Exploratory Data Analyses:
Use the Nepal children’s growth data to practice describing mean and association structures for an outcome (children’s mean weight) over time (age).

1. Download the data from the course web site. Perform and describe standard quality control investigations on weight and age to make sure that the data have no implausible values.
   **Prepare for data import
   
   . clear
   . set memory 40m
   (40960k)
   . set matsize 800

   **Import the Nepal children’s growth data
   
   . use "C:\nepal.stata.dat", clear

   **Make scatter plot of weight vs. age
   
   . scatter wt age, s(Oh)
From the scatter plot of weight vs. age for the Nepal children’s growth data, we can see that there are implausible weights above 80 kg and these observations should be noted as missing value.

**Generate a variable indicating the visit number for each id

.by id: gen time=_n

**Remove observations with weight above 80kg (Q2-4 I used this dataset)

.drop if (wt>80)
(123 observations deleted)

**Or, code observations with weight above 80kg as missing

.replace wt=. if wt>80
(123 real changes made, 123 to missing)

2. **Describe the structure of the data set, including distribution of observation times and a list of baseline and time-varying variables. Are the data balanced? Equally spaced?

This data set contains measurements on Nepalese children. For each of the 200 children, we have measurements at 5 time points, spaced approximately 4 months apart. The original design is balanced and approximately equally spaced, however, the data set is not balanced due to missing values. There are a total of 1000 observations, however, some of the visits contain missing data. After removing the 123 implausible weights above 80 kg, there are 877 complete records on 197 kids. Only 135 children have measurements at all 5 time points.

The baseline variables include child’s id (id), child’s gender (sex), mother’s age (mage), mother’s literacy status (lit), number of kids mother has had that died (died), and mother’s parity (alive). The time-varying variables are child’s age (age), child’s weight (wt), child’s height (ht), child’s arm circumference (arm), breast feeding status (bf), and date of visit (day, month, year).

3. **Plot weight against age and obtain predictions from:

i. **a linear regression

. reg wt age

----------------------------------------------------------------------------
| wt | Coef. Std. Err. t P>|t| [95%, Conf. Interval] |
|-----|----------------------|------------------|
| age | .139823 .0025827 54.14 0.000 .1347539 .144892 |
| _cons | 5.876167 .10909 53.87 0.000 5.662059 6.090276 |

** Obtain fitted value

. predict fit1
(option xb assumed; fitted values)
ii. a linear spline with knots at the 33rd and 67th percentile
**Find the 33rd and 67th percentile for age

. centile age, centile(33, 67)

-- Binom. Interp. --
Variable | Obs Percentile Centile [95% Conf. Interval]
----------------+-------------------------------------------------------------
age | 877 33 27 26 29
| 67 48 46 50.22477

**Create splines
. gen age27=age-27
. replace age27 = 0 if age27<0
(278 real changes made)
. gen age48=age-48
. replace age48 = 0 if age48<0
(576 real changes made)

** Fit a linear spline with knots at the 33rd and 67th percentile

. reg wt age age27 age48

-----------------------------------------------------------------------------
wt | Coef. Std. Err. t P>|t| [95% Conf. Interval]
------------+----------------------------------------------------------------
age | .1730811 .0099874 17.33 0.000 .153479 .1926833
age27 | -.0251327 .0160455 -1.57 0.118 -.056625 .0063596
age48 | -.0498387 .0154251 -3.23 0.001 -.0801133 -.0195642
_cons | 5.160476 .2059649 25.06 0.000 4.756232 5.564721

** Obtain fitted value
. predict fit2
(option xb assumed; fitted values)

iii. a cubic spline with knots at the 33rd and 67th percentile of age.
**Create polynomial terms
. gen age_2=age^2
. gen age_3=age^3
. gen age27_3=age27^3
. gen age48_3=age48^3

** Fit a cubic spline with knots at the 33rd and 67th percentile

. reg wt age age_2 age_3 age27_3 age48_3

-----------------------------------------------------------------------------
wt | Coef. Std. Err. t P>|t| [95% Conf. Interval]
------------+----------------------------------------------------------------
age | .3035956 .0874875 3.47 0.001 .1318846 .4753067
Plot the predicted curves from these three models and compare them. Interpret the output from the models in i. and ii.

** Plot predicted curves

\[ . \text{scatter } wt \text{ age, s(p) || line fit1 fit2 fit3 age, pstyle(p2 p3 >p4) sort||, } \]
\[ \text{legend(lab(2 "Linear Regression") lab(3 "Linear >Spline") lab(4 "Cubic Spline"))} \]

For the linear regression model, the average weight increase is 0.140 kg (95% CI: 0.135, 0.145) per one month increase in age. However, the linear regression model does not capture the non-linear relationship of weight and age. The linear spline and cubic spline model give very similar predicted values. The linear spline model adequately captures the non-linear trend in this case, and it has the following advantages over cubic linear model: easier to interpret and computationally less expensive. Hence the linear spline model is preferred in this case. For the linear spline model, there is 0.173 kg (95% CI: 0.154, 0.193) increase in weight per
one month age increase for children below 27 month of age. For children with age between 27 and 48 month, the weight will increase by 0.148 kg (95% CI: 0.097, 0.199) per one month age increase. For the children older than 48 month, the weight increase is 0.098 kg (95% CI: 0.017, 0.179) per one month increase in age.

4. Create the following spaghetti plots:

i. using all the data
   ** Create smooth Lowess line
   . lowess wt age, gen(smth) nograph
   ** Create spaghetti plot
   . sort id age
   . twoway line wt age, pstyle(p15) connect(L) || line smth age, pstyle(p1) clwidth(thick) sort ||, ytitle("Weight")

   ![Spaghetti plot with Lowess line](image1.png)

ii. using a 20% sample
   ** Create a sample that includes all three data points from 20% of the subjects.
   ** First, reshape to wide format so there is one row for each subject.
   . reshape wide wt age ht arm bf day month year smth, i(id) j(time)
   ** Set seed to 1234 (optional - but it will guarantee matching results)
   . set seed 1234
   ** Generate a random variable from uniform(0,1) distribution.
   . gen random=uniform()
** Now reshape data to long format.

```
. reshape long wt age ht arm bf day month year smth, i(id) j(time)
```

** Find the 20% quantile of the uniform(0,1) variable.

```
. centile random, centile(20)
```

-- Binom. Interp. --

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Percentile</th>
<th>Centile [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>1000</td>
<td>0.1862824</td>
<td>0.1457746 .2215365</td>
</tr>
</tbody>
</table>

** Use subjects with uniform(0,1) < 20th percentile.

```
. twoway line wt age if random<.1862, pstyle(p15) connect(L) || line smth age, pstyle(p1) clwidth(thick) sort ||, ytitle("Weight")
```

This plot shows the same general trend of increase weight with increasing age as the earlier spaghetti plot, but it also shows more clearly the trend for individuals in the data set. Most importantly, it shows that many children are only measured in small time intervals, for example the individual plotted in the top right is only measured between approximately 47 and 63 months of age.

iii. ** a ZAP spaghetti plot

and include a smooth lowess fit on these plots.

** We first construct a ZAP plot using the raw data instead of residuals.

** Obtain median weight for each subject.
. egen medwt=median(wt), by(id)

** Reshape data to wide format

. reshape wide wt age ht arm bf day month year smth, i(id) j(time)
. egen maxmwt=max(medwt)
. egen minmwt=min(medwt)
. egen medmwt=median(medwt)
. egen mwt25=pctile(medwt), p(25)
. egen mwt75=pctile(medwt), p(75)

** Reshape data to long format

. reshape long wt age ht arm bf day month year smth, i(id) j(time)
. gen maxwt=wt if medwt==maxmwt
. gen minwt=wt if medwt==minmwt
. gen mwt=wt if medwt==medmwt
. gen wt25=wt if medwt==mwt25
. gen wt75=wt if medwt==mwt75

** Make a ZAP spaghetti plot using raw data

. scatter wt age, s(oh)|| line maxwt minwt mwt wt25 wt75 age, clwidth(medthick medthick medthick medthick medthick) || line smth age, pstyle(p1) clwidth(thick) sort ||, legend(lab(1 "Weight")) ytitle("Weight")
The ZAP plot again shows that weight increases as a child’s age increases. To make a ZAP plot using the residuals, the above procedure may be repeated, using the residuals from the model of your choice, rather than the raw data.

5. **Group the time (age) variable appropriately and create both a scatterplot matrix and a plot the ACF of the weight data. Include a confidence interval around the ACF estimate. Describe and interpret. (Remember the ACF is based on residuals; use the predictions from the linear regression above to obtain these residuals.)**

** Fit linear regression

```
.reg wt age
```

** Obtain residuals

```
.predict res1,resid
```

** Group time variable based on visit number, create scatterplot matrix and an ACF plot of the weight data

** NOTE: must read in autocor.ado in the working directory

```
.autocor res1 time id
```

<table>
<thead>
<tr>
<th>time1</th>
<th>time2</th>
<th>time3</th>
<th>time4</th>
<th>time5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time2</td>
<td>0.9158 1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time3</td>
<td>0.8901 0.9392 1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time4</td>
<td>0.8804 0.9083 0.9504 1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time5</td>
<td>0.8608 0.8754 0.9257 0.9359 1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

``` +----------+ | acf | |--------| | 1. | .9314421 | | 2. | .9015214 | | 3. | .8761476 | | 4. | .8607549 | +----------+
** Calculate ACF CI

    . keep id time res1
    . reshape wide res1, i(id) j(time)
    . pwcorr res11-res15, obs

    | res11  res12  res13  res14  res15
    |---------------------------------------------
    | res11 | 1.0000          | 185
    |       | 0.9158 1.0000   | 165 176
    | res13 | 0.8901 0.9392 1.0000 | 172 167 180
    |       | 0.8804 0.9083 0.9504 1.0000 | 166 165 167 176
    | res15 | 0.8608 0.8754 0.9257 0.9359 1.0000 | 153 150 153 152 160

The number of independent pairs for each lag \(N(u)\) can be calculated by adding up the elements along diagonal. The standard error of the ACF for each lag \(u\) is \(1/\sqrt{N}\). The code provided in the lab can only be used when there are no missing values in the data.

For lag 1, the number of independent pairs is

    . display 165+167+167+152
    651

For lag 2, the number of independent pairs is
For lag 3, the number of independent pairs is 316

For lag 4, the number of independent pairs is 153

** Plot ACF with 95% CI **

```
. gen lb = acf - 2*se
. gen ub = acf + 2*se
. replace ub = 1 if ub>1
. scatter acf lb ub lag, ylabel(0 (0.5) 1)
```
6. Create and plot a variogram of the weight data. Use this variogram to create an ACF plot that is not based on the grouped time data. You may add this plot to the grouped data ACF from 5. or plot it separately. (Also use residuals as in 5.)

** Must read in variogram.ado, xtdiff.ado, ksmapprox.ado

** Plot variogram of the weight data

```
.variogram res1
```

** Use the variogram output

```
.insheet using vario, names clear
(4 vars, 1610 obs)
```

** Calculate ACF using variogram

** vsmth is the smooth average of sample variogram \( \gamma(u) \)

** vary is the variance \( \sigma^2 \)

** ulag is the lag \( u \)

```
.gen varioacf = 1 - vsmth/vary
(1410 missing values generated)
```

**Make ACF plot using ungrouped data

```
.twoway line varioacf ulag, ylabel(-1(1)1) yline(-1 1)
```
7. Write down a marginal linear regression model for weight on age using an independence model and a uniform (exchangeable) correlation structure. Adjust for gender, height, breastfeeding, mother’s age, and mother’s literacy, and use an appropriate function of age (as determined from 3.). Fit these models with ordinary least squares (OLS) and weighted least squares (WLS), respectively. Make a table comparing the intercept and slope estimates (and their standard errors) from the two models and interpret. Which results do you think are more appropriate? Does the average weight of children at a given age differ according to the mother’s literacy?

\[ Y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + \beta_3 \text{ht}_{ij} + \beta_4 \text{bf}_{ij} + \beta_5 \text{mage}_i + \beta_6 \text{lit}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \]

In matrix notation, the normal linear regression model is \( \mathbf{Y} = \mathbf{X}\beta + \epsilon \), where

\[
\begin{align*}
\mathbf{Y} &= \begin{pmatrix}
\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})^t \\
\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})^t \\
\vdots \\
\mathbf{Y}_{197} = (Y_{197,1}, \ldots, Y_{197,n_{197}})^t
\end{pmatrix}_{877 \times 1}, \\
\mathbf{X} &= \begin{pmatrix}
1 & \text{age}_{1,1} & \ldots & \text{lit}_1 \\
1 & \text{age}_{1,2} & \ldots & \text{lit}_1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & \text{age}_{197,1} & \ldots & \text{lit}_{197}
\end{pmatrix}_{877 \times 7}
\end{align*}
\]

\[
\beta = \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_6
\end{pmatrix}_{7 \times 1}, \quad \epsilon = \begin{pmatrix}
\mathbf{\epsilon}_1 = (\epsilon_{1,1}, \ldots, \epsilon_{1,n_1})^t \\
\mathbf{\epsilon}_2 = (\epsilon_{2,1}, \ldots, \epsilon_{2,n_2})^t \\
\vdots \\
\mathbf{\epsilon}_{197} = (\epsilon_{197,1}, \ldots, \epsilon_{197,n_{197}})^t
\end{pmatrix}_{877 \times 1}
\]

\[ \epsilon_1 \sim \text{MVN}(0, R_i \sigma^2) \]

The \( \epsilon_i \) are independent.
Under an independence model:

\[
R_i = \begin{pmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & \ddots & \\
\vdots & \ddots & \ddots & 0 \\
0 & \ldots & 0 & 1
\end{pmatrix}_{n_i \times n_i}
\]

Under a uniform model:

\[
R_i = \begin{pmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ddots & \\
\vdots & \ddots & \ddots & \rho \\
\rho & \ldots & \rho & 1
\end{pmatrix}_{n_i \times n_i}
\]

**Under an independence model, fit OLS

\[. \text{reg wt age sex ht bf mage lit}\]

**Under a uniform model, fit WLS

\[. \text{xtgee wt age sex ht bf mage lit, i(id) corr(exc)}\]

| wt     | Coef.  | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| age    | 0.0213362 | 0.0054899 | 3.89  | 0.000 | 0.0105761 - 0.0320962 |
| sex    | -0.2694977 | 0.118 | -2.28 | 0.022 | -0.5007734 - 0.38222 |
| ht     | 0.189884 | 0.0085499 | 22.21 | 0.000 | 0.1731265 - 0.2066416 |
| bf     | -0.0677599 | 0.0327353 | -2.07 | 0.038 | -0.1319199 - 0.035998 |
| mage   | 0.0133423 | 0.0098055 | 1.36  | 0.174 | -0.0058761 - 0.0325607 |
| lit    | -0.0677599 | 0.0327353 | -2.07 | 0.038 | -0.1319199 - 0.035998 |
| _cons  | -5.753964 | 0.596781 | -9.64 | 0.000 | -6.923633 - 4.584295 |

The intercept in these models are not interpretable because there are no children in the data set measured at 0 months of age. For one month of age increase, the OLS predicts a decrease in weight (clearly not reasonable) of 0.004 kg and the WLS predicts an increase in weight of 0.021 kg, controlling for gender, height, breastfeeding and mother characteristics. For the WLS model, we assume uniform correlation structure between observations for the same child.
The WLS model with uniform correlation is more appropriate due to the high correlation of the observations for the same subject. I will suggest using a GLS robust estimation, then a poor choice of correlation structure will affect only the efficiency of our inferences for $\beta$, not their validity. Although both models estimate a positive effect of literacy of the mother on the child’s weight, these estimates are not significant, so we do not have enough evidence to answer this question.

8. Write down and fit the marginal model you would need to determine whether or not the rate of increase in weight differs according to whether or not the child’s mother is literate.

$$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + \beta_3 ht_{ij} + \beta_4 bf_{ij} + \beta_5 mage_i + \beta_6 lit_i + \beta_7 lit_i \ast age_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

```stata
gen agelit = age * lit
.xtgee wt age sex ht bf mage lit agelit, i(id) corr(exc)
```

| wt   | Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|------|-------|-----------|-------|------|----------------------|
| age  | .0215329 | .0054793 | 3.93  | 0.000 | .0107935 -.0322722   |
| sex  | -.2712976 | .1183903 | -2.29 | 0.022 | -.5033383 -.392569   |
| ht   | .1890613  | .0085577 | 22.09 | 0.000 | .1722886 .2058341    |
| bf   | -.0677382 | .03258  | -2.08 | 0.038 | -.1315938 -.0038826  |
| mage | .0135778  | .0098362 | 1.38  | 0.167 | -.0057008 .0328655   |
| lit  | -.1056625 | .4811693 | -0.22 | 0.826 | -.1.0.48737 .8374119 |
| litage | .0083558 | .010789 | 0.77  | 0.439 | -.0.127902 .0295019  |
| _cons | -5.69585 | .5986458 | -9.51 | 0.000 | -6.869174 -4.522526  |

The slope on the interaction term between literacy and age may be interpreted as the difference in the average increase in weight associated with a one month increase in age between children whose mothers are literate and children whose mothers are not literate. It appears that the children of literate women in this study gained weight at a slightly faster rate than the children of illiterate women, but again the difference is not statistically significant.

9. Refit the model from 7. using a conditional, random effects model. What random effect specification do you need to include in the model to induce a uniform correlation structure?

$$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + \beta_3 ht_{ij} + \beta_4 bf_{ij} + \beta_5 mage_i + \beta_6 lit_i + U_i + \epsilon_{ij}, \quad U_i \sim N(0, \sigma^2), \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

$U_i$ and $\epsilon_{ij}$ are independent. The normal random effect specification induces a uniform correlation.

```stata
.xtreg wt age sex ht bf mage lit, re i(id)
```
We again get a slightly positive association between weight and age when accounting for within subject correlation.

10. **Show that the OLS estimate of $\beta$ is unbiased using matrix notation.**

   $$E(\beta) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$$

11. **Derive the variance of the $\beta_{OLS}$ estimate using matrix notation.**

   We assume $Y \sim MVN(X\beta, V)$, where $V = \text{Var}(Y)$.

   $$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

   $$\text{Var}(\hat{\beta}_{OLS}) = \text{Var}[(X'X)^{-1}X'Y] = (X'X)^{-1}X'\text{Var}(Y)[(X'X)^{-1}X']'$$

   $$= (X'X)^{-1}X'VX(X'X)^{-1}$$

   If we further assume $\text{Var}(Y) = \sigma^2I$, then

   $$\text{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1}X'VX(X'X)^{-1} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

12. **Show that $\beta_{WLS}$ is optimally efficient when using $W = V^{-1}$. If $V = R\sigma^2$, do we need to know $\sigma^2$ for this result to hold?**

   We assume $Y \sim MVN(X\beta, V)$, where $\text{Var}(Y) = V$. Let $A_{WLS} = (X'WX)^{-1}X'W$ and $A_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}$. Then $\text{Var}(\hat{\beta}_{WLS}) = \text{Var}(A_{WLS}Y)$ and $\text{Var}(\hat{\beta}_{GLS}) = \text{Var}(A_{GLS}Y)$.

   $$\text{Var}(\hat{\beta}_{WLS}) = \text{Var}(A_{WLS}Y) = \text{Var}(A_{GLS}Y + (A_{WLS} - A_{GLS})Y)$$

   $$= \text{Var}(A_{GLS}Y) + \text{Var}((A_{WLS} - A_{GLS})A_{WLS}Y) + 2\text{Cov}(A_{GLS}Y, (A_{WLS} - A_{GLS})Y)$$

   Note that

   $$\text{Cov}(A_{GLS}Y, (A_{WLS} - A_{GLS})Y) = A_{GLS}\text{Var}(Y)(A_{WLS} - A_{GLS})'$$

   $$= (X'V^{-1}X)^{-1}X'V^{-1}V[(X'WX)^{-1}X'W - (X'V^{-1}X)^{-1}X'V^{-1}]'$$

   $$= (X'V^{-1}X)^{-1}X'V^{-1}V[(X'WX)^{-1}X'W] - (X'V^{-1}X)^{-1}X'V^{-1}V[(X'V^{-1}X)^{-1}X'V^{-1}]'$$

   $$= (X'V^{-1}X)^{-1}X'V^{-1}VWX(X'V^{-1}X)^{-1} - (X'V^{-1}X)^{-1}X'V^{-1}VX(X'V^{-1}X)^{-1}$$

   $$= (X'V^{-1}X)^{-1} - (X'V^{-1}X)^{-1} = 0$$
Hence $\text{Var}(\hat{\beta}_{WLS}) = \text{Var}(\hat{\beta}_{GLS}) + \text{Var}(A_{WLS} - A_{GLS})$ and we know that $\text{Var}(A_{WLS} - A_{GLS}) \geq 0$, so $\text{Var}(\hat{\beta}_{WLS}) \geq \text{Var}(\hat{\beta}_{GLS})$ and $V^{-1}$ is optimally efficient.

If $V = R\sigma^2$ we do not need to know $\sigma^2$ for this result to hold because the optimal efficient WLS estimator is

$$
\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y = (X'(R\sigma^2)^{-1}X)^{-1}X'(R\sigma^2)^{-1}Y = (X'R^{-1}X)^{-1}X'R^{-1}Y
$$

which is independent of $\sigma^2$. 