Analysis of a Repeated Measures Bacterial Attachment Study

Matthew D. Austin
Department of Mathematical Sciences
Montana State University

April 30, 1999

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Matthew D. Austin

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

4/30/99

Date

John J Borkowski

Dr. John J. Borkowski
Writing Project Director

Abstract

An analysis will be proposed for a repeated measures study involving comparisons of bacterial attachment rates over time on several different surfaces which are similar to materials used in catheters and shunts. The analysis involves a mixed model using random nested effects and a spatial covariance structure. Model selection techniques will be discussed including model selection criterion, covariance structure selection, and term selection. The *SAS* code from the MIXED procedure will also be presented and discussed in the context of this analysis. The results of the analysis indicate statistically significant differences of bacterial attachment rates on the different surfaces at several time periods in the study. A discussion of the results–including a discussion of of statistical versus practical significance–will be presented along with recommendations and comments on future studies of this type.

Study Design

Introduction

Knowledge of the bacterial attachment rate to surfaces which are inserted into the body is important because of the possibility of blood stream infection. When a foreign surface is inserted into the blood stream, bacteria from both the surface and within the body will begin to attach to the surface. When bacteria adhere faster, the chance of a blood stream infection increases. The data and design presented in this paper is from a thesis study performed at the Montana State University Center for Biofilm Engineering by Jennifer Thompson. This study focuses on the attachment rate of *Pseudomonas aeruginosa* to polystyrene. The surfaces of the polystyrene were the primary interest of the study. Four different types of surfaces of polystyrene were prepared using an oxygen texturing process and a conditioning process. The oxygen texturing introduced oxygen molecules onto the polystyrene which created a different surface. The polystyrene which underwent the oxygenation will be referred to as "textured" surfaces. Calling the oxygenated surface textured is a sort of a misnomer, since the actual topological surface is smoother than the non-oxygenated or "non-textured" surface. The conditioned surfaces were polystyrene which had a coating of *bovine serum albumin*. This process creates a surface with a growth medium for the bacteria which is meant to emulate the environment of the body.

With these two techniques, four combinations of surface types could be created.

1. Non-Textured, Non-conditioned
2. Textured, Non-conditioned
3. Non-Textured, Conditioned
4. Textured, Conditioned

The four combinations will be referred to as treatments one through four for the remainder of this paper.

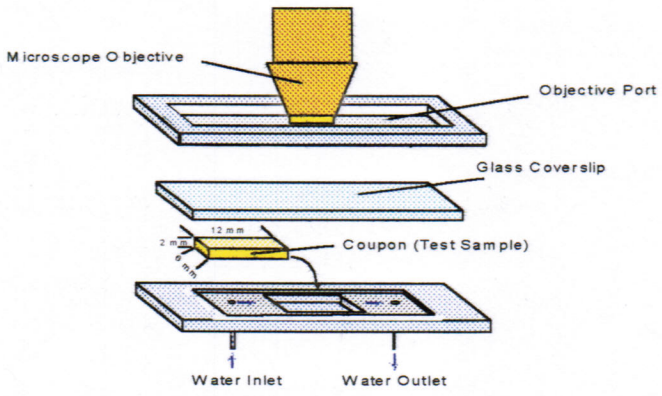

Figure 1: Flow Cell

Methods

The study was conducted on ten days spanning approximately one and a half years. There were three replications of treatments one and two, and two replications for both treatments three and four. The duration of a single replication was 360 minutes. Counts were taken at five minute intervals until 160 minutes, then 10 minute intervals until 240 minutes, and 15 minute intervals until completion. The order of the replications was not randomized. All replications of treatment one were run first, then all replications of treatment two, three and four.

The experiment took place in a flow cell, similar to Figure 1 [3], where a constant flow of bacteria crosses over the polystyrene plate. The cell is a closed system, where the counts are done from pictures taken through a microscope which focuses on the plate through the glass coverslip. The microscope is on a mobile cart which was shared by several experimenters. When the cart was not returned to the flow cell in time for an observation to be recorded, missing data appeared in the data set.

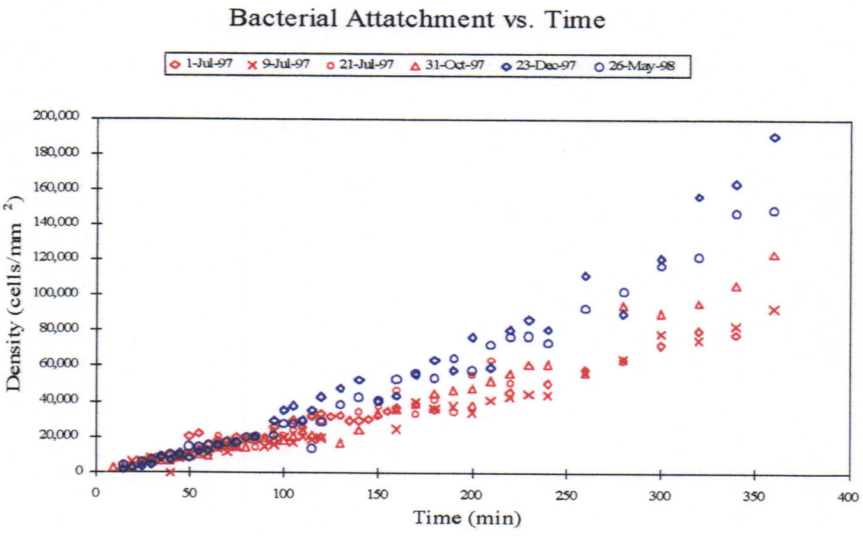

Figure 2: Raw Data [6]

Statistical Analysis

The model in this analysis was fit using *Statistical Analysis Software*. *SAS* is the most widely accepted software for statistical analysis in the United States by both government and industry. The MIXED procedure in *SAS* was used extensively. MIXED allows the researcher to fit both fixed, random, and mixed models with various error structures. The graphics procedure GPLOT was also used for model diagnostics and general plotting of the data.

Because of the design and basic nature of this kind of experiment, some special consideration had to be given to certain specific problems. These problems and how they were addressed in this particular analysis will be discussed in the following sections.

Complications and Considerations

The data from this study exhibited the following:

1. Non-Constant Variance
2. Unequally Spaced Time Series
3. Non-Independent Measurements
4. Missing Data
5. Day of the Count Represents a Random Effect
6. Day effects are Nested within Treatments
7. No Randomization of Treatment Replications

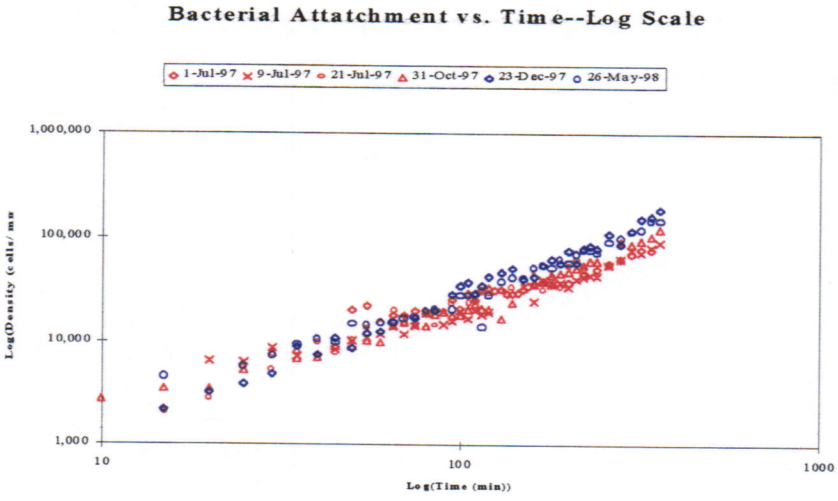

Figure 3: Log Scale

Initial Examination of Data

Figure 2 is a plot of the data from six days of the study. Note how the data is tightly grouped at the beginning times and then starts to separate in later time periods. By fitting a naive model, that is, a model with all factors of interest and simple covariance structure [1], a spread-location plot was constructed. The plot indicated problems with the homogeneity of variance assumption. Several transformations were considered, and a log transformation on both the response and time was selected. Figure 3 is a plot of the data on the dual log scale. Note how there appears to be less spread over time. A second spread location plot indicated that the homogeneity of variance assumption has been met with this transformation.

Selection Criteria

For the bacterial attachment model, both the covariance structure and the terms for the model must be selected. Although the covariance structure is not of direct interest, choosing the appropriate structure will make the inferences from the data more reliable.

The *Corrected Akaike's Information Criterion* , $CAIC$ was selected over all competitors for a selection criterion. The two other major competitors were Akaike's Information Criterion, AIC, and the Bayesian Information Criterion,BIC or SBC. The $CAIC$, AIC and the BIC are all based on the $log - likelihood$ function. The log-likelihood function is a measure of goodness-of-fit of the model.

[1]The simple covariance structure has the form $\sigma^2 I$ where I is the identity matrix of the appropriate dimensions.

$$\sigma^2 \begin{pmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} & \cdots & \rho^{d_{1n_k}} \\ & 1 & \rho^{d_{23}} & \rho^{d_{24}} & \cdots & \rho^{d_{2n_k}} \\ & & 1 & \rho^{d_{34}} & \cdots & \rho^{d_{3n_k}} \\ & & & 1 & & \\ & & & & \ddots & \vdots \\ & & & & & 1 \end{pmatrix}$$

Figure 4: Spatial Power Covariance Structure [1]

[7] However, the log-likelihood will improve with each predictor variable or term added to the model. The AIC, $CAIC$, and BIC add penalization factors to the log-likelihood based on the number of predictors in the model which helps the researcher choose a more parsimonious model.

The AIC is a widely recognized model building tool in many fields, but it will tend to overestimate the size of the model. The BIC and $CAIC$ have a smaller probability of overparameterization than the AIC [8]. However, the BIC is used under the assumption that one of the possible models being fit is the "true" model [5]. In contrast, the AIC and $CAIC$ work under the assumption that the "true" model is not one of the models being fit, and the best alternative is being found. Being more comfortable with the second assumption, the $CAIC$ was chosen as the model building criterion. In some parameterizations of the $CAIC$ the minimum value is optimal. However, the algorithm which SAS uses in the information criterion section of the mixed procedure produces both a maximum and a minimum parameterization. For the covariance and term selection discussion, the maximization of the $CAIC$ will be desired. Along with the criterion, residual diagnostics will also be used to determine the "best" model.

Error Structure

Because of the time series structure of the bacteria counts, independence between observations could not be assumed. Also, the unequally spaced time intervals posed a problem for modeling the error structure. In addition, the error structure had to account for the fact that there were multiple measurements on the same treatment. This type data structure is called a repeated measures design.

When examining biological growth data, it is common to collect more data at the beginning time periods when the growth is more variable and less data in the latter time periods when the growth is expected to have less noise. [2]A special error structure was built into the model to take into account the dependence of one observation on the previous observations and the unequal spacing. The error structure used is called a *spatial power law structure.* The spatial power structure allows the researcher to fit a model where the correlation between observation decays depending on the distance between two observations in time (Figure 4).

By using this structure, the correlation between observations could be estimated. Note that this is a generalization of the one dimensional AR(1)

structure[4]. The structure was fit specifically to match the repeated properties of the data.

The process for selecting the spatial power covariance structure began with building a "full" model. The full model included all terms of possible interest. Several covariance structures, including spherical, compound symmetric, and gaussian, were then analyzed in the full model. The spatial power structure yielded the highest $CAIC$ value and also had the best residual diagnostics.

Model Selection and Diagnostics

The model fitting was also based on maximizing the $CAIC$. The models with the highest $CAIC$ values were then analyzed by checking the residual diagnostics to assure the underlying distributional assumptions had been met (residual versus predicted values and spread location plots). The largest model was fit with linear, quadratic, and cubic centered log-time terms, treatment effects, and interaction effects of the treatments with the different time terms. Other models were fit by removing one term at a time until all time terms were removed. The residual diagnostics were then checked for the models with the best $CAIC$'s. The model with the highest $CAIC$, 412.6, and best residual diagnostics was selected and is explained in the next section. Note that one of the most important effects in this study is the interaction of treatment and time. By specifying this effect in the model statement, MIXED will allow different slopes to be fit for the different treatments. This is important because the primary goal of the study was to find differences in the attachment rate due to treatments. Even if the interaction is not significant, noise in the data may be masking the interaction. One should still check for differences at the important times during the study using the LSMEANS option in the *MIXED* procedure.

Now that the final model has been selected, the last of the complications and considerations can be discussed. Because the treatments were run in a non-randomized order, the inferences from the tests at the given time periods may not be valid. However, by looking at the residuals versus factor plots (figures 4 and 5), an observer can see that the variability did not change over the treatments or over the days on which the experiment was conducted. This indicates that the non-randomization may not discount the inferences on the treatment effects.

Model

$$y_{ijk} = \tau_i + \gamma_{j(i)} + \beta_i t_{ijk} + \varepsilon_{ijk} + \delta_{k(ij)}$$

where

- i = treatment level
- j = day which experiment was run
- k = minutes elapsed during the experiment

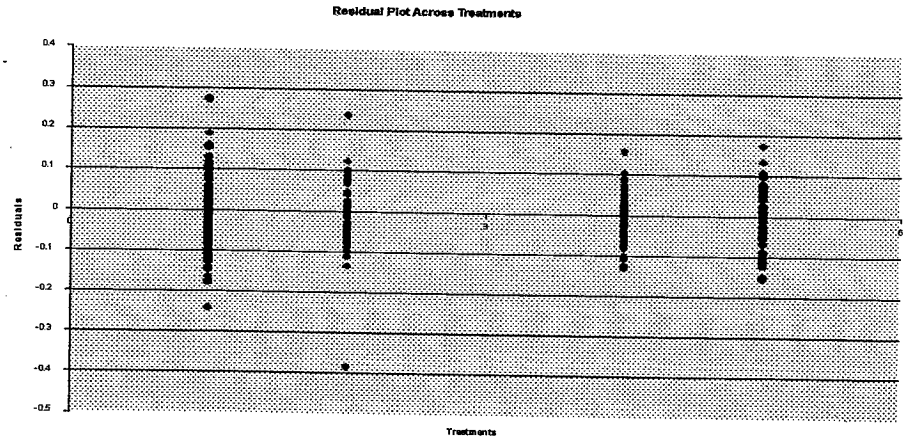

Figure 5: Residual vs Treatment Level

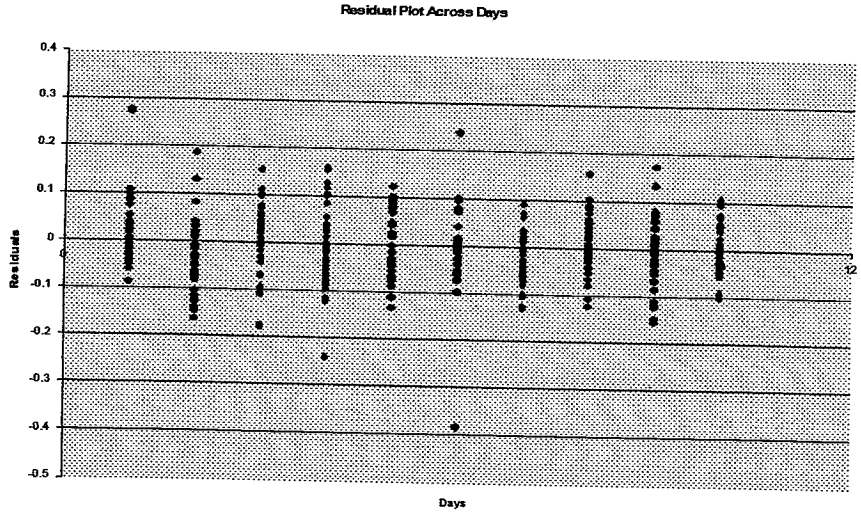

Figure 6: Residual vs Run Order

- y_{ijk} is the log count value at the i^{th} treatment on the j^{th} day taken at time k
- τ_i is the effect of the i^{th} treatment
- $\beta_i t_{ijk}$ is the fixed interaction effect of centered log(time) and the treatment which was run. This allows for different slopes for different levels of the treatment.
- t_{ijk} is the time k on the j^{th} day of the i^{th}
- $\gamma_{j(i)}$ is the effect of the j^{th} day within the i^{th} treatment, considered random because this specific day was not of interest. Consider the day randomly selected from all possible days on which the process could have been run. $\gamma \sim N(0, \sigma^2_\gamma)$
- $\varepsilon_{ijk}$is the random error term for the model, which is assumed to be $N(0, \sigma^2_\varepsilon)$; however, in this model a serial dependence of the error at time t on the error at time $t-1$ was found. Therefore, the model was fit with a spatial power covariance structure.
- $\delta_{k(ij)}$ is the model term that accounts for special error structure.

Specific Code for Example

```
PROC MIXED data=growth method=reml ic;
    CLASS trt day ptime;
    MODEL response = cstime cstime*trt / noint p;
    LSMEANS trt / (cstime) = (-.4986); * 60 minutes;
    LSMEANS trt / (cstime) = (-.0688); * 90 minutes;
    LSMEANS trt / (cstime) = (.23616); * 120 minutes;
    LSMEANS trt / (cstime) = (.66596); * 180 minutes;
    LSMEANS trt / (cstime) = (.97091); * 240 minutes;
    MAKE 'predicted' OUT = pred1;
    REPEATED ptime / type=sp(pow)(time) subject=day(trt) r rcorr;
```

Tricks, Tips, and Suggestions for the Code

Note that time appears in three forms: *time* , *cstime* , and *ptime*. *time* is the original time variable which represents the distance in time that the data was collected. By placing this variable in the specified covariance structure, *SAS* is able to model the spatial power structure using the correct distances in time. *cstime* represents the centered and scaled time used for modeling and model building. *ptime* is the same as *cstime* , but it appears in the class statement so that *SAS* recognizes it in the REPEATED statement. By placing *cstime* in the REPEATED statement, PROC MIXED can align the observations according to the *cstime* regardless of the missing values. This is extremely important and what separates the MIXED procedure from

the GLM procedure. When analyzing repeated measures in GLM with missing data, GLM will drop all runs which have missing data and only analyze those with all the data present. Another disadvantage of using GLM is that it cannot properly handle either the unequally spaced data or the more complex covariance structures.

The subject option in the REPEATED statement specifies the variable on which the covariance matrix is structured. By specifying the subject in this example the covariance matrix is constructed to be a block diagonal matrix with the diagonal elements fitting the spatial power structure to the individual runs of the experiment. The *r* and *rcorr* options instruct MIXED to print out the first block of the covariance matrix and the correlation matrix which represents the first run of the experiment.

The MAKE statement creates *SAS* data sets for output of residuals, predicted values, etc., for use in diagnostics and model fitting. The LSMEANS statement is used to compare the growth curves at different times. Notice that the times are centered and scaled to match *cstime* . A suggestion for finding the centered and scaled times for the LSMEANS statement is to inspect the libraries using the library icon on the tool bar in *SAS* and look in the data set where the centered and scaled values are stored by the STANDARDIZE procedure. If you outputted the data from the STANDARDIZE procedure into a data set with the same name as the original data set created in the DATA step, the original time and the corresponding centered and scaled times are easy to find. The LSMEANS statement was chosen over the CONTRAST statement because of the missing values in the data set.

Results

The primary goal of the study was to determine if the treatments had a significant effect on bacterial growth.

For the primary goal, hypothesis tests were constructed to test for both a significant effect due to the treatments and to find specific differences between bacteria growth due to the different treatments. These tests for differences were conducted at 90, 120, 180, and 240 minutes. The test which tests the overall significance of the treatment effects indicates that treatment effects do exist (Reject $H_o : \tau_i = 0$, P-value<.0001). The hypothesis test was also conducted to verify that the treatments had different interactions with time which indicates that the attachment rate differs for the different treatments(Reject $H_o : \beta_i t_{ijk} = 0$, P-value<.0001).

A family of thirty individual tests for treatment differences were run at the 5%level. Individual tests were run at the 0.17% using the Bonferroni Inequality. These tests indicated that the non-textured non-conditioned runs differed significantly from the textured conditioned runs at all levels of time. Also, the non-textured conditioned runs differed significantly from the textured conditioned runs at all times. These and other differences are listed on the attached table highlighted in red. A point of interest is that the textured non-conditioned runs

Results

Compared Groups	P_{value}	Compared Groups	P_{value}
At 90 minutes		**At 120 minutes**	
1 to 2	0.0425	1 to 2	0.0204
1 to 3	0.6317	1 to 3	0.4711
1 to 4	0.0001	1 to 4	0.0001
2 to 3	0.1811	2 to 3	0.1782
2 to 4	0.0185	2 to 4	0.0138
3 to 4	0.0003	3 to 4	0.0002
At 180 minutes		**At 240 minutes**	
1 to 2	0.0083	1 to 2	0.0049
1 to 3	0.3118	1 to 3	0.2365
1 to 4	0.0001	1 to 4	0.0001
2 to 3	0.1811	2 to 3	0.1866
2 to 4	0.0104	2 to 4	0.0093
3 to 4	0.0001	3 to 4	0.0001

Red is a significant difference, while blue shows a suggestive difference.

Figure 7: Results [6]

were never detected to be significantly different from the non-textured conditioned runs. Notice where differences were not detected does not mean that differences do not exist. This could be due to an insufficient amount of data or to extreme noise in the data.

Suggestions for Future Study

There are two suggestions for future study of similar repeated measures data. The first suggestion is a concerns the data collection. When the experiment was conducted, the microscope was not maintained in the same position on the plate. Because the microscope was shared during the experiment, each measurement was taken from a different area on the plate. There is a spatial aspect to the data which has been ignored in this analysis and may be of importance.

The second suggestion deals with the analysis style. Instead of using a multiple comparison style test, the analysis could have been performed using break points and comparing the slopes of the different treatments. If significant differences were found in the slopes, conclusions could have been made at all time periods, instead of at certain time periods. This type of procedure would require the researcher to have prior knowledge of the behavior of the bacterial attachment to set the break points prior to the data collection.

References

[1] *SAS/STAT Software: Changes and Enhancements through Release 6.11.* SAS Institute, Inc., 1996.

[2] Peter J. Diggle. An approach to the analysis of repeated measurements. *Biometrics*, 44:959–971, 1988.

[3] Center For Biofilm Engineering. *PowerPoint Flowcell Diagram.* Montana State University, 1999.

[4] Ramon C. Littel, George A. Milliken, Walter W. Stoup, and Russell D. Wollfinger. *SAS System for Mixed Models.* SAS Institute, Inc., 1996.

[5] Donna K. Pauler. The schwarz criterion and related methods for model selection in linear regression. Technical report, 1995.

[6] Jenny Thompson. Master's thesis, Center for Biofilm Engineering–Montana State University, 1999.

[7] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S-Plus.* Springer, 1997.

[8] Russ Wolfinger. Covariance structure selection in general mixed models. *Communications in Statistics–Simulations*, 22(4):1079–1106, 1993.