Sample size calculation for longitudinal data of abdominal aortic aneurysm

A Bachelor of Science Thesis in Statistics
by Eric Markarian

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

AstraZeneca

Department of Economics and Statistics

Department of Biostatistics, R&D

Supervisors:

Senior university lecturer Mattias Sundén at University of Gothenburg
Statistical science director Ziad Taib at AstraZeneca

Handelshögskolan, University of Gothenburg, Gothenburg
AstraZeneca, Mölndal
Sweden
April 2013

Abstract

Sample size calculation is a crucial step in all experimental design. In clinical research and drug development activities, it is required in order to be able to demonstrate a presumed statistical effect of a drug or a treatment. Today many sample size calculation algorithms and formulas exist. However, in this work an algorithm based on the results of Liu and Liang (1997) is tested and used to predict the right sample size based on data from a study involving 211 patients with abdominal aortic aneurysm (also known as AAA). In this study the growth of the diameter of the aneurysm was monitored over time and the slope of that growth was calculated. Since no information about treatment effect was provided, a statistically significant reduction of the slope by 20% was chosen to replace the lack. More precisely, we want to calculate the sample size required to demonstrate a desired effect of growth reduction by 20% of a treatment at the statistical power of 80%.

The aim of this work was not only to examine statistically the abdominal aortic aneurysm data from placebo patients and the involved variables but also to evaluate the “longpower” package existing in the programming language R to calculate the sample size for longitudinal data.

The statistical model chosen for this work was a linear mixed model with TIME as a random and fixed variable and logarithm of aneurysm diameter at baseline (AD0) as a fixed variable. Non-equidistant TIME measured the intervals of ultrasound screenings in years whereas AD0 was measured in mm.

The formula of Liu and Liang (1997) using “longpower” package in R computed a required sample size of 420 patients with a power of 80% and reduction of TIME slope by 20%. In order to verify the sample size of 420 a simulation for the control and the treatment groups were run. A two-sample t-test showed statistically significant difference in means of logarithms of aneurysm diameters for simulated control and treatment groups at the significance level of less than 0.1%.

Moreover, a linear mixed model using simulated data for 210 placebo and 210 treatment patients to investigate a cross effect of TIME*TREATMENT as fixed and random variable gave a statistically significant difference between the control and the treatment groups at the significance level of less than 0.1%.

To test the number 2*210 patients, another simulation of 2*105 patients were run. Two-sample t-tests showed statistically significant difference in means of logarithms of aneurysm diameters for these simulated control and treatment groups at the significance level of less than 0.1%. Investigation of the cross effect of TIME*TREATMENT in a linear mixed model showed statistically significance at the significance level of less than 0.1% for the simulation of 2*105 patients.

Although both sample sizes of 2*210 and 2*105 were acceptable from statistical standpoint, power calculations revealed that the sample size of 2*210 gave a power of 73% whereas 2*105 gave only a power of 61%. Finally, the sample size of 420 (2*210) was verified by the simulations.

Contents

Background

Sample size calculation

Prior to designing experiments researchers must know what sample size they should choose to be able to demonstrate a desired effect of a medication or a treatment. Over the course of many years, many different methods and formulas have been developed for this purpose. One of these formulas is Liu and Liang's formula (1997) [3] which is implemented in the "longpower" package of the programming language R. The formula is suited for studies for longitudinal data at equidistant points in time with correlated observations. This function will be applied to abdominal aortic aneurysm data containing non-equidistant measurements for patients who have been ultrasound screened several times over a period of time.

Abdominal aortic aneurysm

Abdominal aortic aneurysm (also known as AAA) is a localized dilatation of the abdominal aorta exceeding the normal diameter by more than 50 percent (normal diameter of the aorta is approximately 20 mm). Mostly, AAA causes no symptoms while it can sometimes cause pain in the abdomen and back. The most dangerous complication of abdominal aortic aneurysms is the rupture of the aneurysm which spills a large amount of blood into the abdominal cavity and can lead to death within a few minutes (See Figure 1).

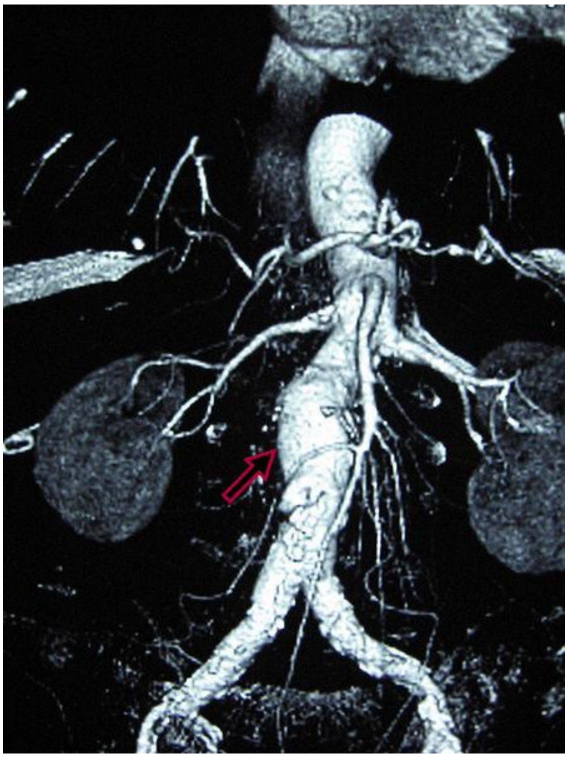


Figure 1: A CT image of an AAA (34 mm in diameter). The red arrow indicates the position of the aneurysm [8].

So far no medication to decrease the growth rate or rupture rate of asymptomatic AAA has been found. However, studies have suggested that therapy with angiotensin converting enzyme inhibitors, beta-blockers, and statins can protect against AAA. Ultimately surgery is needed for a severe (>55 mm in diameter) AAA.

Linear mixed model

A linear mixed model is a linear statistical model containing both fixed and random effects, i.e. mixed effects. Such models are frequently used for statistical analysis and are especially useful when repeated measurements are made on the same person or statistical unit.

A mixed model in matrix notation can be generally represented as:

$$Y = X\beta + ZU + \varepsilon$$

Y is a vector of observations, with mean $E(Y) = X\beta$
β is a vector of fixed effects
U is a vector of random effects with mean $E(U) = 0$ and variance-covariance matrix $Var(U) = G$
ε is a vector of random error terms with mean $E(\varepsilon) = 0$ and variance $Var(\varepsilon) = R$
X and Z are matrices of regressors relating Y to β and U

Nevertheless, the linear mixed model in this work will be represented as:

$$Y_{ij} = \alpha_{i,fixed} + \alpha_{i,random} + \beta_{i,fixed}X_{ij} + \beta_{i,random}Z_{ij} + \varepsilon_{ij}$$

Y_{ij} is a matrix of dependent variable
X_{ij} are matrices of fixed independent variables
Z_{ij} are matrices of random independent variables
$\alpha_{i,fixed}$ is intercept with a fixed part (within the group)
$\alpha_{i,random}$ is intercept with a random part (within the individual) $\sim N(0, \sigma_{intercept}^2)$
$\beta_{i,fixed}$ is slope with a fixed part
$\beta_{i,random}$ is slope with a random part $\sim N(0, \sigma_{slope}^2)$
ε_{ij} are random error terms $\sim N(0, \sigma_{error}^2)$
i is individual
j is time

Longpower

The “longpower” package contains functions for computing sample size for linear models of longitudinal data based on the formula of Liu and Liang (1997) and Diggle et al. (2002) [5].

In this work the formula of Liu and Liang (1997) was chosen to calculate the sample size with a given effect for the linear mixed model.

This is a powerful package since not so many packages which can calculate sample size over a time period with different measurements exist in R. One drawback of the functions of “longpower” is their inability to handle non-equidistant points.

Results

Studied variables

To set up our statistical linear mixed model potential variables needed to be chosen. The pharmaceutical data provided the following variables:

Variable	Description
SUBJECT	Identification number for each patient
RANGE	Range of the aneurysm depending on size at baseline 25<=AD0<35 35<=AD0<45 45<=AD0<55 AD0=>55
AD	Aneurysm diameter
AD0	Aneurysm diameter at baseline
TIME	Time for each measurement
WOMAN	Gender
DIABETES	Occurrence of diabetes type 2
AGE	Age

Table 1: A summary of the variables assumed to be involved in abdominal aortic aneurysm and the response variable (AD).

The study started to assess which of these variables to be included in or excluded from the model. So a series of linear mixed model separately for each variable were run in SPSS for this purpose. RANGE, WOMAN, DIABETES and AGE gave high p-values (much greater than 0.05) and were excluded from the model. It was unexpected that WOMAN was not statistically significant at the significance level of 5% despite the fact that the vast majority of the patients were men. An attempt to categorize the patients according to their age was done with no success of getting an acceptable p-value.

Test the model

Only the variables AD0 (Aneurysm diameter at baseline) and TIME were statistically significant at the significance level of 5%. So a linear mixed model with TIME as fixed and random variable and AD0 as fixed was run. TIME was chosen as a random variable to let a variation in time between the individuals.

The covariance structure chosen here after referring to Littell et al. (2000, [4]) was the default one in SPSS i.e. Variance Components. The reason for the choice was that Variance Components does not need a repeated variable which was lacked in the provided data. In addition Variance Components assumes no special correlation structure between the measurements which was appropriate for our data since the measurements were not equidistant in time.

Moreover logarithmation of AD and AD0 gave much better Information Criteria which made us to use the logarithms of these variables in the rest of the study.

Variables	Schwarz's Bayesian Criterion (BIC)
AD and AD0	5106.096
LOG10_AD and LOG10_AD0	-5014.007

Table 2: The Information Criteria obtained when running the model with AD and AD0 respective LOG10_AD and LOG10_AD0. The lower the BIC value the better the model is.

To evaluate the model the residuals were plotted (Figure 2). As shown the model did not show heteroscedasticity and the residuals seemed normally distributed.

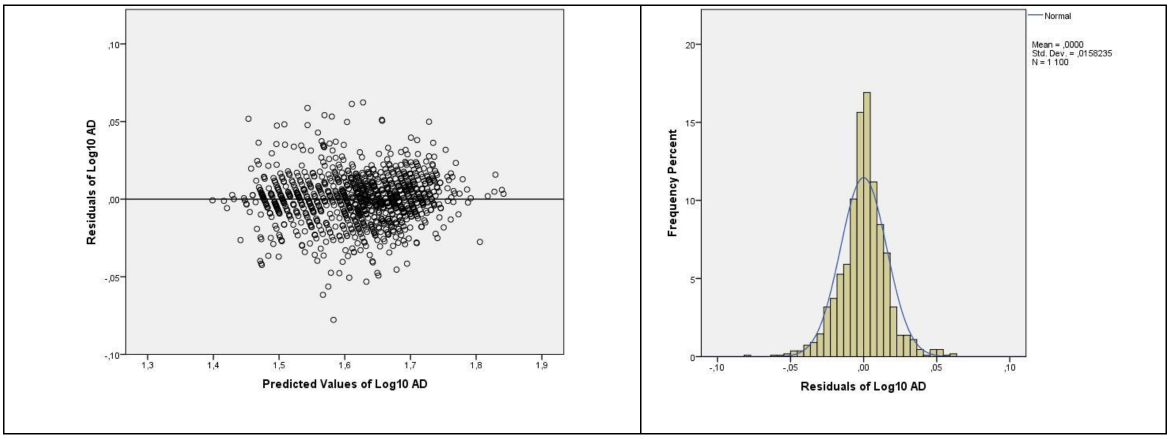


Figure 2: Plotted residuals for the model with dependent variable Log10 _AD and independent variables TIME and Log10_AD0.

Growth rate

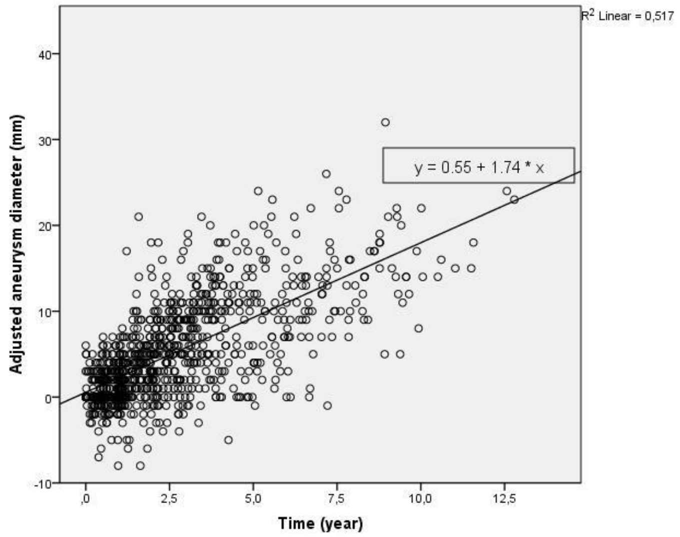


Figure 3: The growth rate of AAA for the patients in the data. For every year the aneurysm grew in average 1.74 mm.

The growth rate of the aneurysm was studied in this group of patients by plotting TIME against ADJUSTED_AD which was simply the difference of AD and AD0. According to Figure 3 for each passed year the aneurysm grew 1.74 mm in average for these patients.

The model

By choosing TIME as random and fixed variable and Log10_AD0 as fixed, the following results were obtained in SPSS.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	,023364	,022676	282,710	1,030	,304	-,021271	,067998
TIME	,023846	,001450	153,547	16,445	,000	,020981	,026710
LOG10_AD0	,983406	,014362	280,894	68,474	,000	,955135	1,011676

a. Dependent Variable: Log10 of aneurysm diameter.

Table 3: The fixed parameter estimates of the linear mixed model.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		,000323	1,626686E-005	19,838	,000	,000292	,000356
Intercept [subject = SUBJECT]	Variance	,000130	2,388897E-005	5,436	,000	9,055928E-005	,000186
TIME [subject = SUBJECT]	Variance	,000300	4,173999E-005	7,180	,000	,000228	,000394

a. Dependent Variable: Log10 of aneurysm diameter.

Table 4: The random variance estimates of the linear mixed model.

No interaction between TIME and Log10_AD0 as fixed variables was defined since TIME started at 0 when measuring AD0.

Our model was defined as:
$\text{Log10\_AD} = \alpha_{\text{fixed}} + \alpha_{\text{random}} + \beta 1_{\text{fixed}}\text{TIME} + \beta 1_{\text{random}}\text{TIME} + \beta 2_{\text{fixed}}\text{Log10\_ AD0} + \varepsilon$

i.e. according to Tables 3 and 4:

$$\text{Log10\_AD} = 0.023364 + \sqrt{0.000130}\ Z1 + 0{,}023846\ \text{TIME} + \sqrt{0.000300}\ Z2\ \text{TIME} \\ + 0{,}983406\ \text{Log10\_ AD0} + \sqrt{0.000323}\ Z3$$

where Z1, Z2 and Z3 were independent random variables distributed N (0,1).

Run “longpower”

“longpower” is a package defined in the programming language R and from this point we stopped using SPSS and the statistical analysis was transferred to R.

Since the data contained only placebo patients and no treatment patients, a strategy of reduction of TIME slope by 20% was used instead. This strategy just implied that the reaching time to the rupture point of 55 mm for the patients was assumed to be delayed by 20%.

Correlation between random intercept and random slope was assumed to be 0 since data about treatment patients was not available. To test this correlation namely 0, different values were inserted in the variable of correlation without changing the sample size significantly. This fact indicated that the correlation between random intercept and random slope was not so important for this data and this model.

Moreover, to assess the number of measurements for the placebo group, the equidistant time t’ until the rupture of aorta in the placebo data was calculated according to:

Log10_AD = 0.023364 + 0,023846 TIME + 0,983406 Log10_ AD0

Log10(55) = 0.023364 + 0,023846 * t’ + 0,983406 * 1.5702 → t’ = 7.2 years

Here 55 mm is the assumed rupture point and 1.5702 is the mean of Log10_AD0 in the placebo data.

In addition, two plots were drawn for a control and a treatment patient to assess the number of measurements for the treatment group in “longpower”. As shown in Figure 4, the control patient passed the rupture line (AD = 55 mm or Log10_AD = 1.74) after approximately 8 years whereas the treatment patient passed the same line after approximately 10 years.

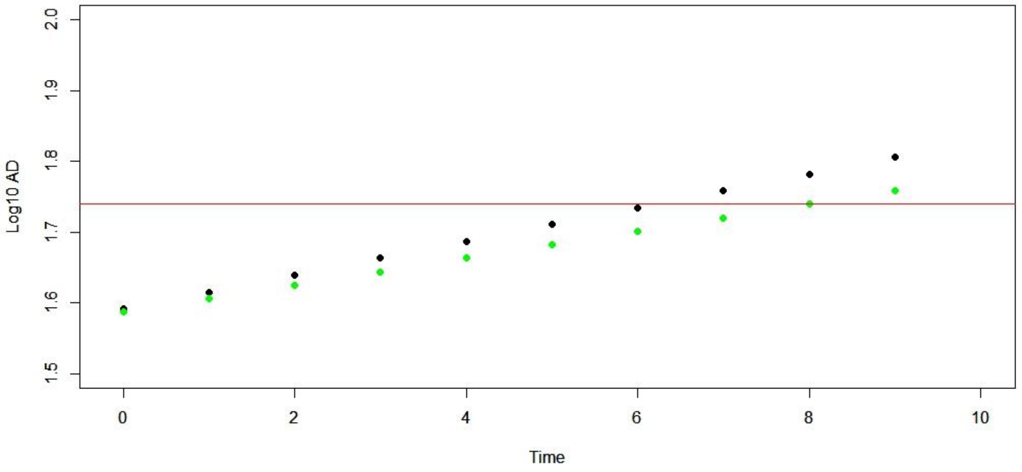


Figure 4: The black dots show the control patient to pass the rupture line (red) after 8 years whereas the green dots, with a slope reduction of 20%, show the treatment patient passing the same line after 10 years.

So, 10 equidistant measurements for the pretended treatment group were assumed. The difference of the slopes of these plots gave the most crucial argument delta in “longpower”.

“longpower” is a straight-forward algorithm [1,3] and only needs to be inserted with the input values. Input arguments in “longpower” are presented in Table 5.

Argument	Description
N	the total sample size
n	sample size per group
delta	group difference (possibly a vector of differences)
u	a list of covariate vectors or matrices associated with the parameter of interest
v	a list of covariate vectors or matrices associated with the nuisance parameter
sigma2	the error variance
R	the variance-covariance matrix for the repeated measures
sig.level	type one error
power	power
alternative	one- or two-sided test

Table 5: The input arguments required in the “longpower” algorithm.

After inserting the variance estimates’ values estimated by SPSS in the “longpower” algorithm [Appendix II], the algorithm calculated a sample size of 420. The R output was:

```
Longitudinal linear model power calculation (Liu & Liang, 1997)

              N = 419.3564
              n = 209.6782, 209.6782
          delta = 0.00477
         sigma2 = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: N is total sample size and n is sample size in each group.
```

An attempt to run the linear mixed model in R was made and variance estimates’ values were extracted. This time the “longpower” algorithm [Appendix III] calculated a sample size of 380. Although the sample size of 380 calculated completely by R is relatively close to the sample size of 420 calculated by SPSS and R but this fact shows that there exists a difference between SPSS and R using different algorithms to model a linear mixed model.

Significance test of the variables

The fixed and random coefficients of the variable TIME and coefficient of Log10_AD0 only as fixed turned out to be significant at the significance level of 5%. The rest of variables: RANGE, WOMAN, DIABETES and AGE were not significant at the significance level of 5%.

An attempt to run Log10_AD0 as a random effect in the linear mixed model was made hoping to obtain a better model. This attempt was not successful implying that there was no correlation in AD0 variable for the patients.

Simulations

In order to evaluate the number 420 calculated by “longpower”, simulations of the model by a sample size of 2*210, number of measurements in time of 10 and two different fixed time slopes representing control and treatment groups were done. The parameter and variance estimates were retrieved from our original data of 211 placebo patients. A two-sample t-test showed a difference in means between these two groups at a statistical significance level of less than 0.1% [Appendix IV].

In SPSS the residuals for 2*210 patients were randomly spread with no heteroskedasticity and normally distributed.

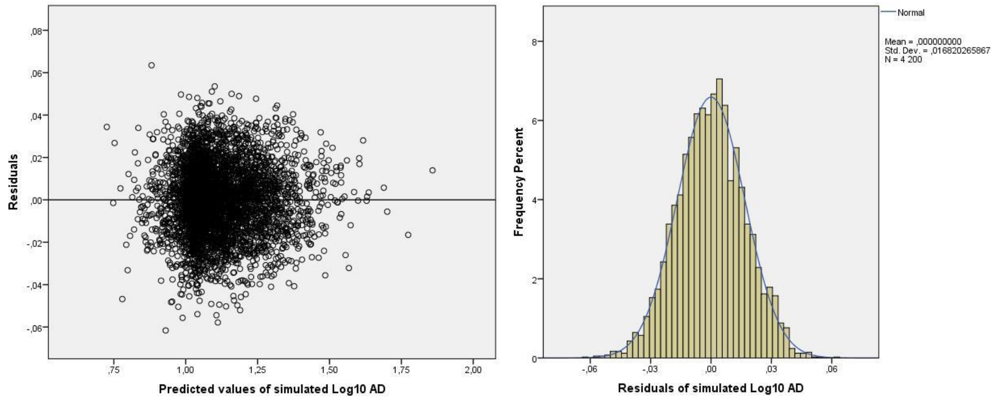
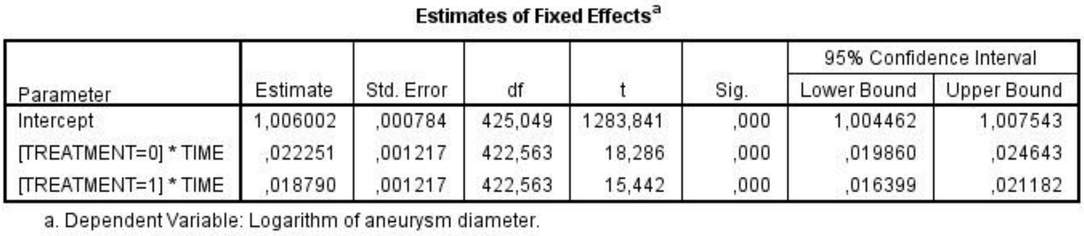


*Figure 5: The residuals for simulated linear mixed model for 2*210 patients.*

By choosing the interaction term TREATMENT*TIME as random and fixed cross variable, the results in Table 6 and 7 were obtained for 2*210 patients in SPSS. As seen in the tables, there existed difference between control and treatment groups at significance level less than 0.1%. This means that 2*210 patients as a sample size is an appropriate sample size to start a similar study with.

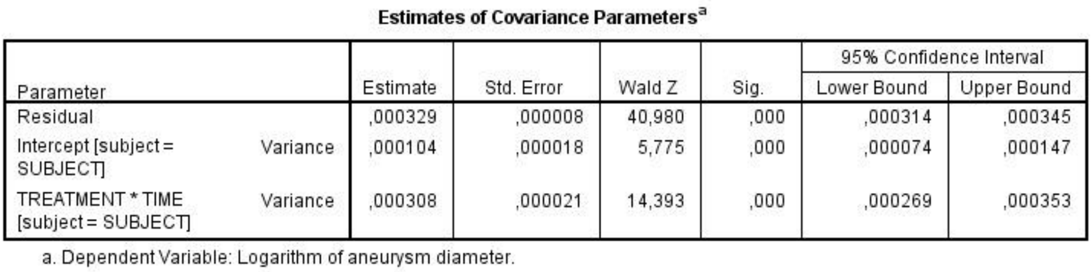


Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1,006002	,000784	425,049	1283,841	,000	1,004462	1,007543
[TREATMENT=0] * TIME	,022251	,001217	422,563	18,286	,000	,019860	,024643
[TREATMENT=1] * TIME	,018790	,001217	422,563	15,442	,000	,016399	,021182

a. Dependent Variable: Logarithm of aneurysm diameter.

*Table 6: The fixed cross effect estimates of the simulated linear mixed model for 2*210 patients were statistically significant. Treatment=0 is for control and Treatment=1 is for the treatment group.*



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		,000329	,000008	40,980	,000	,000314	,000345
Intercept [subject = SUBJECT]	Variance	,000104	,000018	5,775	,000	,000074	,000147
TREATMENT * TIME [subject = SUBJECT]	Variance	,000308	,000021	14,393	,000	,000269	,000353

a. Dependent Variable: Logarithm of aneurysm diameter.

*Table 7: The random variance estimates of the simulated linear mixed model for 2*210 patients were statistically significant.*

To challenge the generated sample size of 2*210 by "longpower" another simulated linear mixed model was run with the same conditions in SPSS with only 2*105 patients. The residuals for 2*105 patients were randomly spread with no heteroskedasticity and normally distributed.

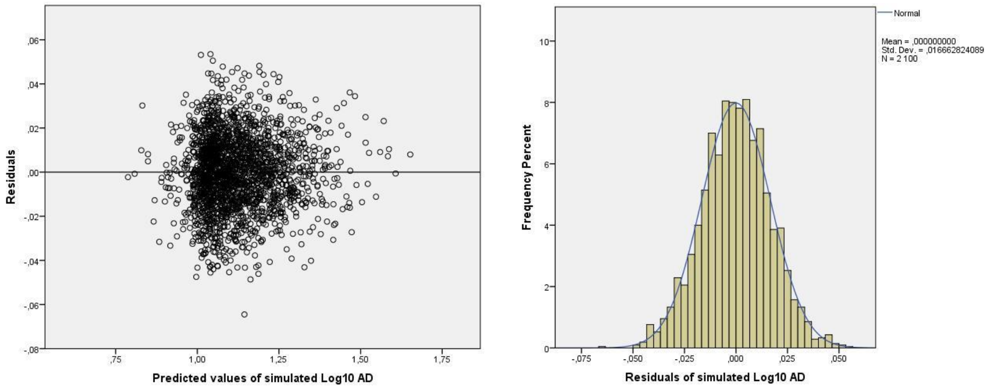
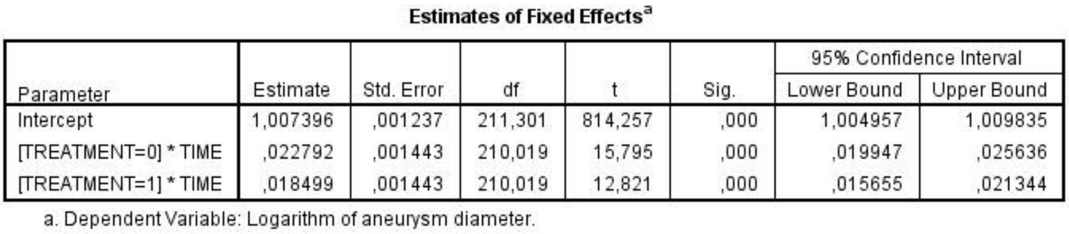


*Figure 6: The residuals for simulated linear mixed model for 2*105 patients.*

TREATMENT*TIME as fixed and random cross variable gave fixed and random effects for 2*105 patients shown in Table 8 and Table 9. As seen, there again existed difference between control and treatment groups at significance level less than 0.1%.

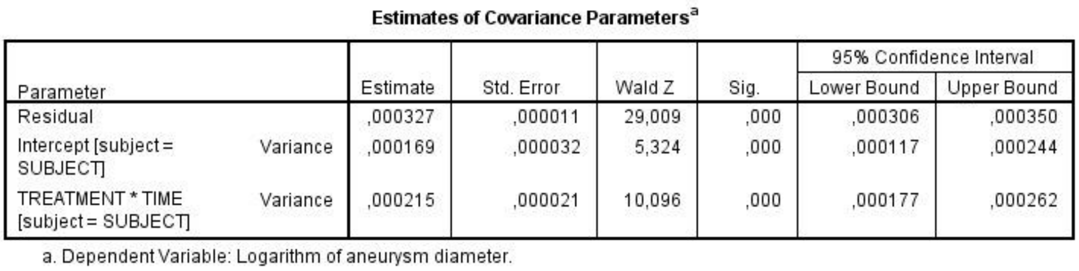


Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1,007396	,001237	211,301	814,257	,000	1,004957	1,009835
[TREATMENT=0] * TIME	,022792	,001443	210,019	15,795	,000	,019947	,025636
[TREATMENT=1] * TIME	,018499	,001443	210,019	12,821	,000	,015655	,021344

a. Dependent Variable: Logarithm of aneurysm diameter.

*Table 8: The fixed cross effect estimates of the simulated linear mixed model for 2*105 patients were statistically significant. Treatment=0 is for control and Treatment=1 is for the treatment group.*



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		,000327	,000011	29,009	,000	,000306	,000350
Intercept [subject= SUBJECT]	Variance	,000169	,000032	5,324	,000	,000117	,000244
TREATMENT * TIME [subject= SUBJECT]	Variance	,000215	,000021	10,096	,000	,000177	,000262

a. Dependent Variable: Logarithm of aneurysm diameter.

*Table 9: The random variance estimates of the simulated linear mixed model for 2*105 patients were statistically significant.*

Since there existed difference between control and treatment groups at significance level less than 0.1%, so even the sample size of 2*105 is acceptable to start a similar study with.

Power calculations

Power calculations both for the calculated sample size by "longpower", 2*210 and our alternative sample size, 2*105 on simulated data were done [Appendices V-VIII].

Power calculations	
Sample size: 2*210	Sample size: 2*105

N = 420.0002
n = 210.0001, 210.0001
delta = 0.00445
sigma2 = 1
sig.level = 0.05
power = 0.7329034
alternative = two.sided
	N = 209.9914
	n = 104.9957, 104.9957
	delta = 0.00456
	sigma2 = 1
	sig.level = 0.05
	power = 0.60751
	alternative = two.sided

*Table 10: Power calculations of sample sizes 2*210 and 2*105 are crucial to choose an appropriate sample size.*

These findings showed how a powerful tool "longpower" is to calculate sample size and power and its ability to predict an appropriate sample size for data before start of an experiment. A desired sample size is followed by a desired power for an experiment.

Discussion

Correlation in scientific data occurs frequently. Longitudinal data is a perfect example of correlated data. The statistical model, the linear mixed model is designed to study such data to give correct estimates for the fixed variables and variance estimates for the random variables. Unfortunately, the correlations in this work were assumed to be zero since the data was collected in non-equidistant time periods. The covariance structure for the linear mixed model was chosen accordingly namely no correlation between measurements for a patient. The longitudinal data was treated as an independent sample in this work.

Abdominal aortic aneurysm (AAA) is especially important to study due to its high mortality rate among people older than 65. Gender, ethnicity and smoking are the biggest risk factors. Ruptured abdominal aortic aneurysms (AAAs) cause 15,000 deaths per year in US and 12,000 in UK [9, 10]. Since patients with risk factors are screened frequently by ultrasound in order to have control over the growth of the aneurysm, makes this group of patients interesting for mixed model analysis.

Another major issue to consider in such analyses is the assessment of the sample size. Sample size calculation is needed prior to the design of the study which makes algorithms created for this purpose of great importance.

The fact that in this work, 420 patients as a sample size was calculated by the “longpower” package in R using a linear mixed model showed the efficiency of this algorithm to calculate a sample size with a desired power prior to an experiment.

A major limitation in this work was the fact that only a data of placebo patients was provided. Data of treatment patients lacked which forced us to make a lot of assumptions in the study. E.g. correlation between random intercept and random slope for the treatment group was assumed to be 0 and the number of measurements for this group was assumed to be 10. In addition the improvement of the treatment group was simulated by reduction of the slope by 20%. Indeed, with a data of treatment patients much more accurate calculations could be done.

Although there were many variables registered in this study, the variables WHITE and SMOKER were missing. These variables are of great importance for emergence and development of abdominal aortic aneurysm [9, 10].

The calculated sample size by “longpower” was 420. Simulations of the model for control and treatment groups with sample sizes of 2*210 and 2*105 gave a difference in means at a statistical significance level of less than 0.1% in two-sample t-tests. In addition, for both simulations the fixed parameter estimates and the random variance estimates were statistically significant when run as linear mixed models. This fact showed that both sample sizes are acceptable until power calculations were done for those sample sizes. The power calculations indicated that a sample size of 2*210 gave much higher power than a sample size of 2*105.

Despite “longpower” being a very powerful tool, it needs at least one more improvement. Namely, one cannot insert in “longpower” function correlations between different measurements when the measurements are non-equidistant. Measurements close to each other are more correlated than measurements far from each other. Although “longpower” has undergone many improvements since its emergence in 1997 but more improvements are clearly expected. Its efficiency to calculate sample size and power will keep it practical for researchers in the near future.

It is important to remember that the results obtained in this work are only computations. These calculations need to be further examined and verified by more clinical experiments and statistical computations.

Future challenges

There is no doubt that sample size calculations and algorithms created for that purpose will continue to be of current interest. More and more of these algorithms will be programmed by statisticians and evaluated by clinical researchers.

The next step of studying abdominal aortic aneurysm is to include as many variables as possible which can be suspected to play a role for the emergence and development of this affliction. Indeed, bigger sample groups of both placebo and treatment patients are needed despite the cost of ultrasound screenings.

The study of abdominal aortic aneurysm will continue to be on the general agenda for the scientists in the future due to aging populations in the world. This affliction will affect millions of people in the future and necessary measures need to be taken.

In the light of this fact, advanced statistician algorithms, software and reasonably less costly screening apparatuses need to be developed and invented. These findings will not only help us to understand the nature of abdominal aortic aneurysm but will also help us to discover new drugs and treatments for AAA patients and other similar diseases and thereby save lives.

Patients and Methods

Data

The dependant variable (AD) was the aneurysm diameter (measured in mm) and the studied independent variables were:

Variable	Description
AD0	Aneurysm diameter at baseline (mm)
AD – AD0	Adjusted aneurysm diameter (mm)
RANGE	25<=AD0<35 == 0 35<=AD0<45 == 1 45<=AD0<55 == 2 AD0=>55 == 3
TIME	Measurements (year)
WOMAN	Woman == 1, Man == 0
DIABETES	Yes == 1, No == 0
AGE	

Table 11: The variables provided in the study of abdominal aortic aneurysm and the response variable (AD).

AD, AD0, ADJUSTED_AD, TIME and AGE were handled as continuous data whereas RANGE, WOMAN and DIABETES as categorical data. No missing values were encountered.

Software

SPSS

SPSS (originally, *Statistical Package for the Social Sciences*) is a computer program used for many different applications, among others survey authoring and deployment, data mining, text analytics and statistical analysis. SPSS version 20 was used in this work to analyze a linear mixed model and to draw respective diagrams. This is powerful software that can do many things and works perfectly with any kind of pharmaceutical data.

Unfortunately, SPSS has a poor export function for texts like tables as a graphic format. It made me first save my tables in other software to be able to save them as graphics. Another limitation was SPSS’ inability to plot a linear mixed model.

R

R is a programming language and software environment for statistical analysis and graphics. Most users agree that R is much more powerful than other popular statistical packages, such as SAS, SPSS and Stata.

R is an extensive programming tool partly because of its user-created packages. This feature makes it possible to do specialized statistical techniques, graphical devices, import/export capabilities, reporting tools and so much more. Importantly, it was possible to plot a linear mixed model in R which makes it easier for the researcher to comprehend and visualize the studied model.

R version 3.0.0 was used in this work to do the statistical calculations and simulations. Especially the "longpower" package written by Liu and Liang (1997) was used for our pharmaceutical data.

Statistical analyses

Linear mixed model

A linear mixed model is a linear statistical model containing both fixed and random effects that allows for correlation between measurements. This model was used in this work because there were mixed effects in the data due to repeated screenings on the same person during time intervals.

The model was run in SPSS choosing the logarithm of aneurysm diameter (Log10_AD) as dependent variable, the logarithm of aneurysm diameter at baseline (Log10_AD0) as fixed and time (TIME) as both fixed and random variable.

The covariance structure was chosen in accordance to Littell et al. (2000, [4]). The covariance structure "Variance components" in SPSS was chosen since our pharmaceutical data lacked a repeated variable and no correlation between the non-equidistant measurements was assumed. To express it statistically:

$\mathrm{Cov}(Y_{ijk}, Y_{ijl}) = \sigma^2_b = 0 \quad , \quad k \neq l$

$\mathrm{Var}(Y_{ijk}) = \mathrm{Cov}(Y_{ijk}, Y_{ijk}) = \sigma^2_b + \sigma^2_w = 0 + \sigma^2_w = \sigma^2_w$

i = individual
j = time
k, l = measurement
σ^2_b = between-individual variance
σ^2_w = within-individual variance

Variance Components:

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}$$

Statistical power

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false i.e. the probability of not committing a Type II error. Since the probability of a Type II error is referred as β, the power is referred as 1 – β.

Power analysis can be used to calculate the minimum sample size required so that one can likely detect an effect of a given size. Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size.

A simple example: Assume

$$\begin{cases} H_0: \mu = 50 \\ H_1: \mu \neq 50 \end{cases}$$

and true value of the mean μ = 52, sample size n = 10 and Type II error β = 0.264.

So the power is: 1 – β = 1 – 0.264 = 0.736 = 73.6% when μ = 52.

Power is a measure of sensitivity of a statistical test and in this example it means that if the true value of mean is really 52, this test will correctly reject the null hypothesis 73.6% of the time. If this value of power is judged to be too low, an increase in sample size n can be considered.

Two-sample t-test

This test was used to compare the means of two simulated control and treatment groups with the same sample size and equal variances. The following hypotheses were considered:

$$\begin{cases} H_0: \bar{X}_1 - \bar{X}_2 = 0 \\ H_1: \bar{X}_1 - \bar{X}_2 \neq 0 \end{cases}$$

The t-statistic to test if the means are different is: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(S_{X_1}^2 + S_{X_2}^2)/n}}$

$\bar{X}_i$ = mean of each group

$S_{X_i}^2$ = variance of each group

i = 1, 2

2n – 2 = degrees of freedom for this test

n = sample size

The denominator of t is the standard error of the difference between two means.

Finally, the null hypothesis is rejected if $t_{obs} > t_{critical}$.

Acknowledgements

I wish to express my gratitude to my supervisor, senior university lecturer Mattias Sundén at Gothenburg University who enthusiastically listened to my questions and results and gave me his comments and guidance throughout the whole work.

I would also like to thank my other supervisor, statistical science director Ziad Taib at AstraZeneca who structured and planned the project and did a final review of the thesis.

And finally, my special thanks to my family, particularly Celina, Leon and Noah, for their support and patience throughout my thesis.

References

Written sources

[1] **M. C. Donohue, S. D. Edland, A. C. Gamst:** Power for linear models of longitudinal data with applications to Alzheimer's Disease Phase II study design, *Division of Biostatistics and Bioinformatics, University of California, San Diego, January 22, 2013.*

[2] **P. J. Diggle, P. Heagerty, K.-Y. Liang, S. L. Zeger:** Analysis of Longitudinal Data, second edition, *Oxford statistical science series 25, 2002.*

[3] **G. Liu, K.-Y. Liang:** Sample Size Calculations for Studies with Correlated Observations, *BIOMETRICS 53, 937-947, September 1997.*

[4] **R. C. Littell, J. Pendergast and R. Natarajan:** Modelling covariance structure in the analysis of repeated measures data, *STATISTICS IN MEDICINE; 19:1793-1819, 2000.*

[5] **M. C. Donohue, A. C. Gamst, S. D. Edland:** Sample size calculation for longitudinal data, *February 15, 2013.*

[6] **J. T. Powell, M. J. Sweeting, L. C. Brown, S. M. Gotensparre, F. G. Fowkes, S. G. Thompson:** Systematic review and meta-analysis of growth rates of small abdominal aortic aneurysms, *Meta-analysis, 2011*.

[7] **K. A. Vardulaki, T. C. Prevost, N. M. Walker, N. E. Day, A. B. M. Wilmink, C. R. G. Quick, H. A. Ashton, R. A. P. Scott:** Growth rates and risk of rupture of abdominal aortic aneurysms, *British Journal of Surgery, 85, 1674-1680, 1998.*

Internet sources

[8] **http://en.wikipedia.org/wiki/File:Abdominal_aortic_aneurysm.JPG** [Accessed 07 September 2012]

[9] **http://www.uptodate.com/contents/abdominal-aortic-aneurysm-beyond-the-basics** [Accessed 24 January 2013]

[10] **http://emedicine.medscape.com/article/1979501-overview** [Accessed 24 January 2013]

Appendix I

(Using SPSS and R)

```
#Program: Calculating delta (difference of slopes of control and
#treatment patients) with a given effect (%20) for an abdominal
#aortic aneurysm data. Delta is an argument in "longpower".
#All parameter estimates were retrieved using SPSS.
#Author: Eric Markarian
#Date: 01.04.2013

#These estimates were retrieved after running linear mixed model in SPSS
fixed.intercept     <- 0.023364   #Fixed intercept coefficient estimate
fixed.time          <- 0.023846   #Fixed time slope coefficient estimate
fixed.log10AD0      <- 0.983406   #Fixed log10AD0 slope coefficient estimate
LOG10AD0            <- 1.5702     #Mean of log10AD0
n                   <- 2*1        #One control and one treatment patient
p                   <- 10         #Number of measurements

#Create a log10AD matrix and define a simple model
log10AD <- matrix(nrow=n, ncol=p)
for(i in 1:n) {
        if (i > n/2)
                fixed.time <- 0.8 * fixed.time          #For treatment patient
        for(t in 1:p) {
                log10AD[i, t] <- fixed.intercept +
                        fixed.log10AD0*LOG10AD0 + fixed.time*t
        }
}

#Split log10AD to control and treatment patients for plotting
log10AD.control     <- log10AD[1:(n/2),]
log10AD.treatment   <- log10AD[((n/2)+1):n,]

#Create time variable and time matrix
time            <- seq(0, 9, 1)
time.matrix     <- matrix(time, nrow=n/2, ncol=p, byrow = TRUE)

#Create a device with specific size with place for 1 figure
dev.new(width=60, height=32)
par(mfrow = c(1,1))

#Plot log10AD.control and log10AD.treatment vs time.matrix
plot(time.matrix, log10AD.control,
        xlim=c(-0.1,10), ylim=c(1.5,2),
        xlab="Time", ylab="Log10 AD", pch=19)
par(new=TRUE)
plot(time.matrix, log10AD.treatment, col="green",
        xlim=c(-0.1,10), ylim=c(1.5,2),
        xlab="Time", ylab="Log10 AD", pch=19)
```

```
abline(h = 1.74, col="red")     #Rupture line

log10AD.control.line <- lm(log10AD.control ~ time)
summary(log10AD.control.line)        #Gave control slope 0.02385

log10AD.treatment.line <- lm(log10AD.treatment ~ time)
summary(log10AD.treatment.line)    #Gave treatment slope 0.01908
```

Appendix II

(Using SPSS and R)

```
#Program: Calculating sample size with a given effect (%20) for an abdominal
#aortic aneurysm data based on the formula of Liu and Liang (1997)
#using "longpower" package in R.
#All variance estimates were retrieved using SPSS.
#Author: Eric Markarian
#Date: 01.04.2013

library(longpower)

#These estimates were retrieved using SPSS
sigma2.i            <- 0.000130          #Variance of random intercept
sigma2.s            <- 0.000300          #Variance of random slope
sigma2.e            <- 0.000323          #Residual variance

#Covariance of slope and intercept
#Correlation between random intercept and random slope was assumed to be 0
covariance.s.i <- 0*sqrt(sigma2.i)*sqrt(sigma2.s)

#t is number of measurements for treatment group
#10 measurements for treatment group were assumed here
t = seq(0, 9, 1)
n = length(t)

#Covaiance structure
covariance.t <- function(t1, t2, sigma2.i, sigma2.s, covariance.s.i){
       sigma2.i + t1*t2*sigma2.s + (t1+t2)*covariance.s.i}
R = outer(t, t, function(x,y){covariance.t(x, y,
       sigma2.i, sigma2.s, covariance.s.i)})
R = R + diag(sigma2.e, n, n)

#A covariate vector associated with the parameter of interest
u = list(u1 = t, u2 = rep(0,n))

#A covariate vector associated with the nuisance parameter
v = list(v1 = cbind(1,1,rep(0,n)), v2 = cbind(1,0,t))

#The slope values were calculated in Appendix I
control.slope        <- 0.02385
treatment.slope      <- 0.01908

#Run the liu.liang.linear.power function to obtain the sample size
print(liu.liang.linear.power(delta=control.slope-treatment.slope, u=u,
       v=v, R=R, sig.level=0.05, power=0.8, alternative="two.sided"), "\n")
```

Appendix III

(Using only R)

```
#Program: Calculating sample size with a given effect (%20) for an abdominal
#aortic aneurysm data based on the formula of Liu and Liang (1997)
#using "longpower" package in R.
#The variance estimates were retrieved after printing the R model below.
#Author: Eric Markarian
#Date: 01.04.2013

library(xlsx)
library(longpower)
library(lme4)
library(languageR)

fil <- read.xlsx("AAA.xlsx", 1)   #Read a .xlsx file's 1st sheet

#Define the linear mixed model with default Variance-Covariance structure (no
#correlation between measurements for the same patient)
model <- lmer(fil$LOG10_AD ~ 1 + fil$LOG10_AD0 + fil$TIME +
        (1 + fil$TIME | fil$SUBJECT), REML = FALSE)
print(model)

#Calculate the residuals and the fitted values
residuals       <- resid(model)
fitted.values   <- fitted(model)

#Create a device with specific size with place for 2 figures
dev.new(width=40, height=16)
par(mfrow = c(1,2))

#Plot fitted values of the model vs residuals
plot(fitted.values, residuals,
  xlab="Fitted values of Log10 AD",
  ylab="Residuals", main="Figure 1", pch=19)
abline(h = 0)

#Create a histogram over the residuals
hist(residuals, density=20, breaks=20, prob=TRUE,
        xlab="Residuals of Log10 AD", ylim=c(0, 35), main="Figure 2")
curve(dnorm(x, mean=mean(x),sd=sd(x)),
    col="blue", lwd=2, add=TRUE)

#Inserted estimates are from the output of the model above
sigma2.i   <- 0.00012363      #Variance of random intercept
sigma2.s   <- 0.00027201      #Variance of random slope
sigma2.e   <- 0.00031904      #Residual variance
```

```
#Covariance of slope and intercept
#Correlation between random intercept and random slope was assumed to be 0
covariance.s.i <- 0*sqrt(sigma2.i)*sqrt(sigma2.s)

#t is number of measurements for treatment group
#10 measurements for treatment group were assumed here
t = seq(0, 9, 1)
n = length(t)

#Covaiance structure
covariance.t <- function(t1, t2, sigma2.i, sigma2.s, covariance.s.i){
        sigma2.i + t1*t2*sigma2.s + (t1+t2)*covariance.s.i}
R = outer(t, t, function(x,y){covariance.t(x, y,
        sigma2.i, sigma2.s, covariance.s.i)})
R = R + diag(sigma2.e, n, n)

#A covariate vector associated with the parameter of interest (random var?)
u = list(u1 = t, u2 = rep(0,n))

#A covariate vector associated with the nuisance parameter (fixed var?)
v = list(v1 = cbind(1,1,rep(0,n)), v2 = cbind(1,0,t))

#The slope values were calculated in Appendix I
control.slope           <- 0.02385
treatment.slope         <- 0.01908

#Run the liu.liang.linear.power function to obtain the sample size
print(liu.liang.linear.power(delta=control.slope-treatment.slope, u=u,
        v=v, R=R, sig.level=0.05, power=0.8, alternative="two.sided"), "\n")
```

Appendix IV

(Using SPSS and R)

```
#Program: Simulations of a linear mixed model with a given sample size with
#two different slopes of time for an abdominal aortic aneurysm data.
#Finally, a two-sample t-test compares these two simulations.
#All parameter and variance estimates were retrieved using SPSS.
#Author: Eric Markarian
#Date: 01.04.2013

#The parameter and variance estimates' values are inserted after running the
#linear mixed model in SPSS. The number of patients is calculated by
#"longpower" package in R.
fixed.intercept      <- 0.023364    #Fixed intercept coefficient estimate
fixed.time           <- 0.023846    #Fixed time slope coefficient estimate
fixed.log10AD0       <- 0.983406    #Fixed log10AD0 slope coefficient estimate
sigma.i              <- sqrt(0.000130)      #Standard deviation of intercept
sigma.s              <- sqrt(0.000300)      #Standard deviation of time slope
sigma.e              <- sqrt(0.000323)      #Standard deviation of standard error
n                    <- 2*105       #Number of control and treatment patients
p                    <- 10          #Number of measurements

#Create a log10AD matrix and define the simulation model
log10AD <- matrix(nrow=n, ncol=p)
for(i in 1:n) {
        random.intercept        <- rnorm(1, 0, sigma.i)         #Random intercept
        random.time             <- rnorm(1, 0, sigma.s)         #Random time
        if (i > n/2)
                fixed.time <- 0.8 * fixed.time          #For treatment group
        for(t in 1:p) {
                log10AD[i, t] <- fixed.intercept + random.intercept +
                        fixed.log10AD0 + (fixed.time + random.time)*t +
                        rnorm(1, 0, sigma.e)   #Random error
        }
        if (i > n/2)
                fixed.time <- fixed.time / 0.8          #Restore fixed.time
}

#Split log10AD to control and treatment groups for plotting
log10AD.control      <- log10AD[1:(n/2),]
log10AD.treatment    <- log10AD[((n/2)+1):n,]

#Create time variable and time matrix
time            <- seq(0, 9, 1)
time.matrix     <- matrix(time, nrow=n/2, ncol=p, byrow = TRUE)

#Create a device with specific size with place for 1 figure
dev.new(width=60, height=32)
par(mfrow = c(1,1))
```

```
#Plot log10AD.control and log10AD.treatment vs time.matrix
plot(time.matrix, log10AD.control,
        xlim=c(-0.1,10), ylim=c(0,2.5),
        xlab="Time", ylab="Simulated Log10 AD", pch=19)
par(new=TRUE)
plot(time.matrix, log10AD.treatment, col="green",
        xlim=c(-0.1,10), ylim=c(0,2.5),
        xlab="Time", ylab="Simulated Log10 AD", pch=19)
abline(h = 1.74, col="red")     #Rupture line

#Two-sample t-test for comparing the means of the control and the
#treatment groups.
print(t.test(log10AD.control, log10AD.treatment, alternative="two.sided",
        var.equal=TRUE))

#Write the output to a file
sink("HelpFile.txt")
array(t(log10AD), c(n*p,1))
sink()
```

Appendix V

(Using SPSS and R)

```
#Program: Calculating delta (difference of slopes of control and treatment
#patients) with a given effect (%20) for a simulated abdominal aortic aneurysm
#data à 2*210 patients. Delta is an argument in "longpower".
#All parameter estimates were retrieved using SPSS.
#Author: Eric Markarian
#Date: 01.04.2013

#These estimates were retrieved after running linear mixed model in SPSS
fixed.intercept          <- 1.006002   #Fixed intercept coefficient estimate
fixed.time.control       <- 0.022251   #Fixed time*control coefficient estimate
fixed.time.treatment     <- 0.018790   #Fixed time*treatment coefficient estimate
n                        <- 2*1        #One control and one treatment patient
p                        <- 10         #Number of measurements

#Create a log10AD matrix and define a simple model
log10AD <- matrix(nrow=n, ncol=p)
for(i in 1:n) {
        if (i > n/2)
                fixed.time.control <- 0.8 * fixed.time.control#For treatment patient
        for(t in 1:p) {
                log10AD[i, t] <- fixed.intercept +
                        fixed.time.control*t
        }
}

#Split log10AD to control and treatment patients for plotting
log10AD.control          <- log10AD[1:(n/2),]
log10AD.treatment        <- log10AD[((n/2)+1):n,]

#Create time variable and time matrix
time            <- seq(0, 9, 1)
time.matrix     <- matrix(time, nrow=n/2, ncol=p, byrow = TRUE)

#Create a device with specific size with place for 1 figure
dev.new(width=60, height=32)
par(mfrow = c(1,1))

#Plot log10AD.control and log10AD.treatment vs time.matrix
plot(time.matrix, log10AD.control,
        xlim=c(-0.1,10), ylim=c(1,2),
        xlab="Time", ylab="Log10 AD", pch=19)
par(new=TRUE)
plot(time.matrix, log10AD.treatment, col="green",
        xlim=c(-0.1,10), ylim=c(1,2),
        xlab="Time", ylab="Log10 AD", pch=19)
abline(h = 1.74, col="red")     #Rupture line
```

```
log10AD.control.line <- lm(log10AD.control ~ time)
summary(log10AD.control.line)                    #Gave control slope 0.02225

log10AD.treatment.line <- lm(log10AD.treatment ~ time)
summary(log10AD.treatment.line)                  #Gave treatment slope 0.0178
```

Appendix VI

(Using SPSS and R)

```
#Program: Calculating power with a given effect (%20) for a simulated
#abdominal aortic aneurysm data à 2*210 patients based on the formula
#of Liu and Liang (1997) using "longpower" package in R.
#All variance estimates were retrieved using SPSS.
#Author: Eric Markarian
#Date: 01.04.2013

library(longpower)

#These estimates were retrieved using SPSS
sigma2.i            <- 0.000104        #Variance of random intercept
sigma2.s            <- 0.000308        #Variance of random slope
sigma2.e            <- 0.000329        #Residual variance

#Covariance of slope and intercept
#Correlation between random intercept and random slope was assumed to be 0
covariance.s.i <- 0*sqrt(sigma2.i)*sqrt(sigma2.s)

#t is number of measurements for treatment group
#10 measurements for treatment group were assumed here
t = seq(0, 9, 1)
n = length(t)

#Covaiance structure
covariance.t <- function(t1, t2, sigma2.i, sigma2.s, covariance.s.i){
        sigma2.i + t1*t2*sigma2.s + (t1+t2)*covariance.s.i}
R = outer(t, t, function(x,y){covariance.t(x, y,
        sigma2.i, sigma2.s, covariance.s.i)})
R = R + diag(sigma2.e, n, n)

#A covariate vector associated with the parameter of interest
u = list(u1 = t, u2 = rep(0,n))

#A covariate vector associated with the nuisance parameter
v = list(v1 = cbind(1,1,rep(0,n)), v2 = cbind(1,0,t))

#The slope values were calculated in Appendix V
control.slope       <- 0.02225
treatment.slope     <- 0.0178

#Run the liu.liang.linear.power function to obtain the power
print(liu.liang.linear.power(N=2*210, delta=control.slope-treatment.slope, u=u,
        v=v, R=R, sig.level=0.05, alternative="two.sided"), "\n")
```