Regression Analysis for Interval-Valued Data

Mousa Negash

431898

July 8, 2018

Abstract

In dit paper worden verschillende regressiemethoden uitgevoerd op symbolische data om vervolgens te oordelen welke methode het meest betrouwbaar is. De data is herschikt in de vorm van intervallen en middelpunten. Hierop werden Deming Regressie en middelpuntsregressie toegepast. De regressiemethode waarbij gebruik wordt gemaakt van empirische, symbolische statistieken (Covariantie-Regressie) geven de laagste standaarderrors in vergelijking met de eerder benoemde regressiemethoden. Om de standaarderrors te verkrijgen zijn er steekproeftrekkingen gegenereerd uit de data. Uit de resultaten valt af te leiden dat Deming Regressie minder presteert dan de andere regressiemethoden die overigens erg veel overeenkomsten vertonen in hun resultaten.

Keywords: symbolic data, linear regression, bootstrap sampling, standard error, (co-)variance, parameter estimation.

Contents

1 Introductie

Er zijn al veel analyses verricht op klassieke data: Data waarbij observaties slechts een enkele punt betreffen. Echter, zijn er ook analyses verricht op symbolische data. Hierbij kunnen de observaties slaan op meerdere waarden, modaal zijn of bevinden de observaties zich in een bepaald interval. Methoden die van toepassing zijn op klassieke data zijn mogelijk niet geschikt voor symbolische data. Hiervoor zijn nieuwe methoden aangesteld die waardevolle statistische analyses creëren. Een methode om deze data te analyseren is aangesteld door Billiard et. al, 2000. De dataset in dit paper omschrijft de hartslag en bloeddruk van meerdere patiënten. De hartslag en bloeddruk zijn gemeten over een bepaalde periode en verschilt logischerwijs over de tijd. Al deze metingen worden opgenomen in een interval. Bij deze methode worden covariantie- en correlatieformules herleid onder de aanname dat mogelijke waarden in de dataset uniform verdeeld zijn over de intervallen. Op basis van deze methode wordt een linear regressiemodel ontwikkeld. Deze aanname is van groot belang en zal ook worden aangehouden in dit paper. Het bestaan van de data uit een verzameling intervallen geeft de mogelijkheid om simpelweg te werken met de middelpunten. Door deze punten kan eenvoudig een lineaire regressie worden uitgevoerd. Verder kan met behulp van de intervallen en hun middelpunten een ander techniek, de zogeheten Deming Regressie, ons analyse uitbreiden. Hier wordt in tegenstelling tot reguliere regressie ook rekening gehouden met de fout die waargenomen wordt op front van de regressor. De ratio tussen de waargenomen fout in de afhankelijke variabele en de regressor speelt dan een rol voor de coëfficiënten die geschat worden. Kortom, er worden drie verschillende regressies toegepast: Deming Regressie, MiddelpuntsRegressie en Covariantie-Regressie (Billard et. al, 2000). Verdere informatie met betrekking tot de regressies wordt verstrekt in de sectie Methodologie. Om te testen welke regressiemethode beter presteert wordt er gekeken naar de standaardfouten die vrijkomen bij het schatten van de parameters. Hiervoor zullen willekeurige steekproeftrekkingen gegenereerd worden om de drie regressies meerdere malen uit te voeren. Bij het vrijkomen van de verschillende parameters is het mogelijk om de standaardfouten te berekenen en achteraf wordt er geconcludeerd op basis van deze standaardfouten welke regressiemethode het beste betrouwbaarheidsinterval geeft.

Onderzoeksvraag:

Welke methode werkt het best om betrouwbare standaardfouten te verkrijgen bij een regressie voor symbolische data?

2 Literatuur

Op de analyse van symbolische data verricht door Billiard en Diday wordt op voortborduurd. Hun werk omvat het toepassen van de Covariantie-Regressie op symbolische data. In Billiard en Diday wordt de symbolische data beschreven door een afhankelijke variabele Y die de hartslag omschrijft en twee afhankelijke variabelen X_1 en X_2 die respectievelijk de systolische en diastolische druk omschrijven. Er wordt gewerkt met een dataset waarvan al deze variabelen per observatie verdeeld zijn over een interval. Voor deze gegevens worden allereerst de symbolische statistieken berekend alvorens de parameters worden geschat bij de Covariantie-Regressie. Voor de formules voor de symbolische statistieken en de parameterschattingen wordt verwezen naar **Sectie 4.3**. De replicatie van de resultaten in Billiard en Diday zijn terug te vinden in de *Appendix*.

Dit paper legt de nadruk op de standaardfouten, waarvan de waarden van groot belang zijn voor een econometrist. Naast het beïnvloeden van de statistische betekenis hebben de standaardfouten ook zeker invloed op de uitspraken die hierbij gedaan kunnen worden. De focus zal voornamelijk liggen op deze contributies. Er wordt geacht dat de lezer algemene kennis heeft van (simpele) lineaire regressiemethoden en (bootstrap) sampling.

3 Data

De dataset die gebruikt wordt is afkomstig van de website van NCAA en omvat het aantal doelpogingen en het aantal doelpunten in de seizoenen 2010-2017. Aangezien voetbal een topsport is en het aantal doelpunten sterk afhangt van het aantal doelpogingen is het interessant om deze data te gebruiken. Deze statistieken betreffen het voetbalteam Florida International University Panthers (FIU) afkomstig uit divisie I. In onderstaand tabel zijn naast het seizoen en de datum, het aantal doelpogingen (SoG) en de doelpunten weergeven. De data is verdeeld in acht seizoenen. Er wordt dus voor acht seizoenen 2010-2017 acht waarnemingen gecreëerd. In tabel 1 is dus te zien dat elke waarneming corresponderend met één seizoen wordt aangegeven met observatie **u**.

Season	Date	SoG	Goals	Season	Date	SoG	Goals	Season	Date	SoG	Goals	Season	Date	SoG	Goals
	9-3	4	1		8-26	7	2		8-24	3	2		8-30	2	1
	9-5	3	0		8-28	12	1		8-31	11	2		9-1	7	5
	9-10	4	2		9-2	4	1		9-2	6	1		9-5	2	1
	9-12	6	2		9-4	5	0		9-7	4	3		9-7	10	4
	9-17	4	2		9-9	11	3		9-9	4	2		9-14	8	1
	9-19	4	1		9-11	9	1		9-14	4	1		9-20	7	0
	9-24	5	1		9-16	7	2		9-16	7	1		9-22	8	2
	10-2	2	0		9-18	12	2		9-22	3	2		9-27	6	2
2010	10-5	3	0	2011	9-23	12	1	2012	9-29	3	1	2013	9-29	9	2
(u = 1)	10-9	7	2	**(u = 2)**	10-2	3	1	**(u = 3)**	10-3	2	0	**(u = 4)**	10-6	15	4
	10-12	6	2		10-8	5	0		10-6	3	1		10-12	4	1
	10-16	1	0		10-11	6	0		10-12	5	2		10-16	5	0
	10-22	8	3		10-16	2	1		10-17	9	2		10-19	7	0
	10-24	4	0		10-19	5	1		10-21	2	1		10-26	5	4
	10-27	4	0		10-22	7	3		10-23	9	0		10-30	5	0
	10-30	8	1		10-29	2	0		10-27	4	1		11-3	4	0
	11-5	5	1		11-4	4	3		10-30	12	5		11-8	4	0
									11-3	8	1				
	8-29	4	1		8-28	4	0		8-26	3	0		8-25	4	1
	8-31	1	0		8-30	3	2		8-28	7	3		8-27	6	3
	9-5	9	1		9-3	8	2		9-2	2	1		9-1	7	2
	9-10	2	0		9-6	12	3		9-4	3	0		9-3	8	3
	9-16	2	1		9-10	9	3		9-9	8	3		9-14	4	2
	9-19	13	2		9-12	2	2		9-17	7	2		9-23	2	2
	9-24	2	2		9-21	9	7		9-20	5	0		9-26	10	5
	9-27	9	2		9-26	10	5		9-24	6	1		9-30	7	4
	10-4	1	0		9-29	6	2		9-30	5	1		7-10	7	4
2014	10-8	12	4	2015	10-3	4	0	2016	10-4	5	1	2017	10-10	13	4
(u = 5)	10-11	7	3	**(u = 6)**	10-10	5	3	**(u = 7)**	10-10	3	0	**(u = 8)**	10-14	11	5
	10-15	6	2		10-17	7	2		10-15	3	2		10-17	6	2
	10-25	11	1		10-20	6	2		10-21	5	1		10-21	14	4
	10-29	3	2		10-27	6	0		10-25	8	2		10-25	10	3
	11-2	3	0		10-31	8	0		10-29	7	4		10-29	7	2
	11-7	6	2		11-7	1	0		11-5	8	2		11-10	4	1
					11-11	8	4		11-9	11	2		11-16	3	2
					11-13	4	1		11-11	7	2		11-19	6	1
					11-15	7	1		11-13	5	0				
					11-19	7	2								

Table 1: FIU Statistieken

De data bestaande uit doelpogingen en doelpunten zijn beiden gebeurtenissen die voorkomen op een bepaald moment in de wedstrijd. Deze statistieken vallen onder count data en volgen dus een Poisson-verdeling. Er wordt ter controle een Poisson Count-Regressie uitgevoerd om te kijken of het aantal doelpogingen daadwerkelijk significant effect heeft op het aantal doelpunten in een wedstrijd. Dit is een sterke aanname doordat er een schot op doel vereist is om een doelpunt te maken.

Dependent Variable: GOALS
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Date: 06/12/18 Time: 20:58
Sample (adjusted): 1 142
Included observations: 142 after adjustments
Convergence achieved after 4 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.385868	0.156825	-2.460493	0.0139
SOG	0.133842	0.019274	6.944134	0.0000

Figure 1: CountRegressie Doelstatistieken FIU

In Figuur 1 is te zien dat het aantal doelpogingen significant effect heeft op het aantal doelpunten. De P-waarde van deze variabele ligt relatief laag (0.0000). Deze regressie is uitgevoerd in EViews en maakt gebruik van Maximum Likelihood schatting.

4 Methodologie

De data (aantal doelpogingen en doelpunten) in Tabel 1 worden onderverdeeld in 8 observaties. Elke waarneming correspondeert met een volledige voetbalseizoen. Per seizoen wordt een interval opgesteld voor de afhankelijke variabelen (aantal doelpunten) en de verklarende variabelen (aantal doelpogingen). Vanzelfsprekend bestaat het interval uit het minimum en maximum aantal doelpunten in dat specifieke seizoen. In deze sectie worden de volgende drie regressies besproken: Deming Regressie, Middelpunt-Regressie en Covariantie-Regressie. Het creëren van de samples en uitvoeren van de regressies wordt gedaan met behulp van de programma MATLAB en EViews. Ten slotte worden er in sectie 4.4 steekproeftrekkingen uit de originele data genomen om de standaardfouten te berekenen die vrijkomen bij elke regressie.

4.1 MiddelpuntsRegressie

Bij middelpuntsregressie wordt van elk waarneming (weergeven als blok) het middelpunt genomen. Het middelpunt van elke blok vormen onze datapunten en door deze punten kan een lijn worden gefit. De middelpunten worden later ook in gebruik genomen door Deming regressie. In figuur 2 zijn de intervallen visueel weergeven als rechthoeken. De data die hiervoor gebruikt is, is afkomstig uit Billiard en Diday en is ook te vinden in de Appendix. Elk hoekpunt van de rechthoek slaat op de getallen in de dataset van de Appendix. Voor elke rechthoek is een middelpunt geïllustreerd en over al deze middelpunten wordt een lijn gefit.

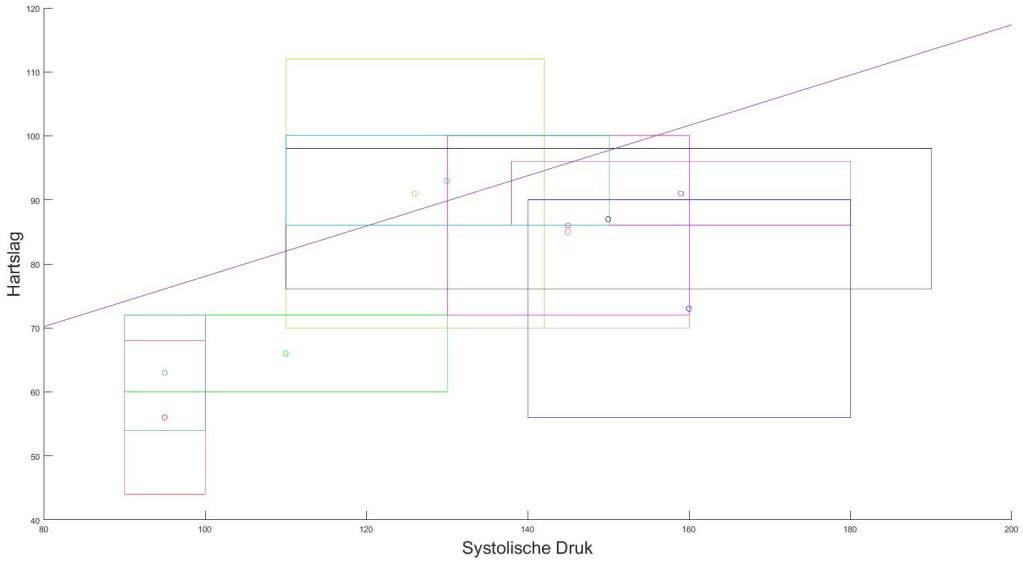

Figure 2: Voorbeeld MiddelpuntsRegressie Billiard en Diday

4.2 Deming Regressie

Deming regressie is uit op het fitten van een lijn voor een tweedimensionale dataset. De regressie wordt uitgevoerd onder de veronderstelling dat er fouten zijn in waarnemingen voor beide assen. In (1) en (2) wordt er rekening gehouden met de meetfouten respectievelijk voor de x-as (horizontale as) en de y-as (verticale as). De meetfouten ϵ_i en δ_i voor observatie i in (1) en (2) zijn onafhankelijk van elkaar. De middelpunten vormen onze waarnemingen en het interval markeert de meetfouten.

$$x_i = X_i + \epsilon_i \tag{1}$$

$$y_i = Y_i + \delta_i \tag{2}$$

Ter illustratie volgt een afbeelding van de gegevens daterend uit 2017 in Tabel 2. Hierin vormen x_i en y_i de coördinaten van het middelpunt.

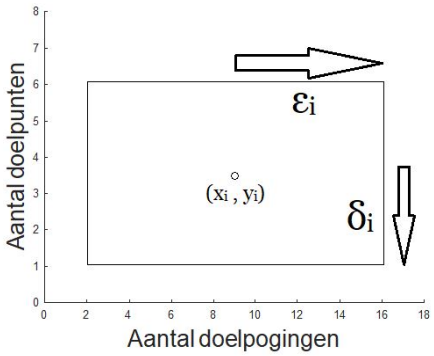

Figure 3: Voorbeeld Deming FIU 2017

Deze regressie neemt aan dat de verhouding in de gemeten fouten constant blijft. Wiskundig is dat als volgt geformuleerd:

$$\lambda_i = \frac{V(\delta_i)}{V(\epsilon_i)} \tag{3}$$

Een vector $\vec{\lambda}$ ter grootte van m observaties kan worden opgesteld bestaande uit: $\vec{\lambda} = [\lambda_1, \lambda_2, .., \lambda_m]$. In (3) staat $V(.)$ voor de variantie van de meetfouten. De volgende twee methodes leiden een constante factor af uit de vector $\vec{\lambda}$.

Gemiddelde $\vec{\lambda}$:

Hierin wordt simpelweg de gemiddelde genomen van de vector $\vec{\lambda}$ om een constante te verkrijgen:

$$\lambda = \frac{\sum_{i=1}^{m} \vec{\lambda}_i}{m}$$

Gewogen gemiddelde $\vec{\lambda}$:

Bij de gewogen gemiddelde worden er wegingen toegekend aan elk element corresponderend met zijn observatie in de vector $\vec{\lambda}$. De weging w_i hangt af van de oppervlakte van het blok.

$$\lambda = \frac{1}{W} \sum_{i=1}^{m} w_i \vec{\lambda_i}$$

Waar W gelijk is aan $\sum_{i=1}^{m} w_i$ en de weging w_i berekend is op de volgende manier:

$$w_i = \frac{A}{A_i}$$

Bij het berekenen van w_i wordt gebruik gemaakt van de oppervlakte van het interval A_i en de totale oppervlakte van alle intervallen opgeteld: $A = \sum_{u=1}^{m} A_i$. Op deze wijze wordt de weging van elk observatie i in de vector $\vec{\lambda}$ in verhouding geplaatst. Een observatie waarvan haar interval een relatief klein oppervlakte bezit krijgt een grotere weging toegekend. Dit is te danken aan het feit dat de variantie dan lager ligt ten opzichte van andere observaties die over een interval bezitten waarvan de oppervlakte relatief groter is.

De geschatte coëfficiënten β_0 en β_1 hangen af van λ op de volgende wijze. Merk op dat er vanaf nu verder wordt gewerkt met de constante λ.

$$\hat{\beta}_1 = \frac{\sigma_y^2 - \lambda\sigma_x^2 + \sqrt{(\sigma_y^2 - \lambda\sigma_x^2)^2 + 4\lambda\sigma_{xy}^2}}{2\sigma_{xy}} \tag{4}$$

In vergelijking (4) staat σ_x^2 en σ_y^2 voor de variantie van x en y. De covariantie tussen x en y wordt genoteerd met σ_{xy}. De schatting van β_0 komt tot stand in vergelijking (5):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{5}$$

Uiteindelijk is het gewenst een lijn te vinden die het best fit:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

4.3 Covariantie-Regressie

Bij de Covariantie-Regressie methode worden de onder- en bovengrens van de intervallen voor elk waarneming genoteerd als $[a_u, b_u]$ waarbij u een observatie voorstelt in de verzameling E. De verzameling E bevat symbolieke objecten en verder is $u = 1, ..., m$ met m het aantal waarnemingen. De intervallen voor de variabelen X en Y zijn als volgt:

$$X_u = \{a_u, b_u\} \; en \; Y_u = \{a_u, b_u\}$$

Met behulp van de volgende symbolische empirische functies kunnen de parameters in de regressie geschat worden.

Empirische symbolische covariantie tussen twee variabelen Y_1 en Y_2:

$$Cov(Y_1, Y_2) = \frac{1}{4m}\sum_{u\in E}(b_{1u} + a_{1u})(b_{2u} + a_{2u}) - \frac{1}{4m^2}[\sum_{u\in E}(b_{1u} + a_{1u})][\sum_{u\in E}(b_{2u} + a_{2u})] \tag{7}$$

Empirische symbolische gemiddelde van Y:

$$\bar{Y} = \frac{1}{2m}\sum_{u\in E}(b_u + a_u) \tag{8}$$

Empirische symbolische variantie van Y:

$$S_Y^2 = \frac{1}{4m}\sum_{u\in E}(b_u + a_u)^2 - \frac{1}{4m^2}[\sum_{u\in E}(b_u + a_u)]^2 \tag{9}$$

Empirische correlatiefunctie tussen twee variabelen X en Y:

$$r(X, Y) = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} \tag{10}$$

Nu alle symbolische data verkregen is, kunnen de parameters worden geschat. Eerst de parameter β_1:

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{S_X^2} = r(X, Y)(S_Y/S_X) \tag{11}$$

Ten slotte kan parameter β_0 geschat worden.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{12}$$

4.4 Standaardfouten met Sampling

In dit paper wordt de nadruk in het onderzoek gelegd op de standaardfouten. Deze fouten geven meer informatie over de betrouwbaarheid van het uitvoeren van een bepaalde regressie. Hierbij wordt de originele data gebruikt om willekeurige steekproeftrekkingen te genereren om elke regressie meerdere malen uit te voeren. In totaal worden 10000 steekproeftrekkingen gegenereerd en vervolgens wordt er voor elke verkregen trekking elk van de drie regressie toegepast. Voor elke willekeurige steekproeftrekking worden er andere parameterschattingen verkregen. Op basis van deze schattingen kunnen de standaardfouten bepaald worden. Het bepalen van deze fouten volgt later in deze sectie. Om steekproeftrekkingen te verkrijgen wordt simpelweg de data opnieuw geconstrueerd waardoor er telkens intervallen worden gegenereerd. Het aantal intervallen dat wordt gegenereerd is gelijk aan het aantal observaties. Het reconstrueren van de intervallen gebeurt als volgt:

Algorithm 1 Bootstrap sampling

1: **Input:** Interval die onder- en bovengrens bevat. Daarnaast ook een natuurlijk getal om de seed vast te stellen.
2: **Initialize:** Een random number generator (RNG) die gebruik maakt van de seed.
3: *Nu worden twee willekeurige waarden gegenereerd op basis van de RNG en zullen deze waarden de nieuwe grenzen voorstellen van ons interval. Merk op dat deze waarden niet gelijk aan elkaar kunnen zijn en de grenzen van het originele interval niet overschrijden. Noem deze getallen* ***randomOne*** *en* ***randomTwo*** *en initialiseer deze waarden op nul.*
4: **while** (**randomOne** == **randomTwo**) *//Conditie waardoor nieuwe waarden blijven gegenereerd worden*
5: **randomOne** = rand×(upperBound - lowerBound) + lowerBound;
6: **randomTwo** = rand×(upperBound - lowerBound) + lowerBound;
7: *Noot dat door de RNG de gegenereerde waarde rand in regel 5 en 6 van elkaar verschillen.*
8: *De variabelen* ***randomOne*** *en* ***randomTwo*** *worden vervolgens afgerond.*
9: **end**
10: a_u = min(**randomOne**,**randomTwo**);
11: b_u = max(**randomOne**,**randomTwo**);
12: **Output:** Interval = $[a_u, b_u]$

Deze constructie van de data vindt 10000 keer plaats en met elke steekproeftrekking worden alle drie de regressies uitgevoerd. In regel 1 is een seed vereist om verschillende reeksen van willekeurige getallen te creëeren. Het getal voor de seed is op 100 geïnitialiseerd en na elke geconstrueerde sample wordt dit getal met één verhoogd. Als gevolg ontstaan er verschillende intervallen per constructie van de data. Merk op dat in regel 8 de twee willekeurige variabelen afgerond worden. Reden voor het uitvoeren van deze handeling is dat de (count) data alleen gehele getallen beschrijft. Voor elke sample zijn $\hat{\beta}_0$ en $\hat{\beta}_1$ geschat en opgeslagen. Deze coëfficiënten worden opgeslagen in $\vec{\beta_0}$ en $\vec{\beta_1}$ beiden ter grootte van 10000. Van beide vectoren kan de variantie berekend worden en

vervolgens ook de standaardfout **SE**:

$$\mathbf{SE}\,(\beta) = \sqrt{V(\vec{\beta})}$$

Waarin V(.) staat voor de variantie en β ($\vec{\beta}$) kan verwijzen naar β_0 of β_1 ($\vec{\beta_0}$ of $\vec{\beta_1}$).

5 Resultaten

Zoals beschreven in de methodologie is voor elk seizoen het minimale en maximale aantal doelpunten en -pogingen bijgehouden in een interval. Een representatie van deze cijfers is te vinden in Tabel 2.

Seizoen	Aantal Doelpogingen		Aantal Doelpunten	
	Minimum	Maximum	Mininmum	Maximum
2010	1	8	0	3
2011	2	12	0	3
2012	2	12	0	5
2013	2	15	0	5
2014	1	13	0	4
2015	1	12	0	7
2016	2	11	0	4
2017	2	14	1	5

Table 2: FIU Intervallen DoelStatistieken

In tabel 2 is er weinig variantie te bekennen in het minimum aantal doelpogingen en -punten. Dit is te danken aan de eigenschappen van de data: Er wordt niet gescoord in alle wedstrijden in één seizoen (met uitzondering van het seizoen 2017). In figuur 4 hieronder valt ook op te merken dat het overgrote deel van de intervallen nul als minimaal aantal doelpogingen heeft. De middelpunten tonen echter een lichte positieve correlatie.

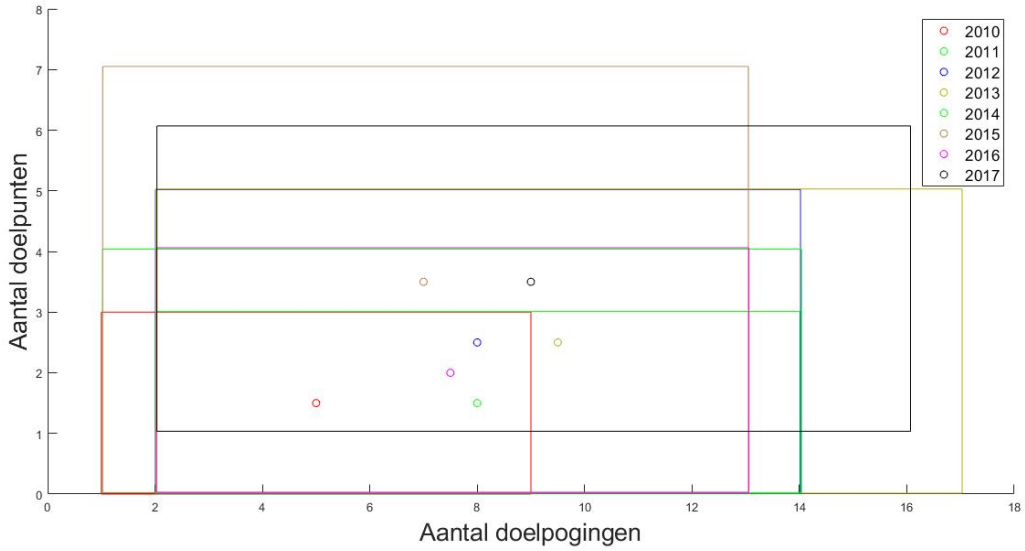

Figure 4: FIU Doelstatistieken

5.1 MiddelpuntsRegressie

Er wordt met behulp van Tabel 2 middelpunten bepaald van elk interval. De middelpunten van elk interval corresponderend met elk voetbalseizoen zijn te vinden in Tabel 3. Op basis van deze gegevens wordt een middelpuntsregressie uitgevoerd.

Seizoen	2010	2011	2012	2013	2014	2015	2016	2017
Middelpunt	(5, 1.5)	(8, 1.5)	(8, 2.5)	(9.5, 2.5)	(7.5, 2)	(7, 3.5)	(7.5, 2)	(9, 3.5)

Table 3: Middelpunten DoelStatistieken

In EViews is de regressie uitgevoerd met deze middelpunten als dataset. De resultaten zijn hieronder te vinden in figuur 5:

Dependent Variable: MIDGOALS
Method: Least Squares
Date: 06/23/18 Time: 18:36
Sample (adjusted): 1 8
Included observations: 8 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.485542	1.674089	0.290034	0.7816
MIDSOG	0.245783	0.214841	1.144023	0.2962

Figure 5: MiddelpuntsRegressie Doelstatistieken FIU

In figuur 5 wordt de coëfficiënt $\bar{\beta}_0$ weergeven door 0.4855 (C) en $\bar{\beta}_1$ door 0.2458 (SOG). De p-waarden (0.7816) en (0.2962) voor respectievelijk β_0 en β_1. Met behulp van deze waarden kan een uitspraak worden gedaan over de significantie van deze variabelen bij een vastgestelde signifi-

cantieniveau.

$$y = 0.5285 + 0.2595x \tag{13}$$

De regressievergelijking in (13) is ook geïllustreerd in onderstaand figuur. Merk op dat x en y in deze sectie staan voor het gemiddeld aantal doelpogingen en het gemiddeld aantal doelpunten. Het is echter geen probleem om *gemiddeld* in dit geval buiten beschouwing te laten. Immers is het gemiddelde genomen van beide variabelen. Aan beide kanten van de vergelijking is dus gedeeld door hetzelfde getal.

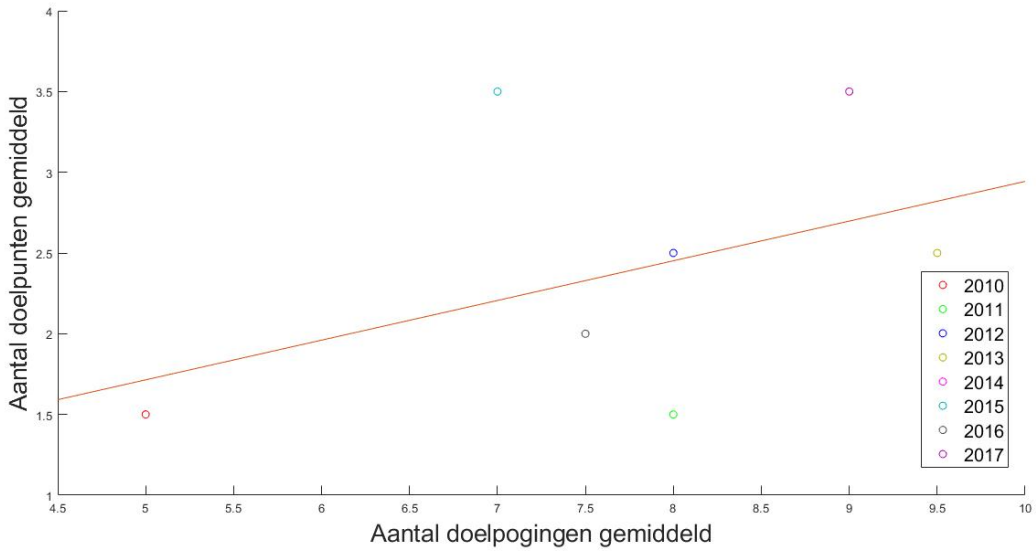

Figure 6: MiddelpuntsRegressie Doelstatistieken FIU

5.2 Deming Regressie

Aan de hand van de intervallen in tabel 2 en de middelpunten in tabel 3 wordt Deming regressie toegepast. Zoals beschreven in de methodologie is de constante λ op twee verschillende manieren berekend. Voor de originele data wordt de regressie dus tweemaal uitgevoerd. In onderstaande tabel zijn de berekeningen van λ weergeven. De uitkomsten van λ zijn berekend volgens de vergelijkingen beschreven in de sectie Methodologie:

Methode		Observatie							
		$u=1$	$u=2$	$u=3$	$u=4$	$u=5$	$u=6$	$u=7$	$u=8$
Gemiddelde λ (0.1870)	λ_u	0.1837	0.0900	0.2500	0.1479	0.1111	0.4050	0.1975	0.1111
Gewogen gemiddelde λ (0.1731)	w_u/W	0.2396	0.1677	0.1006	0.0774	0.1048	0.0653	0.1397	0.1048
	$w_u\lambda_u/W$	0.0440	0.0151	0.0252	0.0114	0.0116	0.0265	0.0276	0.0116

Table 4: Berekening constante λ met twee verschillende methoden

Gemiddelde λ

De coëfficiënten $\hat{\beta}_0$ en $\hat{\beta}_1$ zijn geschat volgens de vergelijkingen (11) en (12):

$$y = -3.5472 + 0.8523x \tag{14}$$

Gewogen gemiddelde λ

Ditmaal wordt de regressie uitgevoerd met λ berekend volgens de gewogen gemiddelde:

$$y = -3.7567 + 0.8828x \tag{15}$$

In figuur 7 zijn beide regressielijnen met alle originele intervallen uit tabel 2 geïllustreerd. Het valt op dat de Deming regressie uitgevoerd met een λ (waarvan de gewogen gemiddelde genomen is) steiler verloopt.

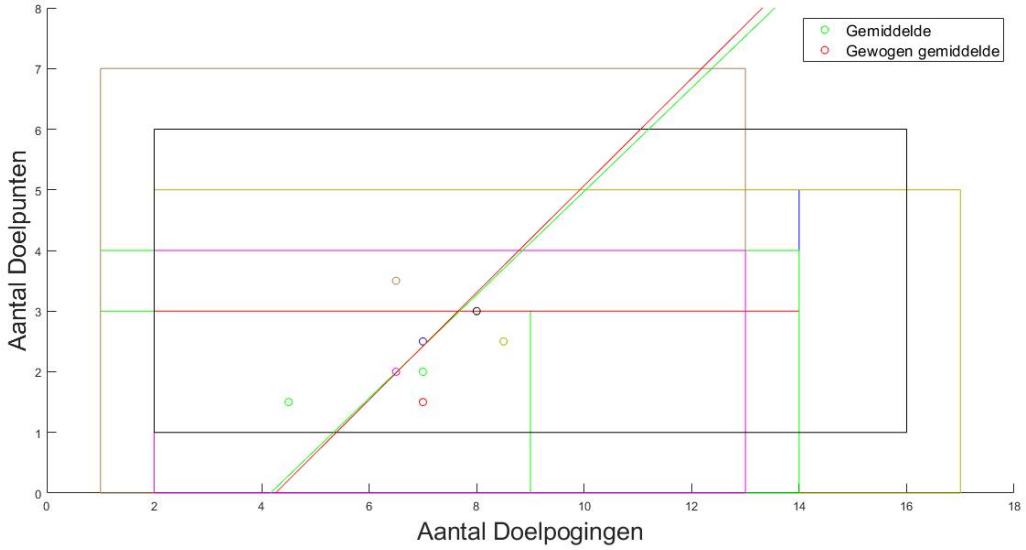

Figure 7: Deming Regressie Doelstatistieken FIU

5.3 Covariantie-Regressie

In deze sectie wordt aan de hand van covariantie-regressie een analyse uitgevoerd op de doelstatistieken van het voetbalteam FIU. Aan de hand van vergelijkingen (7) - (12) zijn de empirische symbolische statistieken berekend. Voor de eenvoudigheid omtrent de calculaties refereert X ook in deze sectie naar het aantal doelpogingen en Y naar het aantal doelpunten. De statistieken zijn hieronder te vinden in tabel 5.

$\bar{X} = 6.8750$	$\bar{Y} = 2.3125$
$S_X^2 = 1.2344$	$S_Y^2 = 0.4336$
$Cov(X, Y) = 0.3204$	$r(X, Y) = 0.4378$
$\hat{\beta}_0 = 0.5285$	$\hat{\beta}_1 = 0.2595$

Table 5: Analyse op basis van Covariantie-Regressie

Er volgt nog een korte toelichting voor tabel 5. In de eerste rij zijn de symbolische gemiddelden ($\bar{X}$ en $\bar{Y}$) berekend van het aantal doelpogingen en -punten. Een rij daaronder idem voor de symbolische variantie (S_X^2 en S_Y^2). In de derde rij worden respectievelijk de covariantie en correlatie tussen de twee variabelen berekend. Ten slotte, staan in de laatste rij de geschatte beta's genoteerd die de volgende vergelijking geeft:

$$y = 0.5285 + 0.2595x \tag{16}$$

Merk op dat deze regressievergelijking overeenkomt met de vergelijking die uitkomt bij de middelpuntsregressie in (13).

5.4 Standaardfouten met Sampling

Voor alle drie de methoden is slechts gebruik gemaakt van één sample. Om een betere inschatting te krijgen hoe betrouwbaar de beta's zijn geschat, zijn er meer samples vereist. De betrouwbaarheid wordt beoordeeld op basis van de verkregen standaardfouten (**Sectie 4.4**).

Voor de Deming Regressie is een lichte aanpassing vereist in algoritme 1. Aangezien de intervallen bestaan uit natuurlijke getallen en elke steekproeftrekking slechts 8 observaties bevatten, is het mogelijk dat de covariantie tussen X en Y in enkele gevallen zeer klein wordt (of nul). Bij Deming Regressie vormt dit een probleem doordat parameter β_1 geschat wordt op basis van deze covariantie. De term σ_{xy} in (4) in de noemer leidt bij relatief kleine waarden of nul tot relatief grote waarden voor β_1 of zelfs oneindig (indien $\sigma_{xy} = 0$). Er wordt gepleit voor een lichte correctie voor de gegenereerde steekproeftrekkingen. Bij het genereren van steekproeftrekkingen voor de Deming Regressie dienen alleen de samples gebruikt te worden waarvan de covariantie tussen het aantal doelpunten en -pogingen een bepaalde grenswaarde overschrijdt. Deze grenswaarde is vastgesteld op een constante **c** maal de covariantie tussen X en Y op basis van de originele data. In het geval dat een steekproeftrekking niet aan deze voorwaarde voldoet telt deze trekking ook niet mee en wordt er ook geen Deming Regressie op toegepast. Als gevolg worden er 10000 steekproeftrekkingen gegenereerd waarvan σ_{xy} relatief hoog genoeg is om parameter β_1 te schatten. De constante c die de ondergrens van de covariantie σ_{xy} bepaalt is gelijkgesteld aan 10. De covariantie tussen de oorspronkelijke X en Y bedraagt 0.3661.

Regressiemethode	$V(\beta_0)$	$SE(\beta_0)$	$V(\beta_1)$	$SE(\beta_1)$
MiddelpuntsRegressie	1.9564	1.3987	0.0494	0.2222
Deming Regressie$_{\bar{\lambda}}$	9.2861	3.0473	0.2108	0.4591
Deming Regressie$_{\bar{\lambda}^*}$	8.7974	2.9660	0.2005	0.4478
Covariantie-Regressie	1.9564	1.3987	0.0494	0.2222

Table 6: Varianties en Standaardfouten van de parameters bij verschillende regressies

Bij het éénmaal uitvoeren van de middelpuntsregressie en de covariantie-regressie bleken de regressievergelijkingen (13) en (16) precies overeen te komen. Na het genereren van de steekproeftrekkingen blijken deze methoden alweer dezelfde resultaten te leveren volgens tabel 6. Deming Regressie levert in vergelijking met de andere regressiemethoden hogere standaardfouten op. De variant waarbij gebruik gemaakt is van de gewogen gemiddelde van λ (aangegeven met $\bar{\lambda}^*$ in tabel 6) levert lagere standaardfouten op in vergelijking met de *reguliere gemiddelde* ($\bar{\lambda}$).

6 Conclusie

In dit paper zijn er verschillende manieren gebruikt om symbolische data te analyseren. De data zelf is op meerdere manieren bewerkt om alvorens een regressie erop toe te passen. Voor Deming Regressie en middelpuntsregressie zijn er intervallen en middelpunten gecreëerd en voor de berekening van de standaardfouten zijn er steekproeftrekkingen gegenereerd. Een overzicht van de resultaten is hieronder weergeven in tabel 7:

	MiddelpuntsRegressie	Deming Regressie		Covariantie-Regressie
		Normale Gemiddelde λ	Gewogen Gemiddelde λ	
Enkelvoudige Regressie	$y = 0.5285 + 0.2595x$	$y = -3.5472 + 0.8523x$	$y = -3.7567 + 0.8828x$	$y = 0.5285 + 0.2595x$
Standaardfouten β	$SE(\beta_0) = 1.3987$ $SE(\beta_1) = 0.2222$	$SE(\beta_0) = 3.0473$ $SE(\beta_1) = 0.4591$	$SE(\beta_0) = 2.9660$ $SE(\beta_1) = 0.4478$	$SE(\beta_0) = 1.3987$ $SE(\beta_1) = 0.2222$

Table 7: Overzicht Resultaten Regressiemethoden inclusief Standaardfouten

Door de restrictie op de covariantie van willekeurige steekproeftrekkingen bij het toepassen van Deming Regressie kunnen de standaardfouten verkregen uit de verschillende methoden niet vergeleken worden. De geselecteerde steekproeftrekkingen die voldoen aan de gestelde eis worden dus ook gebruikt door middelpuntsregressie en covariantie-regressie om een conclusie te kunnen trekken. Wederom zijn de standaardfouten van de parameters weer hetzelfde en vallen ook lager uit dan de standaardfouten verkregen uit Deming Regressie in tabel 7: $SE(\beta_0) = 2.3562$ en $SE(\beta_1) = 0.3648$.

De regressiemethoden middelpuntsregressie en covariantie-regressie zijn op basis van hun lagere standaardfouten de beste regressiemethoden die werken met symbolische data. De standaardfouten die vrijkomen bij de parameterschattingen van Deming Regressie worden lager naarmate de ondergrens voor de covariantie σ_{xy} wordt verhoogd. Dit kan worden gedaan door de constante **c**

te verhogen. Verhoging van deze constante gaat gepaard met langere rekentijd aangezien meerdere steekproeftrekkingen niet meer zullen voldoen aan de strengere covariantie-eis.

Hoewel de data (doelpunten en doelpogingen) in de vorm van count data een interessant rol heeft gespeeld bij het uitvoeren van de regressies, heeft het toch zijn nadelen. Met name de lage covariantie σ_{xy} voor de Deming Regressie vormde een groot probleem en in dit paper werden dan ook maatregelen getroffen die niet noodzakelijk waren wanneer de gegevens meer spreiding hadden. Ook het relatief lage aantal observaties maakte het interpreteren van de parameters lastig. De lezer dient zelf te concluderen of een variabele wel significant is op basis van de P-waarde.

Echter zijn de gelijke resultaten van de middelpuntsregressie en de covariantie-regressie zeer indrukwekkend. In dit paper ontbreekt het bewijs dat deze twee methoden (volledig) identiek zijn. Dit is zeker interessant om te onderzoeken. Verder zijn er wellicht meer manieren om Deming Regressie te verbeteren om lagere standaardfouten te verkrijgen. Hierbij kan λ op een ander manier worden vastgesteld of kunnen de meetfouten een ander weging toegekend krijgen. Voor de eenvoudigheid is er in dit gehele paper gewerkt met slechts één afhankelijke variabele. In praktijk is het natuurlijk gewenst om uitspraken te doen over modellen met meerdere afhankelijke variabelen hetgeen ongetwijfeld de besproken regressiemethoden complexer maken.

7 Referenties

NCSS. 2001. *Deming Regression.* Chapter 303. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf

Billiard et. al. 2000. *Regression Analysis for Interval-Valued Data*

Wei Xu. 2010. *Symbolic Data Analysis: Interval-valued Data Regression.* https://getd.libs.uga.edu/pdfs/xu_wei_201008_phd.pdf

Karen Grace-Martin. *The Analysis Factor: Poisson Regression Analysis for Count Data.* https://www.theanalysisfactor.com/ poisson-regression-analysis-for-count-data/

8 Appendix

In deze sectie wordt aan de hand van covariantie-regressie een analyse uitgevoerd op de data afkomstig uit Billiard en Diday. Aan de hand van vergelijkingen (7) - (12) zijn de empirische symbolische statistieken berekend. De gegevens zijn hieronder beschreven in tabel 8 waarin **u** de observatie voorstelt. Verder is voor elke ander kolom een interval te vinden voor de variabelen Y, X_1 en X_2. Vanzelfsprekend dient het getal voor de komma voor te stellen als ondergrens en het getal achter de komma als bovengrens.

u	Y Pulse Rate	X_1 Systolic Pressure	X_2 Diastolic Pressure
1	44, 68	90, 100	50, 70
2	60, 72	90, 130	70, 90
3	56, 90	140, 180	90, 100
4	70, 112	110, 142	80, 108
5	54, 72	90, 100	50, 70
6	70, 100	130, 160	80, 110
7	**63, 75**	**60, 100**	**140, 150**
8	72, 100	130, 160	76, 90
9	76, 98	110, 190	70, 110
10	86, 96	138, 180	90, 110
11	86, 100	110, 150	78, 100

Table 8: Data replicatie

Merk op dat in tabel 8 observatie $\mathbf{u} = \mathbf{7}$ vetgedrukt is. In Billiard en Diday wordt de regel nageleefd dat de diastolische bloeddruk lager dient te zijn dan de systolische bloeddruk. Dit is voor de zevende observatie niet het geval. In de berekeningen van de symbolische statistieken hieronder in tabel 9 is deze observatie dan ook buiten beschouwing gelaten. In de onderstaande tabel is X beschreven door de systolische bloeddruk en Y door de hartslag. Er vallen verschillen op tussen de symbolische statistieken berekend in de replicatie en in Billiard en Diday zelf. Deze verschillen zijn te danken aan het feit dat in Billiard en Diday gewerkt is met getallen die niet afgerond zijn. Deze gegevens zijn niet beschikbaa voor ons.

$\bar{X} = 131.5$	$\bar{Y} = 79.1$
$S_X^2 = 543.5$	$S_Y^2 = 162.29$
$Cov(X,Y) = 209.85$	$r(X,Y) = 0.7066$
$\hat{\beta}_0 = 28.3221$	$\hat{\beta}_1 = 0.3861$

$\bar{X} = 131.5$	$\bar{Y} = 79.1$
$S_X^2 = 495.41$	$S_Y^2 = 162.29$
$Cov(X,Y) = 194.170$	$r(X,Y) = 0.685$
$\hat{\beta}_0 = 27.639$	$\hat{\beta}_1 = 0.392$

Table 9: Analyse op basis van Covariantie-Regressie. Links de resultaten van op basis van de data in tabel 8 en rechts de resultaten op basis van de niet afgeronde gegevens in Billiard en Diday.

In de eerste twee rijen zijn de symbolische gemiddelden ($\bar{X}$ en $\bar{Y}$) en varianties (S_X^2 en S_Y^2) berekend. In de derde rij worden respectievelijk de covariantie en correlatie tussen de twee variabelen berekend. Ten slotte, staan in de laatste rij de geschatte beta's genoteerd die de volgende vergelijking geeft:

$$y = 28.3221 + 0.3861x$$

De regressievergelijking die tot stand komt uit het paper van Billiard en Diday:

$$y = 27.639 + 0.392x$$