UNIVERSITETET I OSLO

DET MEDISINSKE FAKULTET

Eksamen
ERN4000 – Statistikk
Fredag 6. september 2024 kl 09:00-12:00

Hjelpemidler

- Stata (på den stasjonære maskinen)
- Egen kalkulator (med exp og ln)
- Stata Cheat Sheet
- Lærerbok: Aalen OO (red), Frigessi A, Moger TA, Scheel I, Skovlund E, Veierød MB. Statistiske metoder i medisin og helsefag. Gyldendal Akademisk 2006.
  eller
  Aalen OO (red), Frigessi A, Moger TA, Scheel I, Skovlund E, Veierød MB. Statistiske metoder i medisin og helsefag, 2. utgave. Gyldendal Akademisk 2018.

Kort forklaring til oppgavene

I oppgavene skal du foreta statistisk analyse av noen medisinske problemstillinger. I presentasjonen av resultatene er det tre statistiske mål som er viktige å få frem: estimat (f.eks. gjennomsnitt, andel, odds ratio, regresjonskoeffisient), konfidensintervall og p-verdi.

Datafilen nhanes.dta brukes i oppgave 1, 2 og 3

Datafilen er fra data samlet inn av US National Center for Health Statistics. Data fra 750 deltakere er inkludert med følgende variabler:

Variabelnavn	**Beskrivelse**
Age	Alder (år)
Gender	Kjønn (1=kvinne, 2=mann)
Weight	Vekt (kg)
Height	Høyde (cm)
TotChol	Kolesterol (mmol/L)
Diabetes	Diabetes (1=nei, 2=ja)
PhysActive	Fysisk aktivitet (1=nei, 2=ja)

Oppgave 1 (totalt 30)

a) **(6)** *Gi en beskrivelse av alder, vekt, høyde og kolesterol. Bruk det du synes er mest hensiktsmessig av beskrivende statistikk og presenter resultatene i en tabell.*

	Gjennomsnitt	Standardavvik	Minimun-maksimum
Alder	46.46	17.06	18 - 80
Vekt	81.96	19.74	39 – 167.6
Høyde	168.94	10.27	144.4 – 199.5
Kolesterol	5.02	1.07	2.4 – 12.28

Histogrammene er nokså symmetriske. Gjennomsnitt og median er ganske like.

```
. summarize Age, detail

                            Age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18             18
 5%           20             18
10%           23             18       Obs                 750
25%           32             18       Sum of wgt.         750

50%           46                      Mean           46.46267
                      Largest       Std. dev.      17.05745
75%           59             80
90%           70             80       Variance       290.9565
95%           76             80       Skewness       .1431502
99%           80             80       Kurtosis       2.042872

. summarize Weight, detail

                          Weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%          47.3             39
 5%            55           41.1
10%         58.25           44.4       Obs                 750
25%          66.9           45.4       Sum of wgt.         750

50%         80.65                      Mean           81.95773
                      Largest       Std. dev.      19.73813
75%          93.5            145
90%         108.3          148.2       Variance       389.5936
95%         117.5          149.1       Skewness        .678252
99%         138.7          167.6       Kurtosis       3.524932

. summarize Height, detail

                          Height
-------------------------------------------------------------
      Percentiles      Smallest
 1%         147.9          144.4
 5%         152.5          146.3
10%         155.6          146.4       Obs                 750
25%         161.3          146.6       Sum of wgt.         750

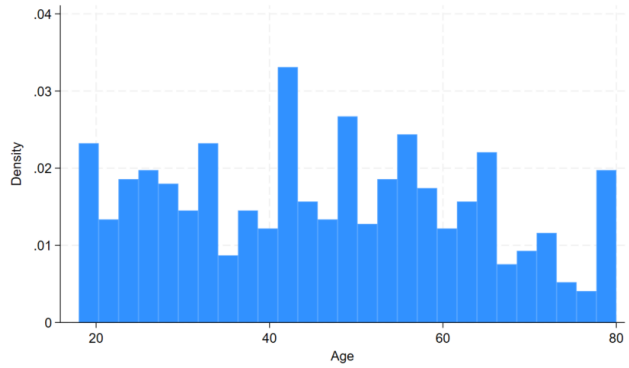
50%         168.4                      Mean           168.9377
                      Largest       Std. dev.      10.26743
75%           176          193.7
90%        182.65          199.4       Variance       105.4202
95%         186.9          199.5       Skewness       .1913956
99%         190.8          199.5       Kurtosis       2.591322

. summarize TotChol, detail
```

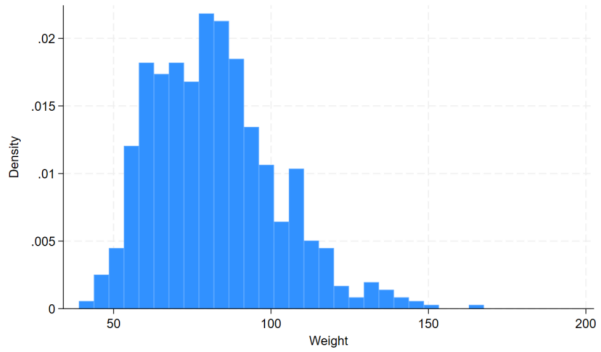
```
                          TotChol
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3.08            2.4
 5%         3.41            2.4
10%         3.65            2.9       Obs                 750
25%         4.27            2.9       Sum of wgt.         750

50%         4.94                      Mean            5.02244
                       Largest        Std. dev.       1.07182
75%         5.66           8.22
90%         6.48           8.22       Variance        1.148799
95%          6.8           9.31       Skewness        .7760964
99%         8.02          12.28       Kurtosis        5.579736
```

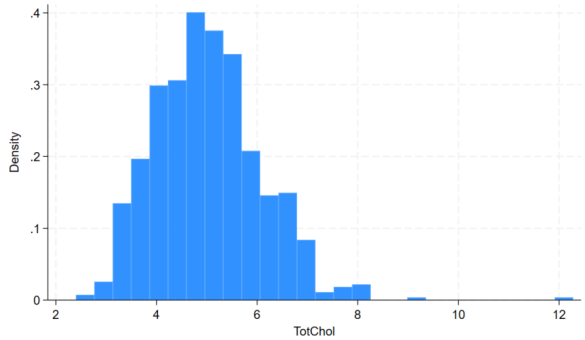
```
histogram Age
```




```
histogram Weight
```




```
histogram TotChol
```



b) **(6)** *Lag tre tabeller: en for kjønn, en for diabetes og en for fysisk aktivitet. Hver tabell skal vise antall personer og prosentandel i hver av variablenes kategorier, samt totalsummen.*

```
. tabulate Gender
```

```
     Gender |      Freq.     Percent        Cum.
------------+-----------------------------------
     female |         368       49.07       49.07
       male |         382       50.93      100.00
------------+-----------------------------------
      Total |         750      100.00

. tabulate Diabetes

   Diabetes |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |         663       88.40       88.40
        Yes |          87       11.60      100.00
------------+-----------------------------------
      Total |         750      100.00

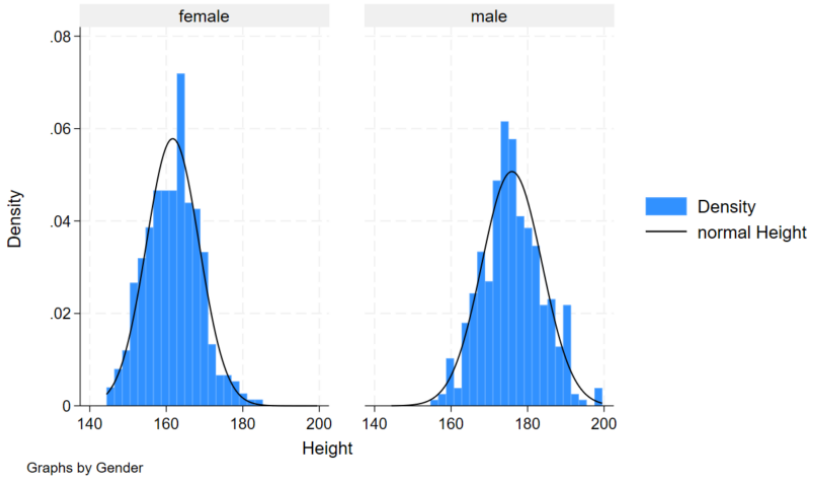
. tabulate PhysActive

 PhysActive |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |         341       45.47       45.47
        Yes |         409       54.53      100.00
------------+-----------------------------------
      Total |         750      100.00
```

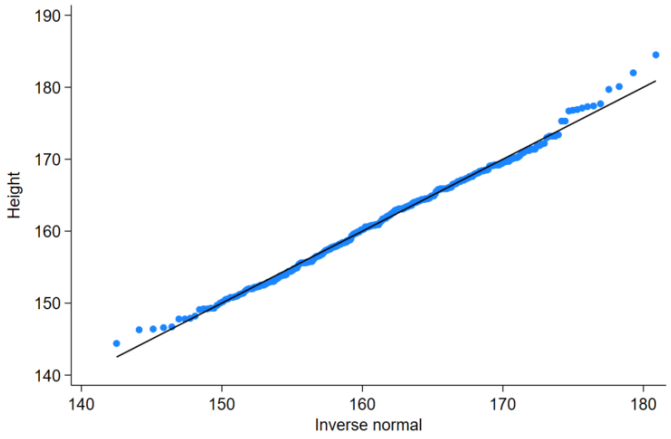
c) **(9)** *Sammenlign høyden for menn og kvinner ved hjelp av en statistisk test. Begrunn valget av test. Kommenter kort det du finner.*

Dataene tilfredsstiller forutsetningene for to-utvalgs t-test: to uavhengige utvalg, noenlunde lik varians og tilnærmet normalfordeling i de to gruppene. To-utvalgs t-test ga signifikant forskjell i høyde hos kvinner og menn, $p<0.001$. Menn var i gjennomsnitt 14.22 cm høyere enn kvinnene, 95% konfidensintervall (13.16, 15.28).

```
. histogram Height, by(Gender) normal
```



```
. qnorm Height if Gender==1
```




```
. ttest Height, by (Gender)

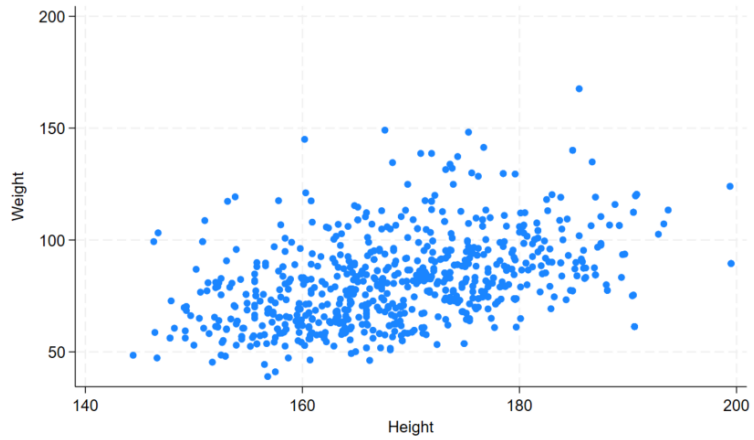
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
  female |     368    161.6951    .3597378    6.900968    160.9877    162.4025
    male |     382    175.9149    .4025288    7.867353    175.1235    176.7064
---------+--------------------------------------------------------------------
Combined |     750    168.9377    .3749137    10.26743    168.2017    169.6737
---------+--------------------------------------------------------------------
    diff |           -14.21981    .5411691                -15.2822   -13.15742
------------------------------------------------------------------------------
    diff = mean(female) - mean(male)                              t = -26.2761
H0: diff = 0                                     Degrees of freedom =      748

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

d) **(9)** *Lag et spredningsdiagram for sammenhengen mellom høyde (x-akse) og vekt (y-akse), og beregn Pearsons korrelasjonskoeffisient for høyde og vekt med konfidensintervall. Kommenter kort det du finner.*

Pearsons korrelasjonskoeffisient er 0.44, 95% konfidensintervall (0.39, 0.50). Det er en moderat grad av lineær sammenheng mellom høyde og vekt, illustrert ved at det i spredningsdiagrammet er en viss grad av samling av datapunktene rundt en tenkt regresjonslinje.

```
. twoway (scatter Weight Height)
```



```
. bootstrap r(rho): corr Height Weight
(running correlate on estimation sample)

warning: correlate does not set e(sample), so no observations will be excluded from
the
         resampling because of missing values or other reasons. To exclude
observations, press
         Break, save the data, drop any observations that are to be excluded, and
rerun
         bootstrap.

Bootstrap replications (50):
.........10.........20.........30.........40.........50 done


Bootstrap results                                       Number of obs = 750
                                                        Replications  =  50

      Command: correlate Height Weight
        _bs_1: r(rho)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .4446559   .0272607    16.31   0.000     .3912258    .4980859
------------------------------------------------------------------------------
```

Oppgave 2 (totalt 24)

Vi fortsetter med datafilen nhanes.dta, og ser på sammenhengen mellom fysisk aktivitet og diabetes. Fysisk aktivitet er eksponering og diabetes er sykdom.

a) (6) *Lag en tabell med diabetes i radene og fysisk aktivitet i kolonnene. I tillegg til å inkludere antall i hver celle skal tabellen prosentueres innen hver kolonne. Kommenter kort det du finner.*

Blant de som ikke er fysisk aktive har 15.84% diabetes og blant de som er fysisk aktive har 8.07% diabetes.

```
. tabulate Diabetes PhysActive, column

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |      PhysActive
  Diabetes |        No        Yes |     Total
-----------+----------------------+----------
        No |       287        376 |       663
           |     84.16      91.93 |     88.40
-----------+----------------------+----------
       Yes |        54         33 |        87
           |     15.84       8.07 |     11.60
-----------+----------------------+----------
     Total |       341        409 |       750
           |    100.00     100.00 |    100.00
```

b) **(9)** *Vi fortsetter med tabellen i punkt a). Beregn relativ risiko (risk ratio) med 95% konfidensintervall og risikodifferansen med 95% konfidensintervall. Kommenter kort det du finner.*

Relativ risiko er 0.51, 95% konfidensintervall (0.34, 0.77)
Risikodifferansen er -0.08, 95% konfidensintervall (-0.12, -0.03).
De fysisk aktive individene har redusert risiko for diabetes sammenlignet med de som ikke er fysisk aktive.

```
. csi 33 54 376 287

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |        33          54  |         87
        Noncases |       376         287  |        663
-----------------+------------------------+-----------
           Total |       409         341  |        750
                 |                        |
            Risk |   .0806846    .1583578  |       .116
                 |                        |
                 |      Point estimate    |    [95% conf. interval]
                 |------------------------+------------------------
 Risk difference |        -.0776732       |   -.1245572   -.0307891
      Risk ratio |         .5095083       |    .3386362    .7666005
 Prev. frac. ex. |         .4904917       |    .2333995    .6613638
 Prev. frac. pop |         .2674815       |
                 +-------------------------------------------------
                              chi2(1) =    10.94  Pr>chi2 = 0.0009
```

c) **(9)** *Hvilken test vil du bruke for å undersøke om det er signifikant sammenheng mellom fysisk aktivitet og diabetes? Begrunn svaret. Utfør testen og oppgi p-verdien fra testen. Kommenter kort det du finner.*

Vi anvender en kjikvadrat test. Forutsetningen er at forventet antall er >5 i alle cellene i 2x2 tabellen, slik det er her. Vi finner at det er signifikant sammenheng, p<0.001, mellom diabetes og fysisk aktivitet (slik resultatene i b også viste).

```
. tabulate Diabetes PhysActive, expected

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

           |      PhysActive
  Diabetes |        No        Yes |     Total
-----------+----------------------+----------
        No |       287        376 |       663
           |     301.4      361.6 |     663.0
-----------+----------------------+----------
       Yes |        54         33 |        87
           |      39.6       47.4 |      87.0
-----------+----------------------+----------
     Total |       341        409 |       750
           |     341.0      409.0 |     750.0
```

```
. tabulate Diabetes PhysActive, chi2

           |       PhysActive
  Diabetes |        No        Yes |     Total
-----------+----------------------+----------
        No |       287        376 |       663
       Yes |        54         33 |        87
-----------+----------------------+----------
     Total |       341        409 |       750

          Pearson chi2(1) =  10.9408   Pr = 0.001
```

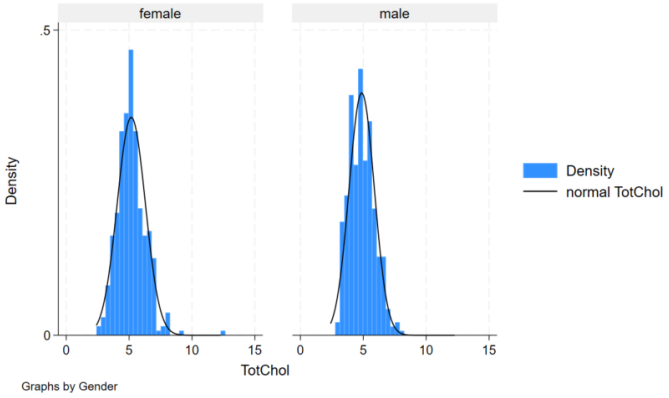
Oppgave 3 (totalt 33)

Vi fortsetter med datafilen nhanes.dta.

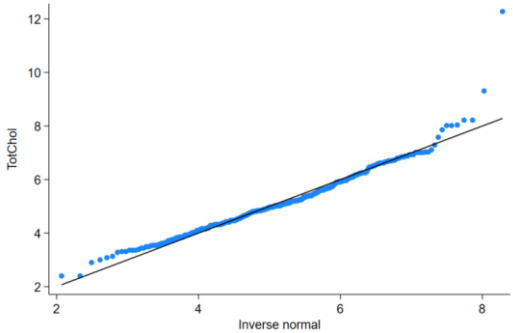
a) (9) Forskerne ville undersøke om kolesterol nivået er forskjellig hos menn og kvinner. Hvilken test vil du bruke for å undersøke dette? Utfør testen og kommenter kort det du finner.

Dataene tilfredsstiller forutsetningene for to-utvalgs t-test: to uavhengige utvalg, noenlunde lik varians og tilnærmet normalfordeling i de to gruppene. To-utvalgs t-test ga signifikant forskjell i kolesterol hos kvinner og menn, $p<0.001$. Kvinnene hadde i gjennomsnitt 0.30 mmol/L høyere kolesterol enn mennene, 95% konfidensintervall (0.15, 0.46).

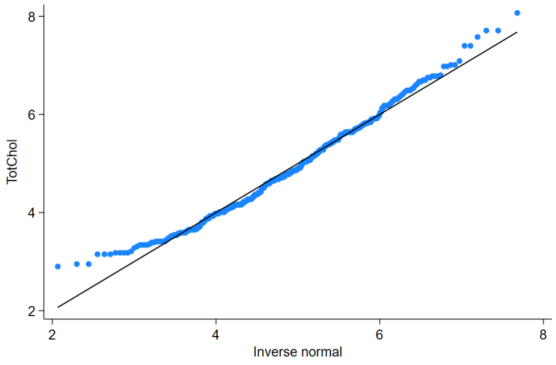
```
. histogram TotChol, by(Gender) normal
```




```
. qnorm TotChol if Gender==1
```



```
. qnorm TotChol if Gender==2
```




```
. ttest TotChol, by (Gender)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
  female |     368    5.177065    .0582461    1.117354    5.062527    5.291603
    male |     382    4.873482    .0514312    1.005214    4.772357    4.974606
---------+--------------------------------------------------------------------
Combined |     750     5.02244    .0391373     1.07182    4.945608    5.099272
---------+--------------------------------------------------------------------
    diff |            .3035835    .0775503                .1513415    .4558256
------------------------------------------------------------------------------
    diff = mean(female) - mean(male)                              t =   3.9147
H0: diff = 0                                     Degrees of freedom =      748

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 0.0000
```

b) **(9)** *Forskerne var interessert i om sammenhengen mellom kjønn og kolesterol endret seg hvis de utførte en multippel regresjonsanalyse og justerte for alder og vekt. Utfør en slik multippel regresjonsanalyse (kolesterol er 'dependent variable'). Er kjønn signifikant i denne modellen? Oppgi regresjonskoeffisienten for kjønn med 95% konfidensintervall. Hvordan fortolkes den estimerte regresjonskoeffisienten for kjønn?*

Sammenhengen mellom kjønn og kolesterol er fortsatt signifikant etter justering for alder og vekt, p=0.001 for kjønn. Regresjonskoeffisienten for kjønn er -0.28, 95% konfidensintervall (-0.44, -0.12). Etter justering for alder og vekt, har kvinnene i gjennomsnitt 0.28 mmol/L høyere kolesterol enn mennene, 95% konfidensintervall (0.12, 0.44).

```
. regress TotChol Gender Age Weight

      Source |       SS           df       MS      Number of obs   =       750
-------------+----------------------------------   F(3, 746)       =     15.69
       Model |  51.0837679         3  17.0279226   Prob > F        =    0.0000
    Residual |  809.366467       746  1.08494164   R-squared       =    0.0594
-------------+----------------------------------   Adj R-squared   =    0.0556
       Total |  860.450235       749  1.14879871   Root MSE        =    1.0416

------------------------------------------------------------------------------
     TotChol | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      Gender |  -.2813936    .082101    -3.43   0.001    -.4425701   -.1202172
         Age |   .0124749   .0022359     5.58   0.000     .0080855    .0168643
      Weight |  -.0001256   .0020789    -0.06   0.952    -.0042068    .0039557
       _cons |   4.877829   .2022171    24.12   0.000     4.480847    5.274811
------------------------------------------------------------------------------
```

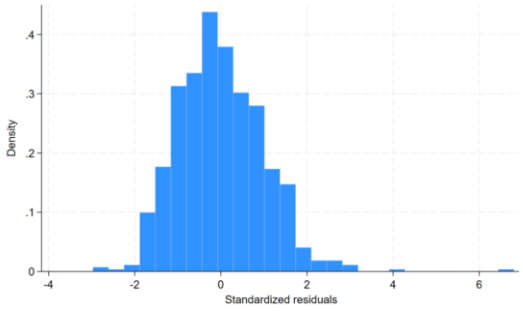
c) **(6)** *Hva er R^2 for den multiple regresjonsmodellen? Hvordan tolkes denne?*

R^2=0.0594. R^2 er forklart varians. Regresjonsmodellen med kjønn, alder og vekt forklarer 5.9% av den totale variasjonen i kolesterol.

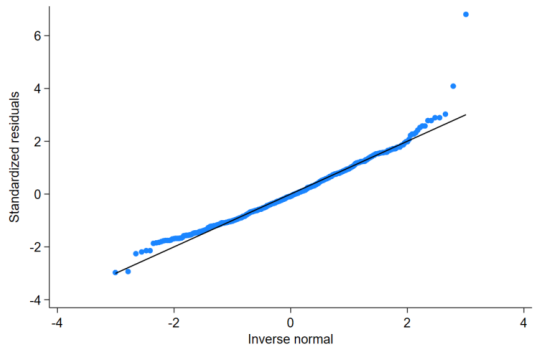
d) **(9)** *Var forutsetningene til stede for å gjøre analysene i b). Begrunn svaret.*

Ja.
[Kommentar: En kvinne har 12.28 i kolesterol og det er henne vi ser som ekstremverdi i figurene i a) og d). Til tross for at utvalget er ganske stort (n=750), vil en så ekstrem verdi kunne påvirke estimatene i noen grad. Siden hun er med i datafila må vi anta at hun tilfredsstiller studiens inklusjons- og eksklusjons kriterier.]

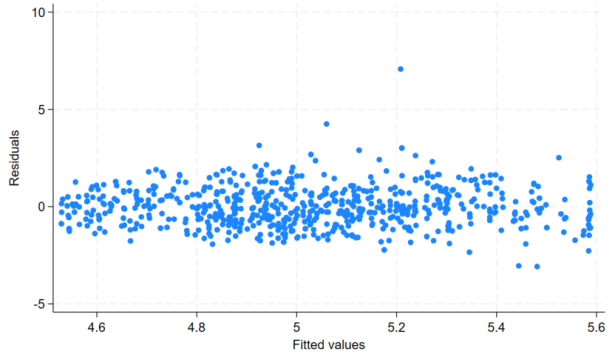
```
. predict stres, rsta
. hist stres
```




```
. qnorm stres
```




```
. rvfplot
(Også OK:
. rvfplot, yline (0))
```



Oppgave 4 (totalt 13)

Multippel sklerose (MS) er en potensielt invalidiserende sykdom. Noen pasienter med alvorlig MS kan miste evnen til å gå, mens andre kan oppleve milde symptomer som nummenhet eller svakhet i lemmer, tretthet eller kognitive problemer.

Forskjellige diagnostiske metoder, inkludert magnetresonanstomografi (MR), brukes til å diagnostisere sykdommen. For å bekrefte sykdommen er det imidlertid nødvendig med detaljert sykehistorie, nevrologisk undersøkelse, blodprøver og cerebrospinalvæskeanalyse.

I en hypotetisk klinikk ble det utført MR for 50 pasienter. Til slutt ble 32 pasienter bekreftet å ha MS, og av disse hadde 25 positive MR-resultater. Blant de som ikke har MS, hadde 3 pasienter positive MR-resultater.

Fyll inn antallene i tabellen nedenfor, og beregn sensitivitet og spesifisitet for MR.

	MS		
MR resultat	*Ja*	*Nei*	*Totalt*
Positiv	*25*	*3*	28
Negativ	7	15	22
Totalt	*32*	18	*50*

Sensitivitet: 25/32 = 0.78
Spesifisitet: 15/18 = 0.83

Maksimal poengsum
Oppgave 1: 30
Oppgave 2: 24
Oppgave 3: 33
Oppgave 4: 13

Krav til bestått
- Grensen for bestått er 65 av 100 mulige poeng.
- Besvarelsen skal vise at kandidaten har bred faglig kunnskap innen hele emnet og god kunnskap innen de mest sentrale områdene. Ingen av disse oppgavene er utenfor de mest sentrale områdene. Dette betyr at manglende/totalt utilfredsstillende besvarelse av enkelte oppgaver ikke kan kompenseres ved svært god besvarelse av andre.
- Bruken av fagterminologi er tilfredsstillende, og besvarelsen er klar og presis.
- Kandidaten viser god faglig vurderingsevne og selvstendighet.