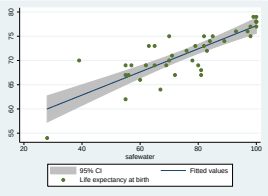
Statistics in Stata

Quantitative data, Group comparisons and Linear regression

Klaus K. Holst

29 Sep 2014



Statistical Methods , One sample t-test

One sample t-test

Used to test simple hypothesis regarding the mean in a single group. Independent samples and data approximately normal distributed (but fairly robust in large samples).

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n$$

Two-sided hypothesis

$$H_0 \colon \mu = \mu_0, \quad H_A \colon \mu \neq \mu_0$$

The analysis should of course be preceeded by graphical and descriptive analysis!

browse, summarize, graph histogram, graph qnorm, graph box, ...

Astronaut data

Trial with 26 astronauts (Bungo et.al., 1985) split into a control group (n=9) and a group (n=17) consuming extra salt and liquid before landing to treat space deconditioning.

```
use http://publicifsv.sund.ku.dk/~kkho/undervisning/
    data/astronaut, clear
```

Puls (beats pr minute) before and after flight for each astronaut.

```
describe
```

```
ut.dta
 obs:            26                          Pulse in two groups of
                                               astronauts before and after
                                               flight
 vars:            3                          27 Sep 2014 13:47
 size:          104
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
salt            byte    %8.0g                 Control group: 0, Salt: 1
pre             byte    %8.0g                 Pre-flight pulse (beats pr
```


Astronaut data (Paste specials directly into Data Editor)

salt	pre	post
1	71	61
1	65	59
1	52	47
1	68	65
1	69	69
1	49	50
1	49	51
1	57	60
1	51	57
1	55	64
1	58	67
1	57	69
1	59	72
1	53	69
1	53	72
1	53	75
1	48	77
0	61	61
0	59	66
0	52	61
0	54	68
0	53	77
0	78	103
0	52	77
0	54	80
0	52	79

One sample t-test

We will just consider the salt treated group.

```
save data/astronaut, replace
drop if salt==0
```

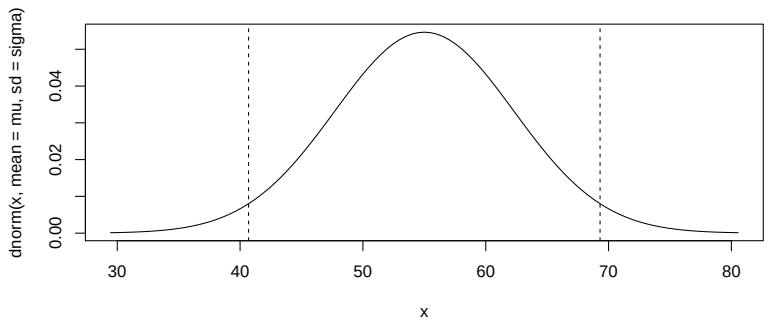
and examine if the pre-flight pulse (population) mean could be 60 beats pr minute

```
summarize pre
return list
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         pre |        17    56.88235    7.296252         48         71

scalars:
                  r(N) =  17
              r(sum_w) =  17
               r(mean) =  56.88235294117647
                r(Var) =  53.23529411764706
                 r(sd) =  7.296252059629454
               r(min) =  48
               r(max) =  71
               r(sum) =  967
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

One sample t-test




```
local sem = r(sd)/r(N)^.5
local tval = (r(mean)-55)/`sem'
display "t-value = " `tval'
display "P-value = " 2*(ttail(r(N)-1,abs(`tval')))
```

```
t-value = 1.063716
P-value = .30324854
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

One sample t-test

```
ttest pre=55
```

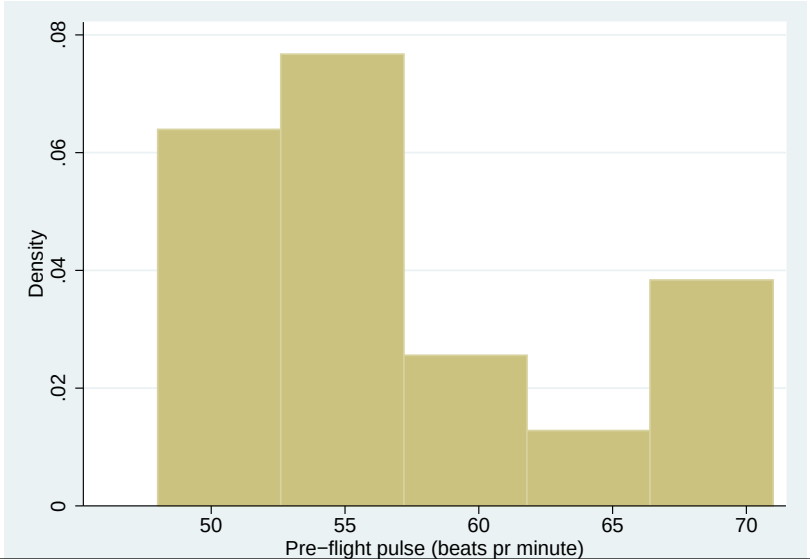
```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |      17    56.88235    1.769601    7.296252    53.13097    60.63374
------------------------------------------------------------------------------
    mean = mean(pre)                                              t =   1.0637
Ho: mean = 55                                    degrees of freedom =       16

    Ha: mean < 55               Ha: mean != 55                 Ha: mean > 55
 Pr(T < t) = 0.8484         Pr(|T| > |t|) = 0.3032          Pr(T > t) = 0.1516
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

One sample t-test

Normality reasonable?

```
hist pre, bin(5)
```



Non-parametric tests, sign-test

Sign test

We can use the sign test to instead formulate our test in terms of the median (without any distributional assumptions).
Two-sided hypothesis

$$H_0\colon \text{median} = m_0, \quad H_A : \text{median} \neq m_0$$

Simply count the number of observations larger than the null median and use this in a binomial test

```
signtest pre=55
```

```
Sign test

        sign |    observed    expected
-------------+------------------------
    positive |           8           8
    negative |           8           8
        zero |           1           1
-------------+------------------------
         all |          17          17
```


Sign-test

```
One-sided tests:
  Ho: median of pre - 55 = 0 vs.
  Ha: median of pre - 55 > 0
      Pr(#positive >= 8) =
         Binomial(n = 16, x >= 8, p = 0.5) =  0.5982

  Ho: median of pre - 55 = 0 vs.
  Ha: median of pre - 55 < 0
      Pr(#negative >= 8) =
         Binomial(n = 16, x >= 8, p = 0.5) =  0.5982

Two-sided test:
  Ho: median of pre - 55 = 0 vs.
  Ha: median of pre - 55 != 0
      Pr(#positive >= 8 or #negative >= 8) =
         min(1, 2*Binomial(n = 16, x >= 8, p = 0.5)) =  1.0000
```


Wilcoxon signed-rank test

Wilcoxon signed-rank test

If we further assume symmetry Wilcoxon signed rank test provides a more powerful test

$$H_0\colon \text{distribution is symmetric around } m_0$$

Rank the observations minus m_0 and check if the ranks of the negative and positive ranks is different.

pre-55:

16 10 -3 13 14 -6 -6 2 -4 0 3 2 4 -2 -2 -2 -7

rank(pre-55):

17 14 5 15 16 2.5 2.5 10.5 4 9 12 10.5 13 7 7 7 1

Wilcoxon signed-rank test

```
signrank pre=55
```

```
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        8          87          76
    negative |        8          65          76
        zero |        1           1           1
-------------+---------------------------------
         all |       17         153         153

unadjusted variance      446.25
adjustment for ties       -2.88
adjustment for zeros      -0.25
                     ----------
adjusted variance        443.12

Ho: pre = 55
             z =   0.523
    Prob > |z| =   0.6013
```

Paired tests, parametric t-test

The primary usage for the one-sample test is in the paired situation. In this situation we cannot use two-sample (independent) test, but must analyze the difference scores!

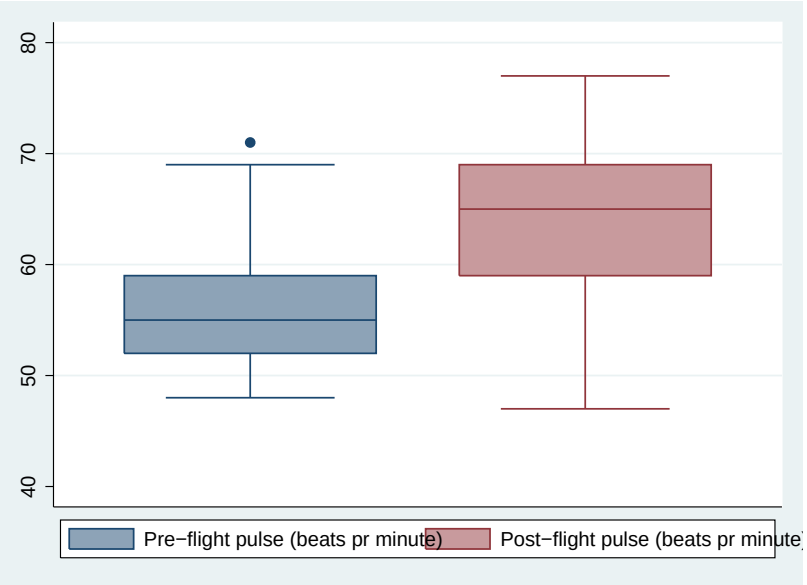
In stata you do not need to calculate the difference but use this syntax for the paired t-test:

```
ttest pre=post
```

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |      17    56.88235    1.769601    7.296252    53.13097    60.63374
    post |      17    63.76471    2.148066    8.856702    59.21101     68.3184
---------+--------------------------------------------------------------------
    diff |      17   -6.882353    2.595078    10.69978   -12.38367   -1.381034
------------------------------------------------------------------------------
     mean(diff) = mean(pre - post)                                 t =  -2.6521
 Ho: mean(diff) = 0                              degrees of freedom =       16

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0087         Pr(|T| > |t|) = 0.0174          Pr(T > t) = 0.9913
```


Paired tests

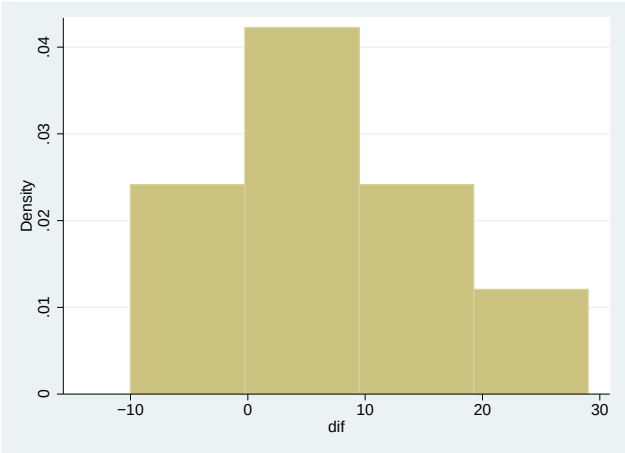
```
graph box pre post
```



Non-parametric tests for paired data

Generally with just a shift in location between post and pre observation we expect symmetrically distributed differences

```
gen dif=post-pre
hist dif
```



Non-parametric tests for paired data

Makes the Wilcoxon-test a good choice (same syntax for the sign-test)

```
signrank pre=post
```

```
Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |        4          29          76
    negative |       12         123          76
        zero |        1           1           1
-------------+---------------------------------
         all |       17         153         153

unadjusted variance      446.25
adjustment for ties       -0.38
adjustment for zeros      -0.25
                     ----------
adjusted variance        445.62

Ho: pre = post
             z =  -2.226
    Prob > |z| =   0.0260
```

Two sample tests

Comparison of two groups, Two-sample t-test

- Assume independent observations within and between groups
- Observations approximately normal distributed within each group (again some robustness)
- Equal variances (can be relaxed easily in stata)

Formally,

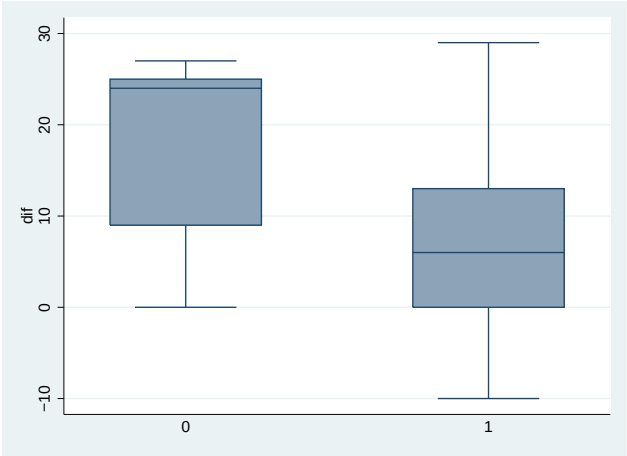
$$Y_{1i} = \mu_1 + \epsilon_{1i}, \quad i = 1, \ldots, n_1$$
$$Y_{2i} = \mu_2 + \epsilon_{2i}, \quad i = 1, \ldots, n_2$$

where independent $\epsilon_{1i}, \epsilon_{2i} \sim \mathcal{N}(0, \sigma^2)$

$$H_0 \colon \mu_1 = \mu_2, \quad H_A \colon \mu_1 \neq \mu_2$$

Two-sample t-test

```
use data/astronaut, clear
gen dif=post-pre
graph box dif, over(salt)
```



Two-sample t-test

Grouping via `by` option and `level` to change CI level

```
ttest dif, by(salt) level(90)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       9    17.44444    3.371083    10.11325    11.17575    23.71313
       1 |      17    6.882353    2.595078    10.69978    2.351649    11.41306
---------+--------------------------------------------------------------------
combined |      26    10.53846    2.255408    11.50037    6.685908    14.39102
---------+--------------------------------------------------------------------
    diff |            10.56209    4.331688                3.151084     17.9731
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.4383
Ho: diff = 0                                     degrees of freedom =       24

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9887         Pr(|T| > |t|) = 0.0225          Pr(T > t) = 0.0113
```


Two-sample comparisons, variance

Not obvious from box-plot that variance homogeniety is fulfilled. Assume instead

$$Y_{1i} = \mu_1 + \epsilon_{1i}, \quad i = 1, \ldots, n_1$$
$$Y_{2i} = \mu_2 + \epsilon_{2i}, \quad i = 1, \ldots, n_2$$

where independent $\epsilon_{1i}, \epsilon_{2i} \sim \mathcal{N}(0, \sigma_i^2)$. We will test

$$H_0 \colon \sigma_1 = \sigma_2, \quad H_A \colon \sigma_1 \neq \sigma_2$$

in some situtations this may even be a primary hypothesis. . .

In stata: `sdtest`, `robvar`

K-sample comparisons, variance

To compare the variance in K groups via Bartlett's test

```
sdtest dif, by(salt)
```

```
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       9    17.44444    3.371083    10.11325    9.670714    25.21817
       1 |      17    6.882353    2.595078    10.69978    1.381034    12.38367
---------+--------------------------------------------------------------------
combined |      26    10.53846    2.255408    11.50037    5.893362    15.18356
------------------------------------------------------------------------------
    ratio = sd(0) / sd(1)                                         f =   0.8934
Ho: ratio = 1                                   degrees of freedom =    8, 16

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.4561         2*Pr(F < f) = 0.9123           Pr(F > f) = 0.5439
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

K-sample comparisons, variance

Robustness to normality assumptions via *Levene's test* or *Brown-Forsythe*

```
robvar dif, by(salt)
```

```
    Control |
  group: 0, |           Summary of dif
    Salt: 1 |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   17.444444   10.113248           9
          1 |   6.8823529    10.69978          17
------------+------------------------------------
      Total |   10.538462   11.500368          26

W0  =  0.00422467   df(1, 24)     Pr > F = 0.94871439

W50 =  0.03418276   df(1, 24)     Pr > F = 0.8548721

W10 =  0.00548156   df(1, 24)     Pr > F = 0.94159416
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

T-test

While no evidence against variance-homogeniety we may a priori want a test that is robust to this assumption:

```
ttest dif, by(salt) unequal
```

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       9    17.44444    3.371083    10.11325    9.670714    25.21817
       1 |      17    6.882353    2.595078    10.69978    1.381034    12.38367
---------+--------------------------------------------------------------------
combined |      26    10.53846    2.255408    11.50037    5.893362    15.18356
---------+--------------------------------------------------------------------
    diff |            10.56209    4.254248                1.596713    19.52747
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.4827
Ho: diff = 0                     Satterthwaite's degrees of freedom =  17.2603

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9882         Pr(|T| > |t|) = 0.0236          Pr(T > t) = 0.0118
```

Very little loss in power so probably always preferable! Also `welch` option (default in R).

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

Comparing several groups

Parametric model for comparison of several groups: `oneway`, `anova` (but we will prefer to use the linear model framework `regress`...)

```
use iris, clear
oneway Sepal_Length Species
```

```
(Edgar Anderson's Iris Data)

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      63.2121333      2   31.6060667    119.26     0.0000
 Within groups         38.9562    147   .265008163
------------------------------------------------------------------------
    Total           102.168333    149   .685693512

Bartlett's test for equal variances:  chi2(2) =  16.0057  Prob>chi2 = 0.000
```

Comparing several groups

```
regress
```

```
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    3.69
       Model |  .434593287     1  .434593287           Prob > F      =  0.0607
    Residual |  5.65360671    48  .117783473           R-squared     =  0.0714
-------------+------------------------------           Adj R-squared =  0.0520
       Total |      6.0882    49   .12424898           Root MSE      =   .3432

------------------------------------------------------------------------------
Sepal_Length |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Petal_Length |   .5422926   .2823153     1.92   0.061    -.0253403    1.109925
       _cons |   4.213168   .4155888    10.14   0.000     3.377571    5.048765
------------------------------------------------------------------------------
```


Comparing several groups

Non-parametric comparison via Kruskal-Wallis test (null-hypothesis that the rank across all groups is the same)

```
kwallis Sepal_Length, by(Species)
```

```
Kruskal-Wallis equality-of-populations rank test

  +-----------------------------+
  |    Species | Obs | Rank Sum |
  |------------+-----+----------|
  |     setosa |  50 |  1482.00 |
  | versicolor |  50 |  4132.50 |
  |  virginica |  50 |  5710.50 |
  +-----------------------------+

chi-squared =    96.761 with 2 d.f.
probability =     0.0001

chi-squared with ties =    96.937 with 2 d.f.
probability =     0.0001
```


Comparing two continuous variables

Rank-correlation (Spearman's ρ `spearman` or Kendall's τ `ktau`), e.g.

```
ktau Petal_Length Sepal_Length
```

```
  Number of obs =     150
Kendall's tau-a =       0.6949
Kendall's tau-b =       0.7185
Kendall's score =    7765
    SE of score =     614.002   (corrected for ties)

Test of Ho: Petal_Length and Sepal_Length are independent
     Prob > |z| =       0.0000  (continuity corrected)
```

```
spearman Petal_Length Sepal_Length
```

```
 Number of obs =     150
Spearman's rho =       0.8819

Test of Ho: Petal_Length and Sepal_Length are independent
    Prob > |t| =       0.0000
```


Comparing two continuous variables

or Pearson correlation (linear correlation)

```
pwcorr Petal_Length Sepal_Length Sepal_Width, star
    (0.05)
```

```
             | Peta~gth Sepa~gth Sepa~dth
-------------+---------------------------
Petal_Length |   1.0000
Sepal_Length |   0.8718*  1.0000
 Sepal_Width |  -0.4284* -0.1176   1.0000
```

with (Sidak) correction for multiple comparisons

```
pwcorr Petal_Length Sepal_Length Sepal_Width, sig sidak
```

```
             | Peta~gth Sepa~gth Sepa~dth
-------------+---------------------------
Petal_Length |   1.0000
             |
             |
Sepal_Length |   0.8718   1.0000
             |   0.0000
             |
 Sepal_Width |  -0.4284  -0.1176   1.0000
```

Comparing two continuous variables

```
use data/hubble, clear
describe
```

```
(Velocity and distance measures of 36 Type Ia super-novae, Hubble Telescope)

Contains data from data/hubble.dta
  obs:            36                          Velocity and distance measures
                                                of 36 Type Ia super-novae,
                                                Hubble Telescope
 vars:             3                          22 Sep 2014 22:27
 size:           864
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
v               double  %9.0g                 v
D               double  %9.0g                 D
sigma           double  %9.0g                 sigma
-------------------------------------------------------------------------------
Sorted by:
```

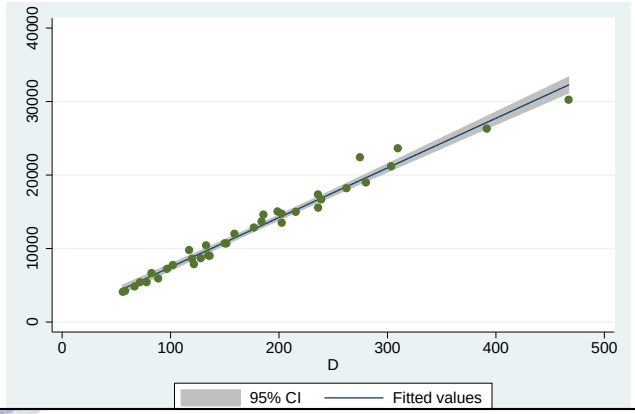

Linear regression

Hubbles law relates velocity and distance from earth

$$v = H_0 D$$

with Hubble constant H_0, with H_0^{-1} approximately equal to the age of the universe (unit 3.085×10^{19} seconds).

```
twoway (lfitci v D) (scatter v D)
```



Linear regression

Linear regression

$$Y = \beta_0 + \beta_X X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

outcome Y and covariate X. Assumptions

- Linearity
- Independence
- Variance homogeneity
- Approximate normal residuals

But no assumptions regarding X.

Stata syntax

```
regress outcome predictor
```


Linear regression

```
regress v D
```

```
      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  1,    34) = 1531.24
       Model |  1.4383e+09     1  1.4383e+09           Prob > F      =  0.0000
    Residual |  31935910.7    34   939291.49           R-squared     =  0.9783
-------------+------------------------------           Adj R-squared =  0.9776
       Total |  1.4702e+09    35  42006068.5           Root MSE      =  969.17

------------------------------------------------------------------------------
           v |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   67.53619     1.7259    39.13   0.000     64.02874    71.04364
       _cons |   711.7957   347.3545     2.05   0.048     5.886355    1417.705
------------------------------------------------------------------------------
```

Linear regression

In this applications it is better to omit the intercept

```
regress v D, noconst
```

```
      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  1,    35) = 7103.41
       Model |  7.2820e+09     1  7.2820e+09           Prob > F      =  0.0000
    Residual |  35880167.3    35  1025147.64           R-squared     =  0.9951
-------------+------------------------------           Adj R-squared =  0.9950
       Total |  7.3179e+09    36   203275674           Root MSE      =  1012.5

------------------------------------------------------------------------------
           v |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   70.66722   .8384642    84.28   0.000     68.96504    72.36939
------------------------------------------------------------------------------
```

```
di "Age of the universe: " (1/_b[D]*3.085e19)
    /(365*60*60*24) " years"
```

```
Age of the universe: 1.384e+10 years
```


Linear regression

X can also be a dummy-variable (values 0 and 1) giving us another alternative way of specifying t-tests. In general

General Linear Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

with k covariates $X_1, \ldots, X_n$, and otherwise same assumptions as in the simple linear regression case

Stata syntax

```
regress y x1 x2 x3
```

- `i.` factor variable
- `b.` base level
- `c.` continuous variable

Linear regression, ANCOVA

Special case: one continuous and one categorical predictor.

A more powerful test in the astronaut exampleq (assuming independence between the group and baseline) obtained by regressing on baseline

```
use data/astronaut, clear
regress post pre salt
```

```
      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  2,    23) =    5.23
       Model |  1034.56337     2  517.281684           Prob > F      =  0.0134
    Residual |  2273.89817    23  98.8651379           R-squared     =  0.3127
-------------+------------------------------           Adj R-squared =  0.2529
       Total |  3308.46154    25  132.338462           Root MSE      =  9.9431

------------------------------------------------------------------------------
        post |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pre |   .4856021   .2637398     1.84   0.079    -.0599852    1.031189
        salt |  -10.73692   4.099835    -2.62   0.015    -19.21807   -2.255765
       _cons |   46.87944   15.45143     3.03   0.006     14.91572    78.84316
------------------------------------------------------------------------------
```


Linear regression

```
ereturn list
```

```
scalars:
                  e(N) =  26
               e(df_m) =  2
               e(df_r) =  23
                  e(F) =  5.232195039475376
                 e(r2) =  .312702250190805
               e(rmse) =  9.94309498381509
                e(mss) =  1034.563367746656
                e(rss) =  2273.898170714883
               e(r2_a) =  .2529372284682663
                 e(ll) =  -95.01740848453022
              e(ll_0) =  -99.89224825715341
              e(rank) =  3

macros:
            e(cmdline) : "regress post pre salt"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
              e(depvar) : "post"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "regress_estat"

matrices:
                  e(b) :  1 x 3
                  e(V) :  3 x 3
```

Factors, categorical predictors

Follow-up study on survival after Acute Myocardial Infarction (AMI)

```
use data/ami, clear
recast byte agecat
label define agecatlabel 0 "-65" 1 "65-75" 2 "75-"
label values agecat agecatlabel
describe
```

```
recast byte agecat

Contains data from data/ami.dta
  obs:         1,878
 vars:            11                          22 Sep 2014 22:48
 size:       107,046
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
obsnr           double  %9.0g                 obsnr
wmi             double  %9.0g                 The hearts ability to pump
status          long    %9.0g      status     status
chf             long    %9.0g      chf        Clinical heart pump failure
```


Factors

We will examine association to between WMI and age defined by the categorical variable agecat

```
regress wmi agecat
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  1,  1876) =   43.60
       Model |  7.25574826     1  7.25574826           Prob > F      =  0.0000
    Residual |  312.176563  1876  .166405417           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0222
       Total |  319.432311  1877  .170182371           Root MSE      =  .40793

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |  -.0772695   .0117018    -6.60   0.000    -.1002193   -.0543197
       _cons |   1.465042   .0138478   105.80   0.000     1.437884    1.492201
------------------------------------------------------------------------------
```


Factors

We need to specify that agecat should be treated as a factor. We do this with the `i.` prefix, and obtain the oneway ANOVA:

```
regress wmi i.agecat
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  2,  1875) =   21.79
       Model |  7.25577095     2  3.62788548           Prob > F      =  0.0000
    Residual |   312.17654  1875  .166494155           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0217
       Total |  319.432311  1877  .170182371           Root MSE      =  .40804

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      65-75  |  -.0774879    .022065    -3.51   0.000    -.1207624   -.0342135
        75-  |   -.154507   .0235706    -6.56   0.000    -.2007344   -.1082795
             |
       _cons |   1.465107   .0149193    98.20   0.000     1.435847    1.494367
------------------------------------------------------------------------------
```

We may also create a single indicator variable for example with the syntax `i(2 3).agecat` (1 when age>65).

Factors

... and the reference can be chosen with the `b.` prefix:

```
regress wmi ib(2).agecat
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  2,  1875) =   21.79
       Model |  7.25577095     2  3.62788548           Prob > F      =  0.0000
    Residual |   312.17654  1875  .166494155           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0217
       Total |  319.432311  1877  .170182371           Root MSE      =  .40804

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
        -65  |   .154507   .0235706     6.56   0.000     .1082795    .2007344
      65-75  |   .077019    .024439     3.15   0.002     .0290885    .1249495
             |
       _cons |    1.3106    .018248    71.82   0.000     1.274812    1.346388
------------------------------------------------------------------------------
```

Twoway ANOVA

To also adjust for sex we simply add sex to the list of covariates (note i. not strictly necessary for sex but convenient wrt output and postestimation):

```
regress wmi i.agecat i.sex
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  3,  1874) =   19.08
       Model |  9.46723213     3  3.15574404           Prob > F      =  0.0000
    Residual |  309.965079  1874  .165402924           R-squared     =  0.0296
-------------+------------------------------           Adj R-squared =  0.0281
       Total |  319.432311  1877  .170182371           Root MSE      =   .4067

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      65-75  |  -.0857886   .0221094    -3.88   0.000    -.1291502    -.042427
        75-  |  -.1751729   .0241635    -7.25   0.000    -.2225631   -.1277826
             |
         sex |
     female  |  -.0766893   .0209733    -3.66   0.000    -.1178228   -.0355558
       _cons |   1.526725   .0224745    67.93   0.000     1.482647    1.570803
------------------------------------------------------------------------------
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

Post-estimation

Stata has built in a number of post-estimation routines which we can call on the last regress (or other model) in memory.
We can use test (and testparm or contrast) to test the overall significance of agecat

```
test 1.agecat 2.agecat
```

```
( 1)  1.agecat = 0
( 2)  2.agecat = 0

       F(  2,  1874) =   26.59
            Prob > F =    0.0000
```

And lincom for computing linear combinations

```
lincom 1.agecat-2.agecat
```

```
 ( 1)  1.agecat - 2.agecat = 0

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

Margins

Obtaining estimates of expected WMI

- The WMI for the reference group (male less than 65 years) can be read off from the intercept (_cons).
- To get the estimated average WMI in the other groups we could chance reference with .b
- Post-estimation with lincom

```
lincom _cons + 1.sex
```

```
 ( 1)  1b.sex + _cons = 0

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.526725   .0224745    67.93   0.000     1.482647    1.570803
------------------------------------------------------------------------------
```

- Or use the margins command

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

Margins

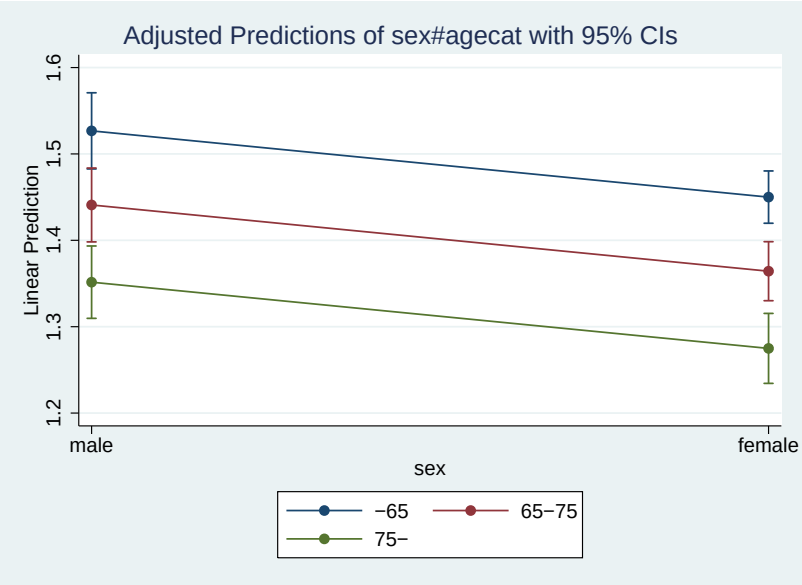
Here we use stata's interaction syntax

```
margins i.sex#i.agecat
```

```
Adjusted predictions                            Number of obs   =       1878
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  sex#agecat |
   male#-65  |   1.526725   .0224745    67.93   0.000     1.482647    1.570803
 male#65-75  |   1.440936   .0217982    66.10   0.000     1.398185    1.483688
   male#75-  |   1.351552   .0213598    63.28   0.000     1.309661    1.393444
 female#-65  |   1.450036    .015431    93.97   0.000     1.419772    1.480299
     female #|
      65-75  |   1.364247   .0174184    78.32   0.000     1.330086    1.398409
 female#75-  |   1.274863   .0206477    61.74   0.000     1.234368    1.315358
------------------------------------------------------------------------------
```

Margins

```
marginsplot
```



Interactions

Interactions can of course be formed by generating relevant multiplications with the `generate` command. But much easier with the `#` operator

Interactions in stata

Main effects with two categorical predictors

```
regress y i.x i.z
```

Main effects and interaction (full factorial)

```
regress y i.x i.z i.x#i.z
```

Interaction between continuous and categorical predictors

```
regress y c.x i.z c.x#i.z
```

- `x##z` expands to `x z x#z`
- `x##z##v` expands to `x v z x#v x#z v#z x#v#z`

Interactions

```
regress wmi i.agecat##i.sex
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  5,  1872) =   11.75
       Model |  9.72300472     5  1.94460094           Prob > F      =  0.0000
    Residual |  309.709306  1872  .165443005           R-squared     =  0.0304
-------------+------------------------------           Adj R-squared =  0.0278
       Total |  319.432311  1877  .170182371           Root MSE      =  .40675

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |
      65-75  |  -.0889243   .0445774    -1.99   0.046     -.176351   -.0014977
        75-  |  -.1437214    .042843    -3.35   0.001    -.2277463   -.0596964
             |
         sex |
     female  |  -.0620542   .0374265    -1.66   0.097    -.1354562    .0113479
             |
  agecat#sex |
      65-75 #|
     female  |   .0067888   .0513824     0.13   0.895     -.093984    .1075615
  75-#female |  -.0515126   .0522536    -0.99   0.324    -.1539939    .0509688
             |
```

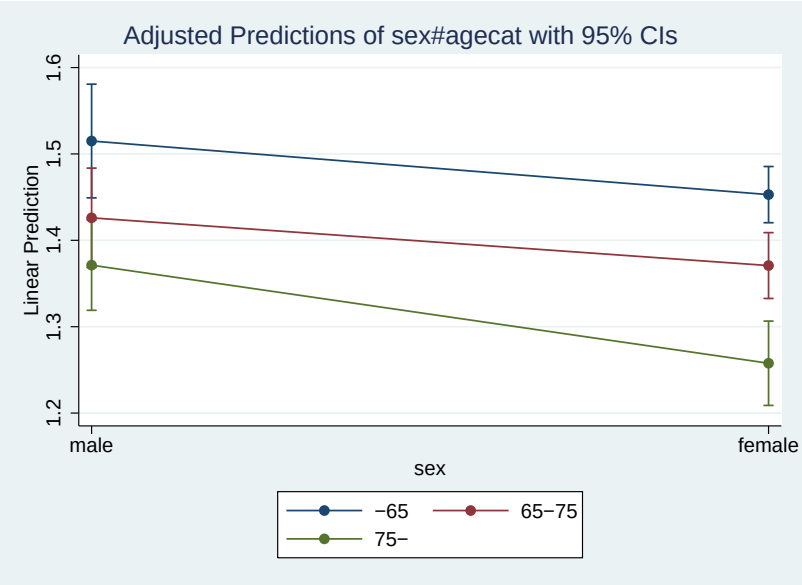

Margins

```
margins sex#agecat
```

```
Adjusted predictions                            Number of obs   =       1878
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  sex#agecat |
   male#-65  |   1.514966   .0335479    45.16   0.000     1.449171    1.580761
 male#65-75  |   1.426042   .0293544    48.58   0.000     1.368471    1.483613
   male#75-  |   1.371245   .0266469    51.46   0.000     1.318984    1.423505
 female#-65  |   1.452912   .0165916    87.57   0.000     1.420372    1.485452
     female #|
      65-75  |   1.370776   .0194351    70.53   0.000     1.332659    1.408893
 female#75-  |   1.257678   .0248925    50.52   0.000     1.208858    1.306498
------------------------------------------------------------------------------
```

Margins

```
marginsplot
```



Interactions

Test for interaction

```
testparm i.agecat#i.sex
```

```
( 1)  1.agecat#2.sex = 0
( 2)  2.agecat#2.sex = 0

       F(  2,  1872) =    0.77
            Prob > F =    0.4618
```

Or with the newer `contrast` function

```
contrast agecat#sex
```

```
Contrasts of marginal linear predictions

Margins      : asbalanced

-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
  agecat#sex |          2        0.77     0.4618
             |
 Denominator |       1872
```


Interactions with a continuous variable

To use continuous variables in the interaction terms we much use the `.c` prefix

```
regress wmi c.age##i.sex
```

```
      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  3,  1874) =   26.30
       Model |  12.9035584     3  4.30118613           Prob > F      =  0.0000
    Residual |  306.528753  1874  .163569238           R-squared     =  0.0404
-------------+------------------------------           Adj R-squared =  0.0389
       Total |  319.432311  1877  .170182371           Root MSE      =  .40444

------------------------------------------------------------------------------
         wmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0050087   .0016303    -3.07   0.002    -.0082062   -.0018113
             |
         sex |
     female  |   .1193504   .1340474     0.89   0.373    -.1435474    .3822482
             |
   sex#c.age |
     female  |  -.0029252    .001903    -1.54   0.124    -.0066574     .000807
             |
       _cons |   1.783095   .1172715    15.20   0.000     1.553098    2.013091
```


Predictions, margins

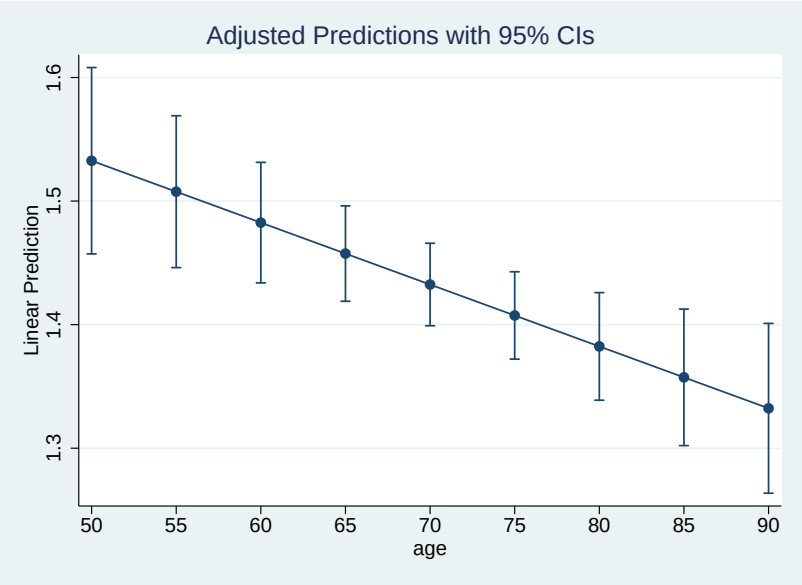
Predicting the age effect in males

```
margins, at(age=(50(5)90) sex=1)
```

```
...
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   1.532658   .0384483    39.86   0.000     1.457252    1.608064
          2  |   1.507614   .0313333    48.12   0.000     1.446162    1.569066
          3  |    1.48257   .0248631    59.63   0.000     1.433808    1.531333
          4  |   1.457527   .0196842    74.05   0.000     1.418921    1.496132
          5  |   1.432483   .0170194    84.17   0.000     1.399104    1.465862
          6  |   1.407439   .0180208    78.10   0.000     1.372096    1.442782
          7  |   1.382396   .0221976    62.28   0.000     1.338861     1.42593
          8  |   1.357352   .0281712    48.18   0.000     1.302102    1.412602
          9  |   1.332308   .0350342    38.03   0.000     1.263598    1.401018
------------------------------------------------------------------------------
```

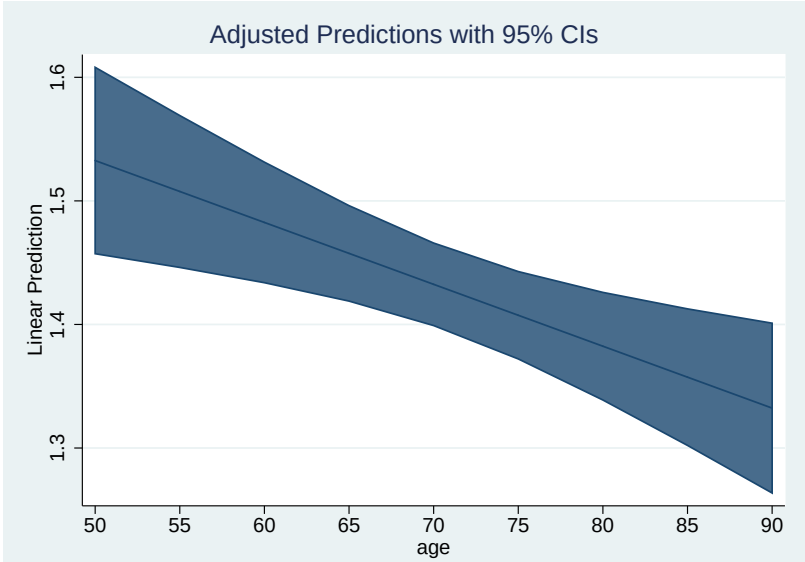
Predictions, margins

```
marginsplot
```



Predictions, margins

```
margins, at(age=(50(5)90) sex=1)
marginsplot, recastci(rarea) plotopts(msymbol(i))
```



Predictions, margins

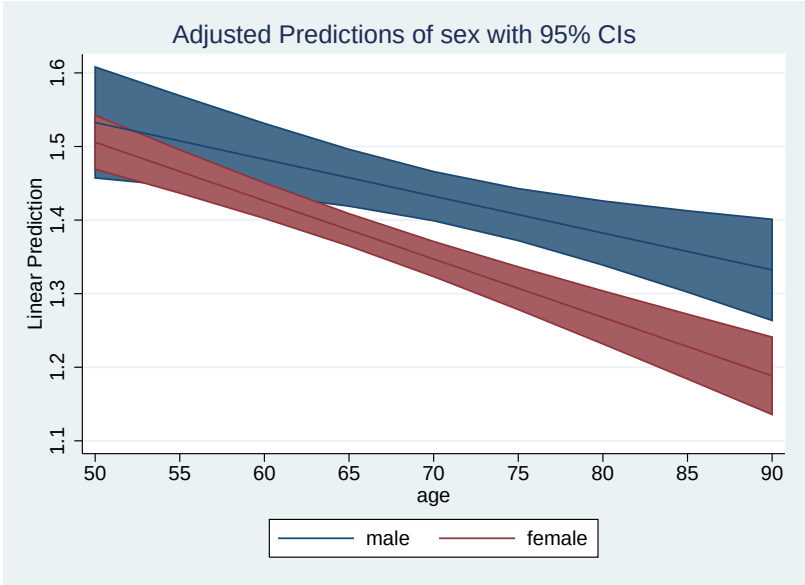
Predicting the age effect in males and females

```
/* margins, at(age=(50(5)90) sex=(1 2)) */
margins sex, at(age=(50(5)90))
```

```
...
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _at#sex |
     1#male  |   1.532658   .0384483    39.86   0.000     1.457252    1.608064
   1#female  |   1.505748   .0186208    80.86   0.000     1.469229    1.542268
     2#male  |   1.507614   .0313333    48.12   0.000     1.446162    1.569066
   2#female  |   1.466079   .0149913    97.80   0.000     1.436677     1.49548
     3#male  |    1.48257   .0248631    59.63   0.000     1.433808    1.531333
   3#female  |   1.426409   .0122849   116.11   0.000     1.402315    1.450503
     4#male  |   1.457527   .0196842    74.05   0.000     1.418921    1.496132
   4#female  |   1.386739   .0111924   123.90   0.000     1.364788     1.40869
     5#male  |   1.432483   .0170194    84.17   0.000     1.399104    1.465862
   5#female  |    1.34707    .012157   110.81   0.000     1.323227    1.370912
     6#male  |   1.407439   .0180208    78.10   0.000     1.372096    1.442782
   6#female  |     1.3074   .0147813    88.45   0.000     1.278411    1.336389
     7#male  |   1.382396   .0221976    62.28   0.000     1.338861     1.42593
   7#female  |    1.26773   .0183671    69.02   0.000     1.231708    1.303753
```


Predictions, margins

```
marginsplot, recastci(rarea) plotopts(msymbol(i))
```



Predictions, margins

Example on quadratic effect via interaction operator

```
use iris,clear
regress Sepal_Length Petal_Length c.Petal_Length#c.
    Petal_Length
```

```
(Edgar Anderson's Iris Data)

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  2,   147) =  312.27
       Model |  82.7025583     2  41.3512792           Prob > F      =  0.0000
    Residual |   19.465775   147  .132420238           R-squared     =  0.8095
-------------+------------------------------           Adj R-squared =  0.8069
       Total |  102.168333   149  .685693512           Root MSE      =   .3639

------------------------------------------------------------------------------
Sepal_Length |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Petal_Length |  -.1643507   .0942709    -1.74   0.083    -.3506521    .0219506
             |
           c. |
Petal_Length#|
           c. |
Petal_Length |    .081463   .0131794     6.18   0.000     .0554175    .1075085
```


Predictions, predict

Instead of the margins command we will use the post-estimation function predict

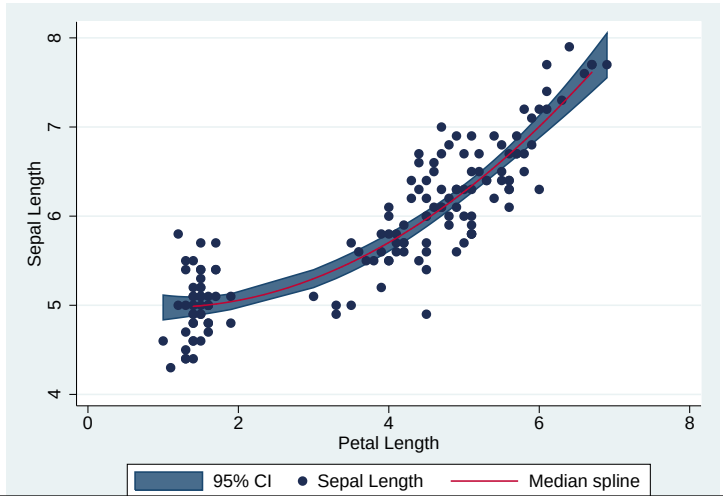
```
sort Petal_Length
capture drop yhat*
predict yhat
predict yhat_se, stdp
capture gen yhatLo = yhat-2*yhat_se
capture gen yhatHi = yhat+2*yhat_se
```

For convenience we could put this in a simple function

```
capture program drop mypredict
program mypredict
   capture drop yhat*
   predict yhat
   predict yhat_se, stdp
   capture gen yhatLo = yhat-2*yhat_se
   capture gen yhatHi = yhat+2*yhat_se
end
```


Predictions, predict

```
twoway (rarea yhatLo yhatHi Petal_Length) (scatter
    Sepal_Length Petal_Length, mcolor(dknavy)) (mspline
     yhat Petal_Length, lcolor(cranberry)), legend(lab
    (1 "95% CI") row(1)) ytitle("Sepal Length")
```



Prediction, out-of-sample

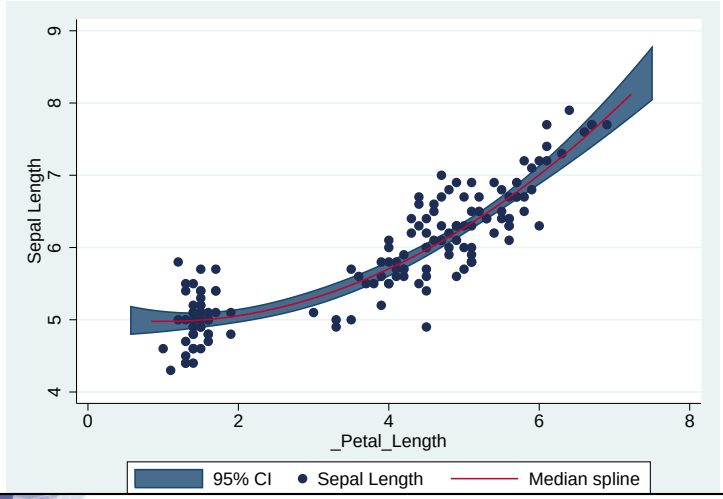
We can also make *out-of-sample* prediction by creating a new dataset

```
clear
set obs 100
gen Petal_Length = 7*(_n)/_N+0.5
gen _Petal_Length = Petal_Length
mypredict
drop Petal_Length
tempfile _tmpdata
save '_tmpdata', replace
use iris, clear
merge 1:1 _n using '_tmpdata'
```

```
obs was 0, now 100
(option xb assumed; fitted values)
 t found)
file /var/folders/f1/6z6dzblx5xjgcbc1_sbksgf00000gn/T//St16314.000008 saved
(Edgar Anderson's Iris Data)

    Result                           # of obs.
    -------------------------------------------
```

Prediction, out-of-sample

```
twoway (rarea yhatLo yhatHi _Petal_Length) (scatter
    Sepal_Length Petal_Length, mcolor(dknavy)) (mspline
     yhat _Petal_Length, lcolor(cranberry)), legend(lab
    (1 "95% CI") row(1)) ytitle("Sepal Length")
```



Categorical data, a very brief look

```
use data/ami, clear
gen CHF=chf-1
```

```
tabulate agecat chf, chi2 exact gamma
```

```
Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 253
stage 1:  enumerations = 0

           |  Clinical heart pump
           |        failure
    agecat |    absent    present |     Total
-----------+----------------------+----------
         0 |       485        263 |       748
         1 |       275        355 |       630
         2 |       136        364 |       500
-----------+----------------------+----------
     Total |       896        982 |     1,878

          Pearson chi2(2) = 176.4462   Pr = 0.000
                    gamma =   0.4832  ASE = 0.031
           Fisher's exact =                 0.000
```


Logistic regression

Same syntax and post-estimation features as for `regress`

```
logistic CHF i.vf age i.sex
```

```
Logistic regression                               Number of obs   =       1878
                                                  LR chi2(3)      =     279.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -1160.2279                       Pseudo R2       =     0.1074

------------------------------------------------------------------------------
         CHF | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          vf |
     present |   3.462275   .7442115     5.78   0.000     2.271941    5.276261
         age |   1.073125    .005394    14.04   0.000     1.062605    1.083749
             |
         sex |
      female |   .9863929    .109582    -0.12   0.902     .7933902    1.226346
       _cons |   .0089841   .0033042   -12.81   0.000     .0043694    .0184726
------------------------------------------------------------------------------
```

Note that we directly obtain OR-estimates. Conditional logistic regression via `clogit`.

Poisson regression

And poisson regression. Here we examine counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group

```
insheet using data/eba1977.csv, delimit(,) clear
encode city, gen(City)
encode age, gen(Age)
describe
```

```
(5 vars, 24 obs)

Contains data
  obs:            24
 vars:             7
 size:           648
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
v1              byte    %8.0g
city            str10   %10s
age             str5    %9s
pop             int     %8.0g
```

Poisson regression

exposure adds the log-offset and eform transforms to the Rate-Ratio scale

```
glm cases i.City i.Age, exposure(pop) family(poisson)
    eform
```

```
Iteration 0:   log likelihood = -59.987835
Iteration 1:   log likelihood = -59.917787
Iteration 2:   log likelihood = -59.917758
Iteration 3:   log likelihood = -59.917758

Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         15
                                                  Scale parameter =          1
Deviance         =  23.44747817                   (1/df) Deviance =   1.563165
Pearson          =  22.56163134                   (1/df) Pearson  =   1.504109

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   5.743146
Log likelihood   = -59.91775758                   BIC             =  -24.22333
```

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN

Poisson regression

```
------------------------------------------------------------------------------
             |                 OIM
       cases |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        City |
    Horsens  |   .7188806   .1304792    -1.82   0.069     .5036871    1.026013
    Kolding  |   .6896672   .1295238    -1.98   0.048     .4772859    .9965533
      Vejle  |   .7616123   .1430714    -1.45   0.147      .527027    1.100614
             |
         Age |
      55-59  |   3.007214   .7466484     4.43   0.000     1.848515    4.892215
      60-64  |   4.565885    1.05763     6.56   0.000     2.899711    7.189442
      65-69  |   5.857403   1.343919     7.70   0.000     3.735991    9.183417
      70-74  |   6.403619   1.506919     7.89   0.000     4.037553    10.15624
        75+  |   4.135687   1.035041     5.67   0.000      2.53231     6.75427
             |
       _cons |   .0035812   .0007171   -28.12   0.000     .0024186    .0053025
     ln(pop) |          1  (exposure)
------------------------------------------------------------------------------
```