**data** wide\_lead;

input ID TRT $ PB1 - PB4;

datalines;

1 P 30.8 26.9 25.8 23.8

2 A 26.5 14.8 19.5 21.0

……………… data omitted ………………

;

**run**;

**proc** **print** data=wide\_lead;

**run**;

| **Obs** | **ID** | **TRT** | **PB1** | **PB2** | **PB3** | **PB4** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 1 | P | 30.8 | 26.9 | 25.8 | 23.8 |
| **2** | 2 | A | 26.5 | 14.8 | 19.5 | 21.0 |
| **3** | 3 | A | 25.8 | 23.0 | 19.1 | 23.2 |
| **4** | 4 | P | 24.7 | 24.5 | 22.0 | 22.5 |
| **5** | 5 | A | 20.4 | 2.8 | 3.2 | 9.4 |

\*Now we'll change the wide formatted data to long formatted;

**data** long\_lead;

set wide\_lead;

week=**0**;

PB=PB1;

output;

week=**1**;

PB=PB2;

output;

week=**4**;

PB=PB3;

output;

week=**6**;

PB=PB4;

output;

drop PB1-PB4;

**run**;

**proc** **print** data=long\_lead (obs=**10**);

**run**;

| **Obs** | **ID** | **TRT** | **week** | **PB** |
| --- | --- | --- | --- | --- |
| **1** | 1 | P | 0 | 30.8 |
| **2** | 1 | P | 1 | 26.9 |
| **3** | 1 | P | 4 | 25.8 |
| **4** | 1 | P | 6 | 23.8 |
| **5** | 2 | A | 0 | 26.5 |
| **6** | 2 | A | 1 | 14.8 |
| **7** | 2 | A | 4 | 19.5 |
| **8** | 2 | A | 6 | 21.0 |
| **9** | 3 | A | 0 | 25.8 |
| **10** | 3 | A | 1 | 23.0 |

**data** long\_lead2;

set wide\_lead;

array APB(**1**:**4**) PB1-PB4;

array Aweek(**1**:**4**) (**0** **1** **4** **6**);

do i=**1** to **4**;

PB = APB[i];

week = Aweek[i];

output;

end;

drop PB1-PB4 Aweek1 - Aweek4 i;

**run**;

**proc** **print** data=long\_lead2 (obs=**10**);

**run**;

| **Obs** | **ID** | **TRT** | **week** | **PB** |
| --- | --- | --- | --- | --- |
| **1** | 1 | P | 0 | 30.8 |
| **2** | 1 | P | 1 | 26.9 |
| **3** | 1 | P | 4 | 25.8 |
| **4** | 1 | P | 6 | 23.8 |
| **5** | 2 | A | 0 | 26.5 |
| **6** | 2 | A | 1 | 14.8 |
| **7** | 2 | A | 4 | 19.5 |
| **8** | 2 | A | 6 | 21.0 |
| **9** | 3 | A | 0 | 25.8 |
| **10** | 3 | A | 1 | 23.0 |

\*First we'll do a simple "spagetti plot" of the data;

**Proc** **SGplot** data = long\_lead;

series x=week y=PB / group =ID LineAttrs= (pattern=**1** color="black");

**run**;



**Proc** **SGplot** data = long\_lead;

series x=week y=PB / group =ID LineAttrs= (pattern=**1**);

**run**;



**Proc** **SGplot** data = long\_lead;

series x=week y=PB / group =ID;

**run**;



\*Now we will include the mean line on the graph;

**proc** **sort** data=long\_lead;

by week;

\*Calculate the mean by week;

**proc** **means** mean data=long\_lead;

by week;

var PB;

output out = MN\_dat mean = mn\_PB;

**proc** **print** data = MN\_dat;

**run**;

\*Stack the mean data onto the long dataset;

**data** stacked;

set long\_lead MN\_dat;

**run**;

| **Obs** | **week** | **\_TYPE\_** | **\_FREQ\_** | **mn\_PB** |
| --- | --- | --- | --- | --- |
| **1** | 0 | 0 | 100 | 26.406 |
| **2** | 1 | 0 | 100 | 19.091 |
| **3** | 4 | 0 | 100 | 19.792 |
| **4** | 6 | 0 | 100 | 22.204 |

**Proc** **SGplot** data = stacked;

series x=week y=PB / group =ID LineAttrs= (pattern=**1** color="black");

series x=week y=mn\_PB / LineAttrs= (pattern=**1** color="blue" thickness=**4**);

**run**;



\*\*We can also create separate plots by TRT group;

**proc** **sort** data=long\_lead;

by TRT;

**Proc** **SGplot** data = long\_lead;

by TRT;

series x=week y=PB / group =ID LineAttrs= (pattern=**1**);

**run**;



\* Here we'll do separate plots (panels) for each TRT group;

**Proc** **SGpanel** data = long\_lead;

PanelBy TRT / columns=**2**;

series x=week y=PB / group =ID LineAttrs= (pattern=**1**);

**run**;



\*Now we will include the mean line on the graph by TRT;

**proc** **sort** data=long\_lead;

by TRT week;

\*Calculate the mean by week;

**proc** **means** mean data=long\_lead;

by TRT week;

var PB;

output out = MN\_TRT\_dat mean = mn\_TRT\_PB;

**proc** **print** data = MN\_TRT\_dat;

**run**;

\*First, let's look at the mean by TRT group;

**Proc** **SGplot** data = MN\_TRT\_dat;

series x=week y=mn\_TRT\_PB / group =TRT LineAttrs= (pattern=**1** thickness=**3**);

**run**;



\*Now we'll look at the means by TRT group with the rest of the data;

\*Stack the mean data onto the long dataset;

**data** stacked\_TRT;

set long\_lead MN\_TRT\_dat;

**run**;

**proc** **sort** data=stacked\_TRT;

by TRT week;

\*First we'll do separate plots for each TRT group with the mean on each;

**Proc** **SGplot** data = stacked\_TRT;

by TRT;

series x=week y=PB / group =ID LineAttrs= (pattern=**1** color="black");

series x=week y=mn\_TRT\_PB / LineAttrs= (pattern=**1** color="blue" thickness=**4**);

**run**;

\* Now we'll combine them onto one plot with two panels;

**Proc** **SGpanel** data = stacked\_TRT;

PanelBy TRT / columns=**2**;

series x=week y=PB / group =ID LineAttrs= (pattern=**1** color="black");

series x=week y=mn\_TRT\_PB / LineAttrs= (pattern=**1** color="blue" thickness=**4**);

**run**;



\*Let's go back to the simple plot, but this time limit our sample.

First, we need to create a new variable to randomly thin our sample;

**proc** **sort** data=long\_lead;

by ID week;

**run**;

**data** long\_lead\_rand;

set long\_lead;

by ID week;

if (first.ID=**1**) then U=ranuni(**37**);

retain U;

**run**;

**Proc** **SGplot** data = long\_lead\_rand;

where U>**0.90**;

series x=week y=PB / group =ID LineAttrs= (pattern=**1** color="black");

**run**;



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\* Example Two (Unbalanced data) \*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth. A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian.;

**data** air\_pol;

input ID Height Age INI\_Height INI\_Age Log\_FEV1;

datalines;

1 1.20 9.3415 1.20 9.3415 0.21511

1 1.28 10.3929 1.20 9.3415 0.37156

1 1.33 11.4524 1.20 9.3415 0.48858

1 1.42 12.4600 1.20 9.3415 0.75142

1 1.48 13.4182 1.20 9.3415 0.83291

1 1.50 15.4743 1.20 9.3415 0.89200

1 1.52 16.3723 1.20 9.3415 0.87129

2 1.13 6.5873 1.13 6.5873 0.30748

2 1.19 7.6496 1.13 6.5873 0.35066

………………………………….Data omitted ………………………………

;

**run**;

\*First, we'll take a couple of snapshots of the data;

**Proc** **sgplot** data = air\_pol;

scatter x=Age y=Log\_FEV1 / group =ID markerattrs=(color=black symbol=plus size=**5**) ;

**run**;



**Proc** **SGplot** data = air\_pol;

series x=Age y=Log\_FEV1 / group =ID LineAttrs= (pattern=**1** );

**run**;



\*Plotting a mean line for unbalanced data is a bit more challenging and requires a loess line smoother;

**proc** **loess** data=air\_pol plots=none;

ods output outputstatistics=out\_pol;

model Log\_FEV1=Age;

**run**;

| **Independent Variable Scaling** | |
| --- | --- |
| **Scaling applied: None** | |
| **Statistic** | **Age** |
| **Minimum Value** | 6.43390 |
| **Maximum Value** | 18.69130 |

| **Optimal Smoothing Criterion** | |
| --- | --- |
| **AICC** | **Smoothing Parameter** |
| -2.82204 | 0.30667 |

\*Note: from here on we use the dataset "out\_pol" and "DepVar" instead of "Log\_FEV1";

**proc** **sort** data=out\_pol;

by Age DepVar;

**run**;

**proc** **sgplot** data=out\_pol;

scatter x=Age y=DepVar/ markerattrs=(color=black size=**5**) ;

series x=Age y=pred/ lineattrs=(color=blue thickness=**5**) ;

**run**;

