Föreläsning 8 och 9: Regressionsanalys

Pär Nyman

1 februari 2016

Det här är anteckningar till föreläsning 8 och 9. Båda föreläsningarna handlar om regressionsanalys, så jag slog ihop dem till ett gemensamt dokument och gjorde samma sak med bilderna till presentationen.

1 Regressionsanalys

Regressionsanalysen är den kvantitativa samhällsvetenskapens i särklass viktigaste verktyg. Oavsett om man skriver en kvantitativ uppsats eller läser en bra forskningsartikel med mycket statistik är det oftast regressionsanalys som används. Det finns hur mycket som helst man kan lära sig om att göra regressioner, men man kommer ganska långt med de grunder vi lär ut på den här kursen. Eftersom regressionsanalys är så pass vanligt vill jag verkligen rekommendera även de som är mer lagda åt kvalitativ forskning att lära sig grunderna. Ni kommer alla att stöta på många regressionsanalyser i era framtida studier och arbetsliv. För att kunna bedöma trovärdigheten i deras resultat är det viktigt att ni förstår vad de har gjort och var svagheterna ligger i deras analys.

Med lite fantasi kan man besvara de flesta typer av frågor med regressionsanalys, men i regel studerar vi kausala samband sådana att värdet på en variabel påverkar värdet på en annan variabel.

Inledningen på dagens första föreläsning är till stor del repetition av den föregående föreläsningen. Anledningen till det är att regressionsanalys är nytt för de flesta av er och ni brukar också uppskatta att vi repeterar just de här bitarna.

1.1 Regressionsekvationen

Att genomföra en regressionsanalys är samma sak som att skatta de olika parametrarna i regressionsekvationen. Om ni minns ”den räta linjens ekvation” ($y = kx + m$) från gymnasiematematiken, så är det här i grunden samma sak. Vi tänker oss att den beroende variabeln (y) är en linjär funktion av en eller flera oberoende variabler (x), så att när x ökar med 1 så förändras y med b. Genom att skatta värdet av interceptet (a) och regressionskoefficienten (b)

kan vi både beskriva sambandet mellan variablerna och göra prediktioner av den beroende variabeln. Regressionsekvationen, vilken även anger våra prediktioner, är dock sällan en perfekt beskrivning av verkligheten. Även om vi känner till värdet på den oberoende variabeln (x) kommer de observerade värdena på den beroende variabeln (y) skilja sig från modellens prediktioner. Vi säger att observationen består av en prediktion ($a + bx$) plus en felterm eller residual (e). Regressionsekvationen kan således skrivas

$$\begin{aligned} y &= a + bx + e \\ \hat{y} &= a + bx \end{aligned} \tag{1}$$

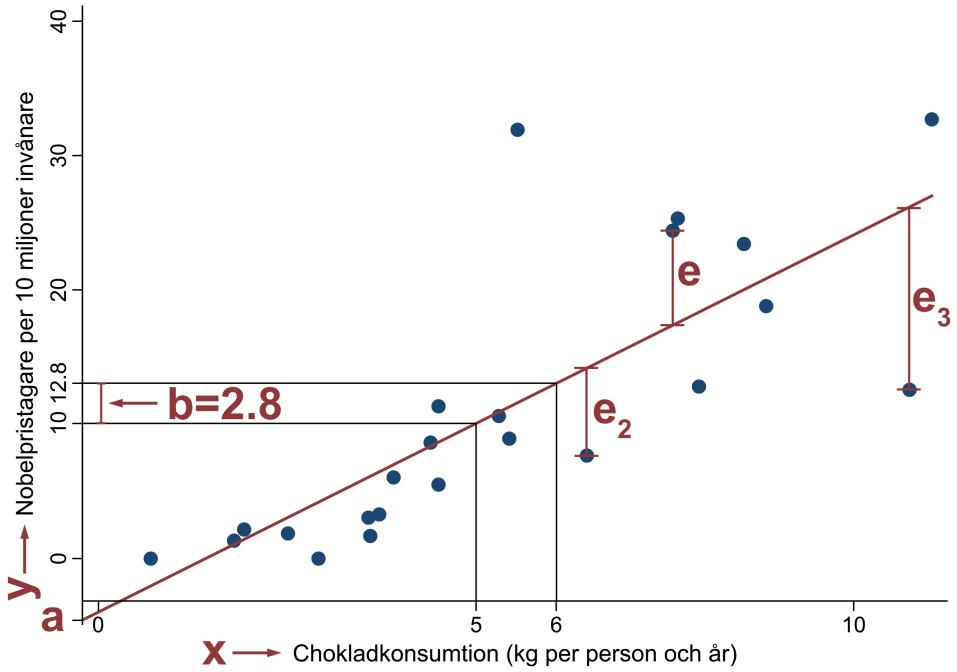
där

y = Beroende variabel
a = Konstant eller intercept
b = Regressionskoefficient
x = Oberoende variabel
e = Felterm eller residual

Ibland skriver man en ekvation för $\hat{y}$ i stället för y. Hatten över y står för att det är en prediktion, en gissning. Eftersom våra prediktioner är identiska med vår regressionsmodell behöver vi inte inkludera någon felterm i ekvationen. Feltermen inkluderas bara när vi har faktiska observationer till vänster om likamedtecknet.

Vi kan också illustrera allt detta grafiskt. Figuren här är hämtad från en kolumn i *The New England Journal of Medicine* och visar den genomsnittliga årskonsumtionen av choklad samt andelen Nobelpristagare för ett urval av länder. De blå prickarna är alltså länder. Som synes har länder med hög chokladkonsumtion (ligger långt åt höger i figuren) också en högre andel Nobelpristagare (är placerade högt upp i figuren). Detta innebär förstås inte att sambandet är kausalt (att chokladkonsumtionen påverkar andelen Nobelpristagare), även om det är vad artikelförfattarna något skämtsamt argumenterar för.

Y är värdet på den beroende variabeln, alltså var observationerna placerar sig på den vertikala y-axeln. X är värdet på de oberoende variablerna och således var observationerna placeras längs den horisontella x-axeln. När $x = 0$ gäller att $\hat{y} = a$. Konstanten är därför det värde vid vilket regressionslinjen skär y-axeln, alltså vår prediktion av y när värdet på x är 0. Regressionskoefficienten är den förväntade förändringen av y när x ökar med 1. I figuren illustreras det med att när chokladkonsumtionen ökar från 5 till 6 kg så stiger antalet Nobelpristagare (per 10 milj invånare) från 10 till 12.8. Residualerna utgörs av varje observations avvikelse från regressionslinjen.



1.2 Passningsmått

När man analyserar olika typer av samband kan det ofta vara av värde att skilja mellan storleken på en kausal effekt och modellens passning. Med en effekts storlek menas hur stor påverkan en viss förändring i den oberoende variabeln (x) har på den beroende variabeln (y), alltså hur stor regressionskoefficienten är. Med modellens passning avses hur väl regressionsmodellen sammanfattar sambandet mellan oberoende och beroende variabel. Intuitivt kan vi förstå en modells passning på följande sätt: ju mindre avståndet är mellan de faktiska observationerna och regressionslinjen, desto bättre är modellens passning. Passning kallas ibland förklaringskraft.

Passningen mäts med olika passningsmått, vilka anger hur väl vår modell beskriver den data vi har observerat. God passning innebär att observationerna ligger nära regressionslinjen, vilket är samma sak som att vi kan göra bra prediktioner med hjälp av vår modell om vi känner till värdena på de oberoende variablerna. Dålig passning innebär att observationerna ligger långt bort från regressionslinjen, så att vi i genomsnitt gör stora fel när vi med hjälp av vår modell försöker gissa värdena på den beroende variabeln.

De två viktigaste passningsmåtten är regressionens standardfel och R^2. Båda dessa mått mäter i grunden samma sak – storleken på residualerna – men uttrycker passningen på olika skalor. Vilket mått som lämpar sig bäst beror helt på vilket syfte man har och vad man vill jämföra med. Ofta är det en bra idé att ange båda måtten.

Standardfelet anger hur långt ifrån regressionslinjen som residualerna

ligger ”i genomsnitt” och är alltid mätt i samma enhet som den beroende variabeln. R^2 kan anta värden mellan 0 och 1 och anger hur stor andel av variationen i den beroende variabeln som vår modell kan förklara, alltså hur mycket vår modell bidrar till att minska residualerna. Nedan följer en lite mer utförlig beskrivning av de båda passningsmåtten.

Regressionens standardfel är *nästan* samma sak som den genomsnittliga avvikelsen från regressionslinjen. Även om det inte är exakt samma sak, är det ofta så man tolkar måttet. Således är det också helt ok för er att uttrycka er så. För att återknyta till exemplet med Nobelpris så var standardfelet i den regressionen 6,6. Vi uttrycker det som att de observerade värdena i genomsnitt avviker från modellens prediktioner med 6,6 Nobelpristagare per 10 miljoner invånare. Precis som med regressionskoefficienterna måste vi alltid bedöma standardfelet i förhållande till skalan. Hade den beroende variabeln varit mätt som antal Nobelpristagare per invånare – i stället för per 10 miljoner invånare – hade standardfelet blivit 0.00000066, men passningen hade fortfarande varit lika bra.

Ni behöver inte känna till eller förstå ekvationen för regressionens standardfel, men en del lär sig ändå mycket av att se den. RSS står för summan av de kvadrerade feltermerna (Residual Sum of Squares), n för antalet observationer och k för antalet oberoende variabler (så att att $n-1-k$ blir antalet frihetsgrader). Ekvationen för regressionsstandardfelet påminner mycket om hur man beräknar en standardavvikelse, men här är vi intresserade av avvikelsen från regressionslinjen $(y_i - \hat{y})$ i stället för avvikelsen från medelvärdet $(y_i - \bar{y})$.

$$\text{Standardfel} = \sqrt{\frac{RSS}{n-1-k}} = \sqrt{\frac{\sum(e_i^2)}{n-1-k}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-1-k}} \qquad (2)$$

Det vanligaste passningsmåttet är R^2, vilket brukar beskrivas som andelen förklarad variation i den beroende variabeln. R^2 kan anta värden mellan 0 (vår modell förklarar ingenting) och 1 (vår modell förklarar 100 procent av variationen i den beroende variabeln). För att återigen använda vårt tidigare exempel, så var R^2 då 0,6. Detta uttrycker vi som att skillnader i chokladkonsumtion kan ”förklara” 60 procent av variationen mellan länder i antalet Nobelpristagare. Anledningen till att jag skriver förklara med citationstecken är att endast en samvariation mellan två variabler sällan betraktas som en fullgod förklaring. Exempelvis har vi inte isolerat sambandet från andra möjliga förklaringar eller funderat särskilt mycket på vilken mekanism som skulle kunna ge upphov till detta samband.

På samma sätt som med standardfelet behöver ni inte känna till eller förstå ekvationen för R^2. TSS står för summan av avvikelserna från medelvärdet (Total Sum of Squares), vilket är samma sak som den totala variationen i den beroende variabeln. För att skapa en intuitiv förståelse för ekvationen kan man tänka ungefär såhär: Ju mindre residualerna är i förhållande till

den totala variationen (RSS är mycket mindre än TSS), desto större andel av variationen kan vi förklara med hjälp av vår modell.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{3}$$

När vi tolkar passningsmåtten är det bra att komma ihåg följande tre saker.

- Vad som är högt och lågt beror som alltid på vad vi har att jämföra med. Studenter har ofta orimligt höga förväntningar på vad våra modeller kan åstadkomma.
- Stirra er inte blinda på passningsmåtten. Vårt mål är sällan att göra de bästa prediktionerna. Vanligare att vi är intresserade av ett kausalt samband.
- Då är det viktigare hur stor effekten är samt huruvida den är statistiskt signifikant, alltså om samvariationen i vårt urval kan bero på slumpen.

När man adderar en variabel till en regressionsmodell kommer R^2 alltid att öka, även om den inte har något med den beroende variabeln att göra. För att korrigera för detta bör man i regel använda ett mått som kallas för justerat R^2. Skillnaden är att justerat R^2 innehåller en korrigering för antalet variabler i modellen (i förhållande till antalet observationer). Justerat R^2 stiger endast om man adderar en variabel som bidrar mer till modellen än vad vi skulle förvänta oss om variabeln inte hade något samband med övriga variabler. Det är vanligt att även justerat R^2 uttrycks som andel av variationen i den beroende variabeln som modellen förklarar. Även om det inte är helt korrekt är det ok för er att tolka båda måtten så. Vill ni vara mer korrekta kan ni lägga till ”justerat för antalet frihetsgrader”.

2 Statistisk signifikans

En fråga har säkert flera av er redan har funderat över: Hur kan vi veta att de regressionsresultat vi finner i urvalet också gäller i den totala populationen som vi i slutändan vill uttala oss om? Hur vet vi exempelvis att den positiva samvariation mellan ideologisk position och inställningen till behovet av ökad jämställdhet mellan könen som vi fann i vårt urval också gäller bland alla röstberättigade svenskar? Med andra ord, hur generaliserar man sina kausala resonemang?

På samma sätt som medelvärdet i ett urval kan användas för att uppskatta medelvärdet i den större populationen kan vi använda regressionskoefficienten i ett urval som en uppskattning av populationens regressionskoefficient. I båda fallen gör vi det dock med ett visst mått av osäkerhet: populationsvärdet är förmodligen något lägre eller något högre än värdet från vårt urval. När vi genomför en regressionsanalys är vi i regel intresserade av om sambandet är

tillräckligt starkt för att vi ska kunna avfärda möjligheten att populationens regressionskoefficient (β) är noll. Om vi kan förkasta möjligheten att $\beta = 0$ så säger vi att sambandet är statististiskt signifikant, vilket är samma sak som att det finns ett samband också i populationen.

Samma resonemang kan användas även om vi inte är intresserade av någon väldefinierad population. De flesta hypoteser vi testar handlar egentligen inte om huruvida det existerar en samvariation i ett urval eller en population, utan snarare om de processer som har orsakat den samvariationen. Kan vi vara säkra på att män tjänar mer än kvinnor på grund av sitt kön, eller kan det vara en slump att männen i vårt urval råkade ha högre inkomster än kvinnorna? Kan det vara en tillfällighet att demokratier är inblandade i färre krig än icke-demokratier, eller måste det vara så att demokratisering minskar risken för väpnad konflikt? Dessa frågor kan omformuleras till om sambandet är statistiskt signifikant. Är vi säkra på att det finns en effekt ($\beta \neq 0$), eller kan samvariationen bero på slumpen?

* * *

Nedan kommer jag i huvudsak prata om generalisering som om den rörde en väldefinierad population, men vi kan lika gärna tänka oss att vi vill generalisera till de mer abstrakta processer som skapat den värld vi observerar. I båda fallen ställer vi oss frågan om huruvida de samband vi observerar i vårt urval kan bero på slumpen. Framställningen skiljer sig en del från hur jag beskrev statistisk signifikans på föreläsningen, men det är egentligen inga skillnader i sak. Föredrar ni föreläsningsversionen så finns den dokumenterad på mina slides.

Ett ofta förekommande fall är att vi är intresserade av att ta reda på om det finns ett samband mellan två variabler x och y i populationen. Men då vi av tids- och resursskäl inte kan undersöka hela populationen tvingas vi nästan alltid arbeta med urval. Grundlogiken skiljer sig inte från hur man gör detta i det univariata fallet, alltså när man vill generalisera beskrivningar. Om det finns ett samband mellan x och y i populationen ska regressionskoefficienten i populationen, som Teorell och Svensson benämner β (för att helt enkelt skilja det från b-värdet i vårt urval), vara skilt från noll. Om β är större än noll har vi ett positivt samband och om β är mindre än noll har vi ett negativt samband. Formellt ställer vi därför ofta upp följande hypoteser om sambandet i populationen som vi vill testa på basis av resultaten i vårt stickprov:

$$H_0 : \beta = 0$$
$$H_{alt} : \beta \neq 0$$

Det enklaste sättet att testa dessa hypoteser är att beräkna ett konfidensintervall kring vår urvalskoefficient (b) och se om detta konfidensintervall

täcker in 0 eller inte. Om intervallet täcker in 0 kan vi inte förkasta H_0 vilket innebär att vi inte vågar tro på att det samband mellan x och y vi ser i vårt urval verkligen också återfinns i populationen. Det är helt enkelt inte statistiskt säkerställt att β skiljer sig från 0. Om konfidensintervallet däremot inte täcker in 0 förkastar vi H_0 och vi säger därmed att vi vågar dra slutsatsen att sambandet som vi har funnit i urvalet också gäller i populationen (effekten är så stor att vi håller det för ytterst osannolikt att värdet i populationen verkligen är 0). Vi säger då att regressionskoeffcienten är statistiskt signifikant.

Tidigare på kursen har ni sett hur man beräknar konfidensintervall för exempelvis proportioner. Logiken för att beräkna konfidensintervall för b-värden är liknande. Precis som i det förra fallet börjar vi med att beräkna felmarginalen. Felmarginalen för en regressionskoefficient får vi helt enkelt genom att multiplicera regressionskoefficientens standardfel med det kritiska värdet för den valda säkerhetsnivån.

$$\text{Felmarginal} = \text{Kritiskt värde} \times \text{Standardfel} \tag{4}$$

Och vad är regressionskoefficientens standardfel för något? Det är ett mått på hur stor osäkerheten är i vår skattning av b-värdet. Anta att vi gjorde ett oändligt antal urval och genomförde samma regressionsanalys på varje urval. Regressionens standardfel är en uppskattning av hur stor standardavvikelsen skulle vara bland alla dessa b-värden. Det är ungefär samma sak som hur långt ifrån "den sanna effekten" b som b-värdena skulle vara i genomsnitt. Notera att koefficienternas standardfel inte har någonting att göra med passningsmåttet regressionsstandardfel.

Precis som i fallet med urvalsproportioner så har vi att väga precision mot säkerhet när vi väljer säkerhetsnivån. När det gällde de valda säkerhetsnivåerna för proportioner så fick vi de kritiska z-värdena för dessa från normalfördelningen. Av anledningar som går utanför den här kursen så bör man dock inte använda sig av normalfördelning när man beräknar felmarginaler för regressionskoefficienter utan av t-fördelningen. Den senare fördelningen är dock väldigt lik normalfördelningen och för n större än 100 är det nästan omöjligt att se skillnad på dem. Givet att vi har bestämt oss för en viss säkerhetsnivå, exempelvis 95 procent, så får vi alltså felmarginalen genom att multiplicera det kritiska t-värdet för denna säkerhetsnivå med b-värdets standardfel.

$$\text{Felmarginal} = t_{kv} * se_b \tag{5}$$

Som alla kan se är logiken densamma som när vi beräknade felmarginaler för proportioner, den enda skillnaden är att vi nu hämtar det kritiska värdet från t-tabellen i stället för från z-tabellen. För att erhålla konfidensintervallet för vårt skattade b-värde subtraherar och adderar vi sedan felmarginalen till vårt b-värde.

$$\text{Konfidensintervall} = b \pm \text{Felmarginalen} = b \pm t_{kv} * se_b \tag{6}$$

Om detta konfidensintervall täcker in 0 så törs vi inte dra slutsatsen att det finns ett samband i populationen, men om det inte täcker in 0 så törs vi dra slutsatsen att ett samband finns. Vi kan alltid testa om det finns ett samband i populationen på detta sätt. Samtidigt är det dock lite omständligt att beräkna konfidensintervallet varje gång man vill testa om det finns ett signifikant samband. Lyckligtvis finns det dock ett enklare sätt. Om det t-värde som vi beräknar eller erhåller är större än det kritiska t-värdet kommer konfidensintervallet inte att täcka in 0. Vi kan beräkna t-värdet genom att dividera regressionskoefficienten med dess standardfel.

$$t = \frac{b}{se_b} \tag{7}$$

Om (det absoluta värdet av) t-värdet är större än det kritiska t-värdet för vår säkerhetsnivå säger vi att sambandet är statistiskt signifikant. Vi vet då att det finns ett samband i populationen och att den observerade samvariationen inte var en tillfällighet.

* * *

Låt oss exempelvis anta att vi har ett urval om 2000 personer. Det kritiska t-värdet för 95 procents säkerhetsnivå är då – precis som i fallet med proportioner – 1,96. Anta att datorn anger att t-värdet för en viss variabel är 2,04. Vi drar då slutsatsen att det finns ett samband i populationen (vi förkastar H_0) då t-värdet faller utanför det kritiska intervallet som avgränsas av -1,96 och +1,96. Om variabelns t-värde däremot hade varit större än -1,96 men mindre än +1,96 hade vi inte vågat dra slutsatsen att det verkligen finns ett samband i populationen.

Vi ska nu återgå till tre exempel från den föregående föreläsningen. Data kommer från SOM-undersökningarna, World Value Survey, Quality of Government-databasen samt svaren på Metod C-enkäten.

1. När vi genomförde en regressionsanalys med ideologisk position som beroende variabel (från 1=vänster till 5=höger) och kön som oberoende variabel (0=man, 1=kvinna) såg vi att kvinnor i genomsnitt står längre åt vänster än män. Regressionskoefficienten var $-0{,}125$ och koefficientens standardfel var 0,027. T-värdet beräknar vi till -4.63 (-0,125/0,027). Eftersom urvalet är stort (n=7329) är det kritiska t-värdet vid 99 procents säkerhetsnivå detsamma som för normalfördelningen: 2,58. Eftersom 4,63 är större än 2,58 (vilket är samma sak som att -4,63 ligger utanför det kritiska intervallet från -2,58 till 2,58) kan vi konstatera att effekten av kön på den ideologiska positionen är statistiskt signifikant vid 99 procents säkerhetsnivå.

2. Regressionsanalysen med korruption som beroende variabel och graden av religiositet som oberoende variabel visade att länder där religion spelar en central roll har en högre grad av korruption än mer sekulära länder. Regressionskoefficienten för religiositet var 1,808 och standardfelet för koefficienten var 0,313. Vi beräknar t-värdet till 5,78 (1,808/0,313). Eftersom det är större än det kritiska värdet på 99 procents säkerhetsnivå (eftersom antalet observationer är ganska litet är t-värdet något större än 2,58, nämligen 2,64) kan vi slå fast att sambandet är statistiskt signifikant på 99 procents säkerhetsnivå.

3. När vi analyserar svaren på Metod C-enkäten ser vi att personer som har en stor tilltro till att regeringen arbetar för det allmänna bästa (oberoende variabel) är mindre benägna än andra att skänka pengar till tiggare (beroende variabel). Men är effekten statistiskt signifikant? Regressionskoefficienten var -0,188 och dess standardfel var 0,156, vilket ger ett t-värde på -1,21 (-0,188/0,156). Eftersom 1,21 är mindre än det kritiska t-värdet på 90 procent, vilket i det här fallet är 1,67, är sambandet inte statistiskt signifikant. Det innebär att vi inte kan utesluta möjligheten att sambandet beror på slumpen och att det inte finns någon effekt av tilltro till regeringen.

2.1 Signifikanstest vid totalundersökningar?

När vi gör beskrivningar är vi ofta genuint intresserade av en existerande population, men så är sällan fallet med en förklaring. Förklaringar handlar nästan uteslutande om mer generella samband mellan variabler, sådana att en ökning av x predicerar en förändring i y. Inte bara i de fall vi känner till – utan även i framtida fall som snart kommer att ske eller i den teoretiska superpopulation som genererar de verkliga fall vi kan studera. Därför är det svårt att tänka sig en totalundersökning när man håller på med förklaringar. Av samma anledning gör vi nästan alltid signifikanstest när vi genomför regressionsanalyser.

Även om vi har studerat alla revolutioner som ägt rum, är det förmodligen revolutioner som fenomen snarare än populationen av inträffade revolutioner vi är intresserade av. Det här kan man uttrycka på lite olika sätt. I samplingslitteraturen pratar man om superpopulationer. Vi kan då tänka oss en superpopulation av alla revolutioner som någonsin kommer att inträffa och betrakta de revolutioner som hittills skett som en dragning ur denna population. Inom andra traditioner pratar man i stället om en datagenererande process (DGP), vilket är en underliggande statistisk modell vars dynamik och kausala samband ger upphov till den värld vi observerar. Detta är ett vanligt uttryckssätt när vi har en följd av observationer över tid, eftersom det är svårt att tänka sig dem som dragna ur en population. Teorell och Svensson (s. 215–218) beskriver detta i termer av epistemologisk probabilism. På grund

av de mätfel vi gör, och alla de mikrosamband vi inte kan observera men som tillsammans skapar vår komplexa värld, kan vi betrakta den verklighet vi observerar som åtminstone delvis genererad av slumpprocesser. Samma typ av resonemang kan förövrigt förklara varför normalfördelningar är så vanliga.

Ett annat sätt att motivera detta, åtminstone i fallet med signifikanstest av regressionskoefficienter, är som en naturlig referenspunkt för vad som utgör ett starkt samband. Ett insignifikant samband är då ett samband som inte är starkare än att det mycket väl skulle kunnat ha genererats av en slumpprocess. Vi behöver alltså inte anta att våra observationer av verkligheten är (delvis) genererade av slumpprocesser, för att använda dessa som referenspunkt.

2.2 Sammanfattning

Låt oss nu summera avsnittet om hur vi kan avgöra om sambandet är signifikant. Notera att alla metoder ger samma slutsatser. För er är de två första metoderna viktigast. Senare i livet kommer ni mest att använda de två sista metoderna.

- T-värdet är högre än det kritiska t-värdet.
- Konfidensintervallet runt koefficienten omsluter inte värdet 0.
- Det står asterisker efter efter regressionskoefficienten. Läs under tabellen för att se vilken säkerhetsnivå de motsvarar. Det är den vanligaste metoden när man läser en regressionstabell.
- P-värdet är mindre än risknivån ($risk = 1 - säkerhetsnivå$). Det är den vanligaste metoden när man tolkar output från ett statistikprogram.

3 Att läsa en regressionstabell

Även om det är fullt möjligt att utföra enklare regressionsanalyser med en simpel miniräknare – vilket man ofta får göra på fortsättningskurser i statistik – beräknar vi sällan regressionsresultaten manuellt. Det vanliga är att vi läser en regressionstabell som andra sammanställt eller att vi läser av den output vi erhåller från vårt statistikprogram. Då är det viktigt att vi vet var vi hittar de värden vi letar efter och vad de kallas.

Regressionstabeller brukar struktureras så att resultaten från regressionsmodellerna anges kolumnvis, så att den första kolumnen motsvarar den första modellen, den andra kolumnen den andra modellen, och så vidare. Varje variabel har en egen rad där dess regressionskoefficient redovisas. Under koefficienterna hittar vi oftast koefficientens standardfel inom parenteser,

Tabell 1: Chokladkonsumtion och Nobelpris

	(1)	(2)
Choklad	2,81***	2,28***
	(0,50)	(0,64)
BNP/cap		0,20
		(0,16)
Konstant	−3,99	−8,42*
	(3,00)	(4,58)
Observationer	23	23
Standardfel	6,60	6,51
R^2	0,60	0,63

Standardfel i parenteser.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

men det kan också vara t-värdet. Oavsett vilket bör det framgå vad som anges i en not under tabellen.

I regel visas vilka regressionskoefficienter som är statistiskt signifikanta med asterisker (*). Olika antal asterisker motsvarar då olika säkerhetsnivåer. Det vanligaste är $* = 90$ procent, $** = 95$ procent och $*** = 99$ procent, men detta bör alltid anges i en not under regressionstabellen. Eftersom vi vill att ni ska kunna tolka statistisk signifikans utan hjälp av asterisker kommer ni inte att se dem så mycket under kursen.

Tabell 1 redovisar resultaten från två regressioner av antalet Nobelpristagare. Den första kolumnen redovisar resultaten från en bivariat regression med endast chokladkonsumtion (kg per invånare och år) som förklarande variabel. I den andra kolumnen har även BNP per capita (tusentals dollar) inkluderats, vilket ger oss en trivariat regression.

4 Multipel regression

Under den förra föreläsningen diskuterades hur man belägger samband mellan två variabler med hjälp av bivariat regressionsanalys. Under dagens föreläsning ska vi titta närmre på hur regressionsanalys kan användas i de fall då vi har fler än en förklaringsfaktor (oberoende variabel). På metodjargong säger vi att vi går från bivariat till multipel regressionsanalys.

Något förenklat kan man säga att det finns tre huvudsakliga skäl att gå från bivariat till multipel analys. Vi kan lyfta in ytterligare oberoende variabler i analysen i syfte att:

- Förbättra förklaringen.
- Isolera sambandet.

- Hitta orsaksmekanism.

Den första anledningen är kanske den mest intuitiva. Få av oss tror på allvar att samhällsvetenskapliga fenomen som krig, demokrati eller inställningen till jämställdhet har en enda förklaring. Ofta tänker vi oss i stället att inställning till jämställdhet kan förklaras av flera faktorer: utbildningsnivå, kön, ålder, politisk ideologi, osv. Med andra ord, om vi på ett tillfredsställande sätt ska kunna förklara variation i inställning till jämställdhet inom en befolkning så måste vi tillåta mer än en oberoende variabel i förklaringsmodellen. Ett sätt att göra detta är genom att skatta en multipel regressionsmodell.

Den andra anledningen är att vi vill isolera sambandet från andra möjliga förklaringar. Att ett samband mellan två variabler är statistiskt signifikant betyder inte att det måste föreligga ett kausalt samband mellan variablerna. En annan möjlighet är att samvariationen mellan de två variablerna beror på en tredje, bakomliggande variabel som påverkar de båda. Exempelvis kanske vi kan tänka oss att rika länder både har råd med mycket choklad och lägger mycket pengar på spetsforskning. Det skulle i så fall kunna förklara varför länder med hög chokladkonsumtion har fått fler Nobelpris. Vi skulle då kalla det ursprungliga sambandet för spuriöst eller skenbart.

Den tredje anledningen är att vi vill identifiera mellanliggande variabler för att förstå hur x påverkar y och därigenom stärka trovärdigheten i vår förklaring. En mellanliggande variabel är helt enkelt en variabel som både påverkas av den oberoende variabeln och i sin tur påverkar den beroende variabeln. Om vi identifierar en mellanliggande variabel är det ursprungliga sambandet mellan den oberoende och den beroende variabeln fortfarande kausalt, men indirekt.

* * *

Givet att vi funnit en (bivariat) kontrafaktisk skillnad och även kan ge argument för den antagna orsaksriktningen blir nästa steg att försöka isolera vårt samband från alternativa variabler som potentiellt kan förklara värdet på såväl vår oberoende som beroende variabel. En kollega vid institutionen (Pär Zetterberg) fann i en regressionsanalys ett bivariat samband mellan en viss typ av könskvoteringslagstiftning till parlament och kvinnliga medborgares politiska intresse. Vi bör då fundera över variabler som föregår både kvoteringslagstiftningen och politiskt intresse i tid och som potentiellt kan förklara utfallen på båda dessa variabler. En sådan variabel skulle kunna vara socioekonomisk utveckling. Risken finns nämligen att det bivariata sambandet är skenbart, eller spuriöst för att använda en mer teknisk term. Det vill säga vi tror att införandet av kvotering ökar kvinnliga medborgares politiska intresse när det i själva verket är så att socioekonomisk utveckling ökar sannolikheten att ett land inför kvotering och ökar benägenheten för kvinnliga medborgare i landet att intressera sig för politik (tack vare ökad utbildningsnivå, osv.).

Om sambandet mellan kvotering och kvinnors politiska intresse skulle minska i styrka och ej längre vara signifikant när vi kontrollerar för socioekonomisk utveckling, då konstaterar vi med andra ord att det bivarata sambandet var spuriöst och alltså icke-kausalt. Om sambandet dock fortfarande är signifikant, då har vi kunnat isolera för socioekonomisk utveckling och vi har i viss mån fått stöd för isoleringskriteriet. Observera dock att det givetvis potentiellt sett finns åtskilliga andra variabler som kan påverka såväl kvoteringslagstiftning som kvinnors politiska intresse. Det är därför viktigt att komma ihåg att vi inte kan kontrollera för allt. Detta är också en av anledningarna till varför teorier är så nyttiga i denna typ av undersökningar. Teorier hjälper oss att identifiera den grupp av bakomliggande variabler som är viktigast att kontrollera för.

Givet att vi har fått visst stöd för isoleringskriteriet kan vi även diskutera det fjärde orsakskriteriet: att belägga orsaksmekanismer. För att kunna leverera en orsaksförklaring menar Teorell och Svensson att vi behöver någon typ av argument för orsaksmekanismen, dvs varför orsak leder till verkan. Eller i ovanstående fall: varför kvoteringslagstiftning leder till ökat politiskt intresse bland kvinnor. Om vi ska tro teorier om genus och politisk representation skulle en sådan mekanism kunna vara att kvotering ökar andelen kvinnor i parlamentet, som kan vara politiska förebilder åt kvinnliga medborgare. På teoretiska grunder tror vi alltså både att kvotering verkligen implementeras effektivt och ökar andelen kvinnor i parlamentet och att dessa kvinnor kan vara politiska förebilder och därmed ökar kvinnors intresse för politik. Till skillnad från när vi försöker isolera ett samband så behöver inte ett försvagat och icke-signifikant samband i trivariat analys innebära att sambandet mellan kvotering och kvinnors politiska intresse är icke-kausalt. Om vi finner ett positivt samband mellan kvotering och andelen kvinnor i parlament och dessutom ett positivt samband mellan andelen kvinnor i parlament och kvinnors politiska intresse, då har vi i stället identifierat den mekanism varigenom effekten mellan kvotering och kvinnors politiska intresse går. Sambandet mellan kvotering och kvinnors politiska intresse är därmed indirekt och i allra högsta grad kausalt.

Det är viktigt att skilja mellan spuriösa och indirekta samband. För medan de förra är icke-kausala är de senare kausala. Att reda ut skillnaden mellan dessa olika samband kommer vi att ägna det mesta av den tid som återstår av denna föreläsning. Ni kommer också att få öva på detta inför seminarium 3 och 4. Här följer dock en kort sammanfattning som ni kan ha med er framöver.

- Om det ursprungliga sambandet kvarstår även efter att vi kontrollerat för en tredje variabel har vi isolerat sambandet (från just den variabeln).
- Om det ursprungliga sambandet försvann eller försvagades efter att vi kontrollerat för en *mellanliggande* variabel var sambandet indirekt. Effekten går via den nya variabeln, vilken fungerar som orsaksmekanism.

- Om det ursprungliga sambandet försvann eller försvagades efter att vi kontrollerat för en *bakomliggande* variabel var det ett skensamband och spuriöst.
- Om det ursprungliga sambandet förstärks efter att vi kontrollerat för en tredje variabel, fungerar den nya variabeln som en suppressorvariabel.
- Orsaksriktningen mellan variabler avgörs huvudsakligen av teoretiska resonemang.

4.1 Den multipla regressionsekvationen

Innan vi studerar resultatet från några multipla regressioner vill jag kort gå igenom den multipla regressionsmodellen. Logiken för multipel regression är densamma som för bivariat regression. Beräkningarna som datorn utför är visserligen mer komplicerade, men de behöver vi lyckligtvis inte bry oss om. Vi ska i stället uppehålla oss vid de substantiella tolkningarna av regressionsresultaten.

Utan att gå in närmare i detalj på den multipla regressionsekvationen ska jag dock först säga några ord om den. Teorell och Svensson går på s. 191 igenom den multipla regressionsekvationen, som på många sätt påminner om den bivariata regressionsekvationen. I det fall när vi har två oberoende variabler kan regressionsekvationen skrivas:

$$y = a + b_1x_1 + b_2x_2 + e \tag{8}$$

Trots något mer komplicerade matematiska beräkningar för att beräkna intercept och regressionskoefficient så är logiken densamma som i bivariat analys. Vi använder minsta kvadratmetoden för att hitta de koefficienter som minimerar summan av de kvadrerade avvikelserna. Tolkningen påminner också mycket om det bivariata fallet: a är det förväntade värdet på y när de oberoende variablerna antar värdet 0. Och regressionskoefficienten b_1 ger den genomsnittliga förändringen i y då x_1 ökas med en enhet och x_2 hålls konstant, medan b_2 ger den genomsnittliga förändringen i y då x_2 ökas med en enhet och x_1 hålls konstant (alternativt ”kontrollerar för x_1”, ”isolerar för x_1”, ”filtrerar bort effekten av x_1”).

4.2 Tre exempel

Vi ska nu återgå till våra tidigare tre exempel och komplettera de bivariata regressionerna med multipla regressioner. I resultaten kommer vi bland annat stöta på indirekta samband, spuriösa samband och suppressorvariabler.

Tabell 2: Beroende variabel: Höger

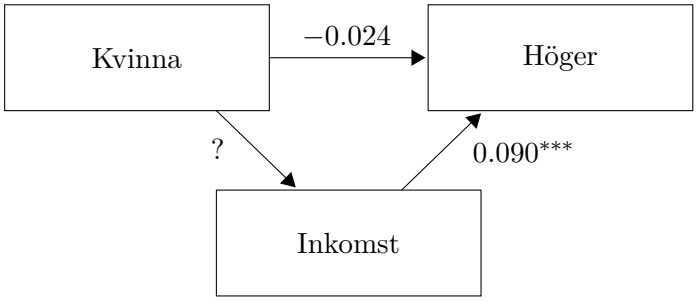
	(1)	(2)	(3)
Kvinna	−0,125	−0,123	−0,024
	(−4,63)	(−4,62)	(−0,87)
Ålder		0,001	
		(0,20)	
Inkomst			0,090
			(17,33)
Konstant	3,074	3,066	2,636
Observationer	7329	7329	7329

t-värden i parenteser.

4.2.1 Exempel 1

Vi har tidigare påvisat ett samband mellan kön och ideologisk position, sådant att kvinnor i genomsnitt är längre åt vänster än män. Sambandet visade sig också vara statistiskt signifikant. I den första kolumnen i Tabell 2 återges resultaten från den bivariata regressionen. Regressionskoefficienten -0,125 kan vi tolka som att kvinnor i gemensnitt står 0,125 skalsteg längre åt vänster än män. Men vad händer med det bivariata sambandet mellan kön och ideologisk position när vi kontrollerar för ålder? Svaret hittar vi i den andra kolumnen. Ålder tycks inte ha någon effekt på den beroende variabeln och det ursprungliga sambandet kvarstår så gott som oförändrat. Regressionskoefficienten är -0,123 och eftersom t-värdet (4,62) är större än det kritiska värdet vid 99 procents säkerhetsnivå (2,58) är sambandet fortfarande statistiskt signifikant. Vi kan säga att vi har isolerat det ursprungliga sambandet från effekten av ålder.

Men vad händer om vi i stället för ålder inkluderar inkomst i modellen? Resultaten återfinns i den tredje kolumnen. Det första vi ser är att effekten av kön har minskat kraftigt. Regressionskoefficienten har krympt till -0,024. Eftersom t-värdet (0,87) är mindre än det kritiska värdet vid 90 procents säkerhetsnivå (1,65) är effekten inte längre statistiskt signifikant. Inkomst har däremot en stor och statistiskt signifikant effekt. Tekniskt sett kan vi tänka oss två möjliga förklaringar till det som skedde. Antingen är sambandet mellan kön och ideologisk position indirekt, så att kön påverkar vilken inkomst man har vilket i sin tur har en effekt på ens ideologiska position. Eller så var det ursprungliga sambandet spuriöst, så att inkomst påverkar både ens kön och ideologiska position. Men eftersom vi finner det osannolikt att ens kön påverkas av ens inkomst kan vi i det här fallet sluta oss till det första alternativet – sambandet var indirekt och inkomst fungerar här som en mellanliggande variabel. Figur 4.2.1 visar hur sambandet kan illustreras



Figur 1: Inkomst är en mellanliggande variabel

med ett kausaldiagram. Asteriskerna visar vilka effekter som är statistiskt signifikanta. Tre asterisker motsvarar 99 procents säkerhetsnivå.

Avslutningsvis kan vi här säga några ord om hur man räknar ut förväntat värde för multipla regressionsekvationer. Anta att ni blir tillfrågade om var ni tror att en man med en inkomst mellan 20 000 och 25 000 kr (4 på inkomstskalan) befinner sig på höger–vänster-skalan. Ni kan då besvara frågan genom att beräkna ett förväntat värde där värdena på de oberoende variablerna är 0 (man=0 på x_1) respektive 4 (värdet på inkomstskalan x_2). Värdena för konstanten ($a = 2.636$), effekten av kön ($b_1 = -0.024$) och effekten av inkomst ($b_2 = 0.090$) hämtar vi från regressionstabellen och stoppar in i ekvationen.

$$\begin{aligned} \hat{y} &= a + b_1 x_1 + b_2 x_2 \\ \text{Höger} &= 2.636 - 0.024 \times 0 + 0.090 \times 4 = 2.996 \end{aligned} \tag{9}$$

Som uträkningen visar blir det förväntade värdet 2.996. Det kan vi tolka som att en man med genomsnittligt hög inkomst förväntas befinna sig i mitten av skalan som går från klart åt vänster (1) till klart åt höger (5). Trots att effekten av kön visade sig vara icke-signifikant är det i regel bäst att ta med det värdet i uträkningen. -0.024 är trots allt vår bästa gissning! Om man trots allt skulle bestämma sig för att utesluta kön ur uträkningen så måste man först genomföra en bivariat analys där inkomst är den enda oberoende variabeln.

4.2.2 Exempel 2

En av Sveriges mest framgångsrika statsvetare, Bo Rothstein, argumenterade på DN Debatt för att religiösa samhällen leder till minskad tillit och ökad korruption. Bosses tes fick också stöd i den bivariata regressionen – ju fler i ett land som anser att religion är en viktig del av deras liv, desto högre tenderar korruptionen i landet att vara. Sambandet var statistiskt signifikant

Tabell 3: Beroende variabel: korruption

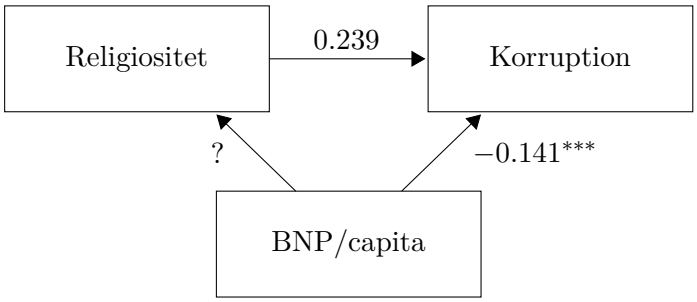
	(1)	(2)
Religiositet	1,808	0,239
	(0,313)	(0,232)
BNP per capita		−0,141
		(0,012)
Konstant	−0,061	6,745
Observationer	78	78

Standardfel i parenteser.

eftersom t-värdet (5,78) var högre än det kritiska värdet på 99 procents säkerhetsnivå (här 2,64).

Eftersom sambandet var statistiskt signifikant beror det förmodligen inte på slumpen, men en rimlig invändning mot Rothsteins argument är att det kan finnas variabler som påverkar både religiositet och korruption. Kanske är det så att ekonomisk utveckling leder till både sekularisering och minskad korruption, vilket skulle kunna förklara varför sekulariserade länder har lägre korruption än religiösa länder. För att undersöka detta genomför vi en regressionsanalys där vi kontrollerar för BNP capita. Resultaten från regressionen presenteras i Tabell 4.

När BNP per capita inkluderas i modellen reduceras den skattade effekten av religiositet kraftigt (från 1,808 till 0,239). Om vi beräknar ett t-värde (0,239/0,232=1,03) ser vi också att effekten inte längre är statistiskt signifikant. Min tolkning är att BNP per capita är en bakomliggande variabel och att det ursprungliga sambandet var spuriöst. Det är också den tolkningen som illustreras i Figur 4.2.2. Vi måste dock vara försiktiga i våra antaganden om orsaksriktningen mellan de tre variablerna – alltså hur pilarna går. Någon skulle kanske invända att ekonomisk utveckling kan påverkas av både kor-



Figur 2: Sambandet mellen religiositet och korruption var spuriöst om vi accepterar pilarnas riktning

Tabell 4: Beroende variabel: Ger till tiggare

	(1)	(2)
Förtroende	−0,188	−0,332
	(−1,20)	(−2,15)
Borgerlig		−0,674
		(−2,95)
Konstant	2,177	2,811
Observationer	55	55

t-värden i parenteser.

ruption och religiositet, vilket gör det långt i från självklart hur sambandet egentligen ser ut. Kanske gör vi säkrast i att bara observera sambanden utan att göra några kraftigare anspråk på kausalitet?

4.2.3 Exempel 3

När vi analyserade Metod C-enkäten fann vi att personer med hög tillit till regeringen var mindre benägna att skänka pengar till tiggare. Effekten var dock inte statistiskt signifikant. Kanske var det bara en slump att de som angav hög tillit också angav en lägre benägenhet att ge? Tabell **??** visar vad som händer om vi adderar en tredje variabel vilken antar värdet 1 om respondenten röstar på ett borgerligt parti och 0 om respondenten röstar på något av de rödgröna partierna.

Resultaten från den trivariata analysen visas i den andra kolumnen. För det första ser vi en tydlig effekt av dummyvariabeln som skiljer borgerliga sympatisörer från övriga. Röstar man på ett borgerligt parti minskar benägenheten att ge pengar till tiggare med 0,674 skalsteg. För det andra har effekten av tillit till regeringen nu vuxit från -0,188 till -0,332. Eftersom t-värdet (2,15) är större än det kritiska värdet vid 95 procents säkerhetsnivå har sambandet dessutom blivit statistiskt signifikant.

Så hur kan de komma sig att effekten av regeringssätt nu är signifikant? Kontrollvariabler som *stärker* sambandet kallas för suppressorvariabler. Teorell och Svensson skriver om detta på s. 194. Order suppressor kommer förmodligen från att variabeln 'trycker tillbaka' eller konstanthåller en del av variationen i den huvudsakliga förklaringsvariabeln som har en motsatt eller svagare effekt på den beroende variabeln.

I det här exemplet fanns det kanske två samband mellan förtroende för regeringen och benägenhet att ge som tog ut varandra i det bivariata fallet. För det första kan det ha funnits en negativ effekt som berodde på att de som litar på regeringen inte känner samma ansvar att själva hjälpa de utsatta – det där är ju upp till regeringen. För det andra kan det ha funnits en positiv

effekt av att de flesta som har förtroende för den nuvarande regeringen också röstar på något av de rödgröna partierna och därför tenderar att vara mer positivt inställda till tiggeri än de som röstar borgerligt. När vi kontrollerar för politisk tillhörighet ”kontrollerar vi bort” den senare effekten, vilket gör att vi bara har en negativ effekt i stället för summan av två effekter med olika tecken. Det är i alla fall en möjlig förklaring till varför borgerlig fungerar som en suppressorvariabel här, men det är förstås inte en berättelse fri från invändningar.

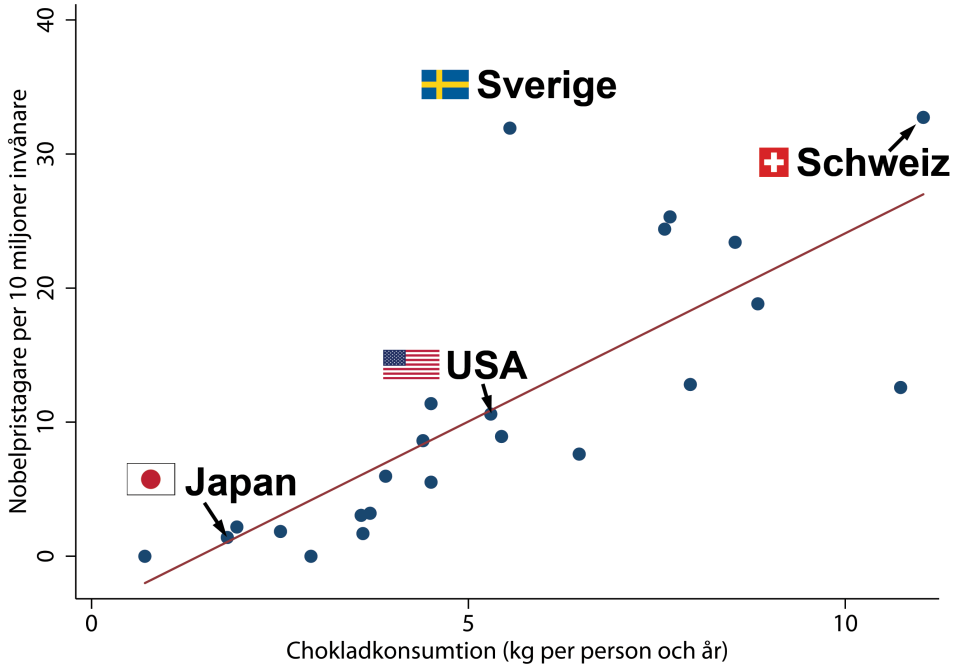
5 Kombinationsstudier

Det förs ständigt en diskussionen om kvantitativa respektive kvalitativa metoders användningsområden och begränsningar. De flesta är nog överens om att uppdelningen ”kvantare” (personer som främst utför extensiva studier) och ”kvallare” (personer som främst utför intensiva studier) är olycklig, särskilt som den leder till onödiga positioneringar och begränsar forskarens möjliga angreppssätt. Valet av metod bör styras av forskningsfrågan och luckorna i tidigare forskning, snarare än vilket läger som forskaren tillhör. Och diskussioner om huruvida en typ av metod alltid är bättre än annan är sällan fruktbara.

En mer konstruktiv utgångspunkt är att se intensiva och extensiva studier som komplementära. Både traditionerna har sina respektive styrkor och svagheter. Det är svårt att visa på samvariation och isolera orsakssamband i en intensiv studie, men i gengäld kan det i dessa studier ofta vara enklare att finna belägg för tidsordning och orsaksmekanism. På motsatt vis är extensiva studier bra på samvariation och isolering men brister ofta i att belägga tidsordning och spåra orsaksmekanismer.

Method of Agreement innebär att vi väljer fall som har samma (likartade) utfall på den beroende variabeln men är så olika som möjligt i alla andra relevanta avseenden. Notera att vi fortfarande saknar belägg för kontrafaktisk skillnad! Endast i *Method of Difference* har vi belägg för kontrafaktisk skillnad, men de intensiva metodernas akilleshäl kvarstår – vi vet fortfarande inte om sambandet är systematiskt eller slumpartat. Extensiva upplägg är därför bättre på att ge belägg för samvariation eller kontrafaktisk skillnad samt isolera andra förklaringar. De har också fördelen att de kan hantera probabilistiska samband bättre än fåfallsstudier (se Teorell & Svensson s.241).

Hela dagen har vi gjort antaganden om orsaksriktningar, sådana att x påverkar y samtidigt som y inte har någon effekt på x. Sådana ataganden brukar vara mer problematiskt än vad det har varit i våra exempel. De kvantitativa metoder som finns för att belägga orsaksriktning kräver bra data med tidsvariation sant ofta andra antaganden, exempelvis om effektens fördröjning (hur lång tid det tar för en förändring i x att resultera i en förändring i y). Och även om extensiva studier kan visa på en orsaksmekanism,



vilket vi har tittat på idag, kan de inte följa en process lika nära som en intensiv studie kan göra. *Utöver* detta fyller intensiva studier även viktiga teoriutvecklande eller hypotesgenererande funktioner, på ett sätt kvantitativa studier mer sällan gör.

För att dra nytta av de relativa fördelarna i båda traditionerna framhåller många kombinationsstudier som ett ideal. Utgångspunkten är att den enda möjligheten för att hitta belägg för alla fyra orsakskriterierna är att kombinera extensiva och intensiva ansatser. Det är förstås tidskrävande att genomföra flera olika delstudier, men kom ihåg att allt inte nödvändigtvis måste göras i samma uppsats eller ens av samma forskare. Man kan också utgå från tidigare forskning för att identifiera vilken typ av studier som fältet är i störst behov av. Den vanligaste typen av kombinerade metoder är förmodligen att välja fall på basis av en extensiv studie. Om syftet är att belägga orsaksriktning och/eller orsaksmekanism bör vi välja fall som passar in i huvudmönstret, det vill säga som är representativa för det samband som vi har funnit. Det kallas ibland för att välja ett fall på regressionslinjen. En annan möjlighet är att använda den intensiva studien till att generera nya (konkurrerande eller komplemetterande) hypoteser om vad som kan förklara ett visst fenomen. Vi väljer då länder som ligger långt ifrån regressionslinjen och alltså inte kan förklaras av våra nuvarande teorier.

Figur 5 visar återigen sambandet mellan chokladkonsumtion och antalet Nobelpris. Om vi skulle välja ett avvikande fall för att leta nya förklaringar till frekvensen av Nobelpris skulle kanske Sverige vara ett bra val. Sveriges höga antal Nobelpris kan inte förklaras av vår chokladkonsumtion, vilket

innebär att det borde finnas en annan variabel som är viktig i det svenska fallet (och därmed kanske även i andra fall). Spelar det möjligtvis någon roll vilket land det är som delar ut prisen? Om vi skulle välja ett illustrativt fall som är representativt för sambandet skulle vi välja något av länderna på regressionslinjen. I detta fall är det förmodligen en god idé att välja Schweiz, eftersom det är enklare att försöka spåra mekanismen i ett land med hög chokladkonsumtion och många Nobelpristagare. Så är det inte alltid. Ibland är det tvärtom enklare att hitta mekanismer i frånvaron av någonting.

6 Några saker att se upp för

Föreläsningen avslutades med sju saker man bör vara vaksam på när man gör en regressionsanalys.

1. Hittills har vi antagit att alla samband är linjära, så att en ökning i x alltid ger samma ökning i $\hat{y}$. Samtidigt vet vi att många samband är avtagande, så att effekten av en ökning i x avtar med värdet på x. Ett klassiskt exempel är hur små inkomstökningar i fattiga länder har en stor påverkan på hälsa och förväntad livslängd, medan samma inkomstökning har en försumbar hälsoeffekt i ett land som Sverige.

   Det vanligaste sättet att hantera avtagande samband är att man logaritmerar den oberoende variabeln. När vi logaritmerar den oberoende variabeln analyserar vi effekten av relativa förändringar (”om x ökar med 100 procent”) i stället för absoluta förändringar (”om x ökar med 10”). Valet att logaritmera kan motiveras både empiriskt (”passningen blir bättre”) och teoretiskt (”det är rimligt att en fördubbling av BNP orsakar en lika stor förändring i förväntad livslängd oavsett BNP-nivå”).

2. Den metod vi använder för att skatta regressionslinjen innebär att observationer med extrema värden – så kallade outliers – kan få en stor betydelse för regressionslinjens lutning. Om extremvärdena kan avfärdas som mätfel bör de exkluderas från modellen, men annars finns det inget självklart svar på hur de ska hanteras. Det bästa är ofta att redovisa resultaten både med och utan outliers.

3. Det är svårt att jämföra regressionskoefficienter med varandra. För det första är variabler ofta mätta på helt olika skalor. En variabel som mäter månadsinkomst kanske antar värden från noll till tiotusentals kr. Den kommer nästan alltid ha en väldigt liten regressionskoefficient, eftersom koefficienten mäter effekten av en förändring av inkomsten med en krona per månad. För det andra kan spridningen skilja sig åt mellan variabler, även om skalorna är jämförbara. I många situationer är en stor effekt ointressant, om nästan alla observationer har samma värden. För det

tredje är det aldrig okomplicerat att jämföra effekter av vitt skilda saker, även om skalorna och spridningen är jämförbar. Hur relevant är det att jämföra helt olika fenomen med varandra, som exempelvis hur ens inkomst förändras av att byta yrke med inkomsteffekten av att läsa ytterligare ett år på universitetet? Men trots svårigheterna är det ofta önskvärt att jämföra effekter med varandra, antingen i en och samma regressionsmodell eller att jämföra en effekt med liknande studier i tidigare forskning. Det viktiga är att vi är försiktiga och medvetna om problemen.

4. Regressionsanalysen säger ingenting om orsaksriktning. När vi bestämmer vilken variabel som är oberoende och vilken variabel som är beroende, men resultaten från en regressionsanalys berättar inte om vårt antagande var korrekt. Att sambandet i själva verket går åt motsatt håll hindrar inte resultaten från att bli statistiskt signifikanta. Vill vi skaffa belägg för orsaksriktningen bör vi använda sunt förnuft, teori och tidigare forskning eller andra metoder. Såväl intensiva studier som experiment är mer lämpade för att visa i vilken riktning ett samband går.

5. Ofta kommer många observationer från ett och samma fall. Vanligast är kanske när vi har data mellan länder och över tid, så att en observation är Sverige 2012 och en annan observation är Sverige 2013. Den typen av data erbjuder unika möjligheter för våra analyser men den skapar också problem. Å ena sidan kan vi, under vissa antaganden som exempelvis hur lång tid det tar för en effekt att äga rum, studera vilken variabel som tycks påverka den andra. Vi kan också välja att endast studera variation över tid, exempelvis genom att inkludera en dummyvariabel per land (eller vad vi nu studerar). På så vis kan vi kontrollera för alla de faktorer som är stabila över tid och som vi inte kan observera. Å andra sidan innebär den typen av data att vi lätt överdriver den statistiska signifikansen i våra resultat. Om vi inte tar hänsyn till att Sverige 2011 och Sverige 2012 knappast är oberoende av varandra, kommer vi få resultat som ser ut att vara mer signifikanta än vad de egentligen är.

6. Är den beroende variabeln dikotom? Vanlig linjär regression är dåligt lämpad för variabler som bara kan anta två värden. Ett av skälen till det är att den kan ge upphov till orimliga prediktioner. Även om den beroende variabel bara kan anta värdena 0 och 1, så kan den linjära regressionsmodellen göra prediktioner som är lägre än 0 och högre än 1. Det vanligaste sättet att hantera dikotoma beroende variabler är genom logistisk regression.

7. Det är svårt att isolera för alla tänkbara förklaringar. Vi vet sällan

vilka alla möjliga bakomliggande förklaringar är. Och även om vi visste det, är det inte självklart hur vi ska mäta dem eller att det ens är praktiskt möjligt. Och även om vi kände till och kunde mäta alla bakomliggande variabler, vet vi inte hur vi ska kontrollera för dem. Den linjära och additiva regressionsekvationen är bara en av många möjligheter. Av dessa skäl finns det nästan alltid en risk för att våra samband är spuriösa, trots att vi gjort vårt bästa för att isolera dem från andra faktorer. Därför är det aldrig fel att visa viss ödmjukhet när man presenterar sina regressionsresultat.