ProjektA_github

Olivia Buhr

2023-09-23

##Övningsprojekt a 18 Samband mellan alkolhol och självmord?

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

För att undersöka hypotesen: alkohol har en påverkan på självmordsförekomst så utför vi multipel linjär regression med självmordsförekomst som responsvariabel och de resterande variablerna som förklarande variabler. Detta görs separat för de två dataseten. Vi undersöker bland annat hur stor multikolinjäritet alkohol har med de andra variablerna och vilken kombination av variabler som tycks vara bäst anpassad, genom att kolla på R^2-värdet och p-värdet för modellen, för att se hur stor inverkan alkoholkonsumtion verkar ha på självmordsförekomsten. Genom att studera dataseten för sig kan vi även se ifall det verkar ha skett någon förändring mellan de två tidsperioderna.

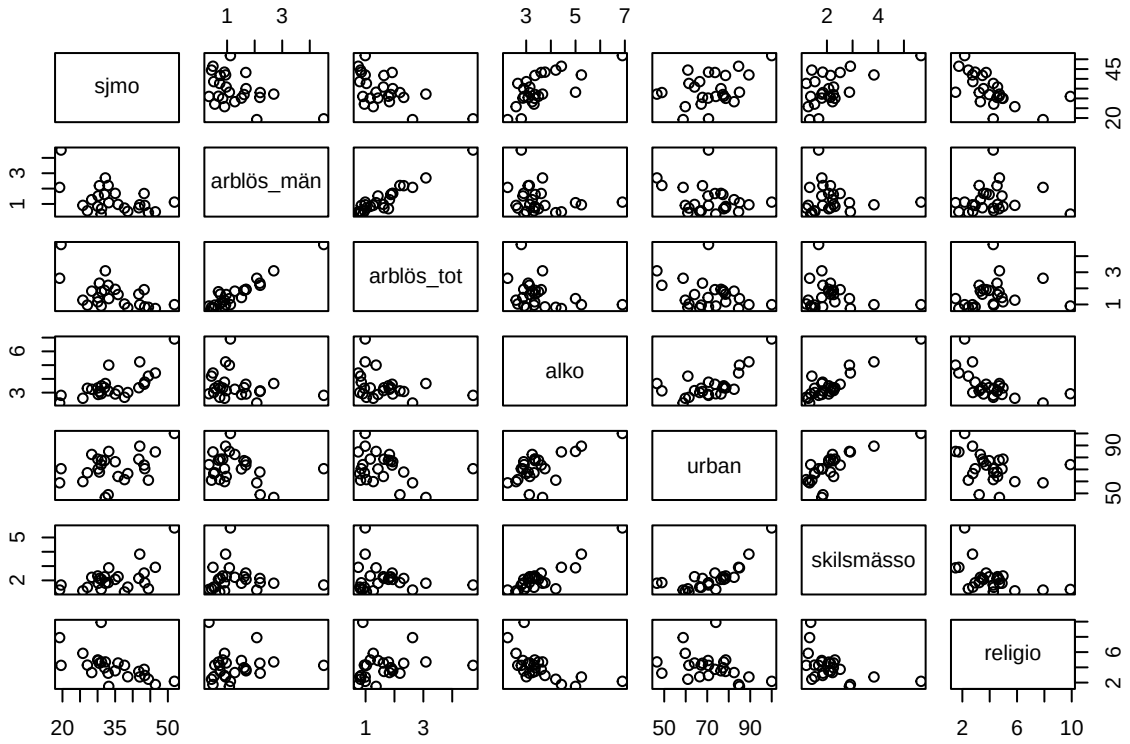
1 Första observation (summary multipel linjär regression) (Sofia)

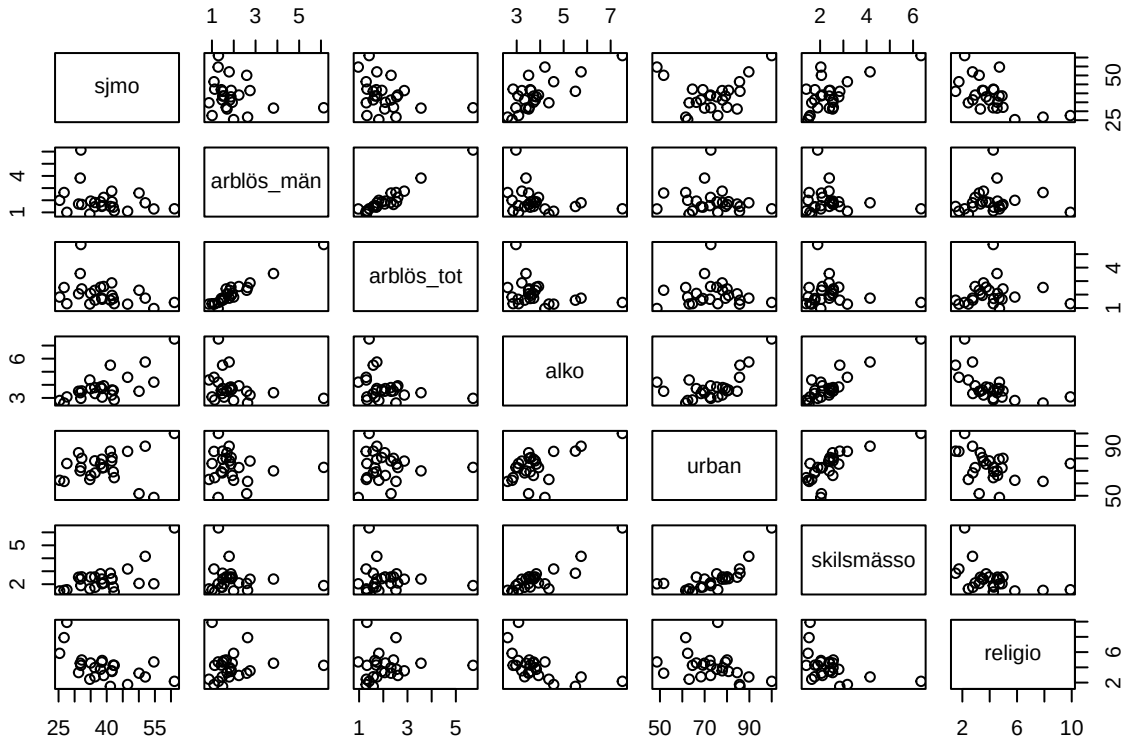
Det första vi gör för att reda ut vilka variabler som har en påverkan på självmordsfrekvens är att observera den insamlade datan åtskilda för de olika perioderna med alla variabler inkluderade. Från multipel regression med självmordsindex som responsvariabel och resterande variabler som förklarande, för tidsperiod 1 och 2

Från summeringen av modell 1 verkar religiositetsindex vara den variabel med lägst p-värde på 0.065. För modell 2 har vi 3 förklarande variabler med lågt p-värde, vilket är religiositetsindex, skilsmässoindex samt urbaniseringsindex med p-värdena 0.041, 0.037 samt 0.060. Modell har även en lägre förklaringsgrad på 0.7, jämfört med förklaringsgraden 0.74 för modell 2.

För båda modellerna har variablerna alkoholindex och skilsmässoindex positiv korrelation med responsvariabeln, medan religiositetsindex, urbaniseringsindex samt total arbetslöshetsindex negativ korrelation.

Någonting som bör noteras är att arbetslöshetsindex för män och totalt har en uppenbar korrelation då antal arbetslösa män har en direkt påverkan på arbetslösheten totalt. Detta kan resultera i missvisande resultat i de linjära regressionerna. För att undersöka detta samband samt andra potentiella fall av kollinearitet mellan de förklarande variablerna undersöker vi alla variabler parvis.


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

[TEXT OM PAIRVISE PLOTS]

Vi ska nu undersöka vilken kombination av förklarande variabler som är bäst för att förklara självmordsförekomsten genom att använda stegvis variabelselektion för att hitta modeller med lägst p-värde. Vi vill även hitta en modell med högt adjusted R^2 och låg korrelation mellan variabler. Vi använde forward selection, backward elimination och en variant där båda används. För att hitta de modeller med högst adjusted R^2 så körde vi alla 2^6 - 1 modeller och plottade respektives adjusted R^2-värde i en scatter plot. Sedan valde vi ut den modellen som låg som maxpunkt på denna kurva. I vissa fall var det flera olika kandidater till bäst adjusted R^2. I dessa fall valde vi att studera alla de olika fallen.

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.