Hjemmeopgave 2

Videregående statistisk.

a)

Vi starter med at indlæse data via følgende kode:

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

Dernæst udfører vi en logistisk regression med backward selection, hvor vi fjerner de forklarende variable som er ikke signifikante på et 5% niveau, som køres via følgende kode:

```
proc logistic data=ssvs.engelske_byer plots=all;
model London=EC1011I EC1012I EC2009I DE2002I DE2003I DE3011I SA1007I
SA1011I SA1008I TE2028I TE2031I/selection=backward;
run;
```

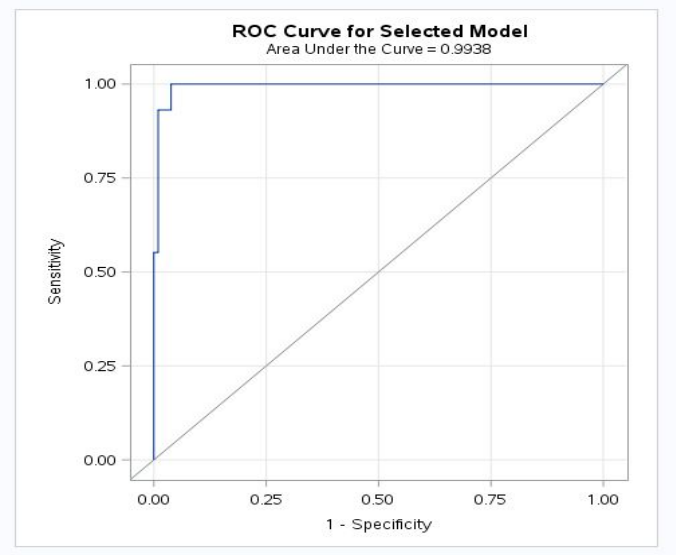
Dette giver følgende output:

Analysis of Maximum Likelihood Estimates

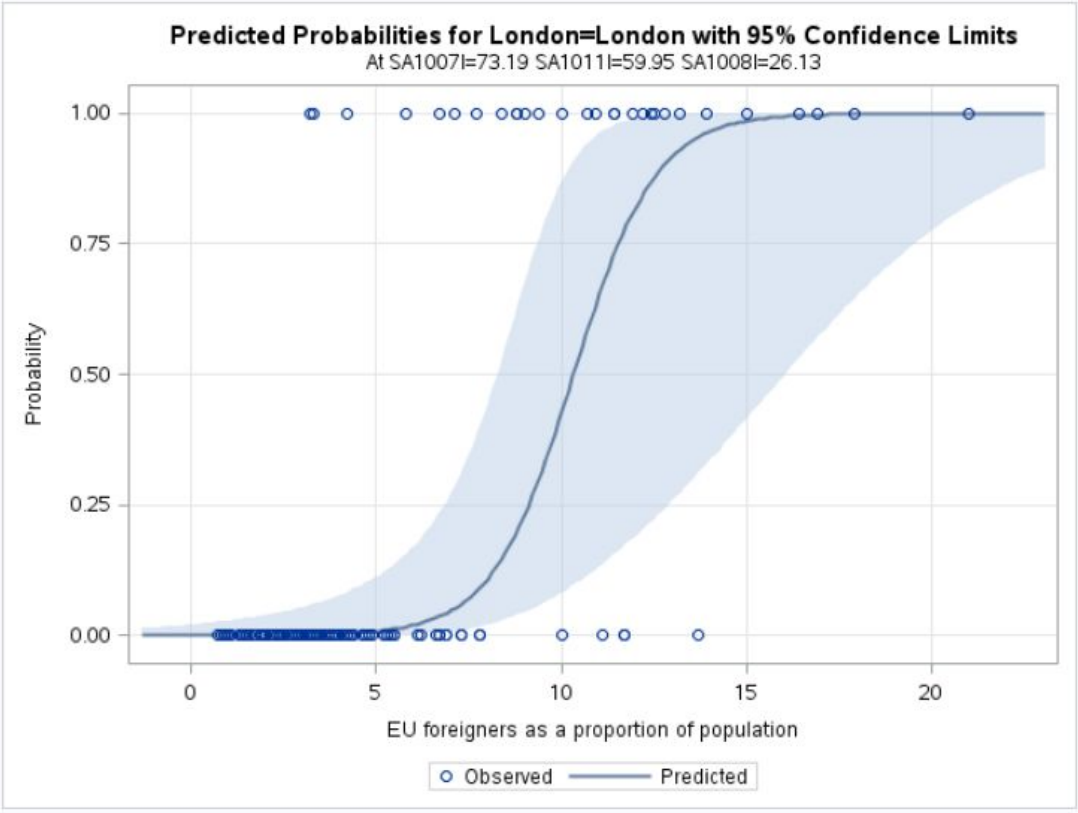
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-815.9	259.9	9.8535	0.0017
DE2002I	1	0.9071	0.3175	8.1607	0.0043
SA1007I	1	7.7899	2.5022	9.6922	0.0019
SA1011I	1	0.3279	0.1160	7.9902	0.0047
SA1008I	1	8.2921	2.6470	9.8135	0.0017

Via vores backward selection fremgår det at følgende forklarende variable er signifikante på et 5% niveau og dermed har en statistisk signifikant effekt på responsvariablen. Dermed haves en signifikant model.

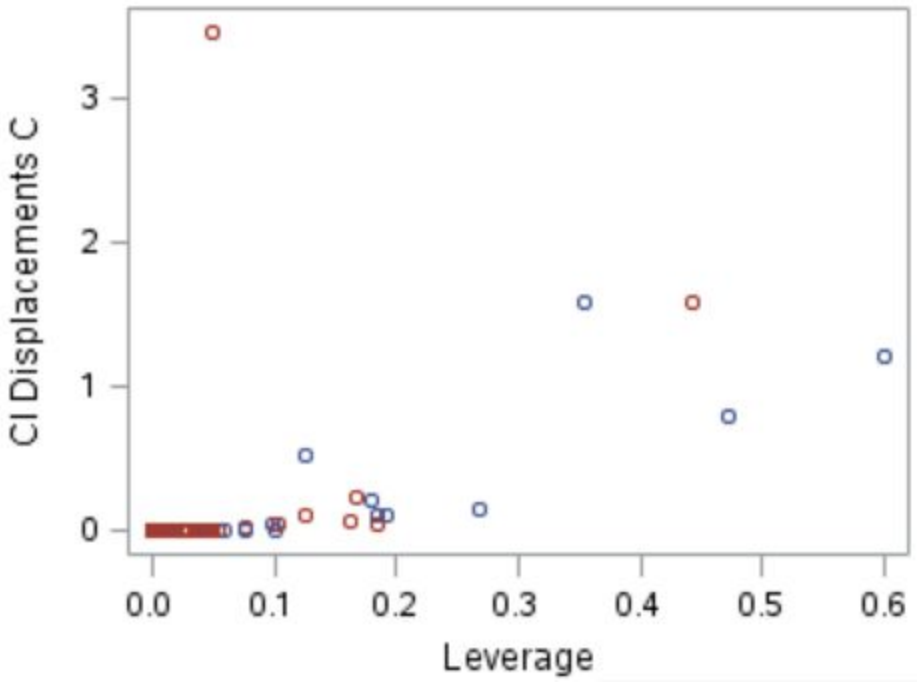
Vi kigger dernæst på fit via følgende ROC-kurve



Det fremgår af ROC-kurven, at modellen har et særdeles godt fit på 0,9938, som siger at modellen, dvs. de forklarende variable, kan forklare 99,3% af variationen i responsvariablen.



Dernæst kigger vi på indflydelsesrige observationer.



Disse specifikke observationer findes ved følgende kode, og tjekke hvilke byer som har den højeste CI Displacement (højest inflydelse):

```
proc logistic data=ssvs.engelske_byer plots=all;
  model London=DE2002I SA1007I SA1011I SA1008I/influence;
  output out=ny c=cidisplacement DFBETAS=DFBETAS;
run;
proc sort data=ny out=nysort;
by cidisplacement;
run;
proc print;
run;
```

bynavn	Population_total	London	Families_with_children	Landsdel	DFBETAS	cidisplacement
Havering	237245	London	2	London	0.11871	0.52815
Hackney	246275	London	2	London	-0.09100	0.78663
Kensington and Chelsea	158650	London	1	London	0.21847	1.20848
Camden	220340	Not London	1	England not London	-0.34741	1.58014
Barking and Dagenham	185910	London	3	London	-0.16450	1.58903
Slough	140210	Not London	3	England not London	1.46207	3.46527

Det bemærkes heraf, at observationer som Kensington and Chelsea, Slough, Camden og Barking er observationer, som falder uden for resten.
De indflydelsesrige observationer er, målt ved leverage, Kensington and Chelsea, som er et velstående område, som tiltrækker folk fra hele verden, som dermed gør, at observationer bliver atypisk. Ligeledes er Camden er meget multikulturel område, som ligeledes forklarer den høje andel af non-EU citizens.

b)
Vi starter med at finde de prædikterede værdier for modellen ift. Om byerne ligger i london eller ej. Dette tjekkes mod det faktiske via følgende kode:

```
proc logistic data=ssvs.engelske_byer ;
model London=DE2002I SA1007I SA1011I SA1008I;
score out=score;
run;
proc freq data=score;
table F_London*I_London;
run;
proc print data=score;
```

```
where F_London ne I_London;
run;
```

Dette resulterer i følgende output:

Obs	DE2002I	DE2003I	DE3002I	DE3011I	EC1011I	EC1012I	EC2009I	SA1007I	SA1008I	SA1011I	SA3005V	TE2028I	TE2031I	by	bynavn	Population_total	London	Families_with_children	Landsdel	F_London	I_London	P_London	P_Not London
38	8.8	10.8	27.6	41.5	14.8	10.4	15.8	69.9	29.8	47.7	1	25.0	25.9	UK102C1	Barking and Dagenham	185910	London	3	London	London	Not London	0.34798	0.65202
42	13.7	13.4	40.5	22.1	9.5	5.9	1.4	14.8	81.7	32.9	3	14.0	59.1	UK107C1	Camden	220340	Not London	1	England not London	Not London	London	0.52553	0.47447
50	3.2	2.6	29.0	30.1	6.3	8.8	6.1	78.3	21.4	74.4	1	33.9	24.2	UK116C1	Havering	237245	London	2	London	London	Not London	0.23638	0.76362
128	11.7	11.3	28.5	39.1	10.0	12.0	10.5	66.5	33.0	54.1	1	23.4	30.6	UK567C1	Slough	140210	Not London	3	England not London	Not London	London	0.98448	0.01552

Det ses, at nogle byer ifølge modellen har meget stor ssh. for at være i London, men faktisk ikke ligger i London. F.eks. en by som Slough har meget stor ssh. for at ligge i london, men ligger ikke i London, dette skyldes formentlig at Slough har en høj andel af non-EU citizens, som er ligeså høj som byer i London.

Dermed kan modellen rammer meget forkert, hvis by er en atypisk i forhold til hvor den ligger.

c)

Vi finder effectplottet via en logistisk regression med *Families_with_children* som responsvariabel mod de signifikante forklarende variable. Dette gøres via følgende kode.

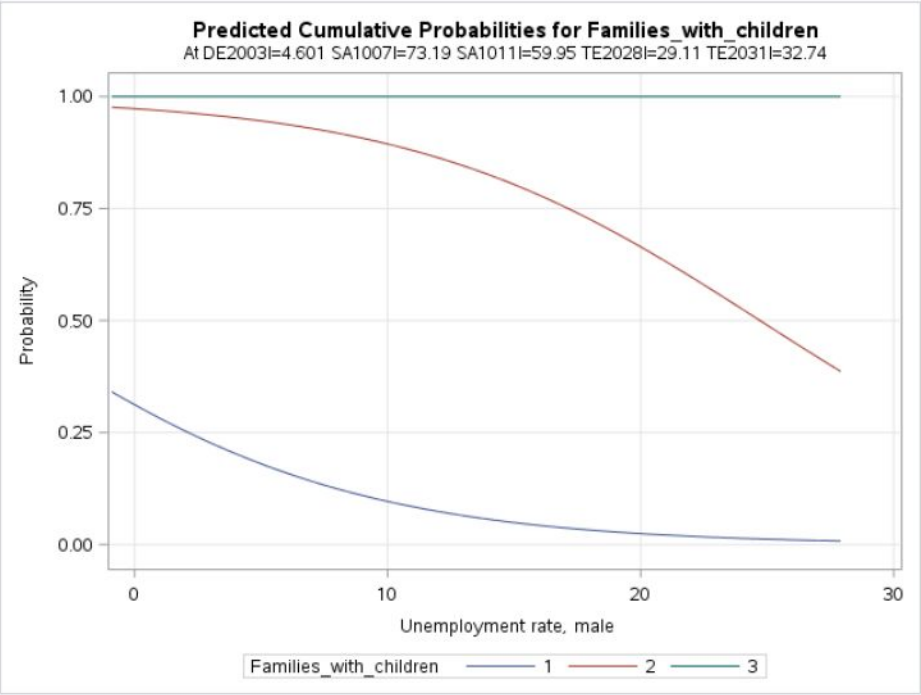
```
proc logistic data=ssvs.engelske_byer plots=all;
model Families_with_children =EC1011I EC1012I EC2009I DE2002I DE2003I
SA1007I SA1011I SA1008I TE2028I TE2031I/selection=backward;
run;
```

Dette giver følgende output:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	5.8049	6.0957	0.9069	0.3409
Intercept	2	1	10.1789	6.1454	2.7435	0.0977
EC1011I		1	-0.1450	0.0725	4.0006	0.0455
DE2003I		1	-0.7759	0.1774	19.1381	<.0001
SA1007I		1	-0.0582	0.0290	4.0203	0.0450
SA1011I		1	-0.2504	0.0514	23.6840	<.0001
TE2028I		1	0.3207	0.1362	5.5455	0.0185
TE2031I		1	0.2112	0.0543	15.1469	<.0001

Og et effectplot



Parallelitetsantagelsen kan umiddelbart ikke tænkes at være opfyldt. Det ville kræve, at arbejdsløshedsraten havde samme påvirkning på antallet af børnefamilier på tværs af kategorierne. Dette virker usandsynligt.
Modellen virker god. Det giver nemlig god mening, at sandsynligheden for at få børn falder når arbejdsløsheden stiger, hvilket er hvad modellen forudsiger, hvorfor modellen må siges at være god.

d)

Vi anvender samme metode som i spg. C via følgende kode:

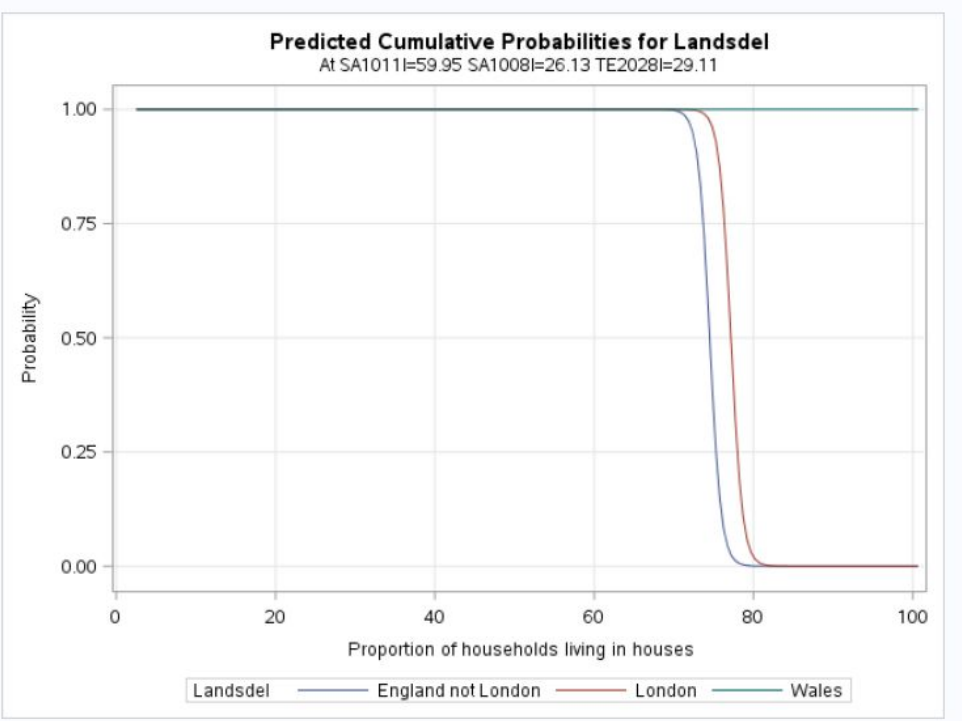
```
proc logistic data=ssvs.engelske_byer plots=all;
model Landsdel =EC1011I EC1012I EC2009I DE2002I DE2003I SA1007I
SA1011I SA1008I TE2028I TE2031I/ selection=backward;
run;
```

Dette giver følgende output:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	England not London	1	140.5	53.6184	6.8697	0.0088
Intercept	London	1	144.2	53.8648	7.1669	0.0074
SA1007I		1	-1.3756	0.5322	6.6806	0.0097
SA1011I		1	-0.1135	0.0404	7.8798	0.0050
SA1008I		1	-1.4802	0.5522	7.1870	0.0073
TE2028I		1	0.2556	0.0786	10.5785	0.0011

Og effectplot:



Modellen virker god. Det giver nemlig god mening, at når andelen af husholdninger som bor i hus stiger, så falder sandsynligheden for at husholdningen bor i London og England og det falder hurtigt, hvilket ligeledes er intuitivt. Dette er hvad modellen forudsiger, hvorfor modellen må siges at være god.

e)

Vi udfører en poisson regression ved at udfører en Generaliserede Lineære Modeller regression og definerer fordelingen til at være en poissonfordeling. Dette gøres via følgende kode:

```
proc genmod data=ssvs.engelske_byer plots=all;
model SA3005V=EC1011I EC1012I EC2009I DE2002I DE2003I DE3011I SA1007I
SA1011I SA1008I TE2028I TE2031I Population_total
/dist=p;
run;
```

Dette giver følgende output:

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6189	13.4835	-30.0460	22.8083	0.07	0.7884
EC1011I	1	0.0514	0.0205	0.0113	0.0915	6.32	0.0119
EC1012I	1	-0.0091	0.0239	-0.0560	0.0377	0.15	0.7031
EC2009I	1	-0.0195	0.0162	-0.0512	0.0121	1.46	0.2265
DE2002I	1	0.0299	0.0325	-0.0339	0.0936	0.84	0.3590
DE2003I	1	-0.0928	0.0386	-0.1685	-0.0171	5.77	0.0163
DE3011I	1	0.0085	0.0236	-0.0377	0.0547	0.13	0.7196
SA1007I	1	0.0767	0.1348	-0.1875	0.3409	0.32	0.5694
SA1011I	1	-0.0153	0.0127	-0.0402	0.0097	1.43	0.2315
SA1008I	1	0.0787	0.1373	-0.1905	0.3479	0.33	0.5666
TE2028I	1	-0.0773	0.0400	-0.1557	0.0010	3.74	0.0531
TE2031I	1	-0.0228	0.0173	-0.0566	0.0111	1.74	0.1873
Population_total	1	0.0000	0.0000	0.0000	0.0000	77.75	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Vi ser, at parameteren for den totale befolkningsstørrelse er en utrolig præcis estimator for antal mord, hvilket ikke er overraskende da der i en by med 500 indbyggere naturligvis er en mindre risiko for mord end i en by med 5 mio. indbyggere.
Ligeledes er det ikke overraskende er estimatet er tæt på 0 da en stigning i indbyggertallet med 1 øger antallet af mord med tæt på 0.