Jeg anvender data fra franske mænd under 40. Først importeres data:

**proc** **import** datafile="C:\Users\asbjp\Documents\KU\Videregående statistik\hjemmeopgave\ESS8e01" out=ud replace DBMS=sav;

**run**;

Derefter dannes et nyt datasæt *minedata* hvor alle undtagen franske mænd under 40 er slettet fra datasættes fra:

\*Sortér data som beskrevet i opgaveformuleringen;

**data** minedata;

set ud;

if agea < **40** then age1=**1**;

if **40**<= agea< **70** then age1=**2**;

if agea =>**70** then delete;

\*som en følge af de næste tre linier fjernes alle respondenter,

der aldrig bruger internet;

if netustm>**5000** then delete;

log\_netustm=log(netustm);

if log\_netustm=**.** then delete;

\*fjerner enkelte med manglende angivelse af køn;

if gndr=**.** then delete;

\*de næste tre liner skal rettes til jf fordelingstabellen;

if cntry ne 'FR' then delete;

if age1 ne **1** then delete;

if gndr ne 1 then delete;

**run**;

***a) Undersøg grafisk om variablene netustm og/eller log\_netustm kan antages at være normalfordelte.***

Den grafiske undersøgelse gennemføres ved at opstille histogrammer for variablene *netustm* (internetbrug) og logaritmen til denne. De holdes op mod en graf for henholdsvis normalfordelingen og logaritmen til normalfordelingen. Det gøres med koden:

**proc** **univariate** data=minedata;

\*Definér de variable vi ønsker beskrevet;

var netustm;

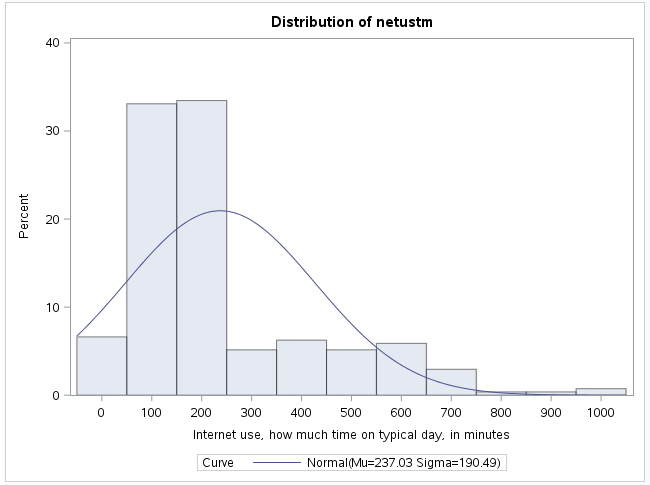
var log\_netustm;

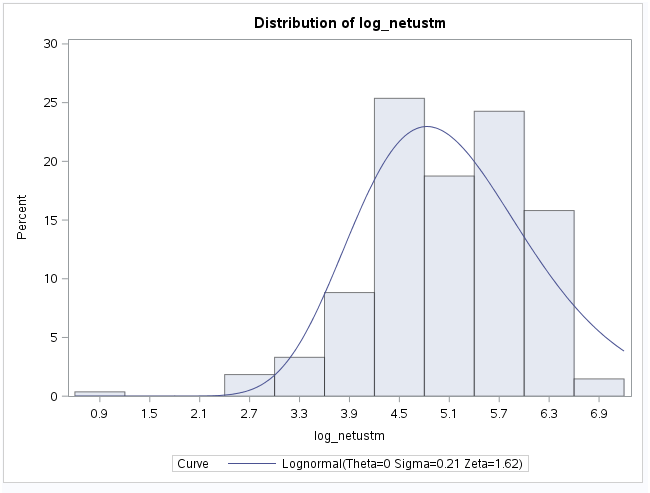
\*Definér histogram med relevant sammenligning;

histogram netustm/normal;

histogram log\_netustm/lognormal;

**run**;

Histogrammet for *netustm* med hvor SAS har tilnærmet en normalfordeling med middelværdi 237,0294 og en standardafvigelse på 190,49163, samt en skewness på 1,476, og dermed er meget højreskæv og kurtosis på 1,96. Her ses det at denne fordeling afviger fra normalfordelingen idet en normalfordeling har kurtsis 3 og og skewness på 0.

Histogrammet for *log\_netustm* med hvor SAS har tilnærmet en normalfordeling med middelværdi 5,144 og en standardafvigelse på 0,8742 samt en skewness på -0,767 og dermed er meget venstreskæv og kurtosis på 2,095. Her ses det at denne fordeling afviger fra normalfordelingen idet en normalfordeling har kurtsis 3 og og skewness på 0.

Dermed vil det ikke være præcist at antage at *neustm* og *log\_neustm* er normalfordelte.

***b) Test om middelværdien af netustm er lig med 197.9 i dit datasæt "minedata". Tallet 197.9 er gennemsnittet i det samlede ESS datamateriale. Benyt eventuelt den logaritmisk transformerede.***

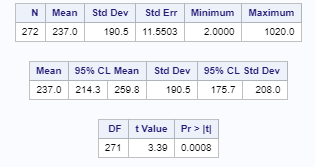
Det vil test om middelværdien af *netustm* er lig OECD gennemsnittet på 197,9. Dette gøres ved at udføre et t-test på mine data. Dette gøres via følgende stump kode:

**proc** **ttest** data=minedata h0=**197.9**;

var netustm;

**run**;

som giver følgende output



Det ses at den empirisk middelværdi bliver 237 med en standardafviglese på 190,5 og maks på 1020 og min på 2. Det ses at t-teststørrelsen bliver 3,39 og p-værdien på 0,8%. Testes der med et signifikansniveau på 5% fås den kritisk værdi til 1,96. Det ses heraf klart at t-teststørrelsen er større end den kritisk værdi og p-værdien er under signifikansniveauet hvormed vi kan afvise nulhypotesen og der er dermed må franske mænd under 40 internet forbrug være signfikant forskelligt fra 197,9.

***c)*** *Udfør en regressionsanalyse, hvor variablen trstep er responsvariabel og trstprl er den forklarende variabel. Kommenter resultatet.*

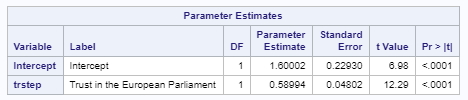
*Trstep* regresseres på *trstprl*. Dette gøres via følgende kode i SAS:

**proc** **reg** data=minedata;

model trstprl=trstep;

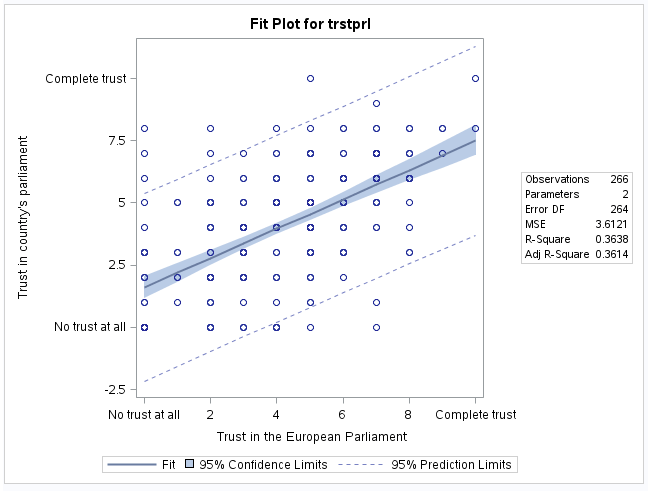
**run**;

Det giver os outputtet:



Vi estimerer en lineære regressionsmodel af følgende type

Dette giver et estimat . Dette skal fortolkes som at alt andet lige effekten af, at tilliden til EU-parlamentet øges med 1 point er, at tilliden til det nationale parlament øges med 0,59 point. Vi estimerer ligeledes som siger at når tilliden til det EU-parlament er 0 er tilliden til det til det nationale parlament positivt, som siger at tilliden til den nationale parlament er højere end EU-parlamentet. Begge estimater er signifikant på 1% niveau.



Det fremgår af plottet af R2=0,36, som betyder at variationen i tilliden til EU-parlamentet kan forklarer 36% af variation til det nationale parlament.

***d)*** *Hvad gør følgende program?*

**data** zero\_trust;

set minedata;

no\_trust\_ep=**1**;

if trstep>**0** then no\_trust\_ep=**0**;

no\_trust\_prl=**0**;

if trstprl=**0** then no\_trust\_prl=**1**;

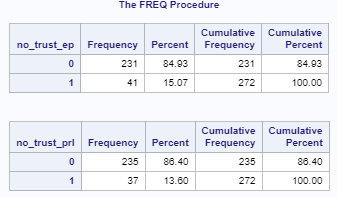
**run**;

Linje 1 genererer et nyt datasæt, der hedder zero\_trust. I linje 2 sættes datasættet til at tage udgangspunkt i datasættet *minedata*, dermed data fra ESS-undersøgelsen for franske mænd under 40. Linje 3 generer en ny variabel no\_trust\_ep, der som udgangspunkt er lig 1. Linje 4 fortæller, at variablen *no\_trust\_ep* skal antage værdien 0, hvis tilliden til Europaparlamentet, variablen *trstep* er større end 0. Linje 5 danner en variabel *no\_trust\_prl*, der som udgangspunkt er lig 0. Linje 6 fortæller, at variablen *no\_trust\_prl* skal antage værdien 1, hvis *trstprl* er lig 0.

Programmet danner altså med udgangspunkt i ESS-undersøgelsen to binære variable, nemlig en binær for ingen tillid til EU-parlamentet, der antager værdien 1, hvis man ikke har tillid til EU-parlamentet, og tilsvarende for det nationale parlament. Måden, de to variable dannes på, er lidt forskellig, da *no\_trust\_ep* som udgangspunkt sættes til 1, mens *no\_trust\_prl* som udgangspunkt sættes til 0.

***e)*** *Undersøg i datasættet zero\_trust, om der er samme andel med værdien 1 af de to variable no\_trust\_ep og no\_trust\_prl. Kan hypotesen testes?*

Første skridt er at beregne procentdelen som ikke har tillid hhv. EU- og det nationale parlament og dernæst teste om disse andele er statistisk forskellige fra hinanden.



Det ses her at 15% ikke har tillid til EU-parlament og 13,6 ikke har tillid til det national parlament. Vi tester om disse værdier er statistisk signifikant via en chi i anden test med følgende kode:

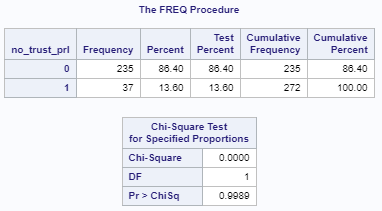
**proc** **freq** data=zero\_trust;

tables no\_trust\_prl/testp=(**0.864**,**0.136**);

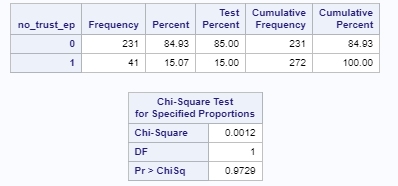
tables no\_trust\_ep/testp=(**0.85**,**0.15**);

**run**;

Der ligges ud med at teste, om den sande værdi for andelen af franske mænd under 40, der ikke har tillid til det nationale parlament, er lig 13,6%



Vi får en *p*-værdi på 0,99 og kan dermed ikke afvise nulhypotesen på et 5% signifikansniveau. Og dermed kan den sande værdi godt være 13,6%. Dernæst testes, om den sande andel, der ikke har tillid til EU-parlamentet, er 15%:



Vi får en *p*-værdi på 0,97 og kan dermed ikke afvise nulhypotesen på et 5% signifikansniveau. Og dermed kan den sande værdi godt være 15%. Ligeldes kan vi nu godt forventet at andelen der har tillid til det nationale og EU parlamentet er ens.

***f****) Test om de to variable no\_trust\_ep og no\_trust\_prl er uafhængige*

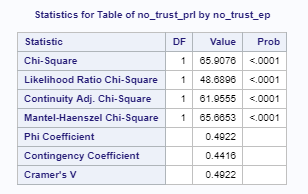
Vi tester om de to variable er uafhængige via følgende stykke kode som udfører en chi-i-anden test

**proc** **freq** data=zero\_trust;

tables no\_trust\_prl\*no\_trust\_ep/chisq;

**run**;

Vi får følgende output



Vi betragter chi-i-anden p-værdien på mindre end 0,1% så på et 5 procents signifikansniveau kan vi altså ikke afvise nulhypotesen om, at *no\_trust\_prl* og *no\_trust\_ep* skulle være uafhængige.

***g****) Test om den gennemsnitlige tid, respondenterne bruger internettet, er den samme for respondenter med no\_trust\_ep=0 som for respondenter med no\_trust\_ep=1,*

Vi udfører et t-test på nulhypotesen at den daglig forbrug af internet er ens for de to grupper. Dette gøres via følgende kode

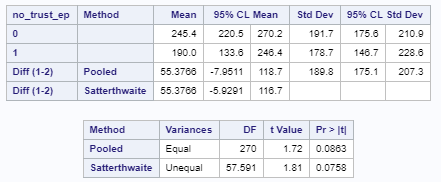
**proc** **ttest** data=zero\_trust;

class no\_trust\_ep;

var netustm;

**run**;

Dette giver følgende output



Givet et 5% signifikansniveau fås en kritisk t-værdi i t fordelingen på 1,96 og testes begge t-teststørrelse ses det at begge størrelser er under den kritisk værdi og dermed kan vi ikke afvise at det daglig internetforbrug ikke er det samme for de to grupper. Det skal dog bemærkes at ved f.eks. et 10% niveau er resultatet signifikant og dermed er vi et grænsetilfælde og det kunne indikere en forskel mellem de to grupper.

***h****) Udfør en regressionsanalyse på datasættet zero\_trust, hvor variablen trstep er responsvariabel og begge de variable trstprl og netustm er forklarende variable. Kommenter resultatet.*

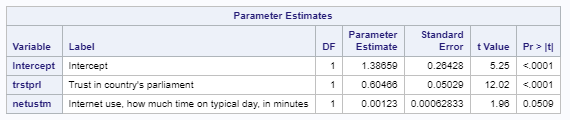
Vi ønsker altså at estimere multiple lineære regressionsmodel:

Dette gøres ligsom i opgave c dog med en ekstra forklarende variable

**proc** **reg** data=zero\_trust;

model trstep=trstprl netustm;

**run**;

vi får følgende resultat.

Dette giver et estimat . Dette skal fortolkes som at alt andet lige effekten af, at tilliden til det nationale parlament øges med 1 point er, at tilliden til EU-parlamentet øges med 0,605 point. Ligeledes estimeres . Dette skal fortolkes som at alt andet lige effekten af, daglig internetforbrug øges med 1 point er, at tilliden til det nationale parlament øges med 0,605 point. Vi estimerer ligeledes som ikke har en fortolkning i en multipel model. og er signifikant på et 1% niveau og er næsten signifikant på et 5% niveau.