ProjektA_github

Olivia Buhr

2023-09-23

Inledning

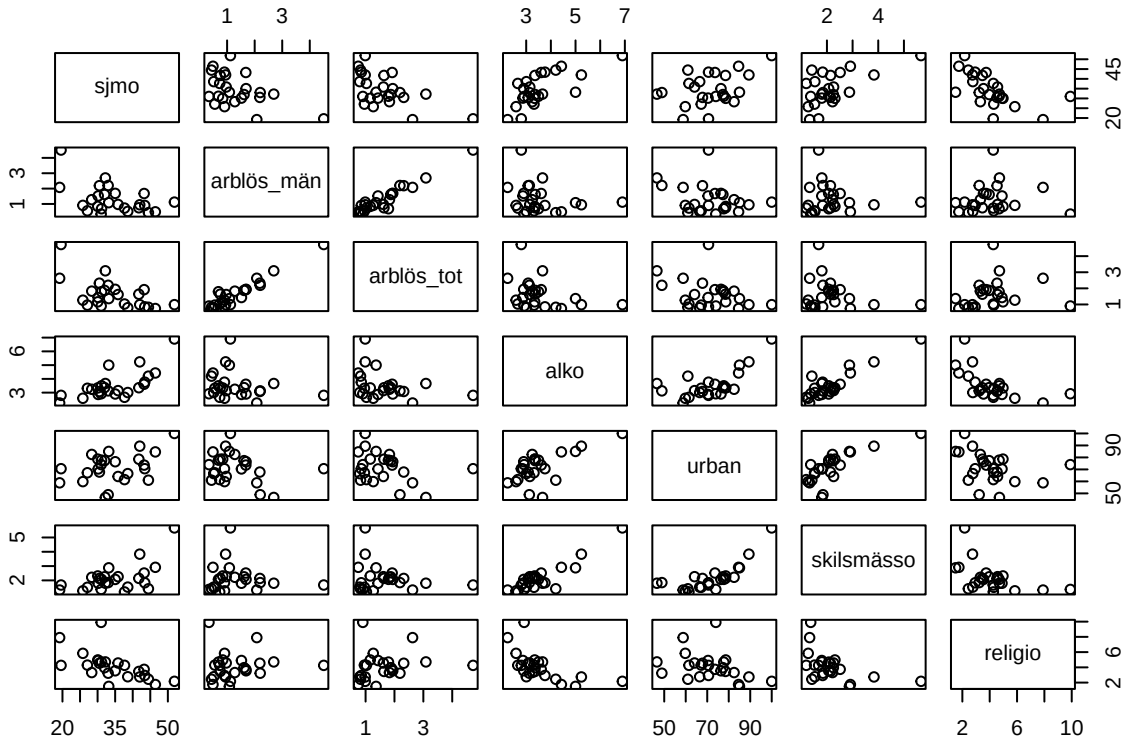
Syftet med detta projekt är att undersöka ifall det finns ett samband mellan alkoholmissbruk och självmordsförekomst. För att göra detta studerar vi två dataset över registerdata från svenska län under 60-talet över tidsperioderna 1963-65 och 1966-1968. Datan har samlats in från 25 län och innehåller 6 variabler som kan tänkas ha en påverkan på självmordsförekomsten i länen. Följande variabler har studerats: (1) Alkoholindex: Årsmedelförbrukningen av alkohol per capita, räknat på befolkning över 15 års ålder. Förbrukningen räknat per capita anses stå i dokumenterat starkt samband med andelen missbrukare, och det är den andelen som man helst skulle velat ha haft som variabel. (2) Skilsmässoindex: procentandel skilsmässor bland män år 1964 resp. 1967. (3) Religiositetsindex: Summan av procentandel statskyrkliga som deltar i gudtjänster och procentandel frikyrkliga. Bara ett värde, dvs gemensamt för båda tidsperioderna. (4) Arbetslöshetsindex totalt: Andelen arbetslösa i medeltal över tidsperioderna. (5) Arbetslöshetsindex män: Motsvarande räknat bara på den manliga delen av befolkningen. (6) Urbaniseringsindex: Procentandel av befolkningen bosatt i tätbefolkat område

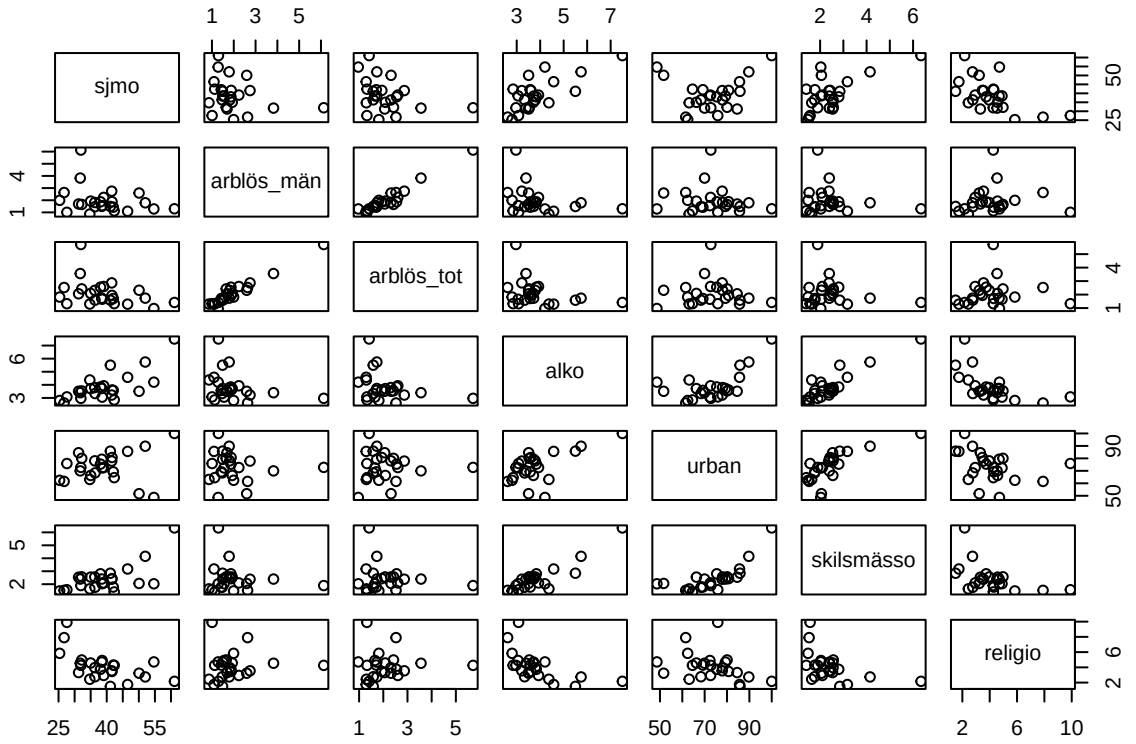
Syfte

För att genomföra vår undersökning kommer vi använda oss av multipel linjär regression. Självmordsfrekvens är vår responsvariabel och övriga förklarande variabler. Då vi är intresserade av att förklara just självmordsfrekvens är det lämpligt att ha just denna variabel som responsvariabel. Det bör dock tilläggas att detta inte implicerar något kausalt samband. Huruvida det är den förklarande variabeln som påverkar självmordsfrekvensen, tvärt om eller en helt annan variabel vet vi inte. Den här undersökningen syftar endast till att förklara förändringen mellan länen. Vi kommer analysera de två perioderna separat då det kan finnas viktiga skillnader i datan. Då vi endast har mätdata från två tidpunkter är det svårt att uttala sig om eventuell förändring över tid. Däremot kan vi upptäcka eventuella skillnader mellan de två tidsperioderna och analysera dessa. För att uppfylla undersökningens syfte kommer vi lägga stor vikt vid valet av multipel linjär regressionsmodell. För att ta fram en så lämplig modell som möjligt kommer vi studera och analysera datan utifrån flera olika aspekter. Vi undersöker bland annat multikolinjäritet för olika kombinationer av variabler, R^2-värden hos olika modeller och p-värden för olika variabler i olika modeller. Då vi är särskilt intresserade av modeller som innehåller alkoholindex som förklarande variabel, men vi undersöker även om det finns andra modeller som kan förklara variationen bättre.

Det vi kom fram till....

Det första steget i undersökningen är att observera den insamlade datan åtskilda för de olika perioderna med alla variabler inkluderade. Vi gör detta med hjälp av parvisa plottar där vi kan studera korrelation mellan alla olika variabler, både respons och förklarande.


sjmo
arblös_män
arblös_tot
alko
urban
skilsmässo
religio



Figur 1 visar parvisa plottar med alla förklarande variabler för dataset 1

Figur 2 visar parvisa plottar med alla förklarande variabler för dataset 1

I Figur 1 kan vi se att den totala arbetslösheten har stark positiv korrelation med arbetslöshet för män. Korrelationen mellan alkoholindex och skilsmässoindex samt urbaniseringsindex och skilmässoindex är också positiv. Det finns även en måttlig negativ korrelation mellan alkoholindex och religiositetsindex samt skilsmässoindex och religiositetsindex. Självmordsfrekvens verkar vara negativt korrelerad med religiositetindex samt positivt korrelerad med skiljsmässoindex och alkoholindex. Eventuell korrelation mellan självmordsfrekvens och resterande variabler är svårare att utröna.

I Figur 2 kan vi se, precis som i figur 1, att den totala arbetslösheten har en stark positiv korrelation med arbetslöshet för män. Korrelationen mellan alkoholindex och skilsmässoindex samt urbaniseringsindex och skilmässoindex är också positiv. Det finns även en måttlig negativ korrelation mellan alkoholindex och religiositetsindex samt skilsmässoindex och religiositetsindex. Till skillnad från period 1 så är självmordsfrekvens och religiositetindex inte lika tydligt negativt korrelerade för period 2. Skiljsmässoindex och alkoholindex verkar dock fortfarande ha en positiv korrelation med självmordfrekvens. Eventuell korrelation mellan självmordsfrekvens och resterande variabler är svårare att utröna, möjligtvis att det kan finnas en negativ korrelation med arbetslöshet total. Vi kan eventuellt se en antydan till ett kvadratiskt samband mellan urbaniseringsindex och självmordsfrekvens.

Den starka korrelationen mellan arbetslöshet total och arbetslöshet män var väntad då antalet arbetslösa män är en delmängd av antalet arbetslösa totalt. Detta innebär att båda dessa variabler förklarar i princip samma sak, vi kommer därför med största sannolikhet inte inkludera båda i vår slutgiltiga modell. Att inkludera starkt korrelerade förklarande variabler i en multipel regressionsmodell kan vara problematiskt då detta kan indikera kolinjäritet. Det räcker dock inte att enbart undersöka plottar, vi behöver även någon sorts mått för att avgöra graden av kolinjäritet mellan variablerna. VIF (variance inflation factor) är ett mått på (multi)kolinjäritet, d.v.s. kolinjäritet mellan två eller flera variabler. Ett VIF-värde nära noll tyder

på en låg kolinjäritet, medan ett högt tyder på en hög kolinjäritet. Ett VIF-värde över fem anses ofta vara för högt, denna gräns kan dock variera beroende på t.ex. forskningsområde eller studie. Problemet med ett allt för högt VIF-värde är att skattningar av parametrar och p-värden kan bli opålitliga. Vi kommer därför studera VIF-värden noga när vi väljer vår modell.

Variabelselektion

Valet av förklarande variabler är helt centralt när vi väljer vilken modell som bäst kan besvara vår frågeställning. Det finns flera metoder för variabelselektion, vi kommer använda oss av forward selection, backward elimination och stepwise regression. Detta är stegvisa metoder som bygger på att välja ut de förklarande variabler som ger högst p-värde i modellen. Det finns dock flera aspekter vi behöver ta hänsyn till, en av dem är förklaringsgraden, R^2. Notera att vi väljer R^2-adjusted istället för R^2 för att undvika överanpassning hos modellen. Då vi endast har sex potentiella förklarande variabler har vi möjligheten att undersöka alla möjliga kombinationer av variabler (63 stycken). Vi kommer därför beräkna R^2-adjusted för samtliga kombinationer av de förklarande variablerna och sedan välja ut de modeller med högst värde. Med dessa metoder kommer vi få flera potentiella kandidater till vår modell. Därefter kommer vi granska respektive kandidat och göra en sammanvägning av R^2-adjusted, p- och VIF-värden för att hitta den bästa modellen. Det bör tilläggas att vi inte kommer använda oss av något mått på prediktionsförmåga hos modellen, detta eftersom syftet är att förklara snarare än att prediktera.

Models

Model 1

Variable selection summary `data_1`, (alla p-gränser = 0.1)

Forward: `alko`, `arbslös_tot`, `religio` (VIF 1.68, 1.13, 1.54)

Backward: `alko`, `arblös_män` `urban` (VIF 1.76, 1.09, 1.81)

Both: `alko`, `arbslös_tot`, `religio` (VIF 1.68, 1.13, 1.54)

Bästa R^2_{ajd} för alla modeller: `alko`, `arblös_tot`, `religio` (VIF 1.68, 1.13, 1.54)
och `arblös_tot`, `urban`, `skilsmässo`, `religio`. (VIF 1.14, 2.57, 2.77, 1.33)

Model 2

3.3 Cooks avstånd

Från plotten ovan ser vi att Jönköpings län som en väldigt inflytelserik observation. Jönköping har väldigt högt religiositetsindex under den tidsperiod 1 (1963-1965), samtidigt som självmordindexet inte verkar påverkas särskilt mycket av detta. I övriga län är religion och självmord negativt korrelerade.