Over-Specified and Under-Specified Correlation Structures: A Simulation Case-Study using SAS

1. Introduction

Mixed Effects models are almost always used to analyze correlated data. One of the most common instances of correlated data can be found in longitudinal datasets where there are both between and within subject correlations. When we fit mixed effect models to analyze said longitudinal data, we assume there are fixed effects as well as random effects and, when we want to estimate fixed effects, we as users need to specify a variance-covariance (var-cov) structure for the random effects. This user specified var-cov structure for the random effects ($\mathbf{R}_i$) can be mis-specified and this can lead to biased coefficient estimates, biased standard errors, and incorrect statistical significance, of the fixed effect parameters. This will lead to Type I and Type II errors, as well.

In this simulation project, I will be looking at the impact of fitting models with different var-cov structures to datasets which are also simulated with different var-cov structures. The associated research question for this project is the following – "How do results of fixed estimates parameters vary when we fit models correctly or incorrectly to simulated data with different var-cov structures?". Specifically, we are interested in looking at the changes in the estimation of fixed effects, the standard error estimates of the fixed effects and the Type I or Type II error for the fixed effects hypothesis testing, when the correlation is over- and under-specified. Over-specification happens when you assume in model fitting, a correlation structure that is more complex than what exists in truth. Under-specification happens when you assume in model fitting, a correlation structure that is simpler than what exists in truth.

For example, assume that the data we have has an uncorrelated var-cov structure. However, when we fit a model to analyze this data, we fit a mixed model with an unstructured var-cov structure. This is an example of an over-specification in model fitting, since we fit a complex var-cov structure to the data than what is required. Again, assume that the data we have has an unstructured var-cov structure. However, when we fit a model to analyze this data, we fit a mixed model with a compound symmetry var-cov structure. This is an example of an under-specification in model fitting, since we fit a simpler var-cov structure to the data than what is actually required.

2. Model Notation

In this project, I will be investigating 3 main cases of simulation, each of which has multiple sub cases. The following general structure will be followed for simulating with all the three cases. The covariates in each model will have a Treatment variable, which is an indicator of 0 = no treatment and 1 = Treatment, a continuous Time variable which is linearly increasing, with time going from 0 to 4 as integers or 0 to 9 as integers depending on the corresponding subcase and, an interaction term between Time and Treatment. The treatment variable will represent a between-subject covariate and the time variable will represent a within subject covariate.

The corresponding linear model for the fixed effect components of each of the subcases is given as follows:

$$Y_{ij} = \beta_0 + \beta_1 Treatment_i + \beta_2 Time_{ij} + \beta_3 Treatment_i * Time_{ij} + \mathcal{E}_{ij}$$

where,

$\mathcal{E}_{ij} \sim N(0, \sigma_e^2)$ and i refers to a subject and j refers to an observation within a subject. I will assume the following values of the beta coefficients to use for simulating the datasets: $\beta_0 = 0.5,\ \beta_1 = 1, \beta_2 = 0.5$ and $\beta_3 = 0.5$. The model error variance σ_e^2 will be assumed to be equal to 1, wherever appropriate. The correlation parameter value will be assumed to be equal to 0.6.

The Mixed Effects Model used for this simulation can be denoted by the following matrix notation:

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{E_i}$$

where,

$\boldsymbol{b_i} \sim N(0, G)$ and $\boldsymbol{E_i} \sim N(0, \sum e)$

$X_i\boldsymbol{\beta}$ is the fixed effect, $Z_i\boldsymbol{b_i}$ is the random effect, and $\boldsymbol{E_i}$ is the error term.

For the purpose of this simulation, I will extend the $\mathbf{E}_i$ to include correlation. Hence, random effects will not be explicitly entered into the model. Therefore, I will be simulating data for each case directly from the var-cov matrices which will either be a compound symmetry (CS) or a first order autoregressive process (AR(1)) or a unstructured (UN) var-cov structure.

3. Simulation

Cases:

The three main cases of simulation can be summarized as follows,

Case 1: Data is simulated with a Compound Symmetry (CS) correlation structure (aka with a Random Intercept).

Case 2: Data is simulated with an Autoregressive (AR (1)) correlation structure.

Case 3: Data is simulated with an Unstructured (UN) correlation structure.

In Table 1, I am summarizing the cases and each of their subcases and their corresponding number of subjects and number of observations within each subject. This table also shows the models to be fitted for each of these subcases.

Table 1: Summary of Cases to be analyzed.

Cases/ Sub-Cases	Subjects (i)	Observations (j)	Time (ij)	Var-Cov Structure	Model Fit
		CASE 1			
Case 1.1	20	1,2, …, 5	0, 1, …, 4	CS	UC, CS, UN
Case 1.2	20	1,2, …, 10	0, 1, …, 9	CS	UC, CS, UN
Case 1.3	200	1,2, …, 5	0, 1, …, 4	CS	UC, CS, UN
Case 1.4	200	1,2, …, 10	0, 1, …, 9	CS	UC, CS, UN
		CASE 2			
Case 2.1	20	1,2, …, 5	0, 1, …, 4	AR (1)	UC, CS, UN
Case 2.2	20	1,2, …, 10	0, 1, …, 9	AR (1)	UC, CS, UN
Case 2.3	200	1,2, …, 5	0, 1, …, 4	AR (1)	UC, CS, UN
Case 2.4	200	1,2, …, 10	0, 1, …, 9	AR (1)	UC, CS, UN
		CASE 3			
Case 3.1	20	1,2, …, 5	0, 1, …, 4	UN	UC, CS, UN
Case 3.2	20	1,2, …, 10	0, 1, …, 9	UN	UC, CS, UN
Case 3.3	200	1,2, …, 5	0, 1, …, 4	UN	UC, CS, UN
Case 3.4	200	1,2, …, 10	0, 1, …, 9	UN	UC, CS, UN

Note: UC refers to uncorrelated var-cov structure; CS refers to compound symmetry var-cov structure; AR(1) refers to first order-autoregressive process var-cov structure and; UN refers to unstructured var-cov structure.

Variance-Covariance Structures:

The general form of the var-cov structures to be simulated from and their corresponding form for the 5x5 case (5 observations within each subject) are given below -

CS Var-Cov Structure:

From the given correlation parameter of 0.6, I can calculate $\sigma_b^2 = 1.5$ since I have, $\frac{\sigma_b^2}{\sigma_b^2+1} = 0.6$. I am also assuming the model error variance to be equal to 1 ($\sigma_e^2 = 1$). Hence, I have $\sigma_e^2 + \sigma_b^2 = 2.5$ for the CS var-cov structure.

$$\begin{pmatrix} \sigma_e^2+\sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_e^2+\sigma_b^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma_e^2+\sigma_b^2 \end{pmatrix} \Rightarrow \begin{pmatrix} 2.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 2.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 2.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 2.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 2.5 \end{pmatrix}$$

AR(1) Var-Cov Structure:

From the given correlation parameter of 0.6, I can set $\phi^2 = 0.6$. We also have, $\sigma_e^2 = \frac{\sigma_z^2}{1-\phi^2}$ and I have assumed the model error variance to be equal to 1 ($\sigma_z^2 = 1$) in this case. From this value I can calculate the $\sigma_e^2 = 2.5$ for the AR(1) var-cov structure.

$$\sigma_e^2 \begin{pmatrix} 1 & \phi & \dots & \phi^n \\ \phi & 1 & \dots & \phi^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^n & \phi^{n-1} & \dots & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 2.500 & 1.937 & 1.500 & 1.162 & 0.900 \\ 1.937 & 2.500 & 1.937 & 1.500 & 1.162 \\ 1.500 & 1.937 & 2.500 & 1.937 & 1.500 \\ 1.162 & 1.5 & 1.937 & 2.500 & 1.937 \\ 0.900 & 1.62 & 1.500 & 1.937 & 2.500 \end{pmatrix}$$

UN Var-Cov Structure:

This var-cov matrix is taken from a randomly generated var-cov structure used for a project by Dr. Nichole Carlson which I am also using for this simulation study. I have posted this var-cov structure in the GitHub page for this project in the folder R within the folder Code. This matrix has a larger magnitude of values in comparison with the CS and AR(1) var-cov structures and therefore I decided to scale the matrix by 2.5/mean(diagonals), so that the average of the diagonal would be closer in magnitude to the average of the CS and AR var-cov diagonals, which is 2.5. This was done so that the estimates across simulations would be more comparable.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} 1.705 & 1.648 & 1.711 & 1.740 & 2.855 \\ 1.648 & 2.126 & 2.135 & 2.195 & 2.225 \\ 1.711 & 2.135 & 2.534 & 2.506 & 2.542 \\ 1.740 & 2.195 & 2.506 & 2.959 & 2.855 \\ 1.786 & 2.225 & 2.542 & 2.855 & 3.175 \end{pmatrix}$$

Iterations:

For each of the sub-cases, I will run 1000 iterations that is, I will be simulating 1000 samples for each of these sub-cases.

Analysis:

For the simulation analysis, I will be fitting a simple multiple linear regression (assuming no correlation) model, a random intercept (CS structure) model and an UN model from the simulations for each of the subcases. From each simulated dataset belonging to each sub-case and from their corresponding model fit combinations, I will calculate means of the parameter estimates based on the average estimates of all the iterations. Additionally, I will calculate the average bias, average standard error, average p-value and the average Type I and Type II errors.

Type II errors will be calculated for all the 3 parameters as the proportion of p-values being more than 0.05 (standard 95% level of confidence) under the Alternative hypothesis.

For the Type I Error, I will set $\beta_3 = 0$ which is the coefficient estimate of the interaction term. Then, I will re-simulate the datasets for each of the sub cases under the null hypothesis of $H_0: \beta_3 = 0$. Therefore, Type I errors for the interaction term will be calculated as the proportion of p-values being less than 0.05 (standard 95% level of confidence) under the Null hypothesis.

Instead of calculating the var-cov structures manually, I have computed the var-cov structures for the AR(1) and UN cases using R. Rest of the analyses has been done is SAS. There have been many similar simulation studies done using R. But according to my knowledge, there have not been many similar simulation studies using SAS. Therefore, in this project, I have mainly used SAS to do this.

4. Results

Case 1

Table 2 below summarizes the results from the data simulated with a CS var-cov structure (Case 1). It has the results for each of the sub-cases where each sub-case has been fitted with a CS, UC, and an UN var-cov structure model.

Beta Estimate

The beta coefficient estimates for the three parameters of time, treatment and the interaction between time and treatment do not vary much when we simulate data with a CS var-cov structure, and they remain mostly similar as we

increase the number of subjects or the number of observations within each subject. This pattern is the same for each of the model fit under each sub-case.

Standard Errors

The standard errors for the three parameters of time, treatment and interaction tend to decrease as we increase the number of subjects and there is a sharp decrease in standard errors as we go from 20 to 200 subjects, when we simulate data with a CS var-cov structure. The standard error for treatment is the largest and it is at max when we simulate data with a CS var-cov structure and fit it with an CS var-cov structure, which is the correct fit. The standard error for treatment seems to be the lowest when we fit the model with an UN var-cov structure and is lower when we fit the model with an UC var-cov structure. This implies that, when we do not fit the model with the correct var-cov structures we are under-estimating the standard errors. This pattern is the same for each of the sub-cases.

P-Value

The coefficient estimates for time and interaction has p-values less than 0.05 for each of the model fits under the first two sub-cases, when we simulate data with a CS var-cov structure. The coefficient estimate of treatment has a p-value greater than 0.05 for the first two sub-cases for each model fit. The corresponding p-values for all the three parameter estimates become less than 0.05 as we increase the number of subjects to 200. For the last case scenario of 200 subjects and 10 observations within each subject, the p-values of all the three parameter estimates becomes 0 in all the model fits.

Bias

The bias of the three parameter estimates of time, treatment and interaction almost always remain very close to 0 for each of the model fits under each of the sub-cases, when we simulate data with a CS var-cov structure. The highest bias is observed for the treatment variable under the second subcase where we have 20 subjects and 10 observations within each subject and this pattern is uniform across all the three model fits for this sub-case.

Type II Error

The Type II error is the highest for the treatment variable with a CS var-cov structure model fit, when we simulate data with a CS var-cov structure which is surprising given it is the correct fit. However, the Type II error for treatment decreases

as we increase the number of subjects and is 0 for the last sub-case when we have 200 subjects and 10 observations within each subject. The Type II error for the other parameters of time and interaction are very low and goes to 0 in the last sub-case.

Type I Error

The Type I Error for the interaction term is the highest as we fit the model with an UN var-cov structure, when we simulate data with a CS var-cov structure and irrespective of what the number of subjects or the number of observations within each subject are, the Type I error always remains the highest for the model fit with the UN var-cov structure. The nature of this Type I error is anti-conservative. The Type I error for the interaction term is the lowest when we fit the model with a UC var-cov structure. The nature of this Type I error is conservative. With the correct model fit as in when we fit the model with a CS var-cov structure we get a Type I error which is midway between conservative and anti-conservative.

Table 2
Simulating data with Compound Symmetry var-cov structure
Fitting the model with a Compound Symmetry (CS), Uncorrelated (UC) and Unstructured (UN) var-cov structure

Fit		Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 1.1, Iterations = 1000, Subjects = 20, Observations = 5							
CS	Time	0.496	0.100	0.001	0.004	0.2	
CS	Treatment	0.981	0.647	0.263	0.019	71.5	
CS	Interaction	0.503	0.141	0.014	-0.003	7.5	6.1
UC	Time	0.496	0.155	0.009	0.004	3.7	
UC	Treatment	0.981	0.537	0.208	0.019	55.1	
UC	Interaction	0.503	0.219	0.059	-0.003	30.9	0.1
UN	Time	0.495	0.089	0.001	0.005	0.6	
UN	Treatment	0.982	0.582	0.239	0.018	63.9	
UN	Interaction	0.505	0.126	0.017	-0.005	7.7	10.3
Case 1.2, Iterations = 1000, Subjects = 20, Observations = 10							
CS	Time	0.500	0.035	0.000	0.000	0	
CS	Treatment	0.963	0.604	0.241	0.037	66.9	
CS	Interaction	0.499	0.049	0.000	0.001	0	4.5
UC	Time	0.500	0.054	0.000	0.000	0	
UC	Treatment	0.963	0.404	0.148	0.037	38.3	
UC	Interaction	0.499	0.076	0.000	0.001	0	0.4
UN	Time	0.500	0.025	0.000	0.000	0	
UN	Treatment	0.953	0.441	0.193	0.047	47.3	
UN	Interaction	0.501	0.036	0.000	-0.001	0	28.4
Case 1.3, Iterations = 1000, Subjects = 200, Observations = 5							

Fit		Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 1.1, Iterations = 1000, Subjects = 20, Observations = 5							
CS	Time	0.502	0.032	0.000	-0.002	0	
CS	Treatment	0.999	0.204	0.001	0.001	0.4	
CS	Interaction	0.499	0.045	0.000	0.001	0	5.9
UC	Time	0.502	0.050	0.000	-0.002	0	
UC	Treatment	0.999	0.173	0.000	0.001	0	
UC	Interaction	0.499	0.070	0.000	0.001	0	0.1
UN	Time	0.501	0.031	0.000	-0.001	0	
UN	Treatment	0.998	0.203	0.001	0.002	0.3	
UN	Interaction	0.500	0.044	0.000	0.000	0	5.8
Case 1.4, Iterations = 1000, Subjects = 200, Observations = 10							
CS	Time	0.500	0.011	0.000	0.000	0	
CS	Treatment	1.010	0.192	0.000	-0.010	0	
CS	Interaction	0.500	0.016	0.000	0.000	0	5.4
UC	Time	0.500	0.017	0.000	0.000	0	
UC	Treatment	1.010	0.131	0.000	-0.010	0	
UC	Interaction	0.500	0.025	0.000	0.000	0	0.4
UN	Time	0.500	0.011	0.000	0.000	0	
UN	Treatment	1.009	0.188	0.000	-0.009	0	
UN	Interaction	0.499	0.015	0.000	0.001	0	6.5

Note: Estimates are averaged over the 1000 iterations; Type I and Type II error are reported in percentages

Case 2

Table 3 below summarizes the results from the data simulated with a AR(1) var-cov structure (Case 2). It has the results for each of the sub-cases where each sub-case has been fitted with a CS, UC, and an UN var-cov structure model.

Beta Estimate

The beta coefficient estimates for the three parameters of time, treatment and interaction do not vary much when we simulate data with a AR(1) var-cov structure and they remain mostly similar as we increase the number of subjects or the number of observations within each subject. This pattern is the same for each of the model fits under each sub-case.

Standard Errors

The standard errors for the three parameters of time, treatment and interaction tend to decrease as we increase the number of subjects and there is a sharp decrease in standard errors as we go from 20 to 200 subjects, when we simulate data with a AR(1) var-cov structure. The standard error also seems to be the largest for the treatment variable irrespective

of the model fit. This pattern is uniform across all model fits for each of the sub-cases. The standard error estimate for Time and the interaction variable is the lowest for the CS model fit. Here we are either over-specifying or under-specifying the var-cov structure and this leads to either under or over-estimate of the standard errors.

P-Value

The coefficient estimates of time and interaction has p-values less than 0.05 for each of the model fits under the first two sub-cases, when we simulate data with a AR(1) var-cov structure. The coefficient estimate of treatment has a p-value greater than 0.05 for the first two sub-cases for each model fit. The corresponding p-values for all the three parameter estimates become less than 0.05 as we increase the number of subjects to 200. For the last case scenario of 200 subjects and 10 observations within each subject, the p-values of all the three parameter estimates becomes 0 in all the model fits.

Bias

The bias of the three parameter estimates of time, treatment and the interaction almost always remain very close to 0 for each of the model fits under each of the sub-cases, when we simulate data with a AR(1) var-cov structure. The highest bias is observed for the treatment variable under the second subcase where we have 20 subjects and 10 observations within each subject and this pattern is uniform for all the model fits under this sub-case.

Type II Error

The Type II error is the highest for the treatment variable with a CS var-cov structure model fit, when we simulate data with a AR(1) var-cov structure for the first two sub-cases. However, the Type II error for treatment decreases as we increase the number of subjects and is 0 for the last sub-case when we have 200 subjects and 10 observations within each subject. The Type II error is the highest for the interaction term with a UC var-cov structure model fit under the first sub-case. However, the Type II error for interaction term decreases to 0 as we increase the number of subjects. The Type II error for the other parameters of time starts very low and then it goes to 0 as we increase the number of subjects to 200 and the number of observations within each subject to 10.

Type I Error

The Type I Error for the interaction term is the highest as we fit the model with a CS var-cov structure, when we simulate data with a AR(1) var-cov structure and irrespective of what the number of subjects or the number of observations within each subject are, the Type I error always remains the highest for the model fit with the CS var-cov structure. Here the Type I error is anti-conservative when we fit the model with CS and UN var-cov structures and is conservative when we fit the model with an UC var-cov structure.

Table 3
Simulating data with AR(1) var-cov structure
Fitting the model with a Compound Symmetry (CS), Uncorrelated (UC) and Unstructured (US) var-cov structure

Fit	Effect	Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 2.1, Iterations = 1000, Subjects = 20, Observations = 5							
CS	Time	0.495	0.095	0.006	0.005	1.8	
CS	Treatment	0.981	0.648	0.269	0.019	69.7	
CS	Interaction	0.503	0.135	0.039	-0.003	12.9	19.9
UC	Time	0.495	0.154	0.021	0.005	9.4	
UC	Treatment	0.981	0.534	0.215	0.019	54.6	
UC	Interaction	0.503	0.218	0.095	-0.003	37.5	4.3
UN	Time	0.495	0.126	0.021	0.005	8.5	
UN	Treatment	0.990	0.629	0.257	0.010	67.0	
UN	Interaction	0.504	0.179	0.081	-0.004	27.8	10.5
Case 2.2, Iterations = 1000, Subjects = 20, Observations = 10							
CS	Time	0.501	0.040	0.000	-0.001	0	
CS	Treatment	0.960	0.562	0.230	0.040	59.8	
CS	Interaction	0.501	0.056	0.000	-0.001	0	29.7
UC	Time	0.501	0.053	0.000	-0.001	0	
UC	Treatment	0.960	0.404	0.157	0.040	39.7	
UC	Interaction	0.501	0.076	0.000	-0.001	0	17.1
UN	Time	0.501	0.053	0.000	-0.001	0.1	
UN	Treatment	0.962	0.488	0.210	0.038	50.5	
UN	Interaction	0.502	0.075	0.004	-0.002	1.2	26.4
Case 2.3, Iterations = 1000, Subjects = 200, Observations = 5							
CS	Time	0.502	0.031	0.000	-0.002	0	
CS	Treatment	0.999	0.205	0.001	0.001	0.5	
CS	Interaction	0.499	0.044	0.000	0.001	0	20.0
UC	Time	0.502	0.050	0.000	-0.002	0	
UC	Treatment	0.999	0.172	0.001	0.001	0.4	
UC	Interaction	0.499	0.070	0.000	0.001	0	4.2
UN	Time	0.502	0.044	0.000	-0.002	0	
UN	Treatment	1.001	0.219	0.002	-0.001	0.4	
UN	Interaction	0.499	0.063	0.000	0.001	0	5.9

Fit	Effect	Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 2.1, Iterations = 1000, Subjects = 20, Observations = 5							
Case 2.4, Iterations = 1000, Subjects = 200, Observations = 10							
CS	Time	0.501	0.013	0.000	-0.001	0	
CS	Treatment	1.011	0.179	0.000	-0.011	0	
CS	Interaction	0.499	0.018	0.000	0.001	0	31.8
UC	Time	0.501	0.017	0.000	-0.001	0	
UC	Treatment	1.011	0.131	0.000	-0.011	0	
UC	Interaction	0.499	0.025	0.000	0.001	0	18.3
UN	Time	0.501	0.023	0.000	-0.001	0	
UN	Treatment	1.011	0.209	0.001	-0.011	0	
UN	Interaction	0.498	0.032	0.000	0.002	0	6.5

Note: Estimates are averaged over the 1000 iterations; Type I and Type II error are reported in percentages

Case 3

Table 4 below summarizes the results from the data simulated with a UN var-cov structure (Case 3). It has the results for each of the sub-cases where each sub-case has been fitted with a CS, UC, and an UN var-cov structure model.

Beta Estimate

The beta coefficient estimates for the three parameters of time, treatment and interaction do not vary much when we simulate data with a UN var-cov structure, and they remain mostly similar as we increase the number of subjects or the number of observations within each subject. This pattern is the same for each of the model fits under each sub-case.

Standard Errors

The standard errors for the three parameters of time, treatment and interaction tend to decrease as we increase the number of subjects and there is a sharp decrease in standard errors as we go from 20 to 200 subjects, when we simulate data with a UN var-cov structure. The standard error estimate seems to be the largest for the treatment variable irrespective of the model fit across all the sub-cases. The standard error estimate for Time is the smallest across all model fits and all sub-cases. The standard error for the interaction variable is the lowest for the CS model fit and is the highest for the UC model fit. Therefore, for the CS model fit the standard error estimate for the coefficient of Time is under-estimated, the standard error estimate for the coefficient estimate of Treatment is over-estimated and the standard error estimate for the coefficient of the interaction term is under-estimated. Therefore, for the UC model fit the standard error

estimate for the coefficient of Time is over-estimated, the standard error estimate for the coefficient of Treatment is over-estimated for the first two sub-cases and is under-estimated for the last two sub-cases and the standard error estimate for the coefficient of the interaction term is over-estimated.

P-Value

The coefficient estimates of time and interaction has p-values less than 0.05 for each of the model fits under the first two sub-cases, when we simulate data with a UN var-cov structure. The coefficient estimate of treatment has a p-value greater than 0.05 for the first two sub-cases for each model fit. The corresponding p-values for all the three parameter estimates become less than 0.05 as we increase the number of subjects to 200. For the last case scenario of 200 subjects and 10 observations within each subject, the p-values of all the three parameter estimates becomes 0 across all the model fits.

Bias

The bias of the three parameter estimates of time, treatment and interaction almost always remain very close to 0 for each of the model fits under each of the sub-cases, when we simulate data with a UN var-cov structure. The highest bias is observed for treatment under the second subcase where we have 20 subjects and 10 observations within each subject and this pattern is uniform across all the three model fits.

Type II Error

The Type II error is the highest for the treatment variable with a CS var-cov structure model fit when we simulate data with an UN var-cov structure for the first two sub-cases. However, the Type II error for the treatment decreases as we increase the number of subjects from 20 to 200. The Type II error for the other parameters of time and interaction is high in the UC var-cov structure model fit in the first sub-case. The Type II error for all the three terms of time, treatment and interaction becomes 0 in the fourth sub-case.

Type I Error

The Type I Error for the interaction term is the highest as we fit the model with a CS var-cov structure across all model-fits and all sub-cases. The nature of this Type I error is anti-conservative. The Type I error for the interaction term is the lowest

as we fit the model with a UC var-cov structure across all the model-fits and all sub-cases. The nature of this Type I error is conservative.

Table 4
Simulating data with Unstructured var-cov structure
Fitting the model with a Compound Symmetry (CS), Uncorrelated (UC) and Unstructured (US) var-cov structure

Fit		Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 3.1, Iterations = 1000, Subjects = 20, Observations = 5							
CS	Time	0.500	0.059	0.000	0.000	0	
CS	Treatment	0.986	0.682	0.265	0.014	76.1	
CS	Interaction	0.499	0.084	0.000	0.001	0.3	20.1
UC	Time	0.500	0.153	0.007	0.000	1.4	
UC	Treatment	0.986	0.529	0.191	0.014	52.6	
UC	Interaction	0.499	0.216	0.049	0.001	31.2	0
UN	Time	0.499	0.079	0.000	0.001	0	
UN	Treatment	0.994	0.521	0.197	0.006	54.9	
UN	Interaction	0.502	0.112	0.009	-0.002	3.8	10.6
Case 3.2, Iterations = 1000, Subjects = 20, Observations = 10							
CS	Time	0.502	0.021	0.000	-0.002	0	
CS	Treatment	0.971	0.664	0.253	0.029	75.6	
CS	Interaction	0.497	0.030	0.000	0.003	0	41.6
UC	Time	0.502	0.053	0.000	-0.002	0	
UC	Treatment	0.971	0.397	0.127	0.029	36.4	
UC	Interaction	0.497	0.074	0.000	0.003	0	4.1
UN	Time	0.501	0.034	0.000	-0.001	0	
UN	Treatment	0.973	0.359	0.140	0.027	36.6	
UN	Interaction	0.497	0.048	0.000	0.003	0	28.0
Case 3.3, Iterations = 1000, Subjects = 200, Observations = 5							
CS	Time	0.501	0.019	0.000	-0.001	0	
CS	Treatment	1.000	0.216	0.001	0.000	0.1	
CS	Interaction	0.499	0.027	0.000	0.001	0	18.6
UC	Time	0.501	0.050	0.000	-0.001	0	
UC	Treatment	1.000	0.172	0.000	0.000	0	
UC	Interaction	0.499	0.070	0.000	0.001	0	0
UN	Time	0.501	0.028	0.000	-0.001	0	
UN	Treatment	1.001	0.182	0.000	-0.001	0	
UN	Interaction	0.499	0.039	0.000	0.001	0	5.4
Case 3.4, Iterations = 1000, Subjects = 200, Observations = 10							
CS	Time	0.500	0.007	0.000	0.000	0	
CS	Treatment	1.008	0.212	0.000	-0.008	0	
CS	Interaction	0.500	0.010	0.000	0.000	0	39.3
UC	Time	0.500	0.017	0.000	0.000	0	
UC	Treatment	1.008	0.131	0.000	-0.008	0	
UC	Interaction	0.500	0.025	0.000	0.000	0	3.1
UN	Time	0.500	0.014	0.000	0.000	0	

Fit		Beta Estimate	Standard Error	P-Value	Bias	Type II Error	Type I Error
Case 3.1, Iterations = 1000, Subjects = 20, Observations = 5							
UN	Treatment	1.007	0.154	0.000	-0.007	0	
UN	Interaction	0.500	0.020	0.000	0.000	0	6.9

Note: Estimates are averaged over the 1000 iterations; Type I and Type II error are reported in percentages

Visualization

In Figures 1 and 2, I am looking at two specific cases of **under-specification** and **over-specification** of the correlation structure, respectively. Figure 1 is based on the model fitted with a CS var-cov structure to the data simulated with an UN var-cov structure, thereby generating an underspecified correlation structure. Figure 2 is based on the model fitted with an UN var-cov structure to the data simulated with a CS var-cov structure, thereby generating an over-specified correlation structure. Here, I am using the first sub-case where we have 20 subjects with 5 observations within each subject and the model fits are computed from the 500th iteration which is at the half-way between all the iterations.

Figure 1:

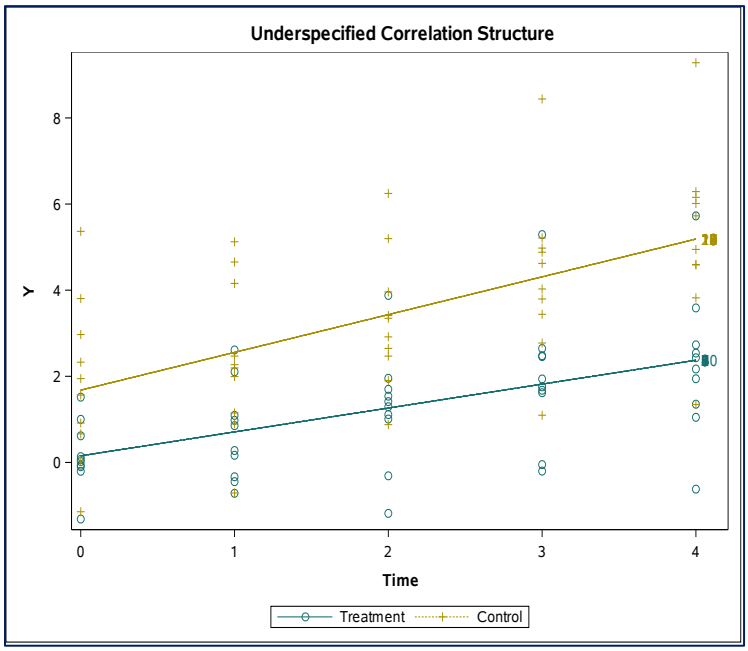
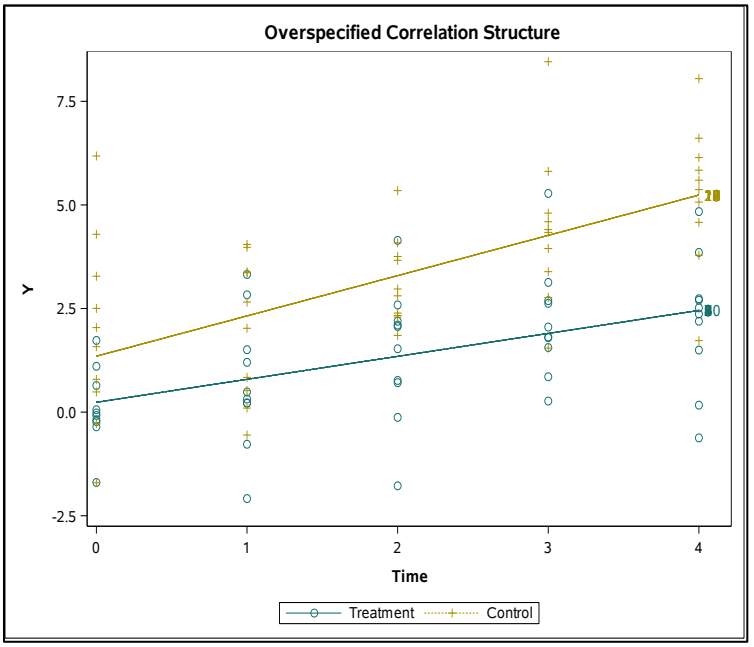


Figure 2:



These figures enable us to see the variation in intercepts for the subjects, but the slopes are determined by the treatment level of each individual. When we have under-specification, we see that the difference between slopes is less than what should have been, and this is because we are fitting a CS var-cov structure model to data simulated with a UN var-cov structure. Had it been the right fit there would have been sharp differences in both slopes and intercept for the subjects

in the treatment and control group. Similarly, when we have overspecification, we see that the differences between slopes is more than what should have been since we are fitting a UN var-cov structure model to data simulated with a CS var-cov structure. Had this been the right fit there would have been sharp differences in mainly intercepts for the subjects in the treatment and control group and not so much for the slopes.

5. Discussion

Limitations

Firstly, I would have liked to compute the AIC for each model fit to determine which model fit based on simulations from different var-cov structures gives the best fit expecting the CS fit from the CS simulation having the lowest AIC in Case 1 and the UN fit from the UN simulation having the lowest AIC in Case 3. I did not compute the AIC since I used *proc glm* for fitting the model with the UC var-cov structure and *proc glm* does not generate AIC. Getting the AIC for each model fit is one of the future works with respect to this project. Secondly, I would like to fit a model with an AR(1) var-cov structure for each of the cases and for Case 2 it would give the correct model fit and would help to compare the results for Case 2.

Conclusion

One of biggest the advantages of using SAS over other statistical packages like R for this type of simulation study was that for the model fit with UN var-cov structures there was no issues with convergence. All my models converged smoothly.

In summary, I found that, when data is simulated with a CS var-cov structure, under-specification of the correlation structure (UC fit) leads to under-estimated standard errors for the treatment variable compared to over-specification of the correlation structure (UN fit) which leads to over-estimated standard errors. But under-specification also leads to higher Type II errors as compared to over-specification. Type I error was also the lowest for the interaction term with UC model fit. Overall, for this case, standard errors, p-values, and Type II errors decreased in magnitude as I increased the number of subjects to 200 and the number of observations within each subject to 10 regardless of which model, I fitted. The Type I error was the highest for the model fit with the UN var-cov structure across all sub-cases. The beta coefficient estimates remained mostly uniform across all the model fits.

From the data simulated with a AR(1) var-cov structure, I found that under-specification of the correlation structure (UC fit) leads to lower standard errors for the treatment variable compared to over-specification of the correlation structure

(UN fit). The Type II error for the Treatment variable is highest with under-specification. For the other variables there were mixed results. Overall, for this case, standard errors, p-values, and Type II errors were the lowest when I increased the number of subjects to 200 and the number of observations within each subject to 10 regardless of which model, I fitted. The Type I error was the highest for the model fit with the CS var-cov structure across all sub-cases. The beta coefficient estimates remained mostly uniform across all the model fits.

From the data simulated with a UN var-cov structure, I found that under-specification of the correlation structure (UC fit) leads to lower standard errors for the treatment variable compared to less under-specification of the correlation structure (CS fit). Also, under-specification leads to lower Type II errors as compared to less under-specification. For the other variables there were mixed results. Overall, for this case, standard errors, p-values, and Type II errors were the lowest when we increased the number of subjects to 200 and the number of observations within each subject to 10 regardless of which model, we fitted. The Type I error was the highest for the model fit with the CS var-cov structure across all sub-cases. The beta coefficient estimates remained mostly uniform across all the model fits.

A correct model fit can be defined as the one where the data is generated from a CS var-cov structure and we fit a model with a CS symmetry var-cov structure as well. This type of model fit should give the best results as the model is correctly specified and this should lead to unbiased estimates of the coefficient(s) and unbiased standard error estimates as well. This project shows different case scenarios where such a best model fit is not achieved and what are the potential consequences we can face because of this.

6. References

Gibbs, P. & Kiernan, K. (2020). Simulating Data for Complex Linear Models. *SAS Global Forum 2020*.

Wicklin, R. Simulating Data with SAS.

Stroup, W. W., Milliken, G.A., Claassen, E.A. & Wolfinger, R.D. SAS for Mixed Models Introduction and Basic Applications.

Wicklin, R. Ten Tips for Simulating Data with SAS.