Statistiska institutionen

LUNDS UNIVERSITET
Ekonomihögskolan

# Alkoholkonsumtion och dess orsaker

*Samband undersökta genom regressionsanalys och paneldata*

Filip Morén
Magnus Berg

Kandidatuppsats
Handledare: Jakob Bergman
Januari 2015

## Abstract

The purpose of this analysis is to use regression models to find out what factors influence the average consumption of alcohol on a national level. The data which is being used come from the WHO and the OECD. The analysis is restricted to countries that are members of the OECD. In order to analyze the material the following types of regression models are investigated: a basic linear model, a logistic model and a model based on panel data. The models are evaluated according to a few criteria: we look to include only significant parameters if possible, obtain an $R^2$ value as high as possible when this measurement is applicable, and we want the assumptions concerning the residuals to be fulfilled. In the end we reach the conclusion that the material used for the analysis could explain the consumption of alcohol only to a small degree. The few significant results we get, however, are in agreement with the findings of previous research. If the country has a monopoly of the selling of alcohol, the amount that people drink on average decreases, and drinking seems to increase in economically good times.

# Innehåll

# 1 Inledning

Att alkohol har en negativ inverkan på individen och samhället är ingen hemlighet. Kostnader som uppstår till följd av vandalisering, vård, förlorad arbetskraft m.m. i alkoholens spår är stora och beslutsfattare har länge, med olika tekniker och resultat, försökt kontrollera befolkningens konsumtion. I Sverige har det gått från helt oreglerat, via motbok till dagens monopol. Att handla sin alkohol på Systembolaget är för svenskar en självklarhet medan det i andra länder är lika självklart att köpa kyld öl på bensinmacken. Staten har olika verktyg för att kontrollera alkoholkonsumtionen i ett land. Exempel på dessa är punktskatter, försäljningsställen och åldersgränser. Dessa variabler är av rent legislativ natur och kan således ändras om nationens beslutsfattare kan komma överens.

I denna uppsats undersöker vi huruvida det går att finna några statistiska samband mellan lagstiftningen i ett land, ekonomiska variabler för detsamma och dess befolknings genomsnittliga alkoholkonsumtion. Denna undersökning görs med olika varianter av regressionsanalys på ett datamaterial som täcker *Organization of Economic Co-ooperation and Developments* (OECD) 34 medlemsstater. Vi hoppas att med denna analys kunna svara på vilken lagstiftning kring alkohol som har störst effekt på konsumtionen. Vidare undersöker vi hur ekonomiska parametrar så som arbetslöshet, konjunktur, BNP per capita m.m. påverkar konsumtionen av alkohol. Dessa parametrar kan inte direkt kontrolleras av lagstiftare på samma sätt som de legislativa men analysen blir fortfarande intressant då den skulle kunna svara på om skillnader i alkoholkonsumtion i stor grad styrs av ekonomiska aspekter eller om lagstiftning kan användas som kontrollmetod

## 1.1 Bakgrund

World Health Organization publicerar med jämna mellanrum *Global status report on alcohol and health* vilket är en sammanställning och genomgång av den globala alkoholkonsumtionen och dess effekter. Det var vid läsandet av 2014 års rapport som idén till denna uppsats föddes. Efter att bara snabbt ögnat igenom rapporten förstod vi att alkoholkonsumtion runt om i världen är ett mycket större problem än vi tidigare trott. Som boende i Sverige har vi vant oss vid en restriktiv alkoholpolicy. Den enda marknadsföring vi ser om alkohol är sådan som förklarar alkoholens skadliga effekter på samhället och individen. Vi är vana vid att rätta oss efter Systembolagets öppettider men har också många gånger besvärats av att behöva gå omvägen via ”systemet” när man går och handlar sin mat. I många andra länder ser det annorlunda ut. En kort tågresa över Öresundsbron och man kan enkelt handla kylda alkoholhaltiga drycker från kiosker dygnet runt.

Enligt WHO är det många länder runt om i världen som arbetar aktivt för att hämma alkoholkonsumtionen (2014, s.60). Man har punktskatter på alkohol, speciella regler för marknadsföring, utbildning om dess risker m.m. Stora resurser satsas på området, och det är förståeligt. Totalt står alkohol för 5,9 % av alla dödsfall, vilket betyder att ungefär 3,3 miljoner människor avlider varje år till följderna av alkoholkonsumtion. Utöver dessa dödsfall står alkohol för 5,1 % av den globala bördan av sjukdomar och skador och är en direkt bidragande faktor till mer än 200 sjukdomar (WHO, 2014, s.2). De skadliga effekterna av alkohol är alltså mycket omfattande och kostsamma, såväl ekonomiskt som socialt. Då hög konsumtion innebär ökad risk för sjukdom (WHO 2014, s.4) undrar vi om det är möjligt att via regressionsanalys se vilka lagar och ekonomiska variabler som har en signifikant effekt på alkoholkonsumtionen i ett land. Det totala drickandet skiljer sig markant mellan länder och även om en stor del säkert beror på kulturella skillnader är det kanske möjligt att dra generella slutsatser om vilka restriktioner som i genomsnitt fungerar väl. Lyckas vi hitta sådana skulle det kunna betyda att det finns sätt att motverka alkoholkonsumtionen som fungerar för flertalet länder och oberoende av kulturella skillnader. Vi intresserar oss även för de ekonomiska variablerna då dessa skulle kunna beskriva en del av konsumtionen som kan vara svår att påverka för beslutsfattare, men kan verka som indikatorer på en eventuell framtida uppgång i befolkningens drickande.

## 1.2 Syfte och frågeställning

Som ovan nämns ämnar vi undersöka vilken typ av lagstiftning och vilka av de ekonomiska variablerna som har en statistiskt signifikant inverkan på den genomsnittliga konsumtionen av alkohol. Resultaten från denna del av analysen skulle också kunna ge en fingervisning om huruvida alkoholkonsumtionen ligger inom beslutsfattares kontroll eller ej. Frågeställningen vi jobbar efter är således:

- Hur påverkar lagstiftning och ekonomiska variabler alkoholkonsumtionen i OECD:s medlemsstater?

# 2 Data

Datamaterialet som används kommer från OECD:s databank (OECD, 2014), WHO:s rapport *Global status report on alcohol and health* (2014). Att data från olika källor kombineras beror på att vissa av de variabler som inkluderas i analysens modeller redovisas väldigt bra av den ena organisationen, medan andra av de variabler som inkluderas i vår undersökning redovisas bättre av den andra. Generellt kan man säga att OECD:s data används för alla variabler som redovisas i exakta siffror, så som pris, konsumtion av alkohol och BNP per capita. Eftersom det krävs en datanivå motsvarande minst en intervallskala för att skatta en modell som inte enbart innehåller kategorivariabler, begränsar vi oss i vår undersökning till OECD-länderna då det finns tillräckligt precis data för dessa länder.

WHO:s data, däremot, är väldigt högupplöst och uppdelad i en mängd kategorier. Där kan man bland annat finna ländernas åldersgränser för att köpa alkohol i barer, vilka länder som har restriktiva lagar angående marknadsföring av alkohol och så vidare. WHO:s data används primärt för att kategorisera datamaterialet. Detta gör att modellerna som skattas under analysen innehåller både kategoriska och kontinuerliga variabler.

## 2.1 Val av variabler

De förklarande variabler som inkluderas i modellerna har valts ut på grund av deras akademiskt bekräftade samband till alkoholkonsumtion. Av de artiklar och rapporter som skrivits i ämnet är det dock svårt att få en klar bild av vilka samband som är starkast och har stor inverkan på konsumtionen av alkohol. Till viss del beror detta på att analysen i många av de undersökningar som görs handlar om ett fåtal individer. Detta gör att resultatet inte nödvändigtvis kan generaliseras till en större population, så som ett helt land. Mer om detta problem finns i diskussionsdelen som avslutar denna uppsats. Målet för denna uppsats är, som tidigare nämnts, att försöka ta reda på vilken av de inkluderade variablerna som bidrar mest till att öka konsumtionen av alkohol och här nedan listas de variabler vi valt, samt referenser till relevant forskning.

- Pris på alkohol och alkoholkonsumtion. I datamaterialet som används har skatt på alkohol använts för att mäta detta. Det normala sättet att reglera priset på alkohol från ett policyperspektiv är att införa alkoholskatt och det ska ha en negativ påverkan på alkoholkonsumtionen (Chaloupka et al., 2002).
- Alkoholkonsumtion och ekonomisk situation. I datamaterialet undersöks sambandet mellan vissa ekonomiska variabler och alkoholkonsumtion. Dessa inkluderar arbetslöshet, BNP-tillväxt och BNP per capita. Kort sagt kan man säga att alkoholkonsumtionen

tycks vara något högre i länder med hög BNP/capita. Dessutom verkar alkoholkonsumtionen öka under högkonjunktur. För mer om hur dessa variabler påverkar alkoholkonsumtion se Ruhm och Black (2002), Ruhm (2005) samt Brenner (1975). OECD:s data används som källa till siffrorna för dessa variabler.

- Alkoholkonsumtion och marknadsföring. Marknadsföring av alkohol tycks ha en positiv effekt på konsumtionen. För studier i området se Smith och Foxcroft (2009) samt Henriksen et al. (2008)

- Alkoholkonsumtion och monopol på försäljning. Som bakgrund till denna uppsats används en rapport som visar att försäljningsmonopol på alkohol har positiva effekter på samhällshälsan på grund av minskad alkoholkonsumtion. Se Norström et al. (2010).

- Ålder då drickande börjar och senare konsumtion av alkohol. Pitkänen et al. (2005) visar att det finns ett negativt samband mellan åldern då människor börjar dricka och hur mycket de dricker senare i livet.

## 2.2 Variabelnamn

För att läsaren enkelt ska kunna tyda våra modeller och resultat följer här en lista med de arbetsnamn variablerna fått samt vad de faktiskt representerar.

- *Registrerad konsumtion* Registrerat antal liter ren alkohol konsumerad per år och person över 15 år. För information om hur denna variabel mäts se OECD (2014)

- *Icke-registrerad konsumtion* Estimerad icke registrerad alkoholkonsumtion, liter ren alkohol konsumerad per år och person över 15 år. Här hamnar t.ex. laglig och olaglig hembryggning/bränning samt icke förtullad importerad alkohol. För information om hur denna variabel mäts se OECD (2014). Bland annat har enkätundersökningar gjorts vilket riskerar att underskatta den faktiska icke-registrerade konsumtionen

- *Monopol* Kategorivariabel som säger om det finns försäljningsmonopol eller ej. Varje land som har monopol på antigen öl, vin, sprit eller samtliga drycker tilldelas en etta. Variabeln mäts på detta sätt på grund av att datamaterialet delas in i väldigt små grupper om man tar hänsyn till att monopolet kan gälla olika dryckestyper.

- *Åldersgräns* Åldersgräns för att köpa öl på barer.

- *Gräns för rattfylleri* Promillegräns för rattfylleri, blodkoncentration.

- *Förbud mot offentligt drickande* Kategorivariabel som delar in länderna i grupper baserat på om det finns en lag som förbjuder drickande på offentlig plats eller ej.

- *BNP/capita* BNP per capita, mätt i köpkraftsjusterade dollar.

- *Logaritmerad BNP/capita* Den naturliga logaritmen av föregående variabel.

- *BNP-tillväxt* Ekonomisk tillväxt som mäts i procentenheter.

- *BNP-gap* Procentuell skillnad mellan potentiell och reell BNP. Ett positivt värde tyder på att landet befinner sig i högkonjunktur och vice versa

- *Arbetslöshet* Arbetslöshet mätt i procent av den arbetsföra delen av befolkningen.

- *Restriktion mot marknadsföring* Kategorivariabel som delar upp länder i grupper som har totalförbud mot marknadsföring av alkoholhaltiga drycker (1) och de som ej har totalförbud (0). Länder som har restriktioner kring marknadsföring av exempelvis sprit men ej mot öl tilldelas en nolla.

- *Punktskatt* Punktskatt på alkohol som läggs på producenterna (destillerierna). Mäts i köpkraftsjusterade dollar per producerad liter hundraprocentig sprit. Denna skatt översätts i förlängningen till priset slutkonsumenten möter, eftersom producenter och återförsäljare måste täcka de extra skattekostnaderna.

I de fall vi analyserar tvärsnittsdata tas samtliga variablers värden från år 2010. Årtalet är valt av den anledning att det är det senaste år då alkoholkonsumtionen finns sammanställd för samtliga länder i urvalet.

# 3 Metod

## 3.1 Linjär regression

Fördelarna med linjär regression har i mångt och mycket att göra med modellens enkelhet samt att många samband faktiskt är just linjära. Parametrarnas tolkning är lätt att förstå och minstakvadratmetoden har många trevliga matematiska egenskaper. Nackdelar med modellen inkluderar bland annat att det måste finnas ett linjärt samband mellan de oberoende och den beroende variabeln. Även om många samband är linjära så finns det även gott om icke-linjära sådana, varför tanken på en linjär modell ibland får överges. En annan nackdel som gör att den linjära modellen inte alltid är direkt applicerbar på ett datamaterial är de antaganden som måste vara uppfyllda för att modellen ska vara passande. Dessa antaganden innebär bland annat att residualerna är normalfördelade, att det inte finns heteroskedasticitet bland residualerna och helst ska det heller inte finnas någon multikolinjäritet mellan de förklarande variablerna. Detta gör att en linjär modell ibland inte passar ett datamaterial som inte bearbetats innan analysen tar sin början.

För att avgöra vilka variabler vi ska inkludera i våra modeller använder vi oss av stegvis regression och bästa delmängdsregression. Stegvis regression finns implementerat i R-paketet *MASS* (Venables & Ripley, 2002) och är en iterativ process som går ut på att man lägger till och tar bort variabler om vartannat för att se om de hjälper till att uppfylla ett givet kriterium. I denna uppsats använder vi oss av *Akaikes infromationskriterium* (AIC). AIC gör en relativ bedömning av en modell där förklaringsgrad och antal variabler vägs in. Den modell med högst förklaringsgrad relativt antal variabler (lägst AIC) anses således vara den bästa. AIC beräknas med följande metod:

$$AIC = e^{\frac{2k}{n}} * \frac{RSS}{n} \qquad (1)$$

där $k$ är antal frihetsgrader, $n$ antal observationer och *RSS* är residualkvadratsumman (Gujarati & Porter, 2009, s.494). Metoden beskriven är något omstridd. Den kan på ett enkelt sätt testa ett stort antal variabler och således ökar sannolikheten att man hittar samband som egentligen inte existerar samtidigt som risken för massignifikans ökar. De variabler som vi testar mot är dock noga utvalda, det finns forskning (presenterad i avsnitt 2.1) som stödjer våra val av förklarande variabler, varför detta inte bör vara något problem.

Bästa delmängdsregression går ut på att, givet de variabler vi har, testa alla olika kombinationer av variabler för att finna en tillfredställande modell. Metoden presenterar den bästa modellen med en förklarande variabel, den bästa med två förklarande variabler o.s.v. Det är sedan upp till statistikern

att själv välja vilket kriterium som ska användas i rangordningen samt hur många variabler man vill ha med i sin modell. Funktionen för bästa delmängdsregression finns i R-paketet *leaps* (Lumley, 2009).

När de båda metoderna är genomförts jämför vi modellen vi fått från den stegvisa regressionen med modellerna från den bästa delmängdsregressionen. Vårt mål är att få en hög förklaringsgrad, signifikanta parametrar och en relativt lätthanterlig modell. Detta innebär att en avvägning får göras, där man noga funderar över varje variabels bidrag till hela modellen. Då frågan om *vilka* variabler som påverkar alkoholkonsumtionen ska besvaras kommer vi så långt det går enbart ha med signifikanta variabler. I denna uppsats anses resultat vara signifikanta om våra p-värden understiger 0,1. Anledningen till att denna höga signifikansnivå väljs är att det finns många kulturella och sociala faktorer som har ett samband med alkoholkonsumtionen. Exempelvis är många skandinaviska högtider, så som midsommar, nära förknippade med stora intag av alkoholhaltiga drycker. I muslimska länder, så som Turkiet, ser högtidsfirandet annorlunda ut. Faktorer som dessa kan vara mycket svåra att mäta och deras påverkan svår att uppskatta, varför de inte förekommer i de data som används. Med tanke på detta anser vi att en signifikansnivå på 0,1 är tillräcklig för uppsatsens syfte.

För att validera modellerna kommer följande steg kontrolleras.

- Homoskedasticitet:
  Vid linjär regression antas residualerna ha samma varians. Först görs en visuell analys där residualerna plottas. Kontrollen fortgår med hjälp av Breusch-Pagans test för heteroskedasticitet (Gujarati & Porter, 2009, s.386) som testar om residualerna av den skattade modellen har samma varians.
- Normalfördelade residualer:
  Residualerna ritas in i ett normalfördelningsdiagram och ett Shapiro-Wilks-test genomförs för att testa antagandet om normalfördelade residualer.
- Multikolinjäritet:
  Om det råder kraftig multikolinjäritet mellan de förklarande variablerna så blir tolkningen av betaskattningarna annorlunda. Man kan inte längre se varje parameter som en marginaleffekt eftersom alla variabler samvarierar. För att testa detta mäts en inflationsfaktor för variansen (VIF).

## 3.2 Logistisk regression

Intuitivt är idén bakom logistisk regression relativt enkel och tilltalande. I grunden antas fortfarande att den beroende variabeln kan predikteras av en linjär kombination av $\beta$-värden och $X$-variabler, men man vill nu använda

de förklarande variablerna för att dela in materialet i kategorier, vilket innebär att den här typen av regressionsanalys typiskt används då den beroende variabeln är av kvalitativ typ. Exempelvis skulle man kunna använda logistisk regression för att undersöka faktorer som påverkar huruvida en bank går i konkurs eller om en behandling av en patient är framgångsrik. Detta innebär att den linjära modellen av betaparametrar och förklarande variabler ska förknippas med en sannolikhet mellan 0 och 1 oavsett vilket värde den antar. Detta innebär naturligtvis att det linjära uttrycket måste transformeras. Hur detta ser ut visas i Appendix A

I denna uppsats delas datamaterialet in i två kategorier, en högkonsumtionskategori och en lågkonsumtionskategori. Ungefär hälften av länderna i datamaterialet klassificeras som högkonsumtionsländer där mycket alkohol konsumeras i genomsnitt, och den andra hälften av länderna har klassats som lågkonsumtionsländer, där folk har en lägre genomsnittlig konsumtion av alkohol. Vad som faktiskt är ett ”högkonsumtionsland” är svårt att säga, men i den här analysen har skiljelinjen dragits vid OECD-ländernas mediankonsumtion. Högkonsumtionsländerna tilldelas värdet 1 och lågkonsumtionsländerna, vilka inkluderar medianlandet, tilldelas värdet 0. Logistisk regression används för att ta reda på vilka av de förklarande variablerna som mest påverkar sannolikheten att hamna i den högkonsumerande kategorin.

Metoden som används påminner på många sätt om tillvägagångssättet vid linjär. För att välja variabler använder vi stegvis och bästa delmängdsregression. Vid valideringen av modellerna skiljer sig dock förfarandena åt. Eftersom man vid logistisk regression skattar sannolikheter för att en observation tillhör en viss grupp kan man inte använda sig av en vanlig förklaringsgrad. Istället används ett så kallat *Pseudo-$R^2$* (Hilbe, 2009, s.243-245). Detta mått är en kvot med följande utseende: $1 - LL_F/LL_C$, där $LL_F$ är log-likelihoodfunktionen för den skattade modellen med alla förklarande variabler och $LL_C$ är log-likelihoodfunktionen för en modell med enbart intercept. När detta mått ska tolkas så bör följande saker hållas i minnet: En sannolikhet antar värden mellan 0 och 1, så en logaritmerad sannolikhet antar värden mellan $-\infty$ och 0. Detta innebär att om en modell är väldigt osannolik, så kommer logaritmen av sannolikheten förknippad med denna modell att anta ett stort negativt värde. För en sannolik modell är förhållandet självfallet motsatt: logaritmen av en sannolikhet förknippad med en trolig modell antar ett värde närmare nollan. Om man jämför modeller som skattats på samma data så kommer alltså den med högst pseudo-$R^2$ att vara bäst, eftersom kvoten man subtraherar då antar ett litet värde. Underförstått är alltså att modellen som skattats, den med de förklarande variablerna, är mer sannolik och bättre än den med enbart intercept.

Många av de antaganden som gäller angående normala linjära modeller är inte längre nödvändiga, men det finns fortfarande en del saker som bör vara uppfyllda för att modellen ska kunna skattas ordentligt.

- Den beroende variabeln ska vara dikotom om binär logistisk regression genomförs. Om ordinal logistisk regression används så ska den beroende variabeln vara av ordinalskala. Att modifiera den beroende variabeln för att nå en lägre datanivå kan innebära stora informationsförluster.
- Feltermerna ska vara oberoende. Om detta inte stämmer så är modellen felspecificerad. (Hilbe, 2009, kap. 4 och 7)
- Det ska finnas ett linjärt samband mellan logoddsen och de förklarande variablerna. Om detta inte stämmer så kommer testerna inte mäta förhållandet på rätt sätt och inga slutsatser kan dras.

När man utvärderar en modell krävs det att man tittar på residualerna och ett förtydligande om vad residualerna faktiskt är i fallet med logistisk regression bör göras här. I den linjära regressionen är både den beroende och de förklarande variablerna av kontinuerlig typ, varför man kan tolka residualerna som en regelrätt avvikelse mellan förväntat och faktiskt y-värde. I det logistiska fallet, då den beroende variabeln antar värdena 0 eller 1 och de förklarande variablerna ofta är kontinuerliga, kan man inte längre tolka residualerna som samma typ av avvikelse. På grund av detta kan man oftast inte använda Q-Q-plot och normalfördelningstest för att utvärdera modellens passform. Dessutom finns det flertalet olika sorters residualer att plocka fram. Den typ som analyseras här är *Pearson Standardized Residuals*, vilka är skillnaden mellan förväntad och faktisk frekvens. Enkelt uttryckt räknas residualerna ut genom att estimera hur många länder det borde finnas i varje kategori och jämföra denna siffra med den faktiska, observerade frekvensen. (Hilbe, 2009, s.271) Det kan kort konstateras att denna typ av residualer är approximativt normalfördelade om man har det som kallas grupperad data, vilket innebär att inga kontinuerliga variabler ingår i modellen. I denna uppsats används dock individuell data i modellen då vissa av våra förklarande variabler är kontinuerliga. För att utvärdera modeller med individuell data används ett Hosmer-Lemeshows test för att se om modellen är väl specificerad. (Hilbe, 2009, s. 249ff) I detta test innebär $H_0$ att modellen passar datamaterialet. Signifikans tyder alltså på en felspecificerad modell. Testets utformning återges nedan.

$$\sum_{l=0}^{1}\sum_{g=1}^{G}\frac{\left(O_{gl}-E_{gl}\right)^2}{E_{gl}} \qquad (2)$$

När Hosmer-Lemeshows test räknas börjar man med att räkna ut varje lands sannolikhet att hamna i högkonsumtionskategorin med hjälp av de skattade betavärdena. $l = 0$ eller $1$ inkluderas för att datamaterialet delas in i två kategorier baserat på hur mycket befolkningen i länderna dricker i genomsnitt. När detta gjorts rangordnas länder utefter deras sannolikheter och delas upp i grupper. Normalt används tio grupper, en för varje decil. De länder som har de tio procent lägsta sannolikheterna blir grupp 1, eller den första decilen. När grupperna är klara tittar man på länderna inom varje decil och räknar ut medelvärdet av deras sannolikheter, $\pi_g$. Denna siffra multipliceras sedan med antalet länder för att få fram det förväntade antalet länder i varje decil, $E_g$, vilket alltså är lika med $N_g * \pi_g$. Denna siffra jämförs sedan med det observerade antalet, $O_g$. Testet jämförs med en $\chi^2$-fördelning med $G - 2$ frihetsgrader.

## 3.3 Paneldata

Regressionsanalys är också applicerbar på paneldata, det vill säga data som innehåller både tid- och tvärsnittsaspekter. Här nedan följer en kort förklaring till hur en modell formuleras i ett av de enklare fallen. En aning mer teknisk förklaring finns i Appendix B, där det också förklaras hur parametrarna skattas. Kortfattat kan man säga att man gör upprepade tvärsnittsobservationer av samma individer under en viss tidsperiod, vilket skapar en panel med data. Detta medför mycket naturligt att det finns flera typer av variation att ta hänsyn till. Residualerna i paneldatamodeller kan skrivas

$$\varepsilon_{it} = \alpha_i + u_{it}$$

U:et betecknar här den vanliga residualen, det vill säga den som uppstår för individ *i*. Ett extra index måste dock läggas till, *t*, eftersom det kommer uppstå en ny residual för individ *i* vid varje tillfälle ett nytt tvärsnitt dras. Alltså: u:et är residualen som hör till individ *i* vid tvärsnittstillfälle *t* och $\alpha$ är individ *i*:s tidsoberoende fel. När man tar fram ett skattat värde kommer man missa det sanna värdet med en viss faktor, och det är denna faktor som är $\alpha_i$. Det totala felet består alltså av en tidsberoende del, som varierar för varje tidpunkt panelen undersöks, och en tidsoberoende del som blir kvar när skattningarna görs.

Vad gäller antaganden bakom modellen som måste vara uppfyllda kan man kort och gott säga att samma sak gäller nu som för linjär regression. Detta är ganska naturligt då man även här skattar en linjär modell. Tidsaspekten som inkluderas i modellen innebär dock att antagandet om oberoende residualer måste undersökas extra noggrant eftersom autokorrelation ofta förekommer.

När man skattar modellens parametrar finns det lite olika vägar att gå. Konventionellt pratar man om fixa effekter eller slumpmässiga effekter.

Fixa effekter innebär att man tänker sig den tidsoberoende delen av feltermen som individuella intercept för alla *i*, och därför endast tar hänsyn till tidsvariationen när man skattar modellens parametrar. Slumpmässiga effekter innebär att man tar hänsyn till dels tidsvariationen, men också den slumpmässiga variationen som uppstår mellan individ *i* och *j*. De individberoende skillnaderna ses alltså som en slumpmässig felkälla. För att avgöra om fixa eller slumpmässiga effekter föreligger kan ett Hausman-test genomföras (Baltagi, 2013, s.76ff). Hur själva testet går till kommer inte förklaras i detalj här, men i korta drag gör man en skattning av både en slumpmässiga effekter- och en fixa effekter-modell och jämför deras parameterskattningar. Om skillnaden är positiv och tillräckligt stor förkastas nollhypotesen att slumpmässiga effekter föreligger. Testet är asymptotiskt $\chi^2$-fördelat. En mer intuitiv beslutsprocess kan baseras på reflektion kring feltermen. Om man misstänker att felet inte bara är vitt brus, utan att det finns någon bakomliggande anledning till skillnaden mellan individerna – det vill säga denna del av den totala variansen som inte är helt slumpmässig – så ska man använda fixa effekter. Ett typexempel på detta är när man gör jämförelsen mellan just länder.

När man bestämt sig för huruvida man ska anta slumpmässiga eller fixa effekter går man vidare med att undersöka om det föreligger tids- eller individuella effekter. Detta testas med ett F-test på liknande sätt som i Hausman-testet. Modeller med antagande om individuella respektive tidseffekter skattas och testas sedan mot varandra.

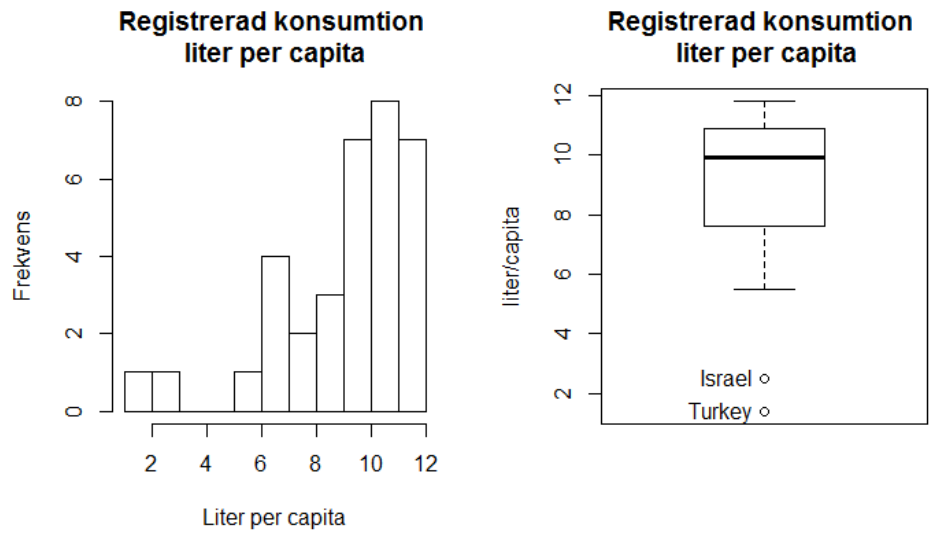
Då det fortfarande är en linjär modell man skattar så sker modellvalideringen på samma sätt som beskrivs i 3.1.

# 4 Analys

Flera tekniker appliceras på datamaterialet. I detta kapitel kommer deras användningsområde, viss matematik, resultat och tolkning av modellernas $\beta$-värden att förklaras. En teoretisk bakgrund till de modeller som används finns att hitta i Appendix.

## 4.1 Beskrivning av data

För analysen sammanställs data från två olika källor: WHO och OECD. Syftet är att förklara vilka av de förklarande variablerna som bidrar mest till den registrerade alkoholkonsumtionen och valet av variabler baseras som tidigare nämnts på de rapporter som studerats innan analysen tar sin början. För att säkerställa att data av tillräckligt god kvalitet begränsas analysen till OECD:s 34 medlemsstater. Figur 4-1 ger en grundläggande beskrivning av de inkluderade ländernas alkoholkonsumtion.




*Figur 4-1. Histogram och lådagram av registrerad konsumtion för OECD:s medlemsländer. Israel och Turkiet skiljer sig tydligt från övriga OECD-länder.*

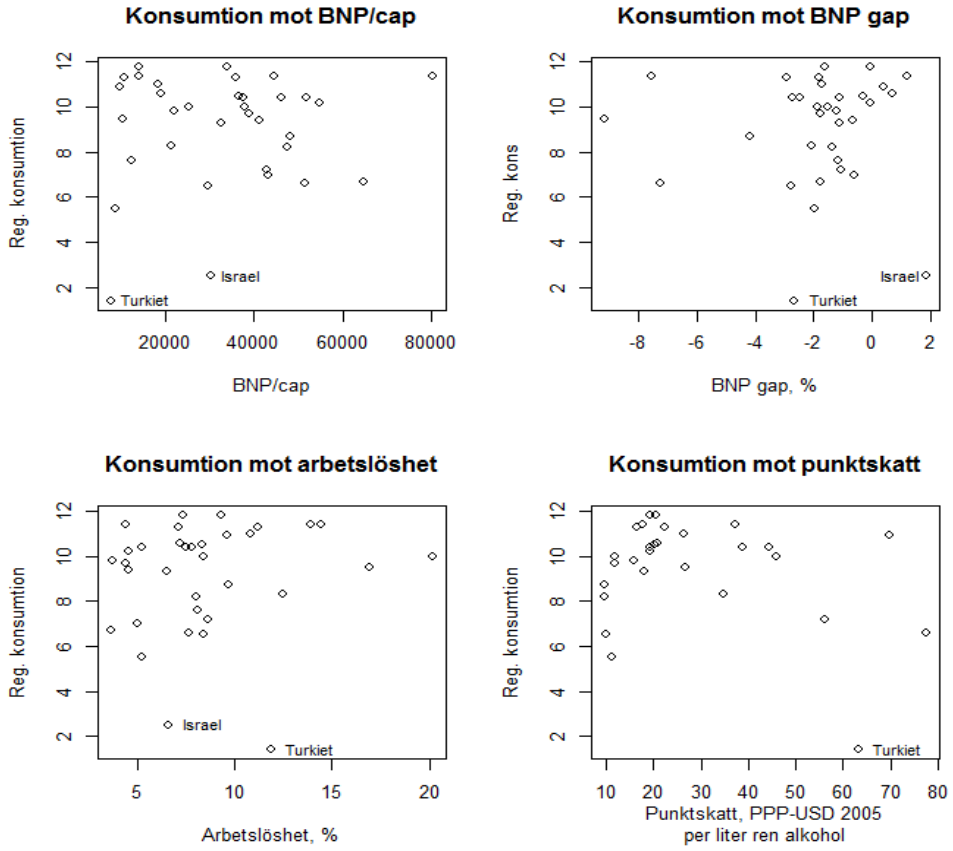
Vår data är som synes relativt homogen och de flesta länders konsumtion ligger mellan fem och tolv liter ren alkohol per person och år. Två av länderna, Turkiet och Israel, klassas som uteliggare och övriga observationer är samlade. Det är svårt att göra något fördelningsantagande utifrån histogrammen i Figur 4-1, men vidare analys med normalfördelningsdiagram och Shapiro-Wilks test visar dock inte på en normalfördelning.

Enligt tidigare gjorda studier skall skattenivån på alkohol ha påverkan på efterfrågan och därmed konsumtionen (Chaloupka et al., 2002). Vi har därför inkluderat en variabel med ländernas alkoholskatter De flesta länderna i undersökningen har alkoholskatter utformade på så sätt att en viss summa läggs på varje liter ren alkohol som produceras. I vår undersökning är det dock sex länder som använder en annan metod för att punktskatta alkohol. Länderna som valt att beskatta alkohol på detta alternativa sätt är Chile, Israel, Japan, Luxemburg, Nederländerna och Norge. I dessa fall läggs skatten på försäljningspriset som en procentsats på det momsgrundande beloppet, med olika procentsatser beroende på dryckens alkoholkoncentration. Detta medför alltså olika skattesatser beroende på om det handlar om öl, vin eller sprit. Då länderna i datamaterialet har skilda tillvägagångssätt för beskattning av alkohol blir en jämförelse länder emellan svår att genomföra på denna punkt. Analys genomförd på den större grupp länder med samma beskattningsmetod visar dock inte på någon signifikant påverkan. För att inte tappa observationer i onödan utesluts denna variabel.

## 4.2 Modeller

### 4.2.1 Linjär regression

För att avgöra huruvida linjära samband föreligger mellan beroende och oberoende variabler görs först en visuell analys av datamaterialet. Variabler plottas mot varandra och i Figur 4-2 finns ett axplock av denna visuella analys.




*Figur 4-2. Registrerad konsumtion mot given variabel. Om linjära samband föreligger så är de i vart fall inte starka. Notera även hur Turkiet och Israel tydligt sticker ut. Israels skattemetod gör att landet ej finns representerat i sista spridningsdiagrammet.*

Utöver de diagram som redovisas i Figur 4-2 görs liknande analys med övriga variabler. Det är dock svårt att urskilja några starka linjära samband. För att se om det finns några lämpliga transformationer för att råda bot på detta problem genomförs ett Box-Cox-test, men detta visar inte på att det skulle finnas några lämpliga transformationer att göra. Som en utgångspunkt i analysarbetet så skattas alltså en linjär regressionsmodell. Modellen estimeras med förfarandet beskrivet i metodkapitlet (3.1). Glädjande nog resulterade den *stegvisa regressionen* och den *bästa delmängdsregressionen* i samma modell.

Resultaten från beräkningarna av den linjära regressionsmodellen redovisas i Tabell 4-1.

*Tabell 4-1. Regressionsmodell skattad med stegvis regression. Regression samt parametrar är signifikanta och modellen beskriver 34 % av alkoholkonsumtionen.*

| **Variabel** | $\widehat{\boldsymbol{\beta}}$ | **Medelfel** | **t-värde** | **Pr(>\|t\|)** | |
|---|---|---|---|---|---|
| Intercept | 7,147e+00 | 1,526e+00 | 4,683 | 6,11e-05 | *** |
| Icke-registrerad konsumtion | 1,295e+00 | 6,711e-01 | 1,930 | 0,0635 | . |
| Monopol | -3,482e+00 | 1,317e+00 | -2,644 | 0,0131 | ** |
| BNP/capita | 5,153e-05 | 2,861e-05 | 1,801 | 0,0821 | . |
| BNP-tillväxt | -3,017e-01 | 1,459e-01 | -2,068 | 0,0476 | * |

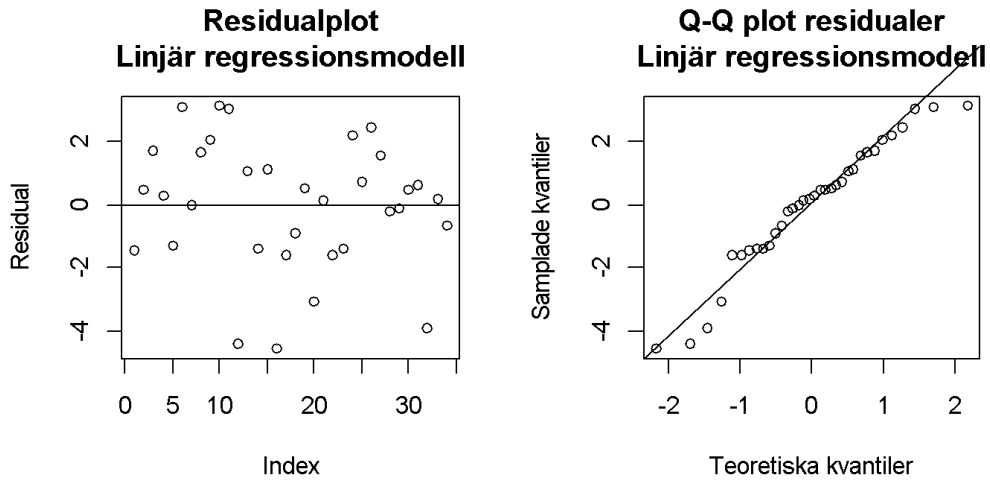
--- Signif. koder: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residualernas medelfel: 2.1661 med 29 frihetsgrader
$R^2$:0,3417
Justerat $R^2$:0,2509
F-värde: 3,12 med 4 och 29 FG. P-värde:0,02988.

Vår modell blir alltså:

$$y_i = 7{,}15 + 1{,}30x_1 - 3{,}48x_2 + 0{,}00005x_3 - 0{,}30x_4 \qquad (3)$$

Registrerad konsumtion ökar alltså med 1,30 liter per varje liter icke registrerad sprit som personen dricker per år. Detta tolkar vi som att tillgången på icke-reglerad alkohol snarare är ett komplement än ett substitut till reglerad alkohol. Den registrerade konsumtionen ökar också med 5 cl per person och år för varje tusen dollar högre BNP/capita landet har. Konsumtionen sjunker med de två resterande variablerna, monopol och BNP-tillväxt. I länder med monopol konsumerar varje person i snitt 3,48 liter mindre 100-procentig sprit per år. Dessutom framgår det att den registrerade konsumtionen sjunker med ungefär 3 dl hundraprocentig sprit per procentenhet ett lands BNP ökat sen tidigare år. Det kan skapa viss förvirring att det finns två förklarande variabler som har med högre BNP att göra och att dessa har koefficienter med olika tecken. Hur mycket vikt som ska läggas vid BNP-variablerna är dock svårt att säga. Den linjära modellen som skattas här är baserad på tvärsnittsdata, och för att få en bättre och mer rättvis bild av sambandet mellan BNP-variablerna och den registrerade alkoholkonsumtionen bör man titta på hur de två variablerna utvecklas över tiden. Ett liknande, något förvirrande resultat nås också av Brenner (1975). Enligt denna undersökning så uppvisar alkoholkonsumtion på kort sikt ett negativt samband till ekonomisk tillväxt, men på lång sikt är förhållandet det motsatta. Enligt Brenner så beror detta till stor del på att olika typer av rusdrycker dricks beroende på tidens ekonomiska läge. Under sämre tider dricks mer starksprit. Denna variabel kommer undersökas närmare och från ett längre tidsperspektiv i den kommande modellen som baseras på paneldata i avsnitt 4.2.4.

Modellspecifikationerna i Tabell 4-1 visar på signifikant regression och signifikanta variabler. Förklaringsgraden är $R^2$=0,3417. Vi lyckas alltså beskriva ungefär 34 % av den registrerade alkoholkonsumtionen med hjälp av våra variabler. BNP/capita och Icke-registrerad konsumtion är signifikanta på nivån 10 %, vilket vi anser vara acceptabelt. Övriga variabler är signifikanta på en signifikansnivå motsvarande minst α=0,05. Modellens residualer återfinns i Figur 4-3 och som synes avviker de inte signifikant från normalfördelningen.




*Figur 4-3. Analys av residualerna av modell (3).*

*Tabell 4-2. Test för normalfördelade residualerav modell (3). Resultatet tyder på normalfördelning i feltermen.*

| **Shapiro-Wilk normalitetstest** | |
|---|---|
| **W** | **p-värde** |
| 0,963 | 0,3112 |

VIF-värdena (Variance Inflation Factor) i Tabell 4-3 är mått på kolinjäritet. Om detta värde är mindre än 2,5 så anser man normalt att det inte finns några problem med multikolinjäritet.

*Tabell 4-3. Test för multikolinjäritet, modell (3).Samtliga värden under 2,5*

| | **Icke-registrerad konsumtion** | **Monopol** | **BNP/cap** | **BNP-tillväxt** |
|---|---|---|---|---|
| **VIF** | 1,4 | 1,5 | 1,6 | 1,1 |

Vi genomför även ett Breusch-Pagan-test för att pröva nollhypotesen att vår modells residualer är homoskedastiska.

*Tabell 4-4. Breusch-Pagan test. Resultatet tyder på att vår modells residualer är homoskedastiska.*

| **Breush-Pagan test för heteroskedasticitet** | | |
|---|---|---|
| **BP** | **FG** | **p-värde** |
| 6,983 | 4 | 0,137 |

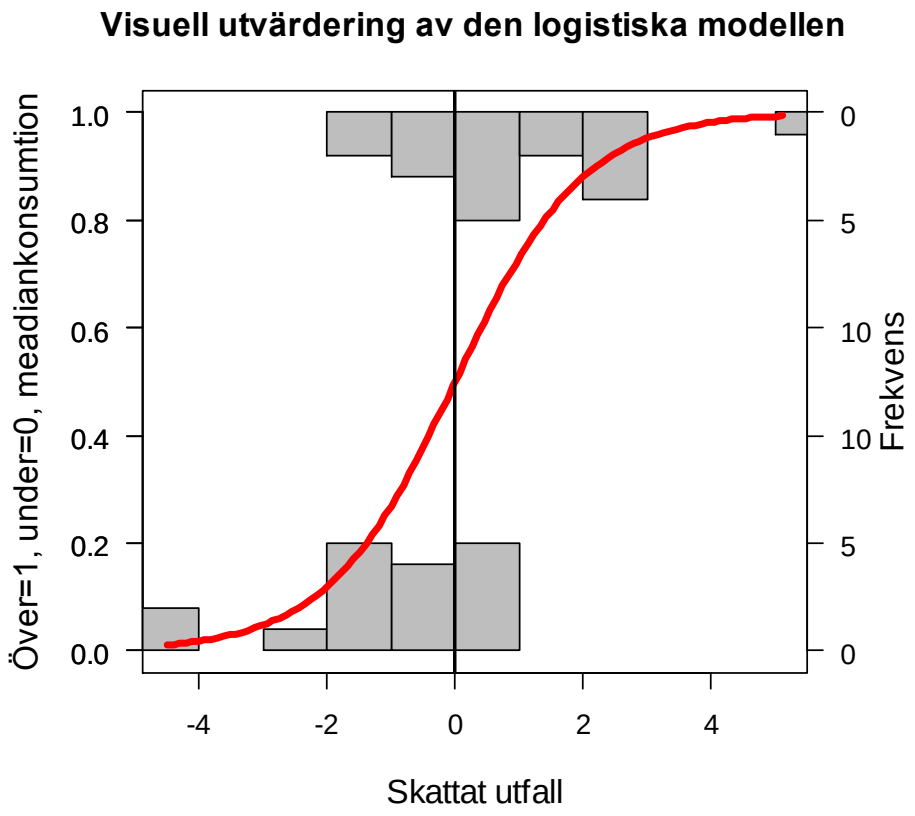
Då denna linjära modell är lätt att tolka så används den i viss mån som referens för att utvärdera huruvida de andra modellerna som skattas innebär en förbättring eller ej.

### 4.2.2 Regression utan extremvärden

I den visuella redovisningen av data i Figur 4-2 kan man se att Turkiet och Israel avviker markant från resterande länder. Vi testar att göra en regressionsanalys då Israel och Turkiet plockats bort. Detta ger dock ingen förbättrad modell.

### 4.2.3 Logistisk regression

Skattandet av den logistiska modellen gick till i enlighet med förfarandet beskrivet i metodkapitlet (3.2). Vi bör dock nämna att i detta fall fick vi vitt skilda resultat beroende på om vi använde oss av *stegvis regression* eller *bästa delmängdsregression*. Modell (4) är den vi fick genom *stegvis regression* och vi valde denna då den visade på högst *pseudo-*$R^2$ (0,23). Med *bästa delmängdsregression* innehåller den bästa föreslagna modellen enbart en förklarande variabel, nämligen åldersgräns. En visuell utvärdering av modellen finns representerad i Figur 4-4 där staplarna ska läsas mot den högra frekvensaxeln, den röda s-kurvan mäter sannolikheter och x-axeln anger värdet på logoddset. Modellens parametrar redovisas i Tabell 4-5.




*Figur 4-4. Predikterade och observerade värden från den logistiska modellen.*

*Tabell 4-5. Sammanställning av den logistiska regressionsanalysen.*

| Variabel | $\hat{\beta}$ | Medelfel | Z-värde | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| Intercept | 6,3235e+00 | 6,8475e+00 | 0,9235 | 0,35576 | |
| Icke-registrerad konsumtion | 1,2035e+00 | 7,8242e-01 | 1,5382 | 0,12401 | |
| Monopol | -2,8935e+00 | 1,6354e+00 | -1,7693 | 0,07685 | . |
| Åldersgräns | -6,4850e-01 | 3,9560e-01 | -1,6393 | 0,10116 | |
| BNP/capita | 6,3760e-05 | 3,8009e-05 | 1,6775 | 0,09345 | . |
| Arbetslöshet | 2,4912e-01 | 1,5061e-01 | 1,6540 | 0,09813 | . |

--- Signif. koder: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Den logistiska modellen blir alltså:

$$\ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = 6{,}32 + 1{,}20x_1 - 2{,}89x_2 - 0{,}65x_3 + 0{,}00006x_4 + 0{,}25x_5 \qquad (4)$$

där $\ln(P_i/(1-P_i))$ är det logaritmerade oddset att ett land skulle tillhöra högkonsumtionsgruppen.

Koefficienterna i Modell 4 visar hur mycket det förväntade logoddset för att ett land hamnar i högkonsumtionskategorin förändras när variabeln ökar

med en enhet. Detta innebär att om man till exempel inför monopol så minskar den skattade sannolikheten för landet att hamna i högkonsumtionsgruppen. Värdet är inte direkt kopplat till sannolikheten, varför det inte är något konstigt med att några koefficienters absolutvärden är större än ett och att några är negativa. Sannolikheten, $P_i$, kan anta värden mellan 0 och 1 vilket innebär att $-\infty < \ln(P_i/(1 - P_i)) < \infty$.

Av de signifikanta variablerna kan man se att införandet av försäljningsmonopol minskar risken att hamna i högkonsumtionsgruppen. På samma vis ökar denna risk när BNP/capita och arbetslöshet ökar. BNP/capita inkluderas även i den linjära modellen (3) och koefficienterna har samma tecken i båda modellerna. Arbetslöshet, däremot, ansågs ej signifikant i det linjära fallet men uppvisar här en positiv signifikant riktningskoefficient. Att icke-signifikanta variabler inkluderas i denna modell beror på att detta var det bästa som gick att åstadkomma med given data, samt att koefficienterna är nära signifikans. Dessa visar att icke-registrerad konsumtion har en positiv påverkan på risken, medan åldersgränsen har en negativ sådan. För att se om modellen är välspecificerad utförs ett Hosmer-Lemeshows-test. Detta återfinns i Tabell 4-6 och som synes är testet ej signifikant och detta tyder på en välspecificerad modell.

*Tabell 4-6. Hosmer-Lemeshows-test. Det ickesignifikanta resultatet tyder på att modellen inte är felspecificerad.*

| $\chi^2$ | FG | p-värde |
|---|---|---|
| 7,63 | 8 | 0,47 |

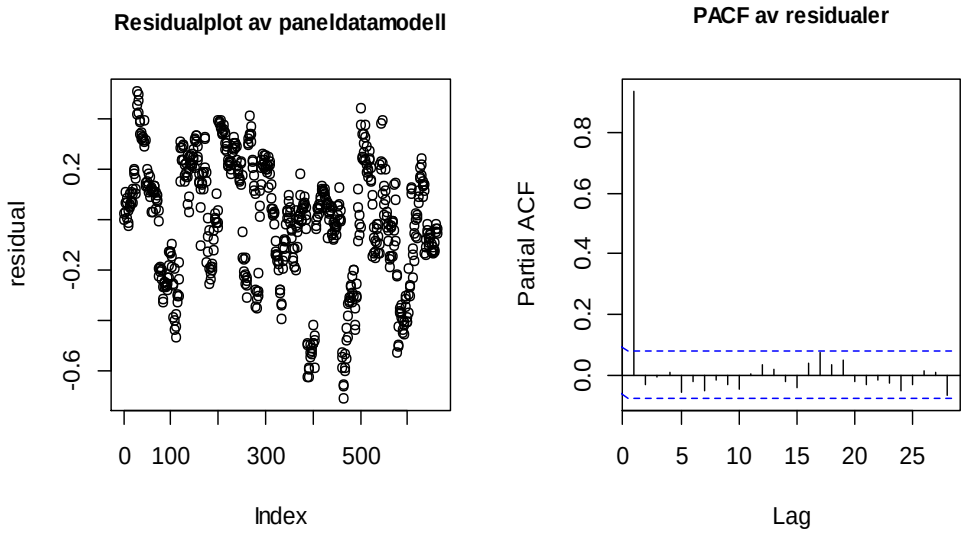
Sammanfattningsvis kan vi säga att en logistisk analys av vårt datamaterial inte ger resultat som överträffar dem i den linjära modellen (3). Det kan också anses aningen konstigt att kategorisera den boroende variabeln som ju är kontinuerlig från början. Detta angreppssätt är lite av ett experiment för att se vilka variabler som påverkar sannolikheten att befolkningen i ett land konsumerar mycket alkohol i genomsnitt. Resultaten stämmer i viss mån överens med dem som påträffas i den linjära modellen så experimentet var inte helt misslyckat. Dock tillför inte denna modell något nytt till analysen.

### 4.2.4 Paneldata

I vår tidigare analys i avsnitt 4.2.1 har vi funnit samband mellan ekonomiska faktorer och alkoholkonsumtion. Vi finner detta intressant och önskar fördjupa analysen. Bland annat fann Ruhm och Black (2002) att det finns samband mellan ekonomiska faktorer och alkoholkonsumtion. För både alkoholkonsumtion och de ekonomiska variablerna finns det årliga observationer som sträcker sig bakåt i tiden i OECDs databas. Dessa används för att sammanställa en panel av data. Den består av variablerna

*BNP/capita, BNP-gap, Arbetslöshet och BNP-tillväxt* som ställs mot *Registrerad Konsumtion*. Senaste observerade värde är för samtliga länder 2010, medan startdatum varierar från 1970-talet och framåt. För länderna Israel, Turkiet och Luxemburg saknas många historiska värden och således utesluts dessa länder ur modellen. Precis som för de andra modellerna som utvärderas i denna uppsats finns det ett paket och kommandon för att få göra paneldataanalys i R (Croissant & Millo, 2008).

Under arbetet med skattningen märker vi snabbt att vår modell lider av en kraftig autokorrelation i residualerna. Detta är relativt vanligt vid analys av paneldata. Strukturen hos datamaterialet gör att alla de problem man kan stöta på vid analys av tidserier, såväl som de problem som uppkommer vid analys av tvärsnittsdata, kan förekomma. Om autokorrelation bland feltermerna förekommer är parameterskattningarna fortfarande konsistenta, men skattningarna blir ineffektiva. De skattade medelfelen blir inte väntevärdesriktiga (Baltagi, 2013, s.96). Det finns olika ansatser för att lösa dessa problem och vi väljer att använda oss av en transformation kallad Cochrane-Orcutt-estimering (Verbeek, 2012, s.114). Tillvägagångssättet innebär att en paneldatamodell skattas, en modell vars residualer illustreras i Figur 4-5. När man analyserar residualerna i figuren syns det att dessa inte uppfyller antagandet om oberoende normalfördelade med samma varians.




*Figur 4-5. Figurerna tyder på ett tydligt beroende och PACF pekar på ett sådant vid tidslagg 1.*

Utifrån detta drar vi slutsatsen att vår felterm har följande struktur: $u_{it} = \rho u_{i(t-1)} + e_{it}$, där ρ anger korrelationen mellan residualen vid tidpunkt *t* och *t-1*. Residualerna används nu för att skatta en AR(1)-modell.

*Tabell 4-7. AR(1)-modell skattad av residualerna tyder på hög autokorrelation.*

| **Koefficient:** | $\widehat{\rho}$ | **Intercept** |
|---|---|---|
| | 0,936 | -0,0001 |
| s.e. | 0,013 | 0,0493 |

Arbetet med transformationen fortgår genom att lagga data på följande vis.

$$Y_{it} - \rho Y_{it-1} = \alpha_i(1-\rho) + \beta_i(X_{it} - \rho X_{it-1}) + u_{it} \qquad (5)$$

Efter att data laggats hoppas man att de nya feltermerna ska vara oberoende normalfördelade med lika varians. Som med alla transformationer finns här vissa nackdelar: i vårt fall blir koefficienter svårare att tolka och vi tappar första observationen för varje land, då $Y_{it=0}$ ej är observerad. Vårt datamaterial består dock av nästan 700 observationer så några förlorade observationer bör inte innebära några större problem.

De numer transformerade data används till att skatta en ny paneldatamodell. Vår slutgiltiga modell finns redovisad i Tabell 4-8 och som synes har bara *Arbetslöshet* och *BNP-gap* tagits med. Eftersom det från början endast fanns fyra variabler att undersöka var det enkelt att manuellt se huruvida de hade i modellen att göra eller ej. Vad gäller *BNP/capita* och *BNP-tillväxt* så blir de för det första ej signifikanta och för det andra är deras effekt på förklaringsgraden högst marginell. I Tabell 4-8 visas också parameterskattningarna, deras signifikansnivåer och modellens förklaringsgrad. Observera att variablerna nu är transformerade.

*Tabell 4-8 Den slutgiltiga skattade modellen av paneldata.*

Effekt="Tid", Modell="Fix"
Obalanserad Panel: n=31, N=632

| **Variabel** | $\widehat{\boldsymbol{\beta}}$ | **Medelfel** | **t-värde** | **Pr(>\|t\|)** | |
|---|---|---|---|---|---|
| Arbetslöshet | -0,061395 | 0,021908 | -2,8025 | 0,005235 | ** |
| BNP-gap | 0,033720 | 0,014040 | 2,4017 | 0,016620 | * |

Signif. koder: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

$R^2$: 0,049481

Justerat $R^2$: 0,047132

F-värde: 15,6692 med 2 och 602 FG, p-värde: 2,3237e-07

Vår paneldatamodell blir:

$$\hat{y}_i = -0{,}06x_1 + 0{,}03x_2 \qquad (6)$$

Tolkningen av modellens koefficienter är här tämligen lik den för en linjär regressionsmodell. Modellen är fortfarande linjär. Enda skillnaden är att den nu dessutom rör sig genom en tidsdimension. När koefficienterna tolkas måste man dock komma ihåg om man använder sig av slumpmässiga eller fixa effekter när man skattar modellen. I detta specifika fall är det en modell med fixa effekter som används, vilket är konventionellt när man är intresserad av att undersöka skillnader mellan länder (Verbeek, 2012). Det som undersöks är alltså hur den beroende variabeln påverkas när de oberoende variablerna varierar över tiden. Alla tidsoberoende delar av feltermen rensas ut. Koefficienterna i modell (6) ska alltså tolkas som att den genomsnittliga konsumtionen av alkohol sjunker när arbetslösheten ökar, och att den genomsnittliga konsumtionen av alkohol ökar när ett land uppvisar ett positivt BNP-gap, vilket innebär att landets reella BNP ligger över deras potentiella BNP. Detta signalerar en högkonjunktur.

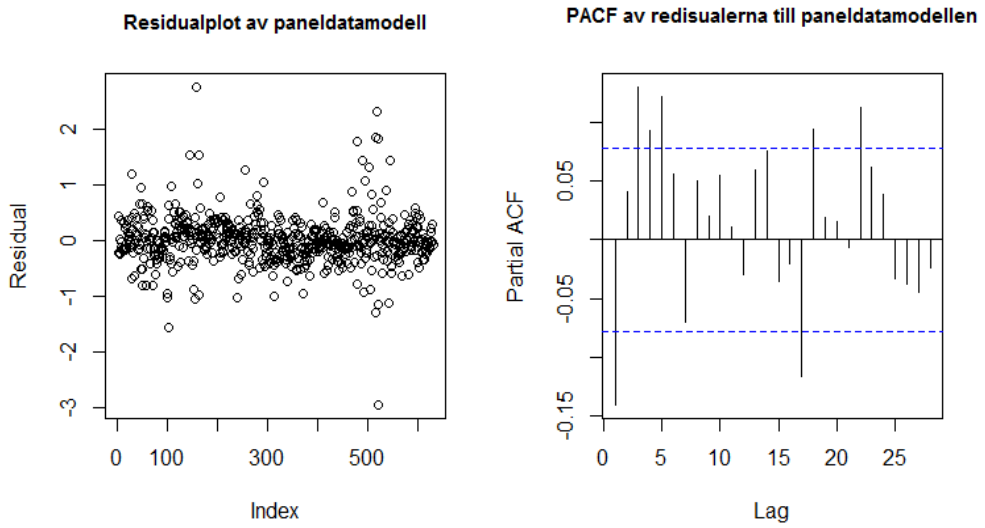
Man kan undra om inte BNP-gap och arbetslöshet egentligen mäter två sidor av samma mynt. I högkonjunktur går arbetslösheten ner och BNP-gapet är positivt. Även om variablerna är negativt korrelerade så visar analys av modellen inga problem med multikolinjäritet. Se Tabell 4-9.

*Tabell 4-9. VIF-värdena för modell (6) tyder inte på några problem med multikolinjäritet.*

| | **Arbetslöshet** | **BNP-gap** |
|---|---|---|
| **VIF** | 1,838 | 1,838 |

Båda koefficienter är signifikanta på minst α=0,05. I Tabell 4-8 syns också att förklaringsgraden ($R^2$) är ungefär 5 %, vilket inte är särskilt högt. Modellen är dock signifikant. När paneldata används erhålls ofta signifikanta modeller beroende på att antalet observationer ökar väldigt snabbt när man tar hänsyn till tidsdimensionen. $SS_E$ delas med väldigt många frihetsgrader varför $MS_E$ blir väldigt liten.

I Figur 4-6 följer en kort analys av residualerna och av de diagram som återfinns nedan kan man snabbt dra slutsatsen att det fortfarande, trots transformation, finns problem bland residualerna.




*Figur 4-6. Visuell residualanalys av paneladatamodellen.*

Precis som vanligt så önskar man ha oberoende feltermer som enbart är vitt brus. Ett signifikant beroende bland residualerna innebär att modellen är något felspecificerad och i vårt fall är modellen just detta, vilket syns på PACF:en i Figur 4-6. Det faktum att en felspecificerad modell ändå analyseras i den här rapporten är att detta fortfarande är den bästa paneldatamodellen som åstadkommits med det datamaterial som använts. Då modellen inte ämnar prediktera framtida konsumtion av alkohol, utan avser beskriva vilken effekt de förklarande variablerna har på alkoholkonsumtionen, kan man fortfarande ha viss behållning av modellens resultat. Parameterskattningarna är trots allt väntevärdesriktiga, även om deras varians inte är det.

Några rader angående modellvalet bör också inflikas. Att man skulle hitta data där det inte finns någon korrelation mellan individernas tidsoberoende fel är inte särskilt sannolikt. Troligtvis så är inte länderna skilda av en samma konstant för alla tidsperioder. När man skattar $y_{it}$ får man med andra ord inte samma $\alpha_i$ i alla tidsperioder, vilket gör att skillnaden mellan länderna är något korrelerad med den beroende variabeln. Detta problem kvarstår oavsett vilka variabler som inkluderas i modellen om endast den data som använts för den här analysen finns att tillgå. Förekomsten av detta problem leder till misstankar om att man kanske skulle använda en modell med slumpmässiga effekter istället. För att ta reda på om rätt modelltyp används så utförs ett *Hausman-test* där $H_0$ innebär att inga fixa effekter förekommer. Se Tabell 4-10.

*Tabell 4-10. Signifikant resultat indikerar att fixa effekter föreligger.*

| Hausman-test | | |
|---|---|---|
| $\chi^2$ | FG | p-värde |
| 198,9724 | 2 | < 2,2e-16 |

För vidare validering undersöks om tidseffekter faktiskt föreligger. Ett Breusch-Pagan-test genomförs för att kontrollera detta. Ett signifikant resultat indikerar att tidseffekter föreligger. Se Tabell 4-11 här nedan:

*Tabell 4-11. Breusch-Pagan-test är signifikant och tyder på att time effekter föreligger*

| Breush-Pagan-test för time effects | |
|---|---|
| $\chi^2$ | p-värde |
| 4,487 | 0,034 |

Slutsatserna angående modell (6) blir att koefficienterna, som blev signifikanta, stämmer överens med tidigare forskningsresultat, men att modellen trots detta bär en del brister. Den lyckas endast förklara cirka 5 % av alkoholkonsumtionens över den tidsperiod som undersöks. Detta betyder enkelt uttryckt att det finns en mängd faktorer som inte tas med, men som har stor betydelse för alkoholkonsumtionen. Dessutom ser vi en signifikant beroendestruktur bland residualerna, trots transformationen av data som ska motverka just detta.

# 5 Slutsats

När modellerna sammanställts och analyserats har vi observerat att den viktigaste variabeln, som lagstiftare kan kontrollera, utan tvekan är monopol på försäljningen. Koefficienten till denna variabel blir alltid signifikant och har en relativt stor påverkan på den registrerade konsumtionen. Vad gäller de ekonomiska variablerna är det svårare att säga vilken som har störst betydelse. Det visar sig dock att BNP/capita inkluderas i alla regressionsmodeller och får signifikanta koefficienter, varför vi tycker det att det är denna variabel som har störst betydelse för den registrerade alkoholkonsumtionen. Rika länder verkar med andra ord dricka mer, åtminstone baserat på vårt datamaterial. Resultatet från undersökningen gjord med paneldata visar också att länder med högkonjunktur, positivt BNP-gap och låg arbetslöshet, verkar dricka mer i genomsnitt. Man kan misstänka att det skulle föreligga multikolinjäritetsproblem här, eftersom båda variabler kan ses som instrument för högkonjunktur, men tester visar att korrelation mellan de båda variablerna inte är så hög att detta innebär definitiva problem.

För att besvara frågan som ställs i inledningen så drar vi slutsatsen att rika länder, och länder som upplever en högkonjunktur, verkar konsumera mer alkohol och politikers bästa instrument för att stävja alkoholkonsumtionen verkar otvivelaktigt vara att instifta ett försäljningsmonopol.

# 6 Diskussion

Som vanligt uppkommer vissa problem när man analyserar data som berör frågor på en samhällelig nivå. Det kan finnas fruktanvärt många aspekter som påverkar utfallet, och många av dessa kan vara kulturellt betingade och otroligt svåra att mäta.

I just detta fall har många av våra problem att göra med riktningen på koefficienterna. Variabeln arbetslöshet, till exempel, får inte samma effekt på den registrerade konsumtionen i de olika modellerna. Den kommer inte med i de linjära modellerna, i paneldata så får den en negativ koefficient och i den logistiska modellen får den en positiv. Detta kan bero på att tvärsnittet tagits vid en något märklig tidpunkt. All data är avser år 2010, då arbetslösheten var relativt hög i många rika länder, där det normalt dricks lite mer än i fattigare länder. Det kan ha varit så att vid just detta tillfälle var arbetslösheten ovanligt hög i många rikare länder där mycket alkohol konsumeras, varför arbetslösheten påverkar sannolikheten att ett land hamnar i högkonsumtionskategorin positivt. Detta är dock bara en gissning. När analysen av paneldata gjordes så fick koefficienten ett negativt värde. Detta verkar på sätt och vis rimligare, dels beroende på att det stämmer

överens med tidigare studier, men också för att det datamaterial som används för analysen av paneldata innehåller mycket fler observationer.

Ett par ord angående monopolvariabeln bör också framföras. Det kan vara så att länder där det finns ett monopol har en historia av restriktiv alkoholpolicy, så som till exempel Sverige har, och att det därför dricks mindre på grund av detta. Monopolet kan alltså vara kulmen av denna generellt restriktiva lagstiftning. Det kanske inte monopolet i sig som gör att det dricks mindre, utan en historia av restriktiv policy rörande alkohol.

Resultaten av våra modeller reser också vissa frågor angående metoden. Den linjära modellen skattas med AIC-metoden i R och ska således vara ”optimal” givet kriterier för AIC. Variabelkombinationen som kommer fram i denna modell innehåller dock långt ifrån alla av variabler som mäts i den data som används. Att många av de variabler som finns i datamaterialet inte kommer med i några av modellerna som prövas kan indikera att vi aldrig kommer kunna modellera alkoholkonsumtionen särskilt väl med det datamaterial vi har. De variabler vi väljer att mäta och inkludera har på intet sätt valts slumpmässigt – det finns uppbackning för deras samband med alkoholkonsumtion i tidigare forskning – men trots detta lyckas de inte förklara konsumtionen särskilt väl. Detta kan möjligtvis ha med källorna att göra. De flesta studier vi funnit undersöker små grupper av individer och det är inte helt säkert att man kan generalisera detta till hela populationer. De aspekter som påverkar några få individers drickande kanske inte är desamma som påverkar en hel befolknings drickande. Vi har ju baserat vårt val av förklarande variabler på dessa studier, så det finns en risk att vi helt enkelt försöker förklara den registrerade alkoholkonsumtionen med faktorer som inte påverkar den nämnvärt egentligen. Hur stor denna risk är dock svårt att bedöma. De resultat vi får stämmer trots allt i mångt och mycket överens med den tidigare forskningen. Det enda entydiga svar vi nått efter all möda är att ett monopol har en negativ effekt på alkoholkonsumtionen. De ekonomiska sambanden är inte alls lika tydliga, men det verkar onekligen vara så att BNP/capita och konjunkturen har ett samband med alkoholkonsumtionen.

I framtiden skulle det vara intressant att göra en undersökning av de sociala och kulturella faktorer som kan påverka länders alkoholkonsumtion. För att göra detta skulle det troligen passa med en mer kvalitativ undersökningsmetod. Dessutom skulle det vara intressant att se hur stora ekonomiska händelser, så som 90-talskrisen i Sverige, påverkat alkoholkonsumtionen.

# 7 Litteraturförteckning


Baltagi, B.H. (2013). *Econometric analysis of panel data*. (5. ed.) Chichester, West Sussex: John Wiley & Sons, Inc..

Brenner, M. H. (1975). Trends in alcohol consumption and associated illnesses. Some effekter of economic changes. *American Journal of Public Health, 65(12),* 1279-1292.

Chaloupka, F. J., Grossman, M., & Saffer, H. (2002). The effekter of price on alcohol consumption and alcohol-related problems. *Alcohol research and health, 26(1),* 22-34.

Croissant & Millo, 2008
Yves Croissant, Giovanni Millo (2008). *Panel Data Econometrics in R: The plm Package*. Journal of Statistical Software 27(2).
URL http://www.jstatsoft.org/v27/i02/.

Gujarati, D.N. & Porter, D.C. (2009). *Basic econometrics*. (5. ed.) Boston: McGrawHill.

Henriksen, L., Feighery, E. C., Schleicher, N. C., & Fortmann, S. P. (2008). Receptivity to alcohol marketing predicts initiation of alcohol use. *Journal of Adolescent Health, 42(1),* 28-35.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton: CRC Press.

Norström, T., Miller, T., Holder, H., Österberg, E., Ramstedt, M., Rossow, I., & Stockwell, T. (2010). Potential consequences of replacing a retail alcohol monopoly with a private licence system: results from Sweden. *Addiction, 105(12),* 2113-2119.

Pitkänen, T., Lyyra, A. L., & Pulkkinen, L. (2005). Age of onset of drinking and the use of alcohol in adulthood: a follow-up study from age 8–42 for females and males. *Addiction, 100(5),* 652-661

Ruhm, C. J., & Black, W. E. (2002). Does drinking really decrease in bad times? *Journal of health economics, 21(4),* 659-678.

Ruhm, C. J. (1995). Economic conditions and alcohol problems. *Journal of health economics, 14(5),* 583-603.

Smith, L. A., & Foxcroft, D. R. (2009). The effect of alcohol advertising, marketing and portrayal on drinking behaviour in young people: systematic review of prospective cohort studies. *BMC public health*, *9*(1), 51.

Verbeek, M. (2012). *A guide to modern econometrics*. Hoboken, NJ: John Wiley & Sons

WHO (2014), *Global status report on alcohol and health 2014*, [Online], nåsvia:http://www.who.int/substance_abuse/publications/global_alcohol_report/en/., läst 2014-09-21.

**Data**

OECD (2014), *OECD Statistics*, [Online], nås via: http://stats.oecd.org/, läst 2014-09-23.

WHO (2014), *Global Information System on Alcohol and Health*, [Online], nås via: http://apps.who.int/gho/data/node.main.GISAH?lang=en, läst 2014-09-28.

# Appendix A Logistisk regression

I detta appendix finns en teoretisk bakgrund till hur man skattar parametrarna i en logistisk regressionsmodell.

Den beroende variabeln, eller responsvariabeln, antas fortfarande kunna uttryckas som en linjär funktion av $\beta$-värden och förklarande variabler. Precis som vanligt läggs en slumpterm till på slutet för att täcka in variationen i y-variabeln.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

I datamaterialet som använts för den här uppsatsen är den beroende variabeln Y indelad i två klasser. Den första klassen utgörs av länder med en hög genomsnittlig konsumtion av alkohol, vilken tilldelats värdet 1, och den andra klassen utgörs av de länder som har låg genomsnittlig konsumtion av alkohol, och denna klass tilldelas värdet 0.

För Y-variabeln gäller alltså:

$$Y_i \in Bernoulli(P_i), \qquad \begin{cases} \Pr(Y_i = 1) = P_i \\ Pr(Y_i = 0) = (1 - P_i) \end{cases}$$

Vissa iakttagelser angående väntevärdet $E(Y_i)$ kan göras redan vid det här laget.

$$E(Y_i) = 0 * (1 - P_i) + 1 * P_i = P_i$$

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i + 0$$

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i = P_i$$

Sannolikheten, P, är en linjär funktion av den oberoende variabeln, och givetvis kan denna sannolikhet endast anta värden mellan 0 och1. Detta innebär att man behöver en transformation, som dels medför att sannolikheten inte bestäms linjärt, men som också begränsar värdet på y-variabeln, som ju endast ska anta värden 0 och 1. Ett sätt att göra detta nås genom att använda den logistiska funktionen som länkfunktion:

$\frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{-Z_i}}$ , där $Z_i = \beta_1 + \beta_2 X_i = \sum_{\forall_i} \sum_{\forall_k} X_{ik} \beta_k$

Oavsett värde på $\beta_1 + \beta_2 X_i = Z_i$ så kommer denna funktion att anta värden mellan 0 och 1. Modellen ändras alltså till:

$$\frac{1}{1 + e^{-Z_i}} = P_i = E(Y_i|X_i)$$

Uttrycket ovan kan lätt skrivas om så att man får:

$$Z_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i$$

Av detta uttryck att döma kan man se att $\beta$-parametrarna och sannolikheten inte har ett linjärt samband. Normalt skattas parametrarna med en variant av maximum likelihood-metoden. Detta kommer dock inte leda till en exakt algebraisk lösning, vilket illustreras här nedan.

Antag att *N* länder ingår i datamaterialet och sannolikheten att varje enskilt land hamnar i högkonsumtionskategorin följer Bernoulli-fördelningen. Sannolikheten att observera *N* Y-värden ges då av

$$\prod_1^N P_i{}^{Y_i}(1-P_i)^{1-Y_i}$$

Om detta uttryck logaritmeras och skrivs om nås följande uttryck:

$$\sum_1^N [Y_i \ln P_i - Y_i \ln(1-P_i) + \ln(1-P_i)]$$

$$= \sum_1^N \left[Y_i \ln\left(\frac{P_i}{1-P_i}\right)\right] + \sum_1^N \ln(1-P_i)$$

Vissa delar i detta uttryck är redan kända, så det går att förenkla detta något innan några faktiska räkningsförsök utförs. Vi vet till exempel att

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i$$

Vidare kan det också enkelt visas att $(1-P_i) = 1 + e^{Z_i}$. Med dessa resultat kan ovanstående summa skrivas om som

$$\sum_1^N Y_i(\beta_1 + \beta_2 X_i) - \sum_1^N \ln\left[1 + e^{(\beta_1+\beta_2 X_i)}\right]$$

Det är nu det blir en aning problematiskt. Om man sätter alla förstaderivator med hänseende på β-parametrarna till noll kommer man tyvärr inse att det inte finns någon exakt algebraisk lösning. Någon form av numerisk optimeringsmetod måste tillämpas. En vanlig sådan är Newton-Raphsons metod, vilken är en iterativ process som utnyttjar Taylorexpansion kring ett gissat värde. Man gissar alltså ett värde för β-parametern ifråga som minimerar den deriverade funktionen,$f(\beta)$. Rimligtvis kommer man inte träffa exakt rätt, varför Taylorexpansionerna kommer in i bilden. Metoden användes traditionellt för att numeriskt hitta rötter till svårare tal med hjälp av linjära approximationer.

När man får sin första skattning av beta använder man denna för att göra en ny gissning av det sanna $\beta$-värdet, som hamnar ännu närmare. Processen beskriven ovan upprepas tills man kommit så långt att det i stort sett inte sker några förändringar av skattningen längre. Här bör det dock tilläggas att när man gör detta i praktiken vill man ju lösa ett ekvationssystem. Målet är att hitta en kombination av $\beta$-värden som maximerar sannolikheten att få de värden man observerat i datamaterialet. I litteraturen skrivs detta ekvationssystem oftast i matrisform, vilket alltså innebär att man löser ut en vektor av betavärden istället för ett enskilt.

**Källor**:

Hilbe, J.M. (2009). *Logistic regression models*. Boca Raton: Chapman & Hall/CRC.

Gujarati, D.N. & Porter, D.C. (2009). *Basic econometrics*. (5. ed.) Boston: McGrawHill.

# Appendix B Regression med paneldata

Som nämns i avsnitt 3.3 så används fixa effekter för jämförelser mellan länder. Hur modellens parametrar skattas förklaras här.

**Fixa effekter**

Normalt sett brukar man skriva modellen på följande sätt:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}$$

Residualerna $u_{it}$ antas här vara oberoende $\mathrm{N}(0, \sigma_u^2)$. Uttrycket ovan gäller dock bara för en enda individ. Om man vill skriva ett generellt regressionsuttryck för vilket *i* som helst får man införa en dummy-variabel:

$$y_{it} = \sum_{j=1}^{N} \alpha_i d_{ij} + \beta x_{it} + u_{it},$$

$$d_{ij} = 1 \; om \; i = j$$

I praktiken går det att estimera parametrarna med vanlig OLS. Detta är dock inte att rekommendera, eftersom man i praktiken får en modell med fruktanvärt många variabler. Istället kan man räkna ut en enkel linjär regression på skillnaden mellan individers värden och deras medelvärden:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}$$

Och

$$\bar{y}_i = \alpha_i + \beta \bar{x}_i + \bar{u}_i$$

Om räknar ut differensen mellan $y_{it}$ och $\bar{y}_i$ får man följande uttryck:

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$$

Om man nu löser ut en skattning av $\beta$-parametern med hjälp av OLS så får man en skattning som är lika med

$$\hat{\beta}_{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_t)(x_{it} - \bar{x}_t)}$$

Denna skattning är konsistent om det gäller att $E\{(x_{it} - \bar{x}_i)u_{it}\} = 0$. Lägg märke till att intercepten $\alpha_i$ försvinner.

Det är också möjligt att skatta Fixa Effekter-modellen genom att lagga modellen en tidsperiod. Uttrycket ser ut på följande sätt efter att modellen laggats:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}$$

$$y_{i,t-1} = \alpha_i + \beta x_{i,t-1} + u_{i,t-1}$$

$$(y_{it} - y_{i,t-1}) = (\alpha_i - \alpha_i) + \beta(x_{it} - x_{i,t-1}) + (u_{it} - u_{i,t-1})$$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it}$$

Detta får alltså samma konsekvenser vad gäller intercepten. De försvinner eftersom de är konstanta över alla tidsperioder.

**Källa:**

Verbeek, M. (2012). *A guide to modern econometrics*. John Wiley & Sons.

# Appendix C Tabell över alkoholkonsumtion

*Tabell 12. Tabell över registrerad, icke-registrerad sam total alkoholkonsumtion i 100 % ren alkohol per person år 2010*

| Land | Registrerad alkoholkonsumtion 2010 | Icke-registrerad konsumtion | Total konsumtion |
|---|---|---|---|
| Österrike | 12,5 | 0,6 | 13,1 |
| Frankrike | 12,0 | 0,4 | 12,4 |
| Irland | 11,6 | 0,5 | 12,1 |
| Tjeckien | 11,4 | 1,2 | 12,6 |
| Estland | 11,4 | 0,8 | 12,2 |
| Luxemburg | 11,4 | 0,5 | 11,9 |
| Tyskland | 11,2 | 0,5 | 11,7 |
| Ungern | 10,8 | 2,0 | 12,8 |
| Portugal | 10,8 | 1,9 | 12,7 |
| Belgien | 10,6 | 0,5 | 11,1 |
| Australien | 10,3 | 1,8 | 12,1 |
| Danmark | 10,3 | 1,0 | 11,3 |
| Slovenien | 10,3 | 1,0 | 11,3 |
| Storbritannien | 10,3 | 1,2 | 11,5 |
| Slovakien | 10,1 | 1,7 | 11,8 |
| Polen | 10,0 | 1,6 | 11,6 |
| Schweiz | 10,0 | 0,5 | 10,5 |
| Spanien | 9,8 | 1,2 | 11,0 |
| Finland | 9,7 | 2,3 | 12,0 |
| Nya Zeeland | 9,6 | 1,6 | 11,2 |
| Nederländerna | 9,3 | 0,5 | 9,8 |
| Sydkorea | 9,0 | 2,5 | 11,5 |
| USA | 8,6 | 0,5 | 9,1 |
| Kanada | 8,2 | 2,0 | 10,2 |
| Chile | 7,9 | 2,0 | 9,9 |
| Grekland | 7,9 | 2,0 | 9,9 |
| Japan | 7,3 | 0,2 | 7,5 |
| Sverige | 7,3 | 2,0 | 9,3 |
| Norge | 6,6 | 1,0 | 7,6 |
| Island | 6,3 | 0,5 | 6,8 |
| Italien | 6,1 | 0,2 | 6,3 |
| Mexiko | 5,4 | 1,8 | 7,2 |
| Israel | 2,7 | 0,3 | 3,0 |
| Turkiet | 1,5 | 0,6 | 2,1 |