Navn/Studie nr.: Jesper Højberg Knudsen – FMW786 og Jeppe Vanderhaegen
Fag: Videregående Statistik - Sommerskole
Dato: 06-07-2022

KØBENHAVNS UNIVERSITET

# Sommerskoleaflevering 1 – Onsdag

## Opgave a)

For at besvare spørgsmålet, køres følgende kode:

```
ods graphics/ imagemap=on  TIPMAX=2200;

libname Data 'C:\Users\Bruger\OneDrive\Skrivebord\Videregaende Statistik
Sommerskole\Data';
run;

**Opgave a**;
proc contents data=data.usa_2022;
run;
**Hele regressionen køres**;
proc reg data=data.usa_2022;
model Violent_crime_rate=share_Black_victim share_White_victim rate_black
income_2018 rate_income_growth unemployment_rate rate_public_health
rate_non_insured_health rate_private_health percent_biden trump_2016
Incarceration_rate rate_killed;
id state;
run;
**Herefter køres regressionen med stepwise selection**;
proc reg data=data.usa_2022;
model Violent_crime_rate=share_Black_victim share_White_victim rate_black
income_2018 rate_income_growth unemployment_rate rate_public_health
rate_non_insured_health rate_private_health percent_biden trump_2016
Incarceration_rate rate_killed/selection=stepwise slstay=0.1 slentry=0.1;
id state;
run;

**Herefter køres regressionen med backwards selection**;
proc reg data=data.usa_2022;
model Violent_crime_rate=share_Black_victim share_White_victim rate_black
income_2018 rate_income_growth unemployment_rate rate_public_health
rate_non_insured_health rate_private_health percent_biden trump_2016
Incarceration_rate rate_killed/selection=b;
id state;
run;

**Herefter køres regressionen med forwards selection**;
proc reg data=data.usa_2022;
model Violent_crime_rate=share_Black_victim share_White_victim rate_black
income_2018 rate_income_growth unemployment_rate rate_public_health
rate_non_insured_health rate_private_health percent_biden trump_2016
Incarceration_rate rate_killed/selection=f;
id state;
run;

**Vi kører nu de udvalgte forklarende varable fra stepwise selectionen**;
proc reg data=data.usa_2022;
model Violent_crime_rate=share_Black_victim rate_killed trump_2016;
id state;
run;
```

Først estimeres modellen for `Violent_crime_rate` som afhængig variabel og med alle variablene beskrevet i opgaven, som forklarende variable.

Dette giver følgende estimater:

| Parameter Estimates | | | | |
|---|---|---|---|---|
| **Variable** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 433.45948 | 1671.22444 | 0.26 | 0.7968 |
| **share_Black_victim** | 288.62980 | 225.85632 | 1.28 | 0.2092 |
| **share_White_victim** | -102.16381 | 148.64642 | -0.69 | 0.4962 |
| **rate_black** | 31.59809 | 395.69657 | 0.08 | 0.9368 |
| **income_2018** | -0.00060009 | 0.00376 | -0.16 | 0.8742 |
| **rate_income_growth** | -259.09761 | 506.19823 | -0.51 | 0.6118 |
| **unemployment_rate** | 19.45967 | 29.96532 | 0.65 | 0.5201 |
| **rate_public_health** | -68.87879 | 1419.19425 | -0.05 | 0.9616 |
| **rate_non_insured_health** | -644.96946 | 2006.53291 | -0.32 | 0.7497 |
| **rate_private_health** | -12.03432 | 1449.52058 | -0.01 | 0.9934 |
| **percent_biden** | -1.80924 | 8.21733 | -0.22 | 0.8269 |
| **trump_2016** | -432.64029 | 862.23268 | -0.50 | 0.6188 |
| **Incarceration_rate** | 0.08905 | 0.18635 | 0.48 | 0.6355 |
| **rate_killed** | 7.85946 | 2.26607 | 3.47 | 0.0013 |

Det ses herved, at kun Rate_killed er signifikant på et 5%-signifikansniveau, mens Share_Black_victim er den variabel, der er næstmest signifikant, men med en p-værdi på 0,209.

Derudover ses det, at modellen har en forklaringsgrad på: R-Square = 0.7009

Herefter er der kørt stepwise, backwards og forwards selektion, hvorved vi vælger modellen med stepwise og en 10%-signifikansgrænse.

Denne finder følgende 3 signifikante variable ”rate_killed, share_Black_victim, trump_2016” og giver estimationen:

| Parameter Estimates | | | | |
|---|---|---|---|---|
| **Variable** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 102.36265 | 79.28239 | 1.29 | 0.2030 |
| **share_Black_victim** | 435.54100 | 78.13681 | 5.57 | <.0001 |
| **rate_killed** | 9.27022 | 1.02663 | 9.03 | <.0001 |
| **trump_2016** | -247.91068 | 131.37083 | -1.89 | 0.0653 |

Det ses herved, at forklaringsgraden stort set ikke er faldet: R-Square = 0.6815. Vi vurderer derved, at denne markant reducerede model er signifikant.

Navn/Studie nr.: Jesper Højberg Knudsen – FMW786 og Jeppe Vanderhaegen
Fag: Videregående Statistik - Sommerskole
Dato: 06-07-2022



## Opgave b)

Følgende kode køres til at besvare opgaven:

```
**Opgave b**
**Oprettelse af samlet variabel**;
%let fvar= share_Black_victim rate_killed trump_2016;

**Outliers**;
proc reg data=data.usa_2022;
model Violent_crime_rate = &fvar.;
output out=ny Rstudent=t covratio=c h=h cookd=d;
id state;
run;

proc sort data=ny out=nysort;
by t;
run;

proc print data=nysort;
var state t c h d;
where abs (t)>2;
run;
**Herefter køres en outliertest**;
*I dette program er benyttet n=51, og abs(t)=2.62093
og 4 estimerede regressionskoefficienter inklusive interceptet;

*P3 er den rigtige signifikanssandsynlighed for sandsynligheden for outlieren*;
data a;
p1=probt( 2.62093,51-1-4);
/*p1 Tager ikke hensyn til, at tallet er valgt som det største af 51 tal*/
p2=2*(1-p1);
p3=1-(1-p2)**51;
/*p3 tager hensyn til tallet er valgt som det største af 51 tal*/
tgraense=tinv( 1- ( 1- ( 1 - 0.05)**(1/51) )/2, 51-1-4);
/*tgraense er en kritisk grænse for outliertests for 51 uafhængige tal*/
tbonf=tinv( 1- 0.05/2/51, 51-1-4);
/*tbonf er en kritisk grænse for outliertests ud fra Bonferroni uligheden for 51
tal*/
run;
proc print;
run;
```

Det et eksternt standardiseret residual, er observationen i's afstand fra den prædikterede værdi, når observation i ikke er med i regressionen. Vi angiver det som en outlier, når sandsynligheden, for at det eksternt standardiseret residual er mindre end 5%.

Vi finder følgende 4 eksternt standardiseret residualer med store t-værdier:

| Obs | state | t | c | h | d |
|---|---|---|---|---|---|
| **1** | Mississippi | -2.51880 | 0.70271 | 0.07502 | 0.11550 |
| **2** | Oklahoma | -2.11494 | 0.88463 | 0.15005 | 0.18383 |
| **50** | Tennessee | 2.37093 | 0.71977 | 0.04526 | 0.06066 |
| **51** | Alaska | 2.62093 | 0.72889 | 0.14313 | 0.25501 |

Vi ser herved, at det numerisk største eksternt standardiserede residual i den reducerede modellen er Alaska med t = 2.62093.

Vi tester sandsynligheden for, at observere en sådan t-værdi:

| Obs | p1 | p2 | p3 | tgraense | tbonf |
|---|---|---|---|---|---|
| 1 | 0.99408 | 0.011845 | 0.45540 | 3.51319 | 3.52164 |

Her beskriver p3 den signifikanssandsynlighed for sandsynligheden for at observere outlieren givet de 50 andre observationer.

Her får vi en sandsynlighed på 45,5% og vi kan dermed ikke ikke afvise, at Alaska er en outlier.

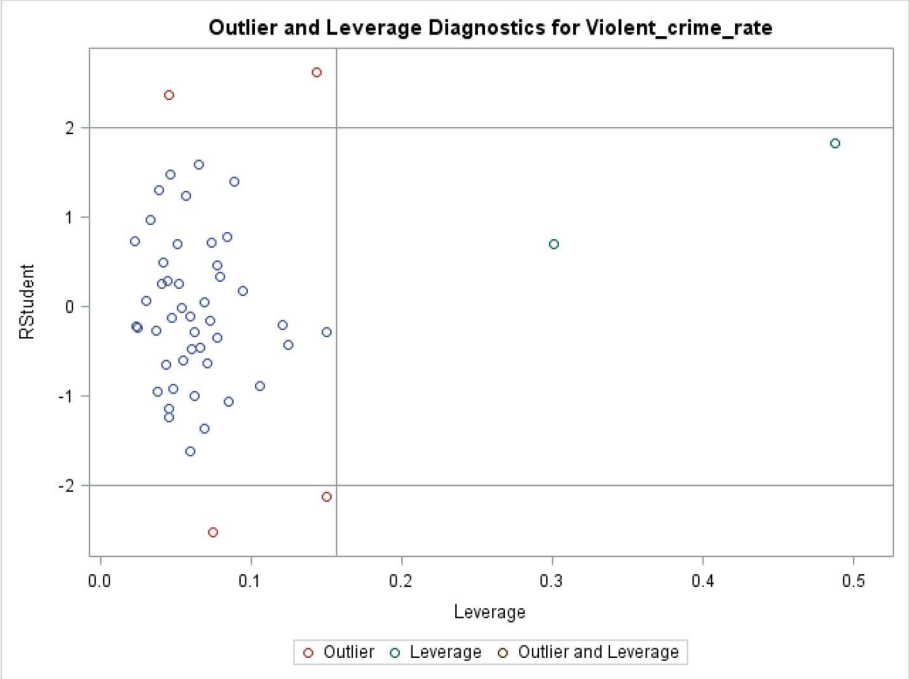
| Obs | state | Violent_crime_rate | share_Black_victim | rate_killed | trump_2016 |
|---|---|---|---|---|---|
| 1 | Mississippi | 278 | 0.45217 | 38.8346 | 0.5794 |
| 2 | Oklahoma | 432 | 0.23200 | 63.1416 | 0.6532 |
| 3 | Maine | 115 | 0.02500 | 29.3608 | 0.4487 |
| 4 | New Jersey | 207 | 0.52212 | 12.1649 | 0.4100 |
| 50 | Tennessee | 595 | 0.27982 | 31.5446 | 0.6072 |
| 51 | Alaska | 867 | 0.08163 | 66.8129 | 0.5128 |

Vi kan herved se, at Alaska har en utroligt høj Violent_crime_rate og rate_killed, men at deres share_Black_victim er meget lavere end alle andre stater. Dette kan skyldes, at næsten ingen sorte bor i Alaska. Da vores estimat på share_Black_victim = 435.5. Derved ville modellen prædiktere en lavere Violent_crime_rate for Alaska pga. den lave share_Black_victim. Dette forklarer hvorfor Alaska er en outlier.

## Opgave c)

```
**Opgave 3**;
proc reg data=data.usa_2022 plots=all;
model Violent_crime_rate = &fvar./partial influence ;
output out=ny Rstudent=t covratio=c h=h cookd=d;
id state;
run;

proc sort data=ny out=nysortt;
by t;
run;

proc print data=nysortt;
var state t c h d;
where abs (t)>2;
run;

proc sort data=ny out=nysortc;
by c;
run;

proc print data=nysortc;
var state t c h d;
where abs (c)>1.1;
run;

proc sort data=ny out=nysortd;
by d;
run;

proc print data=nysortd;
```

```
var state t c h d;
where abs (d)>0.05;
run;

proc sort data=ny out=nysorth;
by h;
run;

proc print data=nysorth;
var state t c h d;
where abs (h)>0.1;
run;
```




Vi kan udlede fra ovenstående diagram, at Discrit of Columbia og New Mexico har størst indflydelse/leverage på estimatorne. Disse er dog ikke outliers. Vores tidligere fundne outliers har ikke nogen stor betydning på estimatorne. Dette kan udledes, fordi de ligger i boksene over og under den samlede gruppering. Samtidig ligger DC og New Mexico langt ude ift leverage, hvor DC også ligger længere oppe mod outlier-delen uden at være dette. Det kan konkluderes, at New Mexico og DC er nogle gode punkter, mens Mississippi, Alaska, Oklahoma og Tennessee dårligere punkter, der dog pga. deres lave leverage ikke påvirker estimatorne betydeligt.

For Alaska kan det udledes, at set bort fra h-værdien, så har Alaska karakteristika af et dårligt punkt. Da den har en stor t-værdi, cov-ratio under 1 og stor cook-værdi., dog med lav leverage. Dette kan modsvares med New Mexico, der har lav t, stor h, c større end 1 og lav d-værdi:

| Obs | state | t | c | h | d |
|---|---|---|---|---|---|
| **47** | Alaska | 2.62093 | 0.72889 | 0.14313 | 0.25501 |
| **50** | New Mexico | 0.69740 | 1.49424 | 0.30069 | 0.05286 |

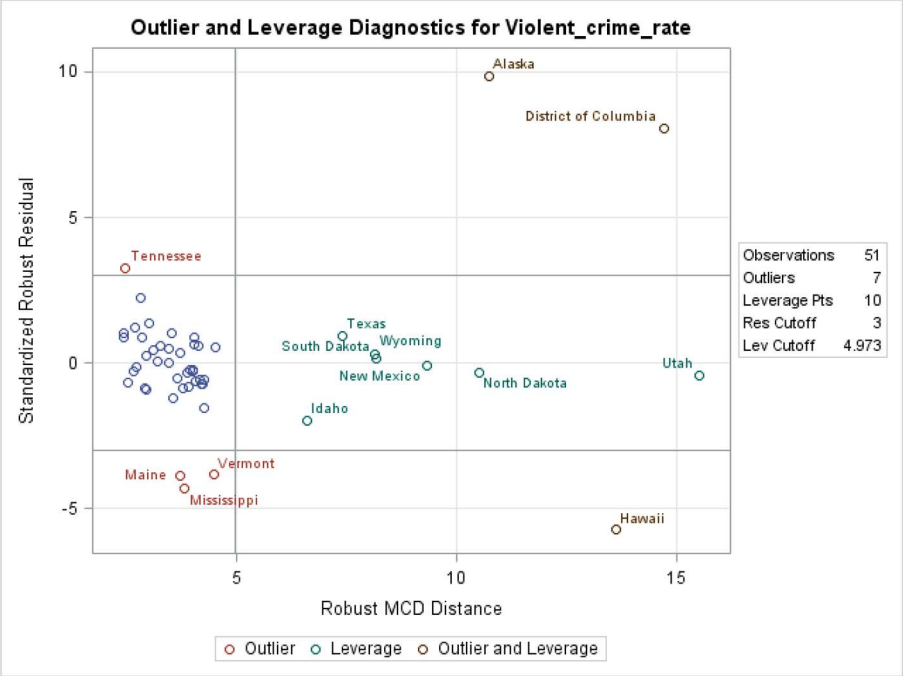
KØBENHAVNS UNIVERSITET

## Opgave d)

Vi kører følgende mode:

```
**Opgave d**;
%let var2 = share_Black_victim share_White_victim rate_black income_2018
rate_income_growth unemployment_rate rate_public_health rate_non_insured_health
rate_private_health percent_biden trump_2016 Incarceration_rate rate_killed;

proc robustreg data=data.usa_2022 method=lts plots=all;
model Violent_crime_rate = &var2./diagnostics(all) leverage;
output out=Robustny outlier=yafvig leverage=xafvig;
id state;
run;
```

Dette giver følgende outlier leverage plot:




Ud af y-aksen ligger leverage. Dette beskriver punktets afstand fra resten af punkterne, h. Stor leverage betyder større variation i observationerne, hvilket sænker variansen.

x-aksen er residualerne størrelse, målt mod de modellens prædikterede værdier når den specifikke observation ikke er med i modellen, t.

Det bedste punkt er Utah, da der er stor leverage, men ikke stor residual. Derimod har Hawaii, DC og Alaska både stor leverage og store residualer og kan dermed trække estimaterne i forkert retning.

KØBENHAVNS UNIVERSITET

## Opgave e)

Kode:

```
**Opgave e**;
proc reg data=robustny plots=all;
model Violent_crime_rate = &var2.;
where yafvig=0;
id state;
run;

**Herefter køres regressionen med stepwise selection**;
proc reg data=robustny plots=all;
model Violent_crime_rate = &var2./selection=stepwise slstay=0.1 slentry=0.1;
where yafvig=0;
id state;
run;

**Dette giver følgende OLS regression**;
proc reg data=robustny plots=all;
model Violent_crime_rate = share_White_victim income_2018 unemployment_rate
percent_biden rate_killed;
where yafvig=0;
output out=robustny1 Rstudent=t covratio=c h=h cookd=d
id state;
run;

proc sort data=robustny1 out=nysortd1;
by d;
run;

proc print data=nysortd1;
var state t c h d;
where abs (d)>0.05;
run;
```

Vi har fjernet de 7 outliers, der blev fundet i opgave d. Derefter er regressionen kørt igen, hvorved følgende estimater er fundet:

| Variable | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| **Intercept** | 852.79522 | 1025.85185 | 0.83 | 0.4124 |
| **share_Black_victim** | -96.81553 | 136.03803 | -0.71 | 0.4822 |
| **share_White_victim** | -395.41216 | 118.11089 | -3.35 | 0.0022 |
| **rate_black** | 541.62671 | 247.25473 | 2.19 | 0.0364 |
| **income_2018** | -0.00866 | 0.00244 | -3.55 | 0.0013 |
| **rate_income_growth** | -205.10559 | 315.01964 | -0.65 | 0.5199 |
| **unemployment_rate** | -51.97181 | 19.88090 | -2.61 | 0.0139 |
| **rate_public_health** | 842.68964 | 962.31220 | 0.88 | 0.3882 |
| **rate_non_insured_health** | 455.87348 | 1280.33701 | 0.36 | 0.7243 |
| **rate_private_health** | 539.65311 | 913.50118 | 0.59 | 0.5591 |
| **percent_biden** | -0.70766 | 4.63939 | -0.15 | 0.8798 |
| **trump_2016** | -533.65154 | 499.96041 | -1.07 | 0.2943 |
| **Incarceration_rate** | -0.13048 | 0.11503 | -1.13 | 0.2657 |
| **rate_killed** | 5.29711 | 1.28547 | 4.12 | 0. |

Det er stadig kun rate_killed, der er signifikant, men forklaringsgraden er steget fra: R-Square = 0.7009 til R-Square = 0.8493. Den er hermed steget med 0,149.

Vi kører herefter stepwise selektion med og en 10%-signifikansgrænse. Dette giver følgende estimater:

| Parameter Estimates | | | | |
|---|---|---|---|---|
| **Variable** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 931.52722 | 149.94735 | 6.21 | <.0001 |
| **share_White_victim** | -391.53114 | 76.11184 | -5.14 | <.0001 |
| **income_2018** | -0.00977 | 0.00168 | -5.82 | <.0001 |
| **unemployment_rate** | -38.60913 | 16.96209 | -2.28 | 0.0286 |
| **percent_biden** | 5.99647 | 1.66340 | 3.60 | 0.0009 |
| **rate_killed** | 4.41523 | 0.76847 | 5.75 | <.0001 |

Dette giver os nu 6 signifikante variable, hvori kun rate_killed går igen fra opgave b. Dette giver nu en forklaringsgrad på R-Square = 0.7925. Der er lidt lavere end estimationen med alle variable.

Vi kan nu se at:

| Obs | state | t | c | h | d |
|---|---|---|---|---|---|
| **39** | New Hampshire | 0.86238 | 1.46145 | 0.28739 | 0.05033 |
| **40** | Kentucky | -1.32609 | 1.06228 | 0.16390 | 0.05633 |
| **41** | Rhode Island | -1.83552 | 0.79012 | 0.11953 | 0.07176 |
| **42** | Arkansas | 2.16852 | 0.63741 | 0.10193 | 0.08105 |
| **43** | Colorado | -1.40082 | 1.18898 | 0.27613 | 0.12168 |
| **44** | Maryland | 1.83685 | 0.85776 | 0.18955 | 0.12379 |

Dette er de stater, der nu har højest Cook's D. Altså de punkter, der påvirker estimaterne mest. Disse har relativt høje t-værdier og høje d-værdier og er dermed relativt dårlige punkter.