1. **Objective:** Analyzing dental data using ordinary least square (OLS) and Generalized Least Square (GLS) in STATA.

2. **Scientific question:** Determine whether there is a difference between boys and girls with respect to the distance and its change over time.

3. **Dataset:** Dental study data set ([http://biosun01.biostat.jhsph.edu/~fdominic/teaching/LDA/dental.dat](http://biosun01.biostat.jhsph.edu/~fdominic/teaching/LDA/dental.dat))
   
   Data description: 27 children, 16 boys and 11 girls were observed at each ages 8, 10, 12 and 14 years. The data set has the following five columns:
   - Column 1: observation number
   - Column 2: child id number
   - Column 3: age
   - Column 4: distance
   - Column 3: gender indicator (0=girl, 1=boy)

   Outcome: distance, is a continuous variables.
   Covariate: age and gender.

4. **STATA output of the analysis**

   ```stata
   . *read the dental data set
   . infile obs id age dist sex using c:\data\dental.dat, clear
      (108 observations read)
   .
   . ****************
   . ******EDA******
   . ****************
   . *scatter plots
   . graph dist age if sex==0, xlab ylab ti("Scatter plot of the data for girls")
      > saving(g1,replace)
   . graph dist age if sex==1, xlab ylab ti("Scatter plot of the data for boys") s
      > aving(g2,replace)
   . graph using g1 g2
   ```
. graph dist if sex==0, box by(age) ti("box-plots of distance vs. age for girls") saving(g3, replace)

. graph dist if sex==1, box by(age) ti("box-plots of distance vs. age for boys") saving(g4, replace)

. graph using g3 g4

. *longitudinal plots
. sort id age
. graph dist age if sex==0, c(L) s(i) xlab ylab ti("Spaghetti plot of the data for girls") saving(g5, replace)

. graph dist age if sex==1, c(L) s(i) xlab ylab ti("Spaghetti plot of the data for boys") saving(g6, replace)

. graph using g5 g6

. *OLS (ignoring the correlation between responses for same person)
. reg dist age if sex==0

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>50.5920455</td>
<td>1</td>
<td>50.5920455</td>
<td>F( 1, 42) = 10.80</td>
</tr>
<tr>
<td>Residual</td>
<td>196.697727</td>
<td>42</td>
<td>4.68327922</td>
<td>Prob &gt; F = 0.0021</td>
</tr>
<tr>
<td>Total</td>
<td>247.289773</td>
<td>43</td>
<td>5.75092495</td>
<td>R-squared = 0.2046</td>
</tr>
</tbody>
</table>

| dist | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|-------|------|------------------------|
| age  | .4795455 | .1459028 | 3.287 | 0.002 | .1851016 | .7739893 |
. reg dist age if sex==1

Source |       SS       df       MS                  Number of obs =      64
---------+------------------------------               F(  1,    62) =   36.65
Model |  196.878125     1  196.878125               Prob > F      =  0.0000
Residual |  333.059375    62   5.3719254               R-squared     =  0.3715
---------+------------------------------               Adj R-squared =  0.3614
Total |  529.93750    63  8.41170635               Root MSE      =  2.3177

------------------------------------------------------------------------------
  dist |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  age |    .784375   .1295657      6.054   0.000       .5253769    1.043373
     _cons |   16.34063   1.454371     11.236   0.000       13.43338    19.24787
------------------------------------------------------------------------------

. reg dis age sex

Source |       SS       df       MS                  Number of obs =     108
---------+------------------------------               F(  2,   105) =   36.41
Model |  375.820875     2  187.910438               Prob > F      =  0.0000
Residual |  541.871254   105  5.16067861               R-squared     =  0.4095
---------+------------------------------               Adj R-squared =  0.3983
Total |   917.69213   107  8.57656196               Root MSE      =  2.2717

------------------------------------------------------------------------------
  dist |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  age |   .6601852   .0977589      6.753   0.000       .4663473    .8540231
   sex |   2.321023   .4448862      5.217   0.000       1.438896     3.20315
     _cons |   15.38569   1.128567     13.633   0.000       13.14795    17.62343
------------------------------------------------------------------------------

. sum age

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
   age |     108          11   2.246493          8         14

. egen mage=mean(age)

. disp mage

11

. gen cage=age-mage

. xi: reg dist cage sex i.sex*cage

i.sex |  Isex_0-1            (naturally coded; Isex_0 omitted)
i.sex*cage |  IsXCag_#            (coded as above)

Source |       SS       df       MS                  Number of obs =     108
---------+------------------------------               F(  3,   104) =   25.39
Model |  387.935027     3  129.311676               Prob > F      =  0.0000
Residual |  529.757102   104  5.09381829               R-squared     =  0.4227
---------+------------------------------               Adj R-squared =  0.4061
Total |  917.69213   107  8.57656196               Root MSE      =  2.2569

------------------------------------------------------------------------------
  dist |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
. *add quadratic term of age to the OLS
. gen age2=cage^2

. reg dis cage age2 sex

| Dist | Coef.  | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|------|--------|-----------|-------|---------|----------------------|
| cage | .6601852 | .0980966 | 6.730 | 0.000   | .465656 .8547144    |
| age2 | .0289352 | .0548377 | 0.528 | 0.599   | -.07981 .1376803   |
| sex  | 2.321023 | .4464228 | 5.199 | 0.000   | 1.43575 3.268296   |
| _cons| 22.50305 | .4396351 | 51.186| 0.000   | 21.63124 23.37486  |

. *we found quadratic term of age is not significant

. *calculate the autocorrelation matrix, determine the correlation structure
. sort id age
. by id: gen num=_n
. autocor dist num id

<table>
<thead>
<tr>
<th></th>
<th>time1</th>
<th>time2</th>
<th>time3</th>
<th>time4</th>
</tr>
</thead>
<tbody>
<tr>
<td>time1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time2</td>
<td>0.6256</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time3</td>
<td>0.7108</td>
<td>0.6349</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>time4</td>
<td>0.5998</td>
<td>0.7593</td>
<td>0.7950</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

acf
1. .685653
2. .7284871
3. .5998338
Autocorrelation Scatterplot

ACF

tau

. *GLS(using independent, exchangeable, exponential(ar1) and unstructured corre
lation)
. *independent correlation(same as OLS)
. xtgee dist age if sex==0, i(id) corr(ind)

Iteration 1: tolerance = 6.603e-15

GEE population-averaged model

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>Number of groups</th>
<th>Link</th>
<th>Obs per group: min</th>
<th>Family</th>
<th>Obs per group: avg</th>
<th>Correlation</th>
<th>Scale parameter:</th>
<th>Wald chi2(1)</th>
<th>Prob &gt; chi2</th>
<th>Deviance</th>
<th>Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gaussian</td>
<td></td>
<td>independent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>.4795455</td>
<td>.1459028</td>
<td>3.287</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>17.37273</td>
<td>1.637755</td>
<td>10.608</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

. xtcorr

Estimated within-id correlation matrix R:

    c1  c2  c3  c4
r1  1.0000
r2  0.0000  1.0000
r3  0.0000  0.0000  1.0000
r4  0.0000  0.0000  0.0000  1.0000

. xtgee dist age if sex==1, i(id) corr(ind)

Iteration 1: tolerance = 2.663e-15

GEE population-averaged model

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>Number of groups</th>
<th>Link</th>
<th>Obs per group: min</th>
<th>Family</th>
<th>Obs per group: avg</th>
<th>Correlation</th>
<th>Scale parameter:</th>
<th>Wald chi2(1)</th>
<th>Prob &gt; chi2</th>
<th>Deviance</th>
<th>Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gaussian</td>
<td></td>
<td>independent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pearson chi2(62):                   333.06      Deviance           =    333.06
Dispersion (Pearson):             5.371925      Dispersion         =  5.371925

| dist | Coef.   | Std. Err. |     z  |     P>|z| | [95% Conf. Interval] |
|-------|---------|-----------|-------|--------|----------------------|
| age   | .784375 | .1295657  | 6.054 | 0.000  | [.5304309 1.038319]  |
| _cons | 16.34063| 1.454371  | 11.236| 0.000  | [13.49011 19.19114]  |

. xtcorr

Estimated within-id correlation matrix R:

c1    c2    c3    c4
r1  1.0000
r2  0.8545 1.0000
r3  0.8545 0.8545 1.0000
r4  0.8545 0.8545 0.8545 1.0000

. *exchangeable correlation
. xtgee dist age if sex==0, i(id) corr(exc)

Iteration 1: tolerance = 7.735e-16

GEE population-averaged model
Number of obs = 44
Group variable: id Number of groups = 11
Link: identity Obs per group: min = 4
Family: Gaussian avg = 4.0
Correlation: exchangeable max = 4
Scale parameter: 4.683279 Wald chi2(1) = 74.22

| dist | Coef.   | Std. Err. |     z  |     P>|z| | [95% Conf. Interval] |
|-------|---------|-----------|-------|--------|----------------------|
| age   | .4795455| .0556625  | 8.615 | 0.000  | [.370449 .5886419]   |
| _cons | 17.37273| .8684316  | 20.005| 0.000  | [15.67063 19.07482]  |

. xtcorr

Estimated within-id correlation matrix R:

c1    c2    c3    c4
r1  1.0000
r2  0.8545 1.0000
r3  0.8545 0.8545 1.0000
r4  0.8545 0.8545 0.8545 1.0000

. xtgee dist age if sex==1, i(id) corr(exc)

Iteration 1: tolerance = 6.146e-16

GEE population-averaged model
Number of obs = 64
Group variable: id Number of groups = 16
Link: identity Obs per group: min = 4
Family: Gaussian avg = 4.0
Correlation: exchangeable max = 4
Scale parameter: 5.371925 Wald chi2(1) = 68.51

| dist | Coef.   | Std. Err. |     z  |     P>|z| | [95% Conf. Interval] |
|-------|---------|-----------|-------|--------|----------------------|
| age   | .4795455| .0556625  | 8.615 | 0.000  | [.370449 .5886419]   |
| _cons | 17.37273| .8684316  | 20.005| 0.000  | [15.67063 19.07482]  |
. xtgee dist age if sex==0, i(id) t(age) corr(ar1)

Iteration 1: tolerance = .00359913
Iteration 2: tolerance = 1.357e-07

GEE population-averaged model                   Number of obs      =        44
Group and time vars:                id age      Number of groups   =        11
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                         AR(1)                     max =         4
Wald chi2(1)                           = 32.30
Scale parameter:                  4.683506      Prob > chi2        =    0.0000

. xtgee dist age if sex==1, i(id) t(age) corr(ar1)

Iteration 1: tolerance = .01108648
Iteration 2: tolerance = .00005707
Iteration 3: tolerance = 4.267e-07

GEE population-averaged model                   Number of obs      =        64
Group and time vars:                id age      Number of groups   =        16
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                         AR(1)                     max =         4
Wald chi2(1)                           = 35.77
Scale parameter:                  5.376588      Prob > chi2        =    0.0000

-----------------------------

dist |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-----------------------------
 age |    .784375   .0947624      8.277   0.000       .5986441    .9701059
 _cons |   16.34062   1.134732     14.400   0.000       14.11659    18.56466
-----------------------------

dist |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-----------------------------
 age |   .4847376   .0852895      5.683   0.000       .3175733     .651902
 _cons |    17.3066   1.113527     15.542   0.000       15.12413    19.48907
-----------------------------

dist |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-----------------------------
 age |   .4847376   .0852895      5.683   0.000       .3175733     .651902
 _cons |    17.3066   1.113527     15.542   0.000       15.12413    19.48907
-----------------------------

. xtcrr

Estimated within-id correlation matrix R:

c1   c2   c3   c4
r1   1.0000
r2   0.4651  1.0000
r3   0.4651  0.4651  1.0000
r4   0.4651  0.4651  0.4651  1.0000

. *exponential correlation
. xtgee dist age if sex==0, i(id) t(age) corr(ar1)

Iteration 1: tolerance = .00359913
Iteration 2: tolerance = 1.357e-07

GEE population-averaged model                   Number of obs      =        44
Group and time vars:                id age      Number of groups   =        11
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                         AR(1)                     max =         4
Wald chi2(1)                           = 32.30
Scale parameter:                  4.683506      Prob > chi2        =    0.0000

. xtcrr

Estimated within-id correlation matrix R:

c1   c2   c3   c4
r1   1.0000
r2   0.8847  1.0000
r3   0.7827  0.8847  1.0000
r4   0.6925  0.7827  0.8847  1.0000

. xtgee dist age if sex==1, i(id) t(age) corr(ar1)

Iteration 1: tolerance = .01108648
Iteration 2: tolerance = .00005707
Iteration 3: tolerance = 4.267e-07

GEE population-averaged model                   Number of obs      =        64
Group and time vars:                id age      Number of groups   =        16
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                         AR(1)                     max =         4
Wald chi2(1)                           = 35.77
Scale parameter:                  5.376588      Prob > chi2        =    0.0000

-----------------------------

_xtcrr

Estimated within-id correlation matrix R:

c1   c2   c3   c4
r1   1.0000
r2   0.8847  1.0000
r3   0.7827  0.8847  1.0000
r4   0.6925  0.7827  0.8847  1.0000

. xtgee dist age if sex==1, i(id) t(age) corr(ar1)

Iteration 1: tolerance = .01108648
Iteration 2: tolerance = .00005707
Iteration 3: tolerance = 4.267e-07

GEE population-averaged model                   Number of obs      =        64
Group and time vars:                id age      Number of groups   =        16
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                         AR(1)                     max =         4
Wald chi2(1)                           = 35.77
Scale parameter:                  5.376588      Prob > chi2        =    0.0000
. xtcorr

Estimated within-id correlation matrix R:

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>0.4657</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>0.2169</td>
<td>0.4657</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>0.1010</td>
<td>0.2169</td>
<td>0.4657</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

. *unstructured correlation
. xtgee dist age if sex==0, i(id) t(age) corr(uns)

Iteration 1: tolerance = .00563941
Iteration 2: tolerance = .00001054
Iteration 3: tolerance = 1.824e-08

GEE population-averaged model                      Number of obs      =        44
Group and time vars:                id age      Number of groups   =        11
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                  unstructured                     max =         4
Wald chi2(1)       =     29.50
Scale parameter:                  4.68365      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
  dist |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
--------+---------------------------------------------------------------
  age    |   .4711862   .0867513      5.431   0.000       .3011568    .6412155
_cons   |   17.46256   .9957054     17.538   0.000       15.51102    19.41411
------------------------------------------------------------------------------

. xtcorr

Estimated within-id correlation matrix R:

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>0.6505</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>0.8411</td>
<td>0.7814</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>0.8453</td>
<td>0.7918</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

. xtgee dist age if sex==1, i(id) t(age) corr(uns)

Iteration 1: tolerance = .00395941
Iteration 2: tolerance = .00004108
Iteration 3: tolerance = 5.253e-06
Iteration 4: tolerance = 5.331e-07

GEE population-averaged model                      Number of obs      =        64
Group and time vars:                id age      Number of groups   =        16
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                  unstructured                     max =         4
Wald chi2(1)       =     57.20
Scale parameter:                  5.372557      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
  dist |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
--------+---------------------------------------------------------------
  age    |   .4711862   .0867513      5.431   0.000       .3011568    .6412155
_cons   |   17.46256   .9957054     17.538   0.000       15.51102    19.41411
------------------------------------------------------------------------------
| dist | Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|------|-------|-----------|------|-----|-------------------|
| age  | 0.7801528 | 0.1031489 | 7.563 | 0.000 | 0.5779846 - 0.982321 |
| _cons | 16.40993 | 1.192964 | 13.756 | 0.000 | 14.07176 - 18.74809 |

```
.xtcorr

Estimated within-id correlation matrix R:

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>0.3833</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>0.6311</td>
<td>0.3867</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>0.2871</td>
<td>0.4803</td>
<td>0.5641</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```