**Opgave om sociale forhold i Bulgarien og Tyskland**

**- til auditoriet tirsdag, onsdag og torsdag i uge 1**

**Datasættet Bulgarien\_Tyskland**

I denne opgave betragtes sociale forhold i større byer i hhv Bulgarien og Tyskland. Data stammer fra Eurostats datasæt ”Urban Audit”. Der er kun oplysninger om disse sociale forhold for ganske få lande, så derfor er valget faldet på netop Bulgarien og Tyskland.

De anvendte responsvariable er:

**stor\_lowwork:** Variablen er dannet ud fra en variabel om "Share of persons living in households with very low work intensity", så værdien 'Ja' er over 59% og 'Nej' betyder under 59%.

**EC2034V** og **lEC2034V:** Employment (jobs) in financial and insurance activities. Forbogstavet "l" angiver, at den er logaritmisk transformeret

**EC3065V:** Share of persons at risk of poverty after social transfers

**social\_transfers:** Variablen **EC3065V** afrundet til hele multipla af 10

**stor\_risk:** Dannet ud fra variablen **EC3065V** så værdien 'Ja' er over 24 og 'Nej' betyder under 24.

**SA1051V:** Average price for buying an apartment - EUR

Som forklarende variable anvendes følgende størrelser:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **DE2009V** | | **DE3002V** | | **DE3008V** | | **SA1008V** | | **SA1012V** | | **TE1026V** | | |  | | --- | | **Foreign-born** | | **One person households** | | **Lone pensioner (above retirement age) households** | | **Number of households living in apartments** | | **Households in social housing** | | **Students in higher education (ISCED level 5-6), total** | |

For disse kontinuert varierende variable, der alle angiver antal, er den logaritmisk transformerede også medtaget i datasættet, idet der blot er sat et ”l” foran variabelnavnet, fx **lDE2009V** **=log(DE2009V)**.

De forklarende variable kan copy-pastes fra linien

**DE2009V DE3002V DE3008V SA1008V SA1012V TE1026V**

resp

**lDE2009V lDE3002V lDE3008V lSA1008V lSA1012V lTE1026V**

Desuden er der dummyvariable for

**Land** med værdierne ’BG’ og ’DE’

**Tyskland** med værdien 0 for Bulgarien og 1 for Tyskland.

Variablen **by** er en kode for byen, mens variablen **bynavn** er byens navn skrevet fuldt ud.

# Tirsdag

1. Den kategoriserede variabel **stor\_lowwork** var tænkt konstrueret, så ca en fjerdedel af observationerne havde værdien 'Ja' og resten værdien 'Nej'. Test ved brug af SAS, om det er lykkedes.

1. Variablen **social\_transfers** er egentlig kontinuert, men da den er afrundet, antager den kun 5 værdier. Test om variablen er ligefordelt over disse 5 værdier.

1. Test om der uafhængighed mellem **stor\_risk** og **stor\_lowwork**. Test ligeledes om der uafhængighed mellem **stor\_lowwork** og **land.**
2. Test om middelværdien af variablen **SA1051V** er lig med 100000. Test ligeledes om middelværdien af den logaritmisk transformerede **lSA1051V** er lig med log(100000)=11.46.
3. Test om middelværdien af variablen **SA1051V** er ens for de to lande. Test ligeledes om middelværdien af den logaritmisk transformerede **lSA1051V** er ens for de to lande
4. Udfør regressionsanalysen med **EC2034V** som responsvariabel og **DE3008V** som forklarende variabel. Vurder ud fra et plot af de to variable mod hinanden om regressionen ser fornuftig ud. Gentag med de logaritmisk transformerede variable **lEC2034V** og **lDE3008V.** Blev det bedre? Test om hældningskoefficienten i denne model er lig med 1.

1. Udvid regressionen for de logaritmisk transformerede variable i f) med dummyvariablen **Tyskland**, der antager værdierne 0 eller 1. Sammenhold med brug af variablen **land** i et kald af PROC GLM.

1. Afprøv ved variabeludvælgelse og signifikanstest, hvilke af de logaritmisk transformerede forklarende variable side 1, der skal med som forklarende variable i regressionsanalysen med **lEC2034V** som responsvariabel. Variablen Tyskland, skal nok også anvendes.

Betragt nu modellen med **lEC2034V** som responsvariabel og de forklarende variable som fundet i h) som forklarende variable. I det følgende skal I ikke tænke over, om modellen er tilfredsstillende, dvs I skal af tidsmæssige grunde springe enhver form for modelkontrol over.

1. Angiv grænserne for, hvornår et eksternt standardiseret residual angiver en outlier i modellen. Test om det numerisk største eksternt standardiserede residual svarer til en outlier. Giver det mening, at netop denne by kunne være en outlier?

1. Vurder om de enkelte byer har stor indflydelse på MK analysens konklusioner og beskriv i givet fald deres indflydelse på estimaterne for regressionskoefficienterne og om der er gode respektive dårlige sæt i materialet. Giver disse resultater geografisk mening?

1. Konstruer de tilføjede variabel diagrammer (partial regression plots) både for den reducerede model i h) og for startmodellen med alle variable i a). Tyder diagrammerne på, at signifikansen og/eller fortegnene på de estimerede regressionskoefficienter skyldes en eller flere indflydelsesrige, evt. atypiske, observationer?

# Onsdag

1. Betragt modellen, hvor den logaritmisk transformerede variabel **lEC2034V** forklares med en enkelt forklarende variabel **lDE3008V,** tirsdag spm f). Brug kun de tyske byer, dvs brugwhere land='DE'; Gentag analysen med M-estimation med Bisquare vægt i et kald af Proc Robustreg med default optioner. Prøv også MAR, hvor der minimeres summen af den numeriske værdi af residualerne. Prøv til sidst også LTS.
2. Betragt modellen med den mange forklarende variable fra tirsdag spm h). Gentag analysen med M-estimation med Bisquare vægt i et kald af Proc Robustreg med default optioner.

1. Gentag analysen med LTS estimation. Udfør OLS estimation på de gode observationer, der identificeres ved LTS estimation. Overvej om både alle outliere og alle indflydelsesrige observationer fra LTS estimationen bør fjernes.

1. Nu kan der vel efterhånden opstilles en passende model, hvor eventuelle insignifikante forklarende variable kan udelades ved tests. Gør det!

**Torsdag**

1. Betragt nu variablen **stor\_risk** som responsvariabel og anvend de logaritmisk transformerede variable samt dummyvariablen Tyskland som forklarende variable i en logistisk regression. Fjern insignifikante forklarende variable. Er modellen tilfredsstillende? Er der byer med stor indflydelse på analysens resultater. Tegn meningsfulde effectplots!

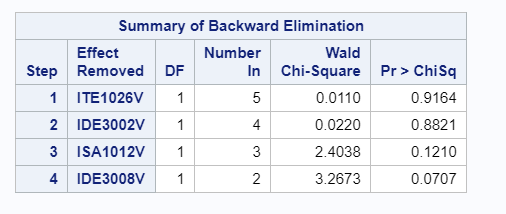
libname ssvs '/courses/d284cd65ba27fe300/Sommerskole 2018/Data';

\*Opgave A torsdag;

proc logistic data=ssvs.bulgarien\_tyskland plots=all;

model stor\_risk=lDE2009V lDE3002V lDE3008V lSA1008V lSA1012V lTE1026V/selection=backward;

run;



Resten af variablene er signifikante

kig på ROC-kurven under all plots.

1. Betragt variablen **social\_transfers** som responsvariabel; denne variable har jf spm b) tirsdag 5 niveauer. Anvend forklarende variable som i a). Fjern insignifikante forklarende variable. Er modellen tilfredsstillende? Er modellen med de logaritmisk transformerede forklarende variable bedre? Er parallelitetshypotesen opfyldt? Tegn meningsfulde effectplots!
2. Anvend **SA1018V** som responsvariabel i en Poissonregression med forklarende variable som i a), men da responsvariablen angiver antal skal de forklarende variable nok ikke logaritmetransformeres. Der kun er observationer for Bulgarien, så dummyvariablen **Tyskland** skal naturligvis ikke anvendes.