Harald Lang 15/2-10

KTH Teknikvetenskap

Några saker att vara observant på vid modellspecifikation med ”linjära modellen”

- **Transformation av beroende variabeln.**
  Ibland är det naturligt att transformera den beroendde variabeln, y-variabeln med t.ex. logaritmen; man använder alltså log(y) som beroende variabel. Det gäller ofta när y naturligt måste vara positiv. Ett exempel är när y är lönen. Säg att vi vill se hur ett extra år på arbetsmarknaden (ett års extra arbetslivserfarenhet) påverkar lönen. Använder vi då log(y) som beroende variabel ser vi hur mycket lönen ökar *relativt* (i procent) med ett extra års erfarenhet (i stället för absolut; kronor per månad).

- **Transformation av oberoende variabler.**
  Det kan också vara relevant att transformera en (eller flera) *beroende* variabler (x-variabler, kovariater). Ta exemplet ovan med lönen som ökar (troligtvis) med arbetslivserfarenheten. Förmodligen ökar lönen mest i början, t.ex. från ett till två års erfarenhet, men mindre – om ens alls – från 9 till 10 års erfarenhet. Man kan då tänka sig att ha både x och x^2 som variabler (x = antal års arbetslivserfarenhet). Då kan beroendet avta med x (koefficienten för x positv, koefficienten för x^2 negativ).

- **Användning av ”dummy-variabler”.**
  En ”dummy-variabel” är en indikator-variabel D som bara antar värdena 0 och 1. Om vi t.ex. vill skatta en löneekvation, som ovan, och har med variablerna utbildning (år), arbetslivserfarenhet (år), kan vi lägga till en ”dummy” för ”kvinna”, t.ex. Dvs. D =1 om personen är kvinna, 0 annars. Koefficienten för D beskriver då hur mycket mer en kvinna har i lön (procentuellt, om vi har logaritmerat lönen, som ovan) än en man med samma utbildning och arbetslivserfarenhet.

  Man kan också kombinera: låt oss säga att vi vill se om det lönar sig mer för kvinnor att utbilda sig än för män. Då kan vi ha med variabeln D·(utbildning). Vi har då två koefficienter, dels en för (utbildning), som gäller för män, dels *summan* av koefficienterna för (utbildning) och D·(utbildning) som är den koefficient som gäller för (utbildning) för kvinnor. Koefficienten för D·(utbildning) mäter alltså hur mycket *mer* det lönar sig för en kvinna att utbilda sig, jämfört med för en man.

- **Koefficienternas tolkning beror på kovariaterna.**
  Koefficienten för en kovariat (x-variabel) beskriver hur mycket den beroende variabeln (y-variabeln) ökar om x-värdet ökar med en enhet *då de övriga kovariaterna hålls fixa.* Exempel: vi håller oss igen till exemplet med lön. Antag att vi skattar de två ekvationerna

  $$\log(y) = a + b_1\cdot(\text{utbildning}) + b_2\cdot(\text{erfarenhet}) + \text{residual}$$

  respektive

  $$\log(y) = a + b_1\cdot(\text{ålder}) + b_2\cdot(\text{erfarenhet}) + \text{residual}$$

  Här är alltså ”erferenhet” arbetslivserfarenhet, ”ålder” personens bilologiska ålder och ”residual” det som också kallas ”felterm”. Jag förväntar mig att b_2 i den första ekvationen är större än b_2 i den andra. Varför det? Jo, tolkningarna är olika. I första ekvationen tolkar vi b_2 som att vi jämför två personer med *lika lång utbildning* där den ene har ett års mer arbetslivserfarenhet. I den andra jämför två personer *med samma ålder* där den ene har ett års mer arbetslivserfarenhet. I det senare fallet är det troligt att den med längre arbetslivserfarenhet har kortare utbildning – hur kan de annars vara lika gamla? Om vi vill mäta om arbetslivserfarenhet lönar sig (ger högre lön) måste vi bestämma oss för hur vi vill mäta: är det ett års längre arbetslivserfarenhet i *stället för* utbildning eller *utöver* utbildning vi menar.

- **Se upp med heteroskedasticitet.**
  I linjär modellen förutsätts att residualerna (feltermerna) har samma varians. Man bör tänka på att specificera modellen så att detta är troligt. Antag att vi vill skatta en ekvation där den beroende variaveln är BNP för ett land, och att vi skattar ekvationen på ett tvärsnitt av länder. Då är det troligt att variansen för residualen för t.ex. USA är betydligt större än för Danmark, pga. skillnaden i storlek. I stället för BNP som beroende variabel bör vi ta BNP per kapita, t.ex.

- **Se upp med multikolinearitet.**
  Om man får väldigt stora skattade standardavvikelser för några koefficienter har man kanske problem med *multikolinearitet.* Det betyder att det finns ett nära linjärt samband mellan några kovariater. Vi förstår att man inte kan ta med samma kovariat två gånger i en ekvation. Men av samma skäl får kovariaterna inte vara (nästan) linjärt beroende. Om vi tar vår evinnerliga löneekvation igen. Antag att vi tar med de tre kovariaterna (ålder), (utbildning) och (erfarenhet). Dessutom har vi ett intercept (konstantterm). Nu är det troligt att (ålder)-(utbildning)-(erfarenhet) är ungefär konstant, dvs, en multipel av interceptet, så vi får

förmodligen väldigt dålig precision i skattningen av koefficienterna. Intuitivt: Givet en persons ålder och utbildning kan vi prediktera hans arbetslivserfarenhet ganska väl, så lägger vi till (arbetslivserfarenhet) i ekvationen kommer den med två gånger.

- **Se upp med utelämnade relevanta variabler.**
  I en regressionsmodell förutsätts att det inte finns någon korrelation mellan resudualen och någon av kovariaterna. Antag t.ex. att vi vill undersöka om rökning påverkar hälsan på något specifikt sätt. Om vi då har (rökning) som en kovariat kanske den koefficienten är mycket signifikant (beroende variabeln är något mått på hälsotillståndet). Men kanske är det så att det som påverkar hälsan är något annat som är korrelerat med rökning. Kanske rökare motionerar mindre än andra, och det är detta som påverkar hälsan. Detta fångas då upp av ekvationen som ett samband mellan rökning och hälsa, fast något sådant inte finns. I stället för ”motion” kanske det är alkoholkonsumtion, eller dålig diet, som är orsaken till hälsotillståndet. Vi måste alltså ha med alla variabler som vi tror är relevanta; rökning, motion, ålder, diet, alkoholkonsumtion… bland kovariaterna. Annars hamnar dessa i redisualen, och då blir residualen korrelerad med ”rökning” (om rökning är korrelerad med någon av de utelämnade variablerna) – se nästa avsnitt. Om man utelämnar en relevant variabel som *inte* är korrelerad med dem som ingår i ekvationen blir skattningen korrekt, men mindre effektiv.

- **Se upp med endogenitet.**
  Det här är ett vanskligt problem. Tekniskt uppstår problemet genom att det finns en korrelation mellan residualen och någon kovariat. Man talar om ”endogenitet” när detta beroende går via den beroende variabeln. Antag att vi vill undersöka hur mycket efterfrågan på kanelbullar ökar om vi sänker priset. Om vi gör ett kontrollerat experiment så är det inget problem (alla försäljare av kanelbullar sänker priset på order under en tidsperiod). Men nu använder vi *observerade* data – priserna varierar litet av olikla skäl. Antag då att vi finner att efterfrågan tycks *minska* när priserna är låga, dvs. i en ekvation där beroende variabeln är försäljningsvolym är koefficienten för kovariaten ”pris” positiv.

  Det vi observerar är förmodligen det omvända. Det är inte bara så att priset påverkar efterfrågan, utan även tvärtom. Kanske efterfrågan på kanelbullar minskar därför att någon populär kändis visar sig föredra croissoner i stället. Folk köper då mindre kanelbullar och börjar äta croissonger i stället. Kanelbullebagarna kontrar då med att sänka priset på kanelbullar. Här är det alltså så att om residualen är liten (någonting händer så att efterfrågan minskar) så sänks priset, och vi får en (negativ, i det här fallet) korrelation mellan residual och kovariaten ”pris”.

- **Se upp med självselektion.**
  Det här igen ett fenomen som skapar en korrelation mellan residualen och någon kovatiat. Antag att Gunnar Englund vill undersöka om de som går på hans föreläsningar i Tillämpad Matematisk Statistik klarar sig bättre på tentan än dem som inte gör det. Antag att han till sin bedrövelse får som resultat att de som inte går på undervisningen klarar sig *bättre* och hans (okunniga) kollegor retar honom för att kans undevisning tydligen *försvårar* inlärning. Men så behöver det inte vara (i högsta grad osannolikt, enligt min mening). I stället kan det vara så att de speciellt studiebegåvade studenterna är de som *väljer* att inte gå på undervisningen. Det är här fråga om en *självselektion.* Vi får en (negativ) korrelation mellan residualen (som innehåller den dolda variabeln ”studiebegåvning”) och kovariaten ”närvaro vid undervisningen”. Det här uppstår när vi inte har kontrollerade experiment där vi själva väljer värden på kovariaterna, utan vi använder *observerade* data.

- **Skilj på prediktion och ”strukturtolkning”.**
  Problemen behandlade ovan gäller när man har en ”strukturtolkning” av ekvationerna, dvs, vi tänker oss att kovariaterna *påverkar* den beroende veriabeln. Om vi däremot bara vill *prediktera* den beroende variabeln så har vi inte dessa problem. Ta sista exemplet: Det är helt OK att *prediktera* att de som inte går på Gunnars föreläsningar kommer att klara tetan bra (förutsatt att vår historia ovan är riktig). Det är först om vi vill påstå att det *beror på* att de inte går på undervisningen som vi får problem.