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

\[ H_0: \mu = \mu_0, \quad H_A: \mu \neq \mu_0 \]

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

\[
\text{browse, summarize, graph histogram, graph qnorm, graph box, ...}
\]
One sample t-test

We will just consider the salt treated group.

```stata
1 save data/astronaut, replace
drop if salt==0
and examine if the pre-flight pulse (population) mean could be 60 beats pr minute
1 summarize pre
2 return list
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre</td>
<td>17</td>
<td>56.88235</td>
<td>7.296252</td>
<td>48</td>
<td>71</td>
</tr>
</tbody>
</table>

scalars:
- \( r(N) = 17 \)
- \( r(\text{sum}_w) = 17 \)
- \( r(\text{mean}) = 56.88235294117647 \)
- \( r(\text{Var}) = 53.23529411764706 \)
- \( r(\text{sd}) = 7.296252059629454 \)
- \( r(\text{min}) = 48 \)
- \( r(\text{max}) = 71 \)
- \( r(\text{sum}) = 967 \)

```stata
1 local sem = r(sd)/r(N)^.5
display "t-value = " 'sem'
2 local tval = (r(mean)-55)/'sem'
display "P-value = " 2*(ttail(r(N)-1,abs('tval')))
t-value = 1.063716
P-value = .30324854
```

Normality reasonable?

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

One-sample t test

```stata
1 ttest pre=55
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre</td>
<td>17</td>
<td>56.88235</td>
<td>1.769601</td>
<td>7.296252</td>
<td>53.13097 60.63374</td>
</tr>
</tbody>
</table>

mean = mean(pre)  
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
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

\[ H_0: \text{median} = m_0, \quad H_A: \text{median} \neq m_0 \]

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

```
sign test
```

<table>
<thead>
<tr>
<th>sign</th>
<th>observed</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>negative</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>zero</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

Wilcoxon signed-rank test

Wilcoxon signed-rank test

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

\[ H_0: \text{distribution is symmetric around} \ m_0 \]

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 -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

One-sided tests:

\[ H_0: \text{median of pre - 55} = 0 \ vs. \ H_a: \text{median of pre - 55} > 0 \]

\[ \Pr(#\text{positive} \geq 8) = \text{Binomial}(n = 16, x \geq 8, p = 0.5) = 0.5982 \]

\[ H_0: \text{median of pre - 55} = 0 \ vs. \ H_a: \text{median of pre - 55} < 0 \]

\[ \Pr(#\text{negative} \geq 8) = \text{Binomial}(n = 16, x \geq 8, p = 0.5) = 0.5982 \]

Two-sided test:

\[ H_0: \text{median of pre - 55} = 0 \ vs. \ H_a: \text{median of pre - 55} \neq 0 \]

\[ \Pr(#\text{positive} \geq 8 \text{ or } #\text{negative} \geq 8) = \min(1, 2\times\text{Binomial}(n = 16, x \geq 8, p = 0.5)) = 1.0000 \]

Wilcoxon signed-rank test

```
signrank pre=55
```

<table>
<thead>
<tr>
<th>sign</th>
<th>obs</th>
<th>sum ranks</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>8</td>
<td>87</td>
<td>76</td>
</tr>
<tr>
<td>negative</td>
<td>8</td>
<td>65</td>
<td>76</td>
</tr>
<tr>
<td>zero</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all</td>
<td>17</td>
<td>153</td>
<td>153</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre</td>
<td>17</td>
<td>56.88235</td>
<td>1.769601</td>
<td>7.296252</td>
<td>53.13097  60.63374</td>
</tr>
<tr>
<td>post</td>
<td>17</td>
<td>63.76471</td>
<td>2.148066</td>
<td>8.856702</td>
<td>59.21101  68.3184</td>
</tr>
<tr>
<td>diff</td>
<td>17</td>
<td>-6.882353</td>
<td>2.595078</td>
<td>10.69978</td>
<td>-12.38367 -1.381034</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>sign</th>
<th>obs</th>
<th>sum ranks</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>4</td>
<td>29</td>
<td>76</td>
</tr>
<tr>
<td>negative</td>
<td>12</td>
<td>123</td>
<td>76</td>
</tr>
<tr>
<td>zero</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>all</td>
<td>17</td>
<td>153</td>
<td>153</td>
</tr>
</tbody>
</table>

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

\[ H_0: \mu_1 = \mu_2, \quad H_A: \mu_1 \neq \mu_2 \]

---

Two-sample t-test

Grouping via \texttt{by} option and \texttt{level} to change CI level

\begin{verbatim}
1 use data/astronaut, clear
2 gen dif=post-pre
3 graph box dif, over(salt)

dif
0 1
−10 0 10 20 30

dif
0 1

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

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[90% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>17.44444</td>
<td>3.371083</td>
<td>10.11325</td>
<td>11.17575 23.71313</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>6.882353</td>
<td>2.595078</td>
<td>10.69978</td>
<td>2.351649 11.41306</td>
</tr>
<tr>
<td>combined</td>
<td>26</td>
<td>10.53846</td>
<td>2.255408</td>
<td>11.50037</td>
<td>6.685908 14.39102</td>
</tr>
<tr>
<td>diff</td>
<td>10.56209</td>
<td>4.331688</td>
<td>3.151084</td>
<td>17.9731</td>
<td></td>
</tr>
</tbody>
</table>

diff = mean(0) - mean(1) \quad t = 2.4383
Ho: diff = 0 \quad degrees of freedom = 24

Ha: diff < 0 \quad Pr(T < t) = 0.9887
Ha: diff \neq 0 \quad Pr(|T| > |t|) = 0.0225
Ha: diff > 0 \quad 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 \( \epsilon_{1i}, \epsilon_{2i} \sim \mathcal{N}(0, \sigma_{1}^2) \). We will test

\[ H_0: \sigma_1 = \sigma_2, \quad H_A: \sigma_1 \neq \sigma_2 \]

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

In stata: \texttt{sdtest, robvar}
K-sample comparisons, variance

To compare the variance in $K$ groups via Bartlett’s test:

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

Variance ratio test:

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>17.44</td>
<td>3.37</td>
<td>10.11</td>
<td>9.67 - 25.21</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>6.88</td>
<td>2.59</td>
<td>10.70</td>
<td>1.38 - 12.38</td>
</tr>
<tr>
<td>combined</td>
<td>26</td>
<td>10.54</td>
<td>2.26</td>
<td>11.50</td>
<td>5.89 - 15.18</td>
</tr>
</tbody>
</table>

Ratio = $sd(0) / sd(1)$

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

Robustness to normality assumptions via Levene’s test or Brown-Forsythe:

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

Control

<table>
<thead>
<tr>
<th>Group</th>
<th>Summary of dif</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salt: 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>17.44</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>6.88</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>10.54</td>
<td>26</td>
</tr>
</tbody>
</table>

W0 = 0.00422467 $df(1, 24)$ $Pr > F = 0.94871439$
W50 = 0.03418276 $df(1, 24)$ $Pr > F = 0.8548721$
W10 = 0.00548156 $df(1, 24)$ $Pr > F = 0.94159416$

T-test

While no evidence against variance-homogeneity 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:

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>17.44</td>
<td>3.37</td>
<td>10.11</td>
<td>9.67 - 25.21</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>6.88</td>
<td>2.59</td>
<td>10.70</td>
<td>1.38 - 12.38</td>
</tr>
<tr>
<td>combined</td>
<td>26</td>
<td>10.54</td>
<td>2.26</td>
<td>11.50</td>
<td>5.89 - 15.18</td>
</tr>
</tbody>
</table>

Diff = mean(0) - mean(1)

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

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>63.21</td>
<td>2</td>
<td>31.61</td>
<td>119.26</td>
<td>0.0000</td>
</tr>
<tr>
<td>Within groups</td>
<td>38.96</td>
<td>147</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>102.16</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bartlett’s test for equal variances: $chi^2(2) = 16.0057$ $Prob > chi^2 = 0.000$
Comparing several groups

### Regress

```
regress
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>.434593287</td>
<td>1</td>
<td>.434593287</td>
<td>Prob &gt; F =</td>
<td>0.0607</td>
</tr>
<tr>
<td>Residual</td>
<td>5.65360671</td>
<td>48</td>
<td>.117783473</td>
<td>Adj R-squared =</td>
<td>0.0714</td>
</tr>
<tr>
<td>Total</td>
<td>6.0882</td>
<td>49</td>
<td>.12424898</td>
<td>Root MSE =</td>
<td>.3432</td>
</tr>
</tbody>
</table>

### Kruskal-Wallis Test

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

Kruskal-Wallis equality-of-populations rank test

<table>
<thead>
<tr>
<th>Species</th>
<th>Obs</th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>50</td>
<td>1482.00</td>
</tr>
<tr>
<td>versicolor</td>
<td>50</td>
<td>4132.50</td>
</tr>
<tr>
<td>virginica</td>
<td>50</td>
<td>5710.50</td>
</tr>
</tbody>
</table>

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

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

### Spearman's and Kendall's Rank Correlation

```
spearman Petal_Length Sepal_Length
```

<table>
<thead>
<tr>
<th>Petal_Length</th>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>0.8819</td>
<td>0.3812</td>
</tr>
</tbody>
</table>

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

```
sidak Petal_Length Sepal_Length
```

<table>
<thead>
<tr>
<th>Petal_Length</th>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>0.8756</td>
<td>0.3812</td>
</tr>
</tbody>
</table>

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

### Pearson Correlation

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

<table>
<thead>
<tr>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>-0.1176</td>
</tr>
</tbody>
</table>

with (Sidak) correction for multiple comparisons

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

<table>
<thead>
<tr>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>-0.1176</td>
</tr>
</tbody>
</table>

### Pearson Correlation

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

<table>
<thead>
<tr>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>-0.1176</td>
</tr>
</tbody>
</table>

with (Sidak) correction for multiple comparisons

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

<table>
<thead>
<tr>
<th>Sepal_Length</th>
<th>Sepal_Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>-0.1176</td>
</tr>
</tbody>
</table>
Comparing two continuous variables

```
1. use data/hubble, clear
2. 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

Hubble's 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 \( 3.085 \times 10^{19} \) seconds).

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

```
Linear regression

Y = \( \beta_0 + \beta_X X + \epsilon \), \( \epsilon \sim N(0, \sigma^2) \)
outcome \( Y \) and covariate \( X \). Assumptions
- Linearity
- Independence
- Variance homogeneity
- Approximate normal residuals
But no assumptions regarding \( X \).

Stata syntax
```
1. regress v D
```

```
Source | SS   df  MS
-------------+------------------------------ 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 application it is better to omit the intercept

```
regress v D, noconst
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7.2820e+09</td>
<td>1</td>
<td>7.2820e+09</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>35880167.3</td>
<td>35</td>
<td>1025147.64</td>
<td>R-squared = 0.9951</td>
</tr>
<tr>
<td>Total</td>
<td>7.3179e+09</td>
<td>36</td>
<td>203275674</td>
<td>Adj R-squared = 0.9950</td>
</tr>
</tbody>
</table>

| v | Coef. | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|---|-------|-----------|-------|---------|----------------------|
| D | 70.66722 | 0.8384642 | 84.28 | 0.000   | 68.96504 72.36939   |

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

Age of the universe: 1.384e+10 years

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 example (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 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1034.56337</td>
<td>2</td>
<td>517.281684</td>
<td>Prob &gt; F = 0.0134</td>
</tr>
<tr>
<td>Residual</td>
<td>2273.89817</td>
<td>23</td>
<td>98.8651379</td>
<td>R-squared = 0.3127</td>
</tr>
<tr>
<td>Total</td>
<td>3308.46154</td>
<td>25</td>
<td>132.338462</td>
<td>Adj R-squared = 0.2529</td>
</tr>
</tbody>
</table>

| post | Coef. | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|------|-------|-----------|-------|---------|----------------------|
| pre  | 0.4856021 | 0.2637398 | 1.84  | 0.079   | -0.0599852 1.031189  |
| salt | -10.73692 | 4.099835  | -2.62 | 0.015   | -19.21807 -2.257675  |
| _cons | 46.87944 | 15.45143  | 3.03  | 0.006   | 14.91572 78.84316   |
```

**Scalars**

```plaintext
e(N) = 26
e(df_m) = 2
e(df_r) = 23
e(F) = 5.232195039475376
```

**Macros**

```plaintext
_e(cmdline) : "regress post pre salt"
e(title) : "Linear regression"
e(marginsok) : "XB default"
e(vce) : "ols"
e(predict) : "regres_p"
```

**Matrices**

```plaintext
_e(b) : 1 x 3
_e(V) : 3 x 3
```
Factors, categorical predictors

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

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

```
use data/ami, clear
regress wmi i.agecat
```

- The hearts ability to pump

```
descibe
```

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

Factors

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

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

```
regress wmi agecat
```

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

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

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

```
use data/ami, clear
regress wmi agecat
```
Two-way 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):

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

### 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`

```stata
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

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

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

### 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 change reference with `.b`
- Post-estimation with `lincom`

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

( 1) 1b.sex + _cons = 0

### Margins

Here we use stata’s interaction syntax

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

Adjusted predictions Number of obs = 1878
Model VCE : OLS
Expression : Linear prediction, predict()
Margins

```bash
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**

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

**Main effects and interaction (full factorial)**

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

**Interaction between continuous and categorical predictors**

```bash
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`

---

Margins

```bash
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
-------------+----------------------------------------------------------------
sex#agecat | 65-75 #| female | 1.514966 .0335479 45.16 0.000 1.449171 1.580761
male#-65 | 1.426042 .0293544 48.58 0.000 1.368471 1.483613
male#65-75 | 1.371245 .0266469 51.46 0.000 1.318984 1.423505
male#75- | 1.452912 .0194351 70.53 0.000 1.332659 1.408893
female #| 65-75 | 1.370776 .0165916 87.57 0.000 1.336871 1.404682
female #| 75- | 1.257678 .0248925 50.52 0.000 1.208858 1.306498
------------------------------------------------------------------------------
```

```
Margins
1 margins sex#agecat
```

### Adjusted predictions

```
Number of obs = 1878
Model VCE : OLS
Expression : Linear prediction, predict()
```

```
| Marg | Delta-method | 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 #| 65-75 | 1.370776 .0165916 87.57 0.000 1.332659 1.408893
female #| 75- | 1.257678 .0248925 50.52 0.000 1.208858 1.306498
```
Margins

```
marginsplot
```

![Adjusted Predictions of sex#agecat with 95% CIs](chart.png)

Interactions

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

<table>
<thead>
<tr>
<th>Test for interaction</th>
<th></th>
<th></th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. agecat#2.sex = 0</td>
<td>0.77</td>
<td>0.4618</td>
<td></td>
</tr>
<tr>
<td>2. agecat#2.sex = 0</td>
<td>0.77</td>
<td>0.4618</td>
<td></td>
</tr>
</tbody>
</table>

Or with the newer contrast function

```
contrast agecat#sex
```

Contrasts of marginal linear predictions

```
Margins : asbalanced
```

<table>
<thead>
<tr>
<th>df</th>
<th>F</th>
<th>P&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.77</td>
<td>0.4618</td>
</tr>
</tbody>
</table>

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
Model  | 12.9035584 | 3  | 4.30118613 | F( 3, 1874) = 26.30
Residual | 306.528753 | 1874 | .163569238 | R-squared = 0.0404
Total | 319.432311 | 1877 | .170182371 | Root MSE = .40444
```

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

Main effects and constant term (intercept) are difficult to interpret without centering the age variable around some meaningful value. Much better to make some predictions...

```
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 | .0150342 | 78.10 | 0.000 | 1.372096 , 1.442782 |
| 7            | 1.382396 | .0130197 | 62.28 | 0.000 | 1.338861 , 1.42593 |
| 8            | 1.357352 | .0110252 | 48.18 | 0.000 | 1.302102 , 1.412602 |
| 9            | 1.332308 | .0090307 | 38.03 | 0.000 | 1.263598 , 1.401018 |
```

Predictions, margins

Predicting the age effect in males

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

| at#sex | Delta-method | Margin | Std. Err. | t | P>|t| | 95% Conf. Interval |
|--------|--------------|--------|-----------|---|------|-------------------|
| 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 | 80.86 | 0.000 | 1.436777 - 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 | 74.06 | 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 - 1.336839 |
| 7#male |              | 1.382396 | .021976 | 62.28 | 0.000 | 1.338861 - 1.42593  |
| 7#female |           | 1.26773 | .0183671 | 69.02 | 0.000 | 1.231708 - 1.303753 |

Adjusted Predictions of sex with 95% CIs
Predictions, margins
Example on quadratic effect via interaction operator

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

(Edgar Anderson's Iris Data)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>82.70</td>
<td>2</td>
<td>41.35</td>
<td>F( 2, 147) = 312.27</td>
</tr>
<tr>
<td>Residual</td>
<td>19.46</td>
<td>147</td>
<td>.1324</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>102.17</td>
<td>149</td>
<td>.6857</td>
<td>Adj R-squared = 0.8069</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sepal_Length</th>
<th>Coef. Std. Err. t P&gt;</th>
<th>t</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petal_Length</td>
<td>-.1644 (.0943) -.17 0.083 -.3507 (.0219)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

1. `sort Petal_Length`
2. `capture drop yhat_*`
3. `predict yhat`
4. `predict yhat_se, stdp`
5. `capture gen yhatLo = yhat-2*yhat_se`
6. `capture gen yhatHi = yhat+2*yhat_se`

For convenience we could put this in a simple function

1. `capture program drop mypredict`
2. `program mypredict`
3. `capture drop yhat_*`
4. `predict yhat`
5. `predict yhat_se, stdp`
6. `capture gen yhatLo = yhat-2*yhat_se`
7. `capture gen yhatHi = yhat+2*yhat_se`
8. `end`

Prediction, predict

1. `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

1. `clear`
2. `set obs 100`
3. `gen Petal_Length = 7*_n/_N+0.5`
4. `gen _Petal_Length = Petal_Length`
5. `mypredict`
6. `drop Petal_Length`
7. `tempfile _tmpdata`
8. `save `_tmpdata`, replace`
9. `use iris, clear`
10. `merge 1:1 _n using `_tmpdata`

Result # of obs.
-----------------------------------------
not matched 50
from master 50 (_merge==1)
from using 0 (_merge==2)
matched 100 (_merge==3)
-----------------------------------------
Prediction, out-of-sample

```
2way (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
2 gen CHF=chf-1
1 tabulate agecat chf, chi2 exact gamma
```

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

```
<table>
<thead>
<tr>
<th>Clinical heart pump failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>agecat</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
```

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

For paired data we can test for marginal homogeneity with the symmetry function.

Logistic regression

```
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 | .0089041 | .0039042 | -12.81| 0.000 | .0043694 - .0184726 |
```

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

Poisson regression

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

Contains data

```
obs: 24
city: 7
size: 648
```

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
Optimization : ML
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

<table>
<thead>
<tr>
<th>cases</th>
<th>IRR Std. Err. z P&gt;</th>
<th>z</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horsens</td>
<td>.7188806 .1304792 -1.82 0.069 .5036871 1.026013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kolding</td>
<td>.6896672 .1295238 -1.98 0.048 .4772859 .996533</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vejle</td>
<td>.7616123 .1430714 -1.45 0.147 .527027 .1.100614</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55-59</td>
<td>3.007214 .7466484 4.43 0.000 1.848515 4.892215</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-64</td>
<td>4.565885 1.05763 6.56 0.000 2.899711 7.189442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-69</td>
<td>5.857403 1.343919 7.70 0.000 3.735991 9.83417</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>6.403619 1.506919 7.89 0.000 4.037553 10.15624</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75+</td>
<td>4.135687 1.035041 4.27 0.000 2.53231 6.75427</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>.0035812 .0007171 -28.12 0.000 .0024186 .0053025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(pop)</td>
<td>1 (exposure)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>