SCB STATISTISKA CENTRALBYRÅN

1984-02-01

PROMEMORIOR FRÅN P/STM

NR 12

REGRESSIONSANALYS FÖR F D STATISTIKSTUDERANDE

AV HARRY LÜTJOHANN

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1984:12. Regressionsanalys för f d statistikstuderande / Harry Lütjohann.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1984-12

SCB STATISTISKA CENTRALBYRÅN

1984-02-01

PROMEMORIÖR FRÅN P/STM

NR 12

REGRESSIONSANALYS FÖR F D STATISTIKSTUDERANDE

AV HARRY LÜTJOHANN

REGRESSIONSANALYS FÖR F D STATISTIKSTUDERANDE

Detta är en kurs i regressionsanalys i första hand för handläggare vid Statistiska centralbyrån. Närmare bestämt förutsätts kursdeltagaren ha läst statistik vid universitetet och lärt sig den statistiska inferensens grunder. Däremot behöver han/hon inte ha specialiserat sig på regressionsanalys.

Kursen betonar inte den formella statistiska inferensen inom regressionsanalysen. Kursen uppehåller sig i stället vid ett antal deskriptiva aspekter på regressionsanalysen. Sex kapitel av sju sysslar med regressionsanalysens modellfria, deskriptiva egenskaper. Denna inriktning hos kursen motiveras av två övertygelser hos författaren. Den ena är att det oftast är i de deskriptiva aspekterna på valet av analysansats, som regressionsanalysens användare begår fel och behöver hjälp. Den andra övertygelsen är att traditionell framställning av regressionsanalysen överbetonar de matematiskt eleganta inferensaspekterna på bekostnad av de mindre fascinerande men i praktiken viktigare deskriptiva aspekterna.

Det föreliggande kursmaterialet bör kompletteras med enkla räkneövningsuppgifter, som kan lösas utan dator.

INNEHÅLLSFÖRTECKNING

1 REGRESSIONSKOEFFICIENTER VID ENKEL LINJÄR REGRESSION

1.1 Anpassning av rät linje medelst minstakvadrat-metoden

Betrakta ett vanligt rätvinkligt koordinatsystem i planet, med X på den vågräta axeln och Y på den lodräta. En rät linje i planet kan betecknas på (bland annat) följande sätt:

$$Y = A + BX,$$

där (A,B) är ett par av konstanter, som bestämmer linjens läge i planet.

De båda konstanternas innebörd är följande.

B = tangens för linjens lutningsvinkel,
A = linjens höjd över X-axeln i den punkt där X=0

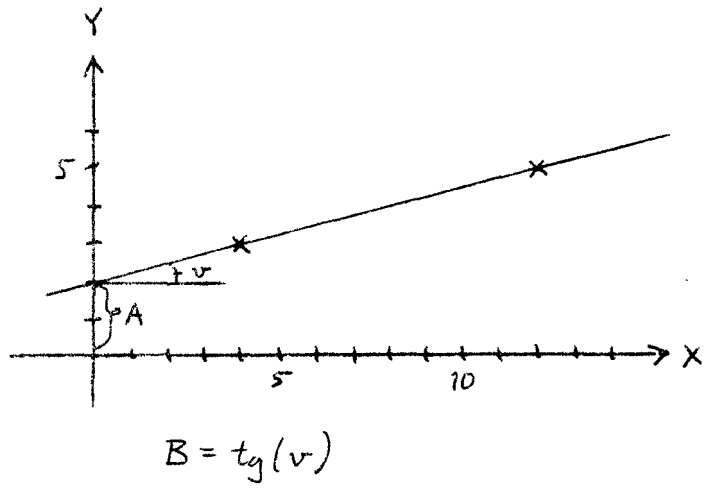
Varje icke lodrätt linje i planet kan anges entydigt medelst konstantparet (A,B).

Exempel 1.1. En rät linje går genom punkterna (X,Y)=(4,3) och (X,Y)=(12,5). Se figur 1.1!

Tangens för lutningsvinkeln är (5-3)/(12-4)=1/4. Där X=0 på linjen är Y=2. Linjen kan således tecknas

$$Y = 2 + (1/4)X.$$

Figur 1.1



Betrakta n observationer (X_i,Y_i), i=1,...,n i två variabler X och Y. Varje observation representeras av en punkt (X_i,Y_i) i koordinatsystemet.

Betrakta två eller flera räta linjer. Vilken linje stämmer bäst med observationerna? Rimligen den där summan av observationernas avstånd från linjen är minst. Men hur ska man mäta avståndet från en punkt till en linje? I vanlig regressionsanalys mäter man avståndet lodrätt, parallellt med Y-axeln.

Betrakta linjen Y=A+BX och punkten (X_i,Y_i). Den punkt på linjen, som ligger lodrätt över eller under (X_i,Y_i), betecknas

$$(X_i,\hat{Y}_i) \qquad \text{där} \qquad \hat{Y}_i = A+BX_i.$$

Avståndet från punkten till linjen är

$$E_i = Y_i-\hat{Y}_i$$

som är positivt om punkten ligger ovanför linjen och negativt om den ligger under.

Tre mått på anpassning ska nu kort diskuteras. Det första är summan av avstånden

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)$$

Detta mått har inget minimum, och det är 0 för flera olika linjer.

Det andra måttet är summan av de absoluta avstånden

$$\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

Detta mått har ett minimum. Men det är matematiskt mindre lätthanterligt än det tredje måttet, som är summan av de kvadrerade avstånden

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Detta är det mått, som användes inom regressionsanalysen.

Exempel 1.2. Det finns tre punkter:

$$(X_1, Y_1) = (4, 3) \ ,$$
$$(X_2, Y_2) = (8, 7) \ ,$$
$$(X_3, Y_3) = (12, 5) \ .$$

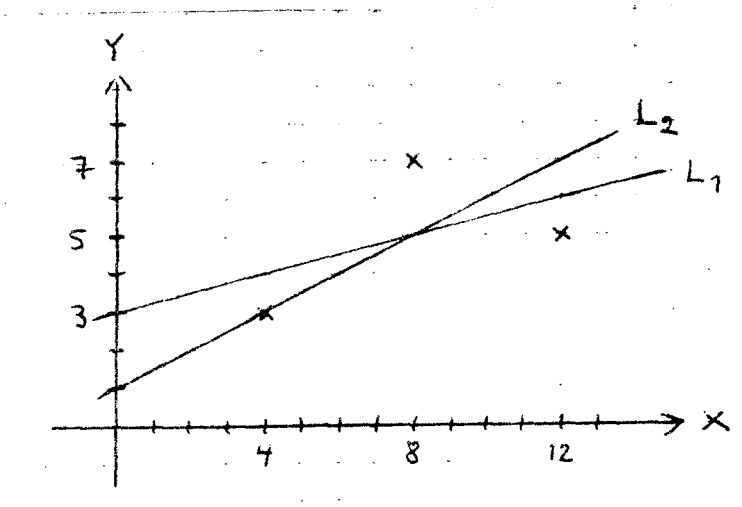
Vilken av följande två linjer passar bäst?

$$L_1: \quad Y = 3 + (1/4)X \ ,$$
$$L_2: \quad Y = 1 + (1/2)X \ .$$

i	X_i	Y_i	$\hat{Y}_i(L_1)$	$E_i(L_1)$	$\hat{Y}_i(L_2)$	$E_i(L_2)$
1	4	3	4	-1	3	0
2	8	7	5	+2	5	+2
3	12	5	6	-1	7	-2

Summan av avstånden är 0 för båda linjerna. Summan av de absoluta avstånden är 4 för båda linjerna. Summan av de kvadrerade avstånden är 6 för L_1 och 8 för L_2. Se också figur 1.2!

Figur 1.2



Enligt både det första och det andra måttet är således L_1 och L_2 likvärdiga. Men enligt det tredje måttet stämmer L_1 bättre med data än L_2 gör. Kanske finns det någon annan linje som enligt det tredje måttet stämmer ännu bättre med data?

Regressionsanalysen tillämpar minstakvadratprincipen, som säger att den linje är bäst, för vilken summan av de kvadrerade lodräta avstånden är minst. (Med "lodrät" menas då egentligen "i Y-riktningen".)

Ur minstakvadratprincipen följer matematiskt att det talpar (A,B), som bestämmer den bästa linjen, kan lösas ut ur ett linjärt ekvationssystem, vars koefficienter är funktioner av data (X_i,Y_i), i=1,...,n. Ekvationssystemet kallas normalekvationerna.

Beteckningar. I fortsättningen betecknar $\bar{X}$ alltid medelvärdet av observationernas X-värden, dvs

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Vidare används Σ utan index och gränser som kort beteckning för summan över i från 1 till n.

Om räta linjen tecknas Y=A+BX, är normalekvationerna

$$n\,A + (\Sigma X)B = \Sigma Y$$
$$(\Sigma X)A + (\Sigma X^2)B = \Sigma XY$$

Normalekvationernas lösning (A,B) kallas regressionskoefficienter.

Man kan visa att följande gäller.

$$B = \Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(X-\bar{X})^2$$
$$A = \bar{Y} - B\bar{X}$$

Exempel 1.3. Samma tre observationer som i exempel 1.2. $\bar{X}=8$, $\bar{Y}=5$.

i	X_i	Y_i	$(X_i-\bar{X})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$
1	4	3	16	+8
2	8	7	0	0
3	12	5	16	0
	24	15	32	+8

$B = 8/32 = 1/4$. $A = \bar{Y}-B\bar{X} = 3$.

Minstakvadratlinjen är $Y = 3 + (1/4)X$. Se figur 1.2!

1.2 Två former för räta linjens ekvation

En rät linje i planet kan betecknas på mer än ett sätt. Låt $\bar{X}$ vara en konstant. (I praktiken väljer man alltid $\bar{X}$ lika med medelvärdet av X-värdena hos n givna observationer (X_i, Y_i), men det som sägs i detta avsnitt 1.2 gäller för godtyckligt $\bar{X}$ och förutsätter inte att det finns några data.)

Givet $\bar{X}$ gäller:

Varje icke lodrät linje i planet kan skrivas

$$Y = C + D(X-\bar{X}) \ ,$$

där konstantparet (C,D) entydigt bestämmer linjens läge i planet. De båda konstanternas innebörd är följande.

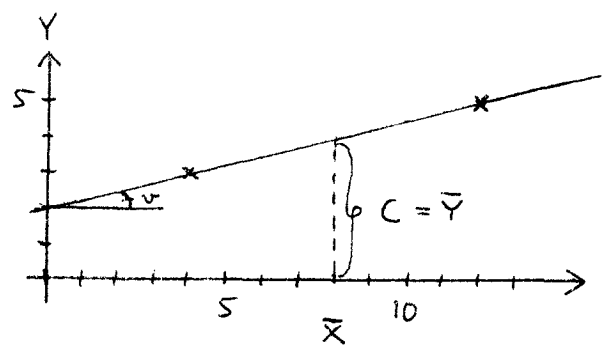
D = tangens för linjens lutningsvinkel,
C = linjens höjd över X-axeln i den punkt där $X=\bar{X}$.

Exempel 1.4. Den räta linjen i exempel 1.1 går genom punkterna (X,Y)=(4,3) och (X,Y)=(12,5). Tangens för lutningsvinkel är 1/4 som i exempel 1.1. Välj $\bar{X}$=8 (= medelvärdet av de nämnda X-värdena). Där $X-\bar{X}$ på linjen är Y=4. Linjen kan således tecknas

$$Y = 4 + (1/4)(X-\bar{X}) \qquad \text{där } \bar{X}=8.$$

Se figur 1.3!

Figur 1.3



$D = B = tg(v)$

Varje rät linje skriven på formen $Y = A+BX$ kan för givet $\bar{X}$ skrivas om

$$Y = (A+B\bar{X}) + B(X-\bar{X}) ,$$

dvs på formen $Y = C+D(X-\bar{X})$, där

$$D = B \quad \text{och}$$
$$C = A + B\bar{X} .$$

Omvänt kan varje rät linje $Y = C+D(X-\bar{X})$ skrivas om

$$Y = (C-D\bar{X}) + DX ,$$

dvs på formen $Y = A+BX$, där

$$B = D \quad \text{och}$$
$$A = C - D\bar{X} .$$

De båda skrivsätten är alltså likvärdiga och är omvändbart entydigt översättbara i varandra för varje valt värde på $\bar{X}$.

1.3 Normalekvationerna i avvikelseform

Betrakta åter n observationer (X_i,Y_i), $i=1,\ldots,n$ med X-medelvärde $\bar{X}$. Minstakvadratmetoden letar upp den räta linje, som stämmer bäst med data i den meningen att summan av de kvadrerade Y-avvikelserna minimeras. Man kan matematiskt bevisa att lösningen är entydig: det finns en och endast en sådan linje. Den kallas regressionslinjen i regressionen av Y på X.

Regressionslinjen kan skrivas $Y = A+BX$ för ett entydigt konstantpar (A,B), som kan beräknas ur normalekvationerna i slutet av avsnitt 1.1 ovan.

Regressionslinjen kan alternativt skrivas $Y = C+D(X-\bar{X})$ för ett entydigt konstantpar (C,D). Man kan kalla detta att skriva regressionslinjen i avvikelseform, eftersom variabeln X här uppträder bara som avvikelse från sitt medelvärde över

observationsmängden. Konstantparet (C,D) kan lösas ut ur ett motsvarande alternativt normalekvationssystem, som ser ut så här:

$$nC = \Sigma Y$$
$$(\Sigma(X-\bar{X})^2)D = \Sigma(X-\bar{X})(Y-\bar{Y})$$

Högerledet i andra normalekvationen kan också skrivas $\Sigma(X-\bar{X})Y$. Detta nya normalekvationssystem kallas normalekvationerna i avvikelseform. Regressionskoefficienterna, som hör till det alternativa skrivsättet, är

$$C = \bar{Y} \quad \text{och}$$
$$D = B \ ,$$

där B är som i slutet av avsnitt 1.1 ovan. Man brukar därför skriva regressionsekvationen i avvikelseform

$$Y = \bar{Y} + B(X-\bar{X}) \ ,$$

där

$$B = \Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(X-\bar{X})^2 \ .$$

Regressionslinjens läge i planet givet $\bar{X}$ bestäms entydigt av koefficientparet $(\bar{Y},B)$, som är två givna funktioner av data.

1.4 Regressionsanalys som asymmetrisk utjämning

Man talar ofta om regressionsanalysen som ett sätt att studera "sambandet mellan variablerna X och Y". Det är då viktigt att notera att regressionsanalysen inte behandlar de båda variablerna symmetriskt. Avståndet mellan punkt och linje mäts lodrätt, parallellt med Y-axeln, och därmed är symmetrin bruten.

Asymmetrin syns också på formeln för regressionskoefficienten B. Om man låter X och Y byta roll och gör "likadant fast tvärtom", så får man inte samma räta linje.

Den regressionslinje, som har härletts här, gäller "regressionen av Y på X". Man kallar X den oberoende variabeln eller (hellre) förklaringsvariabeln. Man kallar Y den beroende variabeln eller (hellre) den analyserade variabeln.

Regressionsanalysens statistiska teori utgår från ungefär följande sätt att resonera.

Till varje värde på variabeln X hör ett (i någon mening) "korrekt" värde på Y. Dessa Y-värden är en linjär funktion av X-värdena så att $Y=A+BX$ för något konstantpar (A,B).

Observationerna, som man har, avspeglar hur Y "beror på" X men är störda på så sätt att de observerade Y-värdena är behäftade med "fel". Däremot är de observerade X-värdena felfria.

För att försöka fånga in den "korrekta" sambandsekvationen ersätter man de observerade Y-värdena med beräknade Y-värden

$$\hat{Y}_i = A+BX_i = \bar{Y} + B(X_i-\bar{X}) \ ,$$

som ligger på den räta linje, som minstakvadratmetoden bestämmer givet de data man har.

De beräknade Y-värdena uppvisar den regelbundenhet (det linjära samband med X), som de observerade Y-värdena "borde" uppvisa men inte gör.

Framställningen här ovan av de teoretiska utgångspunkterna är avsiktligt vag. För att få en statistisk teori måste man precisera sig bättre. Man kan precisera sig olika mycket och bygga in olika starka modellförutsättningar. Ju starkare matematiska antaganden man gör, desto mer omfattande matematisk-statistiska slutsatser kan man dra. Men priset är en ökande risk att antagandena är i praktiken orealistiska.

Regressionskoefficienters medelfel och signifikans är begrepp, som är meningsfulla endast om man gör ganska omfattande modellantaganden. Den här kursen handlar huvudsakligen om sådana egenskaper hos regressionsanalysen, som följer av minstakvadratprincipen och inte kräver några modellantaganden.

2 RESIDUALER OCH KVADRATSUMMOR VID ENKEL LINJÄR REGRESSION

2.1 Uppdelning av observationsvektorn

Betrakta ett givet observationsmaterial (X_i,Y_i), $i=1,\ldots,n$ och den medelst minstakvadratmetoden anpassade räta linjen för regressionen av Y på X. Linjen skrivs i avvikelseform

$$Y = \bar{Y} + B(X-\bar{X})$$

Den går genom punkten $(\bar{X},\bar{Y})$ och har lutningskoefficienten B. (Symbolen Y i ekvationens vänsterled ska inte uppfattas som "värdet av en observation vilken som helst" - dessa värden ligger ju inte på linjen. Symbolen står här för den matematiska variabeln Y i (X,Y)-koordinatsystemet och har inte med data att göra. Beteckningssättet är en aning förvirrande men är fast etablerad praxis.)

De n observerade Y-värdena kan sammanfattas i en observationsvektor

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

Regressionslinjen bestämmer för varje observation ett närmevärde

$$\hat{Y}_i = \bar{Y} + B(X_i-\bar{X})$$

och en residual

$$E_i = Y_i - \hat{Y}_i \quad .$$

Dessa kan sammanfattas i en närmevärdesvektor och en residualvektor

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ . \\ \hat{Y}_n \end{bmatrix} \qquad \mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ . \\ . \\ . \\ E_n \end{bmatrix}$$

Observationsvektorn är summan av närmevärdesvektorn och residualvektorn

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{E}$$

Regressionsanalysen bestämmer uppdelningen av observationerna i två delar så att delarna har följande tre egenskaper. Lägg märke till att egenskaperna inte gäller den enskilda observationen utan hela vektorn.

(1) Närmevärdenas medelvärde är lika med de observerade Y-värdenas medelvärde.

$$\bar{\hat{Y}} = \bar{Y} + B(\bar{X}-\bar{X}) = \bar{Y}.$$

(2) Residualernas medelvärde är noll.

$$\bar{E} = \bar{Y}-\bar{\hat{Y}} = 0.$$

Dessa båda egenskaper följer direkt ur varandra.

(3) Korrelationen mellan förklaringsvariabeln X och residualen E är noll. Korrelationskoefficientens täljare är

$$\Sigma(X_i-\bar{X})(E_i-\bar{E}) = \Sigma(X_i-\bar{X})E_i = \Sigma(X_i-\bar{X})(Y_i-\bar{Y}-B(X_i-\bar{X})) = 0;$$

sista ledet följer ur normalekvationen för B.

Dessa egenskaper hos närmevärden och residualer följer matematiskt ur minstakvadratmetoden. De beror inte på några antaganden därutöver. Omvändningen gäller också. Residualer som uppfyller (2) och (3) implicerar regressionskoefficienter som satisfierar minstakvadratmetodens normalekvationer.

Beteckningar. I fortsättningen betecknar A, B, $\hat{Y}$ och E endast regressionskoefficienter, närmevärden och residualer beräknade enligt minstakvadratmetoden, inte analoga kvantiteter för andra linjer än regressionslinjen.

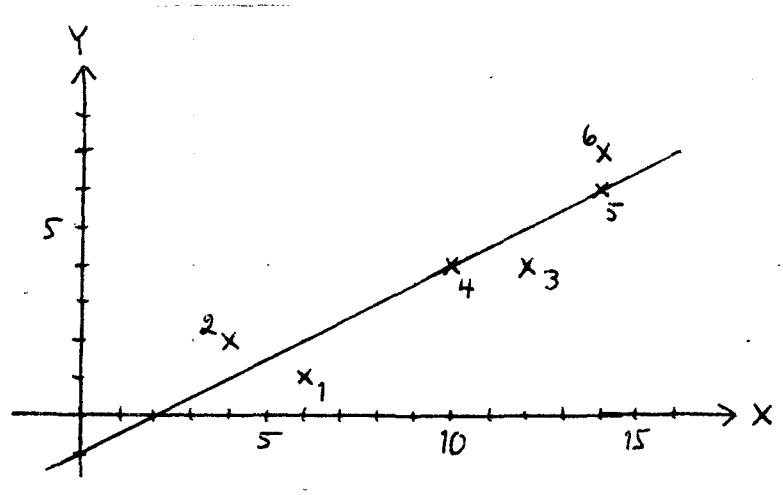
Exempel 2.1. Regressionen av Y på X beräknas för de sex observationerna i följande tabell.

i	X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$	$\hat{Y}$	E	$(X-\bar{X})E$
1	6	1	-4	-3	16	+12	2	-1	+4
2	4	2	-6	-2	36	+12	1	+1	-6
3	12	4	+2	0	4	0	5	-1	-2
4	10	4	0	0	0	0	4	0	0
5	14	6	+4	+2	16	+8	6	0	0
6	14	7	+4	+3	16	+12	6	+1	+4
Σ	60	24	0	0	88	+44	24	0	0

$\bar{X} = 10, \qquad \bar{Y} = 4, \qquad B = +44/88 = 1/2.$

Regressionslinjen är $Y = 4+\frac{1}{2}(X-10)$ som också kan skrivas $Y = -1+\frac{1}{2}X$. Minstakvadratmetodens tre egenskaper verifieras. Se också figur 2.1!

Figur 2.1



2.2 Uppdelning av avvikelsekvadratsumman

Regressionen av Y på X kan betraktas också på följande sätt. Först har man ett observationsmaterial Y_i, i=1,...,n. Variationen mellan observationerna mäts av summan av de kvadrerade avvikelserna från medelvärdet

$$SST = \Sigma(Y_i-\bar{Y})^2$$

där SST betyder "sum of squares, total".

Sedan inför man en förklaringsvariabel X och beräknar regressionen av Y på X. Då får man närmevärden och residualer. Variationen mellan närmevärdena $\hat{Y}_i$ mäts av summan av de kvadrerade avvikelserna från medelvärdet

$$SSR = \Sigma(\hat{Y}_i-\bar{Y})^2$$

där SSR står för "sum of squares, regression". På svenska brukar man kalla SSR den förklarade kvadratsumman eller regressionskvadratsumman.

Den efter regressionsanalysen återstående oförklarade variationen mäts av residualernas (avvikelse-)kvadratsumma

$$SSE = \Sigma E_i^2$$

där bokstaven E syftar på beteckningen E_i. Denna kvadratsumma heter residualkvadratsumman.

Ur minstakvadratmetoden följer matematiskt att

$$SST = SSR + SSE.$$

Regressionsanalysen delar således den totala Y-variationen i två delar, den förklarade kvadratsumman och residualkvadratsumman.

Exempel 2.2. De tre kvadratsummorna beräknas för data i exempel 2.1.

i	$Y-\bar{Y}$	$\hat{Y}-\bar{Y}$	E	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	E^2
1	-3	-2	-1	9	4	1
2	-2	-3	+1	4	9	1
3	0	+1	-1	0	1	1
4	0	0	0	0	0	0
5	+2	+2	0	4	4	0
6	+3	+2	+1	9	4	1
Σ	0	0	0	26	22	4

SST = 26, SSR = 22, SSE = 4.

Kvadratsummeuppdelningen SST = SSR + SSE verifieras.

På ett mer abstrakt matematiskt språk kan kvadratsummeuppdelningen karakteriseras som följer. Beskrivningen torde vara intuitivt tilltalande även för den som inte i grunden förstår den.

Observationsvektorn $\mathbf{Y}$ delas först upp i en medelvärdesvektor $\mathbf{M}$, vars element alla är lika med Y-medelvärdet $\bar{Y}$, och en observationsvektor i avvikelseform $\mathbf{y}$.

Observationsvektorn i avvikelseform $\mathbf{y}$ delas av regressionsanalysen upp i en närmevärdesvektor i avvikelseform $\hat{\mathbf{y}}$ och residualvektorn $\mathbf{E}$, som redan är i avvikelseform, och som vi därför här alternativt betecknar $\mathbf{e}$.

$$\mathbf{Y} = \mathbf{M} + \mathbf{y}, \qquad \mathbf{y} = \hat{\mathbf{y}} + \mathbf{e};$$
$$\mathbf{Y} = \mathbf{M} + \hat{\mathbf{y}} + \mathbf{e}.$$

Närmevärdesvektorn i avvikelseform och residualvektorn är p g a minstakvadratmetoden inbördes ortogonala (och p g a definitionen av ett medelvärde ortogonala mot medelvärdesvektorn). Pythagoras' sats är tillämplig. Den ger kvadratsummeuppdelningen SST = SSR + SSE.

2.3 Determinationskoefficienten

Betrakta ett givet bivariat observationsmaterial och den medelst minstakvadratmetoden anpassade regressionslinjen Y på X. Denna linje är den som stämmer bäst med data. Hur bra stämmer den? Eller omvänt, hur bra stämmer data med en rät linje? Hur nära linjär är punktsvärmen?

Regressionen delar den totala kvadratsumman i två delar, SST = SSR+SSE. Graden av anpassning till en rät linje mäts i regressionsanalysen av den andel av SST, som den förklarade kvadratsumman SSR utgör. Denna kvot kallas determinationskoefficienten och tecknas R^2. Man brukar skriva den som ett minus den andel av variationen, som residualkvadratsumman SSE utgör.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\Sigma E_i^2}{\Sigma (Y_i - \bar{Y})^2}$$

Determinationskoefficienten ligger mellan 0 och 1 och kan antaga båda dessa värden. $R^2=1$ om och endast om punkterna ligger exakt på en (icke lodrät) rät linje.

Exempel 2.3. Samma data som i exempel 2.1 och 2.2.

$$R^2 = 1 - \frac{4}{26} = 0{,}846$$

Jämför åter figur 2.1!

2.4 Regressionsanalys som förklaring av skillnader

Regressionsanalys används ofta för att förklara en variabel Y i termer av en annan variabel X. I vad mån och i vilken mening en variabel egentligen kan förklara en annan variabel är en svår fråga, som vi tillsvidare förbigår. Men givet att en sådan förklaring alls är möjlig och meningsfull, så är följande resonemang värt att beakta.

Regressionsanalysen förklarar inte nivån hos den analyserade variabeln Y. Regressionsekvationen

$$\hat{Y} = \bar{Y}+B(X-\bar{X})$$

säger om genomsnittsnivån hos närmevärdena $\hat{Y}$ bara att den är som den är därför att genomsnittsnivån hos observationerna Y är som den är. Detta är en förklaring av Y i termer av Y och föga upplysande.

Regressionsanalysen förklarar däremot skillnaderna i Y-värde mellan ett godtyckligt par av observationer. Två observationers närmevärden $\hat{Y}_i$ och $\hat{Y}_j$ skiljer sig med en kvantitet

$$\hat{Y}_i-\hat{Y}_j = B(X_i-X_j)$$

som är en för alla par gemensam faktor B gånger skillnaden i X-värden. Detta är en förklaring av skillnader i analysvariabeln Y i termer av skillnader i förklaringsvariabeln X och är inte trivialt.

Regressionsanalysen är således en förklaring av skillnader mellan observationer. Det är följaktligen också en förklaring av variationen mellan observationerna.

Observationerna antages vara störda så att de observerade värdena Y_i avviker från den räta linje de egentligen skulle ligga på. Regressionsanalysen räknar approximativt bort störningen och får kvar närmevärdena $\hat{Y}_i$. Variationen mellan dessa, den förklarade kvadratsumman SSR, är den del av totalvariationen SST som kan tillskrivas förklaringsvariabeln. Relationen

$$SSR = \Sigma(\hat{Y}-\bar{\hat{Y}})^2 = B^2\Sigma(X-\bar{X})^2$$

förklarar (den utjämnade) variationen hos analysvariabeln Y i termer av variationen hos förklaringsvariabeln X. Förklaringsvärdet är andelen R^2 av hela variationen hos analysvariabeln Y.

2.X Bevis (överkurs)

Bevis att normalekvationernas lösning minimerar avvikelsekvadratsumman

Residualen är $E = Y-\bar{Y}-B(X-\bar{X})$.

Avvikelsen från en godtycklig annan rät linje kan tecknas

$$U = Y-(\bar{Y}+\Delta\bar{Y}) - (B+\Delta B)(Y-\bar{Y})$$

där $(\Delta\bar{Y},\Delta B) \neq (0,0)$.

Avvikelsekvadratsumman är

$$\begin{aligned}
\Sigma U^2 &= \Sigma\left[Y-(\bar{Y}+\Delta\bar{Y}) - (B+\Delta B)(X-\bar{X})\right]^2 = \\
&= \Sigma\left[E - \Delta\bar{Y} - \Delta B(X-\bar{X})\right]^2 = \\
&= \Sigma E^2 + n(\Delta\bar{Y})^2 + (\Delta B)^2\Sigma(X-\bar{X})^2 - \\
&\quad - 2(\Delta\bar{Y})\Sigma E - 2(\Delta B)\Sigma(X-\bar{X})E + \\
&\quad + 2(\Delta\bar{Y})(\Delta B)\Sigma(X-\bar{X}) = \\
&= \Sigma E^2 + \\
&\quad + n(\Delta\bar{Y})^2 + (\Delta B)^2\,\Sigma(X-\bar{X})^2 .
\end{aligned}$$

De tre dubbelprodukterna är 0 p g a (2) och (3) i avsnitt 2.1 och definitionen av en avvikelse från ett medelvärde.

Vi ser att

$$\Sigma U^2 \geq \Sigma E^2$$

med likhet om och endast om $\Delta\bar{Y} = \Delta B = 0$.

(Fallet $\Sigma(X-\bar{X})^2 = 0$ är ett urartningsfall, då regressionsanalysen inte fungerar normalt. Alla X-värden är i detta fall lika.)

3 FÖRUTSÄTTNINGSNIVÅER OCH TOLKNINGSNIVÅER

3.1 Tre förutsättningsnivåer vid regressionsanalys

En i sak meningsfull regressionsanalys består inte av data, beräkningsförfaranden och beräkningsresultat. Där måste också ingå förutsättningar, som handlar om datas innebörd. Utan sådana förutsättningar är beräkningsresultaten inte tolkbara, och hela analysen utartar till ren formel- och sifferexercis utan sakinnehåll.

Man kan särskilja tre olika nivåer av förutsättningar för regressionsanalys. Den första nivån innehåller ett absolut minimum av förutsättningar och medger ett absolut minimum av tolkning av analysens resultat. Den andra nivån lägger till fler förutsättningar och medger mer innehållsrik resultattolkning. Den tredje nivån innehåller ännu mer förutsättningar och tolkningsmöjligheter.

Den första förutsättningsnivån kan kallas minimiförutsättningarna. Den andra förutsättningsnivån består därutöver av statistiska förutsättningar, som uttrycks i termer av matematik och sannolikhetslära. Den tredje förutsättningsnivån lägger till kausalförutsättningar (dvs förutsättningar om orsak och verkan) eller eventuellt andra saklogiska förutsättningar. Dessa uttrycks i termer hämtade från en teori om de fenomen, som studeras med hjälp av de data som analyseras.

Man kan tänka sig en analys av data som utesluter de statistiska förutsättningarna, men i så fall inte en regressionsanalys. Den andra förutsättningsnivån innehåller nämligen de förutsättningar, som är specifika för regressionsanalysen.

Exempel 3.1. Följande data föreligger.

i	X	Y	i	X	Y
1	10:00	3:20	6	10:00	2:00
2	12:00	2:85	7	7:50	2:10
3	7:50	2:50	8	7:50	6:05
4	10:00	4:90	9	12:50	3:30
5	8:00	1:60	10	5:00	0:00

Observationerna är tio utlottade barn i en förstaklass i Z-skolan. Variabeln X uppges mäta veckopengen och variabeln Y utgifterna för godis, i båda fallen för vecka V år Å.

(1) Minimiförutsättningarna. Data är korrekta. Alla som har sysslat med statistisk datainsamling vet att detta inte är självklart.

(2) Statistiska förutsättningar. Värdet Y_i kan betraktas som utfallet av en slumpvariabel, vars väntevärde är en linjär funktion av värdet X_i, och alla tio barnen följer en och samma linjära funktion.

(3) Kausalförutsättningar. Det finns ett orsakssamband mellan variablerna. Veckopengens belopp är orsaken och godisutgifterna är verkan. Om X ändras, så ändras (väntevärdet för) Y enligt den gemensamma räta linjen.

(1) Minimiförutsättningarna säger att data är vad det utges för att vara. Mätfel förekommer inte, utom om de kan räknas in under de slumpfel, som nästa förutsättningsnivå till stor del handlar om.

(2) De statistiska förutsättningarna är av teknisk natur. Vi återkommer till dem.

(3) De kausala förutsättningarna säger att varje observation återspeglar ett för alla observationerna gemensamt orsakssamband, där förklaringsvariabeln X är eller mäter orsaken och analysvariabeln Y är eller mäter verkan. Det behöver inte vara så att

utredaren själv kan välja eller ens påverka X. Men hur än X bestäms, så påverkar X i sin tur Y, och det är denna påverkan, som regressionsanalysen formulerar, uppskattar och tolkar.

Regressionsanalysens statistiska förutsättningar har utvecklats av forskare som har använt regressionsanalys inom så olika forskningsområden som astronomi (mätfelsteori) och växtförädling. De är mest naturliga i situationer där man kan experimentera med X för att aktivt studera hur Y påverkas. Men de används också mer allmänt.

De statistiska förutsättningarna kan vara mer eller mindre omfattande. Ett minimum är följande.

För varje observation "i" gäller att Y_i är utfallet av en slumpvariabel, som också brukar betecknas Y_i (en förvirrande men väl etablerad praxis). Alla andra kvantiteter är eller kan betraktas som icke-stokastiska konstanter. Slumpvariabeln Y, har ett väntevärde, och för alla "i" gäller den statistiska modellen

$$E(Y_i) = \alpha + \beta X_i$$

vars parametrar (α, β) är givna okända konstanter som är gemensamma för samtliga observationer "i".

Vidare gäller att varje Y_i har ändlig varians. Slutligen gäller att den mot observationsvektorn **Y** svarande vektorn av slumpvariabler

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}$$

har en n-dimensionell sannolikhetsfördelning utan matematiska konstigheter.

Ur dessa minimala statistiska förutsättningar följer att regressionskoefficienterna

$$(A,B) \text{ eller } (\bar{Y},B)$$

är ett par av slumpvariabler, vars väntevärden är

$$E(B) = \beta,$$
$$E(A) = \alpha \quad \text{eller} \quad E(\bar{Y}) = \alpha + \beta\bar{X}.$$

Koefficientparet (A,B) - eller $(\bar{Y},B)$ - har en bivariat sannolikhetsfördelning. Ju mer man förutsätter om fördelningen för **Y**, desto mer följer om deras fördelning.

Det är brukligt att göra sådana statistiska antaganden utöver de minimala att man ur data kan uppskatta variansen för var och en av regressionskoefficienterna. Man kan då för det första uppskatta parametern β medelst punktestimatorn

$$B = \text{est}(\beta),$$

och för det andra också uppskatta denna skattnings osäkerhet medelst en annan estimator, låt säga

$$\text{est}(\text{var}(B)),$$

som kan beräknas ur data.

En mycket viktig fråga är om de förutsättningar, som man gör, gäller enbart för de observationer "i", som man har, eller mer generellt. I det senare fallet, hur generellt? Vi uppskjuter diskussionen av denna i praktiken helt avgörande fråga till avsnitt 4.4.

3.2 Tre motsvarande tolkningsnivåer

Statistiska data och beräknade kvantiteter talar inte för sig själva. De måste tolkas. Hur innehållsrikt de kan tolkas beror på hur starka förutsättningar man vågar göra. Enbart data ger

inga tolkningar, men data plus förutsättningar utgör tillsammans tillräckligt logisk grund för tolkning. Tolkningens riktighet står och faller med förutsättningarnas riktighet.

Mot var och en av de tre förutsättningsnivåerna i föregående avsnitt svarar en tolkningsnivå, som anger hur innehållsrika tolkningar man har (givit sig) skäl att göra. De tre tolkningsnivåerna kan kallas den deskriptiva, den statistiska och den kausala eller saklogiska.

(1) Minimiförutsättningarna medger deskriptiv tolkning. Regressionslinjen och determinationskoefficienten beskriver det givna observationsmaterialet; arten och graden av samband hos de givna data. Man har inte fog för några tolkningar som går utöver de givna data, varken till andra data eller till några "varför".

Den deskriptiva tolkningsnivån kan förefalla torftig. Men den är ändå inte alltid ointressant. Den här kursen handlar till mycket stor del om tolkningar på denna blygsamma nivå.

(2) Statistiska förutsättningar motiverar statistiska tolkningar. De beräknade regressionskoefficienterna är estimat av motsvarande parametrar i modellen. Estimatens osäkerhet kan (oftast) också beräknas, så att estimaten kan kompletteras med medelfel. Om de statistiska förutsättningarna är tillräckligt omfattande, finns det också grund för sådana statistiska förfaranden som intervallestimation och hypotesprövning. Man kan till exempel ange ett konfidensintervall för parametern β eller testa hypotesen att $\beta = 0$. Sådan hypotesprövning kallas ibland signifikanstestning; skiljer sig det beräknade B signifikant från noll?

(3) Kausalförutsättningar motiverar kausaltolkning, orsakstolkning. För det mesta gör man samtidigt både statistiska och kausala förutsättningar. De statistiska förutsättningarna är av formell natur. De handlar om en modell, som man väljer att ansätta. Men de kausala förutsättningarna handlar om verkligheten själv. De användes för att underbygga en slutsats, som också gäller verkligheten själv.

Antag att regressionslinjen vid regression av Y på X blir

$$Y = 10 + 0{,}3X, \qquad R^2 = 0{,}92,$$

och att koefficienten B = 0,3 är signifikant skild från noll. En omfattande orsakstolkning skulle då vara följande.

(a) Det finns ett linjärt orsakssamband mellan X och Y, vilket svarar för 92 % av variationen i Y i detta material.

(b) Förklaringsvariabeln X påverkar Y så att en enhets ändring av X förorsakar att Y ändras med 0,3 enheter.

Det är mycket viktigt att komma ihåg att orsakstolkningarna på den tredje tolkningsnivån inte följer matematiskt-logiskt ur enbart de matematisk-statistiska förutsättningarna på den andra tolkningsnivån.

Exempel 3.2. Samma data som i exempel 3.1

Regressionsekvationen och determinationskoefficienten beräknas. Se figur 3.1!

Data och regressionslinjen ritas på millimeterpapper. Se figur 3.2!

Figur 3.1. Räkneschema till exempel 3.1-3.2.

i	X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$\hat{Y}$	E	E^2	$(y-\bar{y})^2$
1	10,00	3,20	+1,00	+0,35	+0,350	1,00	3,11	+0,09	0,0081	0,1225
2	12,00	2,85	+3,00	0,00	0,000	9,00	3,64	−0,79	0,6241	0,0000
3	7,50	2,50	−1,50	−0,35	+0,525	2,25	2,46	+0,04	0,0016	0,1225
4	10,00	4,90	+1,00	+2,05	+2,050	1,00	3,11	+1,79	3,2041	4,2025
5	8,00	1,60	−1,00	−1,25	+1,250	1,00	2,59	−0,99	0,9801	1,5625
6	10,00	2,00	+1,00	−0,85	−0,850	1,00	3,11	−1,11	1,2321	0,7225
7	7,50	2,10	−1,50	−0,75	+1,125	2,25	2,46	−0,36	0,1296	0,5625
8	7,50	6,05	−1,50	+3,20	−4,800	2,25	2,46	+3,59	12,8881	10,2400
9	12,50	3,30	+3,50	+0,45	+1,575	12,25	3,77	−0,47	0,2209	0,2025
10	5,00	0,00	−4,00	−2,85	+11,400	16,00	1,80	−1,80	3,2400	8,1225
Σ	90,00	28,50	0,00	0,00	+12,625	48,00	28,51	−0,01	22,5229	25,8600
m	9,00	2,85								

$$B = \frac{+12,625}{48} = +0,263$$

$$A = 2,85 - 0,263 \cdot 9 = 0,483$$

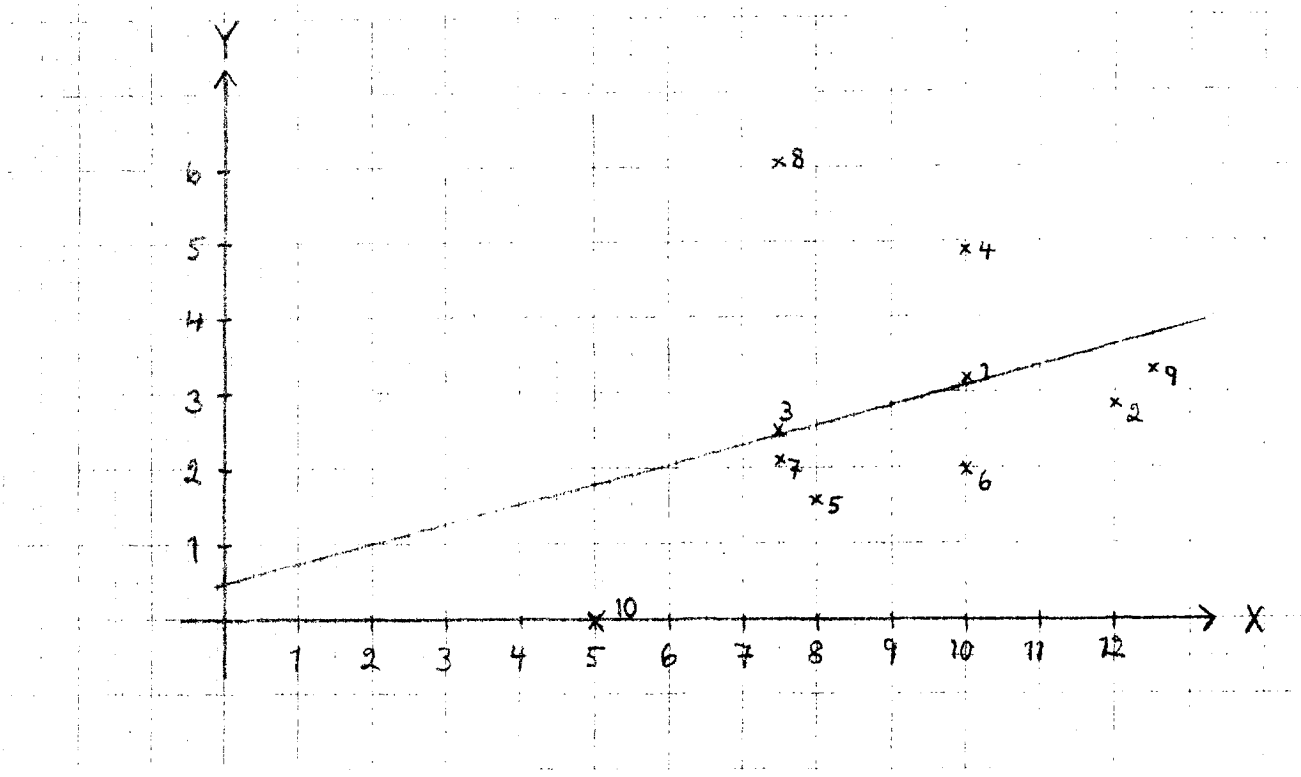
$$\hat{Y} = 2,85 + 0,263(X - 9)$$

$$\hat{Y} = 0,483 + 0,263X$$

$$R^2 = 1 - \frac{22,529}{25,860} = 0,13$$

Figur 3.2

Data och regressionslinje till exempel 3.1-3.2.



(1) Deskriptiv tolkning.

Regressionslinjen är

$$Y = 0{,}483 + 0{,}263X.$$

Determinationskoefficienten är 0,13.
Närmevärden och residualer är som anges i figurerna 3.1 och 3.2.
Kommentar: dålig anpassning.

(2) Statistisk tolkning.

Modellekvationen

$$E(Y) = \alpha + \beta X$$

uppskattas till $Y = 0{,}483 + 0{,}263X$. Slumpvariationen kring linjen är betydande.

(Man kan testa hypotesen H:$\beta = 0$. Det visar sig att regressionskoefficienten B inte är signifikant skild från noll.)

(3) Orsakstolkning.

I: Utan signifikansprövning.

Veckopengen X påverkar godisinköpen Y. En ökning av X med 1 krona medför en genomsnittlig ökning av Y med drygt 26 öre.

II: Med signifikansprövning.

Orsakssamband kan inte påvisas råda mellan veckopeng och godisinköp.

De tre tolkningarna är innehållsligt mycket olika. Som exempel 3.2 visar är dessutom varken den statistiska tolkningen eller orsakstolkningen nödvändigtvis entydig. Tolkningen beror, förutom på tolkningsnivån, på vilka frågor man väljer att ta upp.

Varje tolkningsnivå medger vissa frågor, som kan ställas till data. Man behöver inte ställa alla frågorna. Utan frågor är data stumma. De talar inte för sig själva.

3.3 Tre grader av statistiska förutsättningar

Det finns statistiska standardförutsättningar för linjär regressionsanalys. Dessa kan delas upp i tre delar, som här kommer att kallas "grader". De minimala statistiska förutsättningarna i avsnitt 3.1 ovan omfattar första graden plus litegrann från andra graden.

(1) Första graden av de statistiska standardförutsättningarna handlar om väntevärden.

För varje observation "i" gäller att analysvariabeln Y_i är en slumpvariabel vars väntevärde är en linjär funktion av förklaringsvariabeln X_i,

$$E(Y_i) = \alpha + \beta X_i, \qquad i = 1,.....,n.$$

Parametrarna (α, β) är gemensamma för samtliga observationer.

Man brukar skriva förstagradsförutsättningarna i termer av en slumpvariabel ε_i som är skillnaden mellan faktiskt och förväntat värde på slumpvariabeln Y_i. Förstagradsförutsättningarna lyder då:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad i = 1,....,n.$$

Vi använder i fortsättningen detta skrivsätt.

(2) Andra graden av de statistiska standardförutsättningarna handlar om varianser och kovarianser.

För varje observationer "i" gäller att variansen för analysvariabeln Y_i:s avvikelse från det av X_i bestämda väntevärdet är en (okänd) konstant σ^2.

$$Var(\varepsilon_i) = \sigma^2, \qquad i = 1,.....,n.$$

För varje par av två observationer "i" och "j" gäller att kovariansen mellan Y_i:s och Y_j:s avvikelser från väntevärdena är noll.

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j, \quad i = 1,.....,n, \quad j = 1,...,n.$$

Första och andra graden av standardförutsättningarna medger statistiska slutsatser om väntevärden, varianser och kovarians för regressionskoefficienterna (A,B) - eller ($\bar{Y}$,B).

(3) Tredje graden av de statistiska standardförutsättningarna handlar om fördelningar.

Vektorn av slumpvariabler

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

är multivariat normalfördelad med den väntevärdesvektor och den kovariansmatris, som första och andra graden anger.

Första, andra och tredje graden tillsammans medger statistisk inferens sådan som intervallestimation och hypotesprövning gällande den ansatta statistiska modellens tre parametrar $(\alpha, \beta, \sigma^2)$.

3.4 Utskrift från standardprogrammet SAS

Räknearbete vid regressionsanalys överlåtes numera oftast på en dator. Det finns många standardprogram för regressionsanalys. Den här kursen ger viss vägledning i konsten att läsa utskriften från ett sådant program. Den försöker däremot inte lära ut hur man gör då man kör.

Standardprogrammet ingår i standardprogrampaketet SAS (Statistical Analysis System), som finns vid SCBs datamaskincentral. Programmet heter GLM (General Linear Models) och är mycket flexibelt.

Standardprogrammet SAS-GLM tillämpas på samma data som i exempel 3.1. Resultatet blir detsamma som i exempel 3.2 men med fler decimaler. Figur 3.3 återger de inlästa data.

Figur 3.4 är resultatutskriften. Vissa delar av denna kräver för sin förklaring mer än som hittills anförts i den här kursen.

Figur 3.5 är resultatutskriften kompletterad med sex "ramar".

Ram 1 innehåller analysvariabelns kvadratsummas uppdelning i en "förklarad" (model) och en residual- (error) kvadratsumma.

Ram 2 innehåller determinationskoefficienten.

Ram 3 innehåller regressionskoefficienterna.

Ram 4 innehåller regressionskoefficienternas estimerade medelfel givet statistiska förutsättningar av vad som här har kallats första och andra graden.

Ram 5 innehåller t-värden för test av hypotesen att α respektive β är noll. Dessa test kräver statistiska förutsättningar av första, andra och tredje graden.

Ram 6 innehåller observationsvärden, närmevärden (predicted) och residualer.

Figur 3.3 Inläsning till SAS.

```
1                          S T A T I S T I C A L   A N A L Y S I S   S Y S T E M   8:51 THURSDAY, DECEMBER 3, 1981

NOTE: THE JOB ZSSTMTHP HAS BEEN RUN UNDER RELEASE 79.5 OF SAS
      AT SWEDISH NATIONAL BUREAU OF STATISTICS (00551).

NOTE: SAS OPTIONS SPECIFIED ARE:
      NOOVP


1             DATA;
2             INPUT X Y;
3             CARDS;

NOTE: DATA SET WORK.DATA1 HAS 10 OBSERVATIONS AND 2 VARIABLES. 953 OBS/TRK.
NOTE: THE DATA STATEMENT USED 0.16 SECONDS AND 160K.

14            TITLE ENKEL LINJAR REGRESSION;
15            PROC PRINT;
16            VAR X Y;

NOTE: THE PROCEDURE PRINT USED 0.30 SECONDS AND 160K AND PRINTED PAGE 1.

17            PROC GLM;
18            MODEL Y=X/ P SS2;

NOTE: THE PROCEDURE GLM USED 0.44 SECONDS AND 182K AND PRINTED PAGE 2.

NOTE: SAS USED 182K MEMORY.

NOTE: SAS INSTITUTE INC.
      SAS CIRCLE
      BOX 8000
      CARY, N.C. 27511
```

```
ENKEL LINJAR REGRESSION   8:51 THURSDAY, DECEMBER 3, 1981   1

 OBS      X        Y

   1     10.0     3.20
   2     12.0     2.85
   3      7.5     2.50
   4     10.0     4.90
   5      8.0     1.80
   6     10.0     2.00
   7      7.5     2.10
   8      7.5     6.05
   9     12.5     3.30
  10      5.0     0.00
```

Figur 3.4.

ENKEL LINJAR REGRESSION 8:51 THURSDAY, DECEMBER 3, 1981 2

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	3.32063802	3.32063802	1.18	0.3093	0.128408	58.8953
ERROR	8	22.53936198	2.81742025		STD DEV		Y MEAN
CORRECTED TOTAL	9	25.86000000			1.67851728		2.85000000

SOURCE	DF	TYPE II SS	F VALUE	PR > F
X	1	3.32063802	1.18	0.3093

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > ITI	STD ERROR OF ESTIMATE
INTERCEPT	0.48281250	0.22	0.8350	2.24413429
X	0.26302083	1.09	0.3093	0.24227310

OBSERVATION	OBSERVED VALUE	PREDICTED VALUE	RESIDUAL
1	3.20000000	3.11302083	0.08697917
2	2.85000000	3.63906250	-0.78906250
3	2.50000000	2.45546875	0.04453125
4	4.90000000	3.11302083	1.78697917
5	1.60000000	2.58697917	-0.98697917
6	2.00000000	3.11302083	-1.11302083
7	2.10000000	2.45546875	-0.35546875
8	6.05000000	2.45546875	3.59453125
9	3.30000000	3.77057292	-0.47057292
10	0.00000000	1.79791667	-1.79791667

SUM OF RESIDUALS	0.00000000
SUM OF SQUARED RESIDUALS	22.53936198
SUM OF SQUARED RESIDUALS - ERROR SS	-0.00000000
FIRST ORDER AUTOCORRELATION	-0.10723080
DURBIN-WATSON D	2.07070999

Figur 3.5.

ENKEL LINJAR REGRESSION 8:51 THURSDAY, DECEMBER 3, 1981 2

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

(1)

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	1	3.32063802	3.32063802	1.18	0.3093
ERROR	8	22.53936198	2.81742025		STD DEV
CORRECTED TOTAL	9	25.86000000			1.67851728

(2)

R-SQUARE	C.V.
0.128408	58.8953

				Y MEAN
				2.85000000

SOURCE	DF	TYPE II SS	F VALUE	PR > F
X	1	3.32063802	1.18	0.3093

(3) (5) (4)

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > ITI	STD ERROR OF ESTIMATE
INTERCEPT	0.48281250	0.22	0.8350	2.24413429
X	0.26302083	1.09	0.3093	0.24227310

(6)

OBSERVATION	OBSERVED VALUE	PREDICTED VALUE	RESIDUAL
1	3.20000000	3.11302083	0.08697917
2	2.85000000	3.63906250	-0.78906250
3	2.50000000	2.45546875	0.04453125
4	4.90000000	3.11302083	1.78697917
5	1.60000000	2.58697917	-0.98697917
6	2.00000000	3.11302083	-1.11302083
7	2.10000000	2.45546875	-0.35546875
8	6.05000000	2.45546875	3.59453125
9	3.30000000	3.77057292	-0.47057292
10	0.00000000	1.79791667	-1.79791667

SUM OF RESIDUALS	0.00000000
SUM OF SQUARED RESIDUALS	22.53936198
SUM OF SQUARED RESIDUALS - ERROR SS	-0.00000000
FIRST ORDER AUTOCORRELATION	-0.10723080
DURBIN-WATSON D	2.07070999

4 ENKEL ICKE-LINJÄR REGRESSIONSANALYS

4.1 Direkt anpassning enligt minstakvadratprincipen

Regressionsanalys användes för att studera sambandet $Y = f(X)$ mellan två variabler Y och X. Med minstakvadratmetodens hjälp kan man uppskatta detta samband om det är linjärt. Om det inte är linjärt, går minstakvadratmetoden inte att använda utan en del extra om och men. Man kan gå tillväga på två olika sätt.

Antingen tillämpar man minstakvadratprincipen direkt på det icke-linjära sambandet. Man minimerar avvikelsekvadratsumman

$$\Sigma\left[Y_i - f(X_i)\right]^2$$

Då funktionen $Y=f(X)$ inte är linjär, leder denna ansats inte till någon enkel matematisk lösning. Den kan likväl tillämpas med datorns hjälp. Den kan kallas direkt anpassning enligt minstakvadratprincipen.

Eller också finner man på en transformation av funktionen $Y=f(X)$, som leder till att någon funktion $U=h(Y)$ av analysvariabeln är en linjär funktion

$$U = C+DZ$$

av någon funktion $Z=g(X)$ av förklaringsvariabeln. Man tillämpar minstakvadratmetoden på det linjära sambandet mellan U och Z. Sedan översätter man resultatet tillbaka till i termer av det ursprungliga icke-linjära sambandet mellan Y och X. Detta kan kallas indirekt anpassning medelst minstakvadratmetoden efter lineariserande transformation.

Indirekt anpassning är inte alltid möjlig, eftersom det inte alltid finns någon lineariserande transformation.

Exempel 4.1. Man önskar anpassa en kurva av formen

$$Y = e^{\frac{A}{1-BX}}$$

till n observationer (X_i, Y_i), i=1,...n.

Direkt anpassning innebär att man minimerar avvikelsekvadratsumman

$$\sum_{i=1}^{n} \left[Y_i - e^{\frac{A}{1-BX_i}} \right]^2$$

med avseende på konstanterna A och B.

Indirekt anpassning innebär att man först transformerar den ansatta funktionen till den matematiskt likvärdiga

$$\frac{1}{\log Y} = \frac{1}{A} - \frac{B}{A}X \ ,$$

som också kan skrivas

$$U = C+DZ$$

där U=1/log Y ("log" är e-logaritmen), Z=X, C=1/A och D=-B/A. Sedan beräknar man den linjära regressionen av U på Z. Till sist översätter man regressionskoefficienterna C och D tillbaka till A=1/C och B=-D/C.

Återstoden av detta avsnitt (4.1) ska handla om direkt anpassning.

Direkt anpassning enligt minstakvadratprincipen måste i allmänhet ske med hjälp av beräkning på dator. Man måste ha tillgång till ett datorprogram för icke-linjär regressionsanalys. Sådana program letar sig fram till ett minimum för avvikelsekvadratsumman iterativt. Man måste hjälpa programmet på traven genom att ange lämpliga startvärden på de okända

konstanterna (A,B). Programmet letar sig sedan fram steg för steg till punkter i (A,B)-planet, som är allt bättre i den meningen att de gör avvikelsekvadratsumman allt mindre. Programmet stoppar då det inte kan hitta en bättre punkt (A,B).

Iterativa beräkningsmetoder har två ofrånkomliga svagheter. För det första är de beroende av startvärden och av de inprogrammerade reglerna för hur man väljer ny punkt. För det andra är de beroende av ett inprogrammerat kriterium för när det stegvisa letandet ska avbrytas och en lösning presenteras. Den första svagheten innebär att beräkningen kan råka vilse och aldrig hitta avvikelsekvadratsummans minimum. Den andra svagheten innebär att beräkningarna kan pågå onödigt länge eller brytas för tidigt.

Det finns standardprogram för icke-linjär regression. Sådana program är givetvis omsorgsfullt uttestade. Men icke-linjär regressionsanalys är - till skillnad från linjär regressionsanalys - inte en rutinberäkning vars egenskaper är fullständigt kartlagda.

Exempel 4.2. Följande sju observationer föreligger.

i	X	Y
1	4	2
2	8	3
3	16	3
4	16	5
5	32	4
6	32	5
7	32	7

Man önskar anpassa en funktion av formen

$$Y = AX^B .$$

Standardprogrammet NLIN i standardprogrampaketet SAS tillämpas. (Startvärden hämtas från den indirekta anpassningen i exempel 4.3 nedan.) Resultatet är

$$Y = 1{,}166X^{0{,}440}$$

med residualkvadratsumma 6,7027. Närmevärden och residualer listas i figur 4.1. De återges grafiskt i figur 4.2.

Figur 4.1

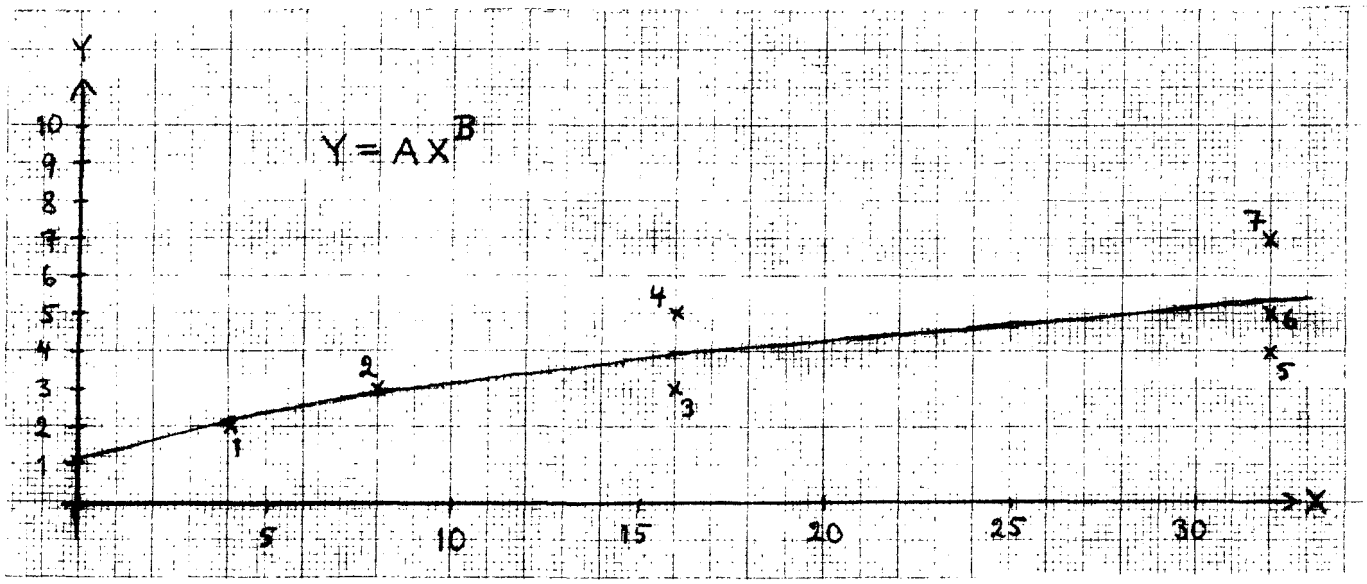
Resultat av direkt anpassning.

ICKELINJÄR REGRESSION

OBS	X	Y	YH	E
1	4	2	2.14515	-0.1452
2	8	3	2.91014	0.0899
3	16	3	3.94793	-0.9479
4	16	5	3.94793	1.0521
5	32	4	5.35582	-1.3558
6	32	5	5.35582	-0.3558
7	32	7	5.35582	1.6442

YH motsvarar $\hat{Y}$.

Figur 4.2



Närmevärden och residualer efter direktanpassning har i allmänhet inte de egenskaper, som omtalades i avsnitt 2.1 ovan. Residualer varken summerar sig till noll eller är okorrelerade med förklaringsvariabeln, om den direktanpassade funktionen är icke-linjär.

4.2 Indirekt anpassning efter linearisерande transformation

Betrakta en icke-linjär funktion

$$Y = f(X)$$

och n bivariata observationer (X_i, Y_i), $i=1,\ldots,n$.

Antag att det finns en linearisерande transformation. Tillämpa den! Resultatet är att nya analys- och förklaringsvariabler definieras:

$$U = h(Y) \text{ respektive } Z=g(X).$$

- Det är också tillåtet att låta den nya analysvariabeln vara en funktion $U=h(Y,X)$ även av förklaringsvariabeln.

Att transformationen är linearisерande innebär att den översätter den givna icke-linjära funktionen $Y=f(X)$ till den linjära funktionen

$$U = C+DZ \ .$$

Data (X_i, Y_i) satisfierar inte $Y=f(X)$ exakt. Följaktligen satisfierar inte heller data (Z_i, U_i) funktionen $U=C+DZ$ exakt.

Beräkna den linjära regressionen av U på Z. Resultatet är närmevärden och residualer

$$\hat{U}_i = C+DZ_i \quad \text{respektive} \quad V_i = U_i - \hat{U}_i$$

som har de egenskaper, som närmevärden och residualer i linjär regression alltid har p g a minstakvadratmetoden. Residualen V summerar sig till noll och är okorrelerad med förklaringsvariabeln Z.

Återtransformationen till i termer av X och Y sker som följer. Transformationen $U=h(Y)$ har en invers $Y=h^{-1}(U)$. - Alternativt: $U=h(Y,X)$ har en invers $Y=h^{-1}(U,X)$. - Den inversa transformationen tillämpas på närmevärdena. Man definierar därigenom närmevärden för den ursprungliga analysvariabeln.

$$\hat{Y} = h^{-1}(\hat{U}).$$

Residualen i ursprunglig skala definieras helt enkelt som differensen

$$E = Y-\hat{Y}.$$

Det finns i allmänhet ingen enkel direkt transformation av residualen V till residualen E, utan översättningen från (U,Z) tillbaka till (Y,X) gäller närmevärdena.

Närmevärdena Y och residualerna E efter indirekt anpassning och återtransformation har i allmänhet inte de egenskaper, som omtalades i avsnitt 2.1 ovan. Residualen varken summerar sig till noll eller är okorrelerad med förklaringsvariabeln.

Exempel 4.3. Samma data som i exempel 4.2. Samma funktion $Y=AX^B$. Logaritmering transformerar $Y=AX^B$ till det linjära sambandet

$$U = C+DZ$$

där

$$U = \log Y \quad \text{och} \quad Z = \log X,$$
$$C = \log A \quad \text{och} \quad D = B\ .$$

Vi beräknar den linjära regressionen av U på Z. Beräkningarna redovisas i figur 4.3. (Vanliga 10-logaritmer har använts.) Resultatet är (där $A = 10^C = 1,128$)

$$Y = 1,128X^{0,443}$$

med residualkvadratsumma (approximativt) 6,7731. Närmevärden och residualer återges approximativt av figur 4.2 ovan.

Närmevärdena vid indirekt anpassning blir i allmänhet inte lika med närmevärdena vid direkt anpassning. I exempel 4.2 och 4.3 är de emellertid så nära lika att skillnaden är svår att tydligt återge grafiskt.

Indirekt anpassning kan beräknas med hjälp av standardprogram för linjär regression, om dessa innehåller tillräckliga faciliteter för transformation av data före och helst även efter regressionsberäkningarna. Figur 4.4 återger resultatet av att låta standardprogrampaketet SAS utföra samma beräkningar som i figur 4.3. (Här har e-logaritmer använts.)

Figur 4.3.

Beräkning av D och C i $U = C + DZ$ samt av $\hat{Y} = h^{-1}(\hat{U})$ och E.

i	X	Y	Z	U	$Z-\bar{Z}$	$U-\bar{U}$	$(Z-\bar{Z})(U-\bar{U})$	$(Z-\bar{Z})^2$	$\hat{U}$	V	$\hat{Y}$	E	E^2
1	4	2	0,6021	0,3010	+0,6020	-0,2848	+0,17144960	0,36240400	0,3190	-0,0180	2,08	-0,08	0,0064
2	8	3	0,9031	0,4771	-0,3010	-0.1087	+0.03271870	0.09060100	0,4524	+0,0247	2,83	+0,17	0,0289
3	16	3	1,2041	0,4771	0	-0,1087	0	0	0,5858	-0,1087	3,85	-0,85	0,7225
4	16	5	1,2041	0,6990	0	+0,1132	0	0	0,5858	+0,1132	3,85	+1,15	1,3225
5	32	4	1,5051	0,6021	+0,3010	+0,0163	+0,00490630	0,09060100	0,7192	-0,1171	5,24	-1,24	1,5376
6	32	5	1,5051	0,6990	+0,3010	+0,1132	+0,03407320	0,09060100	0,7192	-0,0202	5,24	-0,24	0,0576
7	32	7	1,5051	0,8451	+0,3010	+0,2593	+0,07804930	0,09060100	0,7192	+0,1259	5,24	+1,76	3,0976
Σ			8,4287	4,1004	0,0000	-0,0002	+0,32119710	0,72480800	4,1006	-0,0002		+0,67	6,7731
m			1,2041	0,5858								+0,096	

$D = +0,32119710 / 0,72480800 = +0,44314800$

$C = 0,5858 - 0,44314800 \times 1,2041 = +0,05222006$

$\hat{U} = 0,0522 + 0,44314800\,Z$

Figur 4.4

Indirekt anpassning medelst SAS.

INDIREKT ANPASSNING 11:47 TUESDAY, DECEMBER 8

OBS	X	Y	Z	U	UH	V	YH	E
1	4	2	1.38629	0.69315	0.73459	-0.04145	2.08464	-0.0846
2	8	3	2.07944	1.09861	1.04169	0.05693	2.83399	0.1660
3	16	3	2.77259	1.09861	1.34878	-0.25017	3.85272	-0.8527
4	16	5	2.77259	1.60944	1.34878	0.26066	3.85272	1.1473
5	32	4	3.46574	1.38629	1.65587	-0.26958	5.23764	-1.2376
6	32	5	3.46574	1.60944	1.65587	-0.04643	5.23764	-0.2376
7	32	7	3.46574	1.94591	1.65587	0.29004	5.23764	1.7624

INDIREKT ANPASSNING 11:47 TUESDAY, DECEMB

VARIABLE	MEAN	UNCORRECTED SS	CORRECTED SS
E	0.09471653	6.77224441	6.70944586

"Uncorrected SS" motsvarar ΣE^2.

4.3 Determinationskoefficientens definition

Hur väl stämmer data med en viss icke-linjär kurva? Vilken av två kurvor stämmer data bäst överens med? Sådana frågor kan besvaras med hjälp av determinationskoefficienten R^2. Denna koefficient är definierad även vid icke-linjär regression; åtminstone finns det en väl etablerad praxis.

Vid linjär regression kan R^2 definieras på flera olika likvärdiga sätt. Vid icke-linjär regression håller man sig till en av definitionerna, nämligen

$$R^2 = 1 - \Sigma E^2 / \Sigma (Y-\bar{Y})^2 \ .$$

Om residualen E inte har medelvärdet noll, menas här med ΣE^2 kvadratsumman av residualerna sådana de är, inte avvikelsekvadratsumman.

Determinationskoefficientens definition i det icke-linjära fallet kan "motiveras" så här: Man jämför variationen kring den anpassade kurvan med den totala Y-variationen. Ett minus kvoten är den andel av Y-variationen, som "förklaras" av sambandet med X.

Vid indirekt anpassning kan man beräkna två olika determinationskoefficienter. Den ena säger hur stor andel av variationen i den transformerade analysvariabeln U som "förklaras" av det linjära sambandet med den transformerade förklaringsvariabeln Z, låt säga $R^2(U;Z)$. Detta är en alldeles vanlig determinationskoefficient vid linjär regression. Men den handlar inte om variationen hos analysvariabeln i dess ursprungliga skala, och är därför av begränsat intresse.

Den andra determinationskoefficienten vid indirekt anpassning säger hur stor del av variationen i den ursprungliga analysvariabeln Y, som "förklaras" av det icke-linjära sambandet med förklaringsvariabeln X, låt säga $R^2_{ind}(Y;X)$. Detta är den relevanta determinationskoefficienten. Den är direkt jämförbar med determinationskoefficienten vid direkt anpassning, låt säga $R^2_{dir}(Y;X)$.

Eftersom den direkta anpassningen minimerar residualkvadratsumman, så gäller alltid olikheten

$$R^2_{dir}(Y;X) \geq R^2_{ind}(Y;X) \ .$$

Man kan däremot inte vara säker på att $R^2_{dir}(Y;X)$ ska vara större än determinationskoefficienten för den linjära regressionen av Y på X, låt säga $R^2_{lin}(Y;X)$. Huruvida den är det beror på hur väl man har valt icke-linjär funktionstyp.

Exempel 4.4. Samma data som i exempel 4.2 och 4.3. Samma kurva $Y=AX^B$. Variationen hos analysvariabeln Y är SST = 16,8571.

$$R^2_{lin}(Y;X) = 1 - 7,0592/16,8571 = 0,581.$$

$$R^2_{ind}(Y;X) = 1 - 6,7722/16,8571 = 0,598.$$

$$R^2_{dir}(Y;X) = 1 - 6,7027/16,8571 = 0,602.$$

$$R^2(U;Z) = 1 - 0,2944/1,0489 = 0,719 \text{ (avser e-log-skala).}$$

Att, slutligen, som ofta händer,

$$R^2(U;Z) > R^2_{lin}(Y;X)$$

innebär inte att den valda icke-linjära kurvtypen stämmer bättre med data än en rät linje gör. De båda determinationskoefficienter, som här jämförs, handlar om så olika saker att jämförelsen är meningslös.

4.4 Något om extrapolering

Antag att man har anpassat en kurva av en viss typ - till exempel en rät linje - till ett givet bivariat observationsmaterial (X_i,Y_i), $i=1,\ldots,n$. Antag att man drar ut den anpassade kurvan till punkten X_0, där X_0 är större än det största observerade X_i-värdet eller mindre än det minsta observerade X_i-värdet. Man avläser det mot X_0 svarande Y-värdet Y_0 på kurvan och säger: Om X vore X_0, så skulle Y i genomsnitt vara ungefär Y_0. Detta kallas att extrapolera. Hur tillförlitligt är det att göra så?

Exempel 4.5. Samma data som i exempel 4.2 och 4.3. Vi väljer $X_0=64$ och de två alternativa kurvtyperna

$$Y = A+BX \quad \text{respektive} \quad Y = AX^B .$$

Den linjära funktionen anpassas medelst minstakvadratmetoden, vilket ger

$$F_L: \quad Y = 2{,}013 + 0{,}106X .$$

Potensfunktionen medelst direktanpassning är

$$F_P: \quad Y = 1{,}166X^{0{,}440} .$$

Extrapolation till $X_0=64$ ger

$$F_L: \quad Y_0 = 8{,}80 \ ,$$

$$F_P: \quad Y_0 = 7{,}27 \ .$$

Resultatet beror på kurvtypen.

Extrapolering är ett riskabelt förehavande. Två kurvtyper, som det givna materialet stämmer ungefär lika bra med, och som ser nästan lika ut inom det observerade X-intervallet, kan ge avsevärt olika resultat vid extrapolering. Den anpassade funktionen fortsätter sin matematiska bestämda väg även utanför det område där vi har data. Vi vet inte utan vidare att nya data skulle följa samma väg.

Hur mycket hjälper det att tillföra formella förutsättningar? Betrakta åter de tre förutsättnings- och tolkningsnivåerna från kapitel 3!

På den deskriptiva förutsättningsnivån är det ointressant att extrapolera. Andra data än de, som man har anpassat kurva till, ingår överhuvudtaget inte i den rent deskriptiva föreställningsvärlden. Man kan i och för sig extrapolera hur långt man vill, men detta säger ingenting om några som helst nya - eller gamla - data.

På den statistiska förutsättningsnivån kan man föra matematisk-statistiska resonemang om extrapoleringen. Ungefär följande resonemang är brukligt.

Antag att för varje observation väntevärdet för Y är en och samma funktion $E(Y)=f(X)$ av X-värdet. Antag att detta gäller oavsett om observationen är nu given och observerad, nu given men inte observerad, eller icke nu given men möjlig i framtiden eller på annan plats. Vi kan då tolka den extrapolerade anpassade kurvan som statistisk skattning av den sanna funktionen $E(Y)=f(X)$. Detta förutsatt att vi har valt rätt kurvtyp $f(.)$ att anpassa.

Två frågor inställer sig här. (1) Hur kan man veta vilken kurvtyp som är rätt? (2) Hur kan man veta för vilka observationer samma samband gäller? Ingen av dessa båda frågor kan slutgiltigt besvaras på den statistiska förutsättningsnivån, men man kan söka partiella och betingade svar med hjälp av statistisk testning.

På den kausala förutsättningsnivån har man bättre möjligheter att uttala sig i de båda just ställda frågorna. Men svaren måste då ha sin grund inte i data utan i information utöver data.

Ännu några ord bör sägas om extrapolering med enbart statistiska förutsättningar och statistisk tolkning. Vid extrapolering antar man att en given statistisk modell gäller för vissa ännu ej iakttagna observationer. Redan vid anpassningen (estimationen) av den givna kurvtypen antar man ju precis detsamma om samtliga föreliggande observationer. Detta grundläggande antagande kan vara lika felaktigt som det, som ligger till grund för själva extrapoleringen. Hur god grund man har för sitt antagande beror på utomstatistiska faktorer; hur god saklogisk information har vi?

Statistisk metod är inte ett sätt att skapa information ur tomma intet. Statistisk metod är ett sätt att bringa reda i och dra slutsatser ur given information. Osäkerhet om informationens giltighet elimineras inte.

5 LINJÄR REGRESSION MED TVÅ FÖRKLARINGSVARIABLER

5.1 Tre former för planets ekvation

Betrakta ett rätvinkligt koordinatsystem i tre dimensioner (Y,X,Z), där Y mäts lodrätt medan X och Z mäts i två inbördes vinkelräta vågräta riktningar. Ett plan i det tredimensionella rummet kan tecknas (bland annat) så här:

$$Y = A+BX+CZ$$

där (A,B,C) är tre konstanter som bestämmer planets läge i rummet.

De tre konstanternas innebörd är följande.

A = planets höjd över (X,Z)-planet (vars ekvation är Y=0) i punkten (X,Z)=(0,0).

B = tangens för planets lutningsvinkel i X-riktningen, dvs längs skärningen med planet Z=0.

C = tangens för planets lutningsvinkel i Z-riktningen, dvs längs skärningen med planet X=0.

Varje icke lodrätt plan i rummet bestämmer entydigt och bestäms entydigt av de tre konstanterna (A,B,C).

Ett plan kan beskrivas på flera sätt. Låt $\bar{X}$ och $\bar{Z}$ vara två konstanter. (I praktiken väljer man $\bar{X}$ och $\bar{Z}$ lika med medelvärdena av X- respektive Z-värdena hos n givna observationer (Y_i,X_i,Z_i), men det som sägs här i avsnitt 5.1 förutsätter inte detta.)

Givet $(\bar{X},\bar{Z})$ gäller att varje icke lodrätt plan kan skrivas

$$Y = D+F(X-\bar{X})+G(Z-\bar{Z}),$$

varvid taltrippeln (D,F,G) entydigt bestämmer och bestäms av planet. Konstanternas innebörd är följande.

D = planets höjd över (X,Z)-planet i punkten $(X,Z)=(\bar{X},\bar{Z})$.

F = samma som B ovan.

G = samma som C ovan.

Denna andra form för planets ekvation skiljer sig från den första endast i fråga om var man mäter planets höjd över bottenplanet Y=0.

Det finns en tredje form för planets ekvation. Låt $(\bar{X},\bar{Z},T)$ vara tre godtyckliga konstanter. (Vi återkommer till praxis för hur dessa väljes.) Givet de tre konstanterna $(\bar{X},\bar{Z},T)$ kan varje icke lodrätt plan skrivas

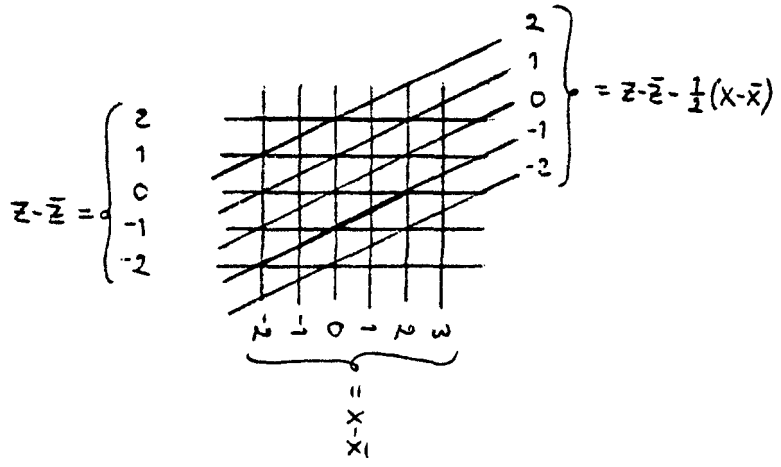
$$Y = M+P(X-\bar{X})+Q\left[Z-\bar{Z}-T(X-\bar{X})\right]$$

varvid de tre konstanterna (M,P,Q) entydigt bestämmer planet och omvänt.

Den tredje varianten innebär, kan man säga, att koordinatsystemet $((X-\bar{X}),(Z-\bar{Z}))$ i bottenplanet byts mot koordinatsystemet $\left[(X-\bar{X}),(Z-\bar{Z})-T(X-\bar{X})\right]$. Det förra koordinatsystemet är ett rätvinkligt rutnät, det senare ett snedvinkligt, vilket utgöres av linjer där $X-\bar{X}$=const respektive $Z-\bar{Z}-T(X-\bar{X})$= const. Varje punkt i bottenplanet bestäms omvändbart entydigt av värdena på de båda koordinaterna i det ena systemet, och likadant av det andra systemets koordinater. Se figur 5.1!

Figur 5.1

Två alternativa koordinatsystem



I praktiken väljer man alltid $(\bar{X},\bar{Z},T)$ som funktioner av n givna observationer (Y_i,X_i,Z_i). Man väljer $\bar{X}$ och $\bar{Z}$ lika med X- respektive Z-medelvärdena. Man väljer T lika med regressionskoefficienten i regressionen av den andra förklaringsvariabeln Z på den första X, så att

$$Z-\bar{Z}-T(X-\bar{X})$$

är residualen i den regressionen. Dessa val av $(\bar{X},\bar{Z},T)$ har fördelar då man ska utreda de deskriptiva egenskaperna hos regression med två förklaringsvariabler.

Samtliga tre sätt att skriva ett plan anger för varje punkt (X,Z) hur högt lodrätt över den planet befinner sig. Skillnaderna mellan varianterna gäller hur punkten i bottenplanet beskrivs. Det lodräta Y-avståndet beskrivs i samtliga tre fall lika.

Exempel 5.1. Sätt $\bar{X}=5$, $\bar{Z}=3$ och $D=(1/2)$. Ett och samma plan kan betecknas på följande sätt:

$$Y = 1 + X + 2Z$$
$$Y = 12 + (X-5) + 2(Z-3)$$
$$Y = 12 + 2(X-5) + 2((Z-3)-(1/2)(X-5))$$

Sätt $(X,Z)=(6,9/2)$. Enligt alla tre skrivsätten blir $Y=16$.

5.2 Normalekvationerna i avvikelseform och residualform

Låt ett antal observationer (X_i,Z_i,Y_i), $i=1,\ldots,n$ vara givna. Minstakvadratprincipen säger att det plan passar bäst, för vilket summan av de kvadrerade avvikelserna i Y-riktningen från punkt till plan är minst.

De tre konstanter, som entydigt bestämmer det bästa planet, kan lösas ut ur ett system av tre linjära ekvationer, normalekvationerna. Hur normalekvationssystemet ser ut beror på hur man har valt att skriva planets ekvation. Vilket plan som passar bäst beror däremot inte på skrivsättet. De tre varianterna ger ett och samma plan skrivet på tre olika men likvärdiga sätt.

Första formen för planets ekvation är $Y=A+BX+CZ$. Normalekvationernas lösning, de partiella regressionskoefficienterna, brukar betecknas så här:

$$A = A_{Y\cdot XZ} \qquad (=\bar{Y}-B\bar{X}-C\bar{Z})$$
$$B = B_{YX\cdot Z}$$
$$C = B_{YZ\cdot X}$$

Residualen brukar betecknas

$$Y-A_{Y\cdot XZ}-B_{YX\cdot Z}X-B_{YZ\cdot X}Z = E_{Y\cdot XZ}$$

Man kan visa att den har medelvärdet noll och är okorrelerad med X och med Z.

Andra formen för planets ekvation är $Y = D+F(X-\bar{X})+G(Z-\bar{Z})$. Vi sätter $\bar{X}$ och $\bar{Z}$ lika med medelvärdena över de givna observationerna av X resp Z. De mot denna ekvationsform svarande normalekvationerna i avvikelseform är

$$\begin{aligned} nD &= \Sigma Y \\ (\Sigma(X-\bar{X})^2)F + (\Sigma(X-\bar{X})(Z-\bar{Z}))G &= \Sigma(X-\bar{X})(Y-\bar{Y}) \\ (\Sigma(X-\bar{X})(Z-\bar{Z}))F + (\Sigma(Z-\bar{Z})^2)G &= \Sigma(Z-\bar{Z})(Y-\bar{Y}) \end{aligned}$$

Man kan visa att följande gäller:

$$\begin{aligned} D &= \bar{Y} \\ F &= B_{YX.Z} \\ G &= B_{YZ.X} \end{aligned}$$

Lägg märke till att normalekvationerna i avvikelseform förenklas om $\Sigma(X-\bar{X})(Z-\bar{Z})=0$ dvs om de båda förklaringsvariablerna är inbördes helt okorrelerade.

I tredje formen för planets ekvation ingår en konstant T. Låt oss sätta $T=B_{ZX}$, där B_{ZX} är koefficienten i regressionen av Z på X, beräknad på de givna observationerna. Låt oss också införa beteckningen

$$E_{Z.X} = (Z-\bar{Z})-B_{ZX}(X-\bar{X})$$

för residualen i den regressionen. Den tredje formen av planets ekvation är med de beteckningarna

$$Y = M + P(X-\bar{X}) + QE_{Z.X}$$

där variabeln $E_{Z.X}$ har medelvärdet noll eftersom den är en residual. Jämför figur 5.1!

De mot den tredje formen för planets ekvation svarande normalekvationerna i residualform är

$$\begin{aligned} nM &= \Sigma Y \\ (\Sigma(X-X)^2)P &= \Sigma(X-\bar{X})(Y-\bar{Y}) \\ (\Sigma E^2_{Z\cdot X})\ Q &= \Sigma E_{Z\cdot X}(Y-\bar{Y}) \end{aligned}$$

Ekvationssystemet blir så pass enkelt därför att X och $E_{Z\cdot X}$ är okorrelerade med varandra; detta emedan $E_{Z\cdot X}$ är en residual efter regression på X.

Man ser omedelbart att

$$M = \bar{Y}$$

och att

$$P = \Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(X-\bar{X})^2 = B_{YX}$$

Den partiella regressionskoefficienten för X i regression av Y på X och residualen $E_{Z\cdot X}$ är således lika med regressionskoefficienten i den enkla regressionen av Y på enbart X.

Man ser också att

$$Q = \Sigma E_{Z\cdot X}(Y-\bar{Y})/\Sigma E^2_{Z\cdot X}$$

Den partiella regressionskoefficienten för residualen $E_{Z\cdot X}$ i regressionen av Y på X och $E_{Z\cdot X}$ är lika med regressionskoefficienten i den enkla regressionen av Y på enbart residualen $E_{Z\cdot X}$.

Genom att lösa ut G ur normalekvationssystemet i avvikelseform kan man också visa att

$$Q = G = B_{YZ\cdot X}\ .$$

Lägg märke till detta! Den partiella regressionskoefficienten $G=B_{YZ\cdot X}$ för Z i regressionen av Y på X och Z är lika med regressionskoefficienten i den enkla regressionen av Y på enbart residualen $E_{Z\cdot X}$ av Z efter regression på X.

Normalt använder man avvikelseformen av planets ekvation och normalekvationerna. Residualformen har här införts därför att man med dess hjälp kan visa hur saker och ting hänger ihop då man utökar mängden förklaringsvariabler.

Exempel 5.2. Följande data föreligger

i	X	Z	Y
1	4	3	18
2	6	5	21
3	8	3	21
4	10	7	33
5	12	7	32

Vi ska beräkna regressionen av Y på X och Z. Beräkningarna redovisas mer detaljerat i figur 5.2.

Normalekvationerna i avvikelseform är

$$\begin{aligned} 5\,\bar{Y} &= 125 \\ 40B_{YX\cdot Z} + 20B_{YZ\cdot X} &= 80 \\ 20B_{YX\cdot Z} + 16B_{YZ\cdot X} &= 52 \end{aligned}$$

Lösningen är

$$\begin{aligned} \bar{Y} &= 25 \\ B_{YX\cdot Z} &= 1 \\ B_{YZ\cdot X} &= 2 \end{aligned}$$

Regressionsekvationen är

$$\begin{aligned} Y &= 25 + 1(X-8) + 2(Z-5) = \\ &= 7 + 1X + 2Z\ . \end{aligned}$$

För att visa ett exempel på att residualformen också ger rätt regressionsplan räknar vi om exemplet i residualform.

Exempel 5.3. Samma data som i exempel 5.2.

Vi ska beräkna regressionen av Y på X och $E_{Z\cdot X}$. Beräkningarna redovisas mer detaljerat i figur 5.2.

Normalekvationerna i residualform är

$$\begin{aligned} 5\bar{Y} &= 125 \\ 40B_{YX} &= 80 \\ 6B_{YZ\cdot X} &= 12 \end{aligned}$$

Lösningen är

$$\begin{aligned} \bar{Y} &= 25 \\ B_{YX} &= 2 \\ B_{YZ\cdot X} &= 2 \end{aligned}$$

Regressionsekvationen är

$$\begin{aligned} Y &= 25 + 2(X-8) + 2\left[Z-5-\tfrac{1}{2}(X-8)\right] = \\ &= 25 + 1(X-8) + 2(Z-5) = \\ &= 7 + 1X + 2Z. \end{aligned}$$

Figur 5.2. Beräkningar till exempel 5.1 och 5.2.

	X	Z	Y	$x = X-\bar{X}$	$z = Z-\bar{Z}$	$y = Y-\bar{Y}$	xz	xy	zy	x^2	z^2
	4	3	18	-4	-2	-7	+8	+28	+14	16	4
	6	5	21	-2	0	-4	0	+8	0	4	0
	8	3	21	0	-2	-4	0	0	+8	0	4
	10	7	33	+2	+2	+8	+4	+16	+16	4	4
	12	7	32	+4	+2	+7	+8	+28	+14	16	4
Σ	40	25	125	0	0	0	+20	+80	+52	40	16
m	8	5	25								

Ex. 5.1

$\Sigma(X-\bar{X})^2 = 40$ $\quad \Sigma(X-\bar{X})(Z-\bar{Z}) = 20$ $\quad \Sigma(X-\bar{X})(Y-\bar{Y}) = 80$

$\Sigma(Z-\bar{Z})^2 = 16$ $\quad \Sigma(Z-\bar{Z})(Y-\bar{Y}) = 52$

Vi beräknar regressionen av Z på X:

$B_{ZX} = +20/40 = +\frac{1}{2}$. $\quad A_{ZX} = 5 - \frac{1}{2} \times 8 = 1$.

Regressionen är $\hat{Z} = 1 + \frac{1}{2}X$.

	X	Z	$\hat{Z}$	$E_{Z \cdot X}$	x	e	y	xe	ey	e^2
	4	3	3	0	-4	0	-7	0	0	0
	6	5	4	+1	-2	+1	-4	-2	-4	1
	8	3	5	-2	0	-2	-4	0	+8	4
	10	7	6	+1	+2	+1	+8	+2	+8	1
	12	7	7	0	+4	0	+7	0	0	0
Σ				0	0	0	0	0	+12	6

Ex. 5.2

$\Sigma(X-\bar{X})^2 = 40$ $\quad \Sigma(X-\bar{X})(E-\bar{E}) = 0$ $\quad \Sigma(X-\bar{X})(Y-\bar{Y}) = 80$

$\Sigma(E-\bar{E})^2 = 6$ $\quad \Sigma(E-\bar{E})(Y-\bar{Y}) = 12$

5.3 Uppdelning av avvikelsekvadratsumman

Regressionen av Y på X och Z fungerar i följande två avseenden exakt analogt med regressionen av Y på enbart X.

(1) Regressionen delar upp varje observationsvärde Y_i i ett närmevärde $\hat{Y}_i$ och en residual E_i. Närmevärdenas medelvärde är lika med de observerade Y-värdenas medelvärde. Residualernas medelvärde är noll. Korrelationen mellan residualen och var och en av förklaringsvariablerna X och Z är noll. Jämför avsnitt 2.1 ovan!

(2) Regressionen delar upp analysvariabelns totala avvikelsekvadratsumma SST i en förklarad kvadratsumma SSR och en residualkvadratsumma SSE. Jämför avsnitt 2.2 ovan!

Närmevärdena är

$$\hat{Y}_i = A_{Y \cdot XZ} + B_{YX \cdot Z} X_i + B_{YZ \cdot X} Z_i$$

och residualerna är

$$E_i = Y_i - \hat{Y}_i \ .$$

De tre kvadratsummorna definieras precis likadant som i avsnitt 2.2 ovan.

Determinationskoefficienten definieras

$$R^2 = 1 - SSE/SST$$

precis som i avsnitt 2.3 ovan för enkel regression.

Exempel 5.3. Samma data som i exempel 5.2.

Närmevärdena $\hat{Y} = 7+1X+2Z$ och residualerna samt de tre kvadratsummorna beräknas. Beräkningarna redovisas mer detaljerat i figur 5.3.

Observerade värden, närmevärden och residualer är

$$\mathbf{Y} = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 33 \\ 32 \end{bmatrix}; \quad \hat{\mathbf{Y}} = \begin{bmatrix} 17 \\ 23 \\ 21 \\ 31 \\ 33 \end{bmatrix}; \quad \mathbf{E} = \begin{bmatrix} +1 \\ -2 \\ 0 \\ +2 \\ -1 \end{bmatrix}$$

Den totala, den förklarade, och residual-kvadratsumman är

$$SST = 194, \quad SSR = 184, \quad SSE = 10$$

Determinationskoefficienten är

$$R^2 = 1-10/194 = 0{,}948.$$

Det verifieras att

$$\Sigma(X-\bar{X})(E-\bar{E}) = \Sigma(Z-\bar{Z})(E-\bar{E}) = 0$$

dvs att residualen är okorrelerad med vardera förklaringsvariabeln.

Figur 5.3. Beräkningar till exempel 5.3.

	X	Z	Y	$\hat{Y}$	$e=E$	$Y-\bar{Y}$	$\hat{Y}-\bar{Y}$	$x=X-\bar{X}$	$z=Z-\bar{Z}$	xe	ze
	4	3	18	17	+1	-7	-8	-4	-2	-4	-2
	6	5	21	23	-2	-4	-2	-2	0	+4	0
	8	3	21	21	0	-4	-4	0	-2	0	0
	10	7	33	31	+2	+8	+6	+2	+2	+4	+4
	12	7	32	33	-1	+7	+8	+4	+2	-4	-2
Σ	40	25	125	125	0					0	0
m	8	5	25	25							

5.4 Stegvis ändring av regressionskoefficienter och kvadratsummor

Betrakta å ena sidan den linjära regressionen av Y på enbart X, å andra sidan den linjära (multipla) regressionen av Y på X och Z. Regressionslinjen respektive regressionsplanet är

$$Y = A_{Y\cdot X} + B_{YX}X$$
$$Y = A_{Y\cdot XZ} + B_{YX\cdot Z}X + B_{YZ\cdot X}Z$$

Att gå över från den första till den andra regressionen är att utöver X införa en ny förklaringsvariabel Z. Detta medför att regressionskoefficienten för X ändras från den enkla (eller totala) regressionskoefficienten B_{YX} till den partiella, $B_{YX\cdot Z}$. Också interceptet ändras, men det lämnar vi därhän. Vad kan man säga om hur och varför B_{YX} ändras till $B_{YX\cdot Z}$?

Avvikelseformen och residualformen av planets ekvation gav oss två olika sätt att skriva det av minstakvadratprincipen utvalda bästa planet. Avvikelseformen skriver regressionsplanet

$$Y = \bar{Y} + B_{YX\cdot Z}(X-\bar{X}) + B_{YZ\cdot X}(Z-\bar{Z})$$

Residualformen skriver

$$\begin{aligned} Y &= \bar{Y} + B_{YX}(X-\bar{X}) + B_{YZ\cdot X}E_{Z\cdot X} = \\ &= \bar{Y} + B_{YX}(X-\bar{X}) + B_{YZ\cdot X}((Z-\bar{Z})-B_{ZX}(X-\bar{X})) = \\ &= \bar{Y} + (B_{YX}-B_{ZX}B_{YZ\cdot X})(X-\bar{X}) + B_{YZ\cdot X}(Z-\bar{Z}) \end{aligned}$$

Eftersom båda varianterna bestämmer samma plan, gäller tydligen sambandet

$$B_{YX\cdot Z} = B_{YX} - B_{ZX}B_{YZ\cdot X}$$

där $B_{YZ.X}$ enligt ovan också kan uppfattas som regressionskoefficienten i regressionen av Y på residualen $E_{Z.X}$.

Sambandet mellan enkel och partiell regressionskoefficient kan också skrivas

$$B_{YX} = B_{YX.Z} + B_{ZX}B_{YZ.X}$$

Denna ekvation brukar ibland klädas i ord ungefär så här: "Den totala inverkan B_{YX} av X på Y är summan av den egentliga inverkan $B_{YX.Z}$ av X och den del av Z:s inverkan $B_{YZ.X}$, som kan tillskrivas X därför att variationen i Z delvis förklaras (B_{ZX}) av variationen i X." Ekvationen i förra stycket utlägges analogt i termer som "rensa bort ur den totala inverkan av X den del som egentligen är inverkan av Z". Varning för förhastade övertolkningar!

Linjär regressionsanalys delar den totala Y-kvadratsumman SST i två delar, den förklarade kvadratsumman SSR och residualkvadratsumman SSE. Låt oss beteckna dessa vid enkel regression

$$SST = SSR(Y;X) + SSE(Y;X)$$

och vid multipel regression

$$SST = SSR(Y;X,Z) + SSE(Y;X,Z).$$

Då man utöver X inför Z som förklaringsvariabel, ändras den förklarade kvadratsumman från SSR(Y;X) till SSR(Y;X,Z). Man kan visa att

$$SSR(Y;X,Z) = SSR(Y;X) + SSR(Y;E_{Z.X})$$

dvs ändringen är en icke-negativ ökning lika med den förklarade kvadratsumman i den enkla linjära regressionen av Y på residualen $E_{Z.X}$.

Residualkvadratsumman ändras samtidigt. Den minskas med samma kvantitet.

Den förklarade kvadratsumman (SSR(Y;X,Z) är den del av analysvariabelns variation, som regressionen "förklarar" genom att tillskriva den samvariationen mellan å ena sidan Y, å andra sidan X och Z. Det vore tilltalande att kunna vidareuppdela denna förklarade kvadratsumma i en del som beror på X plus en del som beror på Z. Men det går inte. Förklaringsvariablerna förklarar gemensamt SSR, men vars och ens separata bidrag kan inte definieras.

Det finns dock ett specialfall, där den förklarade kvadratsumman har en enkel meningsfull uppdelning i en del för vardera förklaringsvariabeln. Det är om och endast om korrelationen mellan de båda förklaringsvariablerna X och Z är exakt 0. Under den förutsättningen gäller att

$$SSR(Y;X,Z) = SSR(Y;X) + SSR(Y;Z).$$

I alla andra fall stöder förklaringsvariablerna varandra i förklaringsarbetet och går samtidigt i vägen för varandra så att deras andelar av förklaringen blir omöjliga att reda ut.

Till sist några ord om stegvis regression. Med stegvis regression menas att beräkna först regressionen av Y på X, sedan regressionen av Y på X och Z, sedan regressionen av Y på X och Z och ännu en förklaringsvariabel, och så vidare - eller att på något annat sätt steg för steg ändra uppsättningen förklaringsvariabler. Man kan antingen själv styra processen eller låta datorn välja variabler efter något inprogrammerat valkriterium. Det finns flera olika sådana kriterier, som alla bygger på uppdelningen i SSR och SSE och hur den ändras.

Stegvis regression är en mycket populär teknik för att leta efter samband i multivariata material. Dess innebörd och statistiska egenskaper är ofullständigt utredda.

Exempel 5.4. Ur jordbruksstatistiken togs data för de 24 länen med avseende på fyra variabler X1, X2, X3 och Y.

X1 = antalet jordbruk med litet antal kor
X2 = antalet jordbruk med medelstort antal kor
X3 = antalet jordbruk med stort antal kor
Y = antalet jordbruk som innehar mjölktank.

Tre regressioner beräknades.

$Y = 316 + 0{,}97X1$
SSR = 10,2 miljoner SSE = 3,6 miljoner $R^2 = 0{,}74$

$Y = 27 - 0{,}03X1 + 0{,}87X2$
SSR = 11,5 miljoner SSE = 2,3 miljoner $R^2 = 0{,}83$

$Y = -23 + 0{,}21X1 + 0{,}55X2 + 1{,}33X3$
SSR = 12,0 miljoner SSE = 1,7 miljoner $R^2 = 0{,}87$

Exemplet visar hur regressionskoefficienter och kvadratsummor kan ändras vid stegvis ändring av uppsättningen förklaringsvariabler. Exemplet torde kunna anses typiskt, även såtillvida att tolkningen är svår.

5.X Bevis (överkurs)

Bevis att normalekvationernas lösning minimerar avvikelsekvadratsumman.

Beviset föres i residualform.

Residualen är $E = Y-\bar{Y}-B_{YX}(X-\bar{X})-B_{YZ.X}\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]$

Avvikelsen från ett godtyckligt annat plan kan tecknas

$$U = Y-(\bar{Y}+\Delta\bar{Y}) - (B_{YX}+\Delta B_{YX})(X-\bar{X}) - \\ - (B_{YZ.X}+\Delta B_{YZ.X})\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right],$$

där $(\Delta\bar{Y},\Delta B_{YX},\Delta B_{YZ.X}) \neq (0,0,0)$

Avvikelsekvadratsumman är

$$\Sigma U^2 = \Sigma\left\{E-\Delta\bar{Y}-\Delta B_{YX}(X-\bar{X})-\Delta B_{YZ\cdot X}\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]\right\}^2 =$$

$$= \Sigma E^2 + n(\Delta\bar{Y})^2 + (\Delta B_{YX})^2\Sigma(X-\bar{X})^2 +$$

$$+ (\Delta B_{YZ\cdot X})^2\Sigma\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]^2$$

Dubbelprodukterna med $\Delta\bar{Y}$ är 0 emedan $\Sigma E = \Sigma(X-\bar{X}) = \Sigma(Z-\bar{Z}) = 0$
Dubbelprodukterna med E är 0 p g a (1) i avsnitt 5.3).
Dubbelprodukten

$$(\Delta B_{YX})(\Delta B_{YZ\cdot X})\ \Sigma(X-\bar{X})\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]$$

är 0 p g a definitionen av B_{ZX}.

Vi ser att

$$\Sigma U^2 \geq \Sigma E^2$$

med likhet om och endast om $\Delta\bar{Y} = \Delta B_{YX} = \Delta B_{YZ\cdot X} = 0$

(Fallet $\Sigma(X-\bar{X})^2 = 0$ är ett urartningsfall analogt med i avsnitt 2.X.ovan.)

(Fallet $\Sigma\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]^2 = 0$ är ett annat urartningsfall. Här är Z en exakt linjär funktion av X. Detta kallas exakt multikollinearitet. Regressionsanalysen fungerar inte normalt. Regressionsplanet är underbestämt.)

5.Y Härledning (Överkurs)

Härledning av regressionen av Y på X ur sambanden mellan Y, X och Z.

De trivariata observationerna (Y,X,Z) åskådliggöres grafiskt i figur 5.4. Likaså de bivariata observationer (Y,X), som man erhåller genom att försumma variabeln (dimensionen) Z.

Regressionsplanet i regressionen av Y på X och Z åskådliggörs grafiskt i figur 5.5. Likaså regressionslinjen i den enkla regressionen av Y på X. Lägg märke till att

$$B_{YX\cdot Z}{=}1 \neq B_{YX}{=}2 \ .$$

Även intercepten är olika, 7 respektive 9.

Vi ska nu härleda den enkla regressionen av Y på X ur dels den multipla regressionen av Y på X och Z, dels hjälpregressionen av Z på X. Vi refererar i fortsättningen genomgående till figur 5.6.

Regressionen av Y på X och Z är i residualform

$$Y = \bar{Y} + B_{YX}(X-\bar{X}) + B_{YZ\cdot X}\left[Z-\bar{Z}+B_{ZX}(X-\bar{X})\right]$$
$$Y = 25 + 2(X-8) + 2\left[Z-5+\tfrac{1}{2}(X-8)\right]$$

Detta är "taket" i den övre delfiguren i figur 5.6.

Regressionen av Z på X är (jfr fig 5.2)

$$Z = 1 + \tfrac{1}{2}X.$$

Den åskådliggöres av en linje i (X,Z)-planet, dvs golvet i den övre delfiguren. Vi ritar in ett lodrätt plan, som skär golvet i denna linje, och betraktar dess skärning med taket.

Skärningslinjen är numeriskt definierad enligt följande.

(1) Ta ett godtyckligt X.
(2) Ta motsvarande $Z=1+\frac{1}{2}X$.
    För detta Z gäller att $Z-5-\frac{1}{2}(X-8)=0$.
(3) Se efter i regressionen av Y på X och Z vilket Y som svarar mot (X,Z).
    Svaret är

$$Y = \bar{Y} + B_{YX}(X-\bar{X}) + B_{YZ\cdot X}[0]$$

$$Y = 25 + 2(X-8) + 2(0) = 9 + 2X.$$

Detta är den enkla regressionen av Y på X.

Skärningslinjen mellan å ena sidan taket

$$Y = \bar{Y} + B_{YX}(X-\bar{X}) + B_{YZ\cdot X}\left[Z-\bar{Z}-B_{ZX}(X-\bar{X})\right]$$

och å andra sidan hjälpregressionen (här i avvikelseform)

$$Z = \bar{Z} + B_{ZX}(X-\bar{X})$$

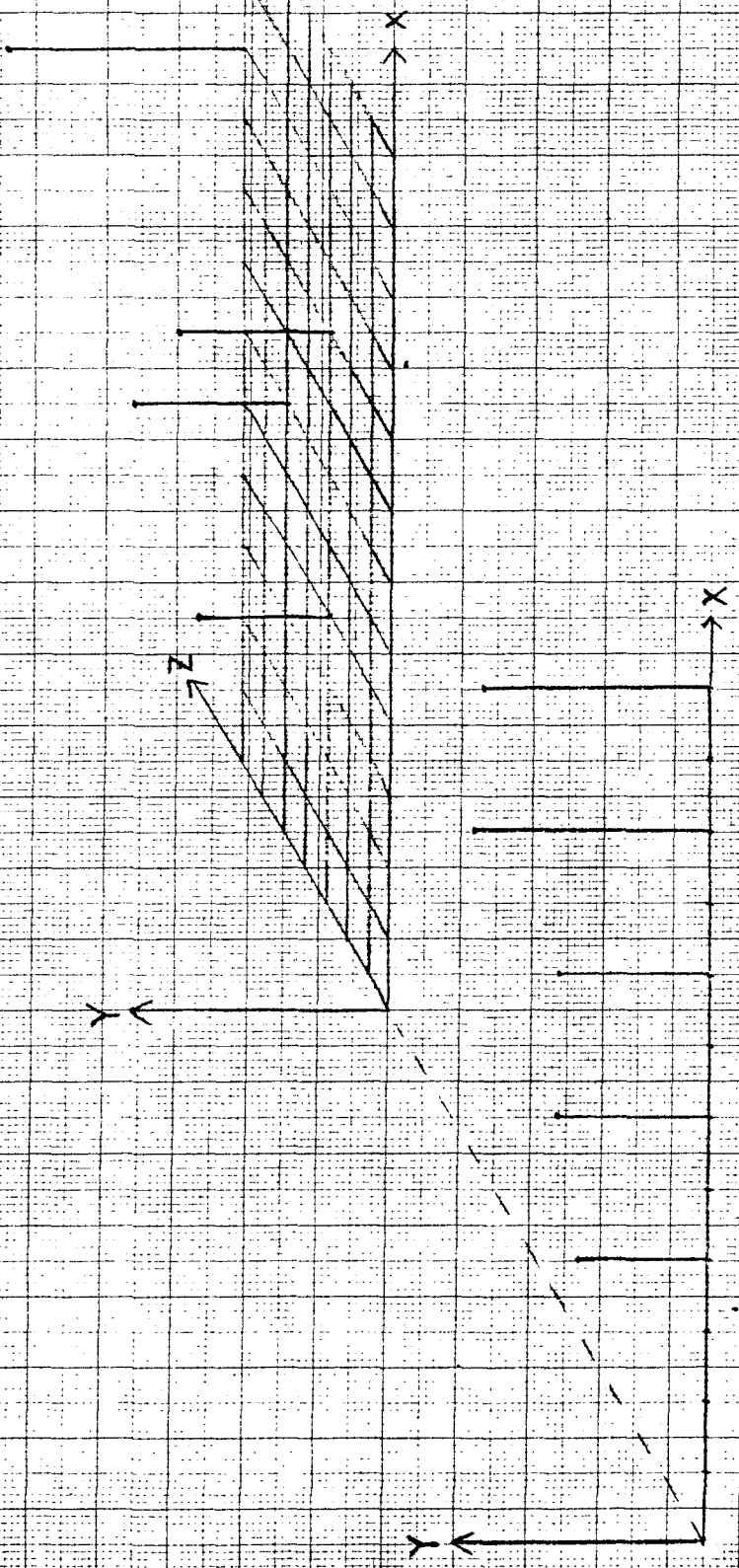
är den enkla regressionen av Y på X

$$Y = \bar{Y} + B_{YX}(X-\bar{X}).$$

Vi projicerar skärningslinjen på den övre delfigurens framsida och får den nedre delfiguren.

Figur 5.4

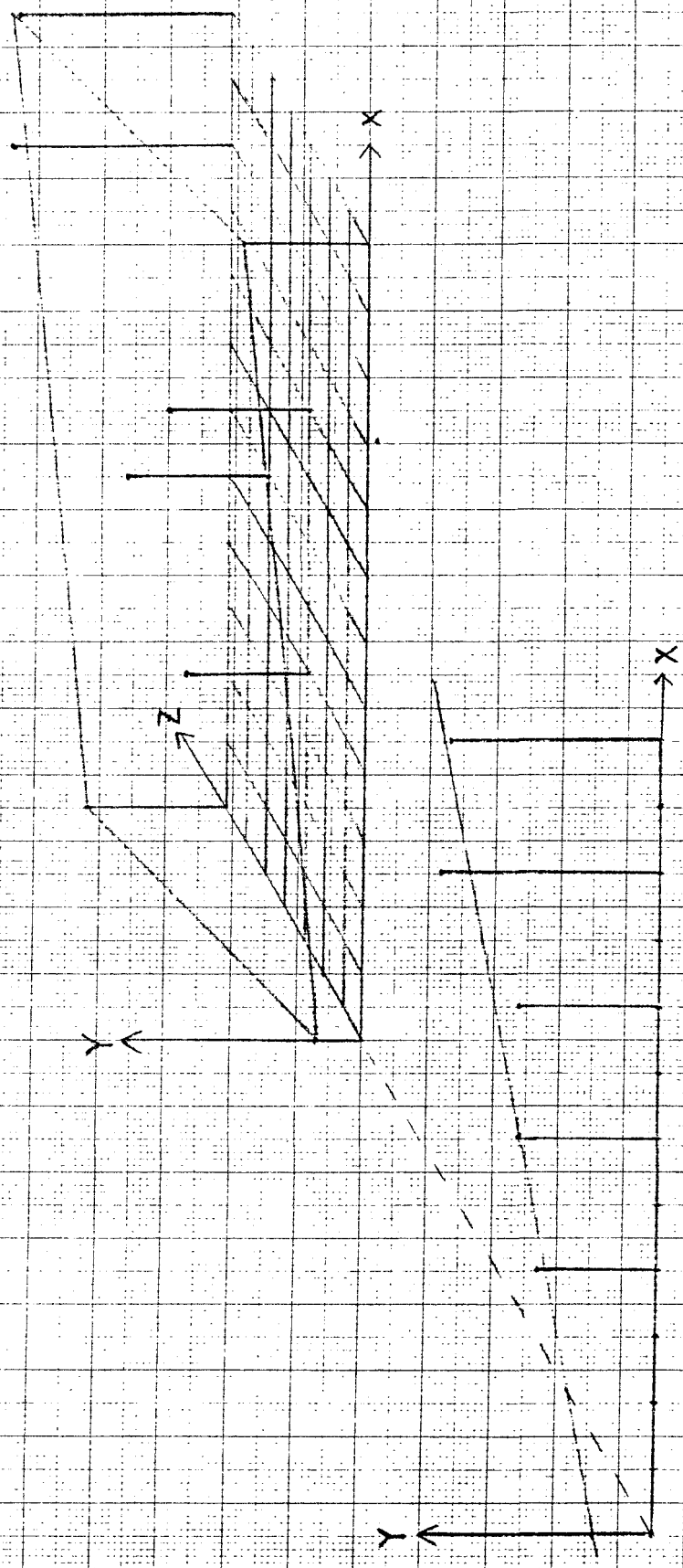
Data i exempel 5.1, 5.2 och 5.3.
Överst: (X,Y,Z).
Nederst: (X,Y).



Den nedre delfiguren erhålles ur den övre genom projektion. De fem punkterna i den övre delfiguren parallellförflyttas till dennas "framsida". Hela "framsidan" parallellförflyttas sedan för att bli lättare urskiljbar.

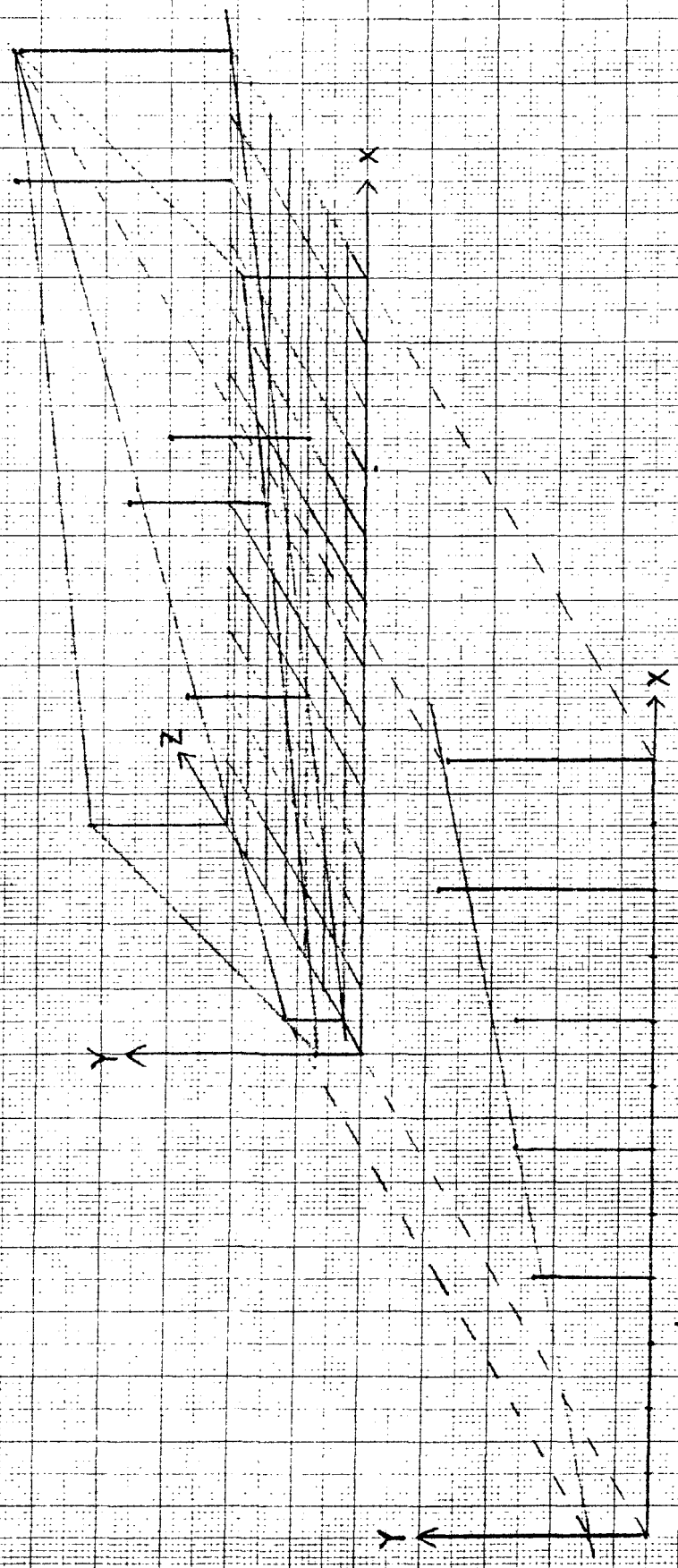
Figur 5.5

Regressioner i exempel 5.1, 5.2 och 5.3.
Nederst: Y = 9 + 2X
Överst: Y = 7 + 1X + 2Z



Figur 5.6

Härledning av regressionen av Y på X ur sambanden mellan Y, X och Z.
Se texten!



6 LINJÄR REGRESSION MED PARALLELL SKIKTNING

6.1 Tre former för parallella räta linjers ekvation

Två eller flera parallella räta linjers ekvationer kan tecknas som en enda ekvation på åtminstone följande olika sätt. Låt linjerna vara numrerade med ett index h = 1, 2, ..., k. Det första skrivsättet är

$$L_h: \; Y = A_h + BX \qquad h = 1, \ldots, k$$

Linjerna har gemensam lutningsvinkel men olika intercept (nivåer). Detta skrivsätt är symmetriskt.

Det andra skrivsättet förutsätter att man tar en linje, säg L_1, till referenslinje och jämför alla de andra med den. Man skriver

$$L_1: \; Y = A + BX$$

$$L_h: \; Y = A + D_h + BX \qquad h = 2, \ldots, k$$

Konstanterna D_2, D_3, ..., D_k anger nivåskillnader i förhållande till referenslinjen L_1. Skrivsättet är osymmetriskt.

Det finns också ett tredje skrivsätt, nämligen

$$L_h: \; Y = C_0 + C_h + BX \qquad h = 1, \ldots, k$$

Här uttryckes k nivåer medelst de k+1 konstanterna C_0, C_1, ..., C_k. Detta skrivsätt är därför inte omvändbart entydigt. Konstanterna bestämmer linjerna entydigt, men k parallella linjer bestämmer inte k+1 konstanter entydigt.

Om i det tredje skrivsättet linjerna ska bestämma konstanterna entydigt, måste konstanterna underkastas ett bivillkor. Ett naturligt sådant bivillkor är $\Sigma C_h = 0$, där summan går från 1 till k.

Exempel 6.1 De tre parallella linjerna (i första skrivsättet)

L_1: Y = 12+3X
L_2: Y = 13+3X
L_3: Y = 8+3X

skrivs i det andra skrivsättet

L_1: Y = 12+3X
L_2: Y = 12+1+3X
L_3: Y = 12-4+3X

och i det tredje skrivsättet med det nämnda bivillkoret

L_1: Y = 11+1+3X
L_2: Y = 11+2+3X
L_3: Y = 11-3+3X

Utan bivillkoret kan linjerna i det tredje skrivsättet skrivas på många andra sätt. Konstanten C_0 kan väljas godtyckligt.

6.2 Indikatorvariabler

Antag att en observationsmängd delas upp i delmängder sådana att varje observation tillhör en och endast en delmängd och ingen delmängd är tom. Låt delmängderna indiceras h = 1, 2, ..., k. Det är då praktiskt att definiera k stycken indikatorvariabler (även kallade dummyvariabler) enligt följande. Låt observationerna vara indicerade j = 1, ..., n.

För h = 1, ..., k gäller:

$$Z_{hj} = \begin{cases} 1 \text{ om observation } j \text{ tillhör delmängd } h \\ 0 \text{ eljest} \end{cases}$$

Det finns således k dummyvariabler $(Z_1\ Z_2\ Z_3\ \ldots\ Z_h\ \ldots\ Z_k)$ som alla är definierade för varje observation. Om t ex observation j = 8 tillhör delmängd h = 2, så är

$$(Z_{18}\ Z_{28}\ Z_{38}\ \cdots\ Z_{k8}) \;=\; (0\ 1\ 0\ \ldots\ 0).$$

För varje observation är precis ett Z_{hj} lika med 1, alla övriga lika med 0.

Indikatorvariabler användes i regressionsanalysen (bland annat) då man till olika delmängder av data vill anpassa räta linjer (eller plan) och vill ha dessa linjer (eller plan) parallella. De k linjerna skrivs enligt det första skrivsättet från 6.1 ovan

$$Y = A_1Z_1+A_2Z_2+\ldots\ldots+A_kZ_k+BX$$

dvs som en linjär regressionsekvation med k+1 förklaringsvariabler, varav k stycken är indikatorvariabler, och - observera! - utan intercept. En sådan regressionsekvation kan beräknas av ett datorprogram för regressionsanalys om detta medger att interceptet stryks ur ekvationen, inte annars. (SAS-GLM medger detta.)

I det andra skrivsättet från 6.1 skrivs de k regressionslinjerna

$$Y = A+D_2Z_2+\ldots\ldots+D_kZ_k+BX$$

dvs som en linjär regressionsekvation med k förklaringsvariabler, varav k-1 stycken är indikatorvariabler. Här ingår ett intercept som vanligt.

Resultatet av en regressionsanpassning med indikatorvariabler för nivåskillnader är oberoende av vilket skrivsätt man har valt. De k bästa parallella räta linjerna (eller planen) är entydigt bestämda av data och påverkas inte av hur man väljer att beskriva dem.

Med det tredje skrivsättet i 6.1 ovan skulle man ansätta k dummyvariabler och intercept. Denna ansats leder till matematisk

underbestämdhet. Datorkörningen avbryts med felutskrift (eller räknar fel).

Exempel 6.2 (Påhittat)

Betrakta dem som börjar i ett visst yrke vid cirka 20 års ålder. Antag att lönen Y utvecklas ungefär linjärt med anställningstiden X enligt följande tre separata regressioner.

$Y = 30+1,03X$ för dem som har enbart grundskola (dummyvariabel Z_1)
$Y = 35+0,98X$ för dem som har enbart fackskola (Z_2)
$Y = 36+1,01X$ för dem som har specialutbildning (Z_3)

Ansatsen med dummyvariabler jämkar ihop dessa tre nästan parallella räta linjer till ekvationen (i första skrivsättet)

$$Y = 30,5Z_1 + 34,7Z_2 + 36,0Z_3 + 1,01X$$

vilket översatt till det andra skrivsättet blir

$$Y = 30,5 + 4,2Z_2 + 5,5Z_3 + 1,01X.$$

De tre intercepten förändras vid hopjämkningen.

Det finns många situationer, där ett observationsmaterial kan delas upp samtidigt efter två "leddar", t ex efter utbildning och efter kön. Man kan då samtidigt ansätta två grupper av indikatorvariabler, i exemplet en för utbildningskategori och en för kön. I sådana situationer är det bara det andra (osymmetriska) skrivsättet som medger beräkning på vanliga datorprogram för regression.

Då ett observationsmaterial delas upp i k delmängder med varsin regressionsekvation, kan man i princip alltid uttrycka alla dessa k ekvationer som en enda ekvation med hjälp av indikatorvariabelsteknik. Detta är emellertid ointressant om de k ekvationerna är helt olika. Av intresse är det då de till delmängd-

erna hörande ekvationerna är (dvs antages vara) i något avseende lika. Det gemensamma är oftast lutningen (lutningarna), men det finns andra varianter också. Indikatorvariabler är en mycket flexibel byggklots då man konstruerar (modell-)ekvationer.

Indikatorvariabler bör inte förväxlas med poängvariabler. Ibland delas ett observationsmaterial in i k delmängder, som har en naturlig inbördes ordning så att man kan numrera delmängderna 1, 2, ..., k. Man kan då införa en poängvariabel P sådan att P_j = h om observation j tillhör delmängd h enligt numreringen. En dylik poängvariabel kan tas med i en regressionsberäkning som förklaringsvariabel och ger då till resultat k parallella regressionslinjer (plan), en för varje delmängd. Poängvariabler är inte detsamma som indikatorvariabler, så ansatserna är inte likvärdiga.

Exempel 6.3 I exempel 6.2 kunde man numrera de tre delmängderna så att

grundskola motsvarar P = 1,
fackskola motsvarar P = 2 och
spec.utb motsvarar P = 3.

Anpassning av en regressionsekvation med P inkluderad, dvs

$$Y = A+FP+BX$$

ger då tre parallella räta linjer

$Y = (A+F) +BX$ för grundskola
$Y = (A+2F)+BX$ för fackskola
$Y = (A+3F)+BX$ för specialutbildade.

Ansatsen med poängvariabler är mycket mindre flexibel än ansatsen med indikatorvariabler. För det första fastlåser den ordningen mellan de k parallella linjerna; de måste komma i ordning efter växande P (eller efter avtagande P). För det andra låser den också fast linjernas inbördes avstånd på så sätt att avståndet i Y-riktning mellan L_3 och L_2 måste vara

lika med avståndet i Y-riktning mellan L_2 och L_1, och så vidare. De tre parallella linjerna i exempel 6.2 uppfyller inte det senare villkoret och måste därför vid poängvariabelsansats ersättas med tre andra linjer.

Vid osymmetrisk indiaktorvariabelsansats lägger man till X ytterligare k-1 nya förklaringsvariabler. (Symmetrisk indikatorvariabelsansats är likvärig med den osymmetriska.) Vid poängvariabelsansats lägger man bara till en ny förklaringsvariabel. Varje skara av k parallella räta linjer, som är möjlig inom ramen för poängvariabelsansatsen, är också möjlig inom ramen för dummyvariabelsansatsen. Men inte tvärtom!

6.3 Anpassning av parallella räta linjer

Antag ett observationsmaterial (X_i, Y_i) i = 1, ..., n. Antag att observationerna kan delas in i $k \geq 2$ delmängder. Antag att man vid enkel regression av Y på X finner ett visst samband. Antag att man finner följande resonemang tillämpligt.

Sambandet mellan X och Y beror på två saker. För det första finns det ett samband mellan X och Y inom varje delmängd av observationer. Dessa samband är inte nödvändigtvis parallella och lika starka. För det andra finns det också ett skenbart samband, vilket kommer till synes därför att delmängderna ligger så i (X, Y)-planet att Y ökar (minskar) med X då man går från en delmängd till en annan. Detta senare samband vill vi rensa bort och renodla sambandet (sambanden?) inom delmängderna.

Givet detta resonemang bör man anpassa en skiktad regression. En skiktad regression åskådliggöres av k stycken parallella linjer, en för var och en av delmängderna.

Skiktad regression beräknas oftast på dator. Skiktningen formuleras därvid i termer av indikatorvariabler. Det andra skrivsättet i avsnitt 6.1 ovan fungerar alltid, men man måste se till att tolka dess resultat rätt.

Exempel 6.4 Åtta observationer (X_i, Y_i) är givna. De indelas i två delmängder.

I		II	
X	Y	X	Y
3	1	9	6
5	4	10	9
9	4	12	7
11	7	13	10

Delmängdernas medelvärden $(\bar{X}, \bar{Y})$ är

I: (7,4) II: (11,8)

Det finns ett positivt samband mellan X och Y dels inom varje delmängd, dels p g a delmängdernas lägen i (X, Y)-planet.

Beräkningarna redovisas i figur 6.1.

Observationer och regressionslinjer återges i figur 6.2.

Exemplets data är så valda att de separata regressionerna inom skikten är parallella. Detta är i allmänhet inte fallet.

Den totala regressionen är

$$Y = -0,804+0,756X.$$

Den skiktade regressionen är en sammanfattning av de två parallella regressionerna inom skikt. I figur 6.1 skrivs den skiktade regressionen enligt det första skrivsättet, symmetriskt och utan gemensamt intercept. Alternativt kan den skrivas enligt det andra skrivsättet, med gemensamt intercept men osymmetriskt, till exempel som följer.

I: $Y = -0,2+0,6X$

II: $Y = -0,2+1,6+0,6X$

Den oskiktade regressionens lutning 0,756 återspeglar såväl de parallella linjernas lutning 0,6 som delmängdernas inbördes läge i planet.

Figur 6.1

Beräkningar till exempel 6.4.

			separat		separat		totalt		totalt	
	X	Y	X-X̄	Y-Ȳ	xy	x^2	X-X̄	Y-Ȳ	xy	x^2
I	3	1	-4	-3	+12	16	-6	-5	+30	36
	5	4	-2	0	0	4	-4	-2	+8	16
	9	4	+2	0	0	4	0	-2	0	0
	11	7	+4	+3	+12	16	+2	+1	+2	4
II	9	6	-2	-2	+4	4	0	0	0	0
	10	9	-1	+1	-1	1	+1	+3	+3	1
	12	7	+1	-1	-1	1	+3	+1	+3	9
	13	10	+2	+2	+4	4	+4	+4	+16	16
ΣI	28	16	0	0	+24	40				
m I	7	4								
ΣII	44	32	0	0	+6	10				
m II	11	8								
Σ	72	48					0	0	+62	82
m	9	6								

Separata regressioner

I: $Y = 4 + \frac{24}{40}(X-7)$; $Y = -0,2 + 0,6X$

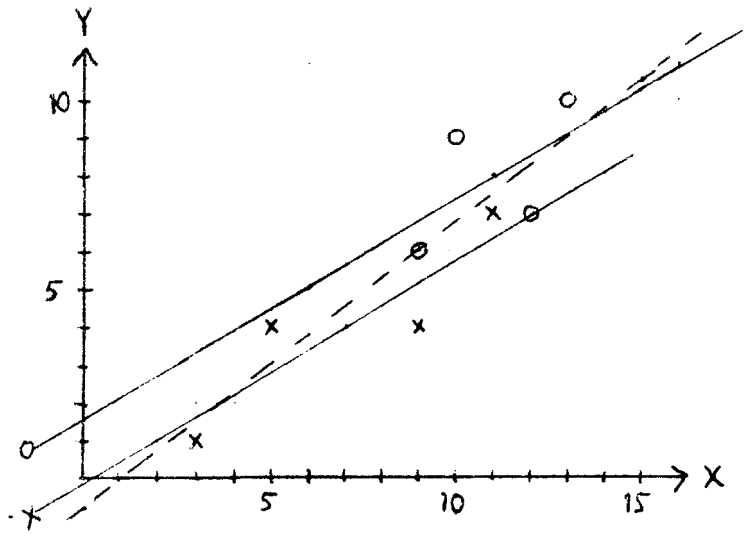
II: $Y = 8 + \frac{6}{10}(X-11)$; $Y = 1,4 + 0,6X$

Total regression (ej skiktad)

$Y = 6 + \frac{62}{82}(X-9)$; $Y = -0,804 + 0,756X$

Figur 6.2.

Observationer och regressionslinjer i exempel 6.4.



Separata: heldragna

Total: streckad

Den skiktade regressionen i exempel 6.4 har också beräknats medelst standardprogrammet SAS-GLM. Därvid har en indikatorvariabel använts, definierad så att Z=0 för delmängd I och Z=1 för delmängd II. Ingångsdata, närmevärden och residualer redovisas i datorutskriften i figur 6.3. Regressionskoefficienterna med mera redovisas i utskriften i figur 6.4.

Figur 6.3

In- och utdata i exempel 6.4

SKIKTAD REGRESSION

OBS	X	Z	Y	YH	E
1	3	0	1	1.6	-0.6
2	5	0	4	2.8	1.2
3	9	0	4	5.2	-1.2
4	11	0	7	6.4	0.6
5	9	1	6	6.8	-0.8
6	10	1	9	7.4	1.6
7	12	1	7	8.6	-1.6
8	13	1	10	9.2	0.8

Figur 6.4.

Utskrift från SAS-GIM, exempel 6.4.

SKIKTAD REGRESSION

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE
MODEL	2	50.00000000	25.00000000	12.50	0.0113	0.833333
ERROR	5	10.00000000	2.00000000		STD DEV	
CORRECTED TOTAL	7	60.00000000			1.41421356	

SOURCE	DF	TYPE II SS	F VALUE	PR > F
X	1	18.00000000	9.00	0.0301
Z	1	3.12195122	1.56	0.2668

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
INTERCEPT	-0.20000000	-0.13	0.9035	1.56843871
X	0.60000000	3.00	0.0301	0.20000000
Z	1.60000000	1.25	0.2668	1.28062485

OBSERVATION	OBSERVED VALUE	PREDICTED VALUE	RESIDUAL
1	1.00000000	1.60000000	-0.60000000
2	4.00000000	2.80000000	1.20000000
3	4.00000000	5.20000000	-1.20000000
4	7.00000000	6.40000000	0.60000000
5	6.00000000	6.80000000	-0.80000000
6	9.00000000	7.40000000	1.60000000
7	7.00000000	8.60000000	-1.60000000
8	10.00000000	9.20000000	0.80000000

SUM OF RESIDUALS	-0.00000000
SUM OF SQUARED RESIDUALS	10.00000000
SUM OF SQUARED RESIDUALS - ERROR SS	-0.00000000
FIRST ORDER AUTOCORRELATION	-0.84800000
DURBIN-WATSON D	3.59600000

6.4 Skiktningens effekter

Betrakta ett observationsmaterial $(X_j\ Y_j)$ $j = 1, \ldots, n$. Regressionen av Y på X är

$$Y = A_0 + B_0 X$$

och determinationskoefficienten är R_0^2.

Dela upp materialet i k delmängder och definiera k-1 däremot svarande indikatorvariabler $(Z_2\ Z_3\ \ldots\ Z_k)$. Regressionen av Y på indikatorvariablerna och X är

$$Y = A + D_2 Z_2 + \ldots + D_k Z_k + BX$$

och determinationskoefficienten är R^2.

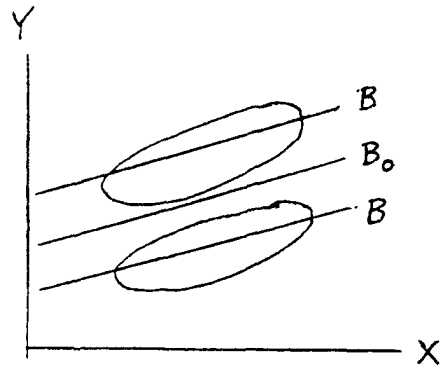
Man kan visa att $R^2 \geq R_0^2$. Vidare blir i allmänhet $B \neq B_0$. Regressionskoefficienten för den "riktiga" första förklaringsvariabeln X ändras då indikatorvariablerna (dvs uppdelningen i delmängder) införes. Varför?

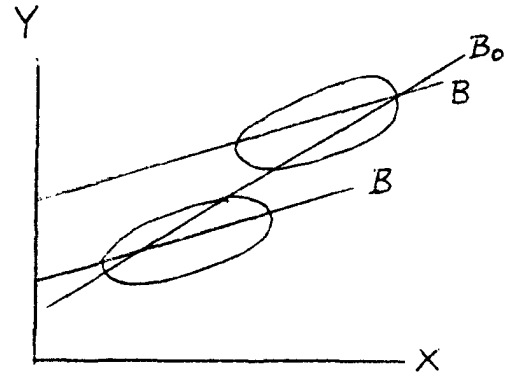
Det som gäller om regressionskoefficienten för X då man inför en ny förklaringsvariabel Z gäller också här. Formellt inför vi k-1 nya förklaringsvariabler. I den mån dessa har "effekt" på Y och i den mån indelningen är "korrelerad" med X så representerar den enkla regressionskoefficienten B_0 delvis också indelningseffekten. Och då ändras B_0 till B när indikatorvariablerna införes; X-effekten renodlas.

Vi har här talat om en indelning och dess effekter på samma sätt som man talar om en ny förklaringsvariabel Z och dess effekter. Detta är tillåtet. En indelning i skikt är korrelerad med X om skiktens X-medelvärden är olika, helt enkelt. Att denna "korrelation" medför att den oskiktade regressionen kan bli missvisande (visavi de parallella linjerna inom skikten) antydes av figur 6.5. Missvisningen behöver inte bli positiv utan kan även bli negativ.

Figur 6.5

Effekten av skiktning





7 STATISTISK INFERENS AVSEENDE REGRESSIONSKOEFFICIENTER

7.1 Modellterminologi och standardantaganden

I detta kapitel behandlas två standardförfaranden för statistisk inferens vid linjärregressionsanalys med en eller flera förklaringsvariabler. Förfarandena bygger på samtliga de statistiska förutsättningar, som nämndes i avsnitt 3.3 ovan. Ungefär den terminologi, som användes nedan, är gängse.

Betrakta först fallet med en enda förklaringsvariabel. Data antages ha genererats i enlighet med den statistiska modellen (för $i=1,\ldots,n$)

$$
\begin{aligned}
&Y_i = \alpha + \beta X_i + \varepsilon_i \\
&E(\varepsilon_i) = 0 \\
&Var(\varepsilon_i) = \sigma^2 \\
&Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ då } i \neq j
\end{aligned}
$$

där X_i är kända konstanter , (α,β,σ^2) är tre okända konstanter , parametrar, och ε_i är icke direkt observerbara slumpvariabler. Härav följer att variablerna Y_i är slumpvariabler. De observerade Y_i-värdena betraktas som utfall av motsvarande slumpvariabler.

I fallet med två förklaringsvariabler är modellekvationen

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$$

och det finns fyra okända parametrar $(\alpha,\beta,\gamma,\sigma^2)$. Och så vidare analogt vid flera förklaringsvariabler.

Slumpvariablerna ε_i antages vara normalfördelade. Denna förutsättning är viktig för inferensteorin.

Det är inte gängse att formulera om modellekvationen så att den svarar mot regressionsberäkningar i avvikelseform.

(Jfr avsnitt 1.3 ovan.) Men det är fullt möjligt att göra det. Modellekvationen blir då

$$Y_i = \mu + \beta(X_i - \bar{X}) + \varepsilon_i$$
$$\text{där } \mu = \alpha + \beta\bar{X}.$$

Konstanten $\bar{X}$ är känd. Den är (i praktiken alltid) medelvärdet av de observerade X_i-värdena, som ju betraktas som kända konstanter.

Statistisk inferens går ut på att estimera och/eller pröva hypoteser om modellens parametrar, i första hand de mot regressionskoefficienterna svarande parametrarna β,γ och så vidare. Två standardförfaranden kommer att behandlas. Det ena är intervallestimation av β. Det andra är prövning av hypotesen att β eller, vid flera förklaringsvariabler, en viss delmängd av parametrarna, är noll. Intervallestimation av β kompletterar punktskattningen B med ett osäkerhetsmått. Hypotesprövning av $\beta=0$ prövar hypotesen att den till β hörande förklaringsvariabeln är irrelevant.

Inferensteorin kan byggas ut så att den också omfattar prognoser av Y-värden för nya observationer. Man måste i så fall förutsätta att den vid analysen av de givna data ansatta modellen gäller också för de nya observationerna. (Jfr avsnitt 4.4 ovan.) Prognoser behandlas inte vidare här.

Den statistiska teori, som inferensen utgår ifrån, kännetecknas av betydande assymetri mellan å ena sidan förklaringsvariablerna (X,Z,...) och å andra sidan analysvariabeln Y. (Jfr avsnitt 1.4 ovan.) Förklaringsvariablerna betraktas som icke slumpmässiga observerade konstanter. Analysvariabeln betraktas som en slumpvariabel. För varje observation finns ett observerat utfallsvärde. På grund av förekomsten av slumpelementet ε_i ger data en "störd" bild av modellekvationen. Den statistiska inferensen försöker genomskåda störningen.

7.2 Medelfelet för en regressionskoefficient

Betrakta först fallet med en enda förklaringsvariabel. Regressionskoefficienten

$$B = \Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(X-\bar{X})^2$$

är en slumpvariabel, och dess beräknade värde är ett utfall. Regressionskoefficienten är en linjär funktion av slumpvariablerna Y_i med vikter bestämda av de kända konstanterna X_i. Eftersom varianser och kovarianser för Y_i är kända, kan variansen för B härledas.

Ur de statistiska standardförutsättningarna i avsnitt 3.3 ovan följer två egenskaper hos slumpvariabeln B.

$$E(B) = \beta$$
$$Var(B) = \sigma^2/\Sigma(X-\bar{X})^2$$

Väntevärdet för regressionskoefficienten B är lika med motsvarande parameter β. Regressionskoefficienten B är således en unbiased estimator av parametern β.

Variansen för B är en känd konstant gånger den okända modellvariansen σ^2. Den senare kan estimeras med hjälp av regressionens residualkvadratsumma. Den gängse unbiased estimatorn är

$$\hat{\sigma}^2 = \Sigma E^2/(n-2)$$

Om man sätter in den i formeln för Var(B) erhåller man

$$\text{est var}(B) = \left[\Sigma E^2/(n-2)\right]/\ \Sigma(X-\bar{X})^2$$

Kvadratroten ur est var(B) är medelfelet för B.

Givet att slumptermerna ε_i har antagits vara normalfördelade, så är också regressionskoefficienten B normalfördelad.

I fallet med två förklaringsvariabler X och Z gäller följande generaliseringar, där β är parametern för X.

$$E(B_{YX\cdot Z}) = \beta$$
$$Var(B_{YX\cdot Z}) = \sigma^2 / \Sigma E^2_{X\cdot Z}$$

Med $E_{X\cdot Z}$ menas här residualen av X efter (deskriptiv) regression på Z i det givna datamaterialet. Vidare gäller att skattningen för modellvariansen är

$$\hat{\sigma}^2 = \Sigma E^2_{Y\cdot XZ}/(n-3)$$

vilket insatt i variansuttrycket ger kvadraten på medelfelet för regressionskoefficienten $B_{YX\cdot Z}$.

I fallet med k≥3 förklaringsvariabler gäller följande. Låt Z beteckna vektorn av alla förklaringsvaribler utom X. Då är $E(B_{YX\cdot Z})$ och $Var(B_{YX\cdot Z})$ som i fallet med två förklaringsvariabler, medan

$$\hat{\sigma}^2 = \Sigma E^2_{Y\cdot XZ}/(n-k-1)$$

(Fallet k=2 ingår egentligen också i detta allmänna fall.)

Standardprogram för regressionsanalys beräknar och trycker medelfelet för varje regressionskoefficient. Se för fallet med en förklaringsvariabel figur 3.5 ram 4! Se för fallet med två förklaringsvariabler figur 6.4 under rubriken "Std error of estimate".

Standardprogrammet kan givetvis inte veta om de nödvändiga statistiska modellförutsättningar är uppfyllda, utan vilka det beräknade medelfelet inte är tolkbart. Medelfelet kan alltid beräknas, tolkbart eller ej.

Medelfelet användes till att tillsammans med punktskattningen definiera ett intervallestimat för β. Intervallestimatorn är intervallet

$$B \pm c\sqrt{\text{est var}(B)}$$

där c är konstant, som beror dels på antalet observationer och förklaringsvariabler, dels på vilken konfidensgrad man önskar. Med 95% konfidensgrad och många fler observationer än variabler är intervallestimatet cirka

$$B \pm 2\sqrt{\text{est var}(B)}$$

Exempel 7.1

Betrakta åter exempel 6.4. Beräkna ett c:a 95% konfidensintervall för parametern β, den modellkonstant som hör till förklaringsvariabeln X. Konstanten c skall enligt tabell här vara inte 2 utan 2,57.

Enligt figur 6.4 är $B_{YX \cdot Z}$=0,600 och dess medelfel 0,200. Konfidensintervallet blir

$$0.600 \pm 2{,}57 \cdot 0{,}200$$

dvs intervallet (0,086-1,114)

Pga det ringa antalet observationer är intervallet brett. Punktskattningens osäkerhet är betydande.

7.3 Hypotesen att en delmängd av parametrarna är noll.

Tre fall kommer att behandlas:
(1) det finns en enda förklaringsvariabel (specialfall av följande fall),
(2) det finns en eller flera förklaringsvariabler, och man vill pröva hypotesen att samtliga variablers parametrar är noll,
(3) det finns två eller flera förklaringsvariabler, och man vill pröva hypotesen att parametrarna för en äkta delmängd av variablerna är noll.

I samtliga fall härleder statistisk teori, givet samtliga

statistiska förutsättningar i avsnitt 3.3 ovan, en testkvantitet F, vars fördelning är känd. Det beräknade F-värdet jämförs med ett tabellvärde, som beror dels på antalen observationer och variabler, dels på vilken nivå man vill testa på. Testets nivå är (som alltid) den risk man tar att förkasta en sann nollhypotes.

Betrakta fallet med en enda förklaringsvariabel. Regressionsanalysen delar upp den totala avvikelsekvadratsumman för Y i en förklarad kvadratsumma SSR(Y;X) och en residualkvadratsumma SSE(Y;X). Den hypotes, som testas, är att $\beta=0$. Givet denna hypotes är SSR och SSE efter division med respektive antal frihetsgrader av jämförbar storlek. Om hypotesen är felaktig, tenderar SSR att vara större.

Testkvantiteten är

$$F = \frac{SSR(Y;X)}{SSE(Y;X)/(n-2)}$$

Den ska jämföras med F-fördelningen med 1 och (n-2) frihetsgrader.

Exempel 7.2

Betrakta åter exempel 3.1 i avsnitt 3.1 och tillhörande datorberäkning i figur 3.5 sist i kapitel 3. Se ram 1 i den figuren! Testkvantiteten är

$$F = \frac{3,321}{22,539/8} = \frac{3,321}{2,817} = 1,18$$

F-värdet 1,18 återfinnes också i programutskriften. Det ska jämföras med det motsvarande tabellvärdet för (1,8) frihetsgrader och, låt säga, 5% risk. Detta värde är 5,32. Eftersom det beräknade F är mindre, kan vi inte förkasta hypotesen att $\beta=0$. Jfr exempel 3.2!

Betrakta nu fallet med $k \geq 2$ förklaringsvariabler och hypotesen att samtliga variablers parametrar är noll.

Testkvantiteten är

$$F = \frac{SSR(Y;X,Z,...)/k}{SSE(Y;X,Z,...)/(n-k-1)}$$

Den ska jämföras med F-fördelningen med (k, n-k-1) frihetsgrader.

Exempel 7.3

Betrakta exempel 6.4 och figur 6.4! Se kvadratsummeuppdelningen överst i datorns utskrift.

$$F = \frac{50/2}{10/5} = \frac{25}{2} = 12,5$$

F-värdet 12,5 finns även i utskriften. Tabellvärdet för 5% signifikansnivå och (2, 5) frihetsgrader är 5,79. Eftersom det beräknade F-värdet är större, förkastar vi hypotesen att parametrarna för X och Z båda är noll. Observera att detta test inte säger något om huruvida det är X eller Z eller båda, som har signifikant inverkan på Y.

Betrakta återigen fallet med $k \geq 2$. Vi ifrågasätter inte $p \geq 1$ förklaringsvaribler X men vill pröva hypotesen att de $q \geq 1$ övriga förklaringsvariablerna Z har parametrar som alla är noll. (Symbolerna X och Z betecknar här och i fortsättningen av detta avsnitt vektorer av variabler.)

I detta fall ska man med residualkvadratsumman jämföra den ökning av den förklarade kvadratsumman, som erhålles, om mängden av i regressionen använda förklaringsvariabler utökas från X till (X,Z). Ökningen är

$$SSR(Y;X,Z) - SSR(Y;X)$$

och kan alternativt beräknas som

$$SSE(Y;X) - SSE(Y;X,Z).$$

Residualkvadratsumman, som mäter den rena slumpvariationen

i modellen oavsett om hypotesen är sann eller ej, är

$$SSE(Y;X,Z).$$

Man ska återigen dividera kvadratsummorna med relevanta frihetsgradsantal. Testkvantiteterna är

$$F = \frac{[SSR(Y;X,Z)-SSR(Y;X)]/q}{SSE(Y;X,Z)/(n-p-q-1)}$$

där q är antalet testade parametrar.

Testförfarandet är tillämpligt även då några av variablerna är indikatorvariabler, förutsatt att dessa är linjärt oberoende av varandra och av alla andra variabler i modellen. Man måste arbeta i vad som i avsnitt 6.1 kallades det andra skrivsättet, så som skedde i beräkningarna i figurerna 6.3 och 6.4.

Exempel 7.4

Data:	X	Z	Y
	1	1	1
	1	3	3
	1	5	5
	3	2	3
	3	4	6
	3	6	9
	5	3	8
	5	5	9
	5	7	10

Vi vill testa hypotesen att Z är irrelevant. Först beräknas regressionen av Y på enbart X. Se figur 7.1! SSR(Y;X)=54. Sedan beräknas regressionen av Y på X och Z. Se figur 7.2! SSR(Y;X,Z)=78 och SSE(Y;X,Z)=4. Frihetsgraderna är q=1 och n-p-q-1=9-1-1-1=6.

$$F = \frac{(78-54)/1}{4/6} = \frac{24}{4/6} = 36$$

Kritiskt 5% värde enligt tabell är 5,99. Hypotesen förkastas. Variabeln Z måste anses ha påvisbar inverkan på Y.

Testkvantiteten F kan alltid beräknas, oavsett om förutsättningar föreligger för att den ska kunna tolkas eller ej. Glöm inte de statistiska och andra förutsättningarna!

Figur 7.1.

Datorutskrift del 1 till exempel 7.4.

FTEST STEG 1 8:55 TUESDAY, APRIL 27, 1982 1

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	54.00000000	54.00000000	13.50	0.0079	0.658537	33.3333
ERROR	7	28.00000000	4.00000000		STD DEV		Y MEAN
CORRECTED TOTAL	8	82.00000000			2.00000000		6.00000000

SOURCE	DF	TYPE II SS	F VALUE	PR > F
X	1	54.00000000	13.50	0.0079

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
INTERCEPT	1.50000000	1.08	0.3177	1.39443338
X	1.50000000	3.67	0.0079	0.40824829

FTEST STEG 1 8:55 TUESDAY, APRIL 27, 1982 2

OBS	X	Z	Y	Y1	E1
1	1	1	1	3	-2
2	1	3	3	3	0
3	1	5	5	3	2
4	3	2	3	6	-3
5	3	4	6	6	0
6	3	6	9	6	3
7	5	3	8	9	-1
8	5	5	9	9	0
9	5	7	10	9	1

Figur 7.2.

Datorutskrift del 2 till exempel 7.4.

FTEST STEG 2 8:55 TUESDAY, APRIL 27, 1982 3

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	78.00000000	39.00000000	58.50	0.0001	0.951220	13.6083
ERROR	6	4.00000000	0.66666667		STD DEV		Y MEAN
CORRECTED TOTAL	8	82.00000000			0.81649658		6.00000000

SOURCE	DF	TYPE II SS	F VALUE	PR > F
X	1	19.20000000	28.80	0.0017
Z	1	24.00000000	36.00	0.0010

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
INTERCEPT	-1.00000000	-1.42	0.2061	0.70546806
X	1.00000000	5.37	0.0017	0.18633900
Z	1.00000000	6.00	0.0010	0.16666667

FTEST STEG 2 8:55 TUESDAY, APRIL 27, 1982 4

OBS	X	Z	Y	Y2	E2
1	1	1	1	1	0
2	1	3	3	3	0
3	1	5	5	5	0
4	3	2	3	4	-1
5	3	4	6	6	0
6	3	6	9	8	1
7	5	3	8	7	1
8	5	5	9	9	0
9	5	7	10	11	-1

7.4 Något om stegvis regression

Stegvis regression är ett förfarande där datorn automatiskt väljer och vrakar bland samtliga förklaringsvariabler tills den hittar den delmängd av förklaringsvariabler, som är bäst enligt de i programmeringen nedlagda kriterierna. Förfarandet kännetecknas av att datorn i varje steg antingen tar in en ny förklaringsvariabel i regressionen eller utesluter en tidigare medtagen förklaringsvariabel ur regressionen.

Standardprogrampaketet SAS innehåller ett program för stegvis regression, som heter STEPWISE. Det finns många liknande datorprogram.

Datorn kan alltid utföra beräkningarna och genomföra en stegvis regression.Datorn kan däremot inte pröva huruvida förfarandet är meningsfullt. Jfr kapitel 3 ovan!

Då nya förklaringsvariabler tas in i ekvationen eller gamla utesluts, ändras de bibehållna förklaringsvariablernas regressionskoefficienter. Jfr kapitel 5 ovan! De bibehållna förklaringsvariablernas regressionskoefficienters medelfel ändras också. Se medelfelsformeln i avsnitt 7.2 ovan!

Stegvis regression är inte ett sätt att medelst statistisk hypotesprövning testa fram rätt uppsättning förklaringsvariabler. Förfarandet kan uppfattas som baserat på statistisk hypotesövning i varje steg, men stegens beroende av varandra gör det alltför invecklat att utreda vilka statistiska egenskaper förfarandet som helhet egentligen har.

Sedan denna reservation har uttalats, ska nu en ofta förekommande strategi för stegvis regression skisseras. Det finns även andra strategier.

1 Fastställ en signifikansivå p_i för inklusion. En förklaringsvariabel tas in i ekvationen i ett visst steg endast om den är signifikant på denna nivå.

2 Fastställ en signifikansnivå p_e för exklusion. En förklaringsvariabel utesluts ur ekvationen i ett visst steg om den inte är signifikant på denna nivå.

3 Bestäm vilka variabler som eventuellt alltid ska vara med i ekvationen. Dessa variablers signifikans prövas inte.

4 Välj den icke inkluderade förklaringsvariabel som, om den inkluderas, mest ökar determinationskoefficienten R^2! Om den är signifikant på nivå p_i och inte just uteslöts i föregående steg, inkluderas den. Gå i annat fall till punkt 7.

5 Uteslut samtliga förklaringsvariabler, som eventuellt inte är signifikanta på nivå p_e.

6 Gå tillbaka till punkt 4.

7 Avbryt körningen.

Ovanstående är en av strategierna i SAS-STEPWISE.

Automatisk stegvis regression torde vara en av de mest brukade och missbrukade statistiska metoderna. Använd den inte mekaniskt utan med eftertanke!

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys.
Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller.
Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognoser av tidsserier.
Åke Holmén (1979-12-20)

4 Handledning i AID-analys.
Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport.
P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt
Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population.
Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling.
Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach.
Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden.
Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata.
Lars-Otto Sjöberg (1984-01-11)

Kvarvarande exemplar av ovanstående nummer kan rekvireras från SCB, P/STM, Elseliv Lindfors, 115 81, eller per telefon 08 14 05 60 ankn 4178.

Garnisonstryckeriet ★ Stockholm ★ 1984