H.9 The linear model (regression)

9.1 What will this chapter tell me?

9.2 An Introduction To the Linear Model (Regression)

9.2.1 The Linear Model With One Predictor

b_1 en b_0 zijn de regressie coëfficiënten. We kunnen de uitkomsten voorspellen aan de hand van: y = ax + b (alleen vullen we andere waarden in dan ax en b, maar principe formule blijft hetzelfde).
Predicted value/Voorspelde waarde = De waarde van een uitkomstvariabele gebaseerd op specifieke waarden van de voorspellende variabele of variabelen die in een statistisch model worden geplaatst.

9.2.2 The Linear Model With Several Predictors

B_0 is de constante. Regressieanalyse is een term voor het aanpassen van een lineair model aan gegevens en het gebruiken ervan om de waarden van een uitkomstvariabele (afhankelijke variabele) te voorspellen op basis van een of meer voorspellende variabelen (onafhankelijke variabelen). Met een voorspellende variabele heet het vaak een eenvoudige regressie, met meerdere variabelen een meervoudige regressie.
Outcome variable/Uitkomstvariabele = Een variabele waarvan we de waarden proberen te voorspellen op basis van een of meer voorspellende variabelen.
Predictor variable/Voorspellende variabele = Een variabele die wordt gebruikt om te proberen waarden van een andere variabele, de zogenaamde uitkomstvariabele, te voorspellen.
Simple regression/Eenvoudige regressie = Een lineair model waarin één variabele of uitkomst wordt voorspeld uit één enkele voorspellende variabele. Het model neemt de vorm aan:

$$Y_1 = (b_0 + b_1 X_{1i}) + \varepsilon_i$$

waarin Y de uitkomstvariabele is, X de voorspeller, b_1 de regressiecoëfficiënt die bij de voorspeller hoort en b_0 de waarde van de uitkomst is wanneer de voorspeller nul is. ε is de error.
Multiple regression/Meervoudige regressie = Een uitbreiding van enkelvoudige regressie waarbij een uitkomst wordt voorspeld door een lineaire combinatie van twee of meer voorspellende variabelen. De vorm van het model is:

$$Y = (b_0 + b_1 X_{1i} + b_2 X_{i}) + \varepsilon_i$$

waarin de uitkomst wordt aangeduid met Y, en elke voorspeller wordt aangeduid met X. Aan elke voorspeller is een regressiecoëfficiënt *b* verbonden, en b_0 is de waarde van de uitkomst wanneer alle voorspellers nul zijn.

9.2.3 Estimating the Model

Het lineaire model is een veelzijdig model om de relatie tussen een of meer voorspellende variabelen en een uitkomstvariabele te beschrijven. Ongeacht het aantal voorspellers kan het model geheel bij een constante (b_0) en bij parameters geassocieerd met elke voorspeller (*bs*) beschreven worden. De parameters worden geschat aan de hand van de methode van Least Squares.
Residual/Residueel = Het verschil tussen de waarde die een model voorspelt en de waarde die is waargenomen in de gegevens waarop het model is gebaseerd. In feite een fout. Wanneer het residu wordt berekend voor elke waarneming in een gegevensverzameling, wordt de resulterende verzameling residuen genoemd.
Residual sum of squares (SS_R) /Residuele som van de kwadraten = Een maat voor de variabiliteit die niet kan worden verklaard door het model dat op de gegevens past. Het is de totale gekwadrateerde

afwijking tussen de waarnemingen en de waarde van die waarnemingen die wordt voorspeld door het model dat op de gegevens past.
Ordinary least squares (OLS) = Een regressiemethode waarbij de parameters van het model worden geschat met behulp van de methode van de kleinste kwadraten.

9.2.4 Assessing the Goodness of Fit, Sums of Squares, R and R^2

Het gemiddelde van de uitkomst is een model waarin geen relatie is tussen de variabelen: als een variabele verandert, blijft de voorspelling voor de andere constant. Als er een gemiddelde van een model genomen wordt, kunnen de verschillen uitgerekend worden tussen de data en het gemiddelde → sum of squares R^2 uitrekenen, representeert hoe goed het gemiddelde is as een model van de geobserveerde uitkomsten. Dan kan een lineair model gemaakt worden en vergeleken worden met het basismodel van 'geen relatie'. Daarnaast is een tweede gebruik van de sum of squares bij het beoordelen van het model de F-toets.
Goodness of fit = Een index van hoe goed een model past bij de gegevens waaruit het werd gegenereerd. De fit is gewoonlijk gebaseerd op de mate waarin de door het model voorspelde gegevens overeenkomen met de werkelijk verzamelde gegevens.
Total sum of squares/Totale som van kwadraten = Een maat voor de totale variabiliteit binnen een reeks waarnemingen. Het is de totale gekwadrateerde afwijking tussen elke waarneming en het totale gemiddelde van alle waarnemingen.
Model sum of squares = Een maatstaf voor de totale variabiliteit waarmee een model rekening kan houden. Het is het verschil tussen de totale som van de kwadraten en de residuele som van de kwadraten.
Mean Squares/Gemiddelde kwadraten = Een maat voor de gemiddelde variabiliteit. Voor elke som van kwadraten (die de totale variabiliteit meet) is het mogelijk gemiddelde kwadraten te creëren door te delen door het aantal dingen dat wordt gebruikt om de som van kwadraten (of een functie daarvan) te berekenen.
F-statistic/F-statistiek = Een toetsstatistiek met een bekende waarschijnlijkheidsverdeling (de F-distributie). Het is de verhouding tussen de gemiddelde variabiliteit in de gegevens die door een bepaald model kan worden verklaard en de gemiddelde variabiliteit die door datzelfde model niet wordt verklaard. Zij wordt gebruikt om de algemene geschiktheid van het model te toetsen bij eenvoudige regressie en meervoudige regressie, en om te toetsen op algemene verschillen tussen groepsgemiddelden bij experimenten.

9.2.5 Assessing individual predictors

Een regressiecoëfficiënt van 0 betekent:

- Een eenheidsverandering in de voorspellende variabele resulteert in geen verandering in de voorspelde waarde van de uitkomst (de voorspelde waarde van de uitkomst is constant)
- Het lineaire model is 'plat'(lijn devieert niet van de horizontale lijn)

Dus als een variabele significant een uitkomst voorspelt moet het een b-waarde hebben dat groter is dan 0.
Berekeing t-toets:

$$t = \frac{b_{geobserveerd} - b_{verwacht}}{SE_b} = \frac{b_{geobserveerd}}{SE_b}$$

De $b_{verwacht}$ is de waarde van b die we verwachten als de nulhypothese waar is → aka 0 → uit de formule.
T-statistic/T-statistiek = Een toetsstatistiek met een bekende waarschijnlijkheidsverdeling (de *t*-verdeling). In de context van het lineaire model wordt deze gebruikt om te toetsen of een *b*-waarde significant verschilt van nul; in de context van experimenteel werk vertegenwoordigt deze *b*-waarde het verschil tussen twee gemiddelden en dus is *t* een toets of het verschil tussen deze gemiddelden significant verschilt van nul.

9.3 Bias In Linear Models?

Generalization/Generalisatie = Het vermogen van een statistisch model om iets te zeggen dat verder gaat dan de reeks waarnemingen die het heeft voortgebracht. Als een model generaliseert, wordt aangenomen dat voorspellingen uit dat model niet alleen kunnen worden toegepast op de steekproef waarop het is gebaseerd, maar op een bredere populatie waaruit de steekproef afkomstig is.

9.3.1 Outliers

Een uitschieter is een geval dat sterk afwijkt van de hoofdtrend in de gegevens en kunnen de schattingen van de regressiecoëfficiënten beïnvloeden (lijn minder stijl en hogere constante).
Unstandardized residuals/Niet-gestandaardiseerde residuen = De residuen van een model uitgedrukt in de eenheden waarin de oorspronkelijke uitkomstvariabele is gemeten.
Standardized residuals/Gestandaardiseerde residuen = De residuen van een model uitgedrukt in standaardafwijkingseenheden (bv z-scores). Gestandaardiseerde residuen met een absolute waarde groter dan 3,29 (gewoonlijk zeggen wij 3) zijn reden tot bezorgdheid omdat in een gemiddelde steekproef zo'n hoge waarde waarschijnlijk niet toevallig voorkomt; indien meer dan 1% van onze waarnemingen gestandaardiseerde residuen heeft met een absolute waarde groter dan 2,58 (gewoonlijk zeggen wij 2.5) dan is er bewijs dat het foutenniveau in ons model onaanvaardbaar is (het model past vrij slecht bij de steekproefgegevens); en als meer dan 5% van de waarnemingen gestandaardiseerde residuen heeft met een absolute waarde groter dan 1,96 (of 2 voor het gemak) dan is er ook bewijs dat het model een slechte weergave is van de werkelijke gegevens.
Studentized residuals/Gestudentiseerde residuen = Een variatie op gestandaardiseerde residuen. Gestudentiseerde residuen zijn het niet-gestandaardiseerde residu gedeeld door een schatting van zijn standaardafwijking die van punt tot punt varieert. Deze residuen hebben dezelfde eigenschappen als de gestandaardiseerde residuen, maar geven gewoonlijk een nauwkeuriger schatting van de foutvariantie van een specifiek geval.

9.3.2 Influential Cases

Je kan ook kijken of bepaalde gevallen een te grote invloed uitoefenen op de parameters van het model (aka als we een bepaald geval zouden schrappen, hoe anders zouden de regressiecoëfficiënten dan zijn)? Er zijn meerdere statistieken om de invloed van een bepaald geval (variabele) te meten:

- Gecorrigeerde voorspelde waarde
- Verwijderd residu
- Studentized deleted residual
- Cook's afstand
- Hefboom/hoedwaarde
- Mahalanobis-afstanden
- DFBeta
- Standardized DFBeta
- DFFit
- Standardized DFFit
- Covariantieverhouding

Adjusted predicted value/Gecorrigeerde voorspelde waarde = Een maatstaf voor de invloed van een bepaald geval van gegevens. Het is de voorspelde waarde van een geval uit een model dat is geschat zonder dat geval in de gegevens op te nemen. De waarde wordt berekend door het model opnieuw te schatten zonder het geval in kwestie, en vervolgens dit nieuwe model te gebruiken om de waarde van het uitgesloten geval te voorspellen. Als een geval geen grote invloed op het model uitoefent, moet de voorspelde waarde vergelijkbaar zijn, ongeacht of het model met of zonder dat geval werd geschat. Het verschil tussen de voorspelde waarde van een geval uit het model wanneer dat geval

werd opgenomen en de voorspelde waarde uit het model wanneer het niet werd opgenomen, is de DFFit.
Deleted residual/Verwijderd residu = een maatstaf voor de invloed van een bepaald geval van gegevens. Het is het verschil tussen de aangepaste voorspelde waarde voor een bepaald geval en de oorspronkelijke waargenomen waarde voor dat geval.
Studentized deleted residual = Een maatstaf voor de invloed van een bepaald geval van gegevens. Dit is een gestandaardiseerde versie van het geschrapte residu.
Cook's distance/Cook's afstand = Een maatstaf voor de algemene invloed van een geval op een model. Cook en Weisberg (1982) hebben gesuggereerd dat waarden groter dan 1 reden tot bezorgdheid kunnen zijn.
Leverage - hat value/Hefboom - hoedwaarde = Hefboomstatistieken (of hoedwaarden) meten de invloed van de waargenomen waarde van de uitkomstvariabele op de voorspelde waarden. De gemiddelde hefboomwaarde is (k+1)/n, waarbij k het aantal voorspellers in het model is en n het aantal deelnemers. De leverage-waarden kunnen liggen tussen 0 en (het geval heeft geen enkele invloed) en 1 (het geval heeft volledige invloed op de voorspelling). Indien geen enkel geval een buitensporige invloed op het model uitoefent, zouden we verwachten dat alle hefboomwaarden dicht bij de gemiddelde waarde liggen. Hoaglin en Welsch (1978) bevelen aan gevallen te onderzoeken met waarden groter dan tweemaal het gemiddelde (2(k+1)/n) en Stevens (2002) beveelt aan driemaal het gemiddelde (3(k+1)/n) te gebruiken als een grenswaarde voor het identificeren van gevallen die een ongepaste invloed hebben.
Mahalanobis distances/Mahalanobis-afstanden = Deze meten de invloed van een geval door te kijken naar de afstand van de gevallen tot het (de) gemiddelde(n) van de voorspellende variabele(n). Men moet op zoek gaan naar de gevallen met de hoogste waarden. Het is niet gemakkelijk om een grenswaarde vast te stellen waarop men zich zorgen moet maken, hoewel Barnett en Lewis (1978) een tabel hebben opgesteld met kritische waarden die afhankelijk zijn van het aantal voorspellers en de steekproefgrootte. Uit hun werk blijkt duidelijk dat zelfs bij grote steekproeven (N = 500) en vijf voorspellers waarden boven 25 reden tot bezorgdheid geven. In kleinere steekproeven (N = 100) en met minder voorspellers (namelijk drie) zijn waarden hoger dan 15 problematisch, en in zeer kleine steekproeven (N = 30) met slechts twee voorspellers moeten waarden hoger dan 11 worden onderzocht.
DFBeta = Een maat voor de invloed van een geval op de waarden van b_i in een regressiemodel. Indien wij een regressieparameter b_i schatten en dan een bepaald geval schrappen en dezelfde regressieparameter b_i opnieuw schatten, dan zou het verschil tussen deze twee schattingen DFBeta zijn voor het geval dat werd geschrapt. Door te kijken naar de waarden van de DFBeta's is het mogelijk gevallen te identificeren die een grote invloed hebben op de parameters van het regressiemodel; de grootte van DFBeta's zal echter afhangen van de meeteenheden van de regressieparameter.
Standardized DFBeta = Een gestandaardiseerde versie van DFBeta. Deze gestandaardiseerde waarden zijn makkelijker te gebruiken dan DFBeta omdat universele afkappunten kunnen worden toegepast. Stevens (2002) stelt voor om te kijken naar gevallen met absolute waarden van meer dan 2.
DFFit = Een maat voor de invloed van een zaak. Het is het verschil tussen de aangepaste voorspelde waarde en de oorspronkelijke voorspelde waarde van een bepaald geval. Als een geval niet invloedrijk is, moet zijn DFFit nul zijn - wij verwachten dus dat niet-invloedrijke gevallen kleine DFFit-waarden hebben. Wij hebben echter het probleem dat deze statistiek afhangt van de meeteenheden van de uitkomst, en dus zal een DFFit van 0,5 zeer klein zijn als de uitkomst varieert van 1 tot 100, maar zeer groot als de uitkomst varieert van 0 tot 1.
Standardized DFFit = Een gestandaardiseerde versie van DFFit.
Covariance Ratio (CVR)/Covariantieverhouding = Een maatstaf voor de invloed van een geval op de variantie van de parameters in een regressiemodel. Als deze ratio dicht bij 1 ligt, heeft het geval zeer weinig invloed op de varianties van de modelparameters. Belsey et al. (1980) bevelen het volgende aan: als de CVR van een geval groter is dan 1 + [3(k + 1)/n] dan zal het schrappen van dat geval de