Chapter 17. Longitudinal Data Analysis

17.1. Longitudinal Study[11]

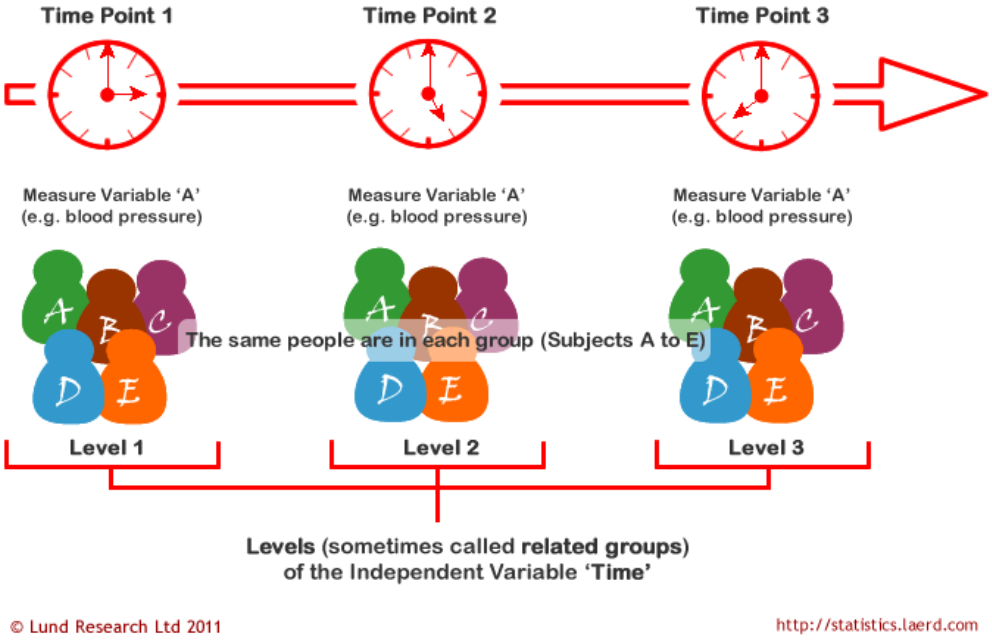
- Participant outcomes (and possibly treatments or exposures) are collected at multiple *follow-up* times (repeated measurements at multiple time points).
- Verify inter-individual differences and how they influence response over time.
- Repeated observations at the individual level exclude the time-invariant unobserved individual differences and observe the temporal order of events (cf. cross-sectional study).
- Cohort: A group of subjects who have shared a particular characteristic or experience during a particular time span
- Prospective (follow-up) / Retrospective (back in time)

[11] See Chapter 10 for manipulating and visualizing the longitudinal data.

- Benefits
  - The timing of disease onset [event] can be correlated with recent changes in exposure and/or with chronic exposure.
  - Multiple follow-up measurements can alleviate recall bias.
  - Measuring the change in outcomes at the individual level provides the opportunity to observe individual patterns of change.
  - The cohort under study is fixed, so changes in time are not confounded by cohort differences. (i.e. Separate cohort and age effects.)
- Challenges
  - There is a risk of bias due to incomplete follow-up or dropout of study participants.
  - Analyzing the correlated data requires a method that can properly account for the intra-subject correlation of response measures.
  - The direction of causality can be complicated by feedback between outcome and exposure (time-varying covariates).

17.2. Longitudinal Data Analysis

- Assuming independence between observations are not appropriate.
  - Measurements within a subject are dependent.
  - Measurements between subjects can be independent.



- Notation
  - Number of subjects: $i = 1, 2, \cdots I$
  - Number of repeated measurements: $j = 1, 2, \cdots J_i$
  - Times of measurement: $t_{ij},\ i = 1, 2, \cdots I,\ j = 1, 2, \cdots J_i$
  - Outcome measured on subject i at time t_{ij}: $y_{ij},\ i = 1, 2, \cdots I,\ j = 1, 2, \cdots J_i$
- Model
  - For subject i, $Y_i = X_i\beta + \varepsilon_i$, where $Var(\varepsilon_i) = \sigma^2 V_i$, $i = 1, 2, \cdots I$.
  - Incorporating all $i = 1, 2, \cdots I$,

$$Y = X\beta + \varepsilon,\ \ Var(\epsilon) = \sigma^2 V = \sigma^2 \begin{bmatrix} V_1 & & \cdots & & 0 \\ & V_2 & & & \\ \vdots & & \ddots & & \vdots \\ & & & V_{I-1} & \\ 0 & & \cdots & & V_I \end{bmatrix}$$

  - The covariates X can be either 1) fixed at the subject level or 2) time-varying.

- Correlation structure[12]

Correlation structure	Description
Unstructured	All elements are unconstrained. ($Var(Y_i) = \Sigma_i = \Sigma$) $n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$ parameters
Compound symmetry (Exchangeable)	$Var(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$
Toeplitz	$Cov(Y_{ij}, Y_{i(j+k)}) = \rho_k$; $1 + (n-1) = n$ parameters $Var(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_1 & 1 \end{bmatrix}$
Banded	Correlation is zero beyond the certain time interval k. i.e. $Cov(Y_{ij}, Y_{i(j+l)}) = 0$ for $l \geq k$ e.g. $k = 2$; $Var(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \rho_1 & 1 \end{bmatrix}$

- Other examples: AR(p), MA(q), exponential

[12] Assume $j = 1, 2, \cdots, n$ for the ith subject

17.3. Random Effects Model vs Generalized Estimating Equation

Approach	Description
Random effects model	- Incorporate correlation structure in the model by introducing random quantities into the mean. - Introduce random subject effects, coming from a distribution. - Change the intercept/slope of the model between individuals. - Estimate between and within subject variance components.
Generalized estimating equation (Marginal model)	- Model the correlation directly. - Separate the mean structure and correlation. - Focus on estimating the main effects (population average effect) and variance matrices. - Estimate a within-subject variance and a covariance matrix.

- Both random effects model and GEE partition total variability into

  1) subject-level and 2) population-level variance.

- Fundamental difference between random effects model and GEE is in the interpretation of the coefficients.
- GEE is "robust": Provide valid asymptotic confidence intervals of β even if the correlation structure in the model is miss-specified through robust SE estimates.

17.4. Random Effects Model

- Fixed effects & random effects

Fixed Effects	Random Effects
Constant across individuals.	Vary across individuals.
Levels of each factor are fixed in advance.	Levels of factor are meant to be representative of a general population of possible levels.
Estimated using least squares or maximum likelihood.	Estimated with shrinkage.
Marginal interpretation	Conditional interpretation

- Mixed effects model if a model contains both fixed and random effects.
- Random effects models are similar mathematically to introducing penalization.
- Using randomness 1) decreases the number of parameters and 2) induces correlation structures.

- Random intercept model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \varepsilon_{ij}$$

subject specific

where $b_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.[13]

  - Random subject effects (intercept) b_i introduces heterogeneity.
  - Assume correlated subject-level errors.
  - Induce a compound symmetry (exchangeable) within-subject correlation structure.

$$Var(Y_i) = \begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 & \sigma_b^2 + \sigma^2 \end{bmatrix} = (\sigma_b^2 + \sigma^2) \begin{bmatrix} 1 & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$$

where $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$.

[13] This is a Gaussian case. Generalized random effects model involves a link function similar to GLM.
The compound symmetry correlation structure will not be induced for any other generalized random effects models.

- Random slope model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i x_{ij} + \varepsilon_{ij}$$

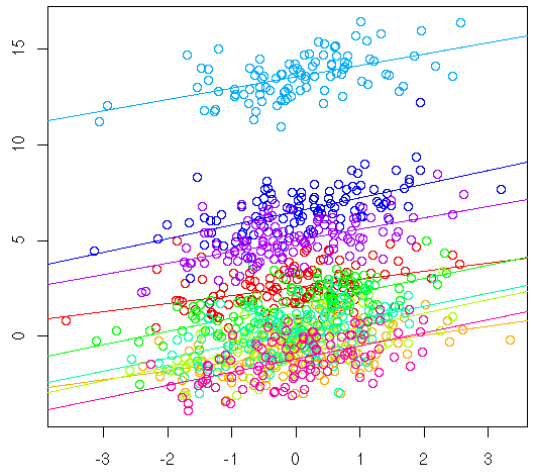
where $b_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Random intercept & slope model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + \varepsilon_{ij}$$

where

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b0}^2 & \sigma_{01} \\ \sigma_{01} & \sigma_{b1}^2 \end{pmatrix}\right] \text{ and } \quad \varepsilon_{ij} \sim N(0, \sigma^2).$$



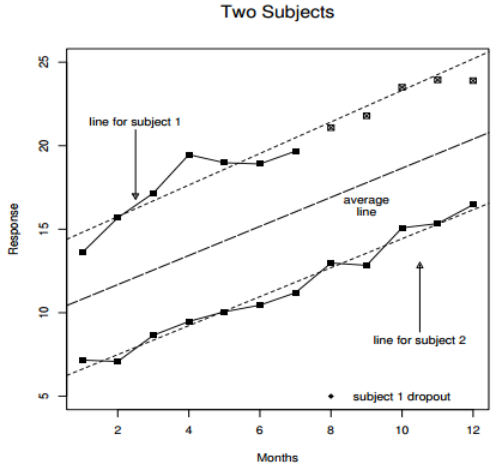
subject specific effect of X on Y

- Interpretation

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + \varepsilon_{ij}$$

  - Time-varying covariates: Effect for an average subject (within-subject covariates)
  - Time-invariant covariates: Individual sharing similar characteristics (between-subject covariates)
  - $E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}$
  - $E(Y_{ij} | b_{0i}, b_{1i}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) x_{ij}$

    (i.e. b_{0i}, b_{1i}: Effect for a particular subject conditional on the random effects)



- Interclass Correlation Coefficient (ICC)

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

ALways less than 1

  - Quantify how strongly observations within a subject resemble each other.
  - Proportion of variability explained by within-subject variation
  - Assess the consistency or reproducibility of quantitative measurements made via multiple visits or by different observers measuring the same quantity.

- PROC MIXED

General Syntax

```
proc mixed data=dataset;
     model dependent-variable = list-of-independent-variables;
     random random-effects <options>;
     repeated repeated-effect <options>;
run;
```

Statement	Description
RANDOM	Specify the effects in the model that represent repeated measurements and impose a particular covariance structure. Multiple RANDOM statements can be added; Effects in the same statement may be correlated, but independent in different statements.
REPEATED	Specify the random effects and their covariance structures. Control the covariance structure of the residuals.

Option	Description
TYPE = `structure`	Specify the covariance structure. TYPE = VC (default), AR, TOEP, UN, CS
SUBJECT = `effect`	Identify the subjects in the mixed model. Complete independence is assumed across subjects.

17.5. Generalized Estimating Equation (GEE)

- Model[14] Mixed effects model is more sensitive to the interpretation: be cautious about wording.

$$Y = X\beta + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma^2 V)$$

Known to be robust, even the assumption is wrong, the conclusion can be trusted.

  - Focus on the marginal distribution (population-averaged effect) of Y rather than on a subject-level conditional distribution.
  - Coefficients are interpreted marginally: Compare subjects based on covariate values.
- Consistent irrespective of the true underlying correlation structure
- Limitations
  - Difficult to assess the goodness-of-fit models due to lack of an inference function
  - Parameter estimates are sensitive to the presence of outliers.
  - Non-convergence and multiple roots problem

[14] This is a Gaussian case. Generalized marginal model involves a link function similar to GLM.

Example

Raw Data

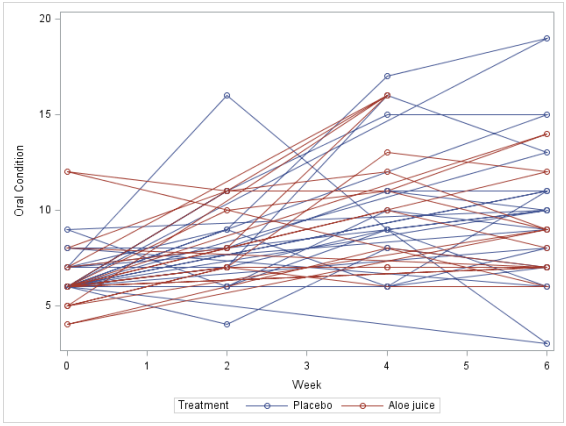
Obs	ID	Treatment	AGE	Initial weight	STAGE	Oral condition in week 0	Oral condition in week 2	Oral condition in week 4	Oral condition in week 6
1	1	Placebo	52	124	2	6	6	6	7
2	5	Placebo	77	160	1	9	6	10	9
3	6	Placebo	60	136.5	4	7	9	17	19
4	9	Placebo	61	179.6	1	6	7	9	3
5	11	Placebo	59	175.8	2	6	7	16	13

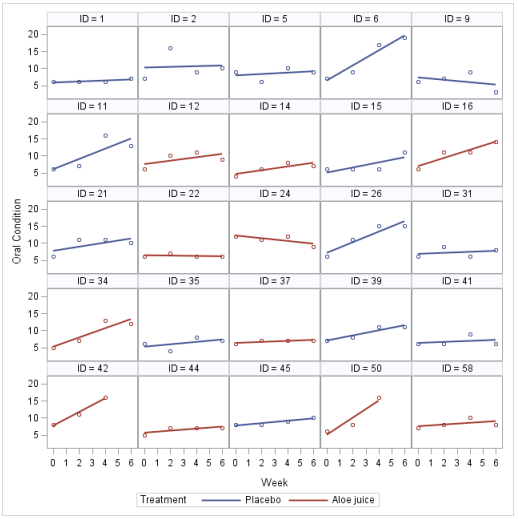
SAS Code

```
* Spaghetti plot;
proc sgplot data=Cancer_long;
series x=week y=totalc /
markers group=trt;
xaxis label="Week";
yaxis label="Oral Condition";
run;

* Individuals;
proc sgpanel data=Cancer_long;
panelby ID / columns=5 rows=5;
reg x=week y=totalc
/group=trt;
colaxis label="Week";
rowaxis label="Oral
Condition";
run;
```

Output



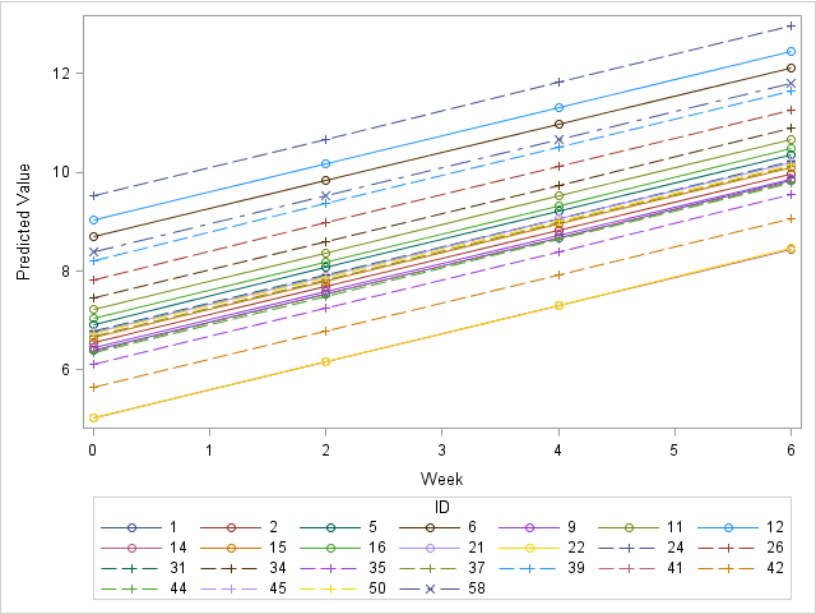


```
* PROC MIXED;
proc mixed data=cancer_long;
class id trt(ref="Placebo");
model totalc = stage weightin
age trt week / solution
outpm=pred;
random intercept / subject=id
type=un;
run;

proc sgplot data=pred;
series x=week y=pred / markers
group=id;
xaxis label="Week";
yaxis label="Predicted Value";
run;
```

Solution for Fixed Effects

Effect	Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		1.0495	3.7169	20	0.28	0.7806
STAGE		0.9116	0.3282	72	2.78	0.0070
WEIGHTIN		0.01032	0.01402	72	0.74	0.4639
AGE		0.04306	0.03266	72	1.32	0.1915
TRT	Aloe juice	-0.4264	0.8543	72	-0.50	0.6192
TRT	Placebo	0	.	.	.	.
WEEK		0.5718	0.1099	72	5.20	<.0001

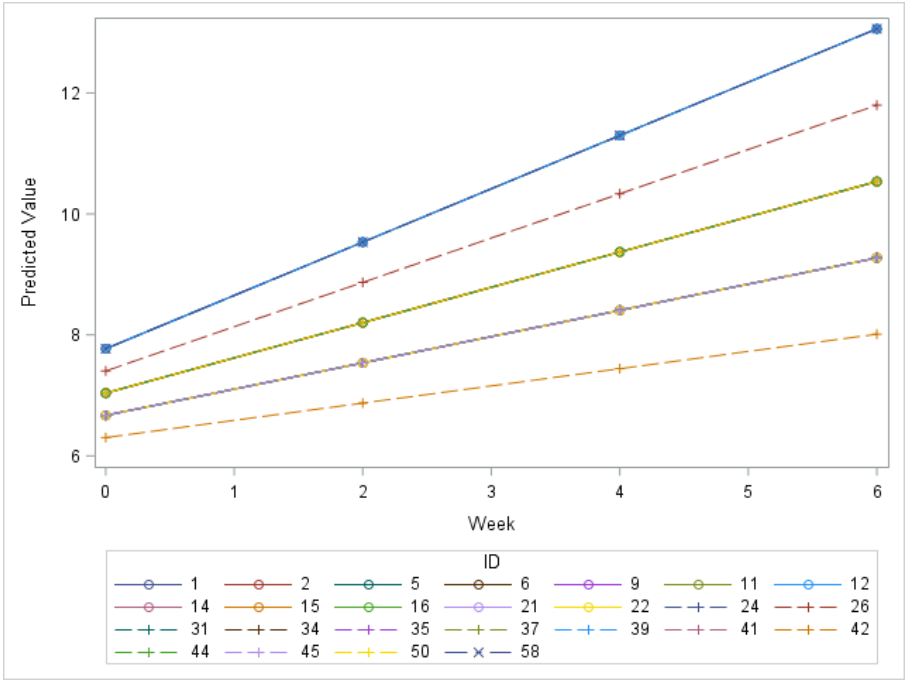


```
* Interaction;
proc mixed data=cancer_long
covtest;
model totalc = stage|week /
solution notest outpm=pred;
     random intercept /
subject=id type=un;
run;

proc sgplot data=pred;
     series x=week y=pred /
markers group=id;
     xaxis label="Week";
     yaxis label="Predicted
Value";
run;
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	6.3027	0.9025	23	6.98	<.0001
STAGE	0.3671	0.4031	71	0.91	0.3655
WEEK	0.2848	0.2053	71	1.39	0.1697
STAGE*WEEK	0.1492	0.09059	71	1.65	0.1041

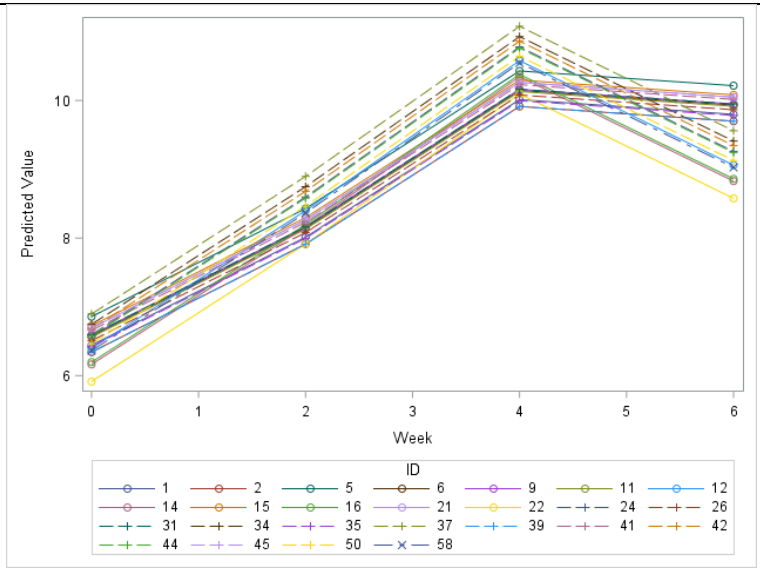


```
* Discrete time;
* Interaction: trt, age;
proc mixed data=cancer_long2
covtest;
class trt(ref="Placebo")
week(ref="0");
model totalc = age trt|week /
solution notest outpm=pred;
random intercept / subject=id
type=un;
run;
```

Solution for Fixed Effects

Effect	Treatment	WEEK	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			5.5762	2.1694	22	2.57	0.0175
AGE			0.01665	0.03383	67	0.49	0.6243
TRT	Aloe juice		-0.1114	1.1826	67	-0.09	0.9253
TRT	Placebo		0	.	.	.	.
WEEK		2	1.5714	0.8775	67	1.79	0.0778
WEEK		4	3.5714	0.8775	67	4.07	0.0001
WEEK		6	3.3571	0.8775	67	3.83	0.0003
WEEK		0	0	.	.	.	.
TRT*WEEK	Aloe juice	2	0.4286	1.3228	67	0.32	0.7470
TRT*WEEK	Aloe juice	4	0.6104	1.3228	67	0.46	0.6460
TRT*WEEK	Aloe juice	6	-0.6938	1.3720	67	-0.51	0.6148
TRT*WEEK	Aloe juice	0	0	.	.	.	.
TRT*WEEK	Placebo	2	0	.	.	.	.
TRT*WEEK	Placebo	4	0	.	.	.	.
TRT*WEEK	Placebo	6	0	.	.	.	.
TRT*WEEK	Placebo	0	0	.	.	.	.

```
proc sgplot data=pred;
series x=week y=pred / markers
group=id;
xaxis label="Week";
yaxis label="Predicted Value";
run;
```



```
* GEE with exchangeable
covariance matrix;
proc genmod data=cancer_long;
class id trt(ref="Placebo");
model totalc = age trt stage
week;
repeated  subject=id /
type=exch covb corrw;
run;
```

Working Correlation Matrix				
	Col1	Col2	Col3	Col4
Row1	1.0000	0.1837	0.1837	0.1837
Row2	0.1837	1.0000	0.1837	0.1837
Row3	0.1837	0.1837	1.0000	0.1837
Row4	0.1837	0.1837	0.1837	1.0000

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		3.3137	1.4753	0.4222	6.2053	2.25	0.0247
AGE		0.0349	0.0226	-0.0093	0.0791	1.55	0.1219
TRT	Aloe juice	-0.1900	0.7530	-1.6658	1.2858	-0.25	0.8008
TRT	Placebo	0.0000	0.0000	0.0000	0.0000	.	.
STAGE		0.8927	0.3214	0.2628	1.5227	2.78	0.0055
WEEK		0.5671	0.1292	0.3138	0.8204	4.39	<.0001