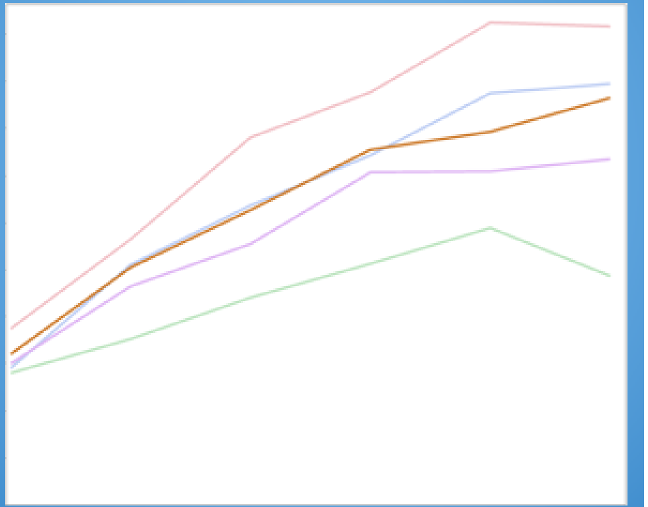
Mixed Models Part 2 - Handling Repeated Measures in Time and Space

Elizabeth A Claassen, PhD
Research Statistician Developer, JMP

Statistical Discovery.™ From SAS.

Regression, Random Coefficients, and Multilevel Models



What we will cover:

- What happens if the slope and intercept are correlated?
- Relationship between models with different names

Simple Linear Regression

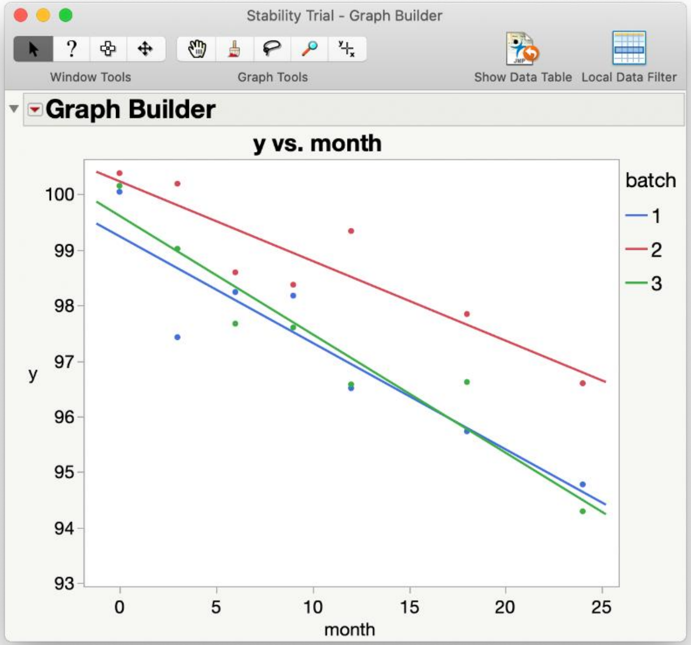
- When we have a continuous predictor, *x*, and a continuous response, *y*, it is time for classic simple linear regression.
- Back in our Geometry classes, we would describe this relationship as $y = mx + b$.
- Statisticians love Greek letters and reinventing the wheel, so the equation is $y = \mu + \beta x$, but the meanings of μ and β are the same as *b* and *m*.
- But in Geometry, we were usually only using two points to define the line, and in Statistics, we (hopefully!!) have more!
- The Least Squares algorithm is used to fit the best line between the points observed.

Extending Regression

- Multiple predictors
  - A factorial treatment design! (Maybe with, maybe without interactions)
  - Referred to as Multiple Regression
- Curvature in the response over the predictor – polynomial regression
- Multiple “subjects” measured
  - Mixed Model territory
    - Random Coefficient models
    - Hierarchical Linear Models
  - Subject-specific regressions (just BLUPs!)
  - Correlation between intercept and slope possible

Stability Trial

- A manufacturer wants to determine the shelf life of a new product.
- They sample 3 batches over several pre-determined times measuring ‘fizziness’ at each time.
- The goal is to find the time when fizziness drops below ‘acceptable’ limits – 90 on the fizziness scale.



Skeleton ANOVA

Experiment Design		Treatment Design		Skeleton ANOVA	
Source	df	Source	df	Source	df
Batch	3-1=2			Batch	2
		Month	1	Month	1
				Batch*Month	2
Obs(Batch)	(7-1)*3=18			Obs(Batch)|Month -> Residual	15
Total	21-1=20			Total	20

ANOVA to Model

Skeleton ANOVA	
Source	df
Batch	2
Month	1
Batch*Month	2
Obs(Batch)|Month-> Residual	15
Total	20

$Y_{ij}=\beta_0+b_{0i}+\beta_1 m_{ij}+b_{1i}m_{ij}+e_{ij}$

Y_{ij} is the j^{th} obs. on the i^{th} batch

β_0 and β_1 are the overall intercept and slope, respectively

m_{ij} is the month of the j^{th} obs. on the i^{th} batch

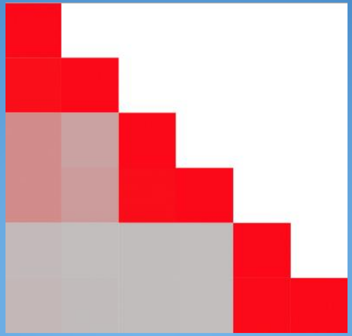
b_{0i} is the batch-specific intercept

b_{1i} is the batch-specific slope

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right)$$

e_{ij} is the residual error and $\sim N(0,\sigma^2)$

Repeated Measures and Longitudinal Data

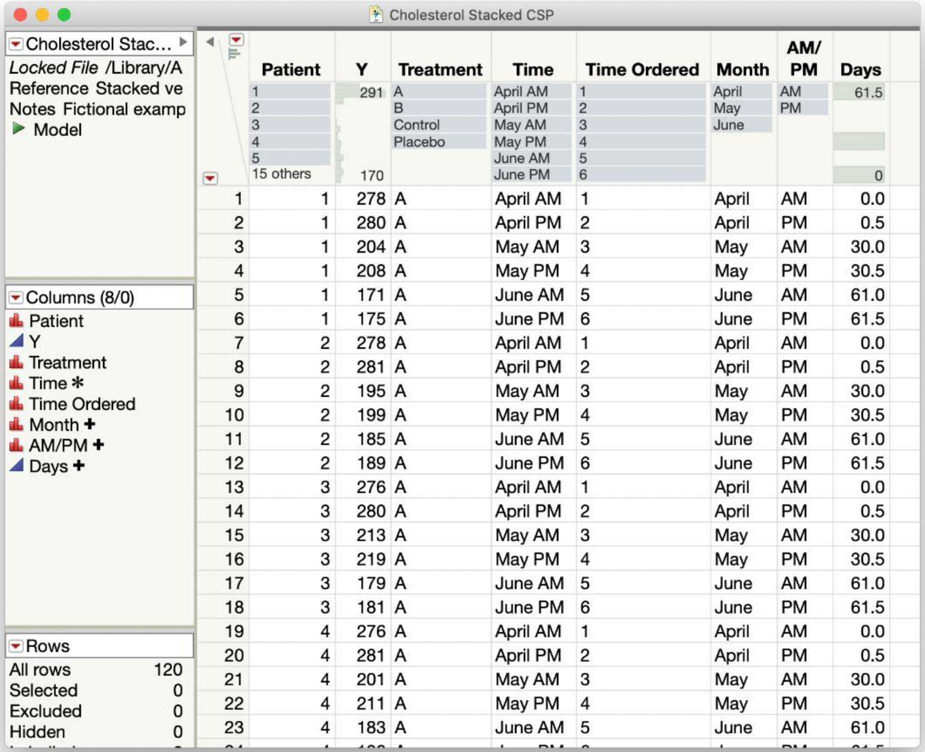


What we will cover:

- Addressing the lack of independence of observations made on the same subject over time
- Assessing the covariance and correlation between time points to select candidate structures
  - Equal variances?
  - Trend across time lags?

Cholesterol Measurements over Time

- In **Cholesterol Stacked CSP.jmp**, there are five subjects in four treatment groups, with measurements taken in the morning and afternoon, once a month, for three months.
- The goal is to fit a model for the response, Y, based on the Treatment and Time (which is composed of Month, and AM/PM).



	Patient	Y	Treatment	Time	Time Ordered	Month	AM/PM	Days
1	1	278	A	April AM	1	April	AM	0.0
2	1	280	A	April PM	2	April	PM	0.5
3	1	204	A	May AM	3	May	AM	30.0
4	1	208	A	May PM	4	May	PM	30.5
5	1	171	A	June AM	5	June	AM	61.0
6	1	175	A	June PM	6	June	PM	61.5
7	2	278	A	April AM	1	April	AM	0.0
8	2	281	A	April PM	2	April	PM	0.5
9	2	195	A	May AM	3	May	AM	30.0
10	2	199	A	May PM	4	May	PM	30.5
11	2	185	A	June AM	5	June	AM	61.0
12	2	189	A	June PM	6	June	PM	61.5
13	3	276	A	April AM	1	April	AM	0.0
14	3	280	A	April PM	2	April	PM	0.5
15	3	213	A	May AM	3	May	AM	30.0
16	3	219	A	May PM	4	May	PM	30.5
17	3	179	A	June AM	5	June	AM	61.0
18	3	181	A	June PM	6	June	PM	61.5
19	4	276	A	April AM	1	April	AM	0.0
20	4	281	A	April PM	2	April	PM	0.5
21	4	201	A	May AM	3	May	AM	30.0
22	4	211	A	May PM	4	May	PM	30.5
23	4	183	A	June AM	5	June	AM	61.0

From ANOVA to Model

One-to-one ANOVA source & Model Parameter

Experiment Design		Treatment Design		Skeleton ANOVA	
Source	df	Source	df	Source	df
“Arm”	4-1=3	Treatment	4-1 = 3	Treatment	3
Patient(Arm)	(5-1)*4=16			Patient(Treatment)	16
		Time	6-1 = 5	Time	5
		Treatment*Time	3*5 = 15	Treatment*Time	15
Measurements (Patient)	20*(6-1) = 100			Measurements|Time, Treatment*Time	100-20=80
Total	4*5*6 – 1 = 119			Total	119

Looks like split-plot????

Statistical Model

Skeleton ANOVA	
Source	df
Treatment	3
Patient(Treatment)	16
Time	5
Treatment*Time	15
Residual	100-20=80
Total	119

$y_{ijk} = \mu + \alpha_i + s_{j(i)} + \tau_k + (\alpha\tau)_{ik} + e_{ijk}$

y_{ijk} is the observation of the j^{th} Subject in the i^{th} Treatment at the k^{th} Time

μ is the intercept

α_i is the i^{th} Treatment effect

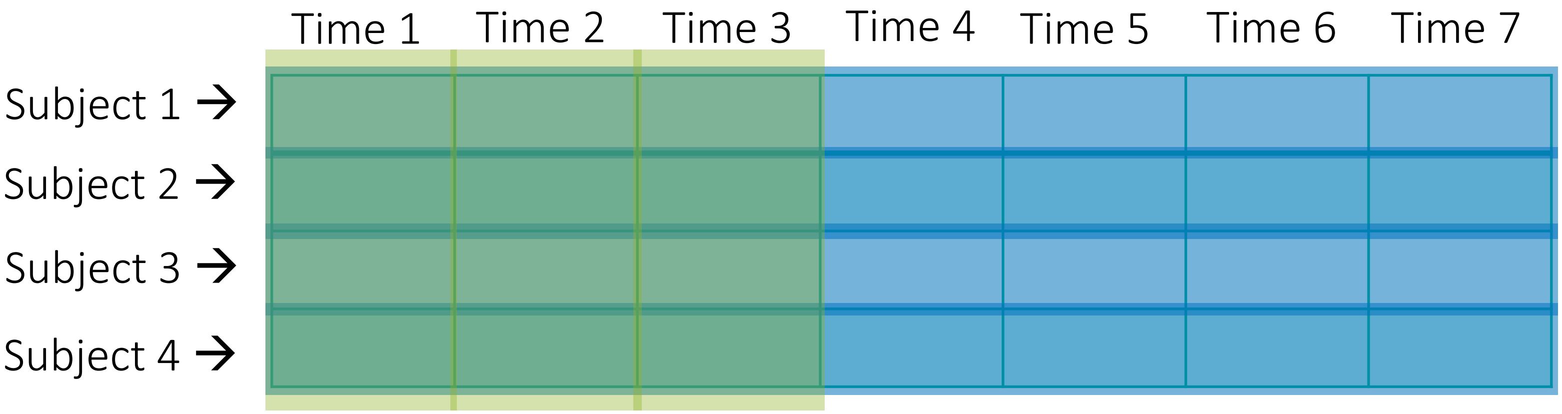
$S_{j(i)}$ is the random effect of the j^{th} Subject in the i^{th} Treatment and $s_{j(i)} \sim N(0,\sigma_s^2)$

τ_k is the effect of the k^{th} Time

$(\alpha\tau)_{ik}$ is the interaction effect of the i^{th} Treatment at the k^{th} Time

e_{ijk} is the residual error and $e_{ijk} \sim N(0,\sigma^2)$

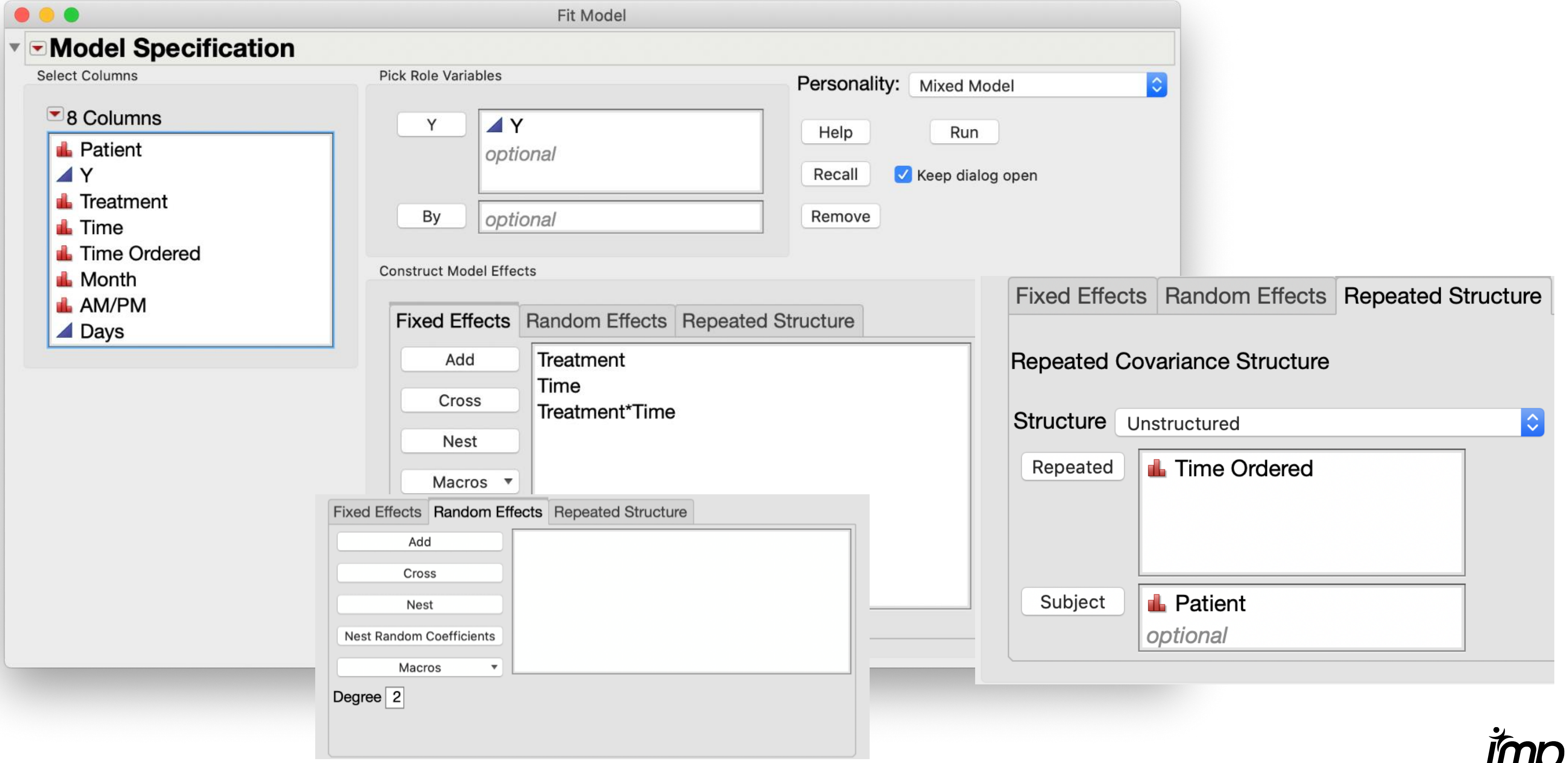
Repeated measures as a split-plot in time…



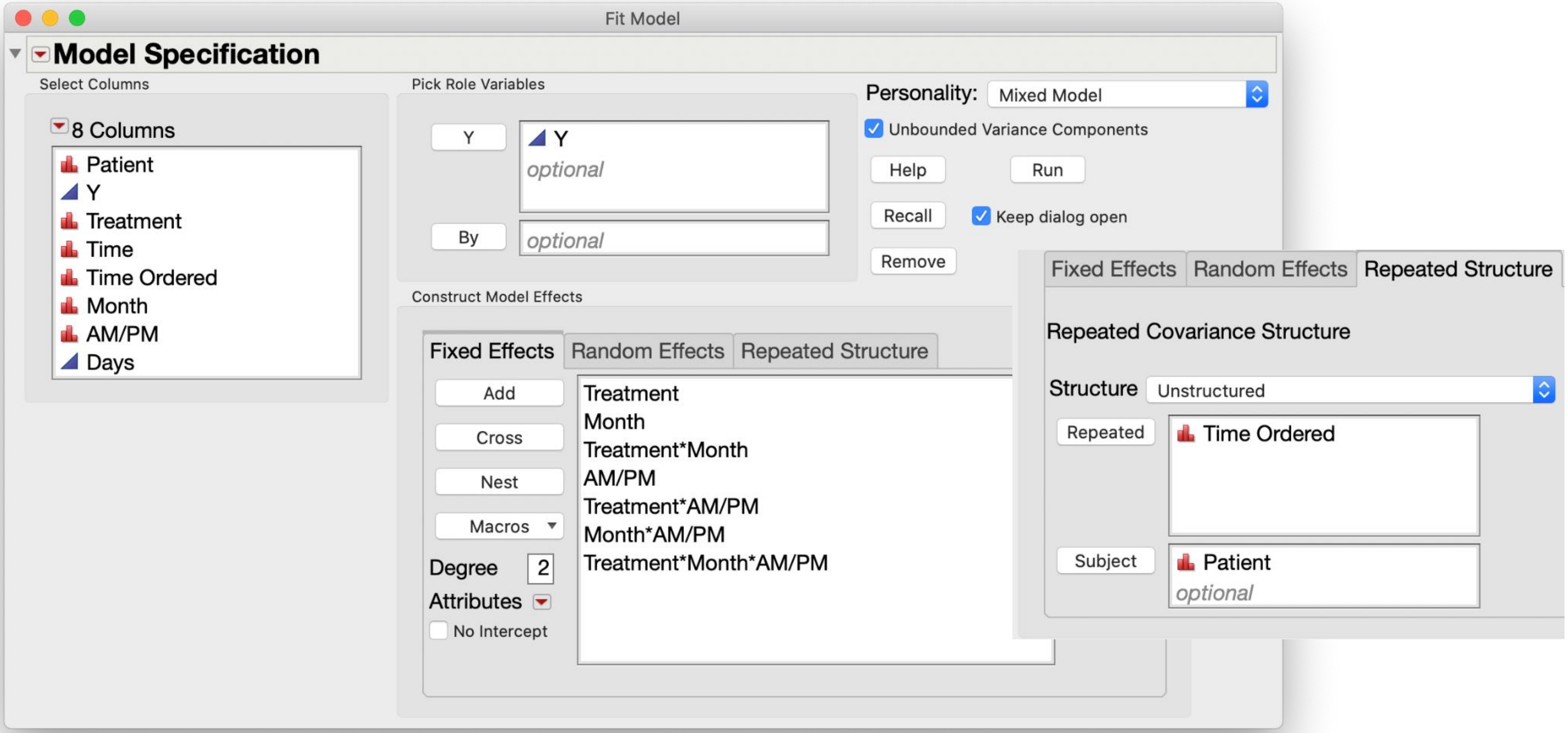
Can we randomize the order of the Subjects?

Can we randomize the order of Time?

Simple Model



All Treatment Interactions Included



Compound Symmetry

Correlation is the same everywhere

- CS is identical to a model with a random effect for the Subject and no other modeling of the correlation over time.
- I.e., CS is identical to Split-Plot-in-Time.
- This structure will fit well when the correlations between time points remain constant over any time lag.

Compound Symmetry Correlation Matrix

1	r	r	r
	1	r	r
		1	r
			1

Example CS Correlation Matrix

1	.97	.97	.97
	1	.97	.97
		1	.97
			1

AR(1)

Correlation decays over time gap size

- AR(1) holds the correlations constant for observations at any two time points of lag 1, and then allows that correlation to decay exponentially as the time lag increases.
- Many statistical software packages require the time points to be at equal intervals, but JMP allows unequal spacing in the time points.

Auto-Regressive(1) Correlation Matrix

1	r	r^2	r^3
	1	r	r^2
		1	r
			1

Example AR(1) Correlation Matrix

1	.90	.81	.73
	1	.90	.81
		1	.90
			1

Toeplitz

Correlation differs over time gap size, but without a pattern

- The Toeplitz pattern has one correlation for all of the lag 1 cells in the correlation matrix, a different correlation, unrelated to the lag 1 correlation, for all of the lag 2 cells, and so on.
- Each diagonal lag band has the same correlation throughout, and there is no trend from one band to the next.

Toeplitz Correlation Matrix

1	r_1	r_2	r_3
	1	r_1	r_2
		1	r_1
			1

Example Toep Correlation Matrix

1	.70	.95	-.60
	1	.70	.95
		1	.70
			1

Antedependent

Correlation differs at lag 1 gaps, but hold a pattern over time

- The Antedependent pattern is harder to see from the matrices.
- Use this when the lag 1 correlations are dissimilar (unlike Toeplitz or AR(1)) but you still have a pattern over time (like AR(1)).
- This structure also works well for unequal spacing for the time measurements.

Antedependent Correlation Matrix

1	r_{12}	$r_{12}r_{23}$	$r_{12}r_{23}r_{34}$
	1	r_{23}	$r_{23}r_{34}$
		1	r_{34}
			1

Example Ante Correlation Matrix

1	.70	-.28	-.16
	1	-.40	-.22
		1	.56
			1

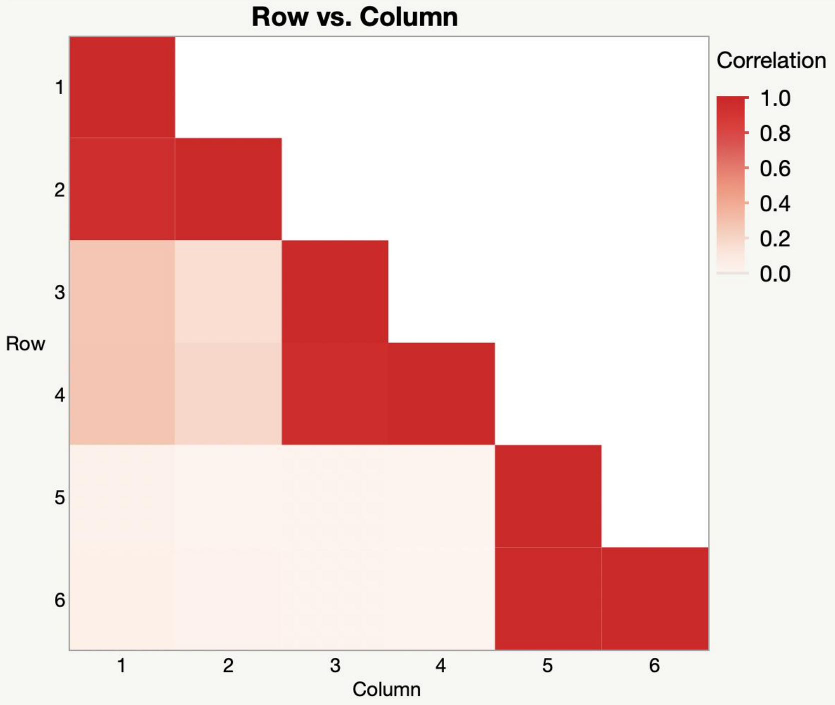
2. Possible Structures?

Correlation Matrix

Row	1	2	3	4	5	6
1	1	.	.	.	.	.
2	0.9662998	1	.	.	.	.
3	0.2849137	0.166762	1	.	.	.
4	0.2880302	0.1997435	0.974226	1	.	.
5	0.0346389	0.0091	0.0184582	0.010723	1	.
6	0.045121	0.021514	0.0152042	0.0106987	0.9898732	1

Correlation Heatmap

Graph Builder



Compound Symmetry?

AR(1)?

Toeplitz?

Antedependent?

Compound Symmetry Unequal / Toeplitz Unequal

Subject: Patient

Covariance Parameter	Estimate
Intraclass Correlation	0.2702698
Variance(1)	18.009356
Variance(2)	19.175518
Variance(3)	55.614518
Variance(4)	55.880836
Variance(5)	66.358405
Variance(6)	68.325809

Subject: Patient

Covariance Parameter	Estimate
Toeplitz Correlation(1)	0.6230225
Toeplitz Correlation(2)	0.2555783
Toeplitz Correlation(3)	0.0962006
Toeplitz Correlation(4)	-0.1112
Toeplitz Correlation(5)	0.2427218
Variance(1)	17.552396
Variance(2)	24.652754
Variance(3)	73.950898
Variance(4)	63.668022
Variance(5)	69.45679
Variance(6)	52.484909

Antedependent Unequal / AR(1)

Subject: Patient

Covariance Parameter	Estimate
Correlation(1, 0)	0.9662988
Correlation(2, 1)	0.1667594
Correlation(3, 2)	0.974226
Correlation(4, 3)	0.010723
Correlation(5, 4)	0.9898732
Variance(1)	18.724233
Variance(2)	19.268185
Variance(3)	56.60329
Variance(4)	57.058035
Variance(5)	63.512277
Variance(6)	65.568482

Subject: Patient

Covariance Parameter	Estimate
AR(1) Days	0.9536532
Residual	44.579921

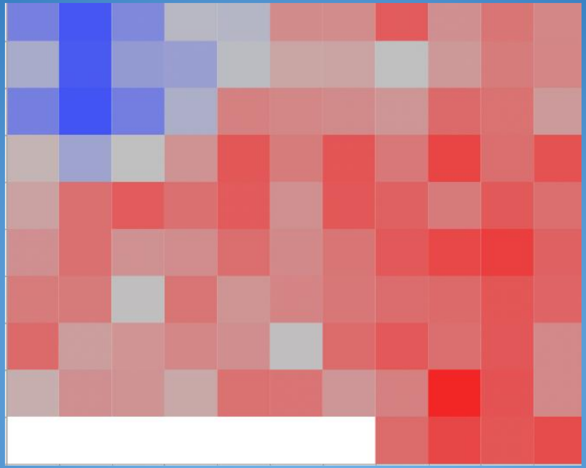
Fit of Repeated Structures

Repeated Structure	# Repeated Parameters	AICc	BIC
CS – Unequal Variances	7	832.8	896.7
CS – Equal Variances	2	832.6	889.9
Toeplitz – Unequal Variances	11	788.0	855.6
Toeplitz – Equal Variances –> did not converge	6		
Unstructured	21	703.8	773.3
Antedependent – Unequal Variances	11	670.1	737.7
Antedependent – Equal Variances	6	659.3	722.0
AR(1)	2	652.6	710.0

Summary

- Need to account for special way that the data are not independent.
- Split-plot-in-time (i.e., Compound Symmetry) is likely an oversimplification.
- Start with the Unstructured to look for patterns:
  - Equal or Unequal Variances across time points?
  - Which candidate correlation structures across time lags?
- Use Fit Statistics and interpretability to choose best structure.

Spatial Models



Modeling Correlation in Space

- Whether in a field, a greenhouse, or on a silicon wafer, observations taken ‘nearby’ each other are often correlated with each other.
- This is why blocking was created!
- But what if blocking doesn’t work, or treatments aren’t being applied?
- Spatial correlation structures, similar to the repeated measures correlation structures, can be used.

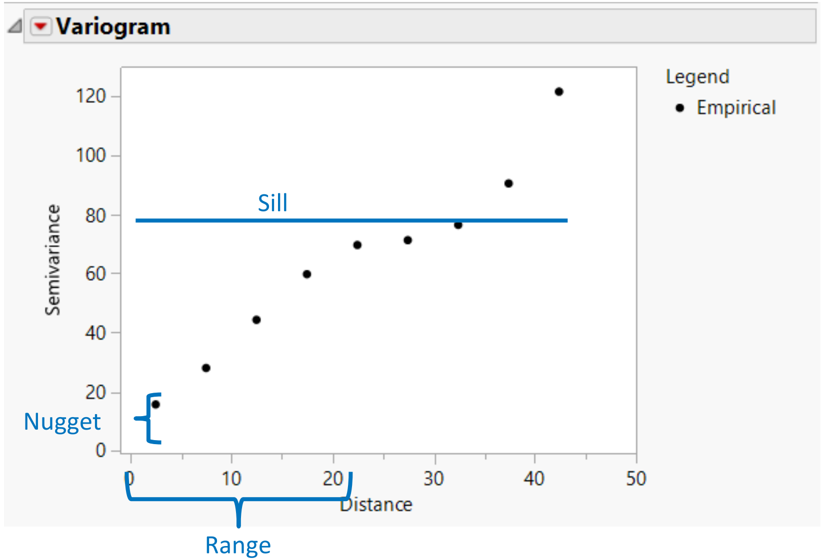
Spatial Covariance Structures

- The covariance/correlation is a function if the distance between the two observations.
- The AR(1) repeated measures structure is a simplified version of the Spatial Power structure.
  - The distance function for AR(1) is $\rho^{|t_i - t_j|}$ - limited to a single dimension
  - Generalizing this to any distance metric d_{ij} yields $\rho^{d_{ij}}$.
  - The d_{ij} distance metric can be multidimensional
- Other functions of the distance metric can be used.
  - Gaussian
  - Exponential
  - Spherical

Special Spatial Terminology

Geostatistics

- Variogram – graphical display of the semivariance as distance increases.
- Nugget – "jump" in semivariance at small distances
- Sill – plateau of the semivariance
- Range – distance to the Sill



Hazardous Waste Example

- Water drainage is important when choosing a storage site for hazardous waste.
- At a potential site, the thickness of a naturally occurring layer of salt could affect water movement.
- The relationship between the salt layer and water movement is believed to be linear.
- Thirty samples were taken at various locations measuring:
  - Salt thickness (covariate)
  - Log-transmissivity of the water
  - North-south (northing) and east-west (easting) coordinates for the location

Skeleton ANOVA

Experiment Design		Treatment Design		Skeleton ANOVA	
Source	df	Source	df	Source	df
		Salt	1	Salt	1
Sample	29			Sample|Salt -> Residual	28
Total	(30-1)=29			Total	29

Skeleton ANOVA	
Source	df
Salt	1
Sample|Salt – Residual	28
Total	29

$Y_i=\beta_0+\beta_1X_i+e_i$

Y_i is log-transmissivity of the i^{th} sample.

β_0 is the intercept.

β_1 is the slope and X_i is the observed salt thickness of the i^{th} sample.

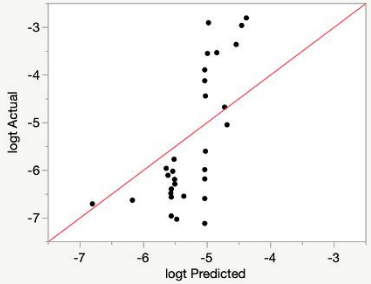
e_i is the residual error and $\mathbf{e}$~N($\mathbf{0}$,$\mathbf{R}$), where $\mathbf{R}$ is a spatial covariance structure.

Exponential Structure Fit

hazardous waste - Fit Mixed

Fit Mixed

Actual by Predicted Plot



Fit Statistics

-2 Residual Log Likelihood	79.486199
-2 Log Likelihood	72.352991
AICc	81.952991
BIC	85.95778

Repeated Effects Covariance Parameter Estimates

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
Spatial Exponential	9.2723023	8.977076	-8.322443	26.867048
Residual	2.5314239	2.0767333	0.8092687	35.89666

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	-5.028738	0.8951974	0.7	-5.62	0.1732	-32.38914	22.331666
salt	-0.017637	0.0058086	24.3	-3.04	0.0056*	-0.029619	-0.005655

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
salt	1	1	24.3	9.2193263	0.0056*

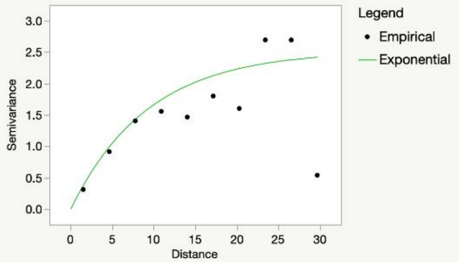
Key Statistics

- Covariance Parameters
  - Spatial Exponential = Range = 9.27
  - Residual = Sill = 2.53

hazardous waste - Fit Mixed

Fit Mixed

Variogram



- Fixed Effects
  - Intercept = -5.029
  - Salt = -0.0176
  - $\widehat{logt} = -5.029 - 0.0176 * salt$

Wrap-Up

- Regression, Random Coefficients, and Multilevel Models
- Repeated Measures
- Spatial

- In regression, intercepts and slopes may be correlated within subjects. Include correlation in model for best results.

- Measures over time on the same subject are correlated
- Time can’t be randomized – is there a trend in correlation over time?

- Measurements taken near in space are correlated

End of Part Two
Thank you!

jmp.com

Statistical Discovery.™ From SAS.