Analyse af en lineær regression med lav R^2-værdi

Denne gennemgang omhandler figur 13 i 'Regn med biologi'. Man kan sagtens lave beregninger på egne data. Forsøgsmæssigt kræver det bare en tommestok tapet fast til væggen, en bog til at måle højden uden sko med og et spirometer. Sæt bogen vinkelret ind på tommestokken med bogryggen nedad. Før bogen ned langs tommestokken til den rammer forsøgspersonens hoved, og mål der højden.

For at se nærmere på data fra den lineære regression figur 13, 'Regn med biologi' kan det være en god idé at lave et residualplot af data. Det foregår på den måde at man bruger forskriften for den lineære regression til at udregne, hvilken værdi de enkelte datapunkter burde have, hvis de skulle følge den lineære model. Herefter udregnes forskellen mellem de observerede data (n_{obs}) og de beregnede data (n_{beregn}). For hvert enkelt datapunkt fremkommer en værdi der udregnes på følgende måde:

Residualværdi = $n_{obs} - n_{beregn}$

Dette indtastes i Excel – indtast som vist i figur 1, højden af forsøgspersonerne i A-kolonnen, vitalkapaciteten (VC) i B-kolonnen (det svarer til n_{obs} i ligningen ovenover), udregningen af n_{beregn} fra forskriften i figur 1.26 i C-kolonnen og residualværdien i D-kolonnen.

	A	B	C	D
1	Højde (cm)	VC (L)	n_{beregn}	Residual (L)
2	172,5	4,1	=A2*0,1022-13,691	=B2-C2

Figur 1. Indtastede formler i Excel. Tallene 0,1022 og -13,691 er hentet fra regressionsmodellen. Bruger man data fra eget forsøg, skal disse tal ændres i forhold til henholdsvis a og b i den lineære regression $y = ax + b$.

Det vil sige at man har en lang liste af sammenhørende værdier af højde og vitalkapacitet i henholdsvis kolonne A og kolonne B. Værdierne i C2 og D2 markeres, og man flytter musen ned i højre hjørne af D2. Der fremkommer et kryds i nederste hjørne, og man trækker ned i kolonnerne, så længe der er data i kolonne A og B. Opstil nu et tomt diagram ved at klikke i en tom celle fx E1 og vælg 'Indsæt' og i 'Diagram' peg på et punktdiagram.

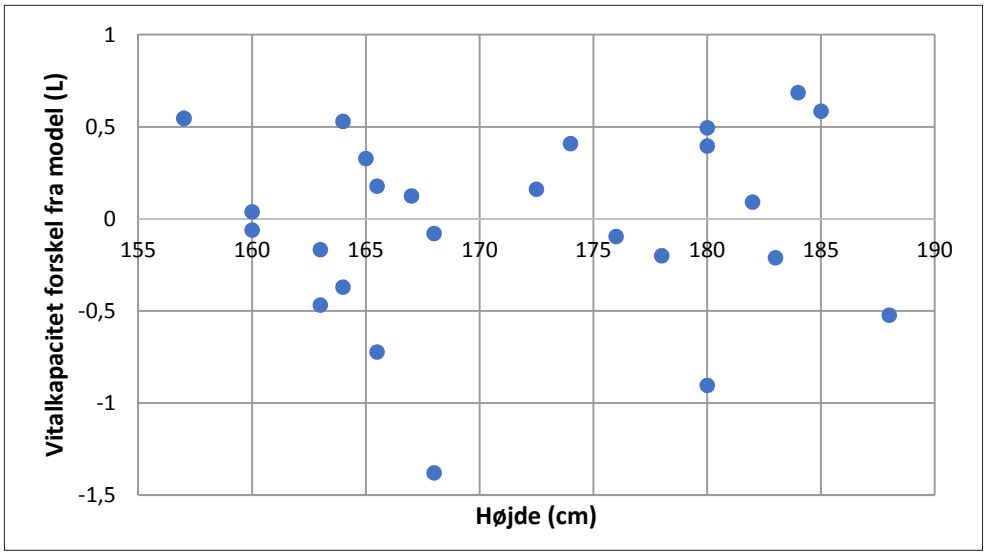
PC: Højreklik på det tomme punktdiagram og klik på 'Vælg data' og derefter 'Tilføj'. Klik herefter på X-serieværdier og markér tallene i kolonne A (ikke overskriften i A1). Derefter klikkes på Y-serieværdier, og tallene i kolonne D markeres. Klik 'OK' og 'OK'.

Mac: Højreklik på det tomme punktdiagram og klik på 'Marker data' og derefter '+' under det hvide felt til venstre i dialogboksen (Excel skriver nu 'Serie1'). Klik med musen i feltet X-serieværdier og markér tallene i kolonne A (ikke overskriften i A1). Derefter klikkes på Y-serieværdier, og tallene i kolonne D markeres. Klik 'OK' og 'OK'.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Begge: I fanebladet 'Design' vælges 'Tilføj dataelement' → 'Aksetitler' → 'Primær vandret' og 'Primær lodret'. På x-aksen rettes 'Aksetitel' til '**Højde (cm)**' og på y-aksen rettes 'Aksetitel' til '**Vitalkapacitet' forskel fra model (L)**'. Derved fremkommer et residualplot som det ses i figur 2.

I figur 2 svarer x-aksen til tendenslinjen i figur 1.26 i 'Regn med biologi'. Man kan nu direkte aflæse om værdierne ligger over x-aksen og dermed tendenslinjen, eller om de ligger under. Det ses at 14 datapunkter ligger over tendenslinjen, og 12 datapunkter ligger under tendenslinjen. Summen af datapunkterne under x-aksen og over x-aksen er tilnærmelsesmæssigt lig med hinanden for sådan laves tendenslinjer, men der er forskel på hvordan datapunkterne er fordelt. Over x-aksen er der mange punkter mellem y = 0 og y = 0,5, mens kun 4 punkter ligger lige over 0,5 der har den højeste værdi for datapunktet (184;0,6862). Der er kun ét datapunkt over 0,6, mens der er 3 punkter under -0,6 og den laveste værdi findes ved (168;1,3768). De enkelte negative værdier ligger derfor længere fra x-aksen end de enkelte positive værdier eller sagt med andre ord: Punkterne over tendenslinjen i figur 1.26 i 'Regn med biologi' ligger tættere på tendenslinjen end punkterne under tendenslinjen, og derved udviser de mere spredning. Det kan til dels forklares med de problemfelter, der blev set bort fra ved opstillingen af den lineære model, nemlig at kropsstørrelse, træning og smertetærskel også betyder noget.



Figur 2. Residualplot på baggrund af figur 1.26 i 'Regn med biologi'.

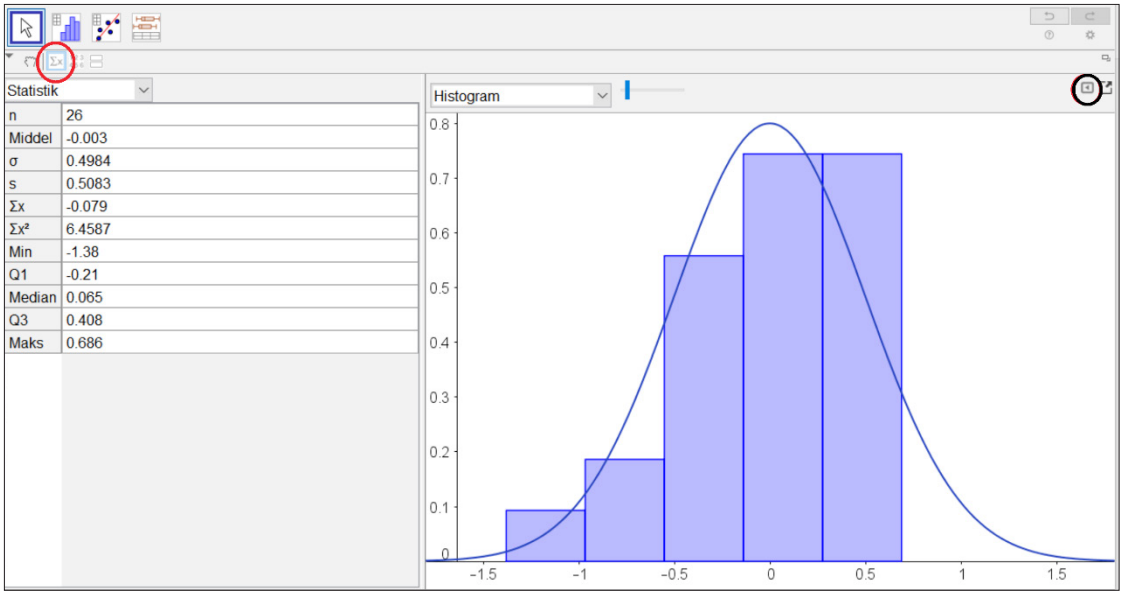
Yderligere analyser

Man kan analysere data yderligere i GeoGebra eller andet CAS-værktøj. Men skal bare huske på at data i GeoGebra bruger punktum i stedet for komma til at adskille decimaler fra heltal. Det rettes på følgende måde: Marker cellerne i D-kolonnen inkl. overskriften og kopiér dem. Åben Word og sæt cellerne ind der. Lav derefter 'Søg og erstat' (pc: tryk Ctrl + h og Mac: tryk CMD + ?) og søg efter ',' (komma) og erstat med '.' (punktum). Tryk på 'Erstat alle'. Kopier de nye celler fra Word og indsæt dem i A-kolonnen i regnearket i GeoGebra. I princippet kan Excel også lave søg og erstat, men det virker dårligt og kan ikke anbefales (cellerne formateres nogle gange til noget som ikke giver mening).

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

I GeoGebra: Klik på 'Enkeltvariabelanalyse' og på tandhjulet øverst til højre og vælg 'Brug sidehoved som titel'. Derved får kolonnen samme navn som overskriften i Excel (her 'Residual (L)'). Klik på 'Analyser'.

Der fremkommer nu et histogram som repræsenterer residualværdierne. Tryk på ikonet i den sorte ring i figur 3 og i fanebladet 'Histogram', vælg 'Normaliseret' og sæt mærke i 'Gausskurve'. Hvis data var fordelt ligeligt omkring middelværdien (eller i vores tilfælde: Ligeligt omkring tendenslinjen), ville histogrammet følge Gausskurven eller normalfordelingskurven som den også kaldes. Man kan justere minimum- og maksimumværdierne i fanebladet 'Graf' – det kan være en fordel at normalfordelingskurven i analyse af residualplot er symmetrisk omkring 0. Normalfordelingskurven i figur 3 er sat til at have en minimumsværdi på -1,8 og en maksimumværdi på 1,8. Klik herefter på statistiktegnet 'Σx' i den røde ring, og der fremkommer en mængde oplysninger om data. Hvordan de skal tolkes, kan man læse om i 'Regn med biologi', men man kan se at der indgår 26 datapunkter (n = 26), og at middelværdien er -0,003 (Middel = 0,003). Egentlig forventes en middelværdi på 0, men idet regressionen der ligger til grund for histogrammet ikke har en høj R^2-værdi, kan man ikke forvente præcist at ramme 0.



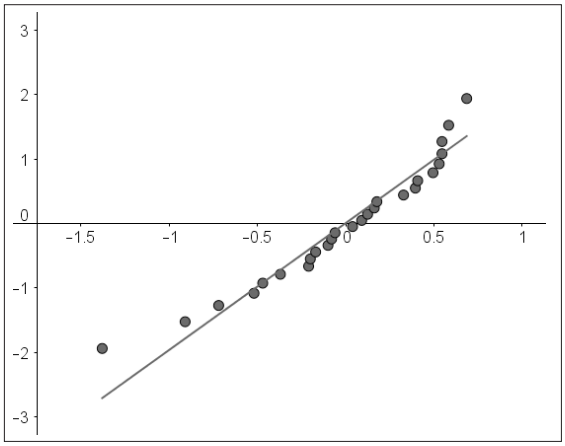
Figur 3. Statistisk behandling i GeoGebra.

Bredden på de enkelte søjler i histogrammet bevirker at det kan være svært at analysere data. Sammenligner man histogrammet med normalfordelingskurven, er det tydeligt at data har en 'hale' til venstre. Man kalder det at kurven har en negativ skævhed, fordi skævheden er til venstre for middelværdien. Modsat er en positiv skævhed at halen er til højre for middelværdien. Forhold ved forsøget, som at det kræver mange kræfter at tømme lungerne, gør at små forsøgspersoner skal arbejde meget for at tømme deres lunger, og derfor stopper de ofte med at puste i spirometeret, før de når deres vitalkapacitet. Det betyder at små personer 'underpræsterer', og det ses som en negativ skævhed i diagrammet.

Ved analyse af andre data eller hvis man havde haft flere målepunkter i denne analyse, kunne man have fået en mere normalfordelt kurve, men hvis spredningen er stor, vil man stadig opleve

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

kurver med lav R^2-værdi (under 0,95), men det betyder bare at variation i biologiske systemer er stor – ikke nødvendigvis at modellen er dårlig.



Figur 4. Kvartilplot i GeoGebra. Skift på diagramvælgeren fra 'Histogram' til 'Kvartilplot'.

Til slut kan man lave et kvartilplot i GeoGebra for at se om data er normalfordelte. Det ses af figur 4 at lave og høje datapunkter ligger over kurven, og helt systematisk ligger datapunkter under kurven inde ved middelværdien. Det bevirker at data ikke kan betegnes som normalfordelte, men netop udviser negativ skævhed.

Til slut kan man lave et kvartilplot i GeoGebra for at se om data er normalfordelte. Fordelen ved dette plot er at selvom det er vanskeligt at forklare, så viser det meget bedre end at sammenligne en klokkeformet kurve med et stolpediagram/histogram om data er normalfordelte.

Hvad vises i et kvartilplot?

I matematikprogrammer er det talanalysen der er fokus, og derfor er der ingen akseangivelser på grafen vist i figur 4. Grafen kaldes et kvartilplot eller et normalfordelingsplot. X-aksen repræsenterer normalfordelingskurven, og y-aksen repræsenterer datasættet. Begge akser er normaliserede, hvilket vil sige, at 0 sættes til middelværdien, værdien 1 betyder middelværdien +1 gange spredningen, og -1 betyder middelværdien -1 gange spredningen. I et datasæt med en middelværdi på 2 og en spredning på 3 betyder 0 på y-aksen derfor 2, 1 betyder 2+3 = 5 og -1 betyder 2-3 = -1.

X-aksen repræsenterer normalfordelingen. Har man som i dette tilfælde, 26 data i et datasæt, inddeles normalfordelingskurven i 26 lige store sandsynligheder. Da sandsynligheden repræsenteres af 'arealet under kurven', vil der være kortere afstand mellem datapunkterne omkring middelværdien hvor den klokkeformede kurve er høj, og langt mellem datapunkterne hvor kurven nærmer sig 0 langt fra middelværdien. Herefter plottes det mindste datapunkt som funktion af det mindste normalfordelingspunkt, det næstmindste mod det næstmindste osv. Til sidst tegnes den bedste rette linje gennem datapunkterne.

Datapunkterne behøver ikke at ligge på linjen for at det er en normalfordeling – de kan sagtens være spredt lidt ud på begge sider af linjen. Det skal dog bære præg af at fordelingen er tilfældig, og ikke at datapunkterne udviser en anden form end en lineær.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Analyse af figur 4

For at data skal være normalfordelte skal de følge en ret linje i kvartilplottet. Det er tydeligt at data har en opadkrumning, idet data i enderne af den rette linje systematisk ligger over linjen og på midten af den rette linje systematisk ligger under linjen. Derfor er data ikke normalfordelte. Man kan se at data udviser en negativ skævhed ved at betragte x-aksen i grafen. Den laveste værdi (-1,4) er længere væk fra middelværdien (0) end den højeste værdi (0,7), og derfor er der tale om en negativ skævhed.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7