

Regression

Datasæt, der består af talpar, repræsenterer typisk en sammenhæng eller i hvert fald noget man vil undersøge om der er en sammenhæng. Det kunne være, at man ville undersøge om der var en sammenhæng mellem højden og vægten af et antal personer og hvordan denne sammenhæng kan beskrives matematisk.

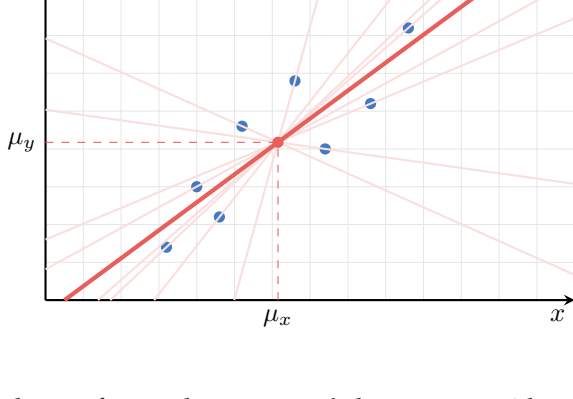
For et datasæt med talpar af n punkter $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ kan der beregnes en regressionslinje på formen:

$$f(x) = a \cdot x + b$$

Dette gøres ved først at bestemme hældningskoefficienten a for regressionslinjen som bestemmes ud fra midtpunktet af datapunkterne (μ_x, μ_y) der beregnes ved:

$$\mu_x = \frac{x_1 + x_2 + \dots + x_n}{n} \quad \text{og} \quad \mu_y = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Hældningskoefficient for regressionslinjen beregnes ved for hvert datapunkt at betragte en linje der går gennem datapunktet og datasættets midtpunkt. Ud fra alle disse linjer beregnes en gennemsnitshældning.



Det bemærkes dog, at for punkterne tæt på datasættets midtpunkt, vil en lille afvigelse fra regressionslinjen have en stor indflydelse på hældningskoefficienten hvilket vil sige, at der er en stor usikkerhed forbundet med hældningskoefficienterne beregnet ved brug af disse punkter. Tilsvarende vil der være en tilsvarende lille usikkerhed forbundet med punkter, der ligger langt væk fra datasættets midtpunkt.

For at håndtere denne varierende usikkerhed vægtes gennemsnittet af de udregnede hældningskoefficienter med den kvadrerede afstand på x -aksen. Den resulterende formel for beregning af regressionslinjens hældningskoefficient er:

$$a = \frac{(x_1 - \mu_x) \cdot (y_1 - \mu_y) + \dots + (x_n - \mu_x) \cdot (y_n - \mu_y)}{(x_1 - \mu_x)^2 + \dots + (x_n - \mu_x)^2}$$

Når hældningskoefficienten a for regressionslinjen er bestemt kan skæringskoefficienten b også bestemmes ved, at regressionslinjen skal gå igennem midtpunktet for datapunkterne:

$$b = \mu_y - a \cdot \mu_x$$

Som det kan ses på udledningen for lineær regression bygger denne på to-punktsformlen for en ret linje. Ud fra to-punktsformlerne for hhv. potensfunktioner, eksponentielle funktioner og logaritmiske funktioner kan også udlede tilsvarende regressionsformler. Disse regressionsmetoder kan således anvendes til at teste og validere matematiske modeller på faktiske datasæt og evt. afklare hvilken type vækst der er tendensen i datasættet.

For at kunne validere en funktion, der er beregnet ud fra regression, og derved bestemme hvor godt den beskriver datasættet, kan man beregne forklaringsgraden R^2 ved:

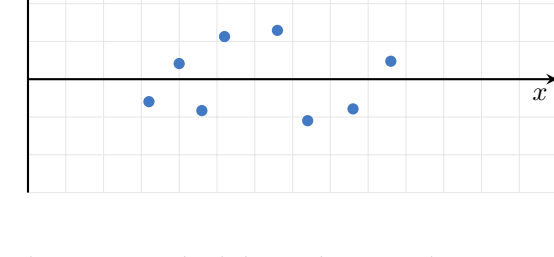
$$R^2 = \frac{(f(x_1) - \mu_y)^2 + \dots + (f(x_n) - \mu_y)^2}{(y_1 - \mu_y)^2 + \dots + (y_n - \mu_y)^2}$$

Forklaringsgraden kan tolkes som hvor stor en andel af datasættet der er forklaret af regressionsfunktionen og er derfor en talværdi i intervallet $0 < R^2 \leq 1$. Jo tættere forklaringsgraden R^2 er på 1 jo bedre passer regressionsfunktionen til datasættet og såfremt $R^2 = 1$ ligger alle punkterne på kurven for regressionsfunktionen.

Ud over at beregne forklaringsgraden er det også vigtigt at undersøge hvilken type afvigelse der er tale om mellem datasættet og regressionsfunktionen. For at undersøge dette beregnes residualerne dvs. afvigelsen for hvert punkt ved:

$$r_1 = y_1 - f(x_1) \quad r_2 = y_2 - f(x_2) \quad \dots \quad r_n = y_n - f(x_n)$$

Residualerne kan plottes ift. de tilsvarende x -værdier i et residualplot som vist nedenfor.

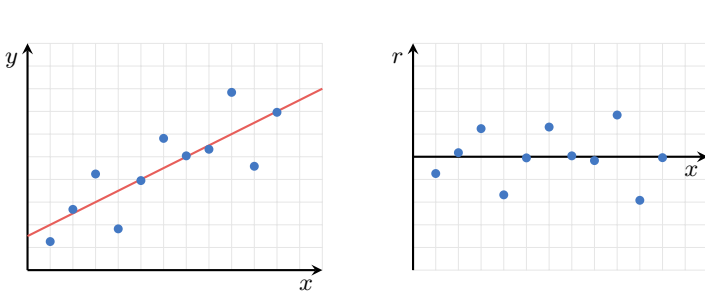


Når man skal analysere et residualplot er der specielt tre ting som man undersøger:

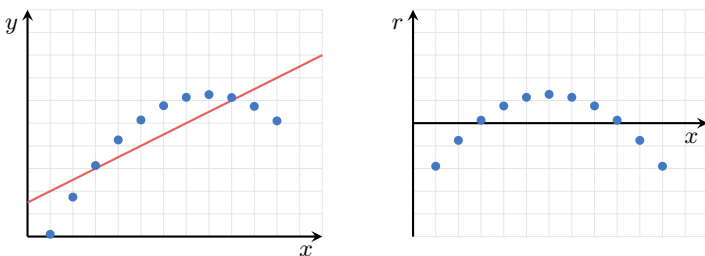
1. Om punkterne for residualerne er tilfældig spredt omkring førsteaksen eller om der forefindes en systematisk afvigelse.
2. Om der forekommer "udstikkere" dvs. datapunkter, der har en residualværdi der er betydeligt større eller mindre end de andre.
3. Om målingerne er jævnt fordelt ud over det interval man undersøger eller om der er en stor andel af punkterne der ligger i et lille område.

For at illustrere vigtigheden af en residualanalyse fremsatte matematikeren Francis Anscombe fire datasæt, der alle frembringer regressionslinjen $f(x) = 0.5 \cdot x + 1.5$ og har forklaringsgraden $R^2 \approx 0.67$, men som har forskellige tydelige fejl i det tilsvarende residualplot.

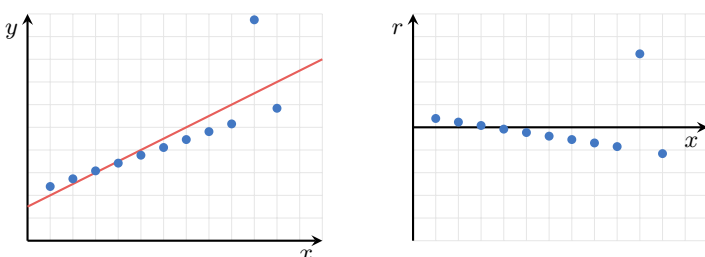
Det første datasæt viser et eksempel på en god overensstemmelse mellem datasættet og regressionsfunktionen.



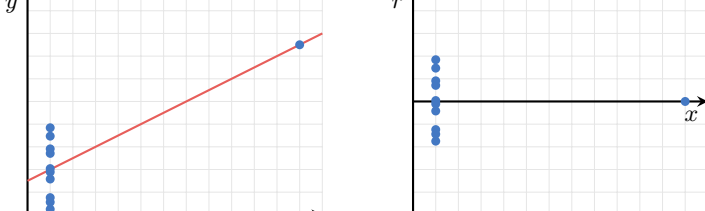
Det andet datasæt viser en tydelig tendens i residualplottet og der er her tale om en systematisk fejl, der indikerer at datasættet ikke har en lineær tendens. Man bør i dette tilfælde undersøge om evt. en potensfunktion eller en eksponentiel funktion vil give en bedre overensstemmelse mellem data og model.



I det tredje datasæt kan det klart ses at der er tale om en udstikker, altså et punkt som har en residual der er klart større end resten af punkterne. Her bør man undersøge om der er tale om en fejlmåling.



Det sidste datasæt viser et eksempel hvor datapunkterne er ujævnt fordelt i det interval der undersøges. I dette tilfælde bør man indsamle flere datapunkter i andre områder af intervallet for at skabe en jævn fordeling.



Geogebra: Regressionsanalyse

Først indtastes data i Geogebra's regneark (spreadsheet).

Hvis man har selve datasættets 1.koordinater indtastet i felterne A1 til A100 og de tilsvarende 2.koordinater i B1 til B100, kan man plotte datasættet som punkter ved:

$$P = (A1 : A100, B1 : B100)$$

Ud fra listen af punkter P kan man

- For regression med en lineærfunktion:

$$f : \text{FitLine}(P)$$

- For regression med en potensfunktion:

$$f : \text{FitPow}(P)$$

- For regression med en eksponentiel funktion:

$$f : \text{FitExp}(P)$$

- For regression med en logaritmisk funktion:

$$f : \text{FitLog}(P)$$

Forklaringsgraden kan beregnes ud fra P og f ved:

$$\text{RSquare}(P, f)$$

Residualplottet kan frembringes ved:

$$R = \text{ResidualPlot}(P, f)$$

Residualspredningen kan herefter beregnes ud fra R ved:

$$\sqrt{\text{Variance}(y(R))}$$

Den største residualværdi beregnes ud fra R ved:

$$\text{Max}(|y(R)|)$$

Tjek om residualerne er normalfordelte ved QQ-plot:

$$\text{NormalQuantilePlot}(y(R))$$