KM 36(1991)
pag 47 - 60

DE KUNST VAN HET VOORSPELLEN

J.C. van Houwelingen

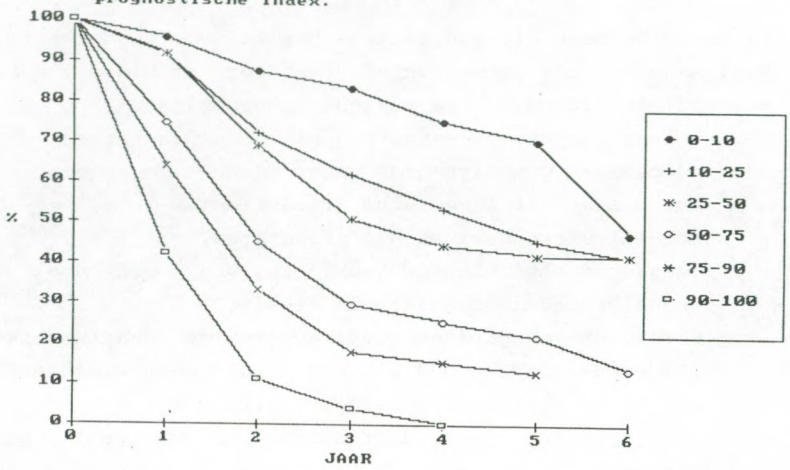
In mijn loopbaan als statisticus ben ik van tijd tot tijd heel nadrukkelijk geconfronteerd met het probleem van het voorspellen. Ik noem u de volgende voorbeelden:

- Transplantaatoverleving bij niertransplantaties. (Gegevens geanalyseerd door Jane Thorogood, Eurotransplant & Medische Statistiek).
- Morbiditeitsonderzoek bij prematuren. (Gegevens uit het POPS-project voor dit doel geanalyseerd door Saskia le Cessie).
- Follow-up van patiënten met gevorderde ovariumkanker. (Gegevens voortkomend uit een langlopende samenwerking met Dr. J.P. Neijt, Oncologie, AZU).
- "Voorspellen" van ontbrekende chromatografische gegevens. (Langlopend contact met Prof.dr. C.L de Ligny).

In een recent artikel (Van Houwelingen et. al. (1989)) werd de overleving bestudeerd van patiënten met gevorderde ovariumkanker gerekend vanaf de start van de behandeling. M.b.v. Cox' regressie is een prognostische index PI voor overleving gecontrueerd, die gebaseerd op de gegevens bij de start van de behandeling. In figuur 1 is de overleving weergegeven in zes groepen van patiënten die geformeerd zijn aan de hand van deze PI.

Afdeling Medische Statistiek
Rijksuniversiteit Leiden
Postbus 9512
2300 RA Leiden

Figuur 1
Overleving in % bij gevorderd ovariumkanker ingedeeld in percentielgroepen van de prognostische index.



De voorspellende waarde van de prognostische index PI lijkt vrij goed, maar natuurlijk dringt de vraag zich op hoe goed de voorspelling is voor nieuwe patiënten uit dezelfde populatie, dan wel voor patiënten uit andere populaties. Dit leidt tot de vraag naar de voorspellende waarde van statistische modellen. Drie problemen komen daarbij aan de orde.

1. Hoe definieer je de "voorspelfout" ? (Eng.: error rate).
2. Hoe schat je de voorspelfout?
3. Zijn de klassieke voorspellers nog te verbeteren en, zo ja, hoe?

Om de verschillende concepten te beschrijven bekijken wij eerst de simpele situatie waarbij $Y_1,\ldots,Y_n,Y_{n+1},\ldots$ een rij trekkingen is uit dezelfde kansverdeling met $E(Y)=\mu$ en variantie$(Y)=\sigma^2$. De eerste $Y_1,\ldots,Y_n$ zijn waargenomen en die willen wij gebruiken om de volgende waarnemingen te voorspellen door middel van de voorspeller $\hat{Y}$. Als voorspelfout hanteren wij de klassieke kwadratische afwijkingen. Dat leidt tot

$$MSE_{ACT} = E(Y_{nieuw}-\hat{Y})^2 = (\mu-\hat{Y})^2 + \sigma^2.$$

Let wel, de nieuwe waarde Y_{nieuw} is hierbij stochastisch, maar $\hat{Y}$ niet. Die is éénmaal bepaald en de waargenomen waarde ervan is onze voorspeller.

Het onderschrift ACT staat voor "actual" om aan te geven dat dit de voorspelfout is die wij graag zouden kennen. Als μ bekend is, zouden wij $\hat{Y}=\mu$ als voorspeller kunnen gebruiken. De voorspelfout wordt dan

$$MSE_{OPT} = \sigma^2.$$

In werkelijkheid is μ meestal onbekend (en σ^2 ook). We willen dan MSE_{ACT} schatten. Een mogelijkheid is om dat retrospectief te doen. Dat levert

$$MSE_{APP} = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i-\hat{Y})^2.$$

Het suffix APP staat voor "apparent". Het verschil tussen MSE_{ACT} en MSE_{APP} wordt in navolging van Efron(1986) het optimisme O genoemd, dus

$$O = MSE_{ACT} - MSE_{APP}.$$

De gebruikelijke voorspeller in deze situatie is $\hat{Y}$; daarvoor is

$$MSE_{APP} = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i-\bar{Y})^2 = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i-\mu)^2 - (\mu-\bar{Y})^2$$

$$MSE_{ACT} = \sigma^2 + (\mu-\bar{Y})^2$$

$$O = 2(\mu-\bar{Y})^2 + \sigma^2 - \frac{1}{n}\Sigma_{i=1}^{n}(Y_i-\mu)^2$$

Als wij ook $Y_1,\ldots,Y_n$ als stochastisch opvatten, dan krijgen wij dat

$$\Omega = E(O) = 2\sigma^2/n.$$

Deze kan worden geschat door $2s^2/n$ met $s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(Y_i-\bar{Y})^2$, de gebruikelijke schatter voor σ^2. Dit leidt tot

$$\hat{MSE}_{ACT} = MSE_{APP} + \frac{2s^2}{n} = (1+\frac{1}{n})s^2$$

als schatting voor MSE_{ACT}. Let wel, dit is slechts een schatting.

Een andere manier om MSE_{ACT} te schatten is door middel van kruisvalidatie, waarbij telkens één waargenomen Y_i uit de andere wordt voorspeld. Het gemiddelde van alle (kwadratische) fouten wordt als schatting van MSE_{ACT} gehanteerd. Met andere woorden,

$\bar{Y}_{(-i)} = \frac{1}{n-1}\Sigma_{j\neq i}Y_j$ "voorspelt" Y_i en

$MSE_{CV} = \frac{1}{n}\Sigma(Y_i-\bar{Y}_{(-i)})^2 = \frac{n}{n-1}s^2$ schat MSE_{ACT}.

Merk op dat het verschil met $\hat{MSE}_{ACT}$ heel klein is.

Na dit hele simpele voorbeeld bekijken wij het voor de praktijk relevantere regressieprobleem. Daarbij zijn $(Y_1,X_1),\ldots,(Y_n,X_n)$ waargenomen en moet de waarde van toekomstige waarnemingen Y_{nieuw} voorspeld worden bij bekende waarde van X_{nieuw}. Hierbij is X een vector van covariaten. Als model voor de relatie tussen Y en X hanteren wij

$E(Y_i|X_i) = \mu_i = X_i'\beta$ met $\dim(\beta)=p$ en

$var(Y_i|X_i) = \sigma^2$.

Het is duidelijk dat de optimale voorspelfout

$$MES_{OPT} = \sigma^2$$

wordt verkregen als het model correct is en β bekend. In werkelijkheid wordt β geschat met de gewone kleinste kwadratenschatter

$$\hat{\beta} = [\Sigma_{i=1}^{n}X_iX_i']^{-1}\Sigma_{i=1}^{n}X_iY_i .$$

De voorspelfout hangt van X_{nieuw} af en wordt gegeven door

$$MSE_{ACT}(X_{nieuw}) = (\mu_{nieuw}-X_{nieuw}'\hat{\beta})^2 + \sigma^2.$$

Om de afhankelijkheid van X_{nieuw} kwijt te raken, middelen wij over X. Er zijn daarbij twee benaderingen, namelijk de "pure replicatie" en de "stochastische X" benadering.

Bij de pure replicatie veronderstellen wij dat bij dezelfde $X_1,\ldots,X_n$ nieuwe waarnemingen worden gedaan. De gemiddelde voorspelfout is dan

$$MSE_{ACT} = \sigma^2 + \frac{1}{n}\Sigma(\mu_i - X_i'\hat{\beta})^2.$$

De klaarblijke fout is

$$MSE_{APP} = \frac{1}{n}\Sigma(Y_i - X_i'\hat{\beta})^2.$$

Als het model correct is, dan is $E(MSE_{ACT})=\sigma^2(1+\frac{p}{n})$ en $E(MSE_{APP})=\sigma^2(1-\frac{p}{n})$, dus $\Omega=E(0)=2p\sigma^2/n$. (Het laatste geldt zelfs algemeen). Een schatter $\hat{\sigma}^2$ van σ^2 kan gebruikt worden om MSE_{ACT} te schatten door $MSE_{APP}+2p\hat{\sigma}^2/n$. Dit is Mallows' C_p (Mallows (1973)). Als het model correct is, kan σ^2 geschat worden door $s^2=\Sigma(Yi-X_i'\hat{\beta})^2/(n-p)$. Dit leidt tot $\hat{MSE}_{ACT}=s^2(1+p/n)$ als eenvoudigste vorm van Mallows' C_p.

Bij de stochastische X benadering wordt verondersteld dat de X'en ook getrokken worden uit een bepaalde verdeling. Dit is de situatie van een observationele studie. In deze situatie kunnen modelfouten zonder veel bezwaar in de residuele variantie σ^2 worden geïncorporeerd. In deze situatie is

$$MSE_{ACT} = \sigma^2 + (\hat{\beta}-\beta)'E(XX')(\hat{\beta}-\beta).$$

Een manier om dit te schatten is om, zoals wij al eerder zagen, de verwachting te bepalen en die te schatten. Onder de veronderstelling dat X een multivariate normale verdeling volgt, geldt dat

$$E(MSE_{ACT}) = \sigma^2(1+\frac{p}{n-p-1}).$$

Dit leidt tot het criterium $S_p=s^2(1+\frac{p}{n-p-1})$ ter beoordeling van de voorspellende waarde van een model.
Het spelletje kan weer worden nagespeeld met kruisvalidatie. Laat $\hat{\beta}_{(-i)}$ de schatter van β zijn, gebaseerd op de waarnemingen (Y_j,X_j) met $j\neq i$. Dan wordt, per definitie,

$$MSE_{CV} = \Sigma(Yi-X_i'\hat{\beta}_{(-i)})^2/n.$$

Dit is bekend als Allen's PRESS (Allen (1971)). Men kan laten zien dat dit bij benadering gelijk is aan $ns^2/(n-p)$. Als $p<<n$, leveren alle criteria ongeveer hetzelfde op. Bij de keuze van een model, bij voorbeeld bij stapsgewijze

regressie, kunnen deze criteria gebruikt worden om het model te selecteren met de grootste voorspellende kracht. Het voordeel van kruisvalidatie boven de andere methodes is dat er nauwelijks veronderstellingen gemaakt behoeven te worden, omdat het voorspellingsproces gewoon wordt nagespeeld. Het nadeel is meer rekenwerk, maar voor gewone regressie zitten de bouwstenen al in de software.

De volgend vraag is of de klassieke kleinste kwadratenvoorspeller nog verbeterd kan worden. Een manier is de ridgeregressie van Hoerl en Kennard(1970), maar die is knap ingewikkeld. Een simpelere methode is voorgesteld door Copas(1983). Hij stelt voor om de voorspeller te laten "krimpen" volgens

$$\hat{Y}_{nieuw} = \bar{Y} + c(X_{nieuw}-\bar{X})'\hat{\beta} \ .$$

In de "pure replicatie"-situatie worden de optimale c_{opt} en zijn schatting $\hat{c}_{heur}$ gegeven door

$$c_{opt} = \frac{\beta'\Sigma(X_i-\bar{X})(X_i-\bar{X})'\beta}{\beta'\Sigma(X_i-\bar{X})(X_i-\bar{X})'\beta + \dim(\beta)\sigma^2} \quad \text{en}$$

$$\hat{c}_{heur} = \frac{SS_{Expl} - \dim(\beta)s^2}{SS_{Expl}} \ .$$

De krimpconstante c kan ook geschat worden door gebruik te maken van kruisvalidatie. Regressie van Y_i op $\hat{Y}_{(-i)}=X_i'\hat{\beta}_{(-i)}$ levert $\hat{c}_{cal}$. ("cal" staat hierbij voor calibratie).

Om uit te proberen hoe dit werkt, is een Monte Carlo-simulatie uitgevoerd met het model

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + e$$

met n=50 waarnemingen en 500 MC-replicaties. De X'en en e zijn onderling onafhankelijk N(0,1) verdeeld. De keuze voor de β's was $\beta_1=\beta=1$, $\beta_0=\beta_2=\beta_3=\beta_4=0$. De resultaten zijn weergegeven in tabel 1.

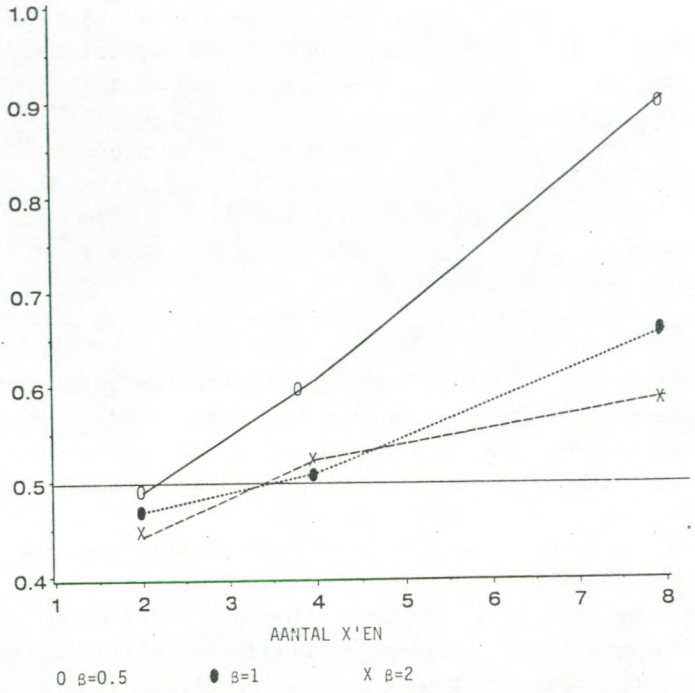
Tabel 1. Monte Carlo resultaten (500 replicaties)

	gemiddelde	standaardafwijking
s^2	1.000	0.211
MSE_{ACT}	1.118	0.074
MSE_{CV}	1.116	0.235
$\hat{c}_{cal}$	0.884	0.056
$MSE(\hat{c}_{cal})_{ACT}$	1.118	0.076
$\hat{c}_{heur}$	0.918	0.036
$MSE(c_{heur})_{ACT}$	1.115	0.073
$\hat{c}_{cal}$ beter dan c=1	57%	
$\hat{c}_{heur}$ beter dan c=1	62%	

Calibratie blijkt een geringe verbetering op te leveren. De theoretische $\hat{c}_{heur}$ doet het hier iets beter, omdat aan alle voorwaarden voor de bepaling ervan is voldaan. Het voordeel van calibratie is dat er nauwelijks veronderstellingen gemaakt behoeven te worden.
De M.C. resultaten voor n=25 bij wisselende waarde van β en het aantal X'en is weergegeven in figuur 2.
Het is duidelijk dat met de beoogde verbetering iets te winnen valt als het aantal verklarende variabelen groot is en de regressieparameters relatief klein zijn.

Bij de beschouwingen hierboven is verondersteld dat van te voren bekend is welke verklarende variabelen in het model zullen worden opgenomen. In de praktijk gaat daar echter meestal een modelbouw-fase aan vooraf waarin met vallen en opstaan het beste statistische model wordt bepaald. De effecten van deze modelconstructie zijn moeilijk te schatten en het spel is ook niet eenvoudig na te spelen omdat de statistische modelbouw meestal niet in hele strakke regels te vatten is. Om toch een mogelijkheid te hebben om een zuivere schatting van de voorspelfout te krijgen zit er niet veel anders op dan de "split sample"-aanpak te hanteren.

Figuur 2: Fractie MC-simulaties waarbij $\hat{c}_{cal}$ het beter doet dan c=1 (n=25)



Daarbij wordt een door loting bepaald deel van de waarnemingen bij het begin van de analyse opzij gezet. Als de statisticus naar beste eer en geweten zijn model bepaald heeft aan de hand van de zogeheten trainings-set, kan daarna de voorspellende waarde bepaald worden met behulp van de apart gezette waarnemingen, de validatie-set. Bovendien kan door calibratie een krimpfactor c bepaald worden die tot een betere voorspelling leidt. Eigenlijk moet dat dan ook weer gevalideerd worden, maar omdat bij het schatten van c wordt niet meer dan een simpele regressie wordt uitgevoerd, durven wij aan te nemen dat de resultaten daarvan goed

reproduceerbaar zijn. Straks zullen wij een voorbeeld van deze split-sample aanpak zien.

De bepaling van de voorspellende waarde van een voorspeller wordt lastiger als de uitkomst variabele Y een binaire variablele is, d.w.z. een variabele die slechts twee waarden kan aannemen. Het is gebruikelijk om die te coderen als 0-1 en de kansen te noteren als

$$Pr(Y=1) = \pi, \qquad Pr(Y=0) = 1-\pi.$$

Als $\hat{\pi}$ een schatter is voor π, dan zijn er verschillende mogelijkheden om de voorspelfout te definiëren. De eerste is de classificatiefout. Als wij voorspellen dat $Y_{nieuw}=1$ als $\hat{\pi}>\frac{1}{2}$, dat $Y=0$ als $\hat{\pi}<\frac{1}{2}$ en loten als $\hat{\pi}=\frac{1}{2}$, dan worden de actuele, optimale en klaarblijkelijke classificatiefout gegeven door

$$CER_{ACT} = E\{Y[\hat{\pi}<\tfrac{1}{2}]+(1-Y)[\hat{\pi}>\tfrac{1}{2}]+\tfrac{1}{2}[\hat{\pi}=\tfrac{1}{2}]\} =$$
$$= \pi[\hat{\pi}<\tfrac{1}{2}]+(1-\pi)[\hat{\pi}>\tfrac{1}{2}]+\tfrac{1}{2}[\hat{\pi}=\tfrac{1}{2}]\}$$

([..]=indicatorfunctie),

$$CER_{OPT} = \min(\pi,1-\pi) \quad \text{en}$$
$$CER_{APP} = \min(\hat{\pi},1-\hat{\pi}).$$

Een ander criterium, dat compatibel is met de maximum likelihood-methode is het MML-criterium (MML staat voor mean minus log likelihood).

$$MML_{ACT} = E\{-Y\ln(\hat{\pi})-(1-Y)\ln(1-\hat{\pi})\} =$$
$$-\pi\ln(\hat{\pi})-(1-\pi)\ln(1-\hat{\pi}).$$

Deze uitdrukking is minimaal voor $\hat{\pi}=\pi$, leidend tot

$$MML_{OPT} = -\pi\ln(\pi)-(1-\pi)\ln(1-\pi) \quad \text{en}$$
$$MML_{ACT} = MML_{OPT} + dist(\hat{\pi},\pi).$$

Het optimisme

$$O = MML_{ACT}-MML_{APP} = (\pi-\hat{\pi})\ln(\hat{\pi}/(1-\hat{\pi}))$$
$$\approx (\pi-\hat{\pi})\ln(\pi/(1-\pi))+(\hat{\pi}-\pi)^2/\{\pi(1-\pi)\}$$

heeft bij benadering verwachting 1/n. Dit leidt dus tot een eenvoudige schatting $\hat{MML}_{ACT}=MML_{APP}+1/n$. Dit staat bekend als Akaike's informatiecriterium. Men kan nagaan dat kruisvalidatie ongeveer hetzelfde resultaat oplevert. Bij

het CER-criterium is de correctie minder eenvoudig en leidt kruisvalidatie tot enigszins vreemde resultaten. Een reden om MML te prefereren.

Wij kunnen beide criteria natuurlijk ook toepassen bij regressie-problemen met binaire Y. Het favoriete model is logistische regressie, waarbij

$Pr(Y_i=1|X_i)=\pi_i$ gegeven wordt door

$\ln(\pi_i/(1-\pi_i))=X_i'\beta$.

De voorspelfout bij de pure replicatie wordt gegeven door

$$MML_{ACT} = -\frac{1}{n}\Sigma_{i=1}^{n}\{\pi_i \ln(\hat{\pi}_i)+(1-\pi_i)\ln(1-\hat{\pi}_i)\}.$$

De optimisme-correctie volgens Akaike is eenvoudig p/n, waarbij p het aantal vrije parameters is. De formule voor CER_{ACT} is analoog, maar de correctie gecompliceerd.

Kruisvalidatie om de "stochastische X"-situatie na te bootsen is natuurlijk weer mogelijk. Het vereist alleen meer rekenwerk, dat versneld kan worden door benaderingen voor $\hat{\beta}_{(-i)}$ te gebruiken. Calibratie is ook mogelijk door logistische regressie van Y_i op $X_i'\hat{\beta}_{(-i)}$. Een klein probleem is dat de constante eigenlijk ook opnieuw berekend moet worden omdat, in tegenstelling tot gewone regressie, $(\bar{Y},\bar{X})$ geen vast punt is. Een voorbeeld van het gebruik van beide criteria CER en MML bij modelbouw wordt gegeven in tabel 2. Wij zien dat bij MML het tweede model een duidelijke verbetering is t.o.v. het eerste en dat kruisvalidatie nagenoeg hetzelfde oplevert als Akaike's correctie. Bij het CER-criterium is dat allemaal minder duidelijk. Reden, mijns inziens, om het MML-criterium te verkiezen.

Tabel 2. Verschillende logistische regressiemodellen voor de POPS-data (premature en/of te lichte babies). (n=1310)
Y = morbiditeit binnen twee jaar
X_1= geboortegewicht
X_2= zwangerschapsduur

	model	
	X_1, X_2	X_1, X_2, X_1^2, X_2^2
dim(β)	3	5
CER_{APP}	0.2473	0.2458
correctie	0.0010	0.0016
$\hat{CER}_{ACT}$	0.2483	0.2474
CER_{CV}	0.2481	0.2466
MML_{APP}	0.5356	0.5180
correctie	0.0023	0.0038
$\hat{MML}_{ACT}$	0.5379	0.5218
MML_{CV}	0.5382	0.5216

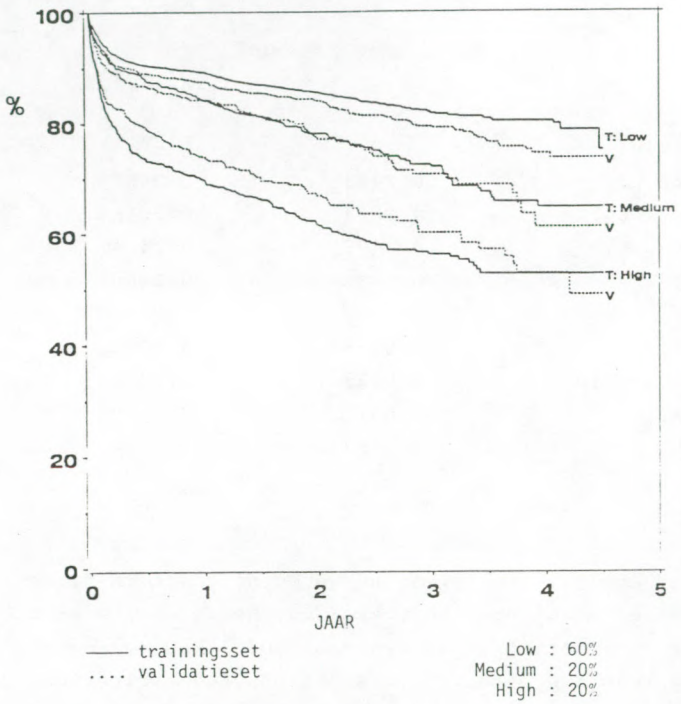
Bij analyse van levensduurgegevens zoals in figuur 1 wordt het allemaal nog ingewikkelder. Het grootste probleem levert de censurering. Cox' regressiemodel kan gehanteerd worden om de relative hazards te schatten. Een criterium voor de voorspelfout is niet eenvoudig te geven. Het is wel mogelijk om de krimp-factor te schatten met

$\hat{c}_{heur}$ = (model-chikwadraat - dim(β))/model-chikwadraat

en die te vergelijken met de calibratie-versie $\hat{c}_{cal}$ zoals die met split-sample kan worden bepaald. Wij besluiten met twee voorbeelden.

Niertranplantatie. De split-sample benadering leidde in een bepaald stadium van de analyse tot figuur 3.

Figuur 3: Niertransplantie overleving in 3 subgroepen gebaseerd op prognostische index uit validatieset.



De groepen zijn geformeerd op grond van de prognostische index $X'\hat{\beta}$ waarbij $\hat{\beta}$ geschat is in de trainingsset. Het is duidelijk dat de groepen in de validatieset dichter bij elkaar liggen dan in de trainingsset. Cox' regressie in de validatie-set leidt tot $\hat{c}_{cal}=0.64$, terwijl het uiteindelijke model in de trainingsset leidt tot $\hat{c}_{heur}=0.94$. Het is duidelijk dat alleen uitgaan van het finale model zelfs na

correctie leidt tot overschatting van de voorspellende waarde. Het uiteindelijke model gedraagt zich meer als het uitgebreide model dat alle parameters schat die bij de aanvang van de analyse in beschouwing werden genomen. Reductie van het model met stapsgewijze regressie leidt niet tot verbetering van de voorspellende waarde, alleen tot vereenvoudiging van de beschrijving.
Overigens is de aanpak bij deze niertransplantiegegevens in een later stadium verder verfijnd, wat leidde tot een betere overeenstemming tussen $\hat{c}_{cal}$ en $\hat{c}_{heur}$.

Bij de ovariumkankergegevens van figuur 1 is een vergelijking gemaakt tussen deze gegevens en de gegevens uit een Deens onderzoek (Lund et al. (1990)). De Nederlandse dataset leidde tot een prognostische index met dim(β)=7 en model-chikwadraat=67. Dat impliceert een $\hat{c}_{heur}$=0.90. Toepassing van deze index op de Deens data levert $\hat{c}$=0.83. De Deense onderzoekers construeerden een index met dim(β)=6, model-chikwadraat=103.3 en $\hat{c}_{heur}$=0.94. Toepassing op de Nederlandse dataset leverde echter slechts een $\hat{c}$=0.62 op. Het lijkt erop dat de Nederlandse index betrouwbaarder is, maar voorzichtigheid bij de interpreatie is geboden.

Referenties

Allen,D.M., Mean square error of prediction as a criterion for selecting variables, Technometrics, 13, 469-475, 1971

Copas,J.B., Regression, prediction and shrinkage, Journal of the Royal Statistical Society, Series B, 45, 311-354, 1983

Efron,B., How biased is the apparent error rate of a prediction rule, Journal of the American Statistical Association, 461-470, 1986

Hoerl,A.E. en Kennard,R.W., Ridge regression. Biased estimation for nonorthogonal problems, Technometrics, 12, 55-67, 1970

van Houwelingen,J.C.,ten Bokkel Huinink,W.W.,van der Burg,M.L.,van Oosterom,A.L. en Neijt,J.P., Predictability of of the survival of patients with advanced ovarian cancer, Journal of Clinical Oncology, 7, 769-773, 1989

van Houwelingen,J.C. en le Cessie,S., Predictive values of statistical models, Statistics in Medicine, 1990

Lund,B.,Williamson,P.,van Houwelingen,H.C. en Neijt,J.P., Comparison of the predictive power of different prognostic indices for overall survival in patients with advanced ovarian carcinoma, Cancer Research, 50, 4626-4629, 1990

Mallows,C.L., Some comments on C_p, Technometrics, 15, 661-675, 1973