University of Alberta

Longitudinal Data Analysis in Social Science Data

by

Xiao Guo

A project submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science in Statistics

Department of Mathematical and Statistical Science

Edmonton, Alberta, Canada
August 2011
Abstract

This project discusses the Generalized Estimating Equation (GEE) model and its application for longitudinal data. For many longitudinal studies involving a discrete (count or categorical) response variable, and for some studies involving a continuous response variable, it is difficult to justify a fully parametric model for the data. In these circumstances, the method of GEE provides consistent estimates for the mean parameter when a model for the correlation may not be reliably specified.

In the second chapter, we begin by discussing Generalized Least Squares (GLS) and then give the theory of the GEE and the methodology of fitting GEE model. GEEs belong to a class of semiparametric regression techniques as they rely on specification of only the first two moments. Under mild regularity conditions, parameter estimates from GEEs are consistent. They are a popular alternative to the likelihood based generalized linear mixed model which is more sensitive to variance structure specification. They are frequently applied in multi-site cohort studies as they can handle many types of unmeasured dependence.

The third chapter discusses project background and the questions of interest, how the dataset is simulated for analysis is presented in Appendix A written by John Simpson.

Finally, this project discusses the application and comparison of GLS and GEE on the Social Science Data, and conclusion is given in the last section.

Keywords: Generalized Linear Squares; Generalized estimating equations; Generalized linear mixed model; Link selection; Marginal model; Repeated measurements; Semiparametric regression; working correlation matrix; Variance-covariance function.
Acknowledgements

I acknowledge my supervisor, Dr. Douglas Wiens, for offering me the opportunity to pursue graduate studies and for investing time and energy to train me. His emphasis on quality, excellent supervision, and persistence have helped me complete my graduate program and this project.

Also my acknowledgement goes to my family, especially my husband Yebin, for their love and support during the past years.
# Table of Contents

1 Introduction  
1.1 Longitudinal Data ................................. 1  
1.2 Notation ........................................... 1  
1.3 Covariance Structure .............................. 2  
1.4 Generalized Least Squares (GLS) Method .......... 4  
1.4.1 GLS Estimations ............................... 4  
1.4.2 Inference of GLS ............................... 5  
1.4.3 Disadvantage of the GLS method .............. 6  

2 Generalized Estimating Equation (GEE) Model .... 7  
2.1 Generalized Linear Model (GLM) .................... 7  
2.1.1 Characteristics of GLM ......................... 7  
2.1.2 GLM & Maximum Likelihood (ML) ............... 8  
2.2 GEE for longitudinal data .......................... 9  
2.2.1 Methodology ................................... 9  
2.2.2 The properties of GEE ......................... 14  

3 Project Background 16  
3.1 Population Growth .................................. 16  
3.2 Data Description ................................... 17  
3.3 Question of Interest ............................... 18  
3.4 Exploring the Longitudinal Data .................. 19  
3.4.1 Summary Statistics ............................. 19  
3.4.2 ANOVA ........................................ 20  
3.4.3 Plot of Data ................................... 22  

4 Data Analysis by GEE 24  
4.1 Main Effect of Agent Type .......................... 24  
4.1.1 Statistical Model ............................... 25  
4.1.2 Output and Interpretation ...................... 25  
4.2 Pattern of Change by Agent Type ................. 28  
4.2.1 Statistical Model ............................... 29  
4.2.2 Hypothesis Tests ............................... 29  
4.2.3 Output in SAS and interpretation ............. 29  
4.3 Pattern of Change by Roundante ................... 32  
4.3.1 Statistical Model ............................... 32  
4.4 Pattern of Change by Spawncost ................... 34  
4.4.1 Statistical Model ............................... 34  
4.4.2 Hypothesis Tests ............................... 35  
4.4.3 Output in SAS and interpretation ............. 35  
4.5 Pattern of Change by Roundante and Spawncost ... 37  
4.6 Conclusion ......................................... 40
Chapter 1

Introduction

1.1 Longitudinal Data

The classical regression model usually has the assumption that the errors are normally and independently distributed in the population with constant variance. But sometimes, one should expect some dependence between the response variables. For example, when you perform several experiments on the same subject, the fact is that the same subject will have some influences on the repeated observations. These data, known as longitudinal data, are defined as the data resulting from the observations of subjects (human beings, animals, or laboratory samples, etc.) which are measured repeatedly over time [12].

In longitudinal study, usually the three scientific questions of interest include not only how the treatment means differ, but also how treatment means change over time and how differences between means of treatments change over time.

In this project, we will mainly introduce and discuss the statistical models such as fixed effect model with GLS method and GEE method for the longitudinal data. More inferences in the field of longitudinal data analysis can be found in [14][5][6] and [15].

In the last chapter, we will apply GLS model and GEE model to the same sociology data and compare the results by the two models and prove that the two models have the similar estimates and arrive at the same inference on the parameters of interest to the sociology data.
1.2 Notation

Before we start our discussion of the approaches, we will consider linear models with regression functions for longitudinal data and some notations. We assume that \( n \) subjects are measured repeatedly over time. Let \( Y_{ij}; \quad i = 1, 2, \ldots, n; j = 1, \ldots, T \) be the response variable on \( j^{th} \) time order for the \( i^{th} \) subject, where observations on different subjects are assumed to be independent, though dependence is allowed among responses observed on the same subject. Each response \( Y_{ij} \) is associated with a \( p \times 1 \) vector of covariates, \( X_{ij} \), through the period of study the covariate may change over time. For example, it can be the age for subject (time varying covariate) or gender for the subject (time stationary covariate). Here we let \( Y_i \) denote the vector of responses for \( i^{th} \) subject and \( X_i \) denote the matrix of which the element is the vector of covariate for the response vector \( Y_i \). The regression models for longitudinal data can be written by:

\[
Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT} \end{pmatrix} = X_i^T \beta + \epsilon_i
\]

where \( Y_i \) and \( X_i \) are defined above, \( \beta \) is the vector of which the elements are the parameter of interest. \( \epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iT}) \) is the vector where the elements \( \epsilon_{ij} \) is the sum of random deviations that cause \( Y_{ij} \) to differ from the mean at time \( j \) for the \( i^{th} \) subject. Note \( \epsilon_1, \ldots, \epsilon_n \) are independent and multivariate normally distributed with variance covariance structure taking the possible forms listed in section 1.3.

1.3 Covariance Structure

One of the challenges for handling the longitudinal data is the repeated observations on the same subject are correlated so that the classical regression methods are no longer practical. In the project, the covariance structures that characterize the correlated data within subjects and the associated correlation structure are defined as

\[
\Sigma_i = \text{COV} [Y_i, Y_i] = E \left( (Y_i - E(Y_i))(Y_i - E(Y_i))^T \right) = (\text{cov} (Y_{is}, Y_{it}))
\]

\[
\Gamma_i = (\text{corr} (Y_{is}, Y_{it}))
\]
These covariance matrices may depend on a vector of distinct unknown parameters $\alpha$ that fully specify the chosen covariance matrix. Although this covariance matrix can differ from subject to subject, we commonly use a common covariance matrix $\Sigma$ and correlation matrix $\Gamma(\alpha) = \Gamma_i, i = 1, 2, \cdots, T$ to approximate the average dependence among repeated observations over subjects.

Therefore, $\Gamma(\alpha)$ is the same for all subjects, and the vector of all unknown parameters $\alpha$ will be estimated from the data.

The followings are different types of covariance matrix $\Sigma$ and the corresponding correlation matrix $\Gamma(\alpha)$ that are often seen with longitudinal data. We restrict our discussion throughout the project to the case of balanced data where all subjects have repeated measurements at the same $T$ time points. \footnote{Section 1.3 is mainly reproduced from Longitudinal and Panel Data Analysis and Applications in Social Science[7].}

1. **Unstructured correlation**: For observations taken at the same time points for different subjects, this assumption places no restriction on the nature of associations among elements of a data vector. If $Y_{is}$ and $Y_{it}$, $s, t = 1, \cdots, T$, are two observations on the same subject where all subjects are observed at the same $T$ time points, and if $\rho_{st}$ represents the correlation between $Y_{is}$ and $Y_{it}$, i.e $\text{corr}(Y_{is}, Y_{it}) = \rho_{st}$, the implied correlation matrix for a data vector with all $T$ observations is the ($T \times T$) matrix

$$
\Gamma(\alpha) = \begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1T} \\
\rho_{21} & 1 & \cdots & \rho_{2T} \\
\vdots & & \ddots & \vdots \\
\rho_{T1} & \cdots & \rho_{T,T-1} & 1
\end{pmatrix}
$$

(1.3)

where $\rho_{st} = \rho_{ts}$ for all $s, t$. Thus, the unstructured working correlation assumption depends on $\frac{T(T-1)}{2}$ distinct correlation parameters. Then, the vector $\alpha$ consists of the distinct possible correlation parameters $\rho_{st}$ for the data vector of maximal size $T$.

2. **One-dependent correlation**: This assumption says that only observations adjacent in time are correlated by the same amount $-1 < \rho < 1$. In principle, this model could be used with any situation; however, for unbalanced data
with different time points, it may not make sense. The model may be written by:

\[ \Gamma(\alpha) = \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ \vdots \\ 0 & \cdots & 0 & \rho & 1 \end{pmatrix} \]  \hfill (1.4)

Then in such a case, the vector \( \alpha = \rho \).

3. **Exchangeable (compound symmetry)**: This assumption says that the correlation between distinct observations on the same subject is the same regardless of when in time the observations were taken, i.e. \( \text{corr}(Y_{is}, Y_{it}) = \rho_{st} = \rho, s \neq t \). In principle, this model could be used with balanced data, ideally balanced data with missing values, and unbalanced data where time points are different for different units. This structure may be written in terms of a single correlation parameter \( 0 < \rho < 1 \):

\[ \Gamma(\alpha) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots \\ \rho & \cdots & \cdots & 1 \end{pmatrix} \]  \hfill (1.5)

4. **Auto-regressive** This assumption says that correlation among observations "tails off" if \(-1 < \rho < 1 \); i.e. \( \text{corr}(Y_{is}, Y_{it}) = \rho_{st} = \rho^{|s-t|} \); the model is

\[ \Gamma(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho^{J-1} \\ \rho & 1 & \cdots & \rho^{J-2} \\ \vdots \\ \rho^{J-1} & \rho^{J-2} & \cdots & 1 \end{pmatrix} \]  \hfill (1.6)

Note the vector \( \alpha = \rho \) for the compound symmetry, AR(1), and one-dependent structure.

### 1.4 Generalized Least Squares (GLS) Method

#### 1.4.1 GLS Estimations

As noted above, the defining characteristic of longitudinal data is that the same response variable is measured repeatedly over time on each subject. So that the ordinary least square applied for estimating the unknown parameters in classical
regression model is not appropriate because the classical regression models assume the independence among the repeated measurements. In longitudinal data cases, ordinary least square can be statistically inefficient, or even leads to misleading inferences.

The method of GLS is introduced to improve upon estimation efficiency when $\text{VAR}(Y_i) = \Sigma_i$ is not a scalar variance covariance matrix. We first assume $Y_i, \ i = 1, 2, \cdots, T$ are independent and multivariate normally distributed and $\alpha$ is unknown.

Then the maximum likelihood estimates of $\beta$ and $\alpha$ boils down to maximize the likelihood function

$$L(y, \beta) = \prod_{i=1}^{n} (2\pi)^{-\frac{T}{2}} \exp \left\{ -\frac{(y_i - \mathbf{X}_i \beta)^T \Sigma_i^{-1} (y_i - \mathbf{X}_i \beta)}{2} \right\}$$

(1.7)

(1.7) is complicated function of $\beta$ and $\alpha$ so that, the solutions to the equation can not be done in a closed form in general. But it is possible to show[7] that ML estimator $\beta$ can be written as

$$\hat{\beta} = \left( \sum_{i=1}^{n} \mathbf{X}_i^T \hat{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \hat{\Sigma}_i^{-1} Y_i$$

(1.8)

where $\hat{\Sigma}_i$ is the covariance matrix for $Y_i$ with the estimator for $\alpha$ plugged in. Note when $\alpha$ is known, (1.8) can be simplified to

$$\hat{\beta} = \left( \sum_{i=1}^{n} \mathbf{X}_i^T \Sigma_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \Sigma_i^{-1} Y_i$$

(1.9)

$\hat{\beta}$ in (1.8) is often called GLS estimator of $\beta$, note the designation "generalized" emphasizes that $\Sigma_i$ is not known and its parameter estimated. Because $\hat{\Sigma}_i$ depends on the estimates of $\alpha$ which in turn depends on the data $Y_i$, it is generally the case that it is not possible to do the calculation in closed form and numerical methods have to be used. More detailed discussion of the GLS theory can also be found in e.g., Amemiya (1985), Greene (2000).

1.4.2 Inference of GLS

Since GLS estimator is based on ML method, large sample theory and the statistical hypothesis such as Wald’s test and likelihood ratio test that are based on large sample theory are also applicable. It may be shown that, approximately, for $n$ large
enough, we have

\[ \hat{\beta} \sim \left\{ \beta, \left( \sum_{i=1}^{n} X_i^T \Sigma^{-1} X_i \right)^{-1} \right\} \]  

(1.10)

For a given matrix of \( L \) of dimension \( (r \times p) \), the linear function of the elements of \( \beta, L\beta \) is asymptotically normally distributed.

\[ L\hat{\beta} \sim N \left\{ L\beta, L \left( \sum_{i=1}^{n} X_i^T \Sigma^{-1} X_i \right)^{-1} L^T \right\} \]  

(1.11)

With \( \Sigma_i \) is replaced by \( \hat{\Sigma}_i \) that depends on the estimated value of \( \hat{\alpha} \), the covariance matrix of sampling distribution of \( \hat{\beta} \) is approximated by

\[ \hat{V}_\beta = \left( \sum_{i=1}^{n} X_i^T \hat{\Sigma}^{-1} X_i \right) \]  

(1.12)

So that we can approximate the sampling distribution of the linear function of \( L\hat{\beta} \) as

\[ L\hat{\beta} \sim N \left\{ L\beta, L \left( \sum_{i=1}^{n} X_i^T \hat{\Sigma}^{-1} X_i \right)^{-1} L^T \right\} = N \left\{ L\beta, L\hat{V}_\beta L^T \right\} \]  

(1.13)

1.4.3 Disadvantage of the GLS method

A drawback of the GLS method is that it is difficult to implement. In practice, certain structures (assumptions) must be imposed on \( \Sigma_i \) so that a feasible GLS estimator can be computed. This approach results in two further difficulties, however. First, the variance covariance structures of the response variable need not be correctly specified. Consequently, the resulting feasible GLS estimator may not be as efficient as one would like. Second, the finite sample properties of feasible GLS estimators are not easy to establish. Consequently, exact tests based on the feasible GLS estimation results are not readily available [10].
Chapter 2

Generalized Estimating Equation (GEE) Model

2.1 Generalized Linear Model (GLM)

2.1.1 Characteristics of GLM

Generalized linear models (GLMs) represent a class of fixed effects regression models for several types of dependent variables. McCullagh and Nelder describe these in great detail [13] and indicate that the term "generalized linear model" is due to Nelder and Wedderburn who described how a collection of seemingly disparate statistical techniques could be unified. GLM is a regression model for responses $Y_1, \ldots, Y_n$ with exponential family distribution and the following features [4]:

1. Mean: The mean of $Y_i$ takes the form of

$$g(E(Y_i)) = x_i^T \beta$$

(2.1)

where $x_i$ is a vector of $p$ covariates and $\beta$ is the vector of $p$ parameters of interest. The link function $g(.)$ is almost always chosen to be strictly increasing or decreasing function of $x_i^T \beta$ for easier interpretation. Here the function $g(.)$ is called link function because it links the mean of $Y_i$ with the covariates $x_i$.

2. Probability distribution: The response $Y_i$ is assumed to follow exponential family class.

3. Variance: The variance of $Y_i$ take the form determined by the function of mean:

$$VAR(Y_i) = \phi V(E(Y_i))$$

(2.2)
where $V(\cdot)$ is the variance function and $\phi$ is a possibly unknown scale parameter.

### 2.1.2 GLM & Maximum Likelihood (ML)

It turns out that the ML estimator of $\beta$ in any GLM solves a set of $p$ equations of the same general form [13]:

$$
\sum_{i=1}^{n} \frac{1}{V(f(x_i^T \beta))} \{Y_i - f(x_i^T \beta)\} f'(x_i^T \beta)x_i = 0 \tag{2.3}
$$

where $f'(\mu) = \frac{df(\mu)}{d\mu}$, the derivative of $f$ with respect to its arguments. Each deviation $\{Y_i - f(x_i^T \beta)\}$ is weighted in accordance with its variance, and here the function $f$ is certainly not a linear function of $\beta$ as in the classical linear model, thus it is not possible to solve (2.3) analytically and the solution must be found using a numerical algorithm.

The estimator of $\beta$ can be made by using Iteratively Reweighed Least Squares (IWLS), the procedures is done as follows:

1. Given a starting value of $\beta^{(0)}$ and evaluate the weights $\frac{1}{V(f(x_i^T \beta))}$.

2. Solve (2.3) with the weights evaluated from the step 1 based on $\beta^{(0)}$. This will give a new estimator of $\beta : \beta^1$.

3. Repeat step 1 and 2, continue estimating $\beta$ until it converges i.e. the two successive $\beta$ are equal.

The **sample distribution** of the estimator of $\beta$ cannot be derived in closed form. However, with the large sample theory approximation, we have the IWLS estimator of $\beta$ that satisfies

$$
\tilde{\beta} \sim N \{ \beta, \phi(\Delta^T V^{-1} \Delta)^{-1} \} \tag{2.4}
$$

Here

- $\Delta$ is a $(n \times p)$ matrix whose $(i, j)$ elements is the derivative of $f(x_i^T \beta)$ with respect to the $j^{th}$ element of $\beta$.

- $V$ is the $n \times n$ diagonal matrix with diagonal elements of $V(f(x_i^T \beta))$. 

8
Standard errors for the elements of $\hat{\beta}$ are then found as the square roots of the diagonal elements of the matrix

$$V_\beta = \hat{\phi} \left( \Delta^T \hat{V}^{-1} \Delta \right)^{-1}$$

(2.5)

The Hypothesis tests of $\beta$ can be tested by using Wald testing procedure. For null hypothesis $H_0 : L\beta = h$, the sampling distribution of $L\hat{\beta}$ can be approximated by

$$L\hat{\beta} \sim N \left( L\beta, L\hat{V}L^T \right)$$

(2.6)

### 2.2 GEE for longitudinal data

As mentioned above, GLMs (McCullagh and Nelder 1989) are standard methods used to fit regression models for univariate data that are assumed to follow an exponential family distribution such as continuous data, count data or data that take on positive values that may have skewed distributions.

Frequently researchers are interested in analyzing data that arise from a longitudinal, repeated measures, where the independent observations of usual typical regression methods is violated. Due to the lack of independence among the repeated measures, it is no longer appropriate to apply GLM to longitudinal data.

In this section, we are concerned with how we might extend GLM to the situation of longitudinal data, where now the responses are vector $Y_i$ of repeated count, binary or other observations on each of $n$ subjects. [11][16][8][18] offer more extensive discusses about the principles and applications of GEE.

#### 2.2.1 Methodology

The approach for extending GLM to longitudinal data is a class of regression models that are known as marginal models. The defining features of marginal models is that marginal models can relate the mean response at each time point, via a link function, to the covariates. Marginal models not only specify a GLM for the longitudinal responses at each time point but also include a model for the within subject association among the responses [16]. The methodology for GEE is listed as follows:

1. Step 1: The marginal response of a data vector $Y_i$ is modelled as a function of time, other covariates, and parameters $\beta$ by using a generalized linear
model type mean structure to represent the mean response of each element of $Y_i$.

$$E(Y_{ij}) = g^{-1}(X_{ij}^T \beta)$$

where $Y_{ij}$ is the response for subject $i$ at time $j$ and also the element of $Y_i$; $X_{ij} = (x_{ij1}, \cdots, x_{ijp})^T$ is the corresponding $p \times 1$ vector of covariates; $\beta = (\beta_1, \cdots, \beta_p)'$ is a $p \times 1$ vector of unknown parameters. The vector $\beta$ characterizes how the cross-sectional response distribution depends on the explanatory variables. Finally, $g(\cdot)$ is the link function.

2. Step 2: The variance of each element of $Y_i$ is modelled by the function of mean that is appropriate for the type of data; i.e., count data are taken to have the poisson variance structure, which says that variance of any element of $Y_i$ is equal to the corresponding model for the mean. These models are often modified to allow for greater variation both within subjects and among subjects by the addition of a dispersion parameter $\phi$.

$$\text{VAR}(Y_{ij}) = V(\mu_{ij}) \phi$$

where $V(\cdot)$ is the variance function and $\phi$ is a possibly unknown scale parameter;

For a normally distributed response, a natural choice is

$$g(\mu_{ij}) = \mu_{ij}, \quad V(\mu_{ij}) = 1, \quad \text{VAR}(y_{ij}) = \phi.$$  \hfill (2.9)

If the response variable is binary, the choice is

$$g(\mu_{ij}) = \log(\frac{\mu_{ij}}{1 - \mu_{ij}}), \quad V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij}), \quad \phi = 1.$$  \hfill (2.10)

If the response variable is Poisson count,

$$g(\mu_{ij}) = \log(\mu_{ij}), \quad V(\mu_{ij}) = \mu_{ij}, \quad \phi = 1.$$  \hfill (2.11)

If the response variable is Gamma distribution,

$$g(\mu_{ij}) = \frac{1}{\mu_{ij}}, \quad V(\mu_{ij}) = (\mu_{ij})^2, \quad \phi = 1$$  \hfill (2.12)

are often used.
3. Step 3: Correlation among observations on the same subject (elements of $Y_i$) is represented by choosing a model, such as the correlation structures corresponding to AR(1), one-dependent or other specifications in the following. Because there is some uncertainty in doing this and no formal way to check it, the chosen model is referred to as the working correlation matrix to emphasize this fact.

$$E(Y_i) = \begin{pmatrix}
g^{-1}(x_{i1}^T\beta) \\
g^{-1}(x_{i2}^T\beta) \\
\vdots \\
g^{-1}(x_{iT_i}^T\beta)
\end{pmatrix} \quad (2.13)$$

$$\text{VAR} (Y_i) = \phi T_i \Gamma_i T_i^T = \Sigma_i \phi \Lambda_i \quad (2.14)$$

where $\Gamma_i$ is the $T \times T$ correlation matrix under one of the correlation matrices dictated in the following, it is assumed that $\Gamma_i$ is properly constructed for the subject’s time pattern. $T_i$ is the diagonal matrix whose diagonal elements are the standard deviations of the $Y_{ij}$ under the model, except for the dispersion parameter $\phi$. $\Sigma_i$ is the covariance matrix for the data vector $Y_i$.

4. Step 4: Estimate the parameter vector $\beta$ and its covariance matrix using GEE method. It turns out that if $Y_i$ follow a multivariate normal distribution, we were led to the estimator for $\beta$

$$\hat{\beta} = \left(X_i^T \hat{\Sigma}_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^T \hat{\Sigma}_i^{-1} Y_i \quad (2.15)$$

where $\hat{\Sigma}_i$ is the covariance matrix with the estimator for $\alpha$ plugged in. And this equation can be rewritten as the following form:

$$\sum_{i=1}^{n} X_i^T \hat{\Sigma}_i^{-1} \left( Y_i - X_i^T \hat{\beta} \right) = 0 \quad (2.16)$$

The estimator for $\beta$ solves a set of $p$ equations for $\beta$ $(p \times 1)$. Comparing (2.16) and (2.3), it is suggested to using a natural approach for fitting model (2.13) and (2.14) because both the estimating equations consist of $p \times 1$ equations that are linear functions of deviations of observations from their assumed mean.
that are weighted in accordance with their covariance (variance in univariate case).

\[ \sum_{i=1}^{n} \Delta^{T} \hat{\Lambda}_i^{-1} \{ Y_i - g^{-1}(\beta) \} = 0_p \]  

(2.17)

where

\[ \Delta = \begin{pmatrix} 
\frac{\partial g^{-1}(x_{i1}\beta)}{\partial \beta} \\
\frac{\partial g^{-1}(x_{i2}\beta)}{\partial \beta} \\
\vdots \\
\frac{\partial g^{-1}(x_{iT}\beta)}{\partial \beta} 
\end{pmatrix} \]  

(2.18)

and \( \hat{\Lambda}_i \) is the matrix \( \Lambda_i \) in (2.13) with an estimator for \( \alpha \) plugged in. To find an estimator of \( \alpha \), we can use GLM to fit the model by assuming that the observations among all the subjects are independent. Then the scale parameter \( \phi \) can be written by

\[ \hat{\phi} = \frac{1}{n \times p - p} \sum_{i=1}^{n} \sum_{j=1}^{T} \frac{Y_{ij} - g^{-1}(x_{ij}\hat{\beta})^2}{V\{g^{-1}(x_{ij}\hat{\beta})\}} \]  

(2.19)

If \( \Lambda_i \) corresponds to the unstructured correlation matrix, then the parameter \( \rho_{jk} \) can be estimated by

\[ \hat{\rho}_{jk} = \frac{1}{n\phi} \sum_{i=1}^{n} r_{ij} r_{ik} \]  

(2.20)

where \( r_{ij} = \frac{Y_{ij} - g^{-1}(x_{ij}\hat{\beta})}{\sqrt{V\{g^{-1}(x_{ij}\hat{\beta})\}}} \) is called the standardized Pearson residuals.

5. Solving the GEE:

The solution is found by iterating between quasi-likelihood methods for estimating \( \beta \) and a robust method for estimating \( \alpha \) as a function of \( \beta \). The procedure is as follows:

(a) Obtain an initial estimator of \( \beta \) by assuming all the subjects are independent. Using this estimator for \( \beta \) to estimate \( \phi \) with (2.19) and then \( \alpha \) for the assumed working correlation and covariance matrix \( \hat{\Lambda}_i \) to form the weights.

(b) Using the fixed weights above, solve the GEE (2.17) to obtain a new estimator \( \hat{\beta} \). The resulting equation may then be solved by a numerical
technique that is an extended version of the IRWLS method used in the ordinary case.

(c) Repeat step 2 and 3 until the two successive result of $\beta$ stay the same.

6. Hypothesis Tests:

- Sampling distribution:
  According to the large sample theory, an approximation to the sampling distribution of the estimator of $\hat{\beta}$ by solving GEE is obtained. For large $n$ subjects, the GEE estimator $\hat{\beta}$ follows

$$\hat{\beta} \sim N \left\{ \beta, \phi \left( \Delta_i^T \Lambda_i \Delta_i \right)^{-1} \right\}$$

(2.21)

For practical use, $\Lambda_i$ and $\Delta_i$ are often replaced by estimates. Then the estimated covariance matrix of $\hat{\beta}$ can be defined as:

$$\hat{V}_{\beta} = \phi \left( \sum_{i=1}^{n} \hat{\Delta}_i^T \hat{\Lambda}_i \hat{\Delta}_i \right)^{-1}$$

(2.22)

- Hypothesis tests:

Now our interest may be to test hypotheses concerning the elements of $\beta$. Consider the null hypotheses of the form

$$H_0 : L\beta = d$$

(2.23)

where $L$ is a $r \times p$ matrix of constants imposing $l$ linearly independent constants on the elements of $\beta$ and $d$ is a $r \times 1$ vector of constants. Because $\hat{\beta}$ is asymptotically normal, we have the large sample approximation

$$L\hat{\beta} \sim N \left( L\beta, L\hat{V}_{\beta}L^T \right)$$

(2.24)

the Wald $\chi^2$ test statistic

$$\chi^2 = \left( L\hat{\beta} - d \right) \left( L\hat{V}_{\beta}L^T \right)^{-1} \left( L\hat{\beta} - d \right)$$

(2.25)

has an asymptotic $\chi^2_l$ distribution if $H_0$ is true.
2.2.2 The properties of GEE

1. One of the appealing properties of GEE in the longitudinal design is the estimator \( \hat{\beta} \) is almost as efficient as the maximum likelihood estimator [16]. Consider the case of linear model for continuous responses assuming to be multivariate normally distributed, it can easily be shown that the GLS estimator of \( \beta \) in linear model can be considered a special case of GEE method. For many longitudinal designs, there is little loss of efficiency when the GEE method is adopted as an alternative of maximum likelihood.

2. A second desirable characteristic of GEE estimator \( \hat{\beta} \) is it yields a consistent estimator of \( \beta \) even if the within-subject associations among the repeated measures have been misspecified. It is shown by Wedderburn that for any choice of weights \( V_i \) in

\[
\sum_{i=1}^{n} X_i^T \hat{V}_i^{-1} (Y_i - X_i^T \hat{\beta}) = 0. \tag{2.26}
\]

GEE estimator \( \hat{\beta} \) is consistent and asymptotically normal. The choice of weights \( V_i = \text{VAR}(Y_i) \) gives the estimator with smallest variance among all estimators in this class. It only requires that the model for the mean response be corrected. This robustness property of GEE is important because the usual focus of a longitudinal study is on changes in the mean response.

We will show in next chapter that when the working correlation is set to identity matrix \( I \) in replacement of \( \Lambda_i \), there is little loss of efficiency on the estimator \( \beta \) for the two settings.

3. Although the GEE method yields a consistent estimator of \( \beta \) under misspecification of the within subject associations, the usual standard errors obtained under the misspecified model for the within subject association are not valid [16]. Fortunately, it turns out that, the modified version of \( \hat{V}_\beta \),

\[
\hat{V}_\beta^R = \left( \sum_{i=1}^{n} \Delta_i^T \tilde{\Lambda}_i^{-1} \Delta_i \right)^{-1} \left( \sum_{i=1}^{n} \Delta_i^T \tilde{\Lambda}_i^{-1} \tilde{S}_i \tilde{\Lambda}_i \Delta_i \right) \left( \sum_{i=1}^{n} \Delta_i^T \tilde{\Lambda}_i^{-1} \Delta_i \right)^{-1} \tag{2.27}
\]

for large \( n \), will provide a reliable estimate of the true sampling covariance matrix of \( \hat{\beta} \) even of the chosen \( \Sigma_i \) is incorrect. In contrast, if the model is
incorrect, $\hat{V}_\beta$ will not provide a reliable estimate.

The alternative estimate $\hat{V}_\beta^R$ is often referred to as the robust estimates of the sampling covariance matrix of $\hat{\beta}$ because it attempts to correct for the possibility that the model assumption on $\Gamma_i$ may be wrong. While $\hat{V}_\beta$ is often referred to as the model-based covariance matrix estimate because it assumes that we are always correct about the assumption on $\Gamma_i$.

No consensus exists on whether to use model-based estimate $\hat{V}_\beta$ or the robust estimate $\hat{V}_\beta^R$ is to be preferred in finite samples in practical problems. If they are different, some people take that as an indication that the original assumption is wrong. Usually there is no recommendation on which estimates to use.
Chapter 3

Project Background

3.1 Population Growth

Population growth in this project is defined as the change in a population over time, and can be quantified as the change in the number of individuals of any species in a population using "per unit time" for measurement. Simple models of population growth include the Malthusian Growth Model and the logistic model.

In the last few centuries, the number of people living on Earth has increased many times over. By the year 2000, there will be 10 times more people on Earth than there were 300 years ago.

How can population grow so fast? Think of a family tree. At the top are 2 parents, and beneath them the children they had. Listed beneath those children are the children they had, and so on and so on, down through each generation. As long as the family members continue to reproduce, the family tree continues to increase in size, getting larger with each passing generation. This same basic idea applies to the world’s population.

Population grows in the same way that money grows when it’s left to compound interest in a bank. With money, growth comes through accumulating interest upon interest. The interest payments you accumulate eventually earn interest, increasing your money. With population growth, new members of the population eventually produce other new members of the population. ¹

In this project, researchers are willing to observe the growth trend of the human population to see how the growth of population is affected by factors such as economy

development, grain yield, agent type (which is explained in next section) and other factors. However, to answer this question in reality takes too much money and time so that a computer program is built by John Simpson to produce the simulated dataset to illustrate the conclusion.

3.2 Data Description

In this study, subjects (to distinguish the population in the dataset from that in statistics, subject is used instead of population through the project) are chosen randomly from two groups, subjects from each group are controlled by different agent types: objectively choosing agent and subjectively choosing agent. Each population is then randomly assigned to different agent type with either objectively choosing agent (SQ) or subjectively choosing agent (NC). Immediately prior to the working of agent type, the size of the population (growth size) is recorded.

Then at every 1 generation (or 10 years), 2 generations, 3 generations, 4 generations, and 5 generations, the growth size is measured again. Intuitively, the growth size is likely to increase if the agent types are working, with the better agent type resulting bigger growth size on average.

Also recorded on each subject is the environment condition and resource such as roundante and an indictor of the spawncost with which the subjects have to cost for next generation.

The variables we will use and be interested in are as follows:

1. Response: growth size

   Although in reality, it is very complicated to evaluate what is the better society, we cannot deny the fact that when having the similar situations or environment conditions, societies with bigger growth size will be seen as successful ones. We make a rough analysis for the dataset before further analysis starts and found that there are no societies of crashing very fast after several rounds or even during the following rounds. Therefore in this project, we do have evidences that the development of the population can be reflected by the size of the population (growth size).

   Since the dataset created by the simulation is huge, we are thinking to make analysis easier by making a sample of the dataset with sample size equal

17
to 100 for each agent type (SQ and NC). The responses or growth size is measured every generation (ten years), sixty years of the population will be considered for the repeated measurement analysis. Because the great value of the population size, the response will be divided by 1000 for calculation. The dataset extracted from stata has been converted to text file so that it can be used in SAS (.dat).

2. Two agent types
One of our research interests from the previous discuss is how the two agent types: the objectively choosing agents and the objectively choosing agents will affect the development of the population which is reflected by the growth size. Agent type has two categories in this project, one is NC, the other one is SQ. Repeated measurement analysis will be used to investigate the effect of the two agent types on the response in the following sections.

3. Two environment conditions
Institutively, in reality we have the common sense that populations with harsher environment conditions have more difficulties to survive or be thriving, does the same conclusion also be revealed in this dataset? To investigate the association between the development of the society and the environment conditions, two environment conditions will be taken into consideration.
Roundante indicator indicates the prices the agent requires for the generation, it also suggests demands or requirements that the populations needs to grow. This indicator is categorized by two cases: roundante 0 means the lower price while roundante 1 means the higher price.
Spawncost indicates the price the agent require for the spawning. Similarly, it also suggests the demands that the society need to help the population grow. This indicator is categorized by two cases: spawncost 0 indicates fewer demands while spawncost 1 indicates higher demands.

3.3 Question of Interest
As mentioned above, the purpose of generating the dataset is to find the relationship between the growth size and time (generations), agent type and other covariates. The researchers had several questions, which are as follows:
1. Is there evidence to suggest that these two agent types differ in their effects on the development of the population, as reflected by growth size in the project?

2. Is there a difference in the pattern of change of growth size between subjects controlled by objectively choosing agent and subjectively choosing agent over a 6 generation period? Does one agent type show results more quickly?

3. What is the difference in growth size at the end of the study? Is there evidence to suggest that one of the agent types produces larger growth size on average than the other after 6 generations?

4. Is there an association between the pattern of change of growth size and spawncost and/or roundante in population controlled by NC agent? This might indicate that this agent type brings about bigger growth size in some kinds of subjects versus others.

5. Is there an association between the pattern of change of growth size and spawncost and/or roundante in population controlled by SQ agent? This might indicate that this agent type brings about bigger growth size in some kinds of subjects versus others.

### 3.4 Exploring the Longitudinal Data

#### 3.4.1 Summary Statistics

Before we take further analysis on the dataset, we consider some summary statistics. Since this was a randomized study, we would expect subjects in the two groups (NC, SQ) to be similar in their characteristics prior to administration of the two agent types. Table 3.1 lists sample mean for growth size at each generation for 100 subjects assigned to NC and 100 subjects assigned to SQ group. The raw sample mean suggests a slight increase at each generation on the growth size in NC group while a possible slight decrease for the first 2 generations followed by a considerable increase for the following generations in the SQ group. Comparing the results for gen 1 to gen 6 (13.633 to 16.674 in NC group and 14.416 to 20.421 in SQ group), the growth size for both the groups increase considerably at the end of the study. The average growth size at the first generation in NC group (13.633) is very close to that in SQ group (14.416) which indicates that subjects in the two agent types are
similar and independent in their characteristic for the first generation as this study is randomized.

### 3.4.2 ANOVA

Researchers are also interested in the difference between the two spawncost groups before the study started as well as the difference between the two roundante groups. The primary goal of the analysis is to determine if any systematic pattern prior to the treatment exists among the subjects to affect the repeated observations. The above analysis will help us summarize the data before further exploration.

The normality checking using Shapiro-Wilks methods shows that the individual observations are normally distributed (growth size at each generation except on the 3th generation) so that we assume that the vectors of such observations have a multivariate normal distribution.

1. Difference on mean growth size at the end

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>192.86</td>
<td>192.86</td>
<td>7.84</td>
<td>0.0056</td>
</tr>
<tr>
<td>Error</td>
<td>198</td>
<td>4871.75</td>
<td>24.604795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>199</td>
<td>5064.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 3.2, it is clear that the mean growth size in the SQ group have a higher average than that in the NC group. From the p value in ANOVA table,
Table 3.3: Parameter Estimations

| Parameter | Estimate | Standard Error | t Value | Pr > | |t| |
|-----------|----------|----------------|---------|-------|
| Intercept | 6.005    | 0.496          | 12.11   | < .0001 |
| NC        | -1.964   | 0.701          | -2.80   | 0.0056 |

Table 3.4: Parameter Estimations (roundante)

| Parameter | Estimate | Standard Error | t Value | Pr > | |t| |
|-----------|----------|----------------|---------|-------|
| Intercept | 13.638   | 0.181          | 75.25   | < .0001 |
| roundante 0 | 1.155 | 0.313          | 3.69    | 0.0003 |

we conclude that at significant level $\alpha = 0.05$, the mean of the growth size for the two groups has a big difference.

While checking the parameter estimation, we observed that group effect is negative(-1.964). This indicates that the effect of agent type (with SQ as control group in R) will decrease on the whole. All suggest that at the end of the study, the SQ group will have a larger growth size than the NC group after 6 generations (60 years).

2. Mean difference for roundante at the beginning

Table 3.4 indicates a significant difference (p value=.0003) of the growth sizes between two groups having different roundante indicators at the beginning of the study. It also provides some evidence that the big difference of growth size as generation grows could come from the big gap when the study starts. This suggests that we need to be very cautious to make the conclusion when we carry out the repeated measurement analysis on the dataset for different agent types.

3. Mean difference for spawncost at the beginning

Similarly, the statistical interest for this section is to investigate if growth size with different spawncost indicators have difference on the growth sizes at the beginning of the study. We apply one way ANOVA analysis of one factor roundante (0,1) on the dataset and obtain the following results:

Table 3.5 indicates that no significant difference (p-value=0.2735) of the
Table 3.5: Parameter Estimations (spawncost)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.190</td>
<td>0.214</td>
<td>66.21</td>
<td>&lt; .0001</td>
<td></td>
</tr>
<tr>
<td>spawncost</td>
<td>-0.335</td>
<td>0.305</td>
<td>-1.10</td>
<td>0.2735</td>
<td></td>
</tr>
</tbody>
</table>

growth size exists between two groups with different spawncost indicators at the beginning of the study. When considering the difference on the pattern of change for the growth size with different spawncost indicators, the intercept or the mean at the first generation will not be considered.

3.4.3 Plot of Data

Figure 3.1 explains how the growth size grows as the generation changes from generation 1 to generation 5 for two different agent types. X-axis indicates the five generations while y-axis denotes the mean of growth size for each agent type at each time point. The plot shows that in the beginning, the two agents type do not show much difference from each other but the trend for each agents type becomes clearer as generation goes. Intuitively, this might suggest that each of agents type has significantly different effects as time goes on, but does not have much different effect between the two agent types. Although the overall trend for the growth size increase across the entire six generations, the relationship between growth size and generation is not linear according to the plot Figure 3.1. All indicate that we need to adopt a model that takes into account the correlation among the repeated measures at generations and a link function that could explain the linear relationship between the covariates and the mean of response variable.
Figure 3.1: Plot of the Raw Data by Agent
Chapter 4

Data Analysis by GEE

In this chapter, we will illustrate the usefulness of GEEs and GLS for the same sociology data set in the previous chapter. The primary purpose for the data set is to find if different agent type will have effect on the growth size, and if different agent types affect the growth size across generations as mentioned previously. Since the response variable is correlated among different generations, several covariance models are considered for the analysis: independence, compound symmetry, unstructured and AR(1). Both GLS method and GEE method will be applied to the dataset with different covariance matrices and we will compare the results to see if GLS method is more sensitive than GEE to the misspecification of covariance matrix.

4.1 Main Effect of Agent Type

Although the Shapiro-Wilks tests indicate that the observations at each generation (except the third generation) looks like normal distribution, thus the observations at entire time points may be assumed to be multivariate normal distribution, the response variable (growth size) varies nonlinearly with generations while the variance is approximately constant. Because of the concerns that observations of growth size from the same subject might be correlated, statistical methods such as GEE are used rather than a traditional general linear model (GLM) because GLM usually ignores the correlations between the repeated measures.
Table 4.1: Parameter Estimates & Empirical S.E for GEE Models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Independence</th>
<th>Compound Symmetry</th>
<th>Unstructured</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Pr &gt;</td>
<td>[Z]</td>
<td>Estimate</td>
</tr>
<tr>
<td>(\beta_0)</td>
<td>2.547</td>
<td>&lt; 0.001</td>
<td>2.547</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td></td>
<td>(0.040)</td>
<td></td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>-0.009</td>
<td>0.757</td>
<td>-0.010</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td></td>
<td>(0.029)</td>
<td></td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>0.024</td>
<td>0.022</td>
<td>0.024</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td></td>
<td>(0.010)</td>
<td></td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>0.024</td>
<td>0.002</td>
<td>0.024</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td></td>
<td>(0.008)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Parameter Estimates & S.E for GLS Methods

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Independence</th>
<th>Compound Symmetry</th>
<th>Unstructured</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Pr &gt;</td>
<td>[Z]</td>
<td>Estimate</td>
</tr>
<tr>
<td>(\beta_0)</td>
<td>2.533</td>
<td>&lt; 0.001</td>
<td>2.533</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td></td>
<td>(0.037)</td>
<td></td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.006</td>
<td>0.779</td>
<td>0.006</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td></td>
<td>(0.024)</td>
<td></td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>0.022</td>
<td>0.017</td>
<td>0.022</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td></td>
<td>(0.007)</td>
<td></td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>0.021</td>
<td>0.0005</td>
<td>0.021</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td></td>
<td>(0.004)</td>
<td></td>
</tr>
</tbody>
</table>

4.1.1 Statistical Model

The model can be written as:

\[
\log(E(y_{ij})) = \beta_0 + \beta_{1agent} + \beta_{2gen} + \beta_{3agent \ast gen} \tag{4.1}
\]

where \(\beta_1\) is the main effect of agent type, \(\beta_2\) is the main effect of generation, \(\beta_3\) is the interaction effect of agent by generation.

4.1.2 Output and Interpretation

Table 4.1 shows the results of fitting model for both agent types together using GEE method with different choice of covariance structures. Table 4.2 provides the results of fitting model for both agent types together using GLS method with different choice of covariance structures. Table 4.5 contains the model based covariance matrix estimates \(\hat{\Sigma}_{\beta}\) and the empirical covariance estimates \(\hat{\Sigma}_{\beta}^{R}\) under AR(1) assumption.
Table 4.3: $R$ for GEE with log link function and normal distribution

<table>
<thead>
<tr>
<th></th>
<th>Gen 1</th>
<th>Gen 2</th>
<th>Gen 3</th>
<th>Gen 4</th>
<th>Gen 5</th>
<th>Gen 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compound symmetry ((\rho=0.371))</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
<td>1.00</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>3</td>
<td>0.37</td>
<td>0.37</td>
<td>1.00</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1.00</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Unstructured</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.44</td>
<td>0.18</td>
<td>−0.16</td>
<td>−0.45</td>
<td>−0.55</td>
</tr>
<tr>
<td>2</td>
<td>0.44</td>
<td>1.00</td>
<td>0.21</td>
<td>0.07</td>
<td>−0.08</td>
<td>−0.14</td>
</tr>
<tr>
<td>3</td>
<td>0.18</td>
<td>0.21</td>
<td>1.00</td>
<td>0.26</td>
<td>0.28</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>−0.16</td>
<td>0.07</td>
<td>0.26</td>
<td>1.00</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td>5</td>
<td>−0.45</td>
<td>−0.08</td>
<td>0.28</td>
<td>0.73</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>6</td>
<td>−0.55</td>
<td>−0.14</td>
<td>0.30</td>
<td>0.83</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>AR(1) ((\rho=0.436))</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.59</td>
<td>0.35</td>
<td>0.21</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>0.59</td>
<td>1.00</td>
<td>0.59</td>
<td>0.35</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.59</td>
<td>1.00</td>
<td>0.59</td>
<td>0.35</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.35</td>
<td>0.59</td>
<td>1.00</td>
<td>0.59</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>0.12</td>
<td>0.21</td>
<td>0.35</td>
<td>0.59</td>
<td>1.00</td>
<td>0.59</td>
</tr>
<tr>
<td>6</td>
<td>0.07</td>
<td>0.12</td>
<td>0.21</td>
<td>0.35</td>
<td>0.59</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Since there is no special rule about which method is to be preferred in practical problems, only the estimates of S.E with empirical method are provided in Table 4.1. Note that S.E for the component of $\hat{\beta}$ are taking the square roots of the diagonal elements of $V^R_{\beta}$.

Compare p value in Table 4.1, the generation effect ($\beta_2$) is moderately significant under independence model; $\hat{\beta}_2 = 0.024$ with a robust S.E of 0.010, yielding a p value for a hypothesis test of $\beta_2 = 0$ of 0.022. The results are also suggestive of an effect of generation under compound symmetry model with p value equal to 0.022, while the generation effect are highly significant under unstructured model and AR(1) model with p value less than 0.001 and equal to 0.006 respectively. In all cases, there does not seem to be any evidence that $\beta_3$ is different from zero.

Given p value equal to 0.757 for independence assumption, 0.733 for compound symmetry assumption, 0.447 for unstructured assumption and 0.995 for AR(1) assumption for estimated $\beta_2$, it appears that we are safe to conclude that there is no enough evidence to suggest the difference in mean response between the two agent types.

The p value equal to 0.002 for both independence assumption, 0.014 for unstructured assumption and 0.005 for AR(1) assumption for the interaction between agents type and generation ($\beta_3$) indicates agent type could affect the development of the subjects by increasing or decreasing the growth size; this suggests that the two agent types at each generation point do not differ significantly, whereas their effects differ from each other as generation changes. This may suggest a different pattern of change of the two agent type effects, which we will verify soon. In conclusion, the two agent types differ in their effects on the development of the societies as generation goes by, which is reflected by growth size in each subject.

Inspecting the results presented in Table 4.2 supports that the estimation of $\beta$ with GLS method under independence, compound symmetry and AR(1) models are largely consistent. For example, p values for $\beta_2$ is 0.017 suggesting that generation effect is moderately significant under independence model, this result is identical to the conclusion drawn under compound symmetry and AR(1) model with p value equal to 0.001 and 0.004 respectively. The estimates for $\beta_2$ under unstructured model is 0.006 with a S.E 0.010, giving p value 0.546 for insignificant result of the generation
Table 4.4: Covariance matrix (Model based & Empirical)

<table>
<thead>
<tr>
<th></th>
<th>Prm1</th>
<th>Prm2</th>
<th>Prm3</th>
<th>Prm4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prm1</td>
<td>0.001697</td>
<td>-0.001006</td>
<td>-0.000373</td>
<td>0.0002204</td>
</tr>
<tr>
<td>Prm2</td>
<td>-0.001006</td>
<td>0.0006610</td>
<td>0.0002204</td>
<td>-0.000144</td>
</tr>
<tr>
<td>Prm3</td>
<td>-0.000373</td>
<td>0.0002204</td>
<td>0.0000980</td>
<td>-0.000058</td>
</tr>
<tr>
<td>Prm4</td>
<td>0.0002204</td>
<td>-0.000144</td>
<td>-0.000058</td>
<td>0.0000375</td>
</tr>
</tbody>
</table>

Covariance Matrix (Empirical)

<table>
<thead>
<tr>
<th></th>
<th>Prm1</th>
<th>Prm2</th>
<th>Prm3</th>
<th>Prm4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prm1</td>
<td>0.001583</td>
<td>-0.001072</td>
<td>-0.000282</td>
<td>0.0002172</td>
</tr>
<tr>
<td>Prm2</td>
<td>-0.001072</td>
<td>0.0008172</td>
<td>0.0002172</td>
<td>-0.000185</td>
</tr>
<tr>
<td>Prm3</td>
<td>-0.000282</td>
<td>0.0002172</td>
<td>0.0001087</td>
<td>-0.000078</td>
</tr>
<tr>
<td>Prm4</td>
<td>0.0002172</td>
<td>-0.000185</td>
<td>-0.000078</td>
<td>0.0000634</td>
</tr>
</tbody>
</table>

We thus conclude that all the fits under GEE method with different candidate covariance structures are qualitative very similar, and since we have no formal method of choosing among the various working correlation assumptions, we feel reasonably confident that results are not too dependent on the correlation assumption by using GEE method. For GLS method, the fits under independence, compound symmetry and AR(1) are qualitative similar but the fit under unstructured model leads to different conclusion, so that there seem to be no evidence that GLS method is robust to the different specification of dependence structure. Therefore, in the following analysis, we will restrict our analysis to GEE method.

The different estimated working correlation structures and estimated $\rho$ are presented in Table 4.3.

4.2 Pattern of Change by Agent Type

The analysis in the previous section indicates the growth size don’t have a significant difference among the two agent types, however, the difference becomes significant as generation goes by. The main goal in this section is to compare the intercept and slope which reveals the rate of change of the mean response over generation among the two agent types. We again consider a regression model with link function
of logarithm that says the logarithm of the mean response is a straight line with separate intercept and slope for each agent type. We will use GEE method to fit the model and compare the estimates of the regression coefficients for different covariance structures.

4.2.1 Statistical Model

Define
\[ t_{ij} = \begin{cases} 
1, & \text{subject } i \text{ at the } j\text{th round} \\
0, & \text{otherwise.} 
\end{cases} \]

\[
\log(Y_{ij}) = \beta_1 + \beta_2 t_{ij}, \quad \text{subject in NC} \tag{4.2}
\]

\[
\log(Y_{ij}) = \beta_3 + \beta_4 t_{ij}, \quad \text{subject in SQ} \tag{4.3}
\]

It is more convenient however entirely equivalent to write the model in an alternative parameterization. Since the question of interest is to compare the rate of change of the mean response over generations between two agents types, we would like to compare the two slopes \( \beta_2 \) and \( \beta_4 \). Define
\[
\delta_i = \begin{cases} 
1, & \text{subject } i \text{ in NC group} \\
0, & \text{otherwise.} 
\end{cases}
\]

Then we could write the model as
\[
\log(Y_{ij}) = \beta_1 + \beta_2 \delta_i + \beta_3 t_{ij} + \beta_4 \delta_i t_{ij} \tag{4.4}
\]

4.2.2 Hypothesis Tests

We can test for \( H_0 : \beta_2 = 0 \) and \( H_0 : \beta_4 = 0 \) respectively to check if the significant variation in the two agents type effect is due to different intercepts or slopes.

4.2.3 Output in SAS and interpretation

Table 4.5 suggests that there is no significant difference in the two intercepts for the two agents type with p value equal to 0.757 for independence assumption, 0.733 for compound symmetry, 0.447 for unstructured and 0.955 for AR(1) , whereas the difference in two slopes is significant with both p values smaller than 0.001 for all the covariance structures. Therefore, we can conclude that difference in two agent types are due to distinctive slopes. That is, there is a difference in the pattern of change of growth size between subjects in NC group and SQ group over every
Figure 4.1: Plot by agent type
Table 4.5: Parameter Estimates & Empirical S.E for GEE Models

| Parameter | Independence Estimate | Pr > |Z| | Compound Symmetry Estimate | Pr > |Z| | Unstructured Estimate | Pr > |Z| | AR(1) Estimate | Pr > |Z| |
|-----------|------------------------|------|---|-----------------------------|------|---|------------------------|------|---|------------------------|------|---|------------------------|------|---|
| $\beta_1$ | 2.528 (0.024) | < 0.001 | 2.528 (0.024) | < 0.001 | 2.320 (0.035) | < 0.001 | 2.532 (0.023) | < 0.001 |
| $\beta_2$ | 0.008 (0.029) | 0.757 | 0.098 (0.029) | 0.733 | 0.034 (0.044) | 0.447 | -0.002 (0.029) | 0.955 |
| $\beta_3$ | 0.048 (0.004) | < 0.001 | 0.048 (0.004) | < 0.001 | 0.108 (0.008) | < 0.001 | 0.053 (0.004) | < 0.001 |
| $\beta_4$ | 0.072 (0.007) | < 0.001 | 0.072 (0.007) | < 0.001 | 0.140 (0.010) | < 0.001 | 0.076 (0.007) | < 0.001 |

generation. Provided that the estimates for $\beta_4$ are larger than $\beta_3$ under all the candidate covariance structures we know SQ group has a larger slope. Because a larger slope corresponds to a faster increasing of growth size, we conclude that the SQ group shows result more quickly, thus quicker agent type effect.

Figure 4.2.3 explains how the logarithm of growth size grows as the generation changes from generation 1 to generation 6 for two different agent types. x-axis indicates the five generations while y-axis denotes logarithm of the mean of growth size for each agent type at each time point. Since logarithm is an increasing function, the interpretations about the effect of agent type and interaction of agent type and growth size will be exactly the same as those from Figure 3.1.

The plot shows that the subject from the two agent types don’t vary considerably in their growth size for the first generation and both SQ and NC groups have an increasing trend of growth size across the entire six generations, while the rate of the change of growth size in SQ group is bigger than that in NC groups which is confirmed from the above output ($\beta_4(0.076) > \beta_3(0.053)$, say AR(1) assumption).

As discussed in section (4.1), since all the fits with different covariance structure give the similar results, only the results under AR(1) will be presented in the following analysis.
4.3 Pattern of Change by Roundante

4.3.1 Statistical Model

Table 4.6 lists sample mean of the growth size at the first generation and the rest generations for NC group and SQ group respectively, grouped by roundante. The raw sample means in NC group suggest a consistent increasing trend in the mean response from the first generation to the last generation for roundante 0. Meanwhile, there is a decrease following an increase, which is then followed by an increase again from the first round to the fifth for roundante 1.

The raw sample means in SQ group suggest a roughly consistent increasing trend in the mean response from the first generation to the last for both roundantes. Based on the observations, we might adopt a model for mean response that allows the possibility of different means at baseline and each following generation. According to the previous analysis for the pattern of change by agent type, again a multivariate normal distribution with a log link function is chosen to model these data using GEE method. The question of interest here is to compare the rate of change of the mean response over generation between roundantes. We next compare the two slopes of each roundante.

Define
\[ \delta_i = \begin{cases} 
1, & \text{subject } i \text{ of roundante 0} \\
0, & \text{otherwise.}
\end{cases} \]

We could write the following model:
\[
\log (y_{ij}) = \beta_{0s} + \beta_{1s}\delta_i + \beta_{2s}t_{ij} + \beta_{3s}\delta_i t_{ij}
\]
(4.5)
Hypothesis Tests

We test for $H_0 : \beta_1 = 0$ and $H_0 : \beta_3 = 0$, respectively in (4.5) to check whether the significant difference in the change of the growth size is due to the difference in intercepts or in slopes, in other words, whether there is a significant association between the pattern of change of the responses and roundantes.

Output in SAS and interpretation

| Parameter | Estimate | S.E | Pr > |Z|  | Parameter | Estimate | S.E | Pr > |Z|  |
|-----------|----------|----|-----|---|-----------|----------|----|-----|---|-----------|----------|----|-----|---|
| $\beta_{0s}$ | 2.492 | 0.018 | < .001 | | $\beta_{0s}$ | 2.531 | 0.030 | < .001 | |
| $\beta_{1s}$ | 0.114 | 0.034 | 0.001 | 0.002 | 0.048 | 0.964 | | $\beta_{2s}$ | 0.050 | 0.007 | < .0001 | 0.093 | 0.010 | < .001 | |
| $\beta_{3s}$ | 0.056 | 0.005 | < .0001 | 0.066 | 0.009 | < .001 | |

The p value for testing $H_0 : \beta_{1s} = 0$ under AR(1) model is less than 0.001, which suggests a highly significant difference in two intercepts for both roundantes in NC group, while the p value for testing $H_0 : \beta_{3s} = 0$ under AR(1) model is less than 0.001, which indicates that the slopes are significantly different for the roundantes in NC group.
Given the p value above, we conclude that, in NC group, both the slopes and intercept for the association between growth size and roundantes are significantly different. Also there is a significant pattern of change between the responses and roundantes.

Considering the larger estimate for roundante 1 (0.056), it might indicate that NC group brings about faster growth size in roundante 1 versus roundante 0. This is because a larger slope corresponds to quicker increasing of growth size.

The p value as shown in Table 4.7 for testing $H_0 : \beta_{12} = 0$ under AR(1) is 0.964, which suggests the lack of significant difference in the two intercepts for both roundantes in SQ group. On the other hand, the p value for testing $H_0 : \beta_{32} = 0$ is less than 0.001, which implies that the slopes for the two roundantes are significantly different.

Therefore, we conclude that roundantes are significantly associated with the pattern of change of the responses in SQ group. Also intercepts shows that in the beginning of the study, the difference of growth sizes between roundantes is insignificant.

Considering the fact that the roundante 0 (0.093) case has a larger estimate than the roundante 1 (0.066) case, it might indicate that SQ groups brings about growth size more swiftly in roundante 0 versus roundante 1. This is because a larger slope corresponds to a faster increasing of growth size, thus a quicker roundante effect.

In addition, here we use the logarithm as the link function, which ensures the inference since it is an increasing function.

### 4.4 Pattern of Change by Spawncost

#### 4.4.1 Statistical Model

Table 4.8 lists a sample growth size at the first generation and the following rounds for the NC and SQ group, grouped by spawncost indicator. The raw sample means in NC group suggest there is a decrease following an increase, which is then followed by an increase again from the baseline to the fifth generation for spawncost 1. While there is a consistent increasing trend in the mean response from the first generation
to the last for spawncost 0 in NC group. This suggests an unstable trend among the five generations for NC group.

Table 4.8: Sample mean growth size at each generation

<table>
<thead>
<tr>
<th>Generation</th>
<th>NC</th>
<th></th>
<th>NC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>spawncost 0</td>
<td>spawncost 1</td>
<td>spawncost 0</td>
<td>spawncost 1</td>
</tr>
<tr>
<td>4</td>
<td>14.592</td>
<td>14.624</td>
<td>15.904</td>
<td>15.656</td>
</tr>
<tr>
<td>5</td>
<td>15.596</td>
<td>15.576</td>
<td>17.772</td>
<td>17.016</td>
</tr>
<tr>
<td>6</td>
<td>17.761</td>
<td>17.590</td>
<td>21.120</td>
<td>19.722</td>
</tr>
<tr>
<td>average over generation 1-6</td>
<td>14.941</td>
<td>15.105</td>
<td>16.479</td>
<td>16.083</td>
</tr>
</tbody>
</table>

Similar to the previous analysis, we chose GEE method to fit the model. Based on the above observations, we adopt a model for mean response that allows different means at the first generation and generation 1 to 6, whereas means at generation 1 to 6 are the same. These might be different over groups (spawncost indicator). Defining

\[ v_{ij} = \begin{cases} 
0, & \text{if } j = 0 \text{ (first generation)} \\
1, & \text{otherwise (generation 1-5)} 
\end{cases} \]

\[ \delta_i = \begin{cases} 
1, & \text{subject } i \text{ of spawncost 0} \\
0, & \text{otherwise} 
\end{cases} \]

we make use of the following model:

\[ y_{ij} = \beta_{0s} + \beta_{1s}\delta_i + \beta_{2s}v_{ij} + \beta_{3s}\delta_i v_{ij} \] (4.6)

4.4.2 Hypothesis Tests

We perform tests for \( H_0 : \beta_{1s} = 0 \) and \( H_0 : \beta_{3s} = 0 \), respectively in (4.6) to answer whether the significant difference in the change of the growth size is due to the difference in intercepts or in slopes, in other words, whether there exists a significant association between the pattern of change of the responses and spawncost.

4.4.3 Output in SAS and interpretation

The p value in Table 4.9 for testing \( H_0 : \beta_{11} = 0 \) in NC group is 0.077, which suggests that the two intercepts for both spawncost indicators are not significantly
Table 4.9: Parameter Estimates & Empirical S.E

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E</th>
<th>Pr &gt;</th>
<th>Z</th>
<th>Estimate</th>
<th>S.E</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.556</td>
<td>0.022</td>
<td>&lt; .001</td>
<td></td>
<td>2.550</td>
<td>0.032</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td>Spawncost0</td>
<td>-0.058</td>
<td>0.033</td>
<td>0.077</td>
<td></td>
<td>-0.037</td>
<td>0.047</td>
<td>0.434</td>
<td></td>
</tr>
<tr>
<td>Gen*Spawncost0</td>
<td>0.060</td>
<td>0.006</td>
<td>&lt; .001</td>
<td></td>
<td>0.084</td>
<td>0.010</td>
<td>&lt; .001</td>
<td></td>
</tr>
<tr>
<td>Gen*Spawncost1</td>
<td>0.048</td>
<td>0.006</td>
<td>&lt; .001</td>
<td></td>
<td>0.067</td>
<td>0.010</td>
<td>&lt; .001</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.4: Plot by spawncost

Figure 4.5: Plot by spawncost
different. The p value for testing $H_0: \beta_{31} = 0$ is $< 0.001$, suggesting that the slopes for the two spawncost indicators are significantly different.

We therefore conclude that there is a strong association between the pattern of changes and spawncost indicators in NC group. Considering the larger estimate for spawncost 0 (0.060), compared with that for spawncost 1(0.048), it might indicate that NC group brings about growth size more swiftly in those subjects who have less requirements for the spawn cost. This is because a larger slope corresponds to faster increasing of growth size for the subject.

The p value for testing $H_0: \beta_{12} = 0$ is 0.434, which suggests that the two intercepts for both spawncost indicators in SQ group are not significantly different. On the other hand, the fact that p value for testing $H_0: \beta_{32} = 0$ is $< 0.001$ suggests that the slopes for the two spawncosts in SQ group are significantly different.

Given the above information, we can conclude that there is a strong association between the pattern of change and spawncost indicator in SQ group. Notice here the result for intercepts coincides with the previous ANOVA analysis for spawncost, which also shows the absence of significant difference between the two spawncost indicators in the growth size at the beginning of the study. Considering the larger estimate for spawncost 0(0.084), compared with that for spawncost indicator 1(0.067), we know that SQ group might bring about effects on growth size more swiftly in those subjects which impose less requirements on the spawncost.

In addition, we use the logarithm as the link function, which ensures the inference for its monotonicity.

### 4.5 Pattern of Change by Roundante and Spawncost

Since both roundante and spawncost have different trends for the growth size, it is reasonable to adopt a model which only allows the possibility of a different mean at the first generation and generation 1 to 5. The means at generations are the same, but might be different by groups (different roundante and spawncost).

#### Model selection

Defining

$$\delta_{i1} = \begin{cases} 1, & \text{subject } i \text{ of roundante } 0 \\ 0, & \text{otherwise} \end{cases}$$
\[
\delta_{i2} = \begin{cases} 
1, & \text{subject } i \text{ of spawncost } 0 \\
0, & \text{otherwise} 
\end{cases}
\]

we write the model as follows

\[
y_{ij} = \beta_0 + \beta_1 \delta_{i1} + \beta_2 \delta_{i2} + \beta_3 v_{ij} + \beta_4 \delta_{i1} v_{ij} + \beta_5 \delta_{i2} v_{ij} + \beta_6 \delta_{i1} \delta_{i2} v_{ij} \tag{4.7}
\]

**Hypothesis Tests**

We can test for \( H_0 : \beta_1 = \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0 \) in (4.7) to investigate whether the significant difference in the change of the growth size is due to the difference in intercepts or in slopes, in other words, whether there is a significant association between the pattern of change of the responses and roundantes and spawncosts.

**Output in SAS and interpretation**

| Parameter | Estimate | S.E | \( Pr > |Z| \) | Estimate | S.E | \( Pr > |Z| \) |
|-----------|----------|-----|---------------|----------|-----|---------------|
| \( \beta_{0s} \) | 2.525 | 0.023 | \_0.001 | 2.547 | 0.036 | \_0.001 |
| \( \beta_{1s} \) | -0.079 | 0.029 | 0.007 | -0.036 | 0.047 | 0.443 |
| \( \beta_{2s} \) | 0.129 | 0.032 | \_0.001 | 0.007 | 0.049 | 0.884 |
| \( \beta_{3s} \) | 0.057(2) | 0.009 | \_0.001 | 0.096(1) | 0.013 | \_0.001 |
| \( \beta_{4s} \) | 0.039(4) | 0.009 | \_0.001 | 0.088(2) | 0.012 | \_0.001 |
| \( \beta_{5s} \) | 0.063(1) | 0.009 | \_0.001 | 0.075(3) | 0.013 | \_0.001 |
| \( \beta_{6s} \) | 0.052(3) | 0.007 | \_0.001 | 0.058(4) | 0.011 | \_0.001 |

The result in Table 4.10 suggests that the difference between intercepts for each roundante is highly significant with \( p \) value equal to 0.007 and the fact that \( p \) value less than 0.001 indicates that there is a highly significant difference between intercept for each spawncost in NC group. While \( p \) values equal to 0.443 and 0.884 respectively for testing \( H_0 : \beta_{12} = \beta_{22} = 0 \) implies that there is no significant difference between intercepts both for each roundante and for spawncost in SQ group.

For slopes, all the combinations are significantly different both in NC group and SQ group. Considering the estimates for each combination, we have a number of conclusions for NC group:

1. The effect on the subjects with roundante 0 and spawncost 1 is the fastest \( (\beta_{51} = 0.063) \).
2. the effect on those subjects with roundante 0 and spawncost 0 ranks the second fastest ($\beta_{31} = 0.057$)

3. the effect on those subjects with roundante 1 and spawncost 1 ranks the third ($\beta_{61} = 0.052$).

4. the slowest effect is with the subjects roundante 1 and spawncost 0 ($\beta_{41} = 0.039$).

Considering the estimates for each combination, we have the following conclusions for SQ group:

1. the SQ group has the fastest effect on subjects with roundante 0 and spawncost 0 ($\beta_{32} = 0.096$)

2. the SQ group has the second fastest effect on subjects with roundante 0 and spawncost 1 ($\beta_{42} = 0.088$).

3. the SQ group has the third fastest effect on subjects with roundante 0 and spawncost 1 ($\beta_{52} = 0.075$).

4. the slowest effect appears on subjects with roundante 1 and spawncost 1 ($\beta_{62} = 0.058$).
4.6 Conclusion

In this project, repeated measurement data are analyzed by GEE method and GLS method. For the classical data analysis, we focus on the analysis before the treatment. Even though the sample data might be multivariate normally distributed, we apply GEE method on discussing the main effect of the two groups (NC and SQ groups) and interactions between agent type and generation since the relationship between the mean response and generation is not linear. The comparison of the estimates of parameters of interest with different choice of covariance matrix implies that GEE method is more stable to the misspecification of the associations among repeated measurements within subjects.

Finally, we discuss the pattern of the change of the response between roundante and spawncost based on GEE model. We also have the following technical conclusion:

- We have strong evidence to conclude that subjects who have more or higher demands on the spawncost take more risks of developing a better society which is reflected by the growing population.

- In the study, we are mainly interested in the pattern of the change in growth size since this will give the researchers more detailed information on the difference of two groups of NC and SQ. Researchers expect that the time for improving the growth size could be reduced as soon as possible, i.e. the pattern of the change will be as high as they expect. We have strong confidence that there is a difference in the pattern of growth size between subjects who think and behave in two ways (NC and SQ) during the six generations (60 years). Subjects who are in the SQ group will develop faster than subjects who take NC group.

- For subjects in the NC group, both spawncost and roundante are risk factors that affects the pattern of the change in growth size indicating that NC group brings about positive effects on growth size more swiftly in those subjects who have less demanding on the spawncost or roundante. While checking the interaction of roundante and spawncost indicator, we confidently conclude that the interaction of roundante and spawncost can affect the pattern of the
change in growth size significantly.

- For subjects in the SQ group, we repeat the same analysis and finally have the converse conclusion of the significance of factor roundante and spawncost, i.e. neither roundante nor spawncost is a significant factor on the pattern of the change in growth size. While checking the interaction of roundante and spawncost, we make the same conclusion as in NC group that the interaction of roundante and spawncost can affect the pattern of the change in growth size significantly. From above analysis, we conclude that NC group is more sensitive to agent type on growth size than SQ group.
Appendix A

Project Background by JS

A.1 Research interest

This project is targeted at better understanding what social worlds are possible. Many people dream of better worlds, but which of these would actually be stable enough to exist in the long term? This is the sort of question that we want to answer in this project. To answer this question in a way that does not resort to armchair theorizing (too removed from reality, influenced by the thinkers own frame of reference, and cognitively taxing) or actual real-world experimentation (too costly, difficult to collect accurate data, ethically questionable) A computer program in C++ is built so that it is capable of producing simulations that model possible worlds. The general methodology used in producing these simulations is what has become know as the ensemble approach to forecasting, as it is used in meteorological science. In this approach a series of simulations is run, each with slight differences in either the initial conditions (e.g. measurements of the weather such as temperature and pressure) or the variables that control the progression of the simulation. By running many of these simulation trials it is hoped that the results will overlap to such a degree that claims can be made about the likelihood of events that are repeatedly produced across many of the simulations actually occurring (e.g. Rain over an area occupied by a city). I co-opt this approach and generate populations of lagents (read: simple models of people) as my set of initial conditions. We are interested in four things: ¹

1. Which types of populations are the most successful? By most successful I mean

¹This chapter is the working of John Simpson from Department of Philosophy in University of Alberta. Email : John.Simpson@ualberta.ca, Johnsipson@gmail.com
some combination of longevity (the population doesn’t crash), size (bigger can be better, but not necessarily), and stability (does the make-up of the population in terms of the behaviours expressed stabilize or fluctuate and how?).

2. What environmental conditions produce the most successful populations? Since we are using an evolutionary model I expect that environmental conditions such as roundante and spawncost (explained in next section) will have a big and significant association with the population survival and composition, especially as conditions for survival or reproduction become harsher.

3. Which behaviours characterize the most successful and/or the least successful population? Each population is made up of agents that amount to complete descriptions of all the behaviours they could express in all the possible situations they might find themselves in. These behaviours are randomly assigned at the start of each trial (seeding of the random number generator allows for identical initial populations to be reproduced as necessary) and then redistributed to agents produced in future generations via a series of coin flips to see which behaviour is inherited from each parent. It should thus be expected that certain behaviours (or possibly collections of behaviours) will rise to higher levels of prominence while others will sink into obscurity to the degrees that they are useful. It is also possible that some behaviours will be irrelevant and these need to be tracked as well.

4. Is one agent type better than another? We have two types of agents: one the models choice the way that ideally rational agents would choose (i.e. From a perfectly/ideally objective perspective) and one that models choice in the way that, perhaps, more realistically captures how actual people make choices (i.e. Framed within a current situation that they must choose whether to try and move away from or not). The objectively choosing agents are called NC (for normal-classic) and the subjectively choosing agents are called SQ (for status-quo).
A.2 Broader Description of the model

Greater detail on the model is provided below. In addition to this we have available the exact C++ code that produces the simulations. Lastly, I will soon have available a fully annotated set of this code for perusal by non-programmers. The model I have developed for a series of ensemble simulation trials to help tease out the rational value of certain behaviours can be most easily described as having three main parts: The agents that represent rational decision making beings, the games that represent the combination of social as well as broader environmental constraints and payoffs, and the broader environment. Each of these will be discussed in turn.

A.2.1 The agents that represent rational decision making beings

The most important core component of the model is the agency template that is used to represent rational decision making beings. Each agent amounts to a complete description of a set of responses to all the possible situations that they might face, given the environment models that they operate in conjunction with. Note that there are actually two different types of agents, each of which represents a different kind of choice procedure. Since both agent types share the same general architecture and it is general architectures that will be presented first the details of their differences will be set aside and discussed later. At the core of each agent type is a base structure that enables the members of each type to perform three specific actions: they can play any $2 \times 2$ game, they are capable of assessing their status of their relationship with other agents in a non-trivial way, and they are able to respond to the actions of others in conjunction with the outcomes of the games that they play with each other. Agents were given the ability to play any $2 \times 2$ game because $2 \times 2$ games are the simplest models of agent interaction that can provide non-trivial decision making scenarios to both players. A standard method of classifying games, where a game is any situation where two or more players operate under a situation of strategic interdependence, is to make explicit the number of players and the total number of options available to each player. The format used to present this information indicates the total number of players via the total number of numbers presented in the classification schema and the number of choices available to each player by the values of these numbers. The numbers are separated by the character '$\times$', which is read as "by" and intended to be interpreted as a multiplication symbol such that
multiplying all the numbers together will result in the number of outcomes available in the game.

Simpler games result in at least one player not really being a player at all, but more of an observer. While there are certainly cases where people are observers against the decisions of others it is likely rarely the case that such observers are without choices, it is simply that these choices do not matter from the perspective of generating outcomes that are meaningfully different and so simpler $1 \times 1$ and $1 \times 2$ games can be modelled by $2 \times 2$ games. More complicated games and scenarios can also be modelled by $2 \times 2$ games although such modeling can result in a loss of information through this simplification. Still there is reason to believe that human beings often perform similar simplifications when facing choice scenarios, reducing decisions into binary choice scenarios. In this way we present ourselves with simple choices such as ”Should I stay or should I go now?” rather than the more complicated choice scenario that they simplify such as ”should I stay and do w or stay and do x or go and do y or go and do z?”. In this way $2 \times 2$ games are reasonable models for simplified real-world choices faced by real-world beings. $2 \times 2$ games also have the advantage of having a heavily explored core set of games with a penumbra of games that have yet to be given much attention. With the notable exceptions of Rappoport’s 198? taxonomy of all the $2 \times 2$ games where neither player is indifferent between possible outcomes and the intense focus that has been given to a handful of $2 \times 2$ games—in particular the Prisoner’s Dilemma, Stag Hunt, Battle of the Sexes, and Chicken (CITE) $2 \times 2$ games remain mostly unexamined. Some of this inattention is justified insofar as there is a large number of $2 \times 2$ games that would uncontroversially be considered.

Mixed strategies are often suggested as a way to solve such games, but they can take a long time to settle into the proper mixing for each player. Allowing the agents within the model I have designed to play any $2 \times 2$ game thus allows them to be used to investigate a wide variety of scenarios, both strictly theoretical/hypothetical and those intended to be more representative of the social environment faced by actual decision making beings. The way that agents are given the capacity to play any game is relatively simple. When each agent is created they are assigned a complete list of responses to all the possible scenarios that they might face (these assignments are random for the initial agents of a population and inherited with 50% probability
from parents during spawning). Since each scenario is basically a choice between
two options from a $2 \times 2$ game under the set of conditions such as who the agent
is playing against this complete list is set up such that it can be indexed by the
scenario presented and the conditions of the choice, which each index value pointing
to a binary switch. The position of the switch that results from the index tells the
agent which of the two actions available to them in the $2 \times 2$ game they are playing
to pursue.

A.2.2 The games that represent constraints and payoffs

The second important feature of Agents within the model is that they are capable of
assessing relationships that hold between them and other agents in a non-trivial way.
The basic assessment of relationships when interacting with other agents within a
system that is intended to model actual decision making scenarios was important to
include because I take it as being obvious that people behave differently in otherwise
identical situations depending on who they are interacting with. Most people simply
do not provide the same level of help and support to every other person that is out
there in the world and that the degree to which they change how they interact
with others depends on what they think, or how they think, of the other person. To
accommodate this every agent is capable of assigning all the agents that they interact
with into one of three categories. These categories can be roughly understood as
"Friend", "Neutral", or "Enemy", but it is important to understand that how agents
assign other agents to these categories may not match onto our understanding of
these terms. Agents simply have three categories and treat all agents within each
category the same. Whether they are nice to all the agents in the "Enemy" category
and mean to everyone in the "Friend" category or vice versa is something that is not
explicitly stipulated in the program. Further, being nice or mean is not necessarily
implemented across all possible situations within each category. It is much more
likely the case that various actions that we as observers might consider to be "nice"
or "mean" are distributed evenly in a random sort of way across across each category
during the start of each trial and so any attempt to anthropomorphize how agents
like each other should be strongly resisted. In addition to the basic relationship
assessment that results from noting that people interact differently with other people
depending on how they like them, I also take it as obvious that people further adjust
their behaviour towards others by taking into account how they believe that they other people see them. People just don’t often treat another person the same way if they believe that the other person hates them as opposed to believing that the other person likes them or at least has a more neutral stance towards them. As a mirror to the three stances that each agent might take towards another agent there are three stances that each agent supposes that the other agent might take towards them. Taking into account what has been said earlier about resisting the attraction of overly anthropomorphizing these agents these three stances can be understood as mapping onto: ”I think that you think that I am a Friend”, ”I think that you think that I am a Neutral Party”, and ”I think that you think that I am an Enemy”. These three stances in combination with the three listed immediately prior result in in nine different possible relationship types that each player can assess as standing between them and the other player whenever a choice is to be made. Having each agent capable of assessing both how they perceive the relationship and how they perceive how the other agent perceives the relationship simulates empathetic preferences (Cite Binmore). This is the ability to take into account how the other person sees the situation and to include that in decision making. Of course, agents always attempt to see the other agent’s perspective only from within their own perspective. Given that the agents initially generated to make up a population are randomly generated an interesting thing to look out for over the course of the trials will be whether or agents evolve to have their perspectives on their relationships and their responses to such perspectives evolve to match those of the other agents that they interact with.

A.2.3 The broader environment

The final important defining characteristic of the agents within my model is their ability to respond to the outcomes brought about by their interaction with other agents and the mechanism that creates this ability. This ability to respond to outcomes simply means that after a game has been played an agent looks at what both they and the other player actually received versus what was available to be received. Based on this assessment the agent is capable of reassigning both the the stance that they take towards the other agent across the three available categories and the stance that they see the agent they just interacted with as taking towards
them. This behaviour is also intended to create the possibility of mimicking the behaviour of actual human beings who respond to the outcomes that they experience as well. Reassignment to a different stance is not necessary, only a possibility; not all outcomes lead to people changing the way people see those they interact with. The results of an assessment of outcomes is determined through a module called AfterPlay. The heart of the AfterPlay module is a three dimensional matrix, internal to each agent, that takes as index values an assessment of whether the agent doing the assessment did better, worse, or the same as some standard drawn from the game, an assessment of whether the agent played against did better, worse, or the same as some standard also drawn from the game, and an assessment of the stance taken regarding the relationship between the two agents when they first entered the game. The standards vary between agent types are one of the ways that the two agent types modelled differ; again, this will be discussed in further detail later. The assessments within an agent type are objective however. Each agent involved in the game knows exactly what the other agent’s standards are, what they received and how this would be assessed by that agent. This is certainly not a perfect mirror of the capabilities of actual human beings, but it is a necessary simplification at this early stage of building this model. The matrix is used twice in the AfterPlay module by each agent. The first time is to see if the stance the assessing agent takes towards the agent just played with will be different in the future. The second time is to see if the stance that the assessing agent takes the agent just played against to be taking towards them will be different in the future. The assessments of better, worse, or no-change regarding the standards draw from the game are used in both assessments and are simply switched from one assessment to the other to represent the change of perspectives. The third index, the stance taken from one agent to the other, changes between assessments. One matrix is used for both assessments for two reasons. First, this is intended to be an extension of the notion of empathetic preferences. Agents assess the effect of outcomes on others, but must do so from their own view of the world. In short, they are doing what they would do if they were in the other agents position and had just experienced the outcomes of the game just played. Second, it helps to keep the AfterPlay module reasonably simple. I use "reasonably" in a rather loose way here. As it currently stands with three ways that the relationship stance can be assessed and three ways that each payoff on the
outcome can be valued there are 27 different possible combinations of indexes to the matrix. Since each cell in the matrix can take on one of three possible values as well (switch to "Friend", switch to "Enemy", switch to "Neutral") there are thus 327 or roughly 7.6 trillion uniquely different matrices that can exist within the AfterPlay module.

A.2.4 Two agents types

Around this three part framework two different agent types have been built, one to represent choice an embedded or framed perspective and one to represent choice from an objective or removed perspective. The first group are called Status-Quo (SQ) choosers and the latter are called Normal/Classic (NC) choosers. NC choosers are intended to represent the way that classical game or decision theory tells us that rational agents are supposed to choose. They see each situation without any personal attachment to it beyond the outcomes that they have available to both them and the other agent that they are playing with. In this way they represent ideally rational agents who are free from the sorts of framing effects that are believed to often influence the decision making of non-ideally rational decision makers. One clear example of a framing effect that is widely believed to be detrimental is the counting of sunk costs (Cite STEELE) and others arise from various psychological studies such as those done and inspired by Kahneman and Tversky (Cite). NC choosing agents achieve their objectivity by evaluating each game in its entirety; simply put, they know exactly which game they are playing and what player-position they have been assigned within the game. Given this knowledge, in conjunction with the knowledge of who they are playing against and what they believe the relationship status between them and the other agent to be, each agent makes a play as determined by their behavioural genetics. Since they are playing a $2 \times 2$ game their choice amounts to a decision between one of two possible actions.

Status-Quo (SQ) choosers are intended to represent something more like the way in which it is supposed that people make choices within a frame of reference. Rather than seeing the situation from a neutral or objective perspective, people often see themselves as already being in one of the possible outcomes. Choice thus becomes less about making the choice that will bring about the best outcome than it is about choosing whether or not to step away from what is seen as the current state of affairs.
or "status quo". SQ agents play the exact same games as the NC agents, but are both randomly assigned an outcome that will act as the status quo for the playing of the game. The agents are both assigned the same status quo outcome and they know exactly which outcome this is. From the assigned status quo position each agent makes an assessment of how the status quo position relates to the other three available outcomes. This status quo assessment consists of each agent taking the payoff of the status quo position and then determining whether or not each of the other available outcomes provides a better, worse, or identical outcome to the status quo position. The choice the SQ agent faces is different from that faced by the NC agent insofar as it is a choice about whether to move off of a default outcome where the NC agent is choosing free of any default position around which to frame their decision. As with the NC agents, SQ agents also have the possibility of choosing across nine different relationship types. The environment that the agents exist within is best seen as consisting of two parts, one physical and the other social. The physical environment is physical in the sense that it affects all agents and sets the rules within which they interact with each other. It is not physical in the sense that it is "real" or tangible in the way that the physical environment that we occupy is, but the constraints that it imposes on the agents within a simulation are just as real for those agents as our physical environment is for us; we must both live by these rules. Within the simulations the physical environment is composed of two things: games and a set of parameters that affects the agents outside games. The games agents play define how the agents may interact with each other. in this way they are a combination of both social and physical environmental conditions from the world we are used to. As mentioned earlier, these games are all $2 \times 2$ games so each agent is confined to a choice between two options within any game. Games also offer payoffs. Both players making a choice results in an outcome, one of the four possible combinations of choices between the two agents that are available, and each player receives a payoff on every outcome. Payoffs are in the form of points that agents can use to spend on things outside the context of the games. There are ten possible point values that an agent may receive, ranging from zero to nine.

A.2.5 The environment conditions

There are two parameters that will be considered in the study:
1. Generational Ante (roundante) - This is the point price that each agent must play in order to have survived the current generation. Agents that can pay this price are given the chance to spawn and allowed to continue into the next generation.

2. Spawning Cost (spawncost) - This is the point price that each agent must pay in order to spawn. Agents that can pay this price and find another agent that is willing to spawn with them will produce a single offspring with their partner. This offspring is generated by a series of virtual coin flips to decide which of each of its characteristics is inherited from which parent. Spawning costs are paid after the generational ante. Agents can spawn as many times as they have the points to pay for. Agents are released into the next generation with whatever packet change they have remaining after spawning as many times as possible.
Appendix B

R code for data analysis

R and SAS code used in this project is based on [17] and [9].

```R
library(lattice)

sdata <- read.table("e:/guox/thesis/sas/pop2.dat", head=T);
names(sdata) <- c("subid", "roundante", "spawcost", "gen", "oldagent", "agent", "pgrowth");
attach(sdata);

mean.NC <- c(mean(sdata$pgrowth[agent==1&gen==1]),
mean(sdata$pgrowth[agent==1&gen==2]),
mean(sdata$pgrowth[agent==1&gen==3]),
mean(sdata$pgrowth[agent==1&gen==4]),
mean(sdata$pgrowth[agent==1&gen==5]),
mean(sdata$pgrowth[agent==1&gen==6]));
mean.NC

mean.SQ <- c(mean(sdata$pgrowth[agent==2&gen==1]),
mean(sdata$pgrowth[agent==2&gen==2]),
mean(sdata$pgrowth[agent==2&gen==3]),
mean(sdata$pgrowth[agent==2&gen==4]),
mean(sdata$pgrowth[agent==2&gen==5]),
mean(sdata$pgrowth[agent==2&gen==6]));
mean.SQ

mean.NC.ave <- mean(mean.NC);
mean.SQ.ave <- mean(mean.SQ);

mean.NC.round0 <- c(mean(sdata$pgrowth[agent==1&gen==1&roundante==0]),
mean(sdata$pgrowth[agent==1&gen==2&roundante==0]),
mean(sdata$pgrowth[agent==1&gen==3&roundante==0]),
mean(sdata$pgrowth[agent==1&gen==4&roundante==0]),
mean(sdata$pgrowth[agent==1&gen==5&roundante==0]),
mean(sdata$pgrowth[agent==1&gen==6&roundante==0]),

52
```
mean(sdata$pgrowth[agent==1&gen==6&roundante ==0]);

mean.NC.round1 <- c(mean(sdata$pgrowth[agent==1&gen==1&roundante ==1]),
mean(sdata$pgrowth[agent==1&gen==2&roundante ==1]),
mean(sdata$pgrowth[agent==1&gen==3&roundante ==1]),
mean(sdata$pgrowth[agent==1&gen==4&roundante ==1]),
mean(sdata$pgrowth[agent==1&gen==5&roundante ==1]),
mean(sdata$pgrowth[agent==1&gen==6&roundante ==1]));
mean.NC.round <- cbind(mean.NC.round0, mean.NC.round1);

mean.SQ.round0 <- c(mean(sdata$pgrowth[agent==2&gen==1&roundante ==0]),
mean(sdata$pgrowth[agent==2&gen==2&roundante ==0]),
mean(sdata$pgrowth[agent==2&gen==3&roundante ==0]),
mean(sdata$pgrowth[agent==2&gen==4&roundante ==0]),
mean(sdata$pgrowth[agent==2&gen==5&roundante ==0]),
mean(sdata$pgrowth[agent==2&gen==6&roundante ==0]));
mean.SQ.round <- cbind(mean.SQ.round0, mean.SQ.round1);

mean.NC.round <- cbind(mean.NC.round0, mean.NC.round1);
mean.SQ.round <- cbind(mean.SQ.round0, mean.SQ.round1);

mean.NC.spawn0 <- c(mean(sdata$pgrowth[agent==1&gen==1&spawncost ==0]),
mean(sdata$pgrowth[agent==1&gen==2&spawncost ==0]),
mean(sdata$pgrowth[agent==1&gen==3&spawncost ==0]),
mean(sdata$pgrowth[agent==1&gen==4&spawncost ==0]),
mean(sdata$pgrowth[agent==1&gen==5&spawncost ==0]),
mean(sdata$pgrowth[agent==1&gen==6&spawncost ==0]));
```r
mean.NC.spawn1 <- c(mean(sdata$pgrowth[agent==1&gen==1&spawncost ==1]),
mean(sdata$pgrowth[agent==1&gen==2&spawncost ==1]),
mean(sdata$pgrowth[agent==1&gen==3&spawncost ==1]),
mean(sdata$pgrowth[agent==1&gen==4&spawncost ==1]),
mean(sdata$pgrowth[agent==1&gen==5&spawncost ==1]),
mean(sdata$pgrowth[agent==1&gen==6&spawncost ==1]));
mean.NC.spawn <- cbind(mean.NC.spawn0, mean.NC.spawn1);

mean.SQ.spawn0 <- c(mean(sdata$pgrowth[agent==2&gen==1&spawncost ==0]),
mean(sdata$pgrowth[agent==2&gen==2&spawncost ==0]),
mean(sdata$pgrowth[agent==2&gen==3&spawncost ==0]),
mean(sdata$pgrowth[agent==2&gen==4&spawncost ==0]),
mean(sdata$pgrowth[agent==2&gen==5&spawncost ==0]),
mean(sdata$pgrowth[agent==2&gen==6&spawncost ==0]));
mean.SQ.spawn1 <- c(mean(sdata$pgrowth[agent==2&gen==1&spawncost ==1]),
mean(sdata$pgrowth[agent==2&gen==2&spawncost ==1]),
mean(sdata$pgrowth[agent==2&gen==3&spawncost ==1]),
mean(sdata$pgrowth[agent==2&gen==4&spawncost ==1]),
mean(sdata$pgrowth[agent==2&gen==5&spawncost ==1]),
mean(sdata$pgrowth[agent==2&gen==6&spawncost ==1]));
mean.NC.spawn <- cbind(mean.NC.spawn0, mean.NC.spawn1);
mean.SQ.spawn <- cbind(mean.SQ.spawn0, mean.SQ.spawn1);

mean.NC.spawn
mean.SQ.spawn

mean.NC.spawn.ave <- c(mean(mean.NC.spawn0), mean(mean.NC.spawn1));
mean.SQ.spawn.ave <- c(mean(mean.SQ.spawn0), mean(mean.SQ.spawn1));

library(nlme);
sdata<- read.table("e:/guox/thesis/sas/pop2.dat", head=T);
names(sdata) <- c("subid", "roundante", "spawcost", 
"gen", "oldagent", "agent", "pgrowth");
sdata$logpgrowth <- log(sdata$pgrowth);
attach(sdata);

#Fit the data with difference variance and covariance matrix by 
#using Generalized 
#Least Square estimates.(GLS)
longg <- groupedData(logpgrowth~agent*gen|subid, data=sdata);
#1. Independence
```

54
fit.in <- gls(logpgrowth ~ agent+gen+(agent)*(gen), data=longg
    ## corr=corCompSymm(, form= ~ 1 | subid)
    ,method="ML"
    );
summary(fit.in) ;

#2. Compound Symmetric
fit.cs <- gls(logpgrowth ~ agent+gen+(agent)*(gen), data=longg,
    corr=corCompSymm(, form= ~ 1 | subid),method="ML"
    );
summary(fit.cs) ;

#3. Unstructure
fit.un <- gls(logpgrowth ~ (agent)*(gen), data=longg,
    corr=corSymm(form = ~ 1 | subid),
    weight=varIdent(form = ~ 1 | gen),method="ML"
    );
summary(fit.un);

#4. AR(1)
fit.ar1 <- gls(logpgrowth ~ (agent)*(gen), data=longg,
    corr=corAR1(, form= ~ 1 | subid) ,method="ML"
    )
summary(fit.ar1)
anova(fit.ar1)

fit.arh1 <- gls(logpgrowth ~ (agent)*(gen), data=longg,
    corr=corAR1(, form= ~ 1 | subid),
    weight=varIdent(form = ~ 1 | gen),method="ML"
    )
summary(fit.arh1)

#Model comparison
anova(fit.cs, fit.in);
anova(fit.cs, fit.un);
anova(fit.cs, fit.ar1);
anova(fit.cs, fit.arh1);
anova(fit.un, fit.arh1);

detach(sdata);

#Model 1:(gen,agent,roundante,spawcost) —using aov function
attach(sdata);
win.graph();

par(cex=.6)
interaction.plot(gen[agent==1], roundante[agent==1],
    logpgrowth[agent==1],
    ylim=c(2.5, 3.2), lty=c(1, 12), lwd=3,
    trace.label="roundante",
    ylab="mean of logpgrowth", xlab="gen");
title("agent=SQ");
dev.copy2eps(file="SQ_GLS.eps");
win.graph();
par(cex=.6)
interaction.plot(gen[agent==2], roundante[agent==2],
logpgrowth[agent==2],
ylim=c(2.5, 3.2), lty=c(1, 12), lwd=3,
trace.label="roundante",
ylab="mean of logpgrowth", xlab="gen")
title("agent=NC")

dev.copy2eps(file="NC/GLS.eps");
detach(sdata);

# Model 1: (gen, agent, roundante) — using the gls function
attach(sdata);
longa <- groupedData(logpgrowth~agent*roundante*gen |
subid, data=sdata);
both.arh1 <- gls(logpgrowth ~(agent)*(roundante)*(gen),
data=longa, corr=corAR1(), form= ~ 1 | subid),
weight=varIdent(form = ~ 1 | gen));
summary(both.arh1);
anova(both.arh1);

detach(sdata);
## ## # Contrasts and interaction contrasts for model 1

## attach(sdata);
## longa$af <- factor(longa$agent, c(1,2));
## longa$rf <- factor(longa$roundante, c(1,2));
## longa$gf <- factor(longa$gen, c(1,2,3,4,5,6));

## contrasts(longa$ef) <- c(-1/2,-1/2);
## contrasts(longa$tf) <- c(-1/2,-1/2);
## contrasts(longa$df) <- c(-1/2, 1/2)
## contrasts(longa$df)
## [,1]
## 1 -0.5
## 2 0.5
## contrasts(longa$ef)

## detach(sdata);
Appendix C

SAS code for data analysis

/***************************************************************************/
/* VARIABLE DESCRIPTION */
/***************************************************************************/
/*                                                                 */
1 Pop id
2 Roundante (= 0 if less, = 1 if more)
3 Spawn cost indicator (= 0 if less, = 1 if more)
4 Generation (6 in total with generation 1 as the baseline)
5 Population size (the response, count)
6 Agent indicator (= 1 if NC, = 2 if SQ)
/***************************************************************************/
/*                                                                 */

The data are not in the correct form for use with the SAS procedures
CORR and DISCRIM we use below. These procedures require that the
data be in the form of one record (line) per experimental unit.
The data in the file dental.dat are in the form of one record per
observation (so that each child has 4 data records).

In particular, the data set looks like

column 1   sub id number
column 3   gen
column 4   response (distance)

We thus create a new data set such that each record in the data
set represents all 4 observations on each child plus spawn cost
identifier. To do this, we use some data manipulation features
of the SAS data step. The second data step does this.

We redefine the values of gen so that we may use gen as an "index"
in creating the new data set POP2. The DATA step that creates
POP2 demonstrates one way (using the notion of an ARRAY) to transform a data set in the form of one observation per record (the original form) into a data set in the form of one record per individual. The data must be sorted prior to this operation; we invoke PROC SORT for this purpose.

In the new data set, the observations at gens 8, 10, 12, and 14 are placed in variables gen1, gen2, gen3, and gen4, respectively.

We use PROC PRINT to print out the first 5 records (so data for the first 5 subjects) using the OBS= feature of the DATA= option.

***********************************************************************************************************************************************
/* Alternatively, replace the above code with the following using spawndisc playdisc proc transpose (thanks to Laine Elliott for noting this):
*/

options ls=80 ps=59 nodate; run;

data pop0; infile 'e:\guox\thesis\sas\pop2.dat' FIRSTOBS=2;
input subid roundante spawncost gen oldagent agent0 pgrowth ;
ga=agent0*gen;
run;

data sdata; set pop0;
if roundante=0 & spawncost =0 then roundspawn=0;
if roundante=0 & spawncost =1 then roundspawn=1;
if roundante=1 & spawncost =0 then roundspawn=2;
if roundante=1 & spawncost =1 then roundspawn=3;
if agent0 =1 then agent =’NC’;
if agent0=2 then agent =’SQ’;
run;

proc print data=sdata; run;
/*Check if it is normally distributed*/
PROC UNIVARIATE PLOT NORMAL;
VAR pgrowth;
histogram pgrowth;
RUN;

data popnorm1; set sdata;
if agent = ’SQ’ then delete;
PROC UNIVARIATE PLOT NORMAL;
  var pgrowth;
  histogram pgrowth;
RUN;

data popnorm2; set sdata;
if agent = 'NC' then delete;

PROC UNIVARIATE PLOT NORMAL;
VAR pgrowth;
histogram pgrowth;
RUN;

proc print data=sdata; run;

proc transpose data=sdata out=tdata prefix=gen;
  by subid roundante spawncost agent notsorted;
  var pgrowth;
run;

proc print data=tdata(obs=20); run;

/*
Analysis of data at each time point examines agent effects separately
at individual observation times and makes no statistical comparisons
among times. No inference is drawn about trends over time, so this
method is not truly a repeated measures analysis. Use of analysis at
each time point is usually at a preliminary stage of data analysis and
is not a preferred method for final publication because it does not
address time effects. SAS statements to obtain analyses at each time
point are:
*/

PROC GLM DATA = tdata;
CLASS agent;
MODEL gen1−gen6 = agent;
MEANS agent/LSD;
ESTIMATE 'agent NC − agent SQ' agent 1 −1 ;
RUN;
title "PROFILE PLOT";
proc glm data=tdata;
  class agent;
  model gen1 gen2 gen3 gen4 gen5 gen6 = agent;
  repeated gen6;
  lsmeans agent / out=means;
  goptions reset=all;

filename output 'e:\guox\thesis\output\agent.eps';
goptions device=pslepsfc gfname=output gsfmode=replace;
ymbol1 interpol=spline width=4 color=blue line=10;
ymbol1 c=blue v=square h=1 i=j width=3;
ymbol2 c=red v=square h=1 i=j width=3;
axis1 order=(13 to 21 by 1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize=10 vsize=8;*/
/* graph procedure goes here */
proc gplot data=means;
  plot lsmean*NAME=agent / vaxis=axis1 haxis=axis2;
run;
quit;

/*Here we need to give the plot to show that there is no difference between the pattern of the two agents.*/

/*MAIN EFFECT OF AGENT TYPE WITH GEE METHOD*/

proc genmod data=sdata;
class subid;
model pgrowth= agent0 gen ga/ dist=normal link=log;
repeated subject=subid / type=IND corrw covb sorted;
run;

proc genmod data=sdata;
class subid;
model pgrowth= agent0 gen ga/ dist=normal link=log;
repeated subject=subid / type=cs corrw covb sorted;
run;

proc genmod data=sdata;
class subid;
model pgrowth=agent0  gen ga/ dist=normal link=log;
repeated subject=subid / type=un corrw covb sorted;
run;
proc genmod data=sdata;
class subid;
model pgrowth=agent0 gen ga/ dist=normal link=log;
repeated subject=subid / type=ar(1) corrw covb sorted;
run;
quit;

title "Question 2.2 SEPERATE SLOPE AND INTERCEPT";
proc genmod data=sdata;
class agent subid;
model pgrowth=agent agent*gen/ dist=normal link=log;
repeated subject=subid / type=IND corrw covb sorted;
run;

proc genmod data=sdata;
class agent subid;
model pgrowth=agent agent*gen/ dist=normal link=log;
repeated subject=subid / type=cs corrw covb sorted;
run;

proc genmod data=sdata;
class agent subid;
model pgrowth=agent agent*gen/ dist=normal link=log;
repeated subject=subid / type=un corrw covb sorted;
run;
proc genmod data=sdata;
class agent subid;
model pgrowth=agent agent*gen/ dist=normal link=log;
repeated subject=subid / type=ar(1) corrw covb sorted;
run;

 Commentary:

/* Question 3 : difference in pop size at the end of the study, check
the effect of the agent on the pop size */

/* 3.1: diff: the difference of the nail length between 12 gen and
0 gen */

data tdata1; set tdata;
diff=gen6-gen1;
/* 3.2: Check if difference is normally distributed*/
title "Question 3.1 NORMALITY OF DIFF";
PROC UNIVARIATE plot NORMAL;
VAR diff;
RUN;
proc print data=tdata1(obs=20); run;
proc sort data=tdata1; by agent; run;

title 'Question 3.2 SIMPLE SUMMARY STATISTICS BY AGENT';
proc means data=tdata1 maxdec=2 n mean std;
by agent; /* statistics computed for each oil type... */
var diff; /* ... on the variable 'visc' */
run;

title 'Question 3.3 PLOT OF THE RAW DATA BY AGENT';
proc plot data=tdata1;
/* request a plot of the raw data */
plot diff*agent;
run;

title 'Question 3.4 PROC GLM ANALYSIS BY AGENT';

proc glm data=tdata1;
class agent;
model diff = agent/ solution;
output out=tdata2 p=yhat r=resid;
/* store fitted values and fitted residuals in dataset called 'oilfit' for later use */
run;

title "RESIDUAL ANALYSIS BY AGENT";
proc univariate data=tdata2 plot normal;
var resid;
/* plot qq-plot of fitted residuals */
run;
proc plot;
plot resid*agent;
plot resid*yhat;
/* two residual plots to check independence and constant variance */
run;

/*Question 4: Spawncost 0 creates more population size at the beginning*/
/*Check if the first gen's data is normally distributed*/

/* 4.1: Check if difference is normally distributed*/
title "Question 4.1 : NORMALITY OF DIFF";
PROC UNIVARIATE plot NORMAL;
VAR gen1;
RUN;

/*4.2: One way ANOVA */
proc sort data=tdata1; by spawncost; run;
title 'Question 4.2: SIMPLE SUMMARY STATISTICS BY SPAWNCOST';
proc means data=tdata1 maxdec=2 n mean std;
/* get simple summary statistics (sample size, sample mean and SD) with
max of 2 decimal places */
by spawncost; /* statistics computed for each oil type... */
var gen1; /* ... on the variable 'visc' */
run;

title 'Question 4.3: PLOT OF THE RAW DATA BY SPAWNCOST';
proc plot data=tdata1;
/* request a plot of the raw data */
plot gen1*spawncost;
run;

title 'Question 4.4: PROC GLM ANALYSIS BY SPAWNCOST';
proc glm data=tdata1;
/* same as 'proc anova' except
'glm' allows residual plots but gives more junk output */
class spawncost;
model gen1 = spawncost / solution;
output out=tdata2 p=yhat r=resid;
/* store fitted values and fitted residuals
in dataset called 'oifit' for later use */
run;

title "Question 4.5: RESIDUAL ANALYSIS BY SPAWNCOST";
proc univariate data=tdata2 plot normal;
var resid;
/* plot qq-plot of fitted residuals */
run;

proc plot;
plot resid*spawncost;
plot resid*yhat;
/* two residual plots to check
independence and constant variance */
run;
/*Check if the first gen’s data is normally distributed*/
/*5.1 is done in question 4*/

/*5.2: One way ANOVA*/
proc sort data=tdata1; by roundante; run;

title 'Question 5.1 : SIMPLE SUMMARY STATISTICS BY roundante';
proc means data=tdata1 maxdec=2 n mean std;
/* get simple summary statistics (sample size, sample mean and SD) with
max of 2 decimal places */
by roundante; /* statistics computed for each oil type... */
var gen1; /* ... on the variable 'visc' */
run;

title 'Question 5.2 : PLOT OF THE RAW DATA BY roundante';
proc plot data=tdata1;
/* request a plot of the raw data */
plot gen1*roundante;
run;

title 'Question 5.3 : PROC GLM ANALYSIS BY roundante';
proc glm data=tdata1;
/* same as 'proc anova' except
'glm' allows residual plots but gives more junk output */
class roundante;
model gen1 = roundante / solution;
output out=tdata2 p=yhat r=resid;
/* store fitted values and fitted residuals
in dataset called 'oifit' for later use */
run;

title "Question 5.4 : RESIDUAL ANALYSIS BY roundante";
proc univariate data=tdata2 plot normal;
var resid;
/* plot qq-plot of fitted residuals */
run;
ods tagsets.simplelatex file="e:\guox\thesis\output\legacy.tex"
stylesheet="e:\guox\thesis\output\sas.sty"(url="sas");
/*ods tagsets.latex file="c:\temp\legacy.tex";*/
proc plot;
plot resid*roundante;

64
plot resid\*yhat;
/* two residual plots to check
independence and constant variance */
run;
/*ods tagsets.latex close;*/
proc print data=tdata2(obs=20); run;

/**
Question 6 Pattern of the change of the pgrowth.
*/
title "Question 6.1: PATTERN OF CHANGE OF SIZE BY ROUNDANTE";
proc genmod data=sdata; by agent;
class subid roundante;
model pgrowth= roundante roundante*gen / dist=normal link=log;
repeated subject=subid / type=IND corrw covb sorted;
run;

proc genmod data=sdata; by agent;
class subid roundante;
model pgrowth= roundante roundante*gen / dist=normal link=log;
repeated subject=subid / type=CS corrw covb sorted;
run;
proc genmod data=sdata; by agent;
class subid roundante;
model pgrowth= roundante roundante*gen / dist=normal link=log;
repeated subject=subid / type=UN corrw covb sorted;
run;
proc genmod data=sdata; by agent;
class subid roundante;
model pgrowth= roundante roundante*gen / dist=normal link=log;
repeated subject=subid / type=ar(1) corrw covb sorted;
run;

title "Question 6.2: PATTERN OF CHANGE OF SIZE BY SPAWNCOST ";
proc genmod data=sdata; by agent;
class subid spawncost;
model pgrowth= spawncost spawncost*gen / dist=normal link=log;
repeated subject=subid / type=IND corrw covb sorted;
run;
proc genmod data=sdata; by agent;
class subid spawncost;
model pgrowth= spawncost spawncost*gen / dist=normal link=log;
repeated subject=subid / type=CS corrw covb sorted;
run;
proc genmod data=sdata; by agent;
class subid spawncost;
model pgrowth= spawncost spawncost*gen / dist=normal link=log;
repeated subject=subid / type=UN corrw covb sorted;
run;
proc genmod data=sdata; by agent;
class subid spawncost;
model pgrowth= spawncost spawncost*gen / dist=normal link=log;
repeated subject=subid / type=ar(1) corrw covb sorted;
run;
data pop6; set tdata;
if agent='NC' then gen1=gen1; gen2=gen2; gen3=gen3; gen4=gen4;
gen5=gen5; gen6=gen6;
proc print data=pop6(obs=101);run;

title "Question 6.3: PATTERN OF CHANGE OF LENGTH BY SPAWNCOST AND ROUNDANTE";
proc genmod data=sdata; by agent;
class subid roundante spawncost;
model pgrowth= spawncost roundante spawncost*roundante*gen /
dist=normal link=log
obstats pscale;
repeated subject=subid / type=IND corrw covb sorted ;
run;
proc genmod data=sdata; by agent;
class subid roundante spawncost;
model pgrowth= spawncost roundante spawncost*roundante*gen /
dist=normal link=log
obstats pscale;
repeated subject=subid / type=CS corrw covb sorted ;
run;
proc genmod data=sdata; by agent;
class subid roundante spawncost;
model pgrowth= spawncost roundante spawncost*roundante*gen /
dist=normal link=log
obstats pscale;
repeated subject=subid / type=UN corrw covb sorted;
run;

proc genmod data=sdata; by agent;
class subid roundante spawncost;
model pgrowth= spawncost roundante spawncost*roundante*gen /
dist=normal link=log
obstats pscale;
repeated subject=subid / type=ar(1) corrw covb sorted;
run;
title "SPAWNCOST AND ROUNDANTE";
proc glm data=tdata1;
  /* same as 'proc anova' except 'glm' allows residual plots but gives more junk output */
class roundante spawncost;
model gen1 = roundante spawncost/ solution;
output out=tdata2 p=yhat r=resid;
  /* store fitted values and fitted residuals in dataset called 'oilfit' for later use */
run;

PROC UNIVARIATE plot NORMAL;
VAR diff;
RUN;
proc print data=tdata1(obs=20); run;

/*plot 1*/
data logsdata; set sdata;
pgrowth=log(pgrowth);
proc transpose data=logsdata out=logtdata prefix=gen;
by subid spawncost agent roundante roundspawn notsorted;
var pgrowth;
run;

title 'PROFILE PLOT FOR DIFFERENT AGENT';
proc glm data=logtdata;
class agent;
model gen1 gen2 gen3 gen4 gen5 gen6 = agent;
repeated gen 6 ;
lsmeans agent / out=means;
/*goptions reset=all;*/

filename output 'e:\guox\thesis\output\logagent.eps';
goptions device=pslepsfc gsname=output gsfmode=replace;
/**symbol1 interpol=spline width=4 color=blue line=10;*/
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
axis1 order=(2.5 to 3 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize= 10 vsize= 8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME=agent / vaxis=axis1 haxis=axis2;
run;
quit;

/*plot 2*/
data logtdata1; set logtdata;
if agent = 'SQ' then delete;
data logtdata2; set logtdata;
if agent = 'NC' then delete;

proc print data=logtdata1 (obs=101); run;
title 'PROFILE PLOT FOR SPAWNCOST IN NC GROUP';
proc glm data=logtdata1;
class spawncost;
model gen1 gen2 gen3 gen4 gen5 gen6 = spawncost;
repeated gen 6;
lsmeans spawncost / out=means;
/*goptions reset=all;*/

filename output 'e:\guox\thesis\output\spawncost_NC.eps';
goptions device=pslepsfc gsname=output gsfmode=replace;
/**symbol1 interpol=spline width=4 color=blue line=10;*/
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
axis1 order=(2.5 to 2.9 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize= 10 vsize= 8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME=spawncost / vaxis=axis1 haxis=axis2;
run;
quit;
/*plot 3*/
proc print data=logtdata2(obs=101); run;
title 'PROFILE PLOT FOR SPAWNCOST IN SQ GROUP';
proc glm data=logtdata2;
class spawncost;
model gen1 gen2 gen3 gen4 gen5 gen6 = spawncost;
repeated gen 6 ;
lsmeans spawncost / out=means;
/*goptions reset=all;*/ 

filename output 'e:\guox\thesis\output\spawncost_SQ.eps';
goptions device=pslepsfc gsfname=output gsfmode=replace;
/*symbol1 interpol=spline width=4 color=blue line=10;*/
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
axis1 order=(2.5 to 3.1 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize= 10 vsize= 8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME_=spawncost / vaxis=axis1 haxis=axis2;
run;
quit;

/*plot 4*/
proc print data=logtdata2(obs=101); run;
title 'PROFILE PLOT FOR ROUNDANTE IN NC GROUP';
proc glm data=logtdata1;
class roundante;
model gen1 gen2 gen3 gen4 gen5 gen6 = roundante;
repeated gen 6 ;
lsmeans roundante / out=means;
/*goptions reset=all;*/ 

filename output 'e:\guox\thesis\output\roundante_NC.eps';
goptions device=pslepsfc gsfname=output gsfmode=replace;
/*symbol1 interpol=spline width=4 color=blue line=10;*/
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
axis1 order=(2.5 to 3.2 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize= 10 vsize= 8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME=roundante / vaxis=axis1 haxis=axis2;
run;
quit;

/*plot 5*/
proc print data=logtdata2(obs=101); run;
title 'PROFILE PLOT FOR ROUNDANTE IN SQ GROUP';
proc glm data=logtdata2;
class roundante;
model gen1 gen2 gen3 gen4 gen5 gen6 = roundante;
repeated gen 6 ;
lsmeans roundante / out=means;
/* goptions reset=all; */

filename output 'e:\guox\thesis\output\roundante_SQ.eps';
goptions device=pslepsfc gsfname=output gsfmode=replace;
/*symbol1 interpol=spline width=4 color=blue line=10; */
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
axis1 order=(2.5 to 3.2 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize= 10 vsize= 8; */
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME=roundante / vaxis=axis1 haxis=axis2;
run;
quit;

/*plot 6*/
title 'PROFILE PLOT FOR ROUNDANTE AND SPAWNCOST IN NC GROUP';

proc glm data=logtdata1;
class roundspawn;
model gen1 gen2 gen3 gen4 gen5 gen6 = roundspawn;
repeated gen 6 ;
lsmeans roundspawn / out=means;
/* goptions reset=all; */
filename output 'e:\guox\thesis\output\roundspawn_NC.eps';
goptions device=pslepsfc gsfname=output gsfmode=replace;
/*symbol1 interpol=spline width=4 color=blue line=10; */
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
symbol3 c=green v=square h=1 i=j width=3;
symbol4 c=purple v=square h=1 i=j width=3;
axis1 order=(2.5 to 3 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize=10 vsize=8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME= roundspawn / vaxis=axis1 haxis=axis2;
run;
quit;

/*plot 7 */
title 'PROFILE PLOT FOR ROUNDANTE AND SPAWNCOAST IN SQ GROUP';

proc glm data=logtdata2;
class roundspawn;
model gen1 gen2 gen3 gen4 gen5 gen6 = roundspawn;
repeated gen 6;
lsmeans roundspawn / out=means;
/*goptions reset=all;*/
/*filename output 'c:\guox\thesis\output\roundspawn_SQ.eps';
goptions device=pslepsfc gsfname=output gsfmode=replace ;
/*symbol1 interpol=spline width=4 color=blue line=10;*/
symbol1 c=blue v=square h=1 i=j width=3;
symbol2 c=red v=square h=1 i=j width=3;
symbol3 c=green v=square h=1 i=j width=3;
symbol4 c=purple v=square h=1 i=j width=3;
axis1 order=(2.5 to 3.2 by 0.1) label=(a=90 'Means') width=2;
axis2 label=('Gen') value=('1' '2' '3' '4' '5' '6') width=2;
/* hsize=10 vsize=8;*/
/* graph procedure goes here */
proc gplot data=means;
plot lsmean*NAME= roundspawn / vaxis=axis1 haxis=axis2;
run;
quit;
proc print data=means(obs=36); run;
Bibliography


