zaključak

Iako smo putem univarijatnog logističkog modela dobili da su sve varijable statistički značajne, multivarijatni logistički model i stepwise procedura pokazali su da to nije tako. Unutar multivarijatnog logističkog modela i stepwise procedure smo primijetili malene razlike u varijablama koje smo dobili kao značajne. Proučavajući razlog toga zaključili smo da postoji koreliranost među varijablama. Nakon što smo uklonili korelirane varijable iz modela, multivarijatni logistički model i stepwise procedura dali su isti rezultat. Kod multivarijatnog logističkog modela (nakon uklanjanja koreliranih varijabli) varijabla Mitoza je na granici da bi ušla u model. P-vrijednost joj je 0,073, a 95% pouzdani interval skoro pa ne sadrži jedinicu ([0,948, 3,325]). Kada bismo promatrali na razini značajnosti od 10% varijabla Mitoza bi sigurno ušla u model. Varijable koje su statistički značajne na razini značajnosti 5% i koje ulaze u model su: Debljina grumena, Marginalna adhezija, Gole jezgre, Normalan kromatin i Normalne jezgrice.

Poglavlje 3

Dodatak

3.1 SAS kod

```
/* Generated Code (IMPORT) */
/* Source File: pod.xlsx */
/* Source Path: /folders/myfolders/Zadace */
/* Code generated on: 8/30/17, 12:15 PM */

    %web drop table(WORK.IMPORT);
    FILENAME REFFILE '/folders/myfolders/Zadace/pod.xlsx';

    PROC IMPORT DATAFILE=REFFILE
DBMS=XLSX
OUT=WORK.IMPORT;
GETNAMES=YES;
RUN;

    PROC CONTENTS DATA=WORK.IMPORT; RUN;

    %web open table(WORK.IMPORT);

    data nova; set WORK.IMPORT;
if Class=4 then do; Class=1;end;
if Class=2 then do; Class=0;end;
run;

    proc print data=nova;
```

```
run;

    data nova1;
set nova;
Debljina grumena = log( Debljina grumena );
Uniformnost velicine stanice = log(Uniformnost velicine stanice );
Uniformnost stanicnog oblika = log( Uniformnost stanicnog oblika );
Marginalna adhezija = log( Marginalna adhezija );
Velicina epitelne stanice = log( Velicina epitelne stanice );
Gole jezgre = log( Gole jezgre );
Normalan kromatin = log( Normalan kromatin );
Normalne jezgrice = log( Normalne jezgrice );
Mitoza = log( Mitoza );
run;

    proc print data=nova1;
run;

    proc ttest data=nova1;
class Class;
var Debljina grumena Uniformnost velicine stanice Uniformnost stanicnog oblika
Marginalna adhezija Velicina epitelne stanice Gole jezgre Normalan kromatin Normalne
jezgrice Mitoza;
run;

    PROC MEANS DATA=nova n mean median std min max;
var Debljina grumena Uniformnost velicine stanice Uniformnost stanicnog oblika
Marginalna adhezija Velicina epitelne stanice Gole jezgre Normalan kromatin Normalne
jezgrice Mitoza;
run;

    /*1*/
title" Univarijatna logisticka regresija-Clump Thickness";
proc logistic data=nova descending;
model Class=Debljina grumena /lackfit rsq outroc=rocgraf;
run;
/*2*/
title" Univarijatna logisticka regresija-Uniformity of Cell Size";
proc logistic data=nova descending;
```

```
model Class=Uniformnost velicine stanice /lackfit rsq outroc=rocgraf;
run;
/*3*/
title” Univarijatna logisticka regresija-Uniformity of Cell Shape”;
proc logistic data=nova descending;
model Class=Uniformnost stanicnog oblika /lackfit rsq outroc=rocgraf;
run;
/*4*/
title” Univarijatna logisticka regresija-Marginal Adhesion”;
proc logistic data=nova descending;
model Class=Marginalna adhezija /lackfit rsq outroc=rocgraf;
run;
/*5*/
title” Univarijatna logisticka regresija-Single Epithelial Cell Size”;
proc logistic data=nova descending;
model Class=Velicina epitelne stanice /lackfit rsq outroc=rocgraf;
run;
/*6*/
title” Univarijatna logisticka regresija-Bare Nuclei”;
proc logistic data=nova descending;
model Class=Gole jezgre /lackfit rsq outroc=rocgraf;
run;
/*7*/
title” Univarijatna logisticka regresija-Bland Chromatin”;
proc logistic data=nova descending;
model Class=Normalan kromatin /lackfit rsq outroc=rocgraf;
run;
/*8*/
title” Univarijatna logisticka regresija-Normal Nucleoli”;
proc logistic data=nova descending;
model Class=Normalne jezgrice /lackfit rsq outroc=rocgraf;
run;
/*9*/
title” Univarijatna logisticka regresija-Mitoses”;
proc logistic data=nova descending;
model Class=Mitoza /lackfit rsq outroc=rocgraf;
run;

    title”Multivarijatna logisticka regresija”;
```

```
proc logistic data=nova descending;
model Class=Debljina grumena Uniformnost velicine stanice Uniformnost stanicnog oblika
Marginalna adhezija Velicina epitelne stanice Gole jezgre
Normalan kromatin Normalne jezgrice Mitoza/lackfit rsq
outroc=rocgraf;
run;

    title"Stepwise"; proc logistic data=nova descending;
model Class=Debljina grumena Uniformnost velicine stanice Uniformnost stanicnog oblika
Marginalna adhezija Velicina epitelne stanice Gole jezgre
Normalan kromatin Normalne jezgrice Mitoza/ selection=stepwise;
run;

    title"Stepwise ROC krivulja";
proc logistic data=nova descending;
model Class=Debljina grumena Uniformnost stanicnog oblika
Marginalna adhezija Gole jezgre Normalan kromatin
Normalne jezgrice /lackfit rsq outroc=rocgraf;
run;

    /*korelacija sve*/
proc reg data=nova;
model Class=Debljina grumena Uniformnost velicine stanice Uniformnost stanicnog oblika
Marginalna adhezija Velicina epitelne stanice Gole jezgre
Normalan kromatin Normalne jezgrice Mitoza / vif tol;
run;

    /*korelacija bez size i shape*/
proc reg data=nova;
model Class=Debljina grumena Marginalna adhezija Velicina epitelne stanice
Gole jezgre Normalan kromatin Normalne jezgrice Mitoza / vif
tol;
run;

    title"Multivarijatna logisticka regresija";
proc logistic data=nova descending;
model Class=Debljina grumena Marginalna adhezija Velicina epitelne stanice
Gole jezgre Normalan kromatin Normalne jezgrice Mitoza/lackfit rsq outroc=rocgraf;
run;
```

```
    title”Stepwise”;
proc logistic data=nova descending;
model Class=Debljina grumena Marginalna adhezija Velicina epitelne stanice
Gole jezgre Normalan kromatin Normalne jezgrice Mitoza/ selection=stepwise;
run;

    title”Multivarijatna logisticka regresija”;
proc logistic data=nova descending;
model Class=Debljina grumena Marginalna adhezija
Gole jezgre Normalan kromatin Normalne jezgrice /lackfit rsq outroc=rocgraf;
run;

    /*CRTANJE*/
/*1*/
title”crtanje”;
proc logistic data=nova descending;
model Class= Debljina grumena;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P za Debljinu grumena”;
proc sgplot data=crtanje;
scatter x=Debljina grumena y=prob;
run;
/*2*/
proc logistic data=nova descending;
model Class= Uniformnost velicine stanice;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P za Uniformnost velicine stanice”;
proc sgplot data=crtanje;
scatter x=Uniformnost velicine stanice y=prob;
run;

    /*3*/

    proc logistic data=nova descending;
model Class= Uniformnost stanicnog oblika;
output out=crtanje predicted=prob xbeta=logit;
```

```
run;
title”graficki prikaz P za Uniformnost stanicnog oblika”;
proc sgplot data=crtanje;
scatter x=Uniformnost stanicnog oblika y=prob;
run;
/*4*/
proc logistic data=nova descending;
model Class= Marginalna adhezija;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P Marginalnu adheziju”;
proc sgplot data=crtanje;
scatter x=Marginalna adhezija y=prob;
run;
/*5*/
proc logistic data=nova descending;
model Class= Velicina epitelne stanice;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P Velicinu epitelne stanice”;
proc sgplot data=crtanje;
scatter x=Velicina epitelne stanice y=prob;
run;
/*6*/
proc logistic data=nova descending;
model Class= Gole jezgre;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P za Gole jezgre ”;
proc sgplot data=crtanje;
scatter x=Gole jezgre y=prob;
run;
/*7*/
proc logistic data=nova descending;
model Class= Normalan kromatin;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P Normalan kromatin”;
proc sgplot data=crtanje;
```

```
scatter x=Normalan kromatin y=prob;
run;
/*8*/
proc logistic data=nova descending;
model Class= Normalne jezgrice;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P za Normalne jezgrice ”;
proc sgplot data=crtanje;
scatter x=Normalne jezgrice y=prob;
run;
/*9*/
proc logistic data=nova descending;
model Class= Mitoza;
output out=crtanje predicted=prob xbeta=logit;
run;
title”graficki prikaz P za Mitozu”;
proc sgplot data=crtanje;
scatter x= Mitoza y=prob;
run;
```

Bibliografija

[1] P. Naik, *Essentials of Biochemistry (for Medical Students)*, Jaypee, 2012.

[2] S. C. Satapathy,J. K. Mandal,S. K. Udgata i V. Bhateja *Information Systems Design and Intelligent Applications*, Springer, 2016.

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

[4] T.Hastie, R. Tibshirani i J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2016.

[5] A.Jazbec, *Odabrana statističke metode u biomedicini*, PMF-MO, nastavni materijali, 2016.

[6] V.Wagner, *Statistički praktikum 2*, PMF-MO, nastavni materijali, 2016.

[7] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), lipanj 2017

[8] Introduction to SAS. UCLA: Statistical Consulting Group, dostupno na https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/, lipanj 2017

[9] http://www.onkologija.hr, kolovoz 2017

Sažetak

U ovom radu napravili smo model predikcije malignosti tumora dojke. Tumor dojke je zloćudna bolest koja napadna sve više i više žena, ali i muškaraca. Oko trećinu svih malignih tumora u žena čini upravo tumor dojke. Također tumor dojke je visko zastupljen među uzrocima smrti žena. U radu smo koristili podatke „Breast Cancer Wisconsin (Diagnostic) Data Set“ koje je prikupio dr. William H. Wolber (University of Wisconsin Hospitals, Madison). Podaci sadrže 699 opservacija i 11 varijabli, odnosno atributa. Pri obradi podataka koristili smo metodu logističke regresije i statistički program SAS. Nakon provedenih analiza došli smo do zaključka da su varijable: Debljina grumena, Marginalna adhezija, Gole jezgre, Normalan kromatin i Normalne jezgrice statistički značajne za model predikcije malignosti tumora dojke.

Summary

In this paper, we have made a model for prediction of breast tumor malignancy. Breast cancer is a malignant disease that attacks more and more women, but also men. About a third of all malignant tumors in women are breast tumors. Also, breast cancer is highly represented among the causes of the death of a woman. In this paper we used the data "Breast Cancer Wisconsin (Diagnostic) Data Set" collected by Dr. William H. Wolber (University of Wisconsin Hospitals, Madison). The data contains 699 observations and 11 variables, respectively attributes. When processing the data we have used the method of logistic regression and the statistical program SAS. After the analysis we have come to the conclusion that the variables : Clump Thickness, Marginal Adhesion, Bare Nuclei, Bland Chromatin and Normal Nucleoli are statistically significant for the model of prediction of breast tumor malignancy.

Životopis

Rođena sam 25.06.1992. godine u Zagrebu. Pohađala sam osnovnu školu "Kustošija", te nakon toga upisujem x.gimanziju "Ivan Supek". 2011. godine upisala sam preddiplomski studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Po završetku istog, 2015. godine nastavljam studij na diplomskom sveučilišnom studiju Matematička statistika na istom fakultetu.