SAMBANDSANALYS

REGRESSION OCH KORRELATION

VT 2014

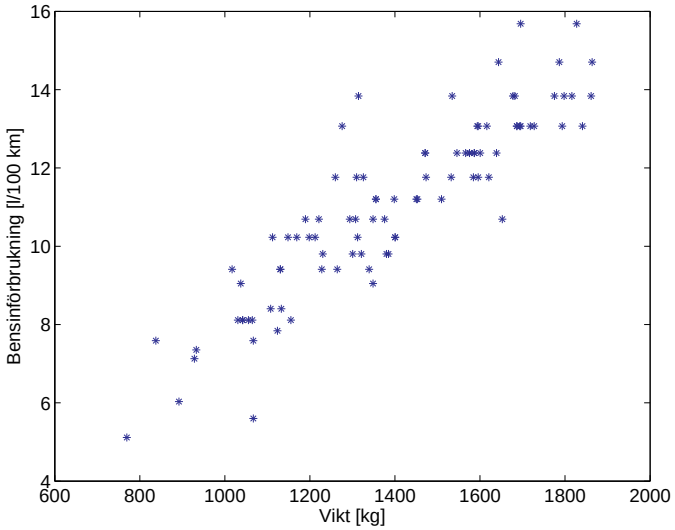
Matematikcentrum
Matematisk statistik

Innehåll

1 Samband mellan två eller flera variabler

Det är ganska vanligt att man gör mätningar på två eller flera variabler och vill undersöka om det finns något samband mellan dem. Vi presenterar två exempel:

> **Exempel 1.1.** För ett slumpmässigt urval av bilar noterar man y-bensinförbrukning i stadskörning (l/100 km) och x-vikt (kg). Data beskrivs i figur 1 där y plottats mot x.



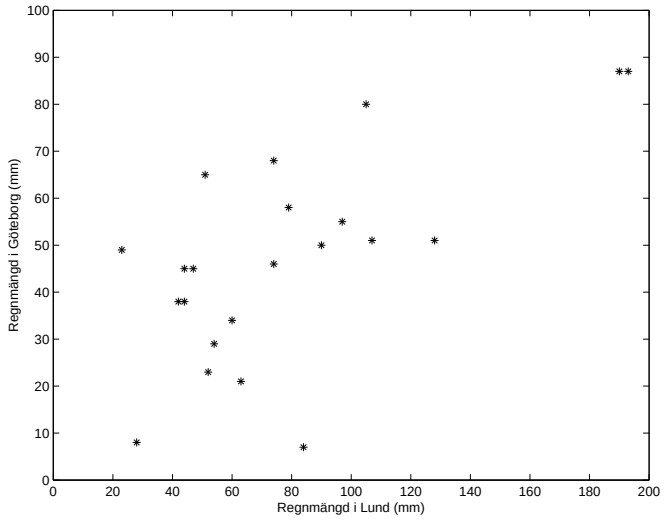
Figur 1: Ett slumpmässigt urval av bilar där y = ”bensinförbrukning i stadskörning” är plottad mot x = ”vikt”.

I detta exempel är det rimligt att tänka sig att y-bensinförbrukning påverkas av x-vikt (och inte tvärt om!). Vi kan alltså försöka beskriva y som en funktion av x, analysen måste naturligtvis också ta hänsyn till att mätningarna påverkas av en slumpmässig störning. Vi gör en *regressionsanalys* där y är responsvariabeln medan x är den förklarande variabeln. Ibland kallas även y för den beroende variabeln medan x är den oberoende variabeln:

$$\underbrace{y}_{\textit{responsvariabel}} = \underbrace{f(x)}_{\textit{regressionsfunktion med förklarande variabel x}} + \underbrace{''slump''}_{\textit{s.v. med fördelning}}$$

När regressionsfunktionen $f(x)$ är linjär med avseende på sina parametrar har vi linjär regression. Från figuren verkar det rimligt att tänka sig ett linjärt samband mellan x och y som beskriver hur stor bensinförbrukning en ”medelbil” av en viss vikt har. Om man, som i vårt exempel, har enbart en förklarande variabel, x, talar man om *enkel linjär regression*. Hela nästa avsnitt kommer att behandla denna viktiga situation.

> **Exempel 1.2.** Månadsnederbörden, d.v.s. den totala mängden nederbörd (mm) under en månad, noterades i Göteborg och Lund under åren 2005 och 2006. I figur 2 markerar varje punkt en månad där Göteborgs nederbörd avläses på y-axeln och Lunds på x-axeln.



Figur 2: Månadsvisa mätningar av nederbörden (mm) där $y =$ ”nederbörd i Göteborg” är plottad mot $x =$ ”nederbörd i Lund”.

Här är det inte självklart att någon av de två uppmätta varaiablerna kan beskrivas som en funktion av den andra. Variablerna är ”likvärdiga” eftersom vi lika gärna skulle kunna byta variabel på axlarna och placera Lundamätningarna på y-axeln och Göteborgsmätningarna på x-axeln. I denna situation är det olämpligt att använda regression, man får nöja sig med att beskriva graden av samband i en *korrelationsanalys*. Vi kommer att studera detta närmare i avsnitt 4.

2 Enkel linjär regression

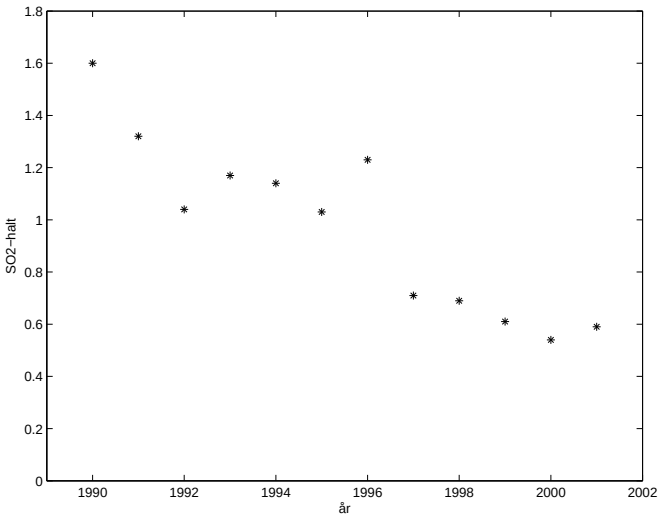
I enkel linjär regression studerar vi en variabel y som beror linjärt av en variabel x men samtidigt har en slumpmässig störning eller avvikelse:

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

där ε_i är den slumpmässiga avvikelsen från linjen.

I detta avsnitt kommer vi illustrera teorin med hjälp av två dataset: mätningarna från exempel 1.1 om bensinförbrukning hos bilar samt mätningar av SO_2-halt i luft.

Exempel 2.1. Inom miljöövervakningsprogrammet EMEP har man under en lång period mätt årsmedelhalter av SO_2 (μg/m^3) i Hoburgen på Gotland. I figur 3 visas halterna under åren 1990-2001 (källa: IVL Svenska Miljöinstitutet AB, www.ivl.se).



Figur 3: Mätningar vid Hoburgen på Gotland y = "SO_2-halt" (μg/m^3) är plottad mot x = "år".

2.1 Intressanta frågeställningar

Det finns en mängd frågeställningar kring den beskrivna situationen som är intressanta:

- Hur ska vi skatta α och β i regressionslinjen $y = \alpha + \beta x$? Lutningen β beskriver hur mycket y ändras då x ökar med en enhet: hur mycket ökar bensinförbrukningen då vikten hos en bil ökar med ett kg? Speciellt intressant är det att undersöka om $\beta = 0$ eftersom det innebär att regressionssambandet då kan reduceras till $y = \alpha$, d.v.s. att y inte beror av x. I data från Hoburgen innebär ett $\beta \neq 0$ att det finns en trend i SO_2-halt.

- Hur stor är variationen kring linjen? Eftersom ε_i beskriver den slumpmässiga avvikelsen från linjen motsvarar det att undersöka hur stor denna avvikelse tenderar att vara - ett mått på detta är $D(\varepsilon_i)$ som vi betecknar σ.

- Givet ett x_0, vad är det **förväntade** värdet på Y? Vi söker alltså $\mu_0 = \alpha + \beta \cdot x_0$, linjens läge i punkten x_0. I bilexemplet kan vi t.ex. vara intresserade av hur stor bensinförbruktingen är i genomsnitt hos bilar som väger 1200 kg. I Hoburgsdata vad förväntad SO_2-halt var 1994.

- Skilj den föregående frågeställningen från följande: Givet ett x_0, vad är en **enstaka** observation av Y, Y_0? Vi vill göra en pediktion av Y-värdet. Det kan t.ex. gälla en prognos av Y för något framtida värde på x. Om vi har en bil som väger 1200 kg, är vi nu intresserade av hur stor bensinförbrukningen är för detta exemplar. I SO_2-exemplet kan vi vilja prediktera halten för år 2002 - inom vilket intervall är det troligt att kommer den att hamna?

- Hur bra passar modellen till data? Är det lämpligt att beskriva sambandet med en linjär funktion eller borde vi ansätta något annat? Denna frågeställning bör man studera först - det är naturligtvis viktigt att den antagna modellen stämmer någorlunda till data innan man detaljstuderar den.

- Hur mycket av den totala variationen i y-led har vi förklarat med modellen? Man kan inte räkna med att modellen ska förklara all variation som finns i mätningarna. Bensinförbrukningen hos en bil beror inte enbart på bilens vikt utan påverkas - förutom av slumpmässig variation - av en mängd andra variabler. Hur stor andel av variation i bensinförbrukning kan beskrivas med hjälp av bilars vikt och hur stor andel av variationen återstår att beskriva? Den återstående variationen kanske delvis kan förklaras med hjälp av andra variabler?

För att kunna hantera dessa frågor gör vi vissa antaganden om den linjära modellen och om våra mätningar $(x_1, y_1), \ldots, (x_n, y_n)$.

2.2 Modellantaganden

Vi använder följande modell där y_i är n st oberoende observationer av

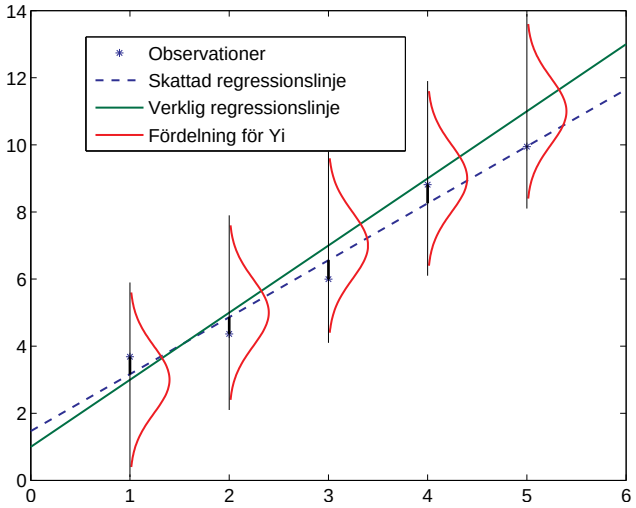
$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad \text{där } \varepsilon_i \in N(0, \sigma), \text{ oberoende av varandra}$$

så observationerna är $Y_i \in N(\alpha + \beta x_i, \sigma) = N(\mu_i, \sigma)$, dvs de är normalfördelade med väntevärde på den okända regressionslinjen $\mu(x) = \alpha + \beta x$ och med samma standardavvikelse σ som avvikelserna ε_i kring linjen har; se figur 4.

Modellen ovan är beskriven i ”kortform”, några förklaringar och kommentarer till den:

- Vi tänker oss att x-värdena är fixa eller uppmätta med ett försumbart mätfel - ofta kan vi själva välja vilka x-värden vi vill studera. Den slumpmässiga variation vi vill modellera finns enbart i y-led. I bilexemplet anses vikten hos en bil inte ha någon större variation; likaså är det uppenbart att x-variablen i Hoburgexemplet - årtalen - är fixa.

- Tidigare har vi haft modeller där mätningarna är observationer av stokastiska variabler ξ_i, vilka hade samma väntevärde μ, men nu är observationernas väntevärde en linjär funktion av x. Beteckningen Y_i är också en naturligare beteckning för den stokastiska variabeln.



Figur 4: Sann regressionslinje, observationer och skattad regressionslinje. Residualerna är markerade som de lodräta avstånden mellan observationerna och den skattade regressionslinjen.

- Att de slumpmässiga avvikelserna från linjen, $\varepsilon_1, \ldots, \varepsilon_n$ är oberoende innebär t.ex. att om en avvikelse råkar bli stor (liten) vid ett visst x-värde ska det inte påverka hur avvikelsen blir vid något annat x-värde. Om SO_2-halten år 1991 är lägre än vad som förväntades enligt linjens läge vid denna tidpunkt ska detta alltså inte påverka hur halten avviker från linjens läge vid t.ex. år 1992.

- För ett fixt x-värde kommer motsvarande y-mätningar att vara normalfördelade kring linjen och standardavvikelsen i den fördelningen är σ; se figur 4. Om vi t.ex. slumpmässigt väljer ut ett antal bilar som alla har vikt 1400 kg och mäter deras bensinförbrukning kommer förbrukningen att fördela sig enligt en normalfördelning med väntevärde $\alpha + \beta \cdot 1400$ och standardavvikelse σ.

- Observera att vi tänker oss att spridningen i normalfördelningarna är den samma oavsett värde på x, d.v.s. σ är konstant. Det innebär t.ex. att modellen inte tillåter att spridningen kring linjen ändrar sig då x-värdet ändras. Det är inte ovanligt i många sammanhang att y-mätningarna uppvisar en större spridning med ökande värde på x; för denna situation kan vi alltså inte direkt använda oss av ovanstående modell.

2.3 Skattningar av parametrarna α, β och σ

För att skatta parametrarna α och β används minsta kvadrat-metoden (MK-metoden). Skattningarna och deras fördelning härleds i appendix i avsnitt 5, här presenteras enbart resultaten.

MK-skattningarna av regressionslinjens lutning, β, och intercept, α, ges av

$$\beta^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \alpha^* = \bar{y} - \beta^* \bar{x}.$$

Eftersom β^* är en linjär funktion av observationerna Y_i ($\beta^* = \sum c_i Y_i$ där $c_i = (x_i - \bar{x})/S_{xx}$), och även α^* en linjär funktion av β^* och observationerna, är dessa skattningar

normalfördelade med väntevärde och standardavvikelse enligt

$$\beta^* \in N(\beta, \frac{\sigma}{\sqrt{S_{xx}}}), \qquad \alpha^* \in N(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}).$$

De två skattningarna är dock *inte* oberoende av varandra. Man kan däremot visa att β^* och $\bar{Y}$ är oberoende[1] av varandra.

Då man ska skatta variansen σ^2 visar det sig lämpligt att studera modellens s.k. *residualer*, $r_1, \ldots, r_n$ där

$$r_i = y_i - (\alpha^* + \beta^* x_i), \quad i = 1, \ldots, n,$$

är residualen för x_i och motsvarar den lodräta avvikelsen mellan det observerade värdet y_i och den skattade linjen, se figur 4. Residualen r_i är ett närmevärde till den slumpmässiga avvikelsen ε_i och eftersom σ^2 är ett mått på spridningen hos ε_i är det rimligt att residualerna kan användas när vi vill skatta variansen.

En väntevärdesriktig skattning av variansen ges av

$$(\sigma^2)^* = s^2 = \frac{Q_0}{n-2}$$

där Q_0 är residualkvadratsumman

$$Q_0 = \sum_{i=1}^{n}(y_i - \alpha^* - \beta^* x_i)^2 = \sum_{i=1}^{n} r_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

För att räkna ut kvadratsummorna S_{xx}, S_{yy} och S_{xy} ”för hand” kan man ha användning av sambanden

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big).$$

Naturligtvis har vi även t.ex. om s_x^2 är stickprovsvariansen för x-dataserien $S_{xx} = (n-1)s_x^2$.

2.4 Konfidensintervall för α och β

Eftersom skattningarna av α och β är normalfördelade får vi direkt konfidensintervall med konfidensgraden $1 - a$ (α är upptagen) precis som tidigare enligt

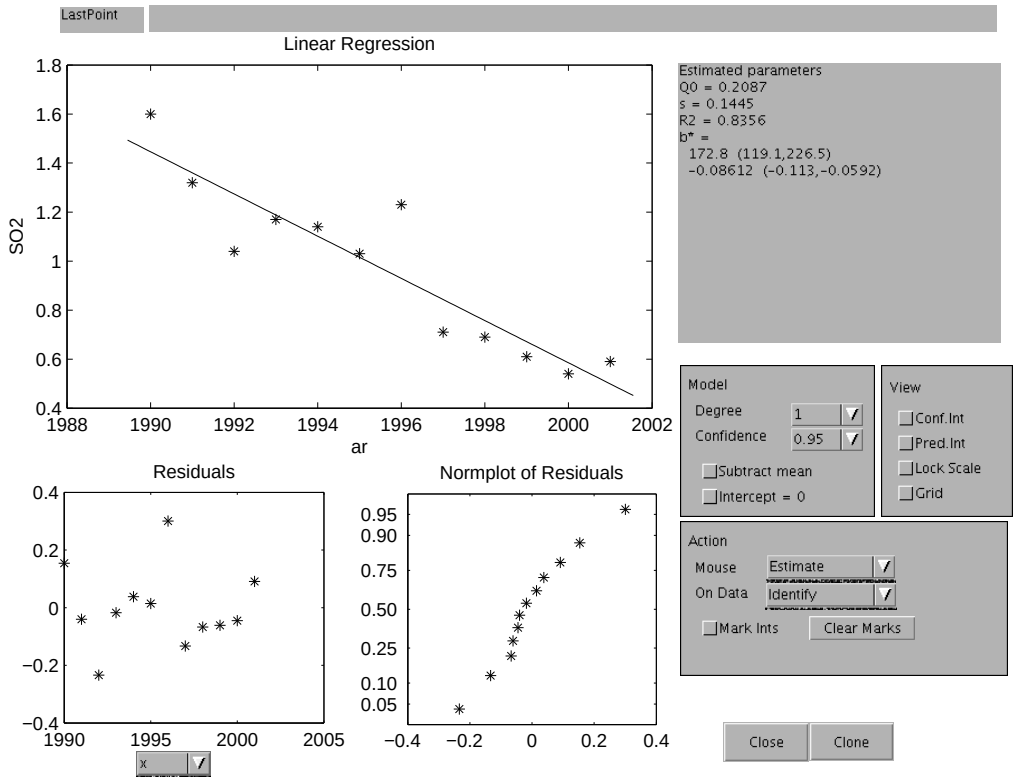
$$I_\beta = \beta^* \pm t_{a/2}(f)d(\beta^*) = \beta^* \pm t_{a/2}(n-2) \cdot \frac{s}{\sqrt{S_{xx}}}$$

$$I_\alpha = \alpha^* \pm t_{a/2}(f)d(\alpha^*) = \alpha^* \pm t_{a/2}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

Om σ skulle råka vara känd används naturligtvis den i stället för s och då även λ- i stället för t-kvantiler.

[1] Vi visar inte här att β^* och $\bar{Y}$ är oberoende av varandra, men det faktum att regressionslinjen alltid går genom punkten $(\bar{x}, \bar{y})$ gör det kanske troligt; om β över- eller underskattas påverkas inte $\bar{Y}$ av detta.

Exempel 2.2. Hoburgsdata i exempel 2 analyserades, med hjälp av rutinen `reggui` i Matlab, och vi fick följande utskrift och figurer.
Rutinerna till reggui kan laddas ner från kursens hemsida.



Figur 5: Regressionsanalys på materialet från Hoburgen; $y =$ "SO_2-halt" är plottad mot $x =$ "år".

Överst till höger i utskriften ges en mängd information, bl.a. skattningar och konfidensintervall för modellens tre parametrar. För att göra det mer åskådligt sammanställer vi resultaten i en tabell:

parameter	skattning	95% konfidensintervall
α	172.8	$(119.1, 226.5)$
β	-0.08612	$(-0.113, -0.0592)$
σ	0.1445	

Vi ser att α skattas till 172.8 μg/m^3 och motsvarande intervall är $I_\alpha = (119.1, 226.5)$. Eftersom α är interceptet med y-axeln motsvaras α i detta exempel av SO_2-halten vid år 0! Det går naturligtvis ej att anta att det linjära sambandet sträcker sig så långt bak, skattningen av α ger oss alltså inte omedelbart någon användbar information. Desto intressantare är lutningen β eftersom den talar om för oss hur mycket SO_2-halten ändras under ett år. Från utskriften ser vi att denna förändring skattas till -0.08612 μg/m^3 per år. Intervallet $I_\beta = (-0.113, -0.0592)$ kan användas för att testa hypotesen $H_0 : \beta = 0$, vilket skulle innebära att SO_2-halt inte påverkas av årtalet (d.v.s. ingen trend i data). Eftersom detta intervall inte täcker över 0 kan vi förkasta hypotesen $H_0 : \beta = 0$ och vi har påvisat (95% säkerhet) en nedåtgående *trend* i SO_2-halt vid Hoburgen.

Vi ser också att σ skattas till 0.1445 (något konfidensintervall för denna storhet ges ej i utskriften). Residualkvadratsumman Q_0 är 0.2087 och det gäller som tidigare att $(\sigma^2)^* = s^2 = \frac{Q_0}{n-2}$ där n är antalet observerade talpar, d.v.s. n =12. Storheten R2 i utskriften kommenteras nedan i avsnittet om förklaringsgraden.

2.5 Skattning av punkt på linjen

För ett givet värde x_0 är Y's väntevärde $E(Y(x_0)) = \alpha + \beta x_0 = \mu_0$, dvs en punkt på den teoretiska regressionslinjen. μ_0 skattas med motsvarande punkt på den skattade regressionslinjen som $\mu_0^* = \alpha^* + \beta^* x_0$. Vi ser direkt att skattningen är väntevärdesriktig samt att den måste vara normalfördelad (linjär funktion av två normalfördelade skattningar). Ett enkelt sätt att bestämma skattningens varians får vi om vi återigen utnyttjar att β^* och $\bar{Y}$ är oberoende av varandra (men inte av α^*)

$$V(\mu_0^*) = V(\alpha^* + \beta^* x_0) = [\alpha^* = \bar{Y} - \beta^* \bar{x}] = V(\bar{Y} + \beta^*(x_0 - \bar{x})) = [\text{ober}] =$$
$$= V(\bar{Y}) + (x_0 - \bar{x})^2 V(\beta^*) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \qquad \Longrightarrow$$
$$\mu_0^* \in N\left(\mu_0, \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right).$$

Vi får således ett konfidensintervall för μ_0 med konfidensgraden $1 - a$ som

$$I_{\mu_0} = \mu_0^* \pm t_{a/2}(f) d(\mu_0^*) = \alpha^* + \beta^* x_0 \pm t_{a/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

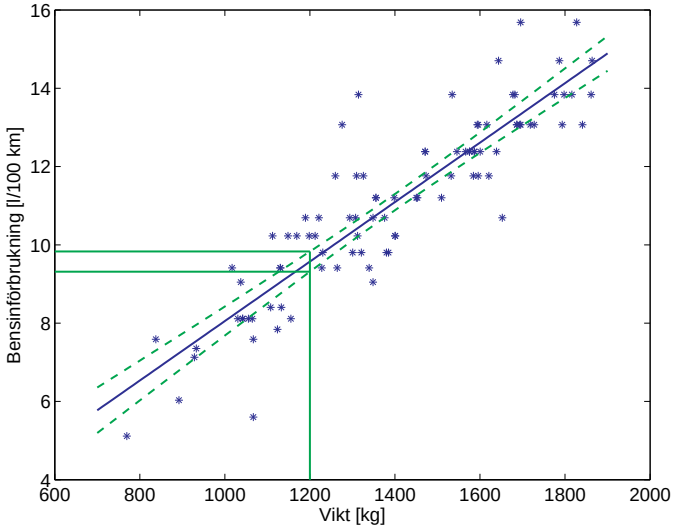
Exempel 2.3. Från exempel 1.1 på sid 3: I ett slumpmässigt urval av bilar avsattes y="bensinförbrukning i stadskörning" som funktion av x="vikt" i en linjär regressionsmodell $Y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \in N(0, \sigma)$. Parametrarna skattas enligt resultaten i avsnitt 2.3 till $\alpha^* = 0.46$, $\beta^* = 0.0076$ samt $\sigma^* = 1.009$.

β är ett mått på hur mycket y beror av x, om vikten ökas med ett kg skattas ökningen av bensinförbrukningen med $\beta^* = 0.0076$ liter per 100 kilometer. Ett 95% konfidensintervall för β blir $I_\beta = (0.0068, 0.0084)$.

Antag att vi är speciellt intresserade av bilar som väger $x_0 = 1200$ kg. En skattning av medelförbrukningen μ_0 för denna typ av bilar blir då $\mu_0^* = \alpha^* + \beta^* x_0 = 9.57$ l/100 km. Ett 95% konfidensintervall för μ_0 blir med ovanstående uttryck $I_{\mu_0} = (9.32, 9.83]$. Detta intervall täcker alltså med sannolikhet 95% den sanna medelförbrukningen för bilar med vikt 1200 kg.

Observera att intervallet inte ger någon information om individuella 1200 kg bilars variation, så det är inte till så mycket hjälp till att ge någon uppfattning om en framtida observation (den 1200 kg bil du tänkte köpa?). Till detta behövs ett *prediktionsintervall*, se nästa avsnitt.

I figur 6 är konfidensintervallen förutom för 1200 kg bilar även plottat som funktion av vikten. I formeln för konfidensintervallet ser man att det är som smalast då $x_0 = \bar{x}$ vilket även kan antydas i figuren. Man ser även att observationerna i regel inte täcks av konfidensintervallen för linjen.



Figur 6: Bensinförbrukning enligt exempel 1.1. Skattad regressionslinje (–), konfidensintervall för linjen som funktion av vikt (- -). Konfidensintervall för linjen då vikten är x_0 =1200 kg är markerat (–).

2.6 Prediktionsintervall för observationer

Intervallet ovan gäller väntevärdet för Y då $x = x_0$. Om man vill uttala sig om *en* framtida observation av Y för $x = x_0$ blir ovanstående intervall i regel för smalt. Om α, β och σ vore kända så skulle intervallet $\alpha + \beta x_0 \pm \lambda_{a/2}\sigma$ täcka en framtida observation Y med sannolikhet $1 - a$.

Eftersom regressionslinjen skattas med $\mu_0^* = \alpha^* + \beta^* x_0$ kan vi få hur mycket en framtida observation $Y(x_0)$ varierar kring den skattade linjen som

$$V(Y(x_0) - \alpha^* - \beta^* x_0) = V(Y(x_0)) + V(\alpha^* + \beta^* x_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right).$$

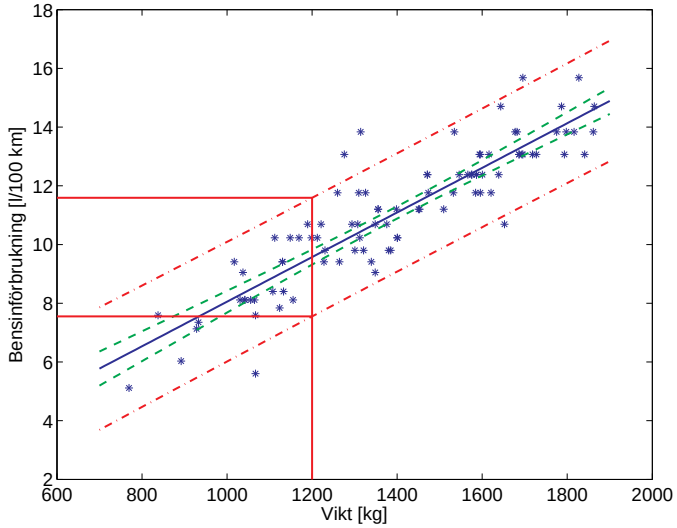
Vi kan alltså få ett *prediktionsintervall* med prediktionsgraden $1 - p$ för en framtida observation som

$$I_{Y(x_0)} = \alpha^* + \beta^* x_0 \pm t_{p/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

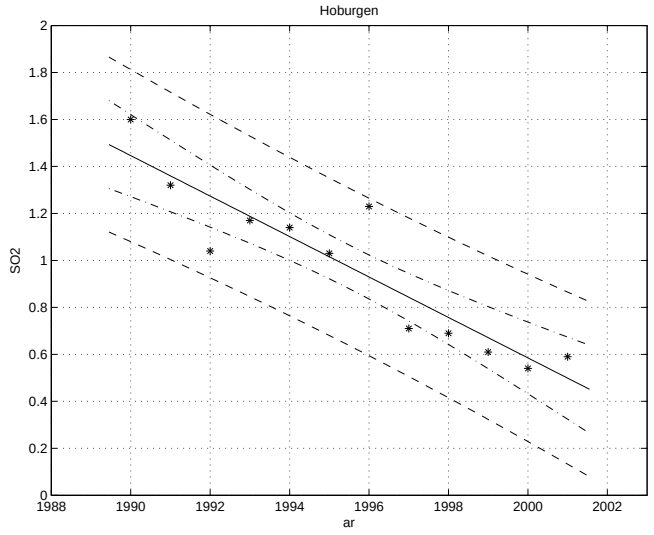
Observera att det bara är ettan i kvadratroten som skiljer mellan prediktionsintervallet och I_{μ_0}.

Exempel 2.4. Ett prediktionsintervall för bensinförbrukningen hos en 1200 kg bil enligt exempel 1.1 blir (7.6, 11.6) vilket är betydligt bredare än intervallet för väntevärdet. I figur 7 ses detta intervall och prediktionsintervallen som funktion av x_0.

Exempel 2.5. Vi anknyter till exemplet med SO_2-halterna igen. I figur 8 är både konfidensintervallet för linjens läge (det inre prick-streckade bandet) samt prediktionsintervallet (det yttre streckade bandet) uttritade som funktion av x_0 i Hoburgsdata.



Figur 7: Bensinförbrukning enligt exempel 1.1. Skattad regressionslinje (–), konfidensintervall för linjen som funktion av vikt (- -), prediktionsintervall för framtida observationer som funktion av vikt (-.). Prediktionsintervall för en framtida observation då vikten är x_0 =1200 kg är markerat (–).



Figur 8: Konfidensintervall för linjens läge (-.) samt prediktionsintervall (- -) för SO_2-halt (μg/m^3).

Vad är SO_2-linjens läge vid år 1996, d.v.s vad är förväntad SO_2-halt detta år? Ett 95% konfidensintervall för linjen beräknas till (0.83, 1.02) (jämför gärna med det inre bandet i figuren vid år 1996). Motsvarande prediktionsintervall (yttre band) för detta år är (0.59, 1.26), den uppmätta SO_2-halten 1996 hade alltså, med 95% sannolikhet, kunnat hamna någonstans mellan 0.59 och 1.26 μg/m^3.

På motsvarande sätt kan man använda prediktionsintervallet för att säga att uppmätt SO_2-halt år 2002, med 95% säkerhet, kommer att hamna någonstans i intervallet (0.03, 0.79) μg/m^3 (gör en försiktig extrapolation i figuren).

2.7 Kalibreringsintervall

Om man observerat ett värde y_0 på y, vad blir då x_0? Man kan lösa ut x_0 ur $y_0 = \alpha^* + \beta^* x_0$ och får

$$x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$$

Denna skattning är inte normalfördelad, men vi kan t.ex använda Gauss approximationsformler för att få en skattning av $d(x_0^*)$ och konstruera ett approximativt intervall

$$I_{x_0} = x_0^* \pm t_{a/2}(n-2)d(x_0^*) = \bar{x} + \frac{y_0 - \bar{y}}{\beta^*} \pm t_{a/2}(n-2) \cdot \frac{s}{|\beta^*|}\sqrt{1 + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{(\beta^*)^2 S_{xx}}}.$$

Ett annat sätt att konstruera kalibreringsintervallet är att dra en linje $y = y_0$ och ta skärningspunkterna med prediktionsintervallet som gränser i kalibreringsintervallet. Ett analytiskt uttryck för detta blir efter lite arbete

$$I_{x_0} = \bar{x} + \frac{\beta^*(y_0 - \bar{y})}{c} \pm \frac{t_{p/2}(n-2) \cdot s}{c}\sqrt{c(1 + \frac{1}{n}) + \frac{(y_0 - \bar{y})^2}{S_{xx}}}$$

$$c = (\beta^*)^2 - \frac{(t_{p/2}(n-2) \cdot s)^2}{S_{xx}}.$$

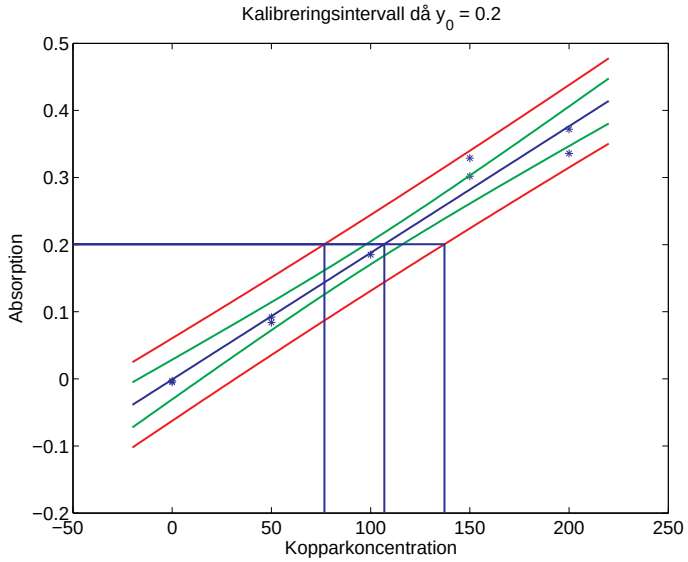
Uttrycket gäller då β är signifikant skild från noll annars är det inte säkert att linjen skär prediktionsintervallen. Grafiskt konstrueras detta intervall enligt figur 9.

2.8 Modellvalidering

2.8.1 Residualanalys

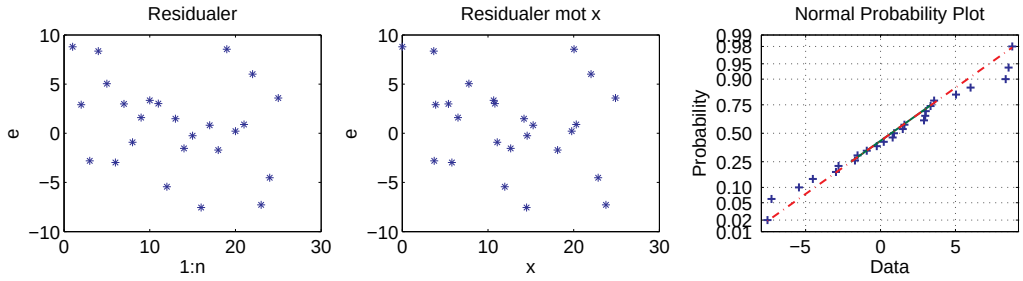
Modellen vi använder baseras på att avvikelserna från regressionslinjen är likafördelade ($\varepsilon_i \in \mathrm{N}(0, \sigma)$) och oberoende av varandra vilket medför att även observationerna Y_i är normalfördelade och oberoende. Dessa antaganden används då vi tar fram fördelningen för skattningarna. För att övertyga sig om att antagandena är rimliga kan det vara bra att studera avvikelserna mellan observerade y-värden och motsvarande punkt på den skattade linjen, d.v.s. de sedan tidigare definierade *residualerna*

$$r_i = y_i - (\alpha^* + \beta^* x_i), \quad i = 1, \ldots, n,$$



Figur 9: Kalibreringsintervall konstruerat som skärning med prediktionsintervall. I försöket har man för ett par prover med kända kopparkoncentrationer mätt absorption med atomabsorptionsspektrofotometri. Kalibreringsintervallet täcker med ungefär 95% sannolikhet den rätta kopparkoncentrationen för ett prov med okänd kopparhalt där absorptionen uppmätts till 0.2.

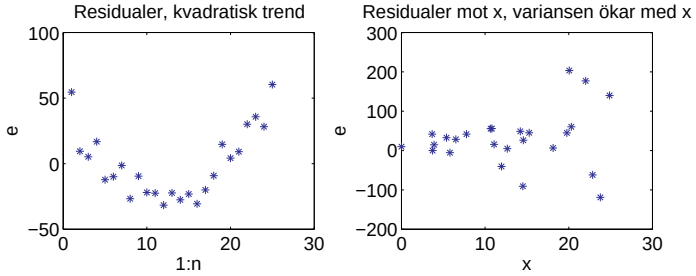
eftersom dessa är observationer av ε_i. Residualerna bör alltså se ut att komma från en och samma normalfördelning samt vara oberoende av dels varandra, samt även av alla x_i. I figur 10 visas några exempel på residualplottar som ser bra ut medan de i figur 11 ser mindre bra ut.



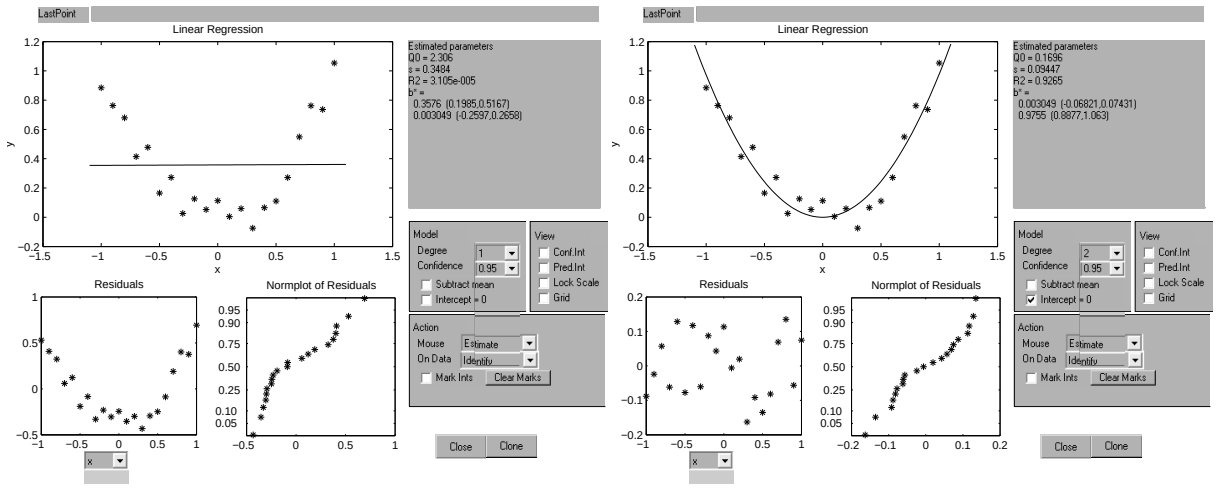
Figur 10: Bra residualplottar. Residualerna plottade i den ordning de kommer, mot x samt i en normalfördelningsplott. De verkar kunna vara oberoende normalfördelade observationer.

Exempel 2.6. Genom att studera graferna i figur 5 kan vi undersöka om den linjära modellen passar bra till Hoburgsdata. Residualplotten (nederst till vänster) visar inte några oroväckande trender och normalfördelningsplotten (nederst till höger) gör det rimligt att avvikelserna (residualerna) är normalfördelade. Sammantaget verkar det linjära modellen med oberoende och normfördelningsantagande vara rimlig i detta fall.

Exempel 2.7. I figur 12a) anpassades modellen $y_i = \alpha + \beta x_i + \epsilon_i$. Residualplotten i nedre vänstra hörnet säger att residualvärdet beror på x. Sambandet är alltså inte linjärt, snarare kvadratiskt. Om vi istället anpassar modellen



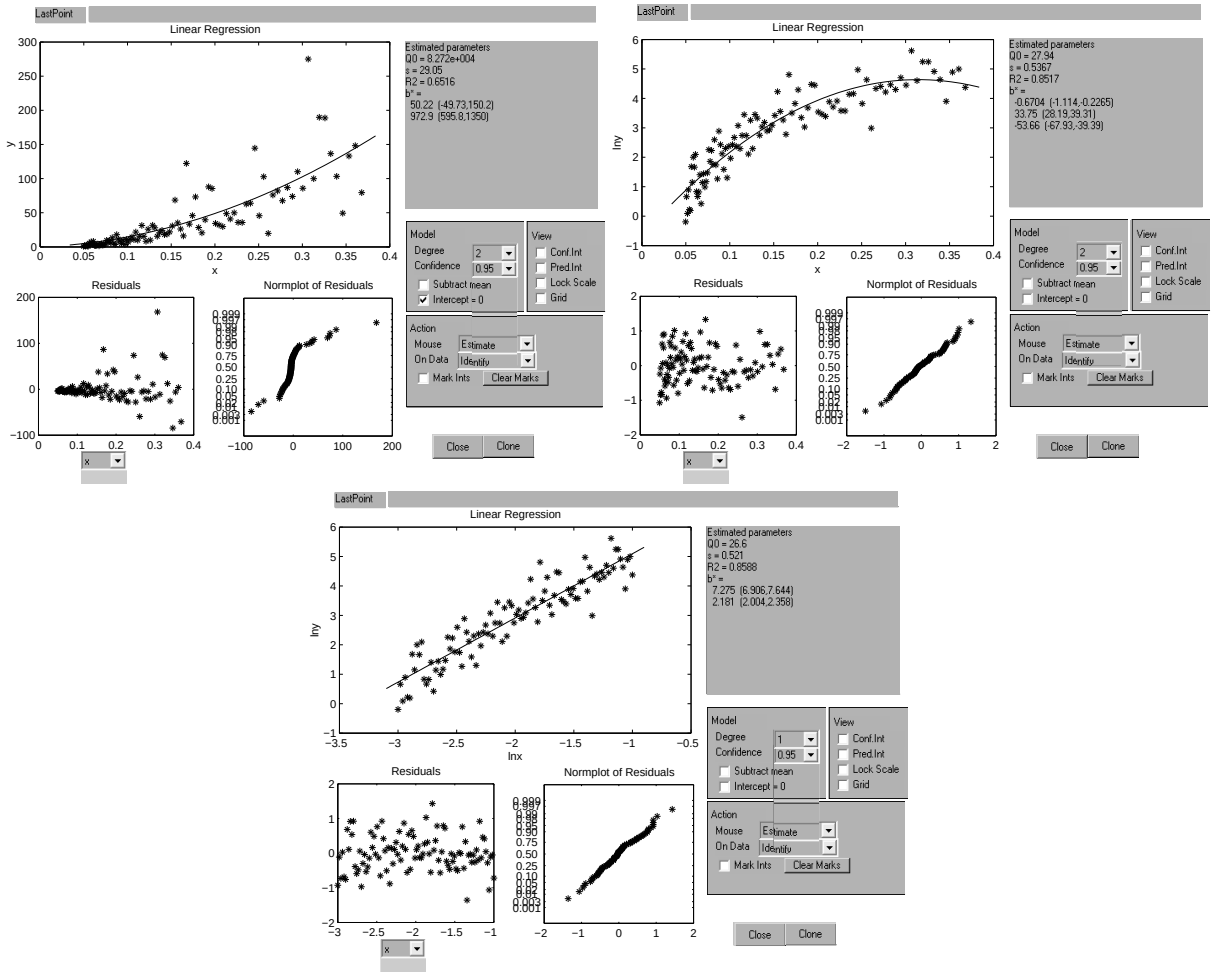
Figur 11: Residualplottar där man ser en tydlig kvadratisk trend i den vänstra figuren och i den högra ser man att variansen ökar med ökat x.



Figur 12: (a) Anpassning av linjär modell till kvadratiska data (vänster). (b) Anpassning av kvadratisk modell till kvadratiska data (höger).

$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ ser residualerna ut som de ska (se figur 12b).

Exempel 2.8. Anpassa den kvadratiska modellen $y_i = \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ (se figur 13a). Anpassningen är dålig eftersom residualernas varians ökar med x. För att åtgärda det anpassar vi istället modellen $\ln y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$

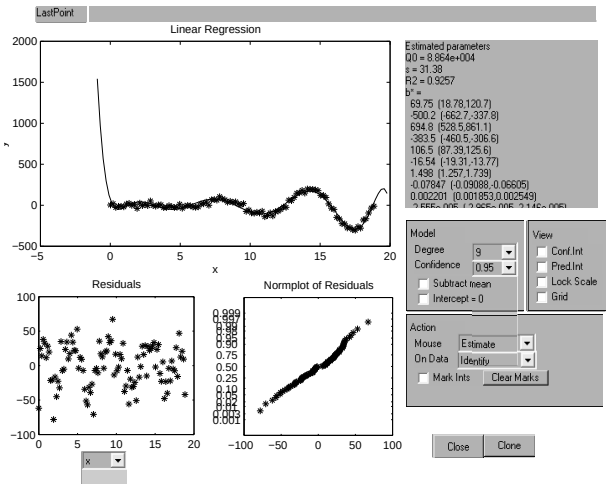


Figur 13: (a) Anpassning av kvadratisk modell (överst till vänster) (b) Anpassning av kvadratisk modell efter logartimering av y (överst till höger) (c) Anpassning av linjär modell efter logartimering av både y och x (underst)

(se figur 13b). Anpassningen är bättre eftersom residualvariansen nu är konstant. Däremot kan vi vara lite tveksamma till en kvadratisk modell eftersom modellen då säger att y ska avta för stora x. Det stämmer inte med observationerna. En bättre transformation är då att istället anpassa modellen $\ln y_i = \alpha + \beta_1 \ln x_i + \epsilon_i$ (se figur 13c). Nu ser residualerna ut som de ska.

Exempel 2.9. Det är inte säkert att det går att hitta en linjär modell eller en enkel tranformation som passar. Anpassa modellen $y_i = \alpha + \beta_1 x_i + \ldots + \beta_p x_i^p + \epsilon_i$ (se figur 14). Trots att vi anpassat ett polynom av högt gradtal finns det fortfarande struktur i residualerna och någon enkel transformation som skulle hjälpa är svårt att tänka ut! Antingen är det inte linjärt eller så är det inte

oberoende, eller båda, kanske är det en tidsserie[2]. Vill man lösa det problemet får man läsa *Stationära stokastiska processer*.



Figur 14: Anpassning av polynom till icke-linjärt samband

2.8.2 Är β signifikant?

Eftersom β anger hur mycket y beror av x är det även lämpligt att ha med följande hypotestest i en modellvalidering

$$H_0: \ \beta = 0$$
$$H_1: \ \beta \neq 0$$

t.ex. genom att förkasta H_0 om punkten 0 ej täcks av I_β. Om H_0 inte kan förkastas har y inget signifikant beroende av x och man kan kanske använda modellen $Y_i = \alpha + \varepsilon_i$ i stället.

2.9 Förklaringsgrad

En vanlig teknik när man analyserar data är att man försöker dela upp den variation som ses i mätningarna på olika variationskällor. Vid enkel linjär regression gäller uppdelningen:

"Total variation" = "variation förklarad av linjen" + "oförklarad variation", där

- "total variation"=$\sum_{i=1}^{n}(y_i - \bar{y})^2$, d.v.s. den variation som finns i y-värdena utan att vi tar hänsyn till x-värdena
- "variation förklarad av linjen"=$\sum_{i=1}^{n}((\alpha^* + \beta^* x_i) - \bar{y})^2$, vilket tolkas som den del av variationen i y-led som beskrivs av den linjära modellen
- "oförklarad variation"=$\sum_{i=1}^{n}(y_i - (\alpha^* + \beta^* x_i))^2$, vilket är identiskt med residualkvadratsumman Q_0 och tolkas som den "återstående" variation vi inte kan förklara med den linjära modellen.

[2]Modellen är i själva verket ickelinjär: $y_i = \sin(x_i) \cdot x_i^2 + \epsilon_i$

Ett mått på hur väl linjen förklarar data är kvoten mellan variation förklarad av linjen och total variation. Denna kvot är *förklaringsgraden*

$$R^2 = \frac{\sum_{i=1}^{n}((\alpha^* + \beta^* x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

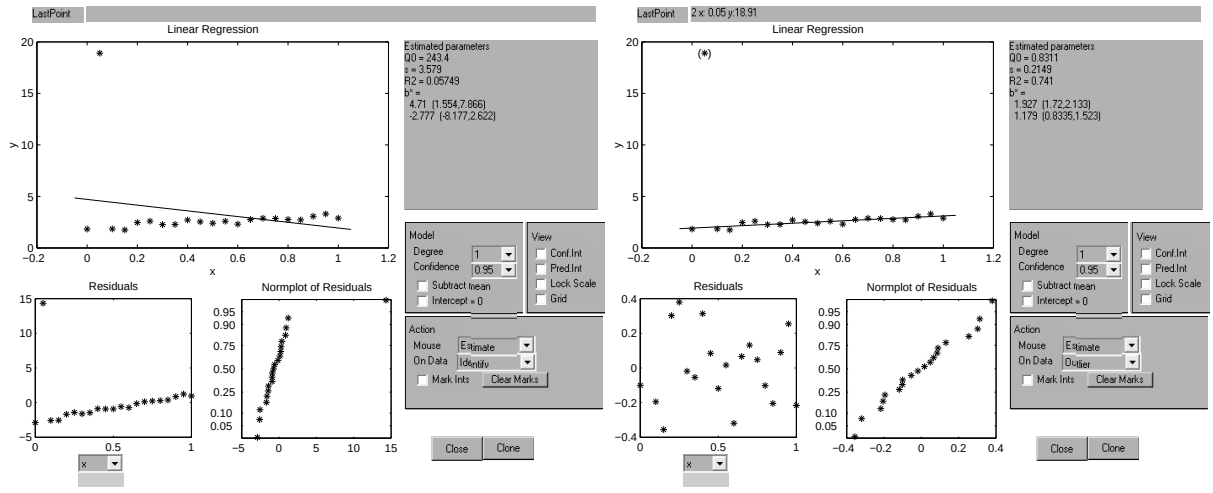
som ligger mellan noll och ett. Om R^2 har ett värde nära ett ligger talparen nära en rät linje - data kan därmed förklaras väl av den linjära modellen. Ett R^2-värde nära noll tyder på att data ej uppvisar ett speciellt linjärt samband och därmed inte förklaras bra av vår linjära modell.

> **Exempel 2.10.** Vid regressionsanalysen på Hoburgsdata i exempel 2.2. blev R^2=0.8356. Huvudparten, 84%, av den variation vi ser i SO_2-halt kan alltså förklaras med den linjärt avtagande trenden i mätningarna.

Förklaringsgraden är identisk med kvadraten på korrelationskoefficienten, se avsnitt 4.

2.10 Outliers

Det är viktigt att vara uppmärksam på *outliers*, dvs enskilda observationer som ligger misstänkt långt från de övriga och som får ett stort inflytande på skattningen av linjen (se figur 15). Outliers kan vara rena felinmatningar, i så fall bör de korrigeras eller plockas bort, men de kan också bero på naturlig variation i data. Då bör man överväga en modell som kan ta hänsyn till den variationen eller använda en mer robust skattningsmetod (ingår ej i denna kurs).



Figur 15: (a) Anpassad modell med en outlier (vänster) (b) Anpassad modell med outliern bortplockad (höger).

2.11 Linjärisering av några icke linjära samband

Vissa typer av exponential- och potenssamband med multiplikativa fel kan logaritmeras för att få en linjär relation. T.ex. fås när man logaritmerar

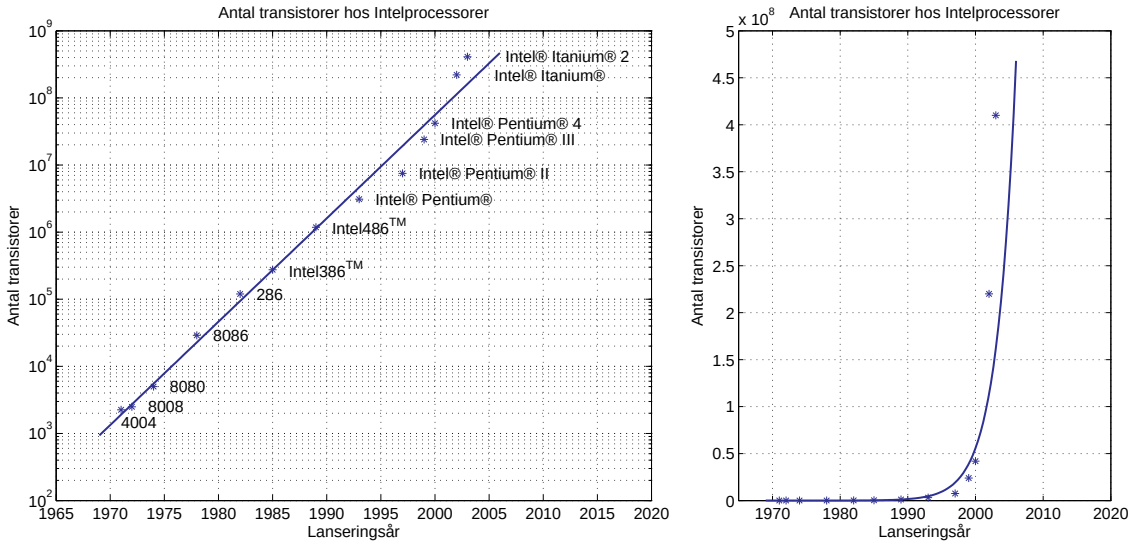
$$z_i = a \cdot e^{\beta x_i} \cdot \varepsilon_i' \qquad \xrightarrow{\ln} \qquad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \cdot x_i + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

ett samband på formen $y_i = \alpha + \beta x_i + \varepsilon_i$. Man logaritmerar således z_i-värdena och skattar α och β som vanligt och transformerar till den ursprungliga modellen med $a^* = e^{\alpha^*}$. Observera att de multiplikativa felen ε_i' bör vara lognormalfördelade (dvs $\ln \varepsilon_i' \in N(0, \sigma)$). En annan typ av samband är

$$z_i = a \cdot t_i^{\beta} \cdot \varepsilon_i' \qquad \xrightarrow{\ln} \qquad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \underbrace{\ln t_i}_{x_i} + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

där man får logaritmera både z_i och t_i för att få ett linjärt samband.

I figur 16 ses ett exempel där logaritmering av y-värdena ger ett linjärt samband.



Figur 16: Antal transistorer på en cpu mot lanseringsår med logaritmisk y-axel i vänstra figuren. Till höger visas samma sak i linjär skala. Det skattade sambandet är $y = 5.13 \cdot 10^{-301} \cdot e^{0.35x}$.

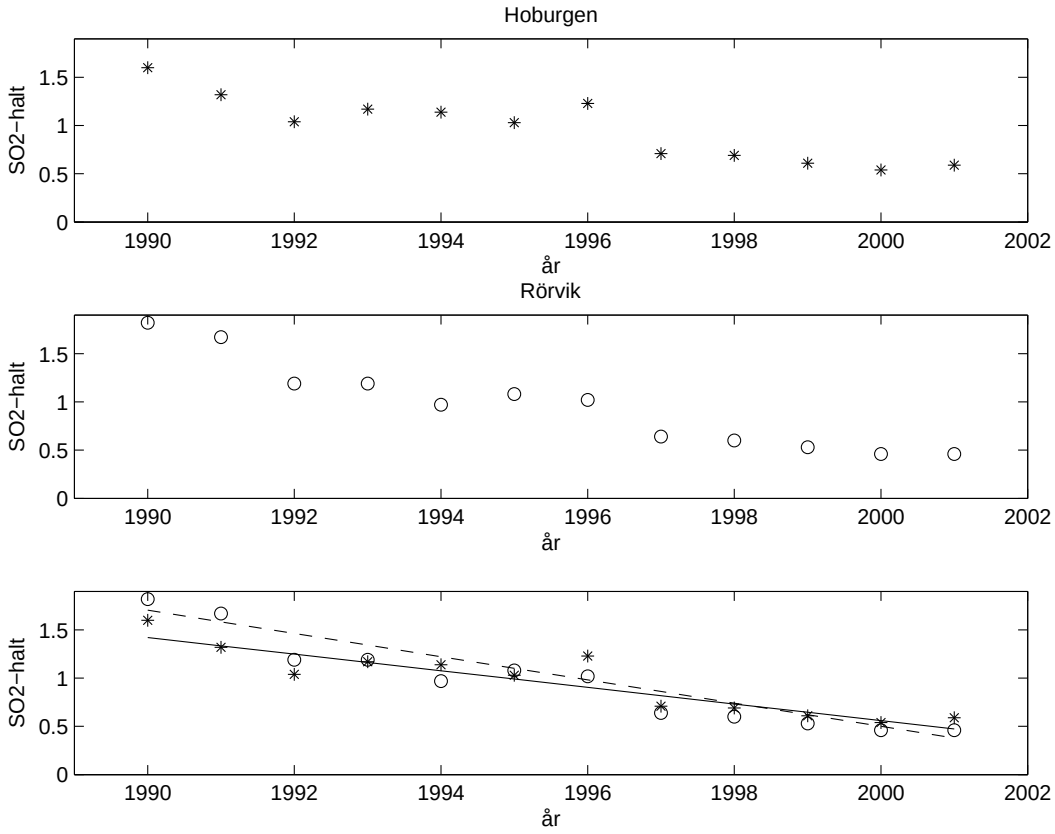
2.12 Jämförelse av två lutningar

Ibland har man en situation där man vill undersöka om regressionssambandet kan vara identiskt för olika grupper. Är t.ex. sambandet mellan blodtryck och ålder det samma för både män och kvinnor? Speciellt intressant kan det vara att studera om den årliga blodtrycksökningen är likartad för de båda könen. Om vi som modell använder två linjära regressionssamband (en för kvinnor och en för män) motsvaras problemet av att jämföra lutningarna i de två sambanden, d.v.s. undersöka om $\beta_{kvinna} = \beta_{man}$. Ett exempel får illustrera metodiken.

Exempel 2.11. SO_2-halten bestämdes inte enbart vid Hoburgen på Gotland utan även vid Rörvik i norra Halland (figur 17). Är trenden i SO_2-halt den samma vid de två mätstationerna eller skiljer den sig åt?

Vi tänker oss att för Hoburgen och mätningarna $(x_1, y_1), \ldots (x_i, y_i), \ldots (x_{n_H}, y_{n_H})$ har vi modellen

$$y_i = \alpha_H + \beta_H \cdot x_i + \epsilon_i, \qquad \epsilon_i \in N\,(0,\, \sigma_H)$$



Figur 17: SO_2-halt vid Hoburgen (överst) samt i Rörvik (mitten). Underst visas mätningarna från båda stationerna med skattade regressionslinjer utritade (heldragen linje för Hoburgen och streckad för Rörvik)

och för Rörvik och mätningarna $(x_1, y_1), \ldots (x_j, y_j), \ldots (x_{n_R}, y_{n_R})$ har vi modellen

$$y_j = \alpha_R + \beta_R \cdot x_j + \epsilon_j, \qquad \epsilon_j \in N\,(0,\, \sigma_R)\,.$$

Genom att göra två separata analyser i Matlab får vi för Hoburgen skattningarna (resultaten är hämtade från exempel 2.2).

$$\alpha_H^* = 172.8; \quad \beta_H^* = -0.08612; \quad \sigma_H^* = 0.1445$$

medan motsvarande för Rörvik är

$$\alpha_R^* = 241.5; \quad \beta_R^* = -0.1205; \quad \sigma_R^* = 0.1436$$

Nu är vi intresserade av hur stor $\beta_R - \beta_H$ är och en **skattning** av denna storhet kan vi få genom $\beta_R^* - \beta_H^* = -0.1205 - (-0.08612) = -0.0344$.

Vill vi göra **konfidensintervall** för differensen $\beta_R - \beta_H$ måste vi ha en uppfattning om "hur bra" denna skattning är, d.v.s. veta variansen för $\beta_R^* - \beta_H^*$. Men från tidigare vet vi att

$$\mathsf{V}(\beta_R^*) = \frac{\sigma_R^2}{S_{Rxx}}$$

där $S_{Rxx} = \sum (x_j - \bar{x})^2$ är kvadratsumman på de x-värden som användes vid Rörviksmätningarna. För Hoburgen har vi på motsvarande sätt

$$\mathsf{V}(\beta_H^*) = \frac{\sigma_H^2}{S_{Hxx}}$$

där S_{Hxx} är kvadratsumman på de x-värden som användes vid Hoburgsmätningarna. Men eftersom x-värdena består av 11 årtal med start 1990 och slut 2001 och vi dessutom mäter vid samma år vid de två stationerna gäller att $S_{Hxx} = S_{Rxx} = 143$.

Om vi dessutom kan anta att $\sigma_H = \sigma_R$ (verkar rimligt i detta exempel) kan vi kalla denna gemensamma standardavvikelse för σ. Detta ger

$$V(\beta_R^* - \beta_H^*) = V(\beta_R^*) + V(\beta_H^*) = \sigma^2(\frac{1}{S_{Rxx}} + \frac{1}{S_{Hxx}}).$$

För att beräkna en skattning av den gemensamma standardavvikelsen gör vi en "poolning" av standardavvikelserna av samma slag som tidigare (observera n-2)

$$\sigma^{2*} = \frac{(n_R - 2) \cdot \sigma_R^{2*} + (n_H - 2) \cdot \sigma_H^{2*}}{(n_R - 2) + (n_H - 2)} =$$

$$\frac{(12 - 2) \cdot 0.1436^2 + (12 - 2) \cdot 0.1445^2}{(12 - 2) + (12 - 2)} = 0.0208.$$

Nu kan vi konstruera ett 95% intervall på välbekant sätt:

$$I_{\beta_R - \beta_H} = (\beta_R^* - \beta_H^* \pm t_{a/2}(n_R - 2 + n_H - 2) d(\beta_R^* - \beta_H^*)) =$$

$$(\beta_R^* - \beta_H^* \pm t_{a/2}(n_R - 2 + n_H - 2)\sqrt{\sigma^{2*}(\frac{1}{S_{Rxx}} + \frac{1}{S_{Hxx}})}) =$$

$$(-0.0344 \pm 2.09 \cdot \sqrt{0.0208(\frac{1}{143} + \frac{1}{143})}) = (-0.0344 \pm 0.0356) = (-0.070,\ 0.0012).$$

Eftersom detta intervall täcker över 0 har vi inte påvisat att det finns en skillnad mellan lutningarna. Dessa mätningar tyder alltså inte på att trenden i SO_2 skiljer sig åt vid de två stationerna.

3 Multipel linjär regression på matrisform

Med matrisnotation kan en allmän linjär regressionsmodell med p st förklarande x-variabler, av typen

$$y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

vare sig den är enkel eller multipel, skrivas

$$y = X\beta + e,$$

där de ingående matriserna har följande form:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{p1} \\ 1 & x_{12} & \ldots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ och } e = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Rent allmänt fås minsta-kvadratlösningen β^* till ett överbestämt ekvationssystem $y = X\beta$ via de så kallade normalekvationerna

$$X^t X\beta = X^t y,$$

som $\beta^* = (X^t X)^{-1} X^t y$. Man bör dock i möjligaste mån undvika att lösa ut β genom att invertera matrisen $X^t X$. Om matrisen är illa konditionerad kan man nämligen få en feltillväxt som gör resultatet helt oanvändbart. En numeriskt sett effektivare och mer stabil lösning fås om man i Matlab använder operatorn \ som kan uppfattas som vänsterdivision.

Det rekommenderade sättet att lösa matrisekvationen ovan är alltså

```
>> b = X\y
```

Skattningen av σ fås genom

$$\sigma^* = s = \sqrt{\frac{Q_0}{n - (p+1)}}$$

där Q_0 kan beräknas antingen som $Q_0 = \mathbf{y}^t\mathbf{y} - \beta^{*t}\mathbf{X}^t\mathbf{y}$, eller genom att uttnyttja att $Q_0 = \sum_{i=1}^{n} r_i^2 = \mathbf{r}^t\mathbf{r}$ där residualerna $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$. Den s.k. *kovariansmatrisen* för β^* ges av $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ vilket innebär att medelfelen $\mathsf{d}(\beta_0^*)$, $\mathsf{d}(\beta_1^*)$, etc, fås som roten ur respektive diagonalelement i $s^2(\mathbf{X}^t\mathbf{X})^{-1}$. Den skattade linjen i punkten $\mathbf{x}_0 = \begin{pmatrix} 1 & x_0^{(1)} & x_0^{(2)} \end{pmatrix}$ ges av $\mu_0^* = \mathbf{x}_0\beta^* \in N\left(\mu_0,\, \sigma\sqrt{\mathbf{x}_0(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0^t}\right)$.

Exempel 3.1. För att undersöka pressningstemperaturens och pressningstryckets inverkan vid tillverkning av en typ av plastkomposit iordningställdes två provbitar för var och en av fem kombinationer av tryck och temperatur. Böjspänningen hos de olika provbitarna av plastkompositen mättes och blev

Böjspänning (y) ($\mathrm{N/mm^2}$)	Temperatur (x_1) (°C)	Tryck (x_2) ($\mathrm{kg/cm^2}$)
152	180	450
150	180	450
103	190	375
99	190	375
88	200	350
89	200	350
122	210	375
120	210	375
162	220	450
161	220	450

Anpassa modellen $y_i = \alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+\epsilon_i$ och gör ett 95 % konfidensintervall för hur mycket böjspänningen ökar då temperaturen ökar med 1 °C. Gör också ett 95 % prediktionsintervall för böjspänningen då temperaturen är 200 °C och trycket $400\,\mathrm{kg/cm^2}$.

Lösning: Skriv om modellen $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ som $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ med

$$\mathbf{y} = \begin{pmatrix} 152 \\ 150 \\ 103 \\ 99 \\ 88 \\ 89 \\ 122 \\ 120 \\ 162 \\ 161 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1\ 180\ 450 \\ 1\ 180\ 450 \\ 1\ 190\ 375 \\ 1\ 190\ 375 \\ 1\ 200\ 350 \\ 1\ 200\ 350 \\ 1\ 210\ 375 \\ 1\ 210\ 375 \\ 1\ 220\ 450 \\ 1\ 220\ 450 \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Parameterskattningar blir

$$\beta^* = \mathbf{X}\backslash\mathbf{y} = \begin{pmatrix} -215.7 \\ 0.41 \\ 0.65 \end{pmatrix} = \begin{pmatrix} \alpha^* \\ \beta_1^* \\ \beta_2^* \end{pmatrix}$$

och, eftersom $Q_0 = \mathbf{r}^t\mathbf{r} = (\mathbf{y} - \mathbf{X}\beta^*)^t(\mathbf{y} - \mathbf{X}\beta^*) = 243.63$,

$$\sigma^* = s = \sqrt{\frac{Q_0}{n-(p+1)}} = \sqrt{\frac{243.63}{10-(2+1)}} = 5.90.$$

Ökningen i böjspänning då temperaturen ökar en grad ges av β_1. För att kunna beräkna konfidensintervall för β_1 behöver vi också beräkna

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{pmatrix} 29.24 & -0.1 & -0.0229 \\ -0.1 & 0.0005 & 0 \\ -0.0229 & 0 & 0.0001 \end{pmatrix}$$

Sedan kan vi få medelfelet $\mathsf{d}(\beta_1^*) = s\sqrt{0.0005}$, där vi tagit andra diagonalelementet i $(\mathbf{X}^t\mathbf{X})^{-1}$. Det första diagonalelementet gäller ju α^* och det tredje β_2^*. Ett konfidensintervall för β_1 med konfidensgrad $1-a$ fås sedan på vanligt sätt som

$$\begin{aligned} I_{\beta_1} &= \big(\beta_1^* \pm t_{a/2}(n-(p+1)) \cdot \mathsf{d}(\beta_1^*)\big) \\ &= (0.41 \pm \underbrace{t_{0.025}(7)}_{2.36} \cdot 5.90\sqrt{0.0005}) \\ &= (0.098,\, 0.722)\,\mathrm{N/mm^2} \text{ per } {}^\circ\mathrm{C}. \end{aligned}$$

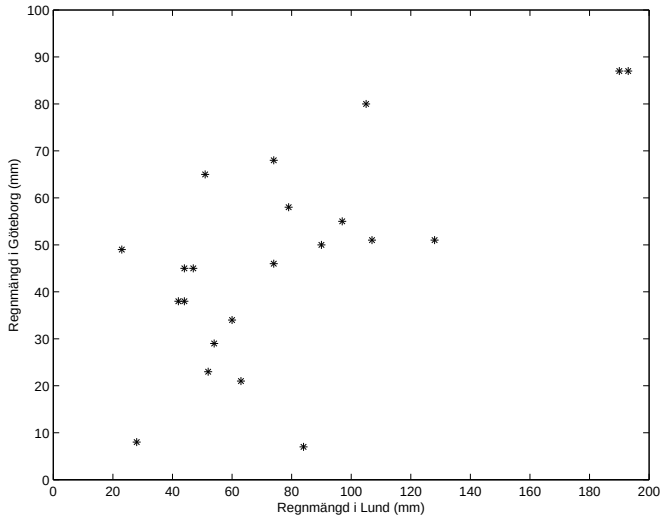
För att göra ett prediktionsintervall för Y_0 då $x_0^{(1)} = 200\,^\circ\mathrm{C}$ och $x_0^{(2)} = 400\,\mathrm{kg/cm^2}$ sätter vi $\mathbf{x}_0 = \begin{pmatrix} 1 & 200 & 400 \end{pmatrix}$ och får skattningen av sambandet till $\mu_0^* = \mathbf{x}_0\beta^* = 124.6$ med medelfelet $\mathsf{d}(\mu_0^*) = s\sqrt{\mathbf{x}_0(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0^t} = 5.90\sqrt{0.1} = 6.187$. Eftersom vi vill ha ett prediktionsintervall, inte ett konfidensintervall, ska vi lägga till en etta under rottecknet så att intervallet ges av

$$\begin{aligned} I_{Y(\mathbf{x}_0)} &= \Big(\mathbf{x}_0\beta^* \pm t_{a/2}(n-(p+1)) \cdot s\sqrt{1+\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}\Big) \\ &= (124.6 \pm \underbrace{t_{0.025}(7)}_{2.36} \cdot 5.90\sqrt{1+0.1}) = (110.0,\, 139.2)\,\mathrm{N/mm^2} \end{aligned}$$

4 Korrelationsanalys

Regressionsanalysen i föregående avsnitt förutsatte att x-variablerna var ”fixa” i den meningen att de var uppmätta med inget eller försumbart mätfel. Om detta inte är uppfyllt är det lämpligare att göra en korrelationsanalys där man inte försöker anpassa någon regressionsfunktion till data utan enbart mäter graden av samband.

Exempel 4.1. I exempel 1.2 på sidan 3 noterades månadsnederbörden, d.v.s. den totala mängden nederbörd (mm) under en månad, i Göteborg och Lund under åren 2005 och 2006. I figur 18 markerar varje punkt en månad där Göteborgs nederbörd avläses på y-axeln och Lunds på x-axeln.



Figur 18: Månadsvisa mätningar av nederbörden (mm) där $y =$ ”nederbörd i Göteborg” är plottad mot $x =$ ”nederbörd i Lund”.

Från figuren tycks det finnas ett positivt samband mellan nederbördsmätningarna från de två städerna - regnar det mycket en månad i den ena staden tenderar det också att göra det i den andra.

4.1 Mått på samband

Som ett mått på samband mellan två variabler X och Y används kovariansen eller korrelationskoefficienten mellan variablerna. Kovariansen definieras som

$$C(X,Y) = E[(X-\mu_x)(Y-\mu_y)],$$

där μ_x och μ_y är väntevärdena för X och Y. Korrelationskoefficienten, ρ_{xy} är den normerade storheten

$$\rho_{xy} = \frac{C(X,Y)}{D(X)\cdot D(Y)},$$

där $D(X) = \sqrt{V(X)}$ är standardavvikelsen för X (och motsvarande för $D(Y)$). För korrelationskoefficienten gäller alltid att $-1 \leq \rho_{xy} \leq 1$.

Tolkning av de två storheterna är oftast enklast då man betraktar motsvarande skattningar. Antag att vi har n mätningar vardera av de två variablerna och därmed de n talparen $(x_1, y_1), \ldots, (x_n, y_n)$. En skattning av kovariansen är då

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

och av korrelationskoefficienten

$$\rho^*_{xy} = r_{xy} = \frac{c_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

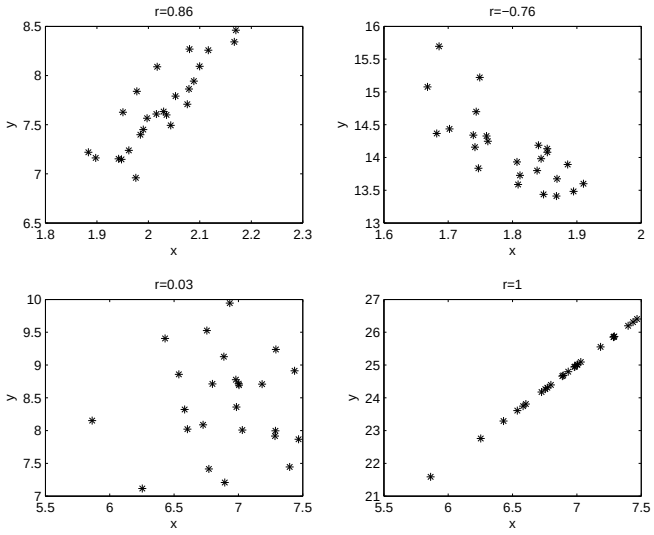
Observera att uttrycket $n - 1$ förkortats bort i sista ledet. För r_{xy} gäller att

$$-1 \leq r_{xy} \leq 1$$

samt att om vi har

positiv samvariation (positiv korrelation) mellan X och Y, d.v.s. $\rho_{xy} > 0$	tenderar	$r_{xy} > 0$
negativ samvariation (negativ korrelation) mellan X och Y, d.v.s. $\rho_{xy} < 0$	tenderar	$r_{xy} < 0$
ingen samvariation (ingen korrelation) mellan X och Y, d.v.s. $\rho_{xy} = 0$	tenderar	$r_{xy} \approx 0$

Om $r_{xy} = 1$ innebär det att x-värdena och y-värdena ligger på en linje med positiv lutning; se figur 19.



Figur 19: Figurerna visar olika grad av samband med tillhörande korrelationskoefficient.

Observera att om r_{xy} ligger nära 0 tyder det på att det inte finns någon samvariation mellan de två variablerna (de är okorrelerade), däremot följer det inte att x och y är oberoende. Om x-värdena och y-värdena däremot är hämtade från normalfördelning är okorrelerad identiskt med oberoende.

4.2 Test av samband

I exemplet med månadsnederbörd från Lund och Göteborg gav beräkningar i Matlab att $r_{xy} = 0.662$. Data tyder alltså på en positiv samvariation - men är värdet på r_{xy}

tillräckligt stort för att vi ska kunna tro på att det **verkligen** finns en samvariation och att det observerade resultatet inte bara är ett utslag av slumpen?

Om $r_{xy} =$ är en skattning av den korrelation, ρ_{xy}, som finns mellan de s.v. X och Y vill vi alltså undersöka om ρ_{xy} är 0. De intressanta hypoteserna är:

$$H_0 : \rho_{xy} = 0 \text{ (inget samband)}; \quad H_1 : \rho_{xy} \neq 0 \text{ (samband)}.$$

För att testa detta används storheten

$$t = r_{xy}\sqrt{(n-2)/(1-r_{xy}^2)}.$$

Om data kommer från en bivariat normalfördelning gäller nämligen att t är t-fördelad med $n-2$ frihetsgrader när H_0 är sann.

Exempel 4.2. Med ett värde $r_{xy} = 0.662$ i nederbördsdata blir

$$t = r_{xy}\sqrt{(n-2)/(1-r_{xy}^2)} = 0.662\sqrt{(23-2)/(1-0.662^2)} = 3.95.$$

Eftersom 3.95 överstiger $t_{0.0005}(21) = 3.82$ innebär det att korrelationen är signifikant skild från 0 på nivå 0.001. Det finns alltså en positiv samvariation mellan de två städernas månadsnederbörd.

4.3 Var försiktig med korrelationskoefficienten!

Det finns en rad ”fallgropar” när man hanterar korrelationskoefficienter. Några exempel:

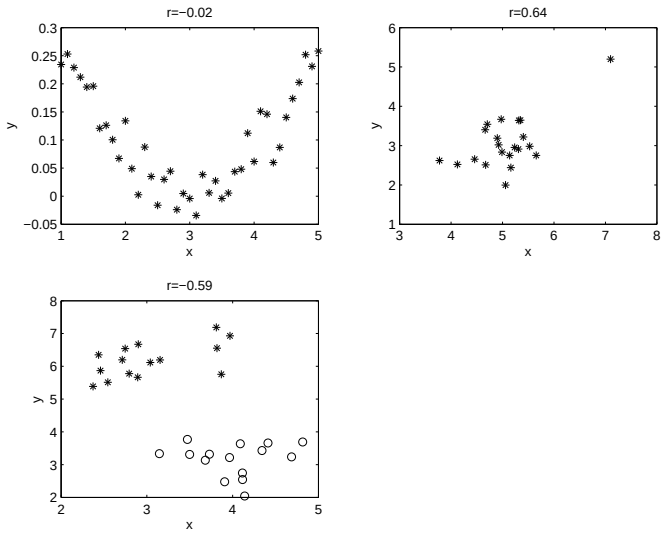
- r_{xy} mäter graden av linjärt samband - i figur 20(a) fås ett värde på r_{xy} som är ungefär 0 eftersom den negativa lutningen i figurens vänstra halva ”tas ut” av den positiva lutningen i andra halvan.

- r_{xy} är känslig för outliers, d.v.s. kraftigt avvikande värden kan starkt påverka värdet på korrelationskoefficenten. Utan outliern i figur 20(b) är r_{xy}=0.24, med outliern blir r_{xy}=0.64 .

- r_{xy} kan bli missvisande då den används på mätningar som naturligt kan delas upp i två grupper (t.ex. kön) och där genomsnittsvärdena för x och y är olika i de två grupperna. I figur 20(c) verkar det inte finnas någon samvariation inom respektive grupp (eller eventuellt en positiv samvariation för ”stjärnorna”) men betraktar man hela materialet - och beräknar okritiskt r_{xy} - tyder korrelationskoefficienten på en negativ samvariation mellan X och Y.

Samtliga dessa ”fällor” kan man förmodligen upptäcka om man alltid tar för vana att plotta sina data och inte bara slentrianmässigt beräknar korrelationskoefficienten.

Viktigare är det att komma ihåg att med korrelationskoefficienten mäter vi (och eventuellt påvisar) ett **statistiskt samband**. Det är därmed inte sagt att det finns ett **orsakssamband** mellan variablerna!



Figur 20: Figurerna visar några situationer där korrelationskoefficienten inte okritiskt kan användas.

Exempel 4.3. Om man för ett antal städer noterar dels antal läkare i staden och dels antalet sjukdagar som stadens innevånare har under ett år kommer man säkert att finna ett positivt samband mellan de två variablerna. Innebär det då att ju fler läkare man har i en stad medför det fler sjukdagar och att vi kan minska antalet sjukdagar genom att minska antalet läkare? Nej, naturligtvis inte; här är det en tredje faktor - antalet invånare i staden - som påverkar de båda undersökta variablerna.

4.4 Anknytning till linjär regression

Korrelationskoefficienten mäter det *linjära* sambandet mellan x och y - alltså borde det kunnna användas även vid linjär regression. I själva verket är kvadraten på korrelationskoefficienten matematiskt identisk med förklaringsgraden som beskrevs i avsnitt 2.9, d.v.s.

$$r_{xy}^2 = R^2.$$

Vid en regressionsanalys - antingen den beskrivs i datorprogram eller i rapporter - anges därför även ofta korrelationskoefficienten. Den är då ett mått på hur ”stor nytta” man har av x-variabeln då man vill förutsäga y. Om r_{xy} är nära 1 (eller -1) betyder det att x och y ligger nästan på en linje och därmed kan y nästan förutsägas direkt utifrån x-värdet. Förklaringsgraden R^2 är då också nära 1. Om däremot värdet på r_{xy} är lågt (vilket ger en låg förklaringsgrad) är sambandet mellan variablerna svagt och y kan näppeligen förutsägas av enbart x.

Test av samband, som beskrivs i avsnitt 4.2, visar sig också vara identiskt med att testa att lutningen $\beta = 0$ (se avsnitt 2.8.2) i regressionsmodellen.

Observera dock - vilket vi redan påpekat - att det finns en skillnad i antagandena om x-värdena när det gäller regressionsanalys respektive korrelationsanalys. För förklaringsgraden R^2 i regressionsanalysen anses x-värdena vara fixa och att vi, i stort sett, kan själva bestämma dess värde. I korrelationsanalysen är däremot x-värdena och y-värdena ”utbytbara”.

5 Appendix: ML- och MK skattningar av parametrarna i enkel linjär regression

5.1 Några hjälpresultat

Vi börjar med ett par användbara beteckningar och räkneregler för de summor och kvadratsummor som kommer att ingå i skattningarna. Då alla summor nedan löper från 1 till n avstår vi från att skriva ut summationsindexen.

Först har vi att en ren summa av avvikelser av ett antal observationer kring sitt medelvärde är noll

$$\sum(x_i - \bar{x}) = \sum x_i - n\bar{x} = [\bar{x} = \frac{1}{n}\sum x_i] = \sum x_i - \sum x_i = 0 \tag{1}$$

Några beteckningar för kvadratiska- och korsavvikelser kring medelvärde

$$S_{xx} = \sum(x_i - \bar{x})^2, \quad S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}), \quad S_{yy} = \sum(y_i - \bar{y})^2$$

där vi känner igen den första och sista från stickprovsvarianserna för x resp. y, $s_x^2 = S_{xx}/(n-1)$ och motsvarande för y. Dessa summor kan skrivas på ett antal former, t.ex kan S_{xy} utvecklas till

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) - \bar{x}\sum(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) \quad \text{eller}$$
$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x}) = \sum(x_i - \bar{x})y_i$$

där sista summan i andra leden blir noll enligt (1). Motsvarande räkneregler gäller för S_{xx} och S_{yy} och vi har sammanfattningsvis

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i \tag{2}$$
$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i(x_i - \bar{x}) \quad \text{och motsvarande för } S_{yy} \tag{3}$$

5.2 Punktskattningar

ML-skattning av α, β och σ^2 då y_i är oberoende observationer av $Y_i \in N(\alpha + \beta x_i, \sigma)$ fås genom att maximera likelihood-funktionen

$$L(\alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1 - \alpha - \beta x_1)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n - \alpha - \beta x_n)^2}{2\sigma^2}} = (2\pi)^{n/2} \cdot (\sigma^2)^{n/2} \cdot e^{-\frac{1}{2\sigma^2}\sum(y_i - \alpha - \beta x_i)^2}$$

Hur än σ väljs så kommer L att maximeras med avseende på α och β då $\sum(y_i - \alpha - \beta x_i)^2$ är minimal, och eftersom det är just denna kvadratsumma som minimeras med MK-metoden så blir skattningarna av α och β de samma vid de två metoderna. Med ML-metoden kan vi dessutom skatta σ^2 varför vi väljer den. Logaritmeras likelihoodfunktionen fås

$$\ln L(\alpha, \beta, \sigma^2) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \alpha - \beta x_i)^2$$

Deriveras denna med avseende på var och en av parametrarna och sedan sättes till noll fås ekvationssystemet

$$\frac{\partial \ln L}{\partial \alpha} = \frac{1}{\sigma^2}\sum(y_i - \alpha - \beta x_i) = 0 \tag{4}$$
$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum(y_i - \alpha - \beta x_i)x_i = 0 \tag{5}$$
$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i - \alpha - \beta x_i)^2 = 0 \tag{6}$$

att lösa med avseende på α, β och σ^2. Eftersom vi kan förlänga de två första ekvationerna med σ^2 och därmed bli av med den kan vi använda dessa till att skatta α och β. (4) och (5) kan formas om till

$$\begin{aligned}\sum y_i &= n\alpha + \beta\sum x_i \\ \sum x_i y_i &= \alpha\sum x_i + \beta\sum x_i^2\end{aligned} \tag{7}$$

Delas första ekvationen med n fås

$$\bar{y} = \alpha + \beta\bar{x} \quad \Longleftrightarrow \quad \alpha = \bar{y} - \beta\bar{x} \tag{8}$$

som vi kan stoppa in i (7) som då blir

$$\begin{aligned}&\sum x_i y_i = \bar{y}\sum x_i - \beta\bar{x}\sum x_i + \beta\sum x_i^2 \quad \Longleftrightarrow \\ &\sum x_i y_i = \beta(\sum x_i^2 - \bar{x}\sum x_i) + \bar{y}\sum x_i \quad \Longleftrightarrow \\ &\beta = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_j^2 - \bar{x}\sum x_j} = \frac{\sum x_i(y_i - \bar{y})}{\sum x_j(x_j - \bar{x})} = [(2)] = \frac{\sum(x_i - \bar{x})y_i}{\sum x_j(x_j - \bar{x})} = [(2) \text{ och } (3)] = \frac{S_{xy}}{S_{xx}}\end{aligned} \tag{9}$$

Detta resultat tillsammans med (8) ger ML-skattningarna av α och β

$$\beta^* = \frac{S_{xy}}{S_{xx}}, \quad \alpha^* = \bar{y} - \beta^*\bar{x}$$

Dessa värden insatta i (6) förlängd med σ^4 ger

$$(\sigma^2)^* = \frac{1}{n}\sum(y_i - \alpha^* - \beta^* x_i)^2$$

som dock inte är väntevärdesriktig utan korrigeras till

$$(\sigma^2)^* = s^2 = \frac{1}{n-2}\sum(y_i - \alpha^* - \beta^* x_i)^2 = \frac{Q_0}{n-2}$$

som är det. Q_0 som är summan av kvadratiska avvikelser från observationerna y_i till motsvarande punkt på den skattade linjen kallas *residualkvadratsumma* och den kan skrivas på formen

$$Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

5.3 Skattningarnas fördelning

Om vi börjar med β^* och utgår från (9)

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})y_i}{\sum x_j(x_j - \bar{x})} = \sum c_i y_i \quad \text{där} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \tag{10}$$

den är alltså en linjär funktion av de normalfördelade observationerna och därmed är skattningen normalfördelad. Väntevärdet blir

$$\begin{aligned} E(\beta^*) &= E(\sum c_i Y_i) = \sum c_i E(Y_i) = \sum c_i(\alpha + \beta x_i) = \frac{1}{S_{xx}} \sum (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \frac{\alpha}{S_{xx}} \sum (x_i - \bar{x}) + \frac{\beta}{S_{xx}} \sum (x_i - \bar{x})x_i = 0 + \beta \frac{S_{xx}}{S_{xx}} = \beta \end{aligned}$$

där vi i näst sista ledet åter använde hjälpresultaten (2) och (3). Skattningen är alltså väntevärdesriktig och dess varians blir

$$V(\beta^*) = V(\sum c_i Y_i) = \sum c_i^2 V(Y_i) = \sum c_i^2 \sigma^2 = \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$$

dvs

$$\beta^* = \frac{S_{xy}}{S_{xx}} \quad \text{är en observation av} \quad \beta^* \in N(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$$

$\alpha^* = \bar{y} - \beta^* \bar{x}$ är även den normalfördelad eftersom den är en linjär funkton av normalfördelningar. Väntevärdet blir

$$\begin{aligned} E(\alpha^*) &= E(\bar{Y}) - \bar{x}E(\beta^*) = E(\frac{1}{n} \sum Y_i) - \bar{x}\beta = \frac{1}{n} \sum (\alpha + \beta x_i) - \bar{x}\beta = \\ &= \frac{1}{n} \sum \alpha + \frac{\beta}{n} \sum x_i - \bar{x}\beta = \alpha + \beta\bar{x} - \bar{x}\beta = \alpha \end{aligned}$$

så även α^* är väntevärdesriktig. Innan vi beräknar dess varians har vi nytta av att $\bar{Y}$ och β^* är oberoende av varandra. Vi visar här att de är okorrelerade, vilket räcker för variansberäkningen. Återigen visar det sig fördelaktigt att uttrycka β^* enligt (10)

$$\begin{aligned} C(\bar{Y}, \beta^*) &= C(\frac{1}{n} \sum Y_i, \sum c_j Y_j) = \frac{1}{n} \sum_i \sum_j c_j C(Y_i, Y_j) = [Y_i \text{ är ober. av } Y_j \text{ då } i \neq j] = \\ &= \frac{1}{n} \sum c_i C(Y_i, Y_i) = \frac{1}{n} \sum c_i V(Y_i) = \frac{\sigma^2}{n} \sum c_i = \frac{\sigma^2}{nS_{xx}} \sum (x_i - \bar{x}) = 0 \end{aligned}$$

där vi återigen känner igen (1) i sista steget. Variansen för α^* blir

$$V(\alpha^*) = V(\bar{Y} - \beta^*\bar{x}) = V(\bar{Y}) + \bar{x}^2 V(\beta^*) - 2\bar{x}C(\bar{Y}, \beta^*) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} + 0$$

dvs

$$\alpha^* = \bar{y} - \beta^*\bar{x} \quad \text{är en observation av} \quad \alpha^* \in N(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}})$$

α^* och β^* är dock inte oberoende av varandra. Kovariansen mellan dem är

$$C(\alpha^*, \beta^*) = C(\bar{Y} - \beta^*\bar{x}, \beta^*) = C(\bar{Y}, \beta^*) - \bar{x}C(\beta^*, \beta^*) = 0 - \bar{x}V(\beta^*) = -\bar{x}\frac{\sigma^2}{S_{xx}}.$$

För variansskattningen och residualkvadratsumman gäller

$$(\sigma^2)^* = s = \frac{1}{n-2}\sum(y_i - \alpha^* - \beta^* x_i)^2 = \frac{Q_0}{f}, \qquad \frac{Q_0}{\sigma^2} \in \chi^2(f)$$

VT 2014

Matematisk statistik
Matematikcentrum
Lunds universitet
Box 118, 221 00 Lund

http://www.maths.lth.se/