College 7: Regressie

Regressie

We blijven bezig dit college bezig met relaties tussen twee numeriek variabele, maar we gaan de asymmetrische kant bekijken. We gaan de ene variabele proberen te voorspellen op basis van de ander. Je moet een afhankelijke en een onafhankelijke variabele aanwijzen. Het heeft te maken met het doel van voorspellen. Voorspellen is een stap op weg naar verklaren van dingen. De correlatie geeft aan hoe sterk het verband is. Vorige keer hebben we de verspreiding van de punten in een scatterplot besproken. Vandaag gaan we ons bezig houden met de lijn in een scatterplot. Deze lijn noemen we de regressielijn.

Wanneer we willen voorspellen hebben we een onderscheid nodig tussen een uitkomst (respons) en voorspellen (predictor). Je gaat proberen om de relatie te vangen met een rechte lijn. Dit is alleen zinvol bij een lineaire relatie. Bij een perfecte voorspelling staan alle punten op de lijn, maar in de sociale wetenschappen hebben we nooit een perfecte voorspelling.

Stappenplan

Wanneer we een regressielijn maken gebruiken we het volgende stappenplan:

1. Bij het beschrijven van een relatie ga je eerst het scatterplot bekijken (vorm, uitbijters, lineaire relatie?). Wanneer je dit niet doet kan je fouten maken en onzinnige uitspraken doen over je data.
2. Als je ziet dat je een lineaire relatie hebt bepaal je hoe sterk de relatie is met de correlatiecoëfficiënt (r) of de verklaarde variantie (r2).
3. Als je dat gedaan hebt ga je aan de slag met de regressielijn. De regressielijn is de beste passende lijn door een puntenwolk. Je kan de vergelijking uitrekenen, als je deze hebt kun je de lijn tekenen. Er is maar één regressielijn voor elk scatterplot.

De best passende lijn door een scatterplot, is de lijn met zo min mogelijk error over de hele range van x waarden. Error is het verschil tussen wat voor waarde je hebt gevonden voor iemand op y en de voorspelling die je zou doen voor die persoon op basis van een regressielijn. Je maakt dus zo min mogelijk fouten. Die fouten noem je error of residu. Dus het verschil tussen wat je gevonden hebt en wat je voorspelt. De error noem je ook wel het residu. Hoe groot een error is bereken je door de geobserveerde waarde - de voorspelling. Error is positief als de geobserveerde waarde hoger is dan de voorspelling. Een error is negatief als de geobserveerde waarde lager is dan de voorspelling. In het boek wordt het vaak de least squares regression line genoemd, dus het is de lijn met de kleinste gekwadrateerde error. We nemen de gekwadrateerde error omdat anders de positieve uitkomsten de negatieve opheffen en andersom.

Regressievergelijking

We willen een manier om de lijn te beschrijven, die doen we met een regressievergelijking. Een rechte lijn valt altijd te beschrijven met een formule: y = a

+bx.

'a' betekend hier intercept. Dit is de waarde op de lijn van y op het moment dat x gelijk is aan 0. 'b' staat voor de helling en hoe schuin de lijn is. Als ik 1 opschuif op x, hoeveel ga ik dan omhoog of naar beneden p y

In het voorbeeld op dia 12 is a = 4 en b = 2. De formule is dus 4 + 2x.

Let op: 'a' valt niet altijd af te lezen. 0 staat namelijk niet altijd weergeven.

De vorm van een regressievergelijking is hetzelfde als die van een rechte lijn, de notatie is anders.

Formules van een regressievergelijking

- rechte lijn: y = a +bx
- voorspelde waarde voor y (regressielijn): $\hat{y} = b_0 + b_1 x$
- $\hat{y}$ = voorspelde waarde van y (dus niet de geobserveerde waarde)
- b_0 = intercept
- $b_1 x$ = regressiegewicht

In alle literatuur gebruiken we b'tjes voor de regressielijn

$$y_i = b_0 + b_1 x_i + e_i$$

- geobserveerde waarde:
- residu = error = $e_i = y_i - \hat{y}$
- regressiecoëfficiënt/ helling: $b_1 = r \frac{s_y}{s_x}$
- Correlatiecoëfficiënt x de standaarddeviatie van y / standaarddeviatie van x
- intercept: $b_0 = \bar{y} - b_1 \bar{x}$
- gemiddelde y – b1 x gemiddelde x

Regressievergelijking voorbeeld (dia 15)

Gegevens:

- r = 0.74 (sterk verband)
- $\bar{x}$ = 7.0 deeltoets a
- $\bar{y}$ = 6.0 deeltoets b
- s_x = 1.43
- s_y = 1.58

De predictor in dit voorbeeld is deeltoets a, deze gaat vooraf aan deeltoets b in de tijd.

Berekening:

$b_1 = 0.74 \frac{\times 1.58}{1.43} = 0.82$

$b_0 = 6.0 - 0.82 \times 7 = 0.26$

$\hat{y} = 0.26 + 0.74x$

Als het cijfer op A met 1 punt toeneemt, neemt het voorspelde cijfer op b met 0.82 toe.

Als een student op deeltoets a een 8 heeft gehaald, haalt deze student een (0.26 + 0.74 x 8) 6.18.

Kenmerken regressie

- De regressielijn loopt altijd door het punt ($\bar{x}, \bar{y}$
- Het intercept is niet altijd af te lezen uit het plot (de x-as loopt niet altijd tot 0).
- Teken r en b_1 (-/+) voor de richting van de relatie. Is de één positief is de ander het ook en andersom
- r^2 gebruiken voor de sterkte van de relatie
- b_1 geeft de steilheid van de lijn aan, niet de sterkte van de relatie. b_1 is afhankelijk van de sd van x en y. als deze ver uit elkaar liggen (y heel groot en x heel klein, wordt het een groot getal).
- Wanneer r 0 is, dan is b_1 ook 0. Als dit zo is, is er geen relatie.

De stijlheid van de regressielijn is afhankelijk van de schaal van x en y. Om de lijn te tekenen kun je gewoon voorspelde waarden voor twee x-waarden bereken. Het is handig om de gemiddelde x en y te nemen als eerste punt en x=0 (intercept) als tweede punt. De lijn kun je dan door die twee punten tekenen. De voorspelling is echter niet perfect. Op individueel niveau zit de (groeps)voorspelling er altijd naast.

Verklaarde variantie

Hoe meer fouten je maakt, hoe onnauwkeuriger je voorspeller. Dat kan je uitdrukken in *verklaarde variantie*. De verklaarde variantie (r²) is een maat voor succes voor de voorspelling en wordt afgeleid van het correlatiecoëfficiënt en wordt berekend door r te kwadrateren. Bij een perfect verband is r²=1=100%. Dit kun je interpreteren als percentage. Met de verklaarde variantie kijk je hoe goed je met de voorspelde x de voorspelde y kan weten. Je wil verschillen/variaties in scores verklaren.

De verklaarde variantie (r²) is de proportie variantie in y die verklaard kan worden door de voorspelling uit x. het gaat erom dat je kan verklaren waarom y niet voor

iedereen hetzelfde is. Hieruit kun je het succes van de voorspelling afleiden. De waarde ligt altijd tussen de 0 en de 1 of 0% en 100%. Hoe hoger het percentage, hoe perfecter het verband. Wanneer er veel spreiding is, ligt de waarde dicht bij de 0.

Wanneer er weinig spreiding is, is de verklaarde variantie dichtbij 1. Hoe meer de geobserveerde punten bij de voorspelde punten liggen, hoe hoger de verklaarde variantie. Dus hoe dichterbij de regressielijn, hoe nauwkeuriger de voorspelling, ook al is de lijn niet steil kan er toch een hoge correlatie zijn.

Perfecte voorspelling:

Bij een perfecte voorspelling is $r^2=1$.

Variantie voorspelde waarden = $s^2_{\hat{y}}$

Variantie geobserveerde waarden = s^2_y

De punten liggen exact op regressielijn dus $s^2_{\hat{y}} = s^2_y$

Verklaarde variantie = $\frac{s^2_{\hat{y}}}{s^2_y} = 1$

Niet perfecte voorspelling:

Bij een niet perfecte voorspelling liggen de punten verspreid rond de lijn, dus $s^2_{\hat{y}} < s^2_y$

Deze formules zijn voor begrip en niet voor berekening!

Haken en ogen

- Een regressie heeft dezelfde haken en ogen als een correlatie.
- We hebben het over lineaire verbanden, doormiddel van een residuenplot kan je kijken naar de lineairiteit. Een residuenplot geeft de error weer. Het is een plaatje van alle afwijkingen van de regressielijn. Het gemiddelde is 0. De x-as is de predictor en de y-as het residu.
- Uitbijters en invloedrijke observaties: kijk naar het scatterplot en kijk of er observaties buiten het patroon vallen. Sommige uitbijters zijn invloedrijk, andere bevestigen de regressielijn en zijn minder invloedrijk. Dia 34: A beïnvloedt de regressielijn het meest, het trekt de lijn naar beneden, terwijl b juist op de lijn ligt en de lop van de regressielijn niet verandert. Het verhoogt wel de verklaarde variantie.
- Extrapolatie: uitspraken over relaties buiten de range van geobserveerde gegevens
- Restricted range probleem: Het restricted range probleem is de onvolledige dekking van het domein. Dit leidt tot een lagere correlatie. Dit kan komen door bijvoorbeeld te weinig informatie of een niet representatieve steekproef. Hier kan je relaties door missen.

Praktijk van correlatie en regressie

Correlatie en regressie worden vaak samen gebruikt. De correlatie geeft de sterkte van het verband, regressie doet de voorspelling. We gebruiken alleen lineaire verbanden en numerieke variabelen.