

Exploratory Data Analysis for Mean Function

Analysis of Pigs Data

I. Read-in Data & Explore Data Structure

- **Increase the memory to hold large data.**

```
. set memory 40m  
(40960k)
```

- **Increase the number of variables you can include in the dataset.**

```
. set matsize 100
```

- **Read-in the data.**

```
** use "c:\data\pigs.stata.dta", clear  
** alternatively click on: file>open>pigs.stata.data
```

- **Short Summary of the data:**

```
. summ  
Variable | Obs Mean Std. Dev. Min Max  
-----+-----  
week1   | 48 25.02083 2.468866 20 31  
week2   | 48 31.78125 2.790383 26.5 39  
week3   | 48 38.86458 3.544158 32.5 48  
week4   | 48 44.39583 3.734483 37 54  
week5   | 48 50.15625 4.534919 38.5 60  
-----+-----  
week6   | 48 56.44792 4.449766 48 67.5  
week7   | 48 62.45833 4.973155 52.5 76  
week8   | 48 69.30208 5.424275 59.5 81.5  
week9   | 48 75.21875 6.335401 64 88  
Id      | 48 24.5 14 1 48
```

- **What variables are in the dataset?**

Id – contains an id number identifying each pig.

week1, week2, ..., week9 – weight of pigs at each of 9 measurement times

- **How many observations are there? Is the data balanced?**

There are 48 observations on each variable, indicating that there is no missing data and that all 48 pigs are measured the same number of times (nine times). Also, we have no additional information about the measurement times except week numbers, indicating that all measurement times are equally spaced for all pigs. Therefore, the data is balanced.

- **Are there any characteristics of interest to the analysis?**

As the weeks increase (pigs get older), notice the increase in both the mean weight as well as in the standard deviations of the weight (“fanning out”). These characteristics are important to model selection.

- **Convert to long format**

For the next exploratory data analysis (EDA) we’ll need the data in the long format.

```
. reshape long week, i(Id) j(time)
(note: j = 1 2 3 4 5 6 7 8 9)
Data wide -> long
-----
Number of obs. 48 -> 432
Number of variables 10 -> 3
j variable (9 values) -> time
xij variables:
week1 week2 ... week9 -> week
-----
```

In the output above, Stata tells us that there are 9 timepoints “j”, and 48 subjects (pigs) “i”, so there will be $48 \times 9 = 432$ observations in the long format dataset. The variable “time” is created to indicate different observations (weights over the different weeks) from the same Id (pig).

- **Convert to Panel Data**

For Stata, we also want to tell Stata that the dataset is longitudinal/”time-series” dataset. To convert an ordinary data into a longitudinal dataset, use the “xtset” command, specifying the subject index and time index.

```
. xtset Id time
panel variable: Id, 1 to 48
time variable: time, 1 to 9
```

- **Describe the pattern of data, especially missing data patterns**

```
. xtodes
Id: 1, 2, ..., 48 n = 48
time: 1, 2, ..., 9 T = 9
Delta(time) = 1; (9-1)+1 = 9
(Id*time uniquely identifies each observation)
Distribution of T_i: min 5% 25% 50% 75% 95% max
9 9 9 9 9 9 9
Freq. Percent Cum. | Pattern
-----+-----
48 100.00 100.00 | 111111111
-----+-----
48 100.00 | XXXXXXXXXXX
```

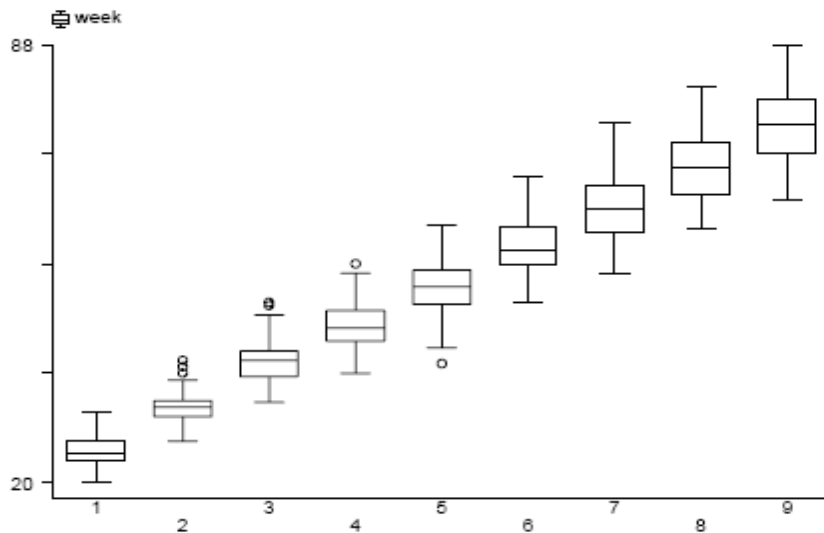
“Delta(time) = 1” means that the time interval between two observations is 1, and this is equally spaced data. From the distribution of T_i we know that there is no missing data, every subject has 9 observations. This is also confirmed by the pattern “111111111” with 100%. The “1’s” indicate where observations occurred during the 9 week period (note that there are 9 “1’s” in a row). If, for example, 6 of the 48 pigs (12.5%) were missing their 4th and 9th observations, stata would have printed out:

```
Freq. Percent Cum. | Pattern
-----+-----
42 87.5 87.50 | 111111111
6 12.5 100.00 | 111011110
-----+-----
48 100.00 | XXXXXXXXXXX
```

II. Explore the Marginal Mean function in the regression model.

- **Explore the marginal relationship between weight and time**

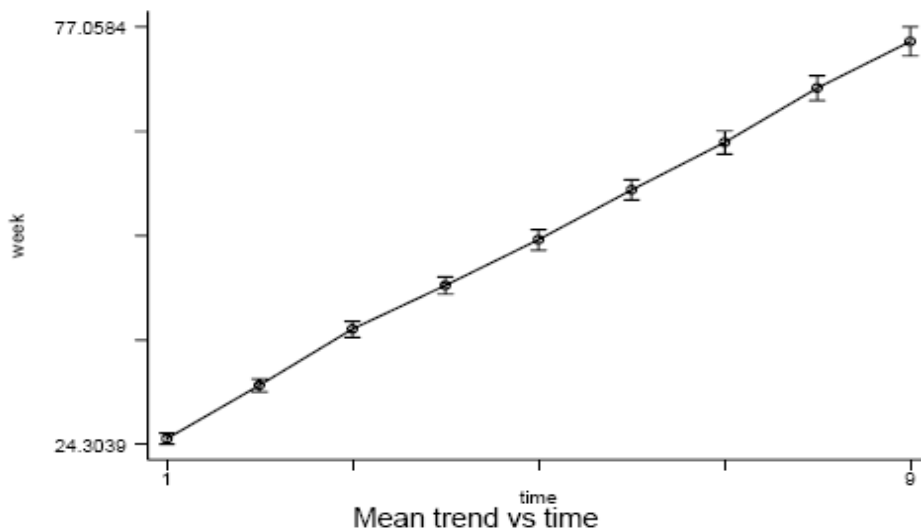
```
** always sort the variable if you want actions by that variable **  
. sort time  
. graph box week, over(time)
```



Observe the increasing trend with time, as well as the increase of spread.

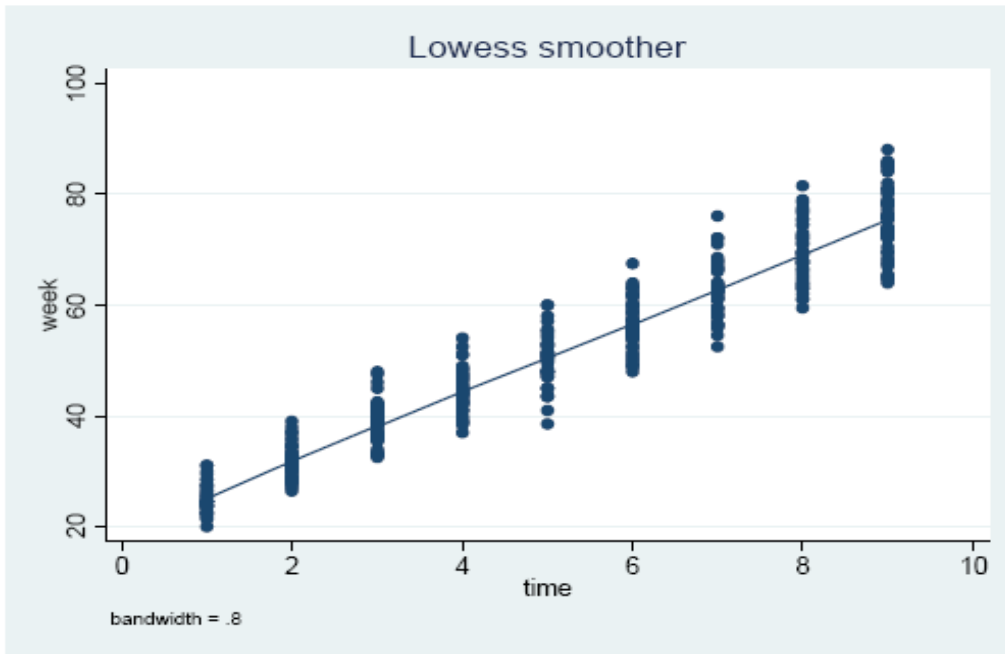
- **Plot the marginal mean trend**

```
. ** run xtgraph.ado in the do-file editor before using the command **  
. xtgraph week, ti("Mean trend vs time") bar(ci)
```



We can see the mean growth trend is positive and quite linear across time. Also, note that the CI's increase as time increases (again, the "fanning out").

- **Plot the lowess smooth curve** (very useful in exploring the mean trend for non-linear data.)
. ksm week time, lowess gen(weeksmth)



The smoothed curve is saved in the variable “weeksmth”.

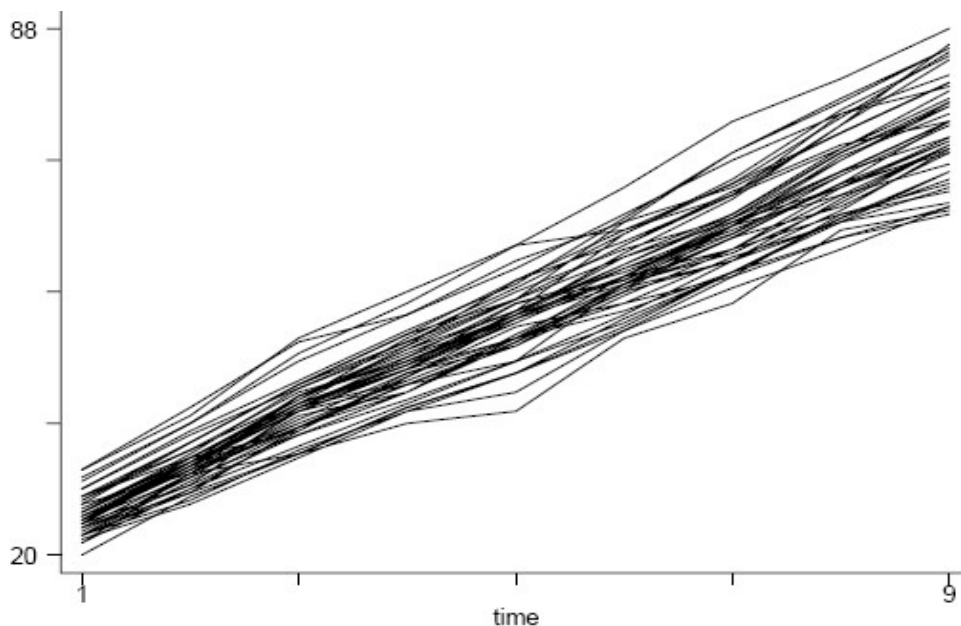
III. Explore the subject-specific mean function

- **Plot a line for growth trajectory of each pig**

To see if each pig gains weight over time, let’s plot the line (spaghetti) plot for the longitudinal relationship between weight and time for each pig. Sorting the data correctly is always important in creating LDA plots.

** we need to plot by time for EACH subject, ie. by id AND time, so we must sort the data accordingly. **

```
. sort Id time
. twoway line week time
```

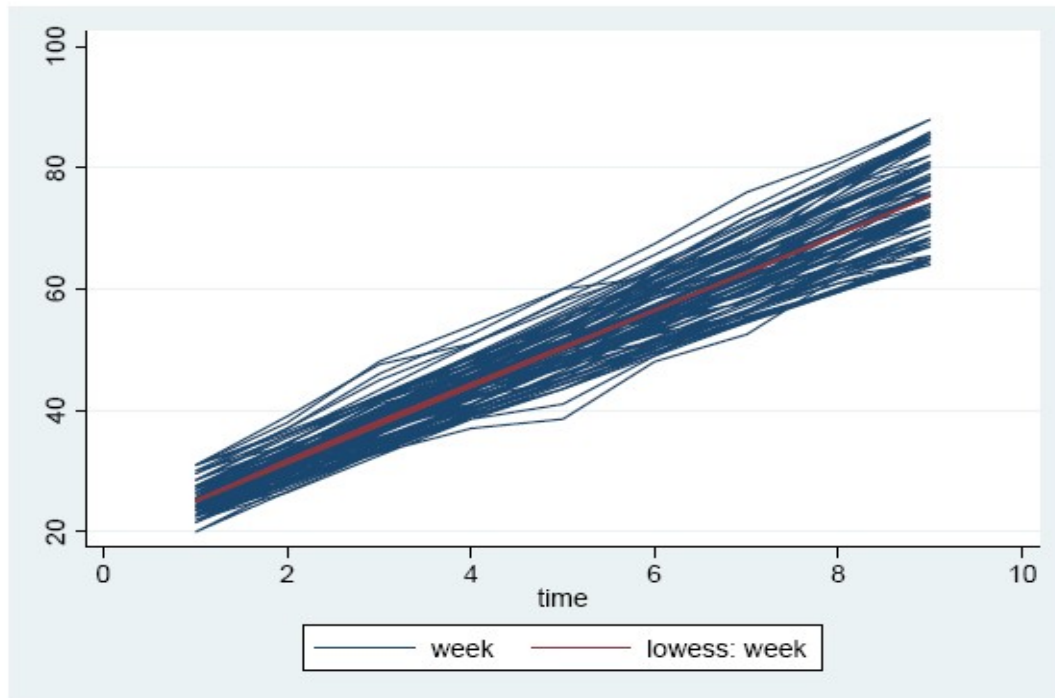


- **What do you conclude from the graph?**

A linear relationship between outcome and time is shown. Also, not much cross-over of these lines indicates that the relative order of pigs, ordered by their weights, remain unchanged over time.

- **Is the marginal growth trend described by the smooth growth curve consistent with individual longitudinal growth trend?**

```
. twoway line week weeksmth time
```



It can be seen that the marginal relationship is consistent with subject-specific longitudinal relationship.

- **Make a ZAP plot**

Remove the time effect on the growth, explore the residual to more directly look at the trajectory (relative order) of growth for each pig.

```
. reg week time
. predict weekrs, resid
```

- **Median residual of each pig**

```
. egen mweekrs=median(weekrs), by(Id)
```

- **Find the pigs that have the minimum, maximum, median, 25th percentile, and 75th percentile median weights.**

Reshape to wide to find "the pig" with specified median residual. If don't use wide format, the percentiles you generated would be among all the 432 observations, instead of among the 48 pigs.

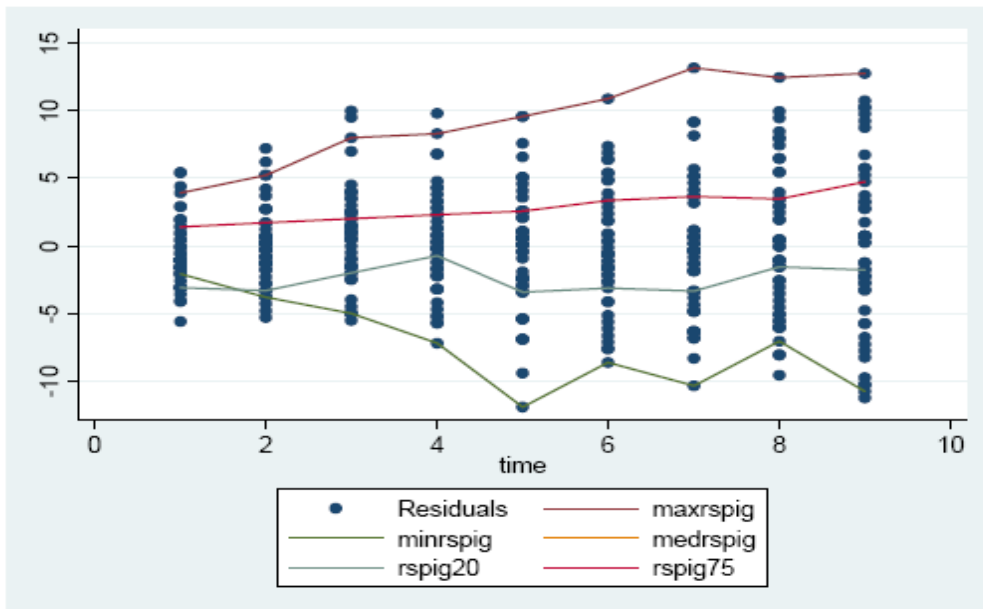
```
. reshape wide weekrs weeksmth week, i(Id) j(time)
. egen maxmrs=max(mweekrs)
. egen minmrs=min(mweekrs)
. egen mrs20=pctile(mweekrs),p(20)
. egen mrs74 = pctile(mweekrs),p(74)
. egen mrs76 = pctile(mweekrs),p(76)
```

Because there are 48 pigs, and 48 is an even number, the 75th percentile is mean of two

values. Therefore, it will not be equal exactly to some data points, and we should use the 74th or 76th percentiles as a surrogate. If you have an odd number of subjects, say 49 pigs, the 75th percentiles would be fine, and the 20th percentile would have a problem. So BE CAREFUL with percentiles. Reshape to long for plotting the residuals for each corresponding pig.

```
. reshape long weekrs weeksmth week, i(Id) j(time)
. gen maxrspig=weekrs if mweekrs==maxmrs
. gen minrspig=weekrs if mweekrs==minmrs
. gen rspig20=weekrs if mweekrs==mrs20
. gen rspig75 = weekrs if mweekrs<=mrs76 & mweekrs>mrs74

. twoway (scatter weekrs time) (line maxrspig minrspig rspig20 rspig75 time)
```



Notice that the pigs with maximum and minimum median residual are not the ones that are furthest away from the general pigs at the beginning, however, they turn out to be two extremes as time goes by. There is some change on the relative orders of the pigs, but not much.