Survival Analysis, Longitudinal Data Analysis, and Joint Modeling of Survival and Longitudinal Models

Kimberly Schveder

Advisor: Dr. Datta

Faculty Sponsor Approval:

Sujay Datta — 05/13/20

Faculty Sponsor Signature — Date

Faculty Reader Approval:

Jun Ge — 5/13/2020

Faculty Reader Signature — Date

Department Chair Approval:

Tim O'Neil — 5/13/2020

Department Chair Signature — Date

Survival Analysis, Longitudinal Data Analysis, and Joint Modeling of Survival and Longitudinal Models

Kimberly Schveder

Advisor: Dr. Datta

Faculty Sponsor Approval:

_____ _____

Faculty Sponsor Signature Date

Faculty Reader Approval:

_____ _____

Faculty Reader Signature Date

Department Chair Approval:

_____ _____

Department Chair Signature Date

Abstract:

Survival Analysis and Longitudinal data analysis are so often used in medical research, such as in clinical trials and in modeling observational data. The joint model that combines parametric survival models and the linear mixed effect longitudinal model into one model. This paper provides a detailed summary of each of these data analysis tools and models. Survival and Longitudinal models are fit onto a few different datasets that involve heart transplant data, VA lung cancer data, and data from the Framingham Heart Study.

Summary:

In the survival analysis section of this paper, the definitions of censoring, truncation, the hazard and survival function, proportional hazards models, the proportional hazards assumption, accelerated failure time model, and frailty are reviewed. Then, methods of estimating hazard and survival functions are described. Such methods include fitting a Kaplan-Meier plot or the Nelson-Aalen Estimation to make a Flemington-Harrison plot, various parametric models, and the semi-parametric Cox PH model can be fit on the survival data. The parametric models described are the exponential distribution model, the Weibull distribution model, the lognormal distribution model, the log-logistic distribution model, Gompertz distribution model, and the Gamma distribution model. Survival Estimation model comparison techniques are discussed, such as the use of the log-rank test and similar tests. The steps of modeling the data are discussed. Such steps include model selection and interpretation, model assessment and diagnostics, working with time-dependent covariates, and working with multiple survival outcomes and competing risks. Data analysis of heart transplant data and a full data analysis of lung cancer data from a VA hospital are completed to demonstrate the theory described.

The longitudinal data analysis section of this paper includes the definitions of longitudinal and cluster data, describes the motivation behind why longitudinal data analysis is used, and distinguishes longitudinal data from time series data and cross-sectional data. The notation and explanations of the common terms in longitudinal data analysis are included. The mixed and fixed linear models are defined and distinguished. Estimation and statistical inference techniques for estimating such linear models are described, including using maximum likelihood estimation, hypothesis testing, confidence intervals and contrast for the linear model coefficient parameter. Mean response modeling techniques are discussed, such as polynomial trends over time and linear splines. Covariance modeling techniques are described, and different types of covariance models (i.e. structured and unstructured models). The use of general linear models in longitudinal data analysis are described. Marginal models that model the mean and the generalized estimating equations which estimate marginal models are discussed. Residual analyses and diagnostics and methods for dealing with missing data and dropout are described. The methods described are applied to the Framingham heart study dataset.

The joint modeling of survival and longitudinal models is described. The discussion of why they are used and beneficial for use in clinical trials is included. If the data to make a survival model and a longitudinal model is available, it might be worth fitting a joint model because this model is overall very efficient and reduces bias in the estimates of the treatment effect. The notation of the combination of the linear mixed effect model and a parametric survival model Weibull are shown and discussed. Methods of estimation of the joint model are discussed, as well as methods for assessing the fit of the model with residuals.

Table of Contents

Survival Analysis..... page 5

- Introduction to Survival Analysis..... page 5
- Statistical Methods and Survival Analysis Concepts Discussion ..... page 5
    - Censoring ..... page 5
    - Truncation ..... page 6
    - The Hazard and the Survival Function ..... page 7
    - Proportional Hazard Models ..... page 7
    - Accelerated Failure Time (AFT) model ..... page 10
    - Frailty ..... page 11
    - Estimating hazard and survival functions ..... page 13
        - Non-parametric and common beginner models ..... page 13
            - Kaplan-Meier Estimator ..... page 13
            - Nelson-Aalen Estimator and Flemington-Harrison ..... page 14
        - Parametric Models ..... page 15
            - Exponential distribution model ..... page 16
            - Weibull distribution model ..... page 16
            - Log-normal distribution model ..... page 17
            - Log-logistic distribution model ..... page 18
            - Gompertz distribution model ..... page 20
            - Gamma distribution model ..... page 20
        - Semi-Parametric Model ..... page 20
            - Cox-Proportional Hazard Model ..... page 20
    - Comparing Methods of Survival Estimation ..... page 22
        - Log-rank test ..... page 22
        - Other tests ..... page 23
    - Methods of Modeling the Data..... page 23
        - Model selection and interpretation..... page 23
            - Covariance adjustment..... page 23
        - Categorical and continuous covariates..... page 25
        - Hypothesis testing for comparing nested models..... page 26
        - The Akaike Information Criterion for Comparing Non-nested models..... page 28
        - Including smooth estimates of continuous covariates in a survival model..... page 28
    - Model assessment and diagnostics ..... page 29
        - Assessing goodness of fit (GOF) using residuals ..... page 29
            - Martingale and Deviance residuals ..... page 29
            - case deletion residuals ..... page 30
        - Checking the PH assumption .... page 31
            - log cumulative hazard plots ..... page 31
            - Schoenfeld residuals ..... page 32
    - Working with time-dependent covariates ..... page 33
    - Working with multiple survival outcomes and competing risks ..... page 38
- Full data analysis ..... page 39
- Conclusions on Survival Analysis ..... page 49

Longitudinal Data Analysis ..... page 49

- Longitudinal and clustered data ..... page 49
- Regression for correlated responses ..... page 50

- Basics of longitudinal data ….. page 52
- Linear models ….. page 57
    - Notation and Distributional assumptions ….. page 57
    - Descriptive methods of analysis ….. page 58
    - Modeling the mean and covariates ….. page 60
- Estimation and Statistical Inference ….. page 61
    - Maximum Likelihood estimation ….. page 61
    - Statistical inference ….. page 63
- Mean Response Modeling ….. page 64
    - Polynomial trends with time ….. page 66
    - Linear splines ….. page 67
    - GLM formulation ….. page 68
- Covariance Modeling ….. page 69
    - Implications of Correlation among Longitudinal Data ….. page 70
    - Unstructured Covariance ….. page 70
    - Covariance Pattern Models (Structured Covariance) ….. page 70
- Linear Mixed Effects Models ….. page 73
    - Definition ….. page 73
    - The two-stage random effects formulation ….. page 74
    - Choice among random effects covariance models ….. page 75
- Fixed Effects vs. Mixed Effects Models ….. page 75
    - Definition: Linear Fixed Effects Models ….. page 75
    - Fixed Effects versus Mixed Effects: Bias-Variance Trade-off….. page 76
- General linear models (GLMs) for Longitudinal Data ….. page 76
- Definition and Feature of GLMs ….. page 76
    - Ordinal Regression ….. page 80
- Residual Analyses and Diagnostics ….. page 80
    - Define Transformed Residuals ….. page 81
    - Define Aggregating Residuals ….. page 81
    - Residuals in the context of GLMs ….. page 82
    - Define: Semi-Variogram ….. page 82
- Marginal Models ….. page 83
- Definition (in general) ….. page 83
- Generalized Estimating Equations (GEE) ….. page 84
- Missing Data ….. page 85
    - Issues with Missing Data ….. page 85
    - Methods for Dealing with dropout ….. page 86
    - Using Multiple Imputation and Weighting Methods ….. page 87
- Longitudinal data analysis in R Framingham heart study data ….. page 88

Joint Modeling of Survival and Longitudinal Models ….. page 103

Acknowledgements ….. page 107

References ….. page 107

Appendix ….. page 109

R code ….. page 111

- Introduction to Survival Analysis

Survival analysis is the study of survival time until there is some ***event*** that happens to individuals or subjects. "We also typically refer to the event as a failure, because the event of interest usually is death, disease incidence, or some other negative individual experience" (Kleinbaum 4). In other words, the event can be anything from such as death, or response to some stimuli. Thus, survival analysis is widely used in the biological and biomedical sciences to model the time to event data, such as time to death data.

When performing survival analysis, there are a few goals that should be kept in mind to stay on track with the analysis. One is to see if there is any relationship between the different explanatory variables to the response variable survival time, usually denoted as *t*. Another is to take the survival data and to fit (an) estimated survival and hazard function(s) with it (Kleinbaum 15). Then, the survival model(s) are analyzed and interpreted. In addition, the different survival and hazard functions can be compared.

The data that will be used is from the data bank in Sneeley's STAT 553 course content at Rice University, in Houston, Texas.[13] This set of datasets include heart transplant data, ovarian cancer data (chemotherapy after surgical treatment), cervical cancer data, brain cancer data, bladder cancer data, lupus nephritis data, primary biliary cirrhosis data, and lastly lung cancer data from a Veterans Administration (VA) hospital. Many of these datasets will be used to show the various models, plots, and calculations throughout the Statistical Methods section and in the Full Data analysis section.

- Statistical Methods and Survival Analysis Concepts Discussion
  - Censoring

Censoring occurs when there is information about an individual's survival time, but the exact survival time is not known. A ***censoring period*** is the interval during which the censoring takes place. In a censoring period, the event may or may not have occurred. Censoring is important and often used because some observations, such as if people are being studied, do not have the event take place before the study ends, or perhaps they did not follow-up or were withdrawn from the study, so the information on the event is not recorded. A "failure" of the event is denoted as 1, or x in Figure 1 below, and a censored observation is denoted as 0 Jenkins (4-5).

There are three types of censoring: right, left, and interval. Right censoring, the most common type, occurs when an observation's length of time from entrance of the censoring period is unknown because the event hasn't occurred (yet) in the censoring period. According to Jenkins[7], "Given entry at time 0 and observation at [survival] time t, we only know that the completed spell is of length $T > t$", where T is some random time. In Figure 1, Subjects B and C are both right censored because the observed individuals do not experience the event during the study period 0 to time 12 or 10, respectively. The observations experienced the event after the study is over. This is not to be confused with subject A in Figure 1, who experienced the event at time 7.

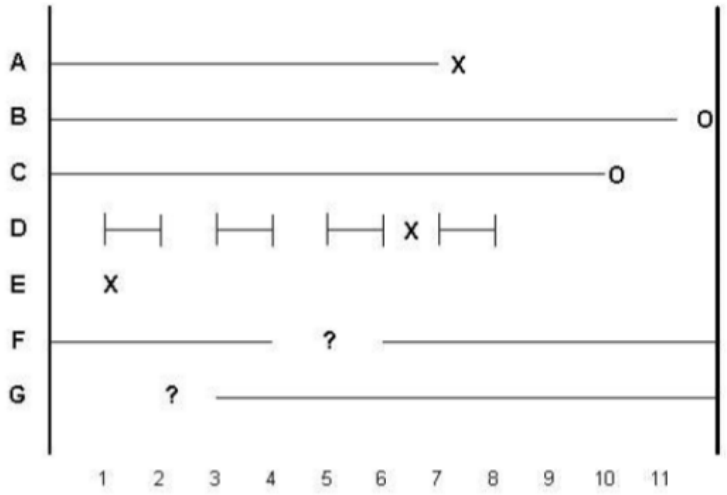


Figure 1: A graph of Censoring and Truncation[12]

Left censoring is the opposite of right censoring, as shown in Figure 1 as subject E, in that the starting time and date of the censoring period was not observed, but the event is known occur before the observation's entrance into the study. Left censoring is used more often in the social sciences, like in psychology, and is also used in the biological and biomedical sciences, though less often than is right censoring. Lastly, interval censoring is when we have an observation that experienced the event in some interval of time on the timeline of the study[12]. But, the exact time(s) are unknown. In Figure 1, subject D is interval censored.

- Truncation

A truncation period is when the event cannot occur. There are three types of truncation: left, right, and interval[12]. Left truncation occurs when observations that have survived past a cut-off point of survival time are included in the study. With left truncation, "a subject is left truncated if it enters the population at risk some stage after the start of the follow-up period". Refer to Figure 1, subject G to see left truncation visually. As said by Stevenson[12], "in a study investigating the date of first BSE diagnosis on a group of farms, the farms that are established after the start of the study are said to be left truncated (the implication here is that there is no way the farm can experience the event of interest before the truncation date)"[12]. Right truncation is very different.

According to Jenkins[7], right truncation occurs when observations that have experienced the event by some survival time are included in the study and the observation leaves the study. Stevenson[12], "For example, in a study investigating the date of first foot-and-mouth disease diagnosis on a group of farms, those farms that are pre-emptively culled as a result of control measures are right truncated on the date of culling." Thus, it is known that the event occurred and will not occur again after the observations in the dataset have left the study.

Lastly, interval Truncation is when "there is no way possible that the event of interest could occur" to a subject within an interval of time, according to Stevenson[12]. Shown in Figure 1, subject F is interval truncated. The subject could not experience the event in-between times 4 and 6.

- The Hazard and the Survival Function

A hazard function is the probability that a subject has survived to time t. This is also called an instantaneous failure rate, "intensity function", or the "force of morality", according to Moore[9], expressed as in equation 1 below.

$$h(t) = \lim_{\delta \to 0} \frac{pr(t<T<t+\delta | T>t)}{\delta} = -\frac{\partial S(t)}{\partial t} \geq 0 \text{ (equation 1)}$$

Equation 1 expresses the probability that a subject has survived to time t, divided by the length of that time interval, $\delta = \Delta t$.

The cumulative hazard, or integrated hazard, is expressed as in equation 2[7].

$$H(t) = \int_0^t h(u)du = -\ln [S(t)] \geq 0 \text{ (equation 2)}$$

The Survival function, which "defines the probability of surviving up to a point t", according to Moore[9], is expressed as in equation 3.

$$S(t) = pr(T > t) = e^{-H(t)} = e^{-\int_0^t h(u)du}, \ 0 < t < \infty \text{ (equation 3)}$$

The survival function, at time 0, will have a value of 1, and "decreases (or remains constant) over time and of course never drops below 0", as stated by Moore[9]. The survival function is, in fact, a right continuous function. See the parametric and semi-parametric distribution sections below, on pages 15 and 20 respectively, for examples of the parametric and semi-parametric plots of each of these three functions for the different types of models mentioned here.

- Proportional Hazard Models

Proportional Hazard (PH) Models are understood to have a separability assumption, by multiplication. The proportional hazard function can be split into two parts, as shown in equation 4.

$$h(t, X) = h_0(t)e^{\beta X} = h_0(t)\ \tau \text{ (equation 4)}$$

The first part in this equation, $h_0(t)$, is the "baseline hazard" that depends *only* on the time t. According to Jenkins[7], "It summarizes the pattern of 'duration dependence', assumed to be common to all persons." Next, the nonnegative variable $\tau = e^{\beta X}$ is dependent only on the covariates or vector of time-independent variables (in the cases discussed here), represented as X.

The proportional hazards function has several assumptions that need to be met and are different for the parametric and semi-parametric models of survival data (discussed below). The ***hazard ratio*** is defined when the hazard models from two observations from a survival dataset, i and j, are proportioned as in equation 5.[7]

$$\frac{\hat{h}(\bar{t}, X_i)}{\hat{h}(\bar{t}, X_j)} = e^{(\beta X_i - \beta X_j)} = e^{(\beta (X_i - X_j))} \text{ (equation 5)}$$

The Proportional Hazard assumption states that the hazard model is constant over time, specifically for the Cox Proportional Hazards model[8]. The PH assumption is more commonly used for the Cox PH model. However, the exponential model always satisfies the PH assumption and

the Weibull model can satisfy the PH assumption in specific cases. (All of these models are defined in further detail below.) Note that, semi-parametric models, such as the Cox PH model, usually come to mind when evaluating the PH assumption.

The PH assumption can be evaluated through the use of plots, hypothesis tests, and other statistics. It is best to use a combination of these when checking the PH model assumption. The following three approaches are checking parallelism through plots, performing a goodness-of fit (GOF) hypothesis test, and/or using time-dependent variables statistics to evaluate the PH assumption[8].

There are two different ways to graphically model the survival models: Log-log plots, a.k.a. $-ln(-ln)$ survivor curves, and observed vs. the expected survival curves. The most widely used method is the $-ln(-ln)$ survivor curves. These curves model the survival data over different categories and combinations of categories of variables in the dataset that the investigator wants to look at. When these survival curves are parallel for the different categories, then we can say the PH assumption is met. Another way is to compare the observed and the predicted survivor curves to see how close they are. If the two curves are close, then the PH assumption is met for the PH model.

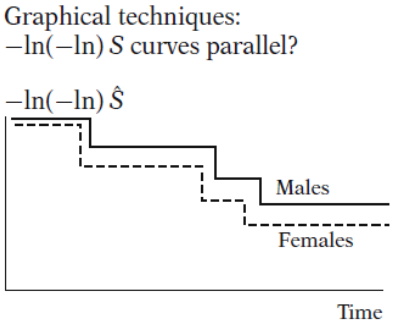


Figure 2: ln(-ln) curve plot example, with sex as the category. Image from Kleinbaum 136.

Plot of $-ln(-ln(\hat{S}))$ survival curves can be positive or negative. When plotting, substate the estimated survival function $\hat{S}$ in for $S(t, X)$ in the equation below, in equation 6.[8]

$$-ln(-ln(S(t,X)) = -\sum_{i=1}^{p} \beta_i X_i + \ln(-\ln(S_0(t))) \quad \text{(equation 6)}$$

After transforming the survival function and plotting $-ln(-ln(\hat{S}))$ vs. survival time, the plot needs to be interpreted, as explained more below.

The plots for each of the categories or subsets of categories should be parallel to conclude the satisfaction of the PH assumption. However, how do we define how parallel the plots should be to be considered parallel? That is, the decision to conclude the PH assumption is met based on parallel plots is *subjective*. The recommended method of dealing with this decision making is to *assume the PH assumption is met* for the appropriate models, unless there is strong evidence that the plots of the different survival curves for each of the categories are not parallel. If the curves for the different categories overlap or will soon cross, then they are not parallel, and the PH assumption is violated.

For example, in Figure 2, the $-ln(-ln(\hat{S}))$ survival curves for each of the categories are plotted vs. survival time for some generic survival data. The curves are grouped by sex. Notice that the two curves do not cross anywhere. Thus, the two survival curves are parallel.

The next graphical approach is the observed vs. the expected survival curves, which are considered to be the visual analog of the goodness-of-fit test. There are two strategies to this[8]. One of the strategies is to create Kaplan-Meier (KM) curves (discussed in further detail below) one-at-a-time to obtain observed plots. The data should be stratified by the categorical predictor variable. Then, for each of the categories of this variable, the observed and expected KM curves are obtained and compared among the different categories[8].

If the observed and the expected plots are very similar and are overlapping for all of the categories, then the PH assumption is met. An example of this situation is shown in Figure 3 below. If the observed and the expected curves for each of the categories are not overlapping, then the PH assumption isn't met. When the predictor variable is numerical, it should be broken up into strata so that the KM expected and observed survival curves make different KM survival curves for each of the strata. The strata are defined depending on the context of survival data and the investigator's question.

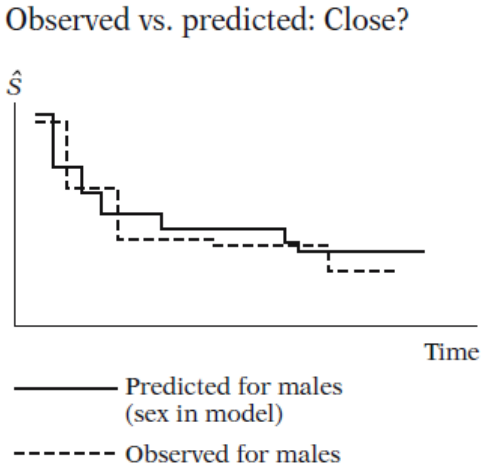


Figure 3: ln(-ln) curve plot example, with sex as the category. Image from Kleinbaum 136.

The other method to producing the observed vs. expected plots is to use stratified Cox PH (CPH) models to get the observation plots, while adjusting for other variables[8]. When you have a numerical predictor variable and need to compute the observed vs. expected survival plots, you can perform a Cox PH model. Such a model will have k categories (or strata, found by using the methods discussed above) and thus k-1 dummy variables in the CPH[8].

An adjusted survival curve plot is then created for each of the given categories using the dummy variables created above. The adjusted survival curve $\hat{S}(t, X_c)$ will be equal to what is shown in equation 7, where X_j are the dummy variables for each of the k categories.

$$\hat{S}(t, X_c) = [\hat{S}_0(t)]^{h(t,X)} = [\hat{S}_0(t)]^{e^{\sum_{j=1}^{k-1} \beta_j X_j}} \text{ (equation 7)}$$

The dummy variables that correspond to each of the categories are placed into the estimated survival curve formula.

Another option, specifically for numeric variables, is to use a PH model that uses the mean of each of the predictor values for each of the k categories ($\overline{X_k}$) in the adjusted survival curve formula[8]. The modified equation for the adjusted survival curve is shown in equation 8.

$$\hat{S}(t, \overline{X_k}) = [\hat{S}_0(t)]^{h(t,X)} = [\hat{S}_0(t)]^{e^{\hat{\beta}\,\overline{X_k}}} \text{ (equation 8)}$$

The investigator would then go through the analysis of these expected vs. observed plots using the same analysis strategy from above. All of these graphical methods are recommended but are best used with more objective tests, such as the GOF test.

What is the GOF test and what are its steps? The GOF test has three steps[8]: (1) Obtain the Schoenfeld residuals (described in further detail below, and defined for every observation that has an event) after running the CPH model; (2) take the failures in the data and rank them according to survival time; (3) take the ranked failure times and the Schoenfeld residuals and run a correlation test on these measures to see if they are correlated or not.

As Kleinbaum[8] mentioned, "The idea behind the statistical test is that if the PH assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time." Thus, the null hypothesis (H_0) of the GOF test is that the PH assumption is met. The alternative hypothesis (H_a) is that the PH assumption is violated. In general, it is best to not reject the H_0 of the GOF test. When the PH assumption is met, then the widely-used, robust, and powerful tests can be used. The p-value needed to verify the PH assumption must be greater than any reasonable significance level, such as $\alpha = 0.05$. Although the H_0 will never be proven to be true, when it is not rejected, then there is not have enough evidence (at the given significance level) that the PH assumption is not met.

- Accelerated Failure Time (AFT) model

The accelerated failure time (AFT) is an alternative to the PH model in describing entire families of survival time distributions, such as the parametric[7]. The AFT model describes the parametric models, whereas the PH model describes mainly semi-parametric models such as the Cox PH model. Furthermore, AFT models are used if the hazard function is a function of time. If not, then the AFT model should not be used. The main advantage of AFT models, when compared to PH models, is that "the effect of covariates on survival can be described in absolute terms (e.g. numbers of years)", according to Stevenson[12]. The Proportional hazards model looks at the effect of covariates on survival only in the relative terms with the use of the hazard ratio.

More precisely, according to Stevenson[12], "the acceleration factor is a ratio of survival times corresponding to any fixed value of S(t)." An acceleration function has a *constant factor* that describes the stretching-out, or contraction of survival functions when comparing one group to another. Letting the constant factor be represented as c, the AFT model equation typically looks something like what is shown in equation 9 and 10.

$$\log(t) = \beta X + \log(\tau) \text{ (equation 9)}$$

which can be written as

$$t = exp^{\beta X}\tau \text{ (equation 10).}$$

by exponentializing both sides[8].

In other words, according to Kleinbaum[8], "The AFT model is additive on the log scale but a multiplicative model with respect to t." In this case, c is equal to $t/\tau = e^{\beta X}$.[12] If $c = t/\tau = e^{\beta X} > 1$, the time will rapidly pass; if $c < 1$, then the time will pass at a slow pace; if $c = 1$, the time passes normally.

The survival functions of dogs and humans are equivalent when the survival time of humans is seven times that of the survival time of dogs. That is, $S_{Dogs}(t) = S_{Humans}(7t)$. Accordingly, "We might say the probability of a dog surviving past 6 years equals the probability of a human

surviving past 42 years because 42 equals 6 times 7", according to Kleinbaum[8]. Dogs have an accelerated lifespan of 7 time that of humans. It can probably be said that seven is the acceleration constant c.

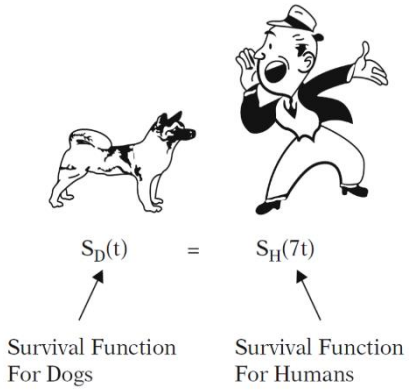


Figure 4: Comparing dog and human AFT, Kleinbuim 266

- Frailty

Frailty is defined as a random variability component in a survival model that is designed to deal with variability due to unobserved factors, such as heterogeneity, in the model[8]. This random variability component it has multiplicative effect on the hazard function such that it follows some distribution[8]. Let frailty be the constant a and the distribution function that it follows be the g(a), with a > 0, and E(g(a)) = 1. The V(g(a)) = o and this parameter o is estimated from the given survival data.

The hazard and survival conditions on frailty are, for the observable characteristics X and survival time t, h(t, X|a) = ah(X, t) and S(X, t|a) = (S(X, t))^a. Observations with an increased hazard: ah(X, t) > h(X, t) and decreased survival S(X, t)^a < S(X, t) have an a > 1. Next, when observations have an increased hazard: ah(X, t) < h(X, t) and decreased survival S(X, t)^a > S(X, t) have an a < 1. Lastly, observations with ah(X, t) = h(X, t) have a = 1 (an average frailty) [8].

Survival functions with frailty models have a conditional survival function S(X, t|a), which represents the individual level. Or survival function can have an unconditional survival function S_U(X, t), which represents the population level. S_U(X, t) is a population average and is related to the S(X, t|a) through integration as shown in equation 11.

$$S_U(t) = \int_0^\infty S(X, t|a)g(a)da \text{ (equation 11)}$$

Hence,

$$h_U(X, t) = \frac{-d[S_U(t)]/dt}{S_U(t)} \text{ (equation 12)}$$

is our instantaneous hazard function corresponding to S_U(t).[8] If a distribution has a frailty a > 0 and a mean of 1, such as gamma or inverse gamma, then it can be used to model frailty.

An example of frailty can be shown through the following scenario: perhaps a researcher is looking at modeling the survival of college students on getting/receiving treatment for a

disease, perhaps diabetes or an eating disorder, at a University. Most of the college students follow a certain demographic at the University, most are between the ages of 16 and 30 and are physically healthy. However, all students have different backgrounds and health issues, such as other disorders and diseases, which might not have been included in the dataset. Some of these students may be more at risk than others of getting the disease that's being studied.

There exists a population average survival function for all such college students, but each student has their individual survival function that introduces a random effect due to the heterogeneity of the students. Hence, frailty will likely need to be factored into the model of the data. Group-level random effects models can be used to analyze the college student data. Individual-level random effect data should not be used.

Shared frailty and unshared frailty are distinct from each other. Despite their differences, unshared and shared frailty both still take into consideration in their model types the variation due to any unobserved factors in the datasets collected by definition. Unshared frailty is the more common type of frailty and has been used up to this point. We have a set of independent observations (taken from a random sample) to construct an unshared frailty model. The population averages are interpreted with the unconditional survival and hazard functions mentioned above[8].

With shared frailty, the clusters of observations, such as people who are relatives or who share a *similar* environment will *share the same frailty*[8]. The shared frailty takes into consideration any within-cluster correlation. The hazard function on shared frailty models is like that of unshared frailty models, but accounts for the covariates in the model. Let there be a total of n_k subjects in a group (cluster) k, and j be each of the individuals (j = 1, 2, …., n_k). Then the conditional hazard for the shared frailty model is shown in equation 13.

$$h_{jk}(t|a_k) = a_k h_{jk}(t), where\ h_{jk}(t) = h(t|X_{jk}) \text{ (equation 13)}$$

Suppose that two individual's are going to be compared and the two individuals have the same frailty. Perhaps $h(t|a_j) = a_j h(t)$ is the jth subject's hazard, for $j = 1, 2, \dots, n$ subjects is given, where the Weibull's hazard function is denoted as $h(t)$ and a_j is the frailty of the jth subject. Note that the Weibull hazard function is "parameterized in terms of the predictor variables and their regression coefficients", according to Kleinbaum[8]. If a pair of observations have the same frailty, then the $a_w = a_k$, where w = k and the hazard ratio is shown in equation 14.

$$\frac{a_k h_k(t)}{a_w h_w(t)} = \frac{h_k(t)}{h_w(t)} \text{ (equation 14)}$$

A good way to deal with two individuals with the same frailty is to compare them. That is, to have one take the "test treatment and the other takes the standard treatment controlling for the other covariates in the model", according to Kleinbaum[8]. Therefore, in the case of a couple individuals with the same frailty being compared, the coefficient estimates from one of the models, such as the model of the second individual, can be used to estimate the conditional hazard ratio.

There may be omitted variables or measurement errors in the regressor variables or in the survival time records, so the unobserved factors might be relevant to these situations[7]. Frailty is important to consider, furthermore, because if the effect of heterogeneity in the data is not taken into account in the model, then the model "will over-estimate the degree of negative duration dependence in the hazard, according to Jenkins[7]. In other words, the hazard function's true proportionate response will be underestimated.

- Estimating hazard and survival functions
  - Non-parametric and common beginner models

What is a non-parametric model and how is it different from a parametric model? According to Stevenson, M.[12], "If the data are consistent with a parametric distribution, then parameters can be derived to efficiently describe the survival pattern and statistical inference can be based on the chosen distribution." When no theoretical distribution, such as any of the parametric models mentioned below, fit the survival data collected, non-parametric models are used to fit the data. Kaplan-Meier and Nelson-Aalen/Flemington-Harrington estimation method are both examples of non-parametric methods.

- Kaplan-Meier Estimator

A Kaplan-Meier estimator is a step function that estimates a survivor probability over the survival dataset. It is a basic way to estimate a survival curve. It is called a step function because, when plotted, the Kaplan-Meier estimator looks like a set of "stairs" going upwards or downwards. To make a Kaplan-Meier estimator, the data set needs to have the ascending survival times, the frequency counts of failures for each of the failure times, the frequencies of the data censored from one failure time to a later failure time, and the risk set of data points that have values at a common point in time.

Let's look at a Kaplan-Meier plot of the heart transplant data. The data is "on 69 patients receiving heart transplants" in the Survival data[13]. The survival status is 1 if a patient is dead and 0 if a patient is alive. The survival time is defined as time after the heart transplant (in days). The following Kaplan-Meier curve of survival time vs. status has confidence intervals around the curve, shown as dashed lines in Figure 5. This Kaplan-Meier models and plot in Figure 4 will be compared to the Flemington-Harrison model and plot to see if there is any difference between the two methods.

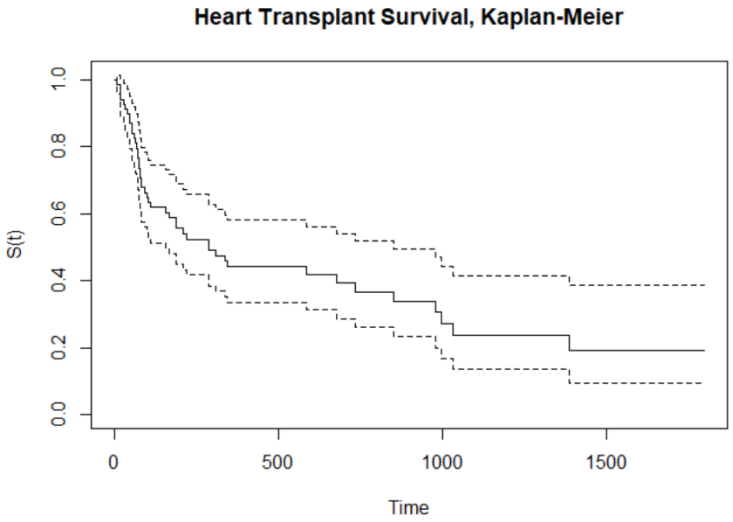


Figure 5: Kaplan-Meier curve of survival time vs. status of heart transplant data

For r_j (the number of individuals at risk at $t_{(j)}$), and d_j (the number of failures at $t_{(j)}$)[12] the general Kaplan-Meier estimator's formula is denoted as in equation 15.

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \times \hat{P}r(T > t_{(j)} | T \geq t_{(j)}) = \prod_{i=1}^{j} \hat{P}r(T > t_{(i)} | T \geq t_{(i)}) = \prod_{i=1}^{j} \frac{(r_j - d_j)}{r_j}, \text{ for } T \geq t_{(j)} \geq 0$$

(equation 15)

where,

$$\hat{S}(t_{(j-1)}) = \prod_{i=1}^{j-1} \hat{P}r(T > t_{(j)} | T \geq t_{(j)}) \text{ (equation 16)}$$

Equation 15 is a product limit formula[8]. The Kaplan-Meier estimator assumes that the censored data is not dependent on the survival time. Lastly, note that the definition of the Kaplan-Meier estimator also explains why the censored data is independent of the failure's cause[12].

- Nelson-Aalen Estimator and Flemington-Harrison

The Nelson-Aalen estimator and the Flemington-Harrison estimator are an alternative to the Kaplan-Meier estimator and plot. According to Stevenson, M.[12], "The Flemington-Harrington estimate of survival can be calculated using the Nelson-Aalen estimate of cumulative hazard using the relationship between survival and cumulative hazard." Recall that the hazard function is denoted as $H(t, X) = -\ln(S(t, X))$. The survivor function thus relates to the hazard function as $S(t, X) = e^{-H(t,X)}$. Now, the estimated hazard function as shown in equation 17

$$\hat{H}(t, X) = \sum_{j}^{t_j \leq t} \frac{d_j}{r_j} \text{ (equation 17)}$$

for the Flemington-Harrison estimator, with d_j and r_j defined as above.

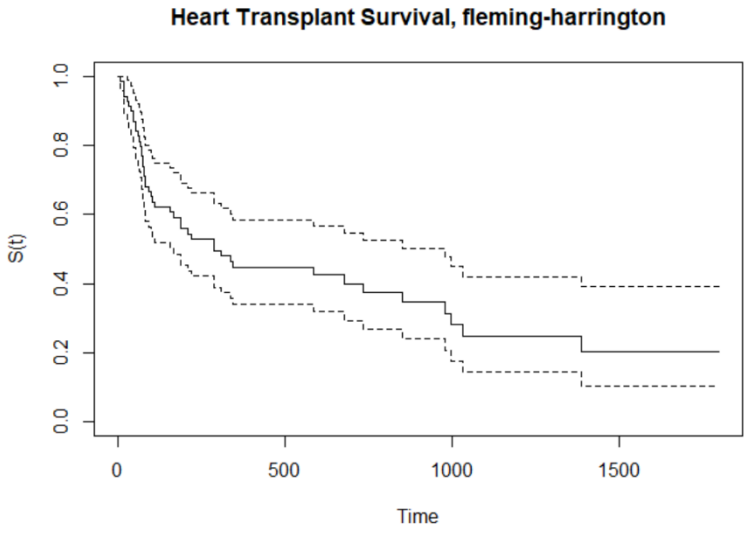


Figure 6: Heart Transplant Fleming-Harrington Survival Plot

With the heart transplant data set that was introduced above, the difference between the Kaplan-Meier and the Flemington-Harrington estimate of survival for the heart transplant patients is not obvious. The two plots in Figures 5 and 6 are very similar. A closer comparison of the two functions give us the results in Figure 6. Notice how, in Figure 7, the Kaplan-Meier (KM) survival estimates are in general a little smaller than Flemington-Harrington (FH) survival estimates. Therefore, the Flemington-Harrington survival estimates indicate that FM is slightly more accurate.

```
> tmp <- as.data.frame(cbind(km = hearttrans.km$surv, fh = hearttrans.fh$surv))
> head(tmp)
         km        fh
1 0.9855072 0.9856118
2 0.9565217 0.9570453
3 0.9420290 0.9426539
4 0.9275362 0.9282626
5 0.9130435 0.9138712
6 0.8985507 0.8994798
> tail(tmp)
          km        fh
59 0.2371029 0.2470240
60 0.1896823 0.2022461
61 0.1896823 0.2022461
62 0.1896823 0.2022461
63 0.1896823 0.2022461
64 0.1896823 0.2022461
```

Figure 7: Compare the Kaplan-Meier and Fleming-Harrison

- Parametric Models

What is a parametric model and how is it different from all the other types of models? Parametric models assume the normality of the survival data. When the survivorship of the subjects in the study follows a predictable pattern, parametric survival distributions should be used. According to Stevenson[12], "An advantage of using a parametric distribution is that we can reliably predict time to event well after the period during which events occurred for our observed data." The common parametric models that are used to describe the survival dataset

include the exponential model, Weibull model, the lognormal model, the log-logistic model, Gompertz model, and the gamma model.

- Exponential distribution model

The Exponential distribution is a very basic parametric model. The exponential hazard (instantaneous) function is shown in equation 17.

$$h(t, X) = \tau \text{ (equation 17)}$$

The constant $\tau > 0$ is equal to $e^{\beta X}$ where β is the vector of regression coefficients and X is the vector of predictors[8]. Both PH and AFT assumption accommodations can be evaluated with this model. Figure 7 shows generic plots of the instantaneous (in the left), the cumulative hazards (in the middle) and the survival distribution (on the right) of the exponential distribution model. Note that most situations do not have a constant instantaneous hazard rate, as shown in the first plot. The exponential distribution is a specific and special case of Weibull distribution.

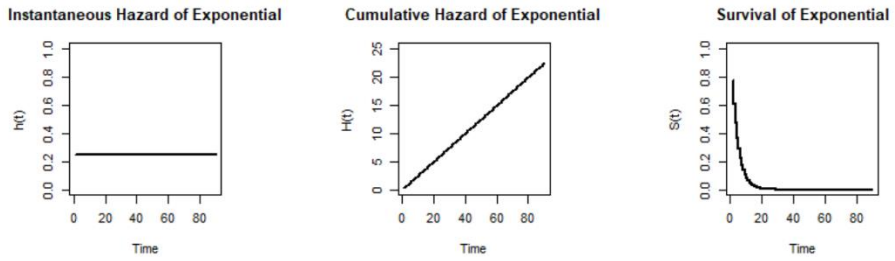


Figure 7: Exponential models

- Weibull distribution model

The Weibull distribution has its instantaneous hazard function denoted as in equation 18.

$$h(t, X) = c\tau t^{c-1} \text{ (equation 18)}$$

where $\mathbf{c}$ is a positive ($c > 0$) constant shape parameter. The positive constant τ is defined in the same way as in the exponential distribution model. Notice the different shapes of the instantaneous hazard plot (on the left), cumulative hazard plot (in the middle), and the survival distribution plot (on the right) for the Weibull models in Figure 8, where $\tau = 0.25$.

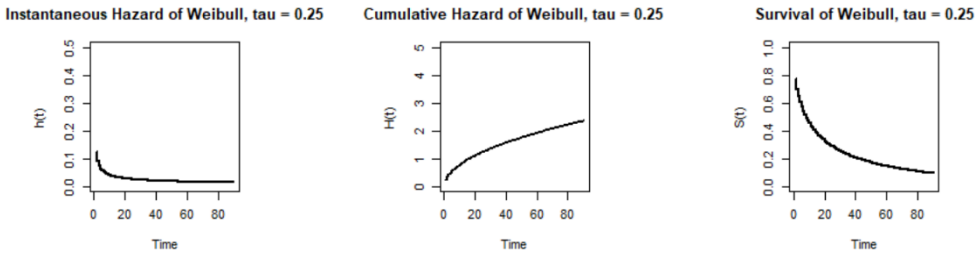


Figure 8: Weibull models for different values of τ

Furthermore, in Figure 9, when $\mathbf{c} > 1$, $\mathbf{h(t, X)}$ monotonically increases with time t (see the dotted and dotted-dashed lines, when c = 1.5 and when c = 3, respectively). When $\mathbf{c} < 1$, as in

Figure 9 when c = 0.5 (see the solid line), $\mathrm{h(t, X)}$ monotonically decreases with time t. Lastly, when c = 1, $\mathrm{h(t, X)} = \tau$, the exponential distribution hazard (see the dashed line).

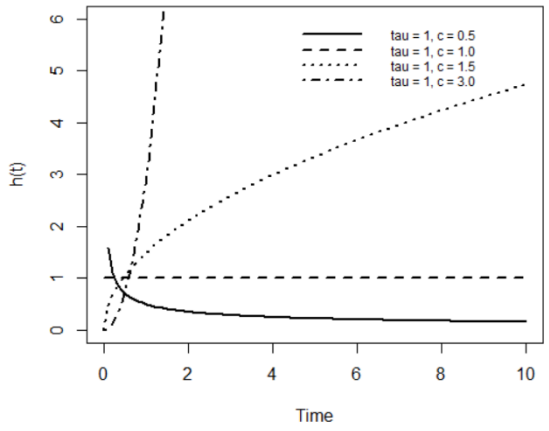


Figure 9: different values of c, for when τ is 1.

Both PH and AFT assumption accommodations can be evaluated for Weibull distribution model (and therefore for the exponential distribution model). The PH assumption will hold if the AFT assumption holds[8]. The Weibull distribution is the only distribution that uses both the AFT and PH models to evaluate the assumptions discussed above, for constant covariates[7]. The survival function of the Weibull is $S(t) = e^{-\tau t^c}$.[8] The PH form of Weibull is $\tau = e^{\beta_0 + \beta_1 * (some\ predictive\ categorical\ treament\ variable)}$ and the AFT model of Weibull is $\frac{1}{\tau^{1/c}} = e^{\alpha_0 + \alpha_1 * (some\ predictive\ categorical\ treament\ variable)}$. When $c = 1$, these PH and AFT models are the exponential distribution's PH and AFT models, respectively.

- Log-normal distribution model

The log-normal distribution has the instantaneous hazard function denoted in equation 19, where $\varphi(.)$ is distributed as N($\mu = 0, \sigma = 1$) and $\mu = \beta^* X$[7].

$$h(t, X) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-[\ln(t)-\mu]^2}{2\sigma^2}}}{1-\varphi(\frac{\ln(t)-\mu}{\sigma})} \text{ (equation 19)}$$

If parameter $\sigma > 1$, $\mathrm{h(t, X)}$ monotonically decreases with time t, as shown in Figure 10 with $\sigma = 2$. If $0 < \sigma \leq 1$, $\mathrm{h(t, X)}$ monotonically increases with time t, as shown in Figure 10 with $\sigma = 0.5$ and $\sigma = 1$.[7] Also note that the log-normal model is an AFT model, meaning the AFT assumption needs to be met.[8]

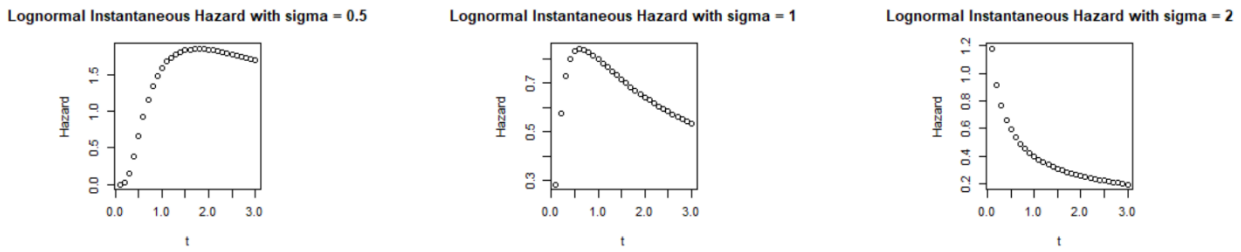


Figure 10, σ varies and μ is constant at 0.

- Log-Logistic distribution model

The Log-Logistic distribution model has the instantaneous hazard function in equation 20, where $\theta = e^{-\beta^* X}$, $g = \theta^l = (e^{-\beta^* X})^l, l = \frac{1}{k}$, and $k > 0$, is the shape parameter of $h(t, X)$[7]

$$h(t,X) = \frac{\theta^{\frac{1}{k}} t^{(\frac{1}{k}-1)}}{k(1+(\theta t)^{\frac{1}{k}})} = \frac{l\theta^l t^{(l-1)}}{(1+(\theta t)^l)} = \frac{lgt^{(l-1)}}{(1+gt^l)} \text{ (equation 20)}$$

The Log-logistic instantaneous hazards are shown in Figure 11, for different values of l. If $l \geq 1$, $\mathrm{h(t, X)}$ monotonically decreases with time t, as shown in Figure 11 when $l = 2$. If $0 < l < 1$, $\mathrm{h(t, X)}$ monotonically increases with time t, as shown in Figure 11 when $l = 0.5$.

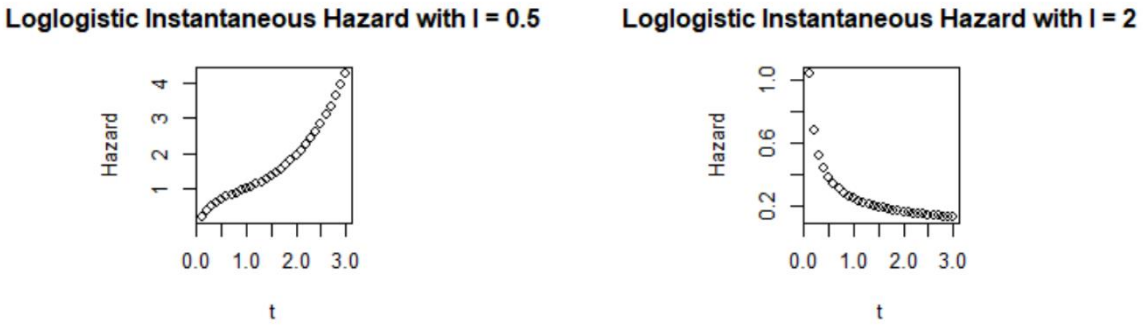


Figure 11, l varies and the center is constant at 0.

Log-logistic is not a PH model, but rather an AFT model; The AFT assumption must be met.[8] One way to check if the AFT assumption for the log-logistic model is met is to also evaluate the proportional odds (PO) assumption. If the PO assumption is met, then the AFT assumption will be met. Similarly, for the PH model with the hazard model, the PO model will have an odds ratio (OR) that remains constant over time[8].

The plot of the survival odds vs. the log of time (ln(t)) needs to be found. The log-logistic survival function is $S(t) = \frac{1}{1+gt^l}$, so the failure function of the logistic function is $1 - S(t) = \frac{gt^l}{1+gt^l}$. The survival odds is defined as by "the probability of not experiencing the event by the time t, divided by the probability of getting the event by time t" = $\frac{P(T \leq t)}{P(T > t)}$ and is calculated as in equation 21.

$$\frac{P(T\leq t)}{P(T>t)} = \frac{S(t)}{1-S(t)} = \frac{\frac{1}{1+gt^l}}{\frac{gt^l}{1+gt^l}} = \frac{1}{gt^l} \text{ (equation 21)}$$

If the log of the survival odds, $\log(survival\ odds) = \ln\left(\frac{S(t)}{1-S(t)}\right) = \ln\left(\frac{1}{gt^l}\right) = \ln(g^{-1} * t^{-l}) = -\ln(g) - l * \ln(t) = -(\ln(g) + l * \ln(t)) = intercept(= -\ln(g)) + slope\ (= -l)$, is plotted vs. $\ln(t)$, and see that the plot is linear with the slope of the lines equal to $-l$, then the loglogistic assumption is met.[8] Alternatively, if the plot of $\ln(t)$ vs.

$$\log(failure\ odds) = \ln\left(\frac{1-S(t)}{S(t)}\right) = \ln(gt^l) = \ln(g) + l * \ln(t) = intercept + slope$$

is linear with slope equal to l, then the log-logistic assumption is met.

If the plot of the $\log(survival\ odds)$ vs. $\ln(t)$ is split into lines by a categorical variable, and the different lines corresponding to the different categories are parallel, then the PO assumption is met.[8] So, if the groups have straight lines (which supports the log-logistic assumption) and parallel survival curves (which supports the PO assumption) then the AFT assumption will hold as well. To find the AFT model for the log-logistic distribution model, take the survival function, $S(t) = \frac{1}{1+gt^l} = \frac{1}{1+(g^{1/l}t)^l}$, and solve for t, as shown in equation 22[7]

$$S(t) = \frac{1}{1+(g^{1/l}t)^l} \rightarrow 1 + (g^{1/l}t)^l = \frac{1}{S(t)} \rightarrow t = \left[\frac{1}{S(t)} - 1\right]^{1/l} * \frac{1}{g^{\frac{1}{l}}} \text{ (equation 22)}$$

We have that the AFT form of the log-logistic model is $\frac{1}{g^{\frac{1}{l}}} = \left(\frac{1}{g}\right)^{\frac{1}{l}}$. This can be reparametrized as $e^{\alpha_0+\alpha_1*(some\ predictive\ categorical\ treament\ variable)}$ to allow a predictor variable, such as a categorical treatment variable, "to be used for the multiplicative scaling of time to any fixed value of $S(t)$", according to Kleinbaum[8]. Then, the acceleration constant factor c can be calculated by taking the ratio of times to some fixed probability of $S(t)$, say w, for the different values or categories of the (categorical) predictive treatment variable. Let $\left(\frac{1}{g}\right)^{\frac{1}{l}} = e^{\alpha_0+\alpha_1*(some\ predictive\ categorical\ treament\ variable)}$ be substituted in the equation for time t above, and letting the predictive categorical variable be binary, the proportion to get is expressed as: $c = \frac{t_2}{t_1} = \frac{\left[\frac{1}{w}-1\right]^{1/l} * e^{\alpha_0+\alpha_1*(some\ predictive\ categorical\ treament\ variable\ =\ 2)}}{\left[\frac{1}{w}-1\right]^{1/l} * e^{\alpha_0+\alpha_1*(some\ predictive\ categorical\ treament\ variable\ =\ 1)}} = \frac{\left[\frac{1}{w}-1\right]^{1/l} * e^{\alpha_0+\alpha_1*(2)}}{\left[\frac{1}{w}-1\right]^{1/l} * e^{\alpha_0+\alpha_1*(1)}} = e^{\alpha_1}$. Note that the constant c can relate to the proportional odds.[8]

According to Kleinbaum[8], we can also find that "the proportional odds form of the log-logistic model can also be formulated by reparametrizing" g, in a similar way as AFT and using the failure odds model (gt^l). If

$$g = e^{\beta_0+\beta_1*(some\ predictive\ categorical\ treament\ variable)},$$

the OR is therefore denoted as

$$OR(some\ predictive\ categorical\ treament\ variable = 2\ vs\ 1) = \frac{t^{l}{}_{*}e^{\beta_0+\beta_1*(2)}}{t^{l}{}_{*}e^{\beta_0+\beta_1*(1)}} = e^{\beta_1}.$$

The relationship between the AFT and PO for the log-logistic model is hence through the coefficients[8]: $\beta_i = -\alpha_i * l$.

- Gompertz distribution model

The Gompertz distribution model has the instantaneous hazard function as shown in equation 23, where ω is the shape parameter of $\mathrm{h(t, X)}$ and can be positive, 0 or negative[7].

$$h(t, X) = \tau e^{\omega t} \text{ (equation 23)}$$

If $\omega > 0$, $\mathrm{h(t, X)}$ monotonically (or exponentially) increases with time t. If $\omega < 0$, $\mathrm{h(t, X)}$ monotonically decreases (exponentially) with time t. When $\omega = 0$, the hazard is constant (it's an exponential model). Gompertz is an AFT model; the AFT assumption must be met. It's a multiplicative scaling of failure time, or a multiplicative model[8]. In the log form, it's an additive failure model.

- Gamma distribution model

The instantaneous hazard model for the gamma distribution is complicated in form. It is expressed in integral form, so it will not be shown. There are two parameters that are used in shape and scaling, respectively: γ and ϑ. According to Jenkins[7], "the hazard function is quite flexible in shape, even including the possibility of a U shaped or so-called 'bath-tub' shaped hazard (commonly cited as a plausible description for the hazard of human mortality looking at the lifetime as a whole)." So, the gamma distribution is a very general distribution, and specific cases of it produce many of the distributions that have been described above. In the case where shape $\gamma = \vartheta$, the scaling parameter, we have the gamma distribution. When the shape $\gamma = 1$, the special case of the gamma distribution model is Weibull. When shape $\gamma = 1$ and scaling $\vartheta = 1$, the model is considered exponential distribution. When shape $\gamma = 0$ the model is log-normal.

- Semi Parametric Model

What is a semi-parametric model and how is it different from a parametric model? Semi-parametric models differ from parametric models in that what is called the baseline hazard (as defined below) is allowed to vary with time[12]. A common semi-parametric model is the Cox proportional hazards (Cox PH) model. According to Stevenson, M.[12], "With the Cox proportional hazards model, the outcome is described in terms of the hazard ratio," unlike the parametric and non-parametric models.

- Cox-Proportional Hazard Model

The Cox PH model is a semi-parametric model that is the most common model used in the health sciences and thus the most popular survival analysis modeling technique[8]. The

instantaneous hazard function is denoted as in equation 24, with β_i being the vector of p regression coefficients and X_i denoted as the p predictors.

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \text{ (equation 24)}$$

The $h_0(t)$ is called the baseline hazard and can vary over time. The hazard is time dependent and $e^{\sum_{i=1}^{p} \beta_i X_i}$ involves X_i's that are not time (t) dependent. In addition, the "Cox PH model is 'robust': it will closely approximate the correct parametric model," according to Kleinbuim, 96). Thus, it is perhaps one of the most important models in survival analysis modeling. A picture of a Cox PH model graphed is shown below in the full data analysis section.

There are three assumptions of the Cox PH model. First, as briefly mentioned above, "the ratio of the hazard function for the two individuals with different sets of covariates does not depend on time" (Stevenson 17). The second assumption is that the time is measured continuously. Lastly, the censored data is assumed to be random.

Taking the hazard function of the Cox PH model, we can estimate the β_i parameter with the maximum likelihood (ML) estimates and denote these estimates as $\widehat{\beta}_i$. These ML estimates will maximize the likelihood function L, which is the joint density of the observations, also denoted by L(β). L(β) considers the probabilities of the subjects that fail and do not include the subjects who are censored, making this a partial likelihood[8].The Cox likelihood function is "based on the observed order of events rather than the joint distribution of events", according to Kleinbaum[8], and is a distribution based only on the outcomes. The Cox PH model is therefore a partial likelihood model.

For example, three people, Ed Apple, Eddie Banana, and Edward Carrot, are each given lottery tickets at three different times. According to Kleinbaum[8], "winning tickets are chosen at times t_j (j = 1, 2, . . .). Assume each person is ultimately chosen and once a person is chosen, he cannot be chosen again (i.e., he is out of the risk set)." Suppose the probability of the order that each person is chosen to get a lottery ticket is in in the order Ed. Apple, Eddie Banana, and then Edward Carrot. This probability is found simply by multiplying the individual probabilities. The probability that Ed. Apple's ticket is chosen prior to Eddie Banana and Edward Carrot is one out of three. Then, Eddie Banana's probability of having his lottery ticket chosen before Edward Carrot is one-half. Lastly, Edward Carrot probability of being chosen last is one. So, the probability is therefore equal to $\frac{1}{3} * \frac{1}{2} * \frac{1}{1} = \frac{1}{6}$. Now, suppose that Ed. Apple gets 5 tickets, Eddie Banana gets 3 tickets, and Edward Carrot gets 4 tickets (a total of 12 tickets distributed), then the probability of the order that each person is chosen to get their lottery tickets is in in the order Ed. Apple, Eddie Banana, and then Edward Carrot is equal to $\frac{5}{12} * \frac{3}{12-5=7} * \frac{4}{7-3=4} = \frac{5*4*3}{12*7*4} = \frac{60}{336}$.

The general approach to finding the likelihood L can be stated in a concise way. If there are k failure times, then the likelihood $L = L_1 * L_2 * \ldots * L_k = \prod_{j=1}^{k} L_j$, or the product of the k likelihood terms. Each of the k L_j's is calculated by using the hazard functions. According to Kleinbaum[8], "L_j = portion of L for the jth failure time given the risk set R(t(j))." Also, the individual likelihoods are calculated by having "the denominator for the term corresponding to

time t_j (j = 1, 2, 3) as the sum of the hazards for those subjects still at risk at time t_j, and the numerator is the hazard for the subject who got the event at t_j" according to Kleinbaum[8]. To obtain the maximum likelihood estimates, take the partial derivative of L and set it to zero: $\frac{\partial lnL}{\partial \beta_i} = 0$, for i = 1, 2, 3, …, p, where we are looking at the number of parameters p.[8] An example of obtaining a maximum likelihood estimate is discussed next.

Perhaps, the data in table in Figure 12 below is the given survival data set for the event of lung cancer for the three men from the example above: Ed. Apple, Eddie Banana, and Edward Carrot.

Subject	Survival Time (in years)	Status (1 for event, 0 for censored)	Smoke? (1 for yes, 0 for no smoking)
Ed. Apple	3	1	1
Eddie Banana	4	1	0
Edward Carrot	6	0	0

Figure 12: survival data set for the event of lung cancer for the three men

Let's say that a Cox PH model is fitted to this data with one covariate, smoke status, $h(t, X) = h_0(t)e^{\beta_1 * Smoke}$ (Kleinbuim, 113), (see the Figure 13 below).

Subject	Hazard
Ed. Apple	$h_0(t)e^{\beta_1}$
Eddie Banana	$h_0(t)e^0$
Edward Carrot	$h_0(t)e^0$

Figure 13: Cox PH model is fitted to this data

Then, the Cox likelihood is the product of the two individuals likelihoods (Ed. Apple and Eddie Banana) who experienced the event of lung cancer: $L = L_1 * L_2 = \left(\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0}\right) * \left(\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0}\right) = \left(\frac{e^{\beta_1}}{e^{\beta_1} + 2 * 1}\right) * \left(\frac{1}{1+1}\right) = \left(\frac{e^{\beta_1}}{e^{\beta_1}+2}\right) * \left(\frac{1}{2}\right) = \left(\frac{e^{\beta_1}}{2*(e^{\beta_1}+2)}\right)$. Notice how, the L_2 has the sum of the two hazards for the subject (Edward Carrot) still at risk in the denominator and Eddie Banana's hazard in the numerator. This is what should be done if there were more subjects who experienced the event. The censored events should also should never have their own individual likelihood term in the multiplication. Notice how, because Edward Carrot was censored, his Likelihood is not accounted for in the likelihood calculation. Also, The Final Likelihood is not dependent on the baseline hazard $h_0(t)$, as stated above, because it cancels out.

- Comparing Methods of Survival Estimation
  - Log-rank test

The log-rank test is used to compare two groups with two different Kaplan-Meier curves, as defined above. It is a test that lets us see if the KM curves are statistically equivalent. The null hypothesis is that there is no difference between the survival curves split by the categories. The alternative hypothesis is that there is a difference in the survival curves. The test statistic for the log-rank test is a chi-square test statistic for a large sample sized dataset. This test statistic uses a

criterion that compares the KM curves and the observed vs. expected cell counts over the categories of outcomes.[8] Such categories to create the log-rank chi-square test statistic are found by using the information of the ordered failure times in the dataset.

- Other tests

Some other tests besides the log-rank test, are Breslow's test, the Cox-Mantel test, and the Peto and Peto modification of Gehan-Wilcoxon test. When the Breslow's test has little censoring and the hazard functions are not parallel, it is more powerful than the log-rank test. The Peto and Peto modification of Gehan-Wilcoxon test is used when "the hazard ratio between groups is not constant" according to Stevenson, M.[12]. Furthermore, "It gives more weight to early failures" and it is similar to Breslow's test[12]. Lastly, the Cox-Mantel test is best used when there is progressive censoring apparent in the data.

o Methods of modeling the data

The methods of modeling data are discussed here. These include model selection and interpretation methods.

- Model selection and interpretation

Next, model selection methods will be discussed before doing a full analysis on a dataset. Such methods are covariate adjustment, categorical and continuous covariate, hypothesis testing for nested models, the Akaike Information Criterion (AIC) for comparing non-nested models and including smooth estimates of continuous covariates in a survival model. Interpretations of these method's results will be discussed.

The data used throughout this section to demonstrate the methods discussed is the brain cancer data. Data was collected on a sample of 30 patients who participated in a "randomized study of radiotherapy with and without a new radiosensitizer (misonidazole)," according to the Survival data[13]. Group is equal to treatment, where group 1 is radiosensitizer, age is age in years at the time of diagnosis, status is censorship status, and survival time is in days.

- Covariance adjustment.

It might be the case where the covariates in the model are included to understand how they affect the survival and to show if there any differences in the models with and without the covariates. According to Moore, D.[9], "If the study is based on observational data, and if there is a primary intervention of interest, then adjustment for potential confounders is essential to obtaining a valid estimate of the intervention effect." It might also be interesting to study other covariates and how they affect survival.

There is a strategy to sort through several potential explanatory variables in the survival model to find the explanatory variables that are significant, regardless of whether the study is observational or experimental. A common strategy to remedy this is using and analyzing the log hazard ratio. If the log hazard ratio treatment effect in the survival model is positive, then "higher hazards are associated with the treatment than with the control," according to Moore,

D.[9] That is, if we have a Cox PH model, with the treatment type (i.e. experimental treatment vs. control) for example, and the coefficient of the treatment is positive, then the "treatment appears to reduce the survival," according to Moore, D.[9], which is likely not a good situation in many contexts. On the contrary, if the coefficient is negative, then lower hazards are associated with the treatment than with the control.

Let's demonstrate the covariance adjustment by looking at a Cox PH model with the brain cancer data. This model will first look at the effect of age on survival unadjusted for the brain cancer of the patients in Figure 14. We see that the estimate of the log hazard ratio age effect is 0.14259. Higher hazards are associated with the treatment than with the control because this estimate is positive. So, therefore, the age appears to reduce survival. This result is unfortunate for the cancer patients who are older. The value of the exponentiated log hazard ratio is also given as 1.15325, suggesting that age is associated with a 15.325% additional risk of death for each year a patient gets older.

```
> coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age)
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$age)

                    coef exp(coef) se(coef)     z        p
brain_cancer$age 0.14259   1.15325  0.03604 3.956 7.63e-05

Likelihood ratio test=21.66  on 1 df, p=3.257e-06
n= 30, number of events= 22
```

Figure 14: Cox PH model with the brain cancer data

Now, the combination of the effect of the age stratified by treatment group on the survival is expressed as another Cox PH model in Figure 15. The age coefficient is still positive. This indicates that, within each treatment group, the treatment is not effective. We can show this by looking at the effect of the two covariates age and treatment group on the survival in a Cox PH model below in Figure 16. It is seen here that the treatment group is not significant, which supports the statement above. The age still effects the chance of survival of the brain cancer patients.

```
> coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age + strata(brain_cancer$group))
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$age + strata(brain_cancer$group))

                   coef exp(coef) se(coef)     z        p
brain_cancer$age 0.14735   1.15876  0.03833 3.844 0.000121

Likelihood ratio test=21.23  on 1 df, p=4.074e-06
n= 30, number of events= 22
```

Figure 15: another Cox PH model

```
> coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age + brain_cancer$group)
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$age + brain_cancer$group)

                     coef exp(coef) se(coef)     z        p
brain_cancer$age   0.14960   1.16136  0.03642 4.107 4.01e-05
brain_cancer$group 0.55271   1.73795  0.44754 1.235    0.217

Likelihood ratio test=23.2  on 2 df, p=9.176e-06
n= 30, number of events= 22
```

Figure 16: another Cox PH model of two covariates age and treatment group

- Categorical and continuous covariates

If a covariate is categorical, it can be represented as a dummy or indicator variable, which "takes on values 0 or 1 depending on which of the two groups a subject belongs to" according to Moore, D.[9] Any categorical variables that have two categories can be coded with just zeros and ones. Categorical variable with three or more categories will need to have two or more dummy variables. One less than the number of categories is the number of dummy variables needed for such situations. Depending on the context of the research question in the study that has such data, this can be arbitrary or can be useful to the context of the question. If they are comparing different treatments, and there are three different categories such as control, treatment 1, treatment 2, then control would likely be a good dummy variable.

A variety of patient data is wanted in the survival model, such as group assignment (i.e. experimental treatment vs. control) in a randomized clinical trial, clinical variables such as blood measurements, BMI, or disease stage indictors, and demographic info such as patient age, gender, or income. Once the set of such k covariates has been chosen, the PH model is fit on the data, with each covariate represented as x_j with β_j $(j = 1, \dots, k)$ as the log hazard ratio for the effect of that covariate parameter on the survival time y_i in equation 25.[9]

$$\log(y_i) = [x_{1i} \quad \dots \quad x_{ki}] \begin{bmatrix} \beta_1 \\ \dots \\ \beta_k \end{bmatrix} = x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k = \mathrm{X}^T\beta \text{ (equation 25)}$$

This equation 25 is true for any PH model.

A log, square root, or other numerical transformations can be completed on any continuous covariates or add in interaction terms to enhance the survival model in a way that will make the covariates predict the survival better, much like linear regression or logistic regression models.

There are certain characteristics of survival models that make them different from linear regression or logistic regression models, such as how covariates can change with time in the survival models. However, at the beginning of model fitting, it is required that the covariates are all fixed and that they cannot change over time[9]. Also, with a proportional hazards model, such as the Cox PH model, there is no intercept term, as it cancels out in the partial likelihood just as the baseline hazard canceled out in the numerator and denominator of the partial likelihood.

Taking into consideration an example using the brain cancer data introduced above, say the covariates treatment group and age are being taken into consideration. Because treatment group has only two groups, it can be converted into a factor class in R with the two factors for the two groups. Making the reference category for group to be zero, the conventional treatment or control group, the following matrix of dummy variables was computed in Figure 17.

```
> group <- relevel(group, ref="1")
> model.matrix(~ group + age)[,-1]
   group0 age
1       1  48
2       0  63
3       0  54
4       1  49
5       0  44
6       1  36
7       0  29
8       1  50
9       1  53
10      1  58
```

Figure 17: matrix of dummy variables for brain cancer data

The first ten observations out of the thirty are shown in Figure 17. Here, just the control is shown for the group, as the dummy variable, with one indicating that the observation belongs to the control group and zero indicating that the observation belongs to the radiosensitizer group. Age is also shown. The intercept column was not shown because there is no intercept term in survival analysis models. The interaction between the group and age can be shown as an added column to the Figure 18. These interaction values in Figure 18 are clearly the product of the group zero status and the age.

```
> model.matrix(~ group + age + group:age)[,-1]
   group0 age group0:age
1       1  48         48
2       0  63          0
3       0  54          0
4       1  49         49
5       0  44          0
6       1  36         36
7       0  29          0
8       1  50         50
9       1  53         53
10      1  58         58
```

Figure 18: matrix of dummy variables for brain cancer data with interaction

- Hypothesis testing for comparing nested models

When comparing two models with each other, the covariates of one model must be a subset (a reduced model) of the covariates of the other (full model), making them nested models that are compared with hypothesis tests[9], such as the *partial likelihood test* or by using the *Wald test*. To compare a full model to a reduced model with the use of such as hypothesis test as the partial likelihood test, the null hypothesis must be defined as the coefficient(s) of a covariate are all zero vs. the alternative that is defined as the covariates are not all zero[9]. The test statistic is a likelihood ratio test statistic in equation 26.

$$2\left(l(\hat{\beta}_{full}) - l(\hat{\beta}_{reduced})\right) \text{ (equation 26)}$$

This equation is "twice the difference between the partial log-likelihood evaluated at the 'full' model and the value at the 'reduced' model," and it is compared to a chi-square distraction that "is the difference in degrees of freedom between the two models" according to Moore, D.[9].

With regards to the brain cancer data, suppose Model A, or a version of a nested model, has only the covariate age in the Cox PH survival model. Model B, another version of a nested model, has just the treatment group variable as the covariate. Now, Model C will have both age and group as the covariates in the survival model. All three of these models will be compare with each other.

The question to ask here is, will the results be the same as those in the covariance adjustment analysis above? That is, will the group still show up to be an insignificant covariate? The results of the test are in Figure 19 below. To see if group belongs in the survival model, we can compare Models A and C. Here, the null hypothesis is that the coefficient for group is zero, whereas the alternative is that the treatment group coefficient is not zero. The likelihood ratio test statistic for this test is equal to

$2\left(l(\hat{\beta}_{full}) - l(\hat{\beta}_{reduced})\right) = 2(-49.71536 - (-50.48475)) = 1.53878$. We compare this to a df = 2-1 = 1 (the difference in the degrees of freedom of models A and C) chi-square distribution. The p-value of the comparison between the test statistic and the chi-square value is 0.2148007, as shown in Figure 20. Therefore, the null hypothesis that the coefficient for group is zero does not get rejected, and it can probably be said that the effect of treatment group is not statistically significant when age is included in the survival model.

```
> modelA.coxph
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$age)

                    coef exp(coef) se(coef)     z        p
brain_cancer$age 0.14259   1.15325  0.03604 3.956 7.63e-05

Likelihood ratio test=21.66  on 1 df, p=3.257e-06
n= 30, number of events= 22
> modelB.coxph
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$group)

                       coef exp(coef) se(coef)     z     p
brain_cancer$group 0.009425  1.009470 0.434224 0.022 0.983

Likelihood ratio test=0  on 1 df, p=0.9827
n= 30, number of events= 22
> modelC.coxph
Call:
coxph(formula = Surv(brain_cancer$"survival time", brain_cancer$status) ~
    brain_cancer$age + brain_cancer$group)

                     coef exp(coef) se(coef)     z        p
brain_cancer$age   0.14960   1.16136  0.03642 4.107 4.01e-05
brain_cancer$group 0.55271   1.73795  0.44754 1.235    0.217

Likelihood ratio test=23.2  on 2 df, p=9.176e-06
n= 30, number of events= 22
```

```
> logLik(modelA.coxph)
'log Lik.' -50.48475 (df=1)
> logLik(modelB.coxph)
'log Lik.' -61.3141 (df=1)
> logLik(modelC.coxph)
'log Lik.' -49.71536 (df=2)
```

Figure 19: covariance adjustment Cox PH analysis

```
> pchisq(likelihoodratioteststat, df = 1, lower.tail = F)
'log Lik.' 0.2148007 (df=2)
```

Figure 20: p-value of the comparison between the test statistic and the chi-square value

This likelihood ratio test can be completed again for Model A's and Model C's significance using this same method. The results all do line up with what was concluded about the age, group, and inclusion of both covariates in the survival models in the covariate adjustment section above. That is, age is statistically significant even when group is included as a covariate in the model, and age should be alone in the model. This occurs without the group covariate, as that provides the best fitted survival model on the brain cancer data.

- The Akaike Information Criterion for Comparing Non-nested models

All too often, there are many possible significant factors to include in the survival analysis model. These significant covariates from the list of the covariates need to be found and selected to be in the final survival analysis model. A good way to do this is to find the Akaike Information Criterion (AIC), computed with the partial log-likelihood at the M.P.L.E ($l(\hat{\beta})$) and with the k parameters in the model, as $AIC = -2l(\hat{\beta}) + 2k =$ goodness of fit + number of parameters[9]. Accordingly, "a 'good' model is one that fits the data well (small value of $-2l(\hat{\beta})$) with few parameters (2k), so that smaller values of AIC should in theory indicate better models," according to Moore, D.[9]. To choose the best model based of AIC, select the model with the smallest value of this measure.

The Bayesian Information Criterion (BIC) is given by $BIC = -2l(\hat{\beta}) + klog(n)$. The BIC can be used as an alternative selection measure to AIC[9]. Mainly, the difference between the AIC and BIC measures is that the penalty for the number of parameters k is k*log(n) in the BIC measure rather than 2*k in the AIC measure. Therefore, the BIC might result in a model with fewer parameters than AIC.

To practically find and select models based of a long list of possible covariates and predictors of survival, the stepwise procedure can be used with the AIC measure as the criterion guiding the selection of the covariates and survival models[9]. Different model's log hazard ratios can be plotted on a forestplot to help aid the comparison and selection of the bet model, by selecting the covariates with the larger log hazard ratio.

Looking at the three Cox PH survival models above for the brain cancer data, models A, B, and C, the AIC scores can be computed, and is in Figure 21 below. The model with the smallest AIC value is Model A, the model with only age as the covariate. Thus, this model is the model with the best fitted one.

```
> AIC(modelA.coxph)
[1] 102.9695
> AIC(modelB.coxph)
[1] 124.6282
> AIC(modelC.coxph)
[1] 103.4307
```

Figure 21: AIC scores for the Cox PH models of the Brain Cancer Data

- Including smooth estimates of continuous covariates in a survival model

If there is a continuous covariate, it may not be related to the log-hazard in the survival as linear, but perhaps in a logarithmic, quadratic, or other nonlinear way. A way to model a nonlinear relationship is to use what are called smoothing splines, which are pieces of polynomial functions that are sewn together at the knots to create a smooth curve. According to Moore, D.[9] , "In survival analysis, an effective method of finding a smoothing spline is via "penalized partial likelihood." The partial log likelihood and a penalty term are to be optimized. Such a likelihood is increased by complex models that have a lot of knots in them due to the improved fit of the survival model. But, with the increase in the number of knots, the penalty term (based on the second derivative of the model) is decreased. Thus, the penalized partial likelihood is balanced with a larger (positive) and smaller (negative) term.

- Model assessment and diagnostics

Now, two different model assessment and diagnostics topics will be discussed. These are assessing goodness of fit using residuals and checking the PH assumption.

- Assessing goodness of fit (GOF) using residuals

When plotting the residuals vs. some other quantity, the covariate value or some other quantity, statisticians can use the observed patterns to diagnose possible problems with the fitted model. According to Moore, D.[9] ,

> "the pattern of the plotted residuals may suggest an alternative model that fits the data better… The survival data evolves over time, and requires special assumptions such as proportional hazards, makes it necessary to develop additional diagnostic residual methods."

Such residual analyses are Martingale and Deviance residuals and case deletion residuals for assessing the goodness of fit of a survival model on the data.

- Martingale and Deviance residuals

Martingale and Deviance residuals work by comparing the binary censoring indicator to the expected value of that indicator under the Cox PH model for each of the observations in the survival dataset. According to Stevenson, M.[12], Martingale residuals are defined as the "difference between the observed number of events for an individual and the conditionally expected number given the fitted model, follow up time, and the observed course of any time-varying covariates." (See the equation below for more details.) These residuals can be plotted against the covariates to check the scale of continuous covariates with no transformations, such as to determine if there is a linear relationship between the covariate values.

According to Moore, D.[9], "if there are no time-dependent covariates and if the survival times are right-censored, [the Martingale residual equation] is given by

$$m_i = \delta_i - \widehat{H_0}(t_i)e^{z_i'\widehat{\beta}}\ "\ (87)\ \text{(equation 27)}$$

In other words, such residuals give the difference between the observed binary censoring value (δ_i) and the "expected value under a particular Cox model" ($\widehat{H_0}(t_i)e^{z_i'\widehat{\beta}}$), as stated by Moore, D.[9]. Martingale and Deviance residuals will add up to zero and will be a very small negative number up to a maximum value of zero. There is an example of Martingale interpretations below.

Because the sum of squares of the Martingale residuals, they should not be used to measure GOF of the survival model on the survival data, the alternative is to use the deviance residual. A shown in equation 28, according to Moore, D.[9], the deviance residual is "defined in terms of the martingale residual

$$d_i = sign(m_i) * \{-2 * [m_i + \delta_i \log(\delta_i - m_i)]\}^{1/2}\text{"}\ \text{(equation 28)}$$

It can be used instead to measure goodness of fit. The sum of the squares of these residuals are of the likelihood ratio test. If these residuals are equally and symmetrically distributed around the value of zero, then this indicates that the survival model is a correct one. In this case, the expected value of the residuals is equal to zero.The martingale residuals vs. age since diagnosis of brain cancer are shown here, as shown in Figure 22. We see that the residuals on the y-axis are evenly and symmetrically around the zero marker. Thus, there is likely no departure from linearity.

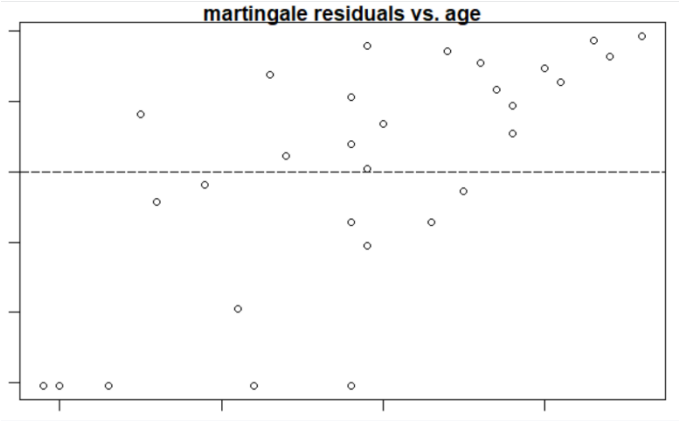


Figure 22: martingale residuals vs. age since diagnosis of brain cancer

- case deletion residuals

When observations in the survival dataset have a very large influence on the value of the parameter estimates on a survival model, we can use the case deletion residuals (a.k.a "jackknife" residuals) to determine and provide a remedy for these influential observations. For each subject, a case deletion residual is the difference in the value of the coefficient using all the data and its value when that subject is deleted from the data set" (Moore, 92). The case-deletion residuals for the Cox PH model on the brain cancer data with age as the covariate is shown below in Figure 23. Such jackknife residuals were used.

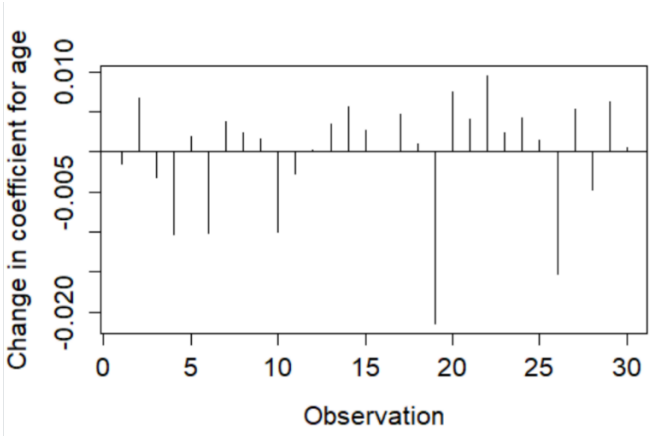


Figure 23: The case-deletion residuals for the Cox PH model on the brain cancer data with age

In Figure 23 above, it is apparent that no single patient changes the estimate of the age coefficient in the Cox PH model by more than 0.01. This estimate is less than 8% of the value of

the coefficient in the Cox PH model (that is, 0.1425855). However, we see that observation 19 has the most influence over the coefficient estimate for age at the time of diagnosis for the patient. Perhaps observation 19 is worth further looking into.

- Checking the PH assumption

We must be able to construct the partial likelihood and be able to cancel out the partial likelihood factors' baseline hazard functions. We must verify the proportional hazards assumption in order to do this. What the proportional hazards assumption tells us is that "the proportional hazards functions are proportional, and hence that the log-hazards are separated by a constant at all time points. Similarly, a categorical variable with many levels will result in parallel log hazard functions," according to Moore, D.[9]. This is only if there is some binary predictor variable (treatment type, sex, etc.) given.

Minor violations of the PH assumption are not likely to affect the inferences of the survival model parameters, as the PH assumption is typically an approximation and hypothesis tests that involve the PH assumption are not typically used or needed. Still, it is important to know the typical assessment methods of this PH assumption (log cumulative hazard plots and Schoenfeld Residuals which are discussed in further detail below) as the assumption is still useful to assess for the survival model.

- Log cumulative hazard plots

According to Stevenson, M.[12], a method of assessing if a parametric model appropriately describes survivorship of a dataset are through the analysis of log cumulative hazard plots. Log cumulative hazard plots are plots "that can help us assess the proportional hazards assumption" between two groups, according to Moore, D.[9]. These are plots of "a function of time (to check for consistency with the exponential distribution) and log cumulative hazard as a function of log time (to check for consistency with the Weibull distribution)", as is said by Stevenson, M.[12]. There is an example of an interpretation of these plots below, for the brain cancer data Cox PH model with the covariate treatment group. The plot of the log cumulative hazards for the Brain cancer treatment group model in Figure 24 shows that the control group and the treatment group overlap at three points. Obviously, the two groups plots are not parallel. Therefore, the PH assumption is not met for this model.

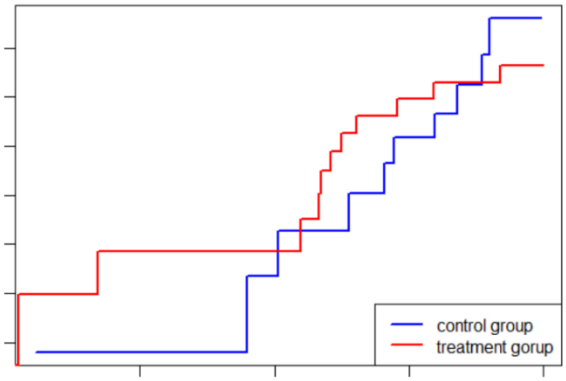


Figure 24: plot of the log cumulative hazards for the Brain cancer treatment group model

- Schoenfeld Residuals

Schoenfeld residual plots can assess the PH assumption by plotting the residuals vs. the covariate to get residual points centered at zero (20)[12]. If the residuals are centered at zero, then the PH assumption is satisfied. These residuals are derived from the partial log-likelihood function and is denoted as in equation 29 and 30[9]:

$$r_i = x_i - \bar{x}(t_i) = x_i - \sum_{k \epsilon R_i} x_k * p(\hat{\beta}, x_k) \text{ (equation 29),}$$

where $p(\hat{\beta}, x_k)$ is an estimate of

$$p(\beta, x_k) = \frac{e^{x_k \beta}}{\sum_{j \epsilon R_i} e^{x_j \beta}} = E(X_i) = \bar{x}(t_i), \text{ (equation 30)}$$

the expected value as the weighted sum of the $p(\beta)$ (96). The r_i are "defined only for the failure (and not the censoring) times," according to Moore, D.[9]. Multiple plots of the Schoenfeld residuals are obtained for cases where there are multiple covariates, one plot for each covariate. An example of using Schoenfeld residuals to check the PH assumption below.

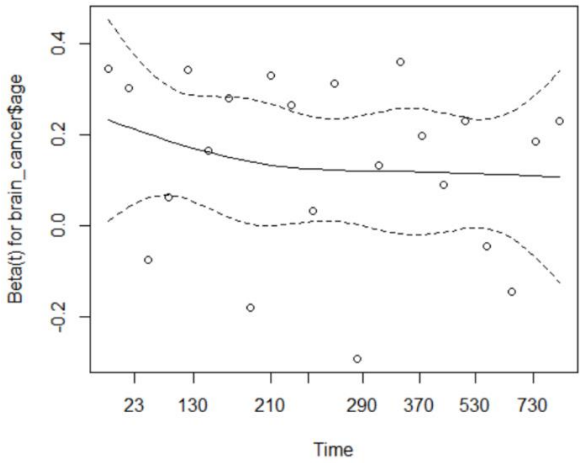


Figure 25: smoothed (loess) curve

The shape of the smoothed (loess) curve is an estimate of the difference parameter (the y-axis, in this case, the Beta(t) for age) as a function of time. The plot in Figure 25 shows the line to be constant or slightly decreasing initially, followed by a steady constant after about 210 days. There is also the 95% confidence interval for this curve. According to the results in Figure 26, the p-value of the proportional hazards function hypothesis test for a constant β is 0.35, for age. The results from this test are found from the plot of the straight line to the residuals vs. the time above in Figure 26. We do not reject the null hypothesis of the PH assumption being met in favor of the alternative hypothesis that the PH assumption is not met. Therefore, it is probably safe to assume that the PH assumption is met for the Cox PH model with just the covariate age in the model.

```
> result.sch.resid
                  chisq df    p
brain_cancer$age 0.883  1 0.35
GLOBAL           0.883  1 0.35
```

Figure 26: the results from this test are found from the plot of the straight line to the residuals vs. the time

- Working with Time Dependent Covariates

A caveat of the partial likelihood theory in the PH Model and the Cox PH model sections is that each of the covariate values need to be determined at time zero and then need to be constant over the time the subject is in the study[9]. According to Stevenson, M.[12], one of the two common types of departure from the proportional hazards assumption is "the influence of a covariate diminishes with time." Such a situation makes the model the time-dependent covariate an "improper" covariate[12]. There are some methods to remedy this issue with any time-dependent covariates to get the appropriate parameter estimates for the survival model.

The data set named jasa will be looked at in this section. The jasa data is a similar data set as the heart transplants plant data set used in the above non-parametric survival sections. It was chosen because there is a time-dependent covariate in it. According to Moore, D.[9],

> "An often cited and extensively studied example of this is the Stanford heart transplant study, published by Clark et al. in the Annals of Internal Medicine in 1971[9]. This study of the survival of patients who had been enrolled into the transplant program appeared to show that patients who received heart transplants lived significantly longer than those who did not. The data are in the 'survival' package in a data set named "jasa" after a journal article that discussed analysis methods for the data" (101).

This data set is good to demonstrate working with time-dependent covariates. This will be demonstrated below.

A Cox PH model was fitted on the jasa data and is shown in Figure 27 below. Notice that transplant and age are both statistically significant because their p-values are nearly zero. However, the "key covariate is 'transplant', which takes the value 1 for those patients who received a heart transplant and 0 for those who do not," according to Moore, D.[9]. The transplant variable's estimated coefficient is -1.717. There are some problems with this covariate that need to be discussed.

```
> summary(result.heart)
Call:
coxph(formula = Surv(futime, fustat) ~ transplant + age + +surgery,
    data = jasa)

  n= 103, number of events= 75

               coef exp(coef) se(coef)      z Pr(>|z|)
transplant -1.71711   0.17958  0.27853 -6.165 7.05e-10 ***
age         0.05889   1.06065  0.01505  3.913 9.12e-05 ***
surgery    -0.41902   0.65769  0.37118 -1.129    0.259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
transplant    0.1796     5.5684    0.1040     0.310
age           1.0607     0.9428    1.0298     1.092
surgery       0.6577     1.5205    0.3177     1.361

Concordance= 0.732  (se = 0.031 )
Likelihood ratio test= 45.85  on 3 df,   p=6e-10
Wald test            = 47.15  on 3 df,   p=3e-10
Score (logrank) test = 52.63  on 3 df,   p=2e-11
```

Figure 27: Cox PH model was fitted on the jasa data

When the covariate's effect varies over time, but the covariate does not vary over time, it is fixed[12]. To remedy fixed time-dependent covariates, the step function PH model, or A.K.A. the piecewise Cox model should be fit on the data. First, the time can be divided into different time intervals. Then, a (Cox) PH model is fit to the survival data, and the coefficients of the covariate's time intervals are compared to see if the coefficients changes over time[12]. If the coefficients are all different, then the coefficients change over time. Therefore, the model is a non-proportional hazards model. If this is the case, then perhaps another model, such as a parametric model, should be fit on the data.

There are some problems with the jasa survival model. Moore, D.[9] had some things to say about the problem with the model developed in Figure 27 above.

> "This result may appear to indicate (as it did to Clark et al. in 1971) that transplants are extremely effective in increasing the lifespan of the recipients. Soon after publication of this result, Gail [21], in an article in the same journal, questioned the validity of the result, and numerous re-analyses of the data followed."

Transplant is a time-dependent covariate. When patients have had heart transplants done, they had lived long enough to get the surgery. That is, patients who are older and live to an age where they can get the heart transplant will live long enough than others who do not live to an age where they can receive the transplant.

A remedy for such a situation is to divide the patients into two groups: those who got the heart transplant before a landmark time, such as 30 days, and those who did not. The former group is called the intervention group and the latter is called the control group. Moore, D.[9] states that the "key requirements of this approach are that (a) only patients who survive up to the landmark are included in the study, and (b) all patients (in particular, those in the comparison group) remain n their originally assigned group regardless of what happens in the future, i.e., after the landmark." With the landmark time of 30 days, the 79 patients who lived 30 days more, 33 of them had a heart transplant surgery within 30 days and the 46 of the others didn't. The 30 people in this latter group did end up having a heart transplant, but they are still in the control group.

This approach to solving the issue of the time-dependent covariate created a variable that has a fixed value in the set of all patients in the set of survivors past 30 days. This binary, TRUE-FALSE variable is called transplant30, as shown in Figure 28 below. First, the observations who lived past 30 days are found, and then those people are filtered so that they must have heart transplant and a wait-time of getting the heart transplant of less than 30 days.

```
> ind30 <- jasa$futime >= 30
> transplant30 <- {{jasa$transplant == 1} & {jasa$wait.time < 30}}
> summary(coxph(Surv(futime, fustat) ~ transplant30 + age + surgery,
+               data=jasa, subset=ind30 ))
Call:
coxph(formula = Surv(futime, fustat) ~ transplant30 + age + surgery,
    data = jasa, subset = ind30)

  n= 79, number of events= 52

                     coef exp(coef) se(coef)      z Pr(>|z|)
transplant30TRUE -0.04214   0.95874  0.28377 -0.148   0.8820
age               0.03720   1.03790  0.01714  2.170   0.0300 *
surgery          -0.81966   0.44058  0.41297 -1.985   0.0472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
transplant30TRUE    0.9587     1.0430    0.5497    1.6720
age                 1.0379     0.9635    1.0036    1.0734
surgery             0.4406     2.2697    0.1961    0.9898

Concordance= 0.618  (se = 0.044 )
Likelihood ratio test= 9.5  on 3 df,   p=0.02
Wald test            = 8.61  on 3 df,   p=0.03
Score (logrank) test = 8.94  on 3 df,   p=0.03
```

Figure 28: transplant30 definition

Then a Cox PH model is fit. Notice how the coefficient of transplant30 is -0.04214 with the p-value = 0.8820. This p-value is not significant at any reasonable level, unlike how it was above. According to Moore, D.[9], the "analysis indicates that there is little or no different in survival between those who got a transplant and those who did not." One limitation with this approach is that the 30-day landmark-time is arbitrary.

When the covariates change with time, "the survival period is divided into a sequence of shorter 'survival spells', each characterized by an entry and exit time, and within which covariate values remain fixed" for each subject, according to Stevenson, M.[12]. Each subject's data is represented by several shorter censored intervals, with possibly one of the intervals ending with the event.

To illustrate this, Moore's example of taking a small number of the subjects in the fasa data set to look at. Three of the six had a transplant and the others did not. In Figure 29 below, subjects 2, 5, 10, 12, 28, and 95, where 2, 5, and 12 did not have a transplant but subjects 10, 28 and 95 did. According to Moore, D.[9], "in this simple data set, all of the patients died within the follow-up time." This data is modelled in a Cox PH model in Figure 29 below.

```
> id <- 1:nrow(jasa)
> jasaT <- data.frame(id, jasa)
> id.simple <- c(2, 5, 10, 12, 28, 95)
> id.simple
[1]  2  5 10 12 28 95
> heart.simple <- jasaT[id.simple, c(1, 10, 9, 6,
 11)]
> heart.simple
   id wait.time futime fustat transplant
2   2        NA      5      1          0
5   5        NA     17      1          0
10 10        11     57      1          1
12 12        NA      7      1          0
28 28        70     71      1          1
95 95         1     15      1          1
> summary(coxph(Surv(futime, fustat)~transplant,
+               data = heart.simple))
Call:
coxph(formula = Surv(futime, fustat) ~ transplant,
 data = heart.simple)

  n= 6, number of events= 6

              coef exp(coef) se(coef)     z
transplant -1.6878    0.1849   1.1718 -1.44
           Pr(>|z|)
transplant     0.15

           exp(coef) exp(-coef) lower .95
transplant    0.1849      5.408    0.0186
           upper .95
transplant     1.838

Concordance= 0.733  (se = 0.077 )
Likelihood ratio test= 2.47  on 1 df,   p=0.1
Wald test            = 2.07  on 1 df,   p=0.1
Score (logrank) test = 2.56  on 1 df,   p=0.1
```

Figure 29: transplant data modelled incorrectly in a Cox PH model

To get the right model, use the partial likelihood function to accommodate the time-dependent covariates. Moore, D.[9] says:

> "The hazard function is given by $h(t) = h_0(t)e^{z_k(t_i)\beta}$, where the covariate $z_k(t_i)$ is the value of the time-varying covariate for the kth subject at time t_i. The modified partial likelihood, in general … where $\varphi_{ki} = e^{z_k(t_i)\beta}$" (104) is denoted as in equation 31.

$$L(\beta) = \prod_{i=1}^{D} \frac{\varphi_{ii}}{\sum_{k \epsilon R_i} \varphi_{ki}}, \text{ (equation 31)}$$

For the time-dependent covariates, in contrast with time-independent covariates, the denominator "has to be recalculated at each failure time" because each subject's value of the covariate(s) might change from one time to the next, according to Moore, D.[9].

For example, subject 2 in Figure 29 above, was the first person to fail, at the future time (futime) t = 5. According to Moore, D.[9], "at this time, all six patients are at risk, but only one, Patient #95, has had a transplant at this time. So the denominator for the first factor is $5 + e^{\beta}$, and the numerator is 1, since it was a non-transplant patient who died" (104). Next, take a look at an example of when there was a transplant. Subject 95 was the first with a transplant and at time 15, there were four patients at risk, where two are transplant patients (#95 and #10). The partial likelihood of this term is $\frac{e^{\beta}}{2+2e^{\beta}}$. The full partial likelihood is $L(\beta) = \frac{1}{5+e^{\beta}} * \frac{1}{4+e^{\beta}} * \frac{e^{\beta}}{2+2e^{\beta}} * \frac{1}{2+e^{\beta}} * \frac{e^{\beta}}{1+e^{\beta}} * \frac{e^{\beta}}{e^{\beta}}$.

There are some more comments on partial likelihood for time-dependent covariates to be mentioned. According to Stevenson, M.[12], "the partial likelihood on which estimation is based has a term for each unique death or event time and involves sums over those observations that are available or at risk at the actual event date." Thus, there is no issue of observations being correlated, and therefore not suitable for a cox model to be fitted on the collection of the observations. The observations are all independent of one another, as the intervals created do not overlap and are different. The covariate values between the event times will not enter the final partial likelihood that the model is based on.

With that said, the Cox PH model can accommodate the time-dependent variables by processing the data prior to the modeling in a "start-stop" format. According to Moore, "the validity of this approach may be derived from the counting process theory of partial likelihoods. Essentially, this approach divides the time data for patients who had a heart transplant into two time periods, one before the transplant and one after" (104). Notice that, in Figure 30 below, "essentially, this approach divides the time data for patients who had a heart transplant into two time periods, one before the transplant and one after," according to Moore, D.[9]. This is done with the "tmerge" function in the survival package in R. The data is then modelled in a Cox PH model, after being modified, as shown in Figure 30 below.

```
> sdata <- tmerge(heart.simple, heart.simple, id=id,
+                 death=event(futime, fustat),
+                 transpl=tdc(wait.time))
> heart.simple.counting <- sdata[,-(2:5)]
> # drop columns 2 thorugh 5
> heart.simple.counting
  id tstart tstop death transpl
1  2      0     5     1       0
2  5      0    17     1       0
3 10      0    11     0       0
4 10     11    57     1       1
5 12      0     7     1       0
6 28      0    70     0       0
7 28     70    71     1       1
8 95      0     1     0       0
9 95      1    15     1       1
```

```
> summary(coxph(Surv(tstart, tstop, death) ~ transpl,
+               data=heart.simple.counting))
Call:
coxph(formula = Surv(tstart, tstop, death) ~ transpl, data
 = heart.simple.counting)

  n= 9, number of events= 6

          coef exp(coef) se(coef)     z Pr(>|z|)
transpl 0.2846    1.3292   0.9609 0.296    0.767

        exp(coef) exp(-coef) lower .95 upper .95
transpl     1.329     0.7523    0.2021      8.74

Concordance= 0.5  (se = 0.082 )
Likelihood ratio test= 0.09  on 1 df,   p=0.8
Wald test            = 0.09  on 1 df,   p=0.8
Score (logrank) test = 0.09  on 1 df,   p=0.8
```

Figure 30: heart transplant Cox PH model

Now, the method above can be applied to the full heart transplant dataset. The modified data and the Cox PH model are shown in Figure 31 below. According to Moore, D.[9], "in the following, we define 'tdata' as a temporary data set, leaving off the dates and transplant-specific covariates. Also, we add 0.5 to the death time on day 0, and break a tied transplant time" (106). Notice that for Subject ID 4, the heart transplant happened for them on day 35. They died on day 38. Regarding the Cox PH model results. There is no evidence that heart transplants increase survival because the p-value for trans is 0.9636 is greater than any reasonable significant level.

```
> #full dataset:
> tdata <- jasa[, -c(1:4, 11:14)]
> tdata$futime <- pmax(.5, tdata$futime)
> indx <- {{tdata$wait.time == tdata$futime} &
+     !is.na(tdata$wait.time)}
> tdata$wait.time[indx] <- tdata$wait.time[indx] - .5
> id <- 1:nrow(tdata)
> tdata$id <- id
> sdata <- tmerge(tdata, tdata, id=id,
+                   death = event(futime, fustat),
+                   trans = tdc(wait.time))
> jasa.counting <- sdata[,c(7:11, 2:3)]
> head(jasa.counting)
  id tstart tstop death trans surgery      age
1  1      0    49     1     0       0 30.84463
2  2      0     5     1     0       0 51.83573
3  3      0    15     1     1       0 54.29706
4  4      0    35     0     0       0 40.26283
5  4     35    38     1     1       0 40.26283
6  5      0    17     1     0       0 20.78576
```

```
> summary(coxph(Surv(tstart, tstop, death) ~
+                 trans + surgery +
+               age, data=jasa.counting))
Call:
coxph(formula = Surv(tstart, tstop, death) ~ trans + surge
ry +
    age, data = jasa.counting)

  n= 170, number of events= 75

            coef exp(coef) se(coef)      z Pr(>|z|)
trans    0.01405   1.01415  0.30822  0.046   0.9636
surgery -0.77326   0.46150  0.35966 -2.150   0.0316 *
age      0.03055   1.03103  0.01389  2.199   0.0279 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
trans      1.0142     0.9860    0.5543    1.8555
surgery    0.4615     2.1668    0.2280    0.9339
age        1.0310     0.9699    1.0033    1.0595

Concordance= 0.599  (se = 0.036 )
Likelihood ratio test= 10.72  on 3 df,   p=0.01
Wald test            = 9.68  on 3 df,   p=0.02
Score (logrank) test = 10  on 3 df,   p=0.02
```

Figure 31: Cox PH model of the modified heart transplant data.

- ### Working with multiple survival outcomes and competing risks

There are often cases where the survival times of each case are dependent on each other. For example, cluster data or the event of interest may violate the independence assumption. According to Moore, D.[9], in the case where the clusters of people are similar and characteristics are dependent on each other, such as schools or a city, "genetic or environmental factors mean that survival times with a cluster are more similar to each other than to those from other clusters, so that the independence assumption no longer holds" (113). The event of interest may occur repeatedly over time, with each event dependent on the previous. For example, seizures or heart attacks may occur several times for a person.

Competing risks can be modeled by a Cox PH (the most popular method) or by a parametric model, or through the use of any model that doesn't use survival, but rather cumulative incidence[8]. The dataset that will be used to show the modeling of competing risks regarding clusters of families is talked about on Moore. According to Kleinbaum[8],

> "Struewing et al. [64], in the Washington Ashkenazi study, examined the effect of mutations of the BRCA gene on risk of breast cancer in an Ashkenazi Jewish population. The original data set consisted of a set of probands who were volunteers of Ashkenazi ancestry. Each proband was genotyped for the BRCA breast cancer gene to determine if she was a mutation

carrier. The proband was also interviewed by the investigators to determine if she had any female first-degree relatives, and the relatives age at the time she developed breast cancer or current age if that relative had never been diagnosed with breast cancer."

This dataset is a subset of 1960 families with at least two female relatives in them and is in the "asaur" package, labeled as "askenazi". Only two women from each family were selected at random (if there were more than 2 females in a family) to be in the study.

```
> ashkenazi[ashkenazi$famID %in% c(1, 11, 93), ]
   famID brcancer age mutant
1      1        0  73      0
2      1        0  40      0
9     11        0  77      0
10    11        0  41      0
85    93        1  89      0
86    93        0  64      0
```

Figure 32: sample of three families in the Ashkenazi dataset

A quick sample of three families in the Ashkenazi dataset are shown in Figure 32 above: family 1, 11, and 93. In family 93, there are two first-degree related women, the older one 89 years and the younger one 64 years. Perhaps, they are mother and daughter. None of them are mutated, but the older women developed breast cancer at age 89.

- Full data analysis

A full data analysis was done on the VA lung cancer dataset. The 137 subjects in this dataset were lung cancer therapy patients at a Veteran Administration (VA) Hospital during a lung cancer treatment trial. The patients with "advanced, inoperable lung cancer were treated with chemotherapy," according to Survival data[13]. The data has some good options for covariates.

There are 8 variables in the dataset: treatment, cell type, survival, status, Karnofsky score, months from diagnosis, age, and prior therapy. Treatment, with two categories (1 = standard and 2 = test), is the treatment that the patient had been assigned in the clinical trial. Cell type is the type of cell with four different types: squamous, small cell, adeno, and large cancer cell. The survival of the patient is in days, and thus is numeric. Status of the subjects are represented in a binary categorical variable, where we have the info on is a patient is dead (1) or was censored (0). A Karnofsky score is a scale of 0 to 100, by tens and is a measure of general performance of the treatment on the patient. Age is measured in years. The prior therapy variable indicates if the patient has had or has not had any prior therapy for their lung cancer.

First, the parametric model Weibull and non-parametric Kaplan-Meier survival functions are calculated for the VA dataset. Kaplan-Meier can be estimated by the Weibull distribution. As seen in Figure 33 below, the Weibull distribution provides an adequate fit to the observed data up to day 1000, then appears to underestimate survivorship.

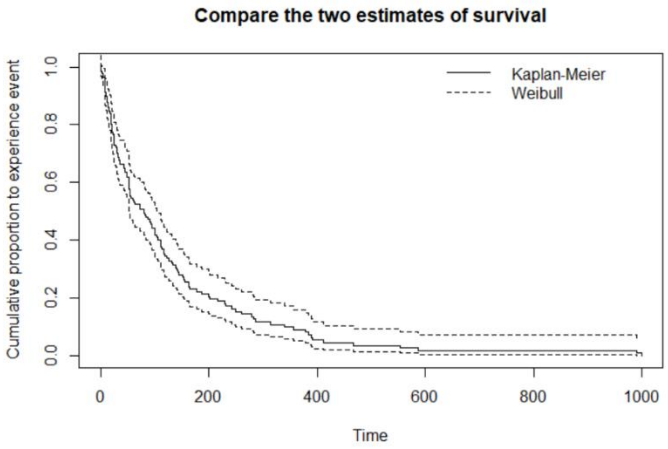


Figure 33: Weibull distribution vs. Kaplan-Meier on VA dataset

Recall that an alternative method of assessing if a parametric model appropriately describes survivorship of a dataset are through the analysis of cumulative hazard plots, as described above in the section on log cumulative hazard plots. As shown in the cumulative hazard plot in Figure 34 below, the left of the Kaplan Meier plots show departure from linearity while the right does not.

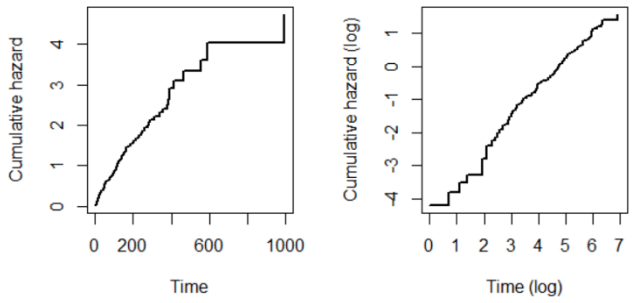


Figure 34: cumulative hazard plot

Next is a case of stratification of the Kaplan-Meier by some categorical variable. Perhaps treatment is stratified. Here is the plot of Kaplan-Meier survival function of days to lung cancer by treatment, shown in the plot in Figure 35 below. T, it should be noted that the median for treatment 1 is 103.5 (with 95% CI of 59 to 132), about twice that of treatment 2, with a median of 52.5 (with 95% CI of 44 to 95). However, this observation will not affect the final answer of whether the two treatment groups produce different results or not. There doesn't appear to be a big difference between the two treatments in Figure 35 below because the lines for the two treatments are relatively close.

```
> va_lung.km
Call: survfit(formula = Surv(Survival, Status) ~ Treatment, data = va_lung,

    type = "kaplan-meier")

              n events median 0.95LCL 0.95UCL
Treatment=1 69     64  103.0      59     132
Treatment=2 68     64   52.5      44      95
```

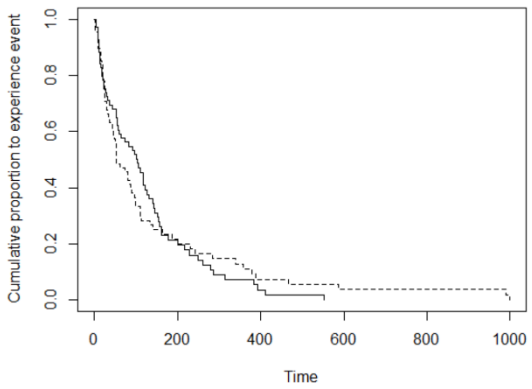


Figure 35: two treatment's Kaplan-Meier curves for VA data

Recall that hypothesis testing can be used to determine if there is a difference between the two treatment's Kaplan-Meier curves or not. Recall that the log-rank (Mantel Haenszel test) and/or the Peto and Peto modification of the Gehan-Wilcoxon test can both be used to achieve this goal. In the "survdiff" function in the survival package in R, "the argument rho = 0 returns the log-rank or Mantel-Haenszel test and rho = 1 returns the Peto and Peto modification of the Gehan-Wilcoxon test. Mantel-Haenszel test," according to Stevenson, M.[12]. Recall that the null hypothesis is there is no difference between the survival curves split by the categories. The alternative hypothesis is that there is a difference in the survival curves.

```
> survdiff(Surv(Survival, Status) ~ Treatment, data = va_lung, na.action = na.omit, rho = 0)
Call:
survdiff(formula = Surv(Survival, Status) ~ Treatment, data = va_lung,
    na.action = na.omit, rho = 0)

             N Observed Expected (O-E)^2/E (O-E)^2/V
Treatment=1 69       64     64.5   0.00388   0.00823
Treatment=2 68       64     63.5   0.00394   0.00823

 Chisq= 0  on 1 degrees of freedom, p= 0.9
```

```
> survdiff(Surv(Survival, Status) ~ Treatment, data = va_lung, na.action = na.omit, rho = 1)
Call:
survdiff(formula = Surv(Survival, Status) ~ Treatment, data = va_lung,
    na.action = na.omit, rho = 1)

             N Observed Expected (O-E)^2/E (O-E)^2/V
Treatment=1 69     32.2     35.4     0.279     0.871
Treatment=2 68     35.2     32.1     0.308     0.871

 Chisq= 0.9  on 1 degrees of freedom, p= 0.4
```

Figure 36: results of the log-rank test for VA data (a) log-rank and (b) Peto and Peto modification

The results of the log-rank test in the first part of Figure 36 has the p-value = 0.9. The null hypothesis will not be rejected from the log-rank test results. It is likely safe to say that there is no difference between the survival curves split by the categories. The results of the Peto and Peto modification of the Gehan-Wilcoxon test in the second part of Figure 36 (b) has the p-value = 0.4. These results are consistent with the log-rank test and the same conclusion stated for the that test would apply to the Peto and Peto modification of the Gehan-Wilcoxon test as well.

Next, the significant covariates should be selected to make the optimal survival model. Let's set contrasts for cell type and prison by setting the reference category for cell type, making cell type 1 (base = 1) the reference category. The same action will be completed for prior

therapy, making absence of a prior therapy the reference category. Next, Karnofsky score is categorized into four different classes based on its quartiles, as shown in Figure 37 below.

```
25% 50% 75%
 40  60  75
```

Figure 37: Karnofsky score categorized into four different classes

So the quartiles of Karnofsky score are 40, 60, and 75. So, we create four categories: when Karnofsky score is less than 40, Karnofsky score is between 40 and 60, Karnofsky score is between 60 and 75, and Karnofsky score is greater than 75.

Assess the influence of Cell type, Prior therapy, Treatment, and Karnofsky score on time to death. As shown in Figure 38 below, the Kaplan-Meier curves for each of the predictor variables by survival time and status provide a lot of areas for analysis. It looks like Cell type's and Karnofsky score's survival curves are different, but Prior therapy's Kaplan-Meier look like they are not different from each other.

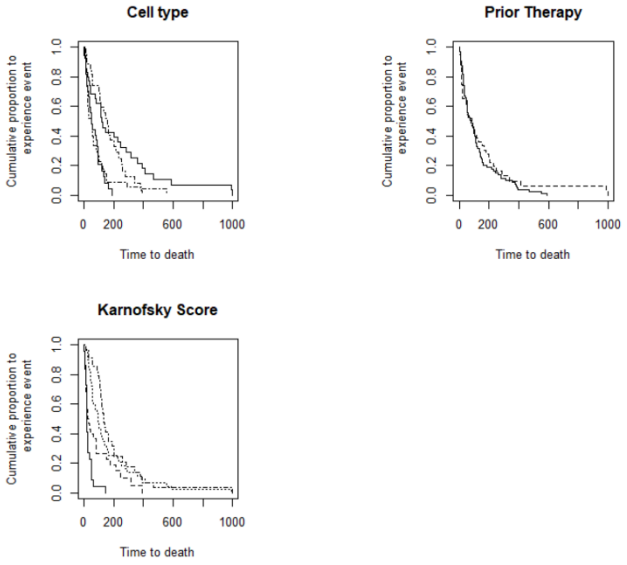


Figure 38: Kaplan-Meier curves for Cell type, Prior therapy, and Karnofsky score

The log rank test was completed for each of the variables mentioned above (Cell type, Prior therapy, Treatment, and Karnofsky score) to see if the various categories have statistically different Kaplan-Meier curves, in Figure 39. These results say that like Cell types' and Karnofsky scores' survival curves are different because the (nearly zero) p-values of each are less than any reasonable significance value. Prior therapy's Kaplan-Meier results demonstrate that the group groups (with prior therapy and without prior therapy) are probably not different from each other. That is, the p-value is 0.5 which is greater than any reasonable significance level. Thus, the null hypothesis that says there is no difference between the two survival curves is not rejected.

```
> survdiff(Surv(Survival, Status) ~ va_lung$Karnofsky.cat,
+          data = va_lung, na.action = na.omit, rho = 0)
Call:
survdiff(formula = Surv(Survival, Status) ~ va_lung$Karnofsky.cat,
    data = va_lung, na.action = na.omit, rho = 0)

                         N Observed Expected (O-E)^2/E (O-E)^2/V
va_lung$Karnofsky.cat=1 22       22     6.46     37.41     42.23
va_lung$Karnofsky.cat=2 30       28    19.84      3.36      4.06
va_lung$Karnofsky.cat=3 50       47    55.17      1.21      2.19
va_lung$Karnofsky.cat=4 35       31    46.53      5.18      8.44

 Chisq= 51  on 3 degrees of freedom, p= 5e-11
```

```
> survdiff(Surv(Survival, Status) ~ va_lung$"Prior therapy",
+          data = va_lung, na.action = na.omit, rho = 0)
Call:
survdiff(formula = Surv(Survival, Status) ~ va_lung$"Prior therapy",
    data = va_lung, na.action = na.omit, rho = 0)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
va_lung$"Prior therapy"=0  97       91     87.4     0.150     0.501
va_lung$"Prior therapy"=10 40       37     40.6     0.323     0.501

 Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

```
> survdiff(Surv(Survival, Status) ~ va_lung$"Cell type",
+          data = va_lung, na.action = na.omit, rho = 0)
Call:
survdiff(formula = Surv(Survival, Status) ~ va_lung$"Cell type",
    data = va_lung, na.action = na.omit, rho = 0)

                          N Observed Expected (O-E)^2/E (O-E)^2/V
va_lung$"Cell type"=1 35       31     47.7      5.82     10.53
va_lung$"Cell type"=2 48       45     30.1      7.37     10.20
va_lung$"Cell type"=3 27       26     15.7      6.77      8.19
va_lung$"Cell type"=4 27       26     34.5      2.12      3.02

 Chisq= 25.4  on 3 degrees of freedom, p= 1e-05
```

Figure 39: log rank tests of the variables Cell type, Prior therapy, and Karnofsky score

Next, the cox proportional hazards multivariable model will be fitted on survival time using prior therapy, Karnofsky score, and cell type. The resulting model is shown in Figure 40 below. The variables Karnofsky score and cell type significantly influence time to death variable. Prior therapy is not statistically significant. Therefore, the variable Prior therapy is dropped from the survival model.

```
> summary(va_lung.cph01)
Call:
coxph(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Prior therapy" + va_lung$"Cell type", data = va_lung,
    method = "breslow")

  n= 137, number of events= 128

                             coef exp(coef) se(coef)      z Pr(>|z|)
Karnofsky.cat           -0.59197   0.55324  0.10576 -5.597 2.18e-08 ***
va_lung$"Prior therapy"2  0.07954   1.08279  0.20611  0.386 0.699577
va_lung$"Cell type"2      0.84429   2.32631  0.25438  3.319 0.000904 ***
va_lung$"Cell type"3      1.33846   3.81318  0.30143  4.440 8.98e-06 ***
va_lung$"Cell type"4      0.36423   1.43940  0.27714  1.314 0.188761
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                          exp(coef) exp(-coef) lower .95 upper .95
Karnofsky.cat                0.5532     1.8075    0.4497    0.6807
va_lung$"Prior therapy"2     1.0828     0.9235    0.7229    1.6218
va_lung$"Cell type"2         2.3263     0.4299    1.4130    3.8300
va_lung$"Cell type"3         3.8132     0.2622    2.1121    6.8844
va_lung$"Cell type"4         1.4394     0.6947    0.8361    2.4779

Concordance= 0.717  (se = 0.022 )
Likelihood ratio test= 55.44  on 5 df,   p=1e-10
Wald test            = 53.9  on 5 df,   p=2e-10
Score (logrank) test = 57.46  on 5 df,   p=4e-11
```

Figure 40: cox proportional hazards multivariable model fitted on survival time using prior therapy, Karnofsky score, and cell type

The new model is then checked to see if it provides a better fit to the data than the old model with prior therapy in it. The resulting Cox PH model is shown in in Figure 41. Removing prior therapy form the model has no effect on model fit because the p-value of the chi-square test is about 1.

```
> va_lung.cph02 <- update(va_lung.cph01, ~. - va_lung$"Prior therapy")
> summary(va_lung.cph02)
Call:
coxph(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type", data = va_lung, method = "breslow")

  n= 137, number of events= 128

                          coef exp(coef) se(coef)      z Pr(>|z|)
Karnofsky.cat       -0.5896    0.5546   0.1053 -5.596 2.19e-08 ***
va_lung$"Cell type"2  0.8317    2.2972   0.2520  3.300 0.000967 ***
va_lung$"Cell type"3  1.3166    3.7308   0.2956  4.454 8.43e-06 ***
va_lung$"Cell type"4  0.3627    1.4371   0.2770  1.309 0.190395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                      exp(coef) exp(-coef) lower .95 upper .95
Karnofsky.cat            0.5546     1.8032    0.4511    0.6818
va_lung$"Cell type"2     2.2972     0.4353    1.4017    3.7646
va_lung$"Cell type"3     3.7308     0.2680    2.0902    6.6593
va_lung$"Cell type"4     1.4371     0.6958    0.8351    2.4731

Concordance= 0.718  (se = 0.022 )
Likelihood ratio test= 55.29  on 4 df,   p=3e-11
Wald test            = 53.93  on 4 df,   p=5e-11
Score (logrank) test = 57.45  on 4 df,   p=1e-11

> # Does va_lung.cph02 provide a better fit to the data than va_lung.cph01?
> x2 <- 2 * (va_lung.cph02$loglik[2] - va_lung.cph01$loglik[2])
> 1 - pchisq(x2,1)
[1] 1
```

Figure 41: new model is then checked to see if it provides a better fit to the data than the old model with prior therapy in it

The scale of continuous covariates with no transformations required for variable Karnofsky score can also be checked. The plot covariate values versus Martingale residuals and va_lung.yi versus covariate values were found and then plotted to see if there is a linear relationship between the covariate values. According to the Figure 42 below, a negative linear relationship is evident between the covariate values and each of the calculated parameters. This shows that the continuous covariate Karnofsky score is linear in its log hazard. Also, the Martingale Residual vs. Karnofsky plot shows that there is no departure from linearity because all of the residuals fall within -3 and 3.

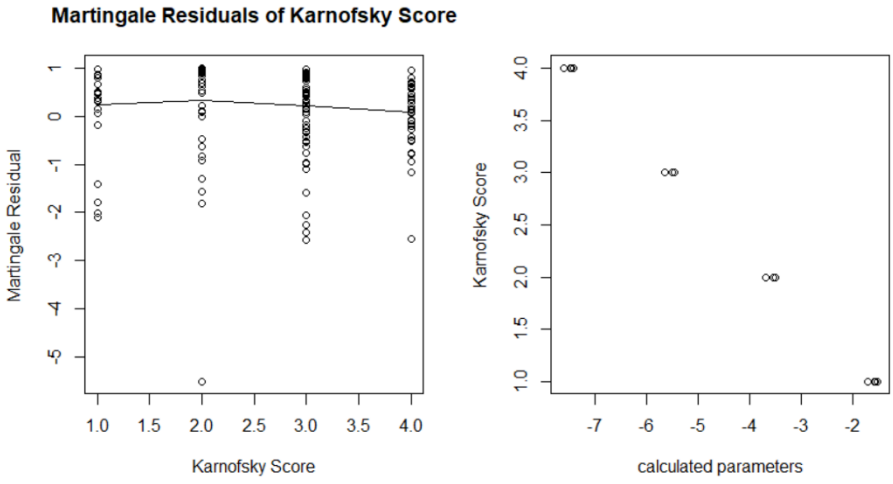


Figure 42: covariate values versus Martingale residuals

Next, any potential interactions in the Cox PH model will be investigated. The interactions in Figure 43 below are between the categorical variables. First, notice that the p-values for all the interaction terms are all not significant at a reasonable significance level. The p-values of the Wald test, Likelihood ratio test, and score (log rank) test for the interaction term Karnofsky.cat * Cell type are all significant (p-value = 4e-10), so the null hypothesis of significant interactions in the model will be rejected. According to Stevenson, M.[12], "The effect of adding an interaction

term should be assessed using the partial likelihood ratio test. All significant interactions should be included in the main-effect model. Wald statistic p-values can be used as a guide to selecting interactions that may be eliminated from the model, with significance checked by the partial likelihood ratio test." We can conclude that none of the interactions should be kept in the final main-effect model.

```
coxph(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type" + (Karnofsky.cat * va_lung$"Cell type"),
    data = va_lung, method = "breslow")

  n= 137, number of events= 128

                                          coef exp(coef)  se(coef)      z Pr(>|z|)
Karnofsky.cat2                       -1.290410  0.275158  0.617365 -2.090 0.036601 *
Karnofsky.cat3                       -1.824534  0.161293  0.554162 -3.292 0.000993 ***
Karnofsky.cat4                       -2.392883  0.091366  0.629439 -3.802 0.000144 ***
va_lung$"Cell type"2                  0.753826  2.125115  0.555440  1.357 0.174728
va_lung$"Cell type"3                  1.052083  2.863608  0.743407  1.415 0.157005
va_lung$"Cell type"4                  0.696987  2.007695  0.743622  0.937 0.348611
Karnofsky.cat2:va_lung$"Cell type"2 -0.258520  0.772194  0.760811 -0.340 0.734011
Karnofsky.cat3:va_lung$"Cell type"2  0.191106  1.210587  0.677961  0.282 0.778033
Karnofsky.cat4:va_lung$"Cell type"2  1.064767  2.900164  0.788924  1.350 0.177130
Karnofsky.cat2:va_lung$"Cell type"3  0.898214  2.455213  0.910129  0.987 0.323688
Karnofsky.cat3:va_lung$"Cell type"3  0.298591  1.347958  0.889525  0.336 0.737116
Karnofsky.cat4:va_lung$"Cell type"3  0.254533  1.289859  0.940519  0.271 0.786675
Karnofsky.cat2:va_lung$"Cell type"4 -0.629433  0.532894  1.030202 -0.611 0.541213
Karnofsky.cat3:va_lung$"Cell type"4 -0.397085  0.672277  0.855775 -0.464 0.642643
Karnofsky.cat4:va_lung$"Cell type"4 -0.004173  0.995836  0.921594 -0.005 0.996388
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                                     exp(coef) exp(-coef) lower .95 upper .95
Karnofsky.cat2                         0.27516     3.6343   0.08205    0.9228
Karnofsky.cat3                         0.16129     6.1999   0.05444    0.4779
Karnofsky.cat4                         0.09137    10.9450   0.02661    0.3137
va_lung$"Cell type"2                   2.12511     0.4706   0.71547    6.3121
va_lung$"Cell type"3                   2.86361     0.3492   0.66700   12.2943
va_lung$"Cell type"4                   2.00770     0.4981   0.46744    8.6232
Karnofsky.cat2:va_lung$"Cell type"2    0.77219     1.2950   0.17383    3.4303
Karnofsky.cat3:va_lung$"Cell type"2    1.21059     0.8260   0.32056    4.5717
Karnofsky.cat4:va_lung$"Cell type"2    2.90016     0.3448   0.61786   13.6131
Karnofsky.cat2:va_lung$"Cell type"3    2.45521     0.4073   0.41246   14.6148
Karnofsky.cat3:va_lung$"Cell type"3    1.34796     0.7419   0.23578    7.7062
Karnofsky.cat4:va_lung$"Cell type"3    1.28986     0.7753   0.20416    8.1492
Karnofsky.cat2:va_lung$"Cell type"4    0.53289     1.8765   0.07075    4.0137
Karnofsky.cat3:va_lung$"Cell type"4    0.67228     1.4875   0.12564    3.5974
Karnofsky.cat4:va_lung$"Cell type"4    0.99584     1.0042   0.16358    6.0625

Concordance= 0.721  (se = 0.022 )
Likelihood ratio test= 69.43  on 15 df,   p=6e-09
Wald test            = 69.27  on 15 df,   p=6e-09
Score (logrank) test = 87.64  on 15 df,   p=3e-12
```

Figure 43: potential interactions in the Cox PH model between categorical variables

Now, the Proportional Hazards assumption needs to be met in order to continue with the model. To test this, a plot of the scaled Schoenfeld residual plots can first be investigated. Recall that "in a `well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0," according to Stevenson, M.[12]. The variability band for Karnofsky score is above zero and there are a lot of outliers above 3. For cell type, the variability band is closer to zero and has a wave or curve similar to Karnofsky score as time increases. These observations both suggest the non-proportionality of hazards.

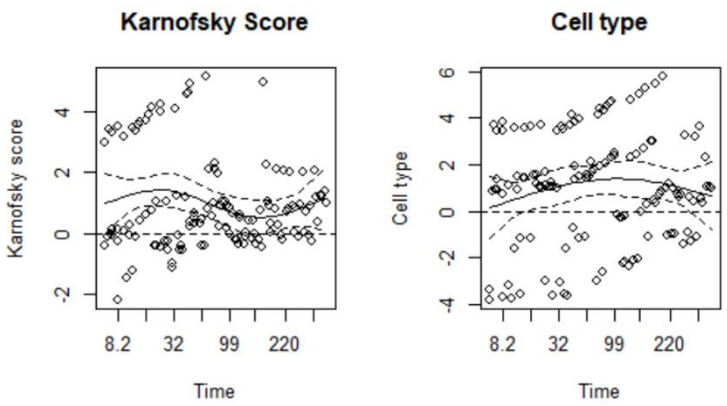


Figure 44: Plot of the scaled Schoenfeld residual plots

Formally test of the proportional hazards assumption for all variables in va_lung.cph01:

```
> cox.zph(va_lung.cph01, global = TRUE)
                     chisq df      p
Karnofsky.cat        16.08  3 0.0011
va_lung$"Cell type"  9.18  3 0.0270
GLOBAL               23.68  6 0.0006
```

Figure 45: test of the proportional hazards assumption

Note that Stevenson, M.[12] mention that "using the cox.zph function, rho is the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. The hypothesis of no correlation is tested using test statistic chisq." In the above example, the significant cox.zph test for Karnofsky.cat and cell type (p-value < 0.05) implies that the proportional hazards assumption has been violated for the Karnofsky.cat and cell type variables.

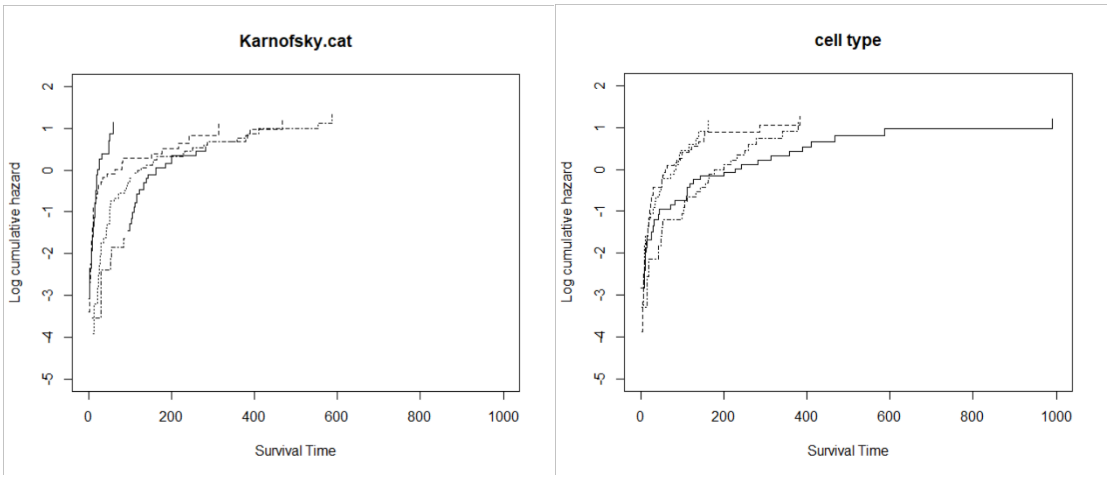


Figure 46: log-cumulative hazards vs. survival time for Karnofsky.cat and cell type.

According to Stevenson, M.[12], "an alternative (and less sensitive) means of testing the proportional hazards assumption is to plot log[-log S(t)] vs time," as discussed above. The log[-log S(t)] plots of the Karnofsky score and cell type vs. time are shown Figure 46 above. These plots obviously have each of the curves for each of the categories in each of the plots. There are crossovers observed between the curves. Thus, we can conclude that the -log[-log S(t)] vs time plots for Cell type are not parallel, not conflicting with the endings of the cox.zph test and the Schoenfeld residual plots.

Now, this violation of the PH assumption needs to be dealt with. Accordingly, "we can produce a separate baseline hazard function for each level of [Karnofsky score]. Note that by stratifying we cannot obtain a hazard ratio for Karnofsky score since the [`Karnofsky score effect'] is absorbed into the baseline hazard," according to Stevenson, M.[12]. Through the comparison of the stratified vs. the original Cox PH models, in Figure 47: the stratified model produces a statistically significant better fit than the original, because the result of comparing the test statistic with the chi-square value gives a p-value of 0.

```
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                 Karnofsky.cat + va_lung$"Cell type",
                 method = "breslow", data = va_lung)
va_lung.cph04 <- coxph(Surv(Survival, Status, type = "right") ~
                 strata(Karnofsky.cat) + va_lung$"Cell type",
                 method = "breslow", data = va_lung)
summary(va_lung.cph04)
Call:
coxph(formula = Surv(Survival, Status, type = "right") ~ strata(Karnofsky.cat) +
    va_lung$"Cell type", data = va_lung, method = "breslow")

  n= 137, number of events= 128

                       coef exp(coef) se(coef)     z Pr(>|z|)
va_lung$"Cell type"2 0.9050    2.4719   0.2697 3.356  0.00079 ***
va_lung$"Cell type"3 1.2234    3.3986   0.3057 4.002 6.27e-05 ***
va_lung$"Cell type"4 0.3478    1.4160   0.2854 1.219  0.22296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     exp(coef) exp(-coef) lower .95 upper .95
va_lung$"Cell type"2     2.472     0.4045    1.4572     4.193
va_lung$"Cell type"3     3.399     0.2942    1.8669     6.187
va_lung$"Cell type"4     1.416     0.7062    0.8093     2.478

Concordance= 0.612  (se = 0.029 )
Likelihood ratio test= 20.39  on 3 df,   p=1e-04
Wald test            = 19.5  on 3 df,   p=2e-04
Score (logrank) test = 20.6  on 3 df,   p=1e-04
> # Compare the original model with the stratfied model:
> x2 <- 2 * (va_lung.cph04$loglik[2] - va_lung.cph01$loglik[2])
> 1 - pchisq(x2, 1)
[1] 0
```

Figure 47: comparison of the stratified vs. the original Cox PH models

Parameterizing Karnofsky score as a time dependent covariate would be one option for dealing with non-proportionality of hazards and retaining the ability to quantify the effect of Karnofsky score. Below in Figure 48 is a plot of the Kaplan-Meier survival curves for each Karnofsky score, adjusting for the effect of cell type.

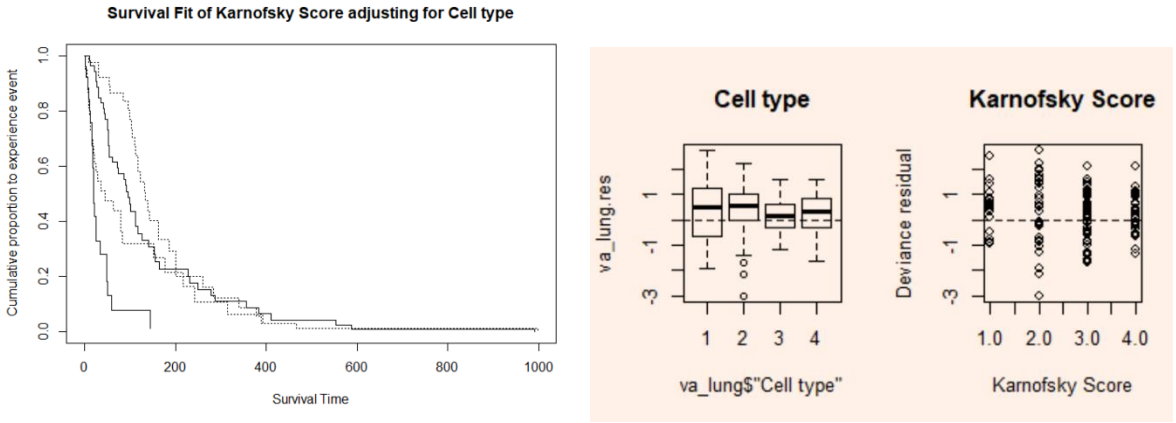


Figure 48: comparison of the stratified vs. the original Cox PH models

Now, a few of parametric models will be fit on the data. These models will be compared and interpreted with their AFT scores. The best model (whether it's Weibull or Cox PH) will be selected using the AIC value to compare the models and find the smallest measures. Here, the

Figures are in Table 1 below, we have the results for the Cox PH model, exponential, Weibull, lognormal, and log-logistic, in cells a, b, c, d, and e, respectively. In part (f) of Table 1, we see the AIC values for each of the five models. The AIC for the Cox PH model is the smallest, indicating that this model provides the best fit with the data.

```
> va_lung.cph04 <- coxph(Surv(Survival, Status, type = "right") ~
+                          strata(Karnofsky.cat) + va_lung$"Cell type",
+                        method = "breslow", data = va_lung)
> summary(va_lung.cph04)
Call:
coxph(formula = Surv(Survival, Status, type = "right") ~ strata(Karnofsky.cat) +

    va_lung$"Cell type", data = va_lung, method = "breslow")

  n= 137, number of events= 128

                      coef exp(coef) se(coef)     z Pr(>|z|)
va_lung$"Cell type"2 0.9050    2.4719   0.2697 3.356  0.00079 ***
va_lung$"Cell type"3 1.2234    3.3986   0.3057 4.002 6.27e-05 ***
va_lung$"Cell type"4 0.3478    1.4160   0.2854 1.219  0.22296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     exp(coef) exp(-coef) lower .95 upper .95
va_lung$"Cell type"2     2.472     0.4045    1.4572     4.193
va_lung$"Cell type"3     3.399     0.2942    1.8669     6.187
va_lung$"Cell type"4     1.416     0.7062    0.8093     2.478

Concordance= 0.612  (se = 0.029 )
Likelihood ratio test= 20.39  on 3 df,   p=1e-04
Wald test            = 19.5  on 3 df,   p=2e-04
Score (logrank) test = 20.6  on 3 df,   p=1e-04
```

a.

```
> #Exponential model:
> va_lung.exp01 <- survreg(Surv(Survival, Status, type = "right") ~
+                  Karnofsky.cat + va_lung$"Cell type",
+                  dist = "exp", data = va_lung)
> summary(va_lung.exp01)

Call:
survreg(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type", data = va_lung, dist = "exp")
                      Value Std. Error      z       p
(Intercept)          3.7069     0.3297 11.24 < 2e-16
Karnofsky.cat        0.5702     0.0988  5.77 7.8e-09
va_lung$"Cell type"2 -0.8312     0.2376 -3.50 0.00047
va_lung$"Cell type"3 -1.2608     0.2659 -4.74 2.1e-06
va_lung$"Cell type"4 -0.3566     0.2665 -1.34 0.18085

Scale fixed at 1

Exponential distribution
Loglik(model)= -718.6   Loglik(intercept only)= -751.2
        Chisq= 65.22 on 4 degrees of freedom, p= 2.3e-13
Number of Newton-Raphson Iterations: 5
n= 137

> shape.exp = 1/ va_lung.exp01$scale
> shape.exp
[1] 1
```

b.

```
> # Weibull model:
> va_lung.wei01 = survreg(Surv(Survival, Status, type = "right") ~
+                Karnofsky.cat + va_lung$"Cell type",
+                dist = "weib", data = va_lung)
> summary(va_lung.wei01)

Call:
survreg(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type", data = va_lung, dist = "weib")
                      Value Std. Error      z       p
(Intercept)          3.7528     0.3154 11.90 < 2e-16
Karnofsky.cat        0.5635     0.0938  6.01 1.9e-09
va_lung$"Cell type"2 -0.8314     0.2231 -3.73 0.00019
va_lung$"Cell type"3 -1.2771     0.2500 -5.11 3.3e-07
va_lung$"Cell type"4 -0.3676     0.2502 -1.47 0.14182
Log(scale)           -0.0641     0.0668 -0.96 0.33732

Scale= 0.938

Weibull distribution
Loglik(model)= -718.2   Loglik(intercept only)= -748.1
        Chisq= 59.85 on 4 degrees of freedom, p= 3.1e-12
Number of Newton-Raphson Iterations: 6
n= 137

> shape.wei = 1 / va_lung.wei01$scale
> shape.wei
[1] 1.066158
```

c.

```
> va_lung.lognorm01 = survreg(Surv(Survival, Status, type = "right") ~
+                         Karnofsky.cat + va_lung$"Cell type",
+                        dist = "lognormal", data = va_lung)
> summary(va_lung.lognorm01)

Call:
survreg(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type", data = va_lung, dist = "lognormal")
                      Value Std. Error      z       p
(Intercept)          2.6490     0.3229  8.20 2.3e-16
Karnofsky.cat        0.6836     0.0956  7.15 8.7e-13
va_lung$"Cell type"2 -0.6110     0.2508 -2.44  0.0149
va_lung$"Cell type"3 -0.7523     0.2869 -2.62  0.0087
va_lung$"Cell type"4  0.0776     0.2884  0.27  0.7878
Log(scale)           0.0962     0.0627  1.53  0.1248

Scale= 1.1

Log Normal distribution
Loglik(model)= -719.8   Loglik(intercept only)= -749.5
        Chisq= 59.35 on 4 degrees of freedom, p= 4e-12
Number of Newton-Raphson Iterations: 4
n= 137

> shape.lognorm = 1 / va_lung.lognorm01$scale
> shape.lognorm
[1] 0.9082912
```

d.

```
> # Loglogistic model
> va_lung.loglogistic01 = survreg(Surv(Survival, Status, type = "right") ~
+                              Karnofsky.cat + va_lung$"Cell type",
+                             dist = "loglogistic", data = va_lung)
> summary(va_lung.loglogistic01)

Call:
survreg(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
    va_lung$"Cell type", data = va_lung, dist = "loglogistic")
                      Value Std. Error      z       p
(Intercept)          2.8777     0.3250  8.86 < 2e-16
Karnofsky.cat        0.6469     0.0885  7.31 2.8e-13
va_lung$"Cell type"2 -0.7385     0.2545 -2.90  0.0037
va_lung$"Cell type"3 -0.8603     0.2753 -3.13  0.0018
va_lung$"Cell type"4 -0.0520     0.2782 -0.19  0.8518
Log(scale)           -0.5040     0.0739 -6.82 9.1e-12

Scale= 0.604

Log logistic distribution
Loglik(model)= -717.3   Loglik(intercept only)= -750.3
        Chisq= 65.98 on 4 degrees of freedom, p= 1.6e-13
Number of Newton-Raphson Iterations: 4
n= 137

> shape.loglogistic = 1 / va_lung.loglogistic01$scale
> shape.loglogistic
[1] 1.655352
```

e.

```
> #Compare the three models using AIC:
> extractAIC(va_lung.cph04)
[1]   3.0000 664.3012
> extractAIC(va_lung.exp01)
[1]    5.000 1447.221
> extractAIC(va_lung.wei01)
[1]    6.000 1448.331
> extractAIC(va_lung.lognorm01)
[1]    6.000 1451.596
> extractAIC(va_lung.loglogistic01)
[1]    6.000 1446.549
```

f.

Table 1:

Next, the AFT can be used to describe the various parametric models shown in Table 1 parts b, c, d, and e above, starting with the Weibull model. What is the effect of Cell type of 2 on

survival time (after adjusting for the effect of presence of Karnofsky.cat)? Here the psm function is used, in Figure 49, which is a function in the rms library, to develop an AFT model. The psm function is a modification of survreg and is used for fitting the accelerated failure time family of parametric survival models. A patient with a Cell type of 2 cuts the patient survival time in about half (by a scale of 0.44), according to Figure 49 above. What does this mean in terms of calendar time? In Figure 50 below, it can be concluded that patients with a Cell type of 2 remained on the cancer treatment program for about an additional 19 days. These patients are being compared with those treated in the other three levels of the cell type.

```
> library(rms)
> va_lung.aft01 <- psm(Surv(Survival, Status, type = "right") ~
+                            Karnofsky.cat + va_lung$"Cell type",
+                      dist = "weibull", data = va_lung)
> va_lung.aft01
Parametric Survival Model: Weibull Distribution

 psm(formula = Surv(Survival, Status, type = "right") ~ Karnofsky.cat +
     va_lung$"Cell type", data = va_lung, dist = "weibull")

                      Model Likelihood     Discrimination
                         Ratio Test           Indexes
 Obs         137    LR chi2      59.85    R2       0.354
 Events      128    d.f.             4    Dxy      0.436
 sigma 0.9379476    Pr(> chi2) <0.0001    g        0.880
                                          gr       2.410

               Coef    S.E.   Wald Z Pr(>|Z|)
 (Intercept)    3.7528 0.3154 11.90  <0.0001
 Karnofsky.cat  0.5635 0.0938  6.01  <0.0001
 va_lung=2     -0.8314 0.2231 -3.73  0.0002
 va_lung=3     -1.2771 0.2500 -5.11  <0.0001
 va_lung=4     -0.3676 0.2502 -1.47  0.1418
 Log(scale)    -0.0641 0.0668 -0.96  0.3373

> exp(va_lung.aft01$coefficients[3])
va_lung=2
0.4354304
```

Figure 49: VA lung aft model

```
> log.t <- as.numeric(va_lung.aft01$coefficients[1] +
+                       (va_lung.aft01$coefficients[3]* 1))
> exp(log.t)
[1] 18.56623
```

Figure 50: results of patients with a Cell type of 2

IV. Conclusions on Survival Analysis

Survival analysis is useful with time to event data, as in the VA lung cancer dataset's scenario. There are parametric models, such as the Weibull model, semi-parametric such as the Cox PH model, and non-parametric models, such as the Kaplan-Meier modeling approach to fit the time to event data and make inferences from it. Now, a different type of time-dependent models will be explored: Longitudinal Data models.

Longitudinal Data Analysis

- **Longitudinal and clustered data**

The main goal of performing a longitudinal data analysis study is to directly study change of the same individuals over time and to find the variables that influence that change in response over time. This is done by taking repeated measurements of the subjects[2]. Within-individual change can be captured due to the repeated measurements taken from individuals and modeled with a longitudinal model.

Longitudinal Data Analysis Is Different from Other Types Analysis Techniques

Longitudinal study design differs from a cross sectional study design. In a cross sectional study design the response is recorded at a single instance in time (that is, not over time) and does not "provide any information about how individuals change during the corresponding period," according to Fitzmaurice et. al[2], with a variety of other covariates, such as age, sex, BMI, genetics, etc. A longitudinal study will look at a group of similar individuals or example. According to Hand et al, "One of the attractive features of repeated measures data is that (for numerical data, at least) they can be displayed in a graphical plot which is readily interpretable, without requiring a great effort" (1). Longitudinal data analysis takes a look at how each individual's responses change over time, within in individual's experiences, not between multiple individuals' experiences[2]. Data is also different from time series data in that for longitudinal data, each unit has several observations recorded within a relatively short amount of time[5]. A time series looks at trends of a bunch of data collectively over a long period of time.

Hence, the data is often clustered in longitudinal data. According to Fitzmaurice et. al[2], "clusters are composed of the repeated measurements obtained from a single individual at different occasions." The observations in such a data set will often be positively correlated with each other on a variety of covariates. Such positive correlation will be accounted for in the longitudinal data analysis. Additionally, longitudinal data have a "***temporal order***". That is, second measurements within a cluster must come after the first measurement on an individual, third measurements within a cluster must come after the second measurement on an individual, etc. According to Fitzmaurice et. al[2], "we might reasonably expect that measurements on units within a cluster are more similar than the measurements on units indifferent clusters" (4). Correlation within the same cluster can express the degree of clustering. Because most statistical methods require the independence assumption to be met that there will be some dependence among the longitudinal data that's measured with the correlation measure, longitudinal models (as well as other statistical models), must show how the lack of independence is accounted for.

A few items need to be clarified before going into more detail. First, each observation collected from the subjects are all measuring the same items. Then, the measurements can be measured on different occasions than the other observations, even though the ideal situation occurs when all measurements are measured at the same time. Lastly, according to Hand et. al.[5], "the sequential nature of the observations means that particular kinds of covariance structures are likely to arise, unlike more general multivariate situations, where there may be few or no indications of the structure." The next section introduces how regression models are used on the modeling of longitudinal data.

- Regression for correlated responses

Linearity vs. Non-linearity

Regression models broadly refers "to any model that describes the dependence of the mean of a response variable on a set of covariates in terms of some model for a continuous response variable," according to Fitzmaurice et. al[2]. For the response variables that are not

simply continuous, such as binary response variables or categorical response variables, the logistic regression and Poisson or log-linear regression models can be used, respectively. ***Linearity*** has a very specific definition. When a model for the mean (or a transformed mean) is linear, it is linear in the regression parameters. That is, the conditional mean or expected value of Y, given X, E(Y|X), is linear in the regression parameters. It is equal to a linear combination of the covariates. For example, $\mathrm{E(Y|X)} = \beta_0 + \beta_1 \mathrm{X} + \beta_1 \mathrm{X}^2$, is an example of this. Any non-linear looking models are $E(Y|X) = \beta_0 + e^{\beta_1 X}$ or $E(Y|X) = \frac{\beta_0}{\beta_1 + e^{\beta_2 X}}$. In these cases, "the mean is non-linear in the regression parameters," according to Fitzmaurice et. al[2]. Any nonlinearity of the mean response $\mathrm{E(Y|X)}$ on the regression parameters can be transformed into a linear model.

Transformations to deal with non-linearity

Transformations such as the log transformation in the case of Poisson regression, a log transformation on the covariates, and/or by including polynomials of the covariates can be used on longitudinal models to aid with non-linearity issues. According to Fitzmaurice et. al[2], "The inclusion of transformed covariates in no way violates the "linearity" of the regression model; that is, the model is still linear in the regression parameters." Covariates in regression models can be categorical or numeric. In addition, according to Fitzmaurice et. al[2], "the mean response, or any suitable transformation of the mean, can be related to a continuous covariate in a curvilinear or non-linear fashion by simply taking an appropriate transformation of the covariate or by the inclusion of polynomials (e.g., time and time 2)."

A regression model can also contain interactions between numeric covariates and other numeric covariates, categorical covariates and other categorical covariates, and categorical covariates and numeric covariates, according to Fitzmaurice et. al[2]. Such covariances are displayed in a covariance matrix that is arbitrary and "allows individuals to be measured on different numbers of occasions," according to Hand et. al.[5]. So, in other words, individuals can have different covariate values and even if their measurements in the study are incomplete, the observation can still be used in the longitudinal data analysis model.

So, regression models will be discussed in further detail in this paper. A major perk of using regression models is that they "can often be used to distinguish within- and between-subject trends in the response (e.g., "longitudinal" versus "cross-sectional" effects of age)." There are a few other points to note about regression models. One of them is that regression models are typically created or written in such a way that the parameters of the regression model, according to Fitzmaurice et. al[2],

> "have interpretations that bear directly on the scientific question of main interest. For example, in a regression model for data from a longitudinal clinical trial, a particular regression coefficient can be given an interpretation in terms of the constant rate of change in the mean response over time in one of the treatment groups."

Another interesting regression model is when there is a continuous numeric response variable with only categorical covariates. This is called an analysis of variance (ANOVA) model.

- Basics of longitudinal data
    - Objectives and features

The development and persistence of disease can be understood through longitudinal models. A longitudinal study design allows the discovery of heterogeneity factors between the individuals in a data set. According to Fitzmaurice et. al[2],

> "The distinguishing feature of longitudinal studies is that the study participants are measured repeatedly throughout the duration of the study, thereby permitting the direct assessment of changes in the response variable over time. In cross-sectional studies, where measurements are obtained at only a single point in time, it is not possible to assess individual changes on the basis of a single snapshot of the individual's response taken at a given time."

Therefore, at least two observations taken at different times of the response variable, are made on a few or more individuals in the study. Sometimes longitudinal study designs require that the individuals in the study have a fixed number of repeated measurements on them during a set of common time points. Thus, as a result of this requirement, longitudinal data analysis can help find the "within-individual changes on the response variable," according to Fitzmaurice et. al[2].

Mainly, the objective of longitudinal data analysis is to "describe trends in these within-individual changes in the response and to relate these changes to selected covariates (e.g. treatment group)," according Fitzmaurice et. al[2]. That is, the behavior of the individuals in the dataset should be quantified and described such that the individual can be compared to itself in response to some treatment variable categories[5]. Such comparisons are called the mentioned within-individual changes and have some specific methods to modeling and analyzing them.

Such within-individual changes extend to the general "response trajectories" as time increases, from the more specific "difference scores" measured as time increases, which is proportional to a constant rate of change/slope, of a linear response trajectory. To smooth out and summarize these within-individual changes in the response over time of the study, curvilinear and piecewise trajectories, and other response trajectories can be used. There are two steps of modeling with-individual changes with longitudinal data analysis methods. Step number one is the within-individual change is summarized based on repeated measures over the study time that the individual experiences. Difference scores or response trajectory methods can be used to provide this summary. Step number two is to take the estimates of within-individual changes and compare or relate them to "inter-individual difference in selected covariates," according to Fitzmaurice et. al[2]. There is an easier way to show these two steps[2]; a single longitudinal model can be used to model both steps at the same time.

According to Hand et. al.[5], a model that fits well over a timed sequence of responses is called response feature analysis, which is based on analyzing response features. A "response feature[s] … summarizes some aspect of the response over time." For example, the difference between the mean pre- (maybe day 0 of pre-) and post- (maybe day 0 of post-) treatment scores

for the same time units can be considered as a response feature. Although the main disadvantage of feature analysis is a loss of power or degrees of freedom, this method is flexible. That is, even if the individuals in the study have their measurements taken at different times, or are missing some measurements, the feature analysis can still produce summarizing features that are relatively easily interpretable simple univariate analyses[5].

Intro to Missing Data Issues

Missing data is a big deal in longitudinal data analysis. The ideal data set is a balanced data set, with all measurements properly taken a recorded for each of the individuals in the longitudinal study. If the mean response of each individual is plotted at each of the n occasions, then differences in mean response time between each occasion of the individuals can be calculated. The differences in mean response time can be calculated by subtracting the current from the prior means calculated for each of the occasions.

However, a quite common and a major problem in the collection and modeling of longitudinal data is when a subject misses one or more measurements in the repeated measurement schedule, and if this happens with multiple individuals. This leaves the longitudinal data set and resulting analysis unbalanced over time. According to Fitzmaurice et. al[2],

> "One of the consequences of lack of balance and/or missing data is that it requires some care to recover within-individual change…When data are missing,…then a plot of the mean response over time can be mis-leading; changes over time may reflect the pattern of missingness or the attrition, and not within-individual change" (24).

The situation is even worse when a subject has missing data and apparently has different responses from the other subjects who do not have missing data and/or remain in the longitudinal study. There are a few remedies for individuals with missing data, such as imputation.

Common Notation of longitudinal data and models

The common notation for longitudinal data with i = 1 ,…, N rows of individual observations at the jth occasion (a.k.a., the repeated measurements), where j = 1, …., n, for balanced data needs to be described. It is assumed, in this ideal model, that the repeated measures are taken at the same time intervals for each of the individuals. Random variable Y_{ij} has an observed value y_{ij}. According to Fitzmaurice et. al[2], "Given that we have n repeated measures of the response variable on the same individual, we can group these into a nx1 response vector, denoted by $Y_i = \begin{pmatrix} Y_{i1} \\ \dots \\ Y_{in} \end{pmatrix} = (Y_{i1} \quad \dots \quad Y_{in})^{T}$". The true mean response μ_{ij} is found by taking the expected value of the response vector, or the long-run or weighted average, Y_i: $\mu_{ij} = E(Y_{ij})$.

The conditional mean response at the jth repeated measure is what μ_{ij} is also commonly referred to as. The conditionality here means that the mean response is dependent on the

covariates in the model. Two variables are dependent if the conditional distributions of one of them depends on the other. According to Fitzmaurice et. al[2], "two variables are said to be independent if the conditional distribution of one of them does not depend on the other" (27). When variables are dependent, "the response of an individual on one occasion is very likely to be predictive of the response of the same individual at a future occasion."

Dependence can also be measured between two variables with correlation, which recall is dependent on variance and covariance. According to Fitzmaurice et. al[2], "the conditional variance of Y_{ij} is $\sigma_j^2 = E\{(Y_{ij} - E(Y_{ij}))\}^2 = E(Y_{ij} - \mu_{ij})^2$." The denoted σ_j is the conditional standard deviation for random variable Y_{ij}. This measure is not to be confused with the conditional covariance, $\sigma_{jk} = E\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}$, where Y_{ij} and Y_{ik} are separate occasion. This measure described can be either positive or negative and is a "measure of the linear dependence between Y_{ij} and Y_{ik}." All of the variances and covariations between each of the n repeated measurements in the Y_i dataset can be displayed in a variance-covariance matrix (a.k.a. the var-cov matrix) in equation 32, where $\sigma_{n1} = \sigma_{1n}$, by symmetry of the matrix.

$$Cov\left(\begin{pmatrix} Y_{i1} \\ \dots \\ Y_{in} \end{pmatrix}\right) = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix} \text{ (equation 32)}$$

Conditional correlation can be used as a measure, without any variability or units of measurement. This occurs because the two random variables are dependent on the covariance between and the standard deviations of the two random variables. The formula for the correlation is shown in equation 33.

$$\rho_{jk} = \frac{E\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}}{\sigma_j \sigma_k} = \frac{\sigma_{jk}}{\sigma_j \sigma_k} \text{ (equation 33)}$$

This measure is between the values of -1 and 1, where -1 is perfect negative correlation (a perfectly negative line going through data points in the coordinate plane), 0 is no correlation (a random scattering of points on the plane), and 1 is perfect positive correlation (a perfectly positive line going through data points in the coordinate plane). Although, in the case of longitudinal data, the values of ρ_{jk} are typically between 0 and 1 because the observations will only be positively associated or not associated with each other. The correlation matrix takes the above var-cov matrix and divides it me the $\sigma_j \sigma_k$ of each entry. After doing so, the resulting matrix is shown in equation 34, where $\rho_{n1} = \rho_{1n}$, according to Fitzmaurice et. al[2].

$$Corr\left(\begin{pmatrix} Y_{i1} \\ \dots \\ Y_{in} \end{pmatrix}\right) = \begin{bmatrix} 1 & \cdots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \cdots & 1 \end{bmatrix} \text{ (equation 34)}$$

Sources of correlation

In longitudinal data analysis, the data is positively correlated; the correlation measurements decrease as the time intervals increase in length; between repeated measurements of an individual, the correlations don't typically approach zero, even in cases with long time intervals between measurements of responses; and when the time interval between repeated measurements is small, the correlation between those two repeated measures rarely is near one[2]. There are three possible sources of variability impacting the correlation measure between an individual's repeated measurements. They are between-individual heterogeneity, within-individual biological variation, and measurement error.

Between-individual heterogeneity, as shown in Figure 51 part (a), occurs when some individuals in a longitudinal dataset response consistently higher or lower on average than the others in the data set. In other words, the covariates are being used to draw any distinctions between individuals[5]. According to Fitzmaurice et. al[2], "The central idea that has been introduced here is that each individual's underlying propensity to respond-whether it be "high," "medium," or "low," and whether it be due to genetic, environmental, social, or behavioral factors (or some combination of these factors)-is shared by all of the repeated measures obtained on that individual." Thus, an observation with a mean response type ("high," "medium," or "low") will have their subsequent repeated responses be similar to the previous response types.

When a treatment or intervention is given to an individual, it is predicted to lead to an improvement in the response variable. As a consequence, the different individuals in the dataset will show "different gains over time," according to Fitzmaurice et. al[2] in their response trajectories. The remedy to this issue of between-individual heterogeneity is with individual-specific random effects, where some of the covariates or effect's coefficients in the regression equations are assumed to vary randomly.

In contrast, ***within-individual biological variation*** is when genetic biological variability of health outcomes of an individual "is an important source of variability that has an impact on the correlation among longitudinal responses," according to Fitzmaurice et. al[2]. That is, "the pattern of change over occasion – the profile of expected scores" is being considered for each individual, according to Hand et. al.[5]. See Figure 51 part (b) for a generic visualization of within-individual biological variation. On some individual, a sequence of repeated measures "will vary around some homeostatic set point in a random manner," according to Fitzmaurice et. al[2]. Inherent within-individual biological variability is what such variability is named and is apparent in most measured biological parameters, such as blood pressure, heart rate, and serum cholesterol.

Such parameters represent biological processes, or interactions between such processes, that go up and down in cycles over time. Thus, when the time period between repeated measurements is small, the individual will have very similar measurements of deviation for those responses that are shown by observing the response trajectories. In this case of within-individual biological variation, "successive random deviations cannot be assumed to be independent of one

another" and the measurements taken close together will be highly positively correlated, according to Fitzmaurice et. al[2]. Any misspecification of response trajectories between individuals with different response trajectories over time will increase each of the individual's in the datasets response trajectory over time.

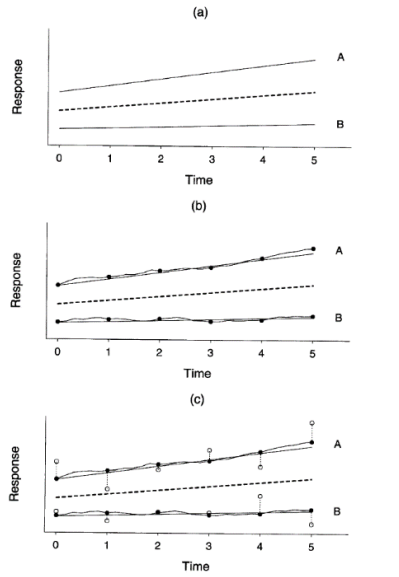


Fig. 2.4 Graphical representation of the cumulative impact of three sources of variability in longitudinal data: (a) between-individual heterogeneity, (b) within-individual biological variation (where • denotes repeated measure free of measurement error), and (c) measurement error (where ○ denotes observed repeated measure with measurement error).

Figure 51: Graphical representation of different variability in longitudinal data

Measurement error (represented by a generic plotted data set in Figure 51 (c)) is typically random in nature and is the most common and ubiquitous type of error on all data collection methods. According to Fitzmaurice et. al[2], "Although this source of variability can account for some of the within-subject variation in many health outcomes, it should not be confused with the inherent (within-individual) biological variability of these outcomes." Notice the difference between Figure 51 part (b) and (c) below. In part (c), there are hollow points that denote observed response measures that do not align with the expected black dots. In part (b) of Figure 51, the perfectly straight line does not match up perfectly with the line that shows the random effects due to biological generic factors that affect the response measurement.

The measurement procedure may have random error on some individual, and the response measurements taken at the same time to rule out the potential inherent biological variability. These measurements shouldn't be assumed to have the same values. Reliability is a coefficient that refers to the precision of the (repeated) measurement procedure, taken under the same conditions. Fitzmaurice et. al[2] says

> "The statistical definition of reliability then expresses the relative magnitude of the variability of the true scores to the overall variability of the data. That is, reliability is defined as the proportion of the total or overall variability that is due to individual-to-individual variability in the true scores."

Precision is often measured with the standard error of measurement, or the variance of the measurement errors. Any unreliability shrinks "the correlation among the repeated measures closer to zero."

- Linear models
    - Notation and Distributional assumptions

The notation of the linear models for longitudinal data is for the n repeated measurements, or the n_i observed responses, for the ith individual in the dataset. The $n_i \leq n$. According to Fitzmaurice et. al[2],

> "In addition to missing data, there may be mistimed measurements, in the sense that measurements are not obtained at the planned n occasions; instead, they are obtained some time before or after the intended measurement occasions. Thus both the number and the timing of the repeated measurements may not be common for all subjects" (50).

The vector of the time-ordered responses, with the n_i repeated measures of the response variable for the N independent individuals, from above is shown in equation 35 below.

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}_{n_i * 1}, i = 1, \dots, N \text{ (equation 35)}$$

Although unlikely, the number of repeated measures might be the same for all the individuals in the study and there might be no missing values in the dataset. When this occurs, "it is not necessary to include the index i in n_i (since $n_i = n$ for i = 1,…,N)," according to Fitzmaurice et. al[2]. Such as response vector is related to a vector of covariates, as will be shown next.

A px1 vector of covariates, for each response Y_{ij} at the ith subject and jth occasion, X_{ij}, for i =1,….,N; j=1,…, n_i, is displayed in equation 36 below:

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ \vdots \\ X_{ijp} \end{pmatrix}_{p*1} \text{ (equation 36)}$$

The p rows in this covariate vector correspond to each of the different covariates. So, according to Fitzmaurice et. al[2], "X_{i1} is a p x 1 vector whose elements are the covariate values associated with the response variable for the ith subject at the 1st measurement occasion, X_{i2} is a p x 1 vector whose elements are the covariate values associated with the response variable for the ith subject at the 2nd measurement occasion, and so on." This vector might contain covariates that do not change during the study or covariates that change over time. All of these covariate vectors can be grouped into an $n_i * p$ matrix, X_i. Such a matrix is denoted as in equation 37.

$$X_i = \begin{pmatrix} {X_{i1}}^T \\ \dots \\ {X_{in_i}}^T \end{pmatrix}_{n_i * p} = \begin{bmatrix} X_{i11} & \cdots & X_{i1p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{bmatrix}_{n_i * p} \text{ (equation 37)}$$

The rows of this matrix are the covariates associated with the responses at each of the different n_i repeated measurement occasions. The columns of this matrix are the p distinct covariates.

The regression model relating Y_i to X_i is denoted as in equation 38, where $\beta = (\beta_1, \dots, \beta_p)^T$ is a px1 vector of the unknown regression parameters.

$$Y_i = \begin{pmatrix} Y_{i1} \\ \dots \\ Y_{in_i} \end{pmatrix}_{n_i \times 1} = X_i\beta + \varepsilon_i = \begin{bmatrix} X_{i11} & \cdots & X_{i1p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{bmatrix}_{n_i \times p} \begin{pmatrix} \beta_1 \\ \dots \\ \beta_p \end{pmatrix}_{p \times 1} + \begin{pmatrix} \varepsilon_{i1} \\ \cdots \\ \varepsilon_{in_i} \end{pmatrix}_{n_i \times 1} \quad \text{(equation 38)}$$

The $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})^T$ random vector of errors that are "associated with the corresponding elements of the vector of responses on the ith subject" and has an $n_i \times 1$ dimension. The random vector of errors is assumed to be distributed as $\varepsilon_i \sim N(0, \Sigma_{n \times i})$, with $\Sigma_{n \times i}$ a submatrix of the unstructured covariance matrix Σ_n (defined as discussed below) for all n potential repeated measurement occasions[5].

Next, ***the joint multivariate normal distribution*** (or ***MVN***) is quite useful in the modeling to longitudinal data. Fitzmaurice et. al[2] said that "with n_i repeated measures on the ith individual, we have a vector of responses and need to consider their joint probability distribution." A MVN distribution is a solution to modeling longitudinal data because it is a "natural *extension* of the ***univariate normal distribution*** for a single response to a vector of responses." Such a distribution has a density denoted as in equation 39.

$$f(y_i) = f(y_{i1}, \dots, y_{in_i}) = (2\pi)^{-n_i/2}|\Sigma_i|^{-1/2}\, e^{\frac{-1}{2}(y_i-\mu_i)^T\Sigma_i^{-1}(y_i-\mu_i)} \quad \text{(equation 39)}$$

The notation here means that Y_i is given X_i, with $\mu_i = E(Y_i) = E(Y_i|X_i) = (\mu_{i1}, \dots, \mu_{in_i})^T$, and $|\Sigma_i|$ is the determinant of Σ_i, or a.k.a. the generalized variance. In addition, $j = 1, \dots, n_i$ and $-\infty < y_i < \infty$.

A major assumption of the MVN distribution is normality: "the association between any pair of responses is linear," according to Fitzmaurice et. al[2]. Independence of the errors or residuals from the others and equal variance of the residuals are also important assumptions. According to Fitzmaurice et. al[2],

> "in summary, in longitudinal studies the repeated measurements on the same individual are inherently dependent or correlated. This lack of independence can be accounted for by considering the multivariate distribution of the entire vector of repeated measurements (given the covariates)."

Repeated measurements are correlated, but the observations (each of the rows of the data vector) or individuals in the dataset are all independent of each other.

- Descriptive methods of analysis

Longitudinal data models can be shown visually with the use of time plots, with the measurement times for each individual on the x-axis and the responses on the y-axis2. When multiple individual's time plots are placed on the same plot, the data points will likely overlap, sometimes making the plot a hard to read and interpret, especially for one individual. Fitzmaurice et. al[2] noted that "the most extreme case of this problem arises when the response variable is binary; then it is impossible to discern any information about time trends from the resulting time plot due to the completely overlapping data points."

Generally, time plots that display the mean response are more informative, with all plotted points connected together with straight lines. Such time plots can be "enhanced by including standard error bars for the mean response at each occasion," according to Fitzmaurice et. al[2]. Also, when a covariate is numerical, at least two reference categories should be chosen (such as low, medium, or high) and the numeric observations for that covariate are placed into one of the reference categories bins.

When the data is unbalanced (which is typically the case; see the section about missing values below) and a time plot is wanted, the data can be smoothed with ***smoothing techniques***. Such techniques need a smoothing or bandwidth parameter that, according to Fitzmaurice et. al[2], "controls the amount of smoothing." When choosing such a parameter, there is a tradeoff that comes into play for each option available: the tradeoff between precision (the amount or level of variance) and bias. According to Fitzmaurice et. al[2], "All smoothing techniques must compromise in some way, and the goal is to find an appropriate trade-off between these two competing forces: increased bias versus decreased variance of the estimated mean response trend over time." There are two types of such smoothing techniques and smoothing parameters for each type: parametric methods that deal with balanced data and nonparametric data that deal with unbalanced data.

An example of a parametric smoothing technique is the running average or ***moving average***. This measure is used to determine the "systematic neighborhood of values used to estimate the mean response at time t," according to Fitzmaurice et. al[2]. This is done by observing the order of the running average measurement. When the moving average's order is higher, the more smoothed out and the less wiggly the mean and trend. On the contrary, When the moving average's order is lower, the less smoothed out and the wigglier the mean and trend.

The moving average (S_t) is defined as in equation 40, when the longitudinal data is complete and balanced.

$$S_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=-k}^{k} w_j y_{i,t+j} \text{ (equation 40)}$$

The t denotes any time, k is a positive integer, with the order of the moving average represented by 2k+1, and w_j is a set of weights, where $\sum_{j=-k}^{k} w_j = 1$ and $w_j > 0$. According to Fitzmaurice et. al[2], when the weights are not equal, "they are chosen so that they decrease symmetrically about some maximum value." So, symbolically, $w_j = w_{-j}$ and $w_0 > w_1 > \cdots > w_k$. Therefore, the observations that are obtained at nearly the same or at a nearby time will have the "greatest impact or 'weight' in the calculation of the mean or average response at time t," according to Fitzmaurice et. al[2]. As a result, moving averages work best with observations in a dataset that are nearly equally separated over the time span of the study, as the best smoothing occurs in this scenario.

When data is unbalanced over time and/or collected in irregular time intervals, a nonparametric method called ***locally weighted regression*** (or ***lowess***) is employed. This method is

similar to the parametric methods, such as the common moving average approach, in that they "attempt to trace the salient features of the mean response as a function of time while making only minimal assumptions about the form of the relationship," according to Fitzmaurice et. al[2]. The main difference in the case of lowess in comparison to moving average method is that it uses a fitted straight line on the data within a window centered at time t. In other words, lowess uses "a robust regression technique that gives more weight to observations close to the center of the window and that also down-weights potential outliers" (70) and doesn't use the simple weighted average of observations in the window.

- Modeling the mean and covariates

Substantive vs. nuisance parameters

There are two types of parameters that need to be distinguished from each other that are important in the modeling of mean and covariates: ***substantive parameters*** and ***nuisance parameters***. Substantive parameters are regression parameters (β) that "summarize important aspects of the research questions," according to Fitzmaurice et. al[2]. For example, the β are the change in the mean response over time for a covariate. Parameters that summarize secondary parts of the regression model, for example the summary of covariance and correlation, are known to be nuisance parameters. They are parameters that the researcher is likely to not have any intrinsic interest in them in regards to helping answers the research question.

Modeling Mean Response over time

Now, when modeling the mean response over time, two broad approaches need to be assessed to find the right fit for the research question***: response profiles*** and ***parametric or semi-parametric curves.*** Response profile analysis "allows arbitrary patterns in the mean response over time," according to Fitzmaurice et. al[2]. In this analysis type, there is no assumption of a specific time trend and measurement times are regarded as levels of a discrete factor. When observations in the data are "measured at the same set of occasions and the number of occasions is usually small," response profile analysis will be the appropriate approach to modeling the mean over time, according to Fitzmaurice et. al[2].

Parametric or semi-parametric curves can be used as an analysis method to model the mean response over time. A parametric curve (such as a linear or quadratic trend) or a semi-parametric curve (such as a piece-wise linear) can be used. With these methods, there is no need for the observations in the datasets to have their measurement times happen at the same time or to have the same count of repeated measurements[2].

Types of Covariance

Although the covariance (which the correlation is based off) of each of the repeated measure's responses in the longitudinal data is not typically of intrinsic interest, such a statistic must not be ignored or included in the modelling procedure. When correlation (and thus covariance) is accounted for in the longitudinal regression model for the repeated measures, the

efficiency or the precision is increased and the regression parameters can therefore be estimated in the output of such a model. A longitudinal regression model with covariance/correlation accounted for (even in cases where there is missing data) produces correct standard errors and the regression parameters can have valid inferences made about them from the investigator[2].

Unstructured covariance, covariance pattern models, and random effects covariance structures are three approaches that can be applied to modeling covariance among repeated measures. ***Unstructured covariance*** occurs when an "arbitrary pattern of covariance among the repeated measures" is allowed to exist among the data, according to Fitzmaurice et. al[2]. There are $n \times (n-1)/2$ pairwise covariances/correlations that can be estimated among the n repeated measures. This method is the preferred method. ***Covariance pattern models*** occur when the "modeling the covariance place structure on the covariance matrix" is employed, according to Fitzmaurice et. al[2]. Finally, the ***random effects covariance structure*** occurs where "the correlation among repeated measurements is accounted for by the inclusion of a single individual specific random effect," according to Fitzmaurice et. al[2]. This is done by using the univariate repeated measures ANOVA model. Historically, the latter-most method has been the method that has been used. See the covariance modeling section below for more details on the various covariance pattern models.

- Estimation and Statistical Inference

A generalized linear regression model for the mean response vector is how all of the models for longitudinal data will be expressed from now on. That is, the expression will be $E(Y_i|X_i) = X_i\beta$. According to according to Fitzmaurice et. al[2], "the response vector, Y_i, is assumed to have a conditional distribution that is multivariate normal with covariance matrix $Cov(Y_i|X_i) = \Sigma_i(\theta) = \Sigma_i$." The θ is a q×1 covariance parameter vector. Both θ and β are assumed to be unknown parameters that must be estimated.

  - Maximum Likelihood estimation

The maximization of the likelihoods of θ and β is a common way to estimate θ and c. This method is called the Maximum Likelihood (ML) estimation. According to Fitzmaurice et. al[2], "the fundamental idea behind ML estimation is really quite simple and is conveyed by its name: use the estimates of the θ and β the values that are most probable (or most 'likely') for the data that have actually been observed." The likelihood function that is maximized is the joint probability of the response variates which are evaluated at each of the observed values in a fixed set of values. This function is a function of β and Σ_i. The maximum likelihood estimates will be denoted as $\hat{\beta}$ and $\hat{\Sigma}_i = \Sigma_i(\hat{\theta})$.

The observations in the data will all be independent of each other, collected through a "series of cross-sectional studies that are repeated at n occasions," according to Fitzmaurice et. al[2]. The covariates of the model and the mean response are related via the linear regression model denoted as: $E(Y_{ij}|X_{ij}) = {X_{ij}}^T\beta$. The values of the parameters in this regression equation that maximize he joint normal probability function of all of the observations in the data set is the

wat to find the maximum likelihood estimates of the vector β. The univariate normal density function is $f(y_{ij}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2}(y_{ij}-\mu_{ij})^2/\sigma^2}$, with $-\infty < y_{ij} < \infty$. With the independence assumption of the observations satisfied, the likelihood function of the univariate normal density is $\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij})$. The log-likelihood function l is more commonly used to find the MLEs. The equation for the univariate normal density is as follows in equation 41.

$$l = \log\{\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij})\} = -\frac{K}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - {X_{ij}}^T\beta)^2/\sigma^2 \text{ (equation 41)}$$

According to Fitzmaurice et. al[2], in equation 41, l "will be evaluated at the observed numerical values of the data, with respect to the regression parameters, β" (91). In l, the value K is the total number of observations ($K = n \times N$). According to Fitzmaurice et. al[2], "obtaining the maximum likelihood estimate of β is equivalent to finding the ordinary least squares (OLS) estimate of β; that is, the value of β that minimizes the sum of the squares of the residuals" (91). Such a solution for the MLE of β is in equation 42.

$$\hat{\beta} = \{\sum_{i=1}^{N}\sum_{j=1}^{n}(X_{ij}{X_{ij}}^T)\}^{-1}\{\sum_{i=1}^{N}\sum_{j=1}^{n}(X_{ij}y_{ij})\} \text{ (equation 42)}$$

The above explanations for the univariate method of finding the MLE can be extended to correlated observations and multivariate data, such as in the case of longitudinal data. In longitudinal data, the n_i repeated measures taken on each individual should not be assumed to be independent of each other. "Thus the log-likelihood function, l, can be expressed as a sum of the individual multivariate normal probability density functions for Y_i given X_i," according to Fitzmaurice et. al[2]. The log-likelihood for such longitudinal data is expressed as in equation 43.

$$l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{n}\log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)^T{\Sigma_i}^{-1}(y_i - X_i\beta) \text{ (equation 43)}$$

This is the expression that will be maximized to find the MLEs. A solution for the MLE of β is shown in equation 44.

$$\hat{\beta} = \{\sum_{i=1}^{N}({X_i}^T{\Sigma_i}^{-1}X_i)\}^{-1}\sum_{i=1}^{N}({X_i}^T{\Sigma_i}^{-1}y_i) \text{ (equation 44)}$$

Some notes on $\hat{\beta}$ are that is an unbiased and consistent estimate for β; "that it provides a valid estimate
of β even when the multivariate normal distribution assumption does not hold" (Fitzmaurice 93). Although Σ_i is assumed to be known, it rarely will be known. According to Ftizmaurice,

> "Once the ML estimate of θ has been obtained [via an estimate from the data through a computer algorithm], we then simply substitute the estimate of $\Sigma_i(\theta)$, say $\hat{\Sigma}_i = \Sigma_i(\hat{\theta})$ into the generalized least squares estimator of β given by (4.4) to obtain the maximum likelihood (ML) estimate of β" (93).

Note that Σ_i will likely be biased in cases with finite samples of data. This can be taken into consideration through the use of Restricted maximum Likelihood (REML) Estimation (Fitzmaurice 101). Such an Generalized Least Squares (GLS) estimator is denoted as in equation 45.

$$\hat{\beta} = \left\{\sum_{i=1}^{N}({X_i}^T{\hat{\Sigma}_i}^{-1}X_i)\right\}^{-1}\sum_{i=1}^{N}({X_i}^T{\hat{\Sigma}_i}^{-1}y_i) \text{ (equation 45)}$$

- Statistical inference

Confidence intervals and ***hypothesis tests*** are the common ways to make inferences about β. This GLS estimator $\hat{\beta}$ has an estimated covariance matrix, according to Hand et. al.[5]denoted as in equation 46, with the test statistic for the given hypothesis, given a contrast $r \times p$ matrix C.

$$\hat{V}_\beta = \widehat{Cov}(\hat{\beta}) = \left\{\sum_{i=1}^{N} (X_i{}^T \hat{\Sigma}_i{}^{-1} X_i)\right\}^{-1} \text{ (equation 46)}$$

The $C_{r\times p}\beta_{p\times 1} = c_{r\times 1}$, is $\left(C\hat{\beta} - c\right)^T \left(C\hat{V}_\beta C^T\right)^{-1} \left(C\hat{\beta} - c\right) \dot{\sim} \chi_r^2$ (57).

The standard error of the estimated single component of β, $\hat{\beta}_k$, is the "square-root of the diagonal element of $\hat{V}_\beta$ corresponding to $\hat{\beta}_k$," according to Fitzmaurice et. al[2] The standard error is expressed as $\sqrt{\widehat{Var}(\hat{\beta}_k)}$. Now, the test statistic of the test of the null hypothesis of the true value of the single component of β, β_k, is equal to zero ($H_0: \beta_k = 0$), versus the alternative hypothesis that β_k is not equal to zero ($H_a: \beta_k \neq 0$) is the Wald Statistic as shown in equation 47.

$$Z = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \text{ (equation 47)}$$

This wald test statistic Z is compared with the standard normal distribution value (two-sided or one-sided), depending on the significance level, such as 1.96 for the 5% significance level (in two sided case). 95% confidence intervals of the β_k can be obtained simply by the equation $\hat{\beta}_k \pm 1.96 * \sqrt{\widehat{Var}(\hat{\beta}_k)}$.

Next, confidence intervals and hypothesis tests can be used to test whether certain linear combinations, or ***contrasts***, or the components of the β, such as $[\beta_1, \beta_2, \beta_3]^T$ are related to or equal. Perhaps the vector of the known wrights to test, C, is defined, for example, as $[0, 1, -1]$. Then, the null hypothesis $H_0: C\beta = 0 = [\beta_1, \beta_2, \beta_3]^T [0, 1, -1] = \beta_2 - \beta_3$ can be tested vs. $H_A: C\beta \neq 0$. The ***Wald test statistic*** is denoted as in equation 48.

$$Z = \frac{C\hat{\beta}}{\sqrt{C * \widehat{Cov}(\hat{\beta}_k) * C^T}} \text{ (equation 48)}$$

Compared to a value from the standard normal distribution. An equivalent wald hypothesis test statistic that "helps to motivate how the Wald test is readily generalized when C has more than one row, thereby allowing the simultaneous testing of a single multivariate hypothesis," according to Fitzmaurice et. al[2], is denoted in equation 49.

$$W^2 = \left(C\hat{\beta}\right)^T \left\{C * \widehat{Cov}\left(\hat{\beta}_k\right) * C^T\right\}^{-1} \left(C\hat{\beta}\right) \sim \chi^2_{df=r(=1\ in\ this\ case)} \text{ (equation 49)}$$

The C could be a matrix, such as C= $\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$, so the $C\beta = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \end{bmatrix}$, and the null hypothesis would therefore be: $H_0: C\beta = 0$ or $H_0: \beta_1 = \beta_2 = \beta_3$, with the $H_A: C\beta \neq 0$ or

H_A: $At\ least\ one\ \beta_k$ is not equal to the others. Lastly, the 95% confidence interval for $C\beta$ will be expressed as in equation 50.

$$C\hat{\beta} \pm 1.96 * \sqrt{C * \widehat{Cov}(\hat{\beta}_k) * C^T} \text{ (equation 50)}$$

The Wald test's alternative is called the likelihood ratio test, which compares the maximized log-likelihood values for two nested models. One of the nested models is the full model, whereas the other model is the reduced, with the reduced model not including one or more of the β_k components of β that is being tested. Let the likelihoods of the full and reduced models be denoted as $\hat{l}_F$ and $\hat{l}_R$ respectively. The test statistic for the likelihood ratio test is $G^2 = 2 * (\hat{l}_F - \hat{l}_R)$ and compared with $\chi^2_{df=(number\ of\ parameters\ in\ full\ model - number\ parameters\ in the\ reduced\ model}$. Some other actions that can be completed with the likelihood ratio test are construction of the likelihood-based confidence intervals (through the use of a profile log-likelihood, $l_p(\beta_k)$, with test statistic in equation 51, and hypothesis testing about the covariance parameters.

$$2 * \left(l_p(\hat{\beta}_k) - l_p(\beta_k)\right) \sim \chi^2_{df=1}) \text{ (equation 51)}$$

Also, when the response variable is discrete, the likelihood ratio tests are more useful and advantageous than using the Wald test[2].

- Mean Response Modeling

The mean response in longitudinal data can be modeled through the use of response profiles. The ideal situation for modeling the mean response occurs when there is one covariate predictor variable that is categorical or discrete and is predicting the numeric response variable. There should be "no specific a priori pattern for the differences in the response profiles between groups can be specified" in such an ideal situation, according to Fitzmaurice et. al[2]

The sequence of the mean responses over time is called a response profile and the analysis of such a graph is called profile analysis. Recall that the main goal of profile analysis is to determine if the patterns of change in the mean responses are over time across the different groups of the categories in the predictive covariate. A variety of hypothesis tests can be formed to test the mean responses over time are different from one another Fitzmaurice et. al[2].

There are three main questions that can be explored with the use of profile analysis on the mean response. These questions are the following:

1. Are the mean response profiles among the different groups parallel or not? If they are, then that tells us that the responses are similar among the different categories.
    a. "This is a question that concerns the group × time interaction effect," according to Fitzmaurice et. al[2]
    b. See Figure 52 below, part (a) for a visualization of the parallel profiles of mean response.

c. This is the most common question in analysis in scientific research questions in context.
    i. Most common in observation studies and randomized trial experiments as the primary question of interest[2].
d. The null and alternative hypotheses:
    i. There are G ≥ 2 groups in the categorical/discrete predictive covariate.
    ii. The analysis of the mean response *profiles* of comparing two groups, such as the treatment and the control group using the notation for the mean response profile, taken at n repeated measurement times, of the treatment group and the control group, respectively $\mu(g) = \{\mu_1(g), \dots, \mu_n(g)\}^T$, with g = 1,…,G.
    iii. The mean difference in responses between the treatment and control groups is denoted as $\Delta_j(g) = \mu_j(g) - \mu_j(g)$, for j = 1,…, n and g = 1,…,(G-1).
    iv. The null hypothesis is $H_{01}: \Delta_1(g) = \cdots = \Delta_n(g)$ vs. alternative H_{A1}: at least one mean different is not equal to the others and thus the profiles of the mean responses for the different groups are not parallel.
    v. This "test of the null hypothesis of no group × time interaction effect has (G-1) ×(n-1) degrees of freedom," according to Fitzmaurice et. al[2].

2. Next, the question might be: if the population mean response times are parallel, are the mean response profiles *constant*? Is the sequence of mean responses parallel flat lines, in other words, for each of the categories?
    a. The above response is a question of if there is a time effect.
    b. See Figure 52 below, part (b) for a visualization of the parallel flat line profiles of mean response.
    c. This question is rare and may not have any scientific relevance.
    d. Not common in randomized trial experiments, where the question of interest in "on the comparison of the mean response at each occasion averaged over the groups" according to Fitzmaurice et. al[2].
        i. Can be of interest in observation studies[2].
3. Lastly, if the population mean response times are parallel, are the mean response profiles at the same level? That is, are they the same line and overlap at all time points?
    a. The above response is a question of if there is a group effect.
    b. See Figure 52 below, part (c) for a visualization of the mean response profiles at the same level.
    c. This question is rare and may not have any scientific relevance.
        i. Not usually of interest in randomized trial experiments[2]
        ii. Can be of interest in observation studies[2].

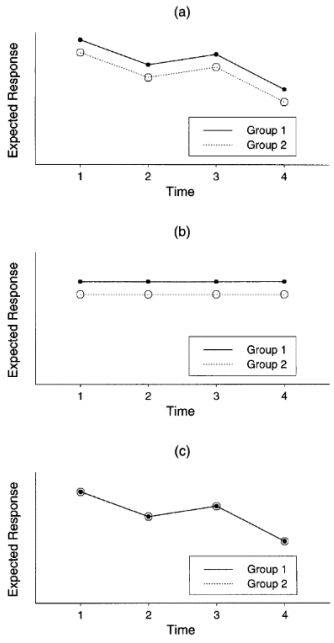


Fig. 5.2 Graphical representation of the null hypotheses of (a) no group × time interaction effect, (b) no time effect, and (c) no group effect.

Figure 52: visualization of the profiles of mean response.

According to Fitzmaurice et. al[2], "The appropriate scientific hypotheses in any particular study must be derived from the relevant scientific issues in that investigation." Randomized trial longitudinal data and observational study longitudinal data are both different. Both affect the hypotheses that will be tested in different ways and are chosen based on the context of the data and study. With regards to data from a randomized trial experiment, the mean response at the first measurement occasion will be independent of the treatment assignment. Thus, all initial mean responses will be the same despite the treatment. In an observation study, because the data is not assumed to be randomly collected, the mean responses at the initial time will not be assumed to be the same. There are four different strategies to handle the baseline response[2].

Response profiles can be found by the use of general linear model (GLM), with $E(Y_i|X_i) = \mu_i = X_i\beta$, with n repeated measures for N subjects and G groups in the predicative covariate. There needs to be G×n parameters to express a model of the G mean responses. The methods that are used here are similar to the discussion of the null and alternative hypotheses above for the first test that is of interest, with X_i being some matrix or vector of contrasts[2].

- Polynomial trends with time

The mean response can be modeled with several models. ***Parametric models***, such as linear or quadratic models, or ***semi-parametric models***, such as the piece-wise linear can be fit on the longitudinal data to model the mean response[2].

Linear trends

A linear trend over time is the simplest model to fit on the longitudinal data. According to Fitzmaurice et. al[2], "in this model the slope for time has direct interpretation in terms of a constant change in the mean response for a single-unit change in time." Perhaps there is a situation where two groups (treatment and control) are being compared. Fitzmaurice et. al[2] said that "if the mean response changes in an approximately linear fashion over the duration of the study, we can adopt the following linear trend model: $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Group_i + \beta_4 Time_{ij} \times Group_i$." The ith individual that was assigned to the treatment group is when $Group_i = 1$. When $Group_i = 0$, the ith individual is assigned to the control group. Next, because $Time_{ij}$ has the ij indices, there might be mistimed measures for two different individuals i and i', where $Time_{ij} \neq Time_{i'j'}$. Now, the mean model for subjects in the treatment group is $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3(1) + \beta_4 Time_{ij} \times (1) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) Time_{ij}$. The mean model for the subjects in the control group is $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3(0) + \beta_4 Time_{ij} \times (0) = \beta_1 + \beta_2 Time_{ij}$ [2].

Quadratic trends

Quadratic trends over time is a higher-order polynomial model that models the monotonically increasing mean responses over time. If the same situation is used from above, and the mean response is better approximated with a quadratic trend, $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2 + \beta_4 Group_i + \beta_5 Time_{ij} \times Group_i + \beta_6 Time_{ij}^2 \times Group_i$. Now, the mean model for subjects in the treatment group is $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2 + \beta_4(1) + \beta_5 Time_{ij} \times (1) + \beta_6 Time_{ij}^2 \times (1) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Time_{ij} + (\beta_3 + \beta_5) Time_{ij}^2$. The mean model for the subjects in the control group is $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2 + \beta_4(0) + \beta_5 Time_{ij} \times (0) + \beta_6 Time_{ij}^2 \times (0) = \beta_1 + \beta_2 Time_{ij} + \beta_3 Time_{ij}^2$ [2]. In the hypothesis testing, the βs that are with the higher order covariates, the squared terms are tested first, then the interaction effects (if any), and then all of the other βs. Note that as the degree of the higher order covariate in the quadratic model, the βs tend to get harder to interpret in the context of the research question.

- Linear splines

A solution to when the βs are hard to interpret in the higher-level quadratic models is using the semi-parametric methods. The common semi-parametric methods are called ***linear splines***. Linear splines model nonlinear trends in longitudinal data in mean responses. According to Fitzmaurice et. al[2], "The basic idea behind linear spline models is remarkably simple: divide the time axis into a series of segments and consider a model for the trend over time that is comprised of piecewise linear trends, having different slopes within each segment but joined or tied together at fixed times." Knots are where the locations meet. If the signs of the regression slopes for the line segments are positive, then as time increases, the mean response will increase. If the signs of the regression slopes for the line segments are negatives, then as time decreases, the mean response will decrease.

In the same situation above, the mean response over time, modeled in a piecewise way, is denoted as $E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 (Time_{ij} - t^*)_+ + \beta_4 Group_i + \beta_5 Time_{ij} \times Group_i + \beta_6 (Time_{ij} - t^*)_+ \times Group_i$. When a truncated line function $(x)_+$ is "defined as a function that equals x when x_is positive and is equal to zero otherwise," according to Fitzmaurice et. al[2]. When $Time_{ij} > t^*$, $(Time_{ij} - t^*)_+ = (Time_{ij} - t^*)$ and when $Time_{ij} \leq t^*$ $(Time_{ij} - t^*)_+ = 0$. So, for the treatment group the mean response expression is:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 (Time_{ij} - t^*)_+ + \beta_4 (1) + \beta_5 Time_{ij} \times (1) + \beta_6 (Time_{ij} - t^*)_+ \times (1) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Time_{ij} + (\beta_3 + \beta_6)((Time_{ij} - t^*)_+).$$

For the control group, the mean response expression is:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 (Time_{ij} - t^*)_+ + \beta_4 (0) + \beta_5 Time_{ij} \times (0) + \beta_6 (Time_{ij} - t^*)_+ \times (0) = \beta_1 + \beta_2 Time_{ij} + \beta_3((Time_{ij} - t^*)_+).$$

So when $Time_{ij} > t^*$, the mean response for the treatment group is:

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Time_{ij} + (\beta_3 + \beta_6) \left((Time_{ij} - t^*) \right) = (\beta_1 + \beta_4) + (\beta_3 + \beta_6)(t^*) + (\beta_2 + \beta_3 + \beta_5 + \beta_6) Time_{ij},$$

and for the control group:

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 \left((Time_{ij} - t^*) \right) = \beta_1 - \beta_3 t^* + (\beta_2 + \beta_3) Time_{ij}.$$

When $Time_{ij} \leq t^*$, the treatment group's mean response expression is[2]:

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Time_{ij} + (\beta_3 + \beta_6)(0) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Time_{ij},$$

and for the control group,

$$E(Y_{ij}) = \beta_1 + \beta_2 Time_{ij} + \beta_3 (0) = \beta_1 + \beta_2 Time_{ij}$$

The group comparison test has a "null hypothesis of no group differences in patterns of change over time can be expressed as $H_0: \beta_5 = \beta_6 = 0$," according to Fitzmaurice et. al[2]. The null hypothesis that tests if there are no differences in the group man responses before t^* is denoted $H_0: \beta_5 = 0$. If there are K knots in a linear spline, then there's K+1 slopes. If k = 1,…, K+1, then complex non-linear patterns for the changes in the mean responses can be accommodated for with the inclusion of "a sufficient number of variables, $(Time_{ij} - t_k{}^*)_+$, with knots located at $t_k{}^*$."[2]

- GLM formulation

A general linear model can be used to express both the linear spline models and the polynomial models. As soon as the covariance of Y_i is specified, the restricted maximum likelihood estimation of β, as well as the confidence intervals and hypothesis tests can be found[2].The general linear model for the treatment group (as in the situation above) for a

quadratic trend is expressed as $E(Y_i|X_i) = \mu_i = X_i\beta =$
$\begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & 1 & t_{i1} & t_{i1}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & t_{in_i} & t_{in_i}^2 & 1 & t_{in_i} & t_{in_i}^2 \end{bmatrix}_{n_i \times 6} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_6 \end{bmatrix}_{6\times 1} = \begin{bmatrix} (\beta_1 + \beta_4) + (\beta_2 + \beta_5)t_{i1} + (\beta_3 + \beta_6)t_{i1}^2 \\ \dots \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5)t_{in_i} + (\beta_3 + \beta_6)t_{in_i}^2 \end{bmatrix}_{n_i \times 1}$
. The general linear model for the control group for a quadratic trend is expressed as $E(Y_i|X_i) = \mu_i = X_i\beta = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & t_{in_i} & t_{in_i}^2 & 0 & 0 & 0 \end{bmatrix}_{n_i \times 6} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_6 \end{bmatrix}_{6\times 1} = \begin{bmatrix} \beta_1 + \beta_2 t_{i1} + \beta_3 t_{i1}^2 \\ \dots \\ \beta_1 + \beta_2 t_{in_i} + \beta_3 t_{in_i}^2 \end{bmatrix}_{n_i \times 1}$. For the spline model, treatment group general linear model is expressed as

$$E(Y_i|X_i) = \mu_i = X_i\beta = \begin{bmatrix} 1 & t_{i1} & (t_{i1} - t^*)_+ & 1 & t_{i1} & (t_{i1} - t^*)_+ \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & t_{in_i} & (t_{in_i} - t^*)_+ & 1 & t_{in_i} & (t_{in_i} - t^*)_+ \end{bmatrix}_{n_i \times 6} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_6 \end{bmatrix}_{6\times 1}.$$

For the control group, the GLM is $E(Y_i|X_i) = \mu_i = X_i\beta =$
$\begin{bmatrix} 1 & t_{i1} & (t_{i1} - t^*)_+ & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & t_{in_i} & (t_{in_i} - t^*)_+ & 0 & 0 & 0 \end{bmatrix}_{n_i \times 6} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_6 \end{bmatrix}_{6\times 1}$[2].

- Covariance Modeling

Approaches must be considered for the covariance modeling of longitudinal data. Recall that longitudinal data is positively correlated. This is due to the fact that when covariance is accounted for in the appropriate longitudinal model, the measure will reduce the variability of the estimate and will increase the validity of change over time for the repeated measures in the dataset. The two main parts of the data that are modeled (interdependently) typically are the conditional mean response over time and, on the same individuals, the conditional covariance on repeated measures[2].

According to Fitzmaurice et. al[2], "the choice of model for the covariance should be based on a 'maximal' model for the mean that minimizes any potential misspecification of the model for the mean." For balanced longitudinal data analysis designs this choice of maximal model is a very straightforward process. When there are situations where there are ample covariates, it may not be very realistic to consider a mean response model that is saturated. Higher-order covariate interactions over time can be considered in the selection of covariates in the final model, based on subject matter grounds, such as through the use of the maximized REML log-likelihood test (given and full and reduced model[2]).

For example, since the compound symmetry model is nested within the Toeplitz model, as discussed below, the full model is the latter and the reduced model is the former. This test will test the null hypothesis of the reduced model being the better, more concise model (or rather that the full and reduced models produce very similar results). In addition, the AIC or BIC values can be used to compare models that are not nested within the other[2].

- Implications of Correlation among Longitudinal Data

Perhaps there is a simple longitudinal study design of that measures the change in health outcome of before and after treatment, therefore stating that there are two repeated measurements. The difference score of the observed before and after response for every individual is $Y_{i2} - Y_{i1}$ and the variance is shown in equation 52 below.

$$V(Y_{i2} - Y_{i1}) = V(Y_{i1}) + V(Y_{i2}) - 2Cov(Y_{i1}, Y_{i2}) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$$ (equation 52)

When $Y_{i2} \perp Y_{i1}$, $V(Y_{i2} - Y_{i1}) = V(Y_{i1}) + V(Y_{i2}) - 2 * 0 = \sigma_1^2 + \sigma_2^2$. Note that, according to Fitzmaurice et. al[2], "provided that the correlation among repeated measures is positive, the variability of the within-individual differences is always smaller than the variability of the between-individual differences."

Read the missing value section below for the definitions of Missing at Random (MAR) and Missing Completely at Random (MCAR) values. Also note that there needs to be a joint distribution of the response vector and covariance model when MAR values occur in the likelihood estimation of the β regression parameters. This is not necessary when there are MCAR values in the longitudinal dataset[2].

- Unstructured Covariance

Unstructured covariance doesn't make assumptions about the variance-covariance matrix $Cov(Y_i)$. It is a very flexible and robust covariance "structure", since the matrix will have no structure. An unstructured covariance matrix is when the covariances of the elements in the matrix are arbitrary and homogeneous, and also symmetric and positive-definite (there's no redundancy of the repeated measurements). Such a matrix is normally found when the n repeated measurement count is small and when all the research subjects are measured at the same times. An unstructured covariance matrix is denoted as in equation 53.

$$Cov(Y_i) = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix}$$ (equation 53)

This covariance matrix will have $\frac{n\times(n+1)}{2}$ and will hare $\frac{n\times(n-1)}{2}$ pairwise covariances. There is a potential drawback "of assuming an unstructured covariance: the number of covariance parameters to be estimated grows rapidly with the number of measurement occasions," according to Fitzmaurice et. al[2]. That is, having no assumptions of the $Cov(Y_i)$ matrix is typically costly[5]. There are plenty of methods that can help with resolve any issues with the unstructured covariance matrix: the structured covariance matrix pattern models.

- Covariance Pattern Models (Structured Covariance)

One of the main problems with the unstructured covariance matrix is that, as stated by Fitzmaurice et. al[2]:

> "If too little structure is imposed on the covariance, there will be too many covariance parameters to be estimated from the limited amount of data available, and this will adversely affect the precision with which the main parameters of interest, β, can be estimated. As a result, imposing too little structure on the covariance can result in weaker inferences concerning β."

Through the use of covariate pattern models, such as compound symmetry, Toeplitz, Autoregressive, banded, exponential, and hybrid models, structure can be fit to the covariance data in the covariance matrix.

The ***covariance compound symmetry matrix*** is denoted as in equation 54, where σ^2 is a constant compound symmetry covariance and $\rho = corr(Y_{ij}, Y_{ik}) \geq 0$ (positively correlated), for each j and k.

$$Cov(Y_i) = \sigma^2 \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix} \text{ (equation 54)}$$

The longitudinal responses can be expressed as $Y_{ij} = X_{ij}^T\beta + b_i + \varepsilon_{ij}$, with the random effect denoted as the b_i. According to Fitzmaurice et. al[2], "In an experiment where the within-subject factor is randomly allocated to subjects, randomization arguments can be made to show that the constant variance and constant correlation conditions hold." This condition mentioned is known as sphericity or as circularity, and all usual ANOVA F-tests or T-tests remain valid[5]. Notice that, in this case, compound symmetry holds because $V(Y_{i2} - Y_{i1}) = V(Y_{i1}) + V(Y_{i2}) - 2Cov(Y_{i1}, Y_{i2}) = (\sigma^2) + (\sigma^2) - 2\sigma^2\rho = 2\sigma^2(1 - \rho)$ is a constant.

A major flaw of the covariance compound symmetry is that the assumption of the correlation between a pair of repeated measurements is the same as the correlations between other pairs. This is a bold assumption because the time intervals between the repeated measurement times are likely to be different and that is not appealing for most longitudinal studies due to the rigid assumptions and restrictions. There are plenty of alternative covariance structures to choose from, however.

A similar covariance structure to the covariance compound symmetry matrix is called the ***Toeplitz covariance pattern***. It is denoted as in equation 55, with $\rho_k = corr(Y_{ij}, Y_{ij+k})$ (for all j and k) and a constant variance σ^2 assumed to be across all repeated measurements.

$$Cov(Y_i) = \sigma^2 \begin{bmatrix} 1 & \cdots & \rho_{n-1} \\ \vdots & \ddots & \vdots \\ \rho_{n-1} & \cdots & 1 \end{bmatrix} \text{ (equation 55)}$$

It is also assumed that there is the same correlation between a pair of responses equally separated in time. The first-order ***autoregressive*** covariance is a specific case of the Toeplitz covariance pattern matrix.

Autoregressive is defined as, according to Hand et. al.[5], "when the measurements on an individual have been made in sequence over time [and when] the errors may be serially correlated [as a result]." (Serial correlation being defined as a sequential pattern in the responses over time). Hence, the repeated measurements of the ith individual might have a relationship between the error components of the error vectors. Note that the autoregressive process is often used in time series, a cousin of longitudinal data analysis.

The autoregressive covariance structure matrix is denoted as in equation 56, with $\rho^k = corr(Y_{ij}, Y_{ij+k})$ (for all j and k) and a constant variance σ^2 assumed to be across all repeated measurements.

$$Cov(Y_i) = \sigma^2 \begin{bmatrix} 1 & \cdots & \rho^{n-1} \\ \vdots & \ddots & \vdots \\ \rho^{n-1} & \cdots & 1 \end{bmatrix} \text{ (equation 56)}$$

According to Fitzmaurice et. al[2], "Because the autoregressive covariance has a Toeplitz form, this structure is only appropriate when the measurements are made at equal (or approximately equal) intervals of time." As the difference in the pairs of repeated measures increase, the correlation values will decrease over time.

The *banded covariance structure matrix* is denoted as in equation 57, where $corr(Y_{ij}, Y_{ij+k}) = 0$, for $k \geq 3$. So, as a result, $\rho_2 = \rho_3 = \cdots = \rho_{n-1} = 0$.

$$Cov(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ 0 & \rho_1 & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \text{ (equation 57)}$$

The constant variance σ^2 is assumed to be across all repeated measurements. As the difference in the pairs of repeated measures increase, the correlation values will decrease over time to zero.

Exponential covariance structure is best when the repeated measurements are not equally spaced over time. This is because repeated measurements are not assumed to be equally spaced over time. This covariance structure "assumes that the correlation is one if measurements are made repeatedly at the same occasion (or replicate measurements on an individual can be obtained at the same occasion), and that the correlation decreases rapidly to zero as the time separation between measurements increases," according to Fitzmaurice et. al[2]. The structure of the model is $Cov(Y_{ij}, Y_{ik}) = \sigma^2 e^{-\theta|t_{ij} - t_{ik}|}$ for positive correlation $\rho = e^{-\theta} \geq 0$, $\theta = -\log(\rho) \geq 0$ and constant variance σ^2. The flaws of the covariance structure is that there is an unrealistic assumption of the responses being recorded and measured with no error or missing values. The correlation values will decrease over time to zero[2].

Hybrid models combine compound symmetry models and autoregressive models. This model type has the covariance structure $Cov(Y_i) = \Sigma_1 + \Sigma_2$. The matrices are shown in equations 58 and 59 below.

$$\Sigma_1 = \sigma_1^2 \begin{bmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & & \vdots & & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \cdots & 1 \end{bmatrix} \text{ (equation 58)}$$

and

$$\Sigma_2 = \sigma_2^2 \begin{bmatrix} 1 & \rho_2^{|t_{i1}-t_{i2}|} & \rho_2^{|t_{i1}-t_{i3}|} & \dots & \rho_2^{|t_{i1}-t_{in}|} \\ \rho_2^{|t_{i2}-t_{i1}|} & 1 & \rho_2^{|t_{i2}-t_{i3}|} & \dots & \rho_2^{|t_{i2}-t_{in}|} \\ \rho_2^{|t_{i3}-t_{i1}|} & \rho_2^{|t_{i3}-t_{i2}|} & 1 & \dots & \rho_2^{|t_{i3}-t_{in}|} \\ \vdots & & \vdots & & \vdots \\ \rho_2^{|t_{in}-t_{i1}|} & \rho_2^{|t_{in}-t_{i2}|} & \rho_2^{|t_{in}-t_{i3}|} & \dots & 1 \end{bmatrix}$$ (equation 59)

The variance is $V(Y_{ij}) = \sigma_1^2 + \sigma_2^2$ and the covariance is $Cov(Y_{ij}, Y_{ik}) = \rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2$. So therefore, the correlation is denoted as in equation 60.

$$Corr(Y_{ij}, Y_{ik}) = \frac{\rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$ (equation 60)

This implies that when $|t_{ij} - t_{ik}| = 0$, or when the repeated measurements are completed at the same occasion, where the correlation is denoted as in equation 61[2].

$$Corr(Y_{ij}, Y_{ik}) = \frac{\rho_1\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$ (equation 61)

- Linear Mixed Effects Models
    - Definition

When a subset of the regression parameters has random variance between individuals in the data, then a linear mixed effect model should be fit on the data. This model will account for the heterogeneity in the dataset and population. According to Fitzmaurice et. al[2],

> "that is, individuals in the population are assumed to have their own subject-specific mean response trajectories over time and a subset of the regression parameters are now regarded as being random. The distinctive feature of linear mixed effects models is that the mean response is modeled as a combination of population characteristics, that are assumed to be shared by all individuals, and subject specific effects that are unique to a particular individual."

These effects are called the fixed effect and random effects, respectively, and when characteristics of both are combined then the effect is mixed.

The expected value of this model is expressed as $E(Y_i|X_i) = X_i\beta$, where the combination of β parameters are fixed effects and effects dependent on the subjects. The random effects are taken into consideration in the random effects' covariance structure denoted as $Cov(Y_i|X_i) = \Sigma_i$. "With the inclusion of random effects, the covariances among the repeated measures can be expressed as functions of time," according to Fitzmaurice et. al[2]. To get predictions of individual's growth trajectories, the linear mixed effects model can be used. Such model is flexible at accommodating imbalances in longitudinal datasets. A linear mixed effect model also accounts for the covariance in the repeated measures. The random intercept model is an example of an application of the linear mixed effect model[2].

The linear mixed effects model is expressed as in equation 62.

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i$$ (equation 62)

The fixed effects vector is a (p×1) and is β. The random effects vector is (q×1) and is denoted b_i, and is independent of the X_i covariates and is distributed as MVN($E(b_i) = 0, Cov(b_i) = G$). The two matrices of covariances are X_i, which is an $n_i \times p$ matrix, and Z_i, which is an $n_i \times q$ matrix

and is a design matrix that links the vector of random effects b_i to the $n_i \times 1$ vector Y_i. The vector of random errors ε_i is $n_i \times 1$ is MVN($E(\varepsilon_i) = 0, Cov(\varepsilon_i) = Cov(Y_i|b_i) = R_i = \sigma^2 I_{n_i}$). Also, the conditional, or subject-specific, mean of the responses is $E(Y_i|b_i) = X_i\beta + Z_i b_i$ and the marginal (population-averaged mean) of the responses is denoted as in equation 63.

$$E(Y_i) = \mu_i = E\big(E(Y_i|b_i)\big) = E(X_i\beta + Z_i b_i) = X_i\beta + 0 = X_i\beta \text{ (equation 63)}$$

The marginal covariance of Y_i is denoted as in equation 64.[2]

$$Cov(Y_i) = cov(Z_i b_i) + cov(\varepsilon_i) = Z_i cov(b_i) {Z_i}^T + R_i = Z_i G {Z_i}^T + \sigma^2 I_{n_i} \text{ (equation 64)}$$

Perhaps we have a linear mixed effect model with the intercepts and slopes that have random variance among the individuals. This model is denoted as in equation 65, for the ith individual and the jth measurement occasion, where $j = 1, \dots, n_i$. for the ith individual and the jth measurement occasion, where $j = 1, \dots, n_i$.

$$Y_{ij} = \beta_1 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij} \text{ (equation 65)}$$

Or, this model can be expressed as in equation 66.

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix} \beta + \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix} b_i + \varepsilon_i \text{ (equation 66)}$$

The individuals in this dataset will "vary not only in their baseline level of response (when $t_{i1} = 0$), [and] in terms of their changes in the mean response over time," according to Fitzmaurice et. al[2].

If there is the example from above that compares the treatment to the control group, and if the mean response changes in a linear way, the linear mixed effects model can be denoted as:

$$Y_{ij} = \beta_1 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + (\beta_3 + \beta_4 t_{ij})Group_i + \varepsilon_{ij}.$$

This expression comes with the assumption that the means of the slopes and the intercepts vary with the group, where the treatment group is Group = 1, with the design matrix $X_i = \begin{bmatrix} 1 & t_{i1} & 1 & t_{i1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 1 & t_{in_i} \end{bmatrix}$ and the control group is Group = 0, with $X_i = \begin{bmatrix} 1 & t_{i1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 0 & 0 \end{bmatrix}$. But, $Z_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}$ for both the treatment and control groups[2].

The $V(Y_{ij}) = V({X_{ij}}^T\beta + {Z_{ij}}^T b_i + \varepsilon_{ij}) = 0 + V({Z_{ij}}^T b_{ij} + \varepsilon_{ij}) = V(b_{1i} + b_{2i}t_{ij} + \varepsilon_i) = V(b_{1i}) + 2t_{ij}cov(b_{1i}, b_{2i}) + {t_{ij}}^2 V(b_{2i}) + \sigma^2$. When $t_{ij} \geq 0$, the $V(Y_{ij})$ will increase for this quadratic linear mixed effect model when $cov(b_{1i}, b_{2i}) \geq 0$ and will decrease when $cov(b_{1i}, b_{2i}) < 0$. Lastly, the $cov(Y_{ij}, Y_{ik}) = cov({X_{ij}}^T\beta + {Z_{ij}}^T b_i + \varepsilon_{ij},\ {X_{ik}}^T\beta + {Z_{ik}}^T b_i + \varepsilon_{ik}) = cov({Z_{ij}}^T b_i + \varepsilon_{ij},\ {Z_{ik}}^T b_i + \varepsilon_{ik}) = V(b_{1i}) + (t_{ij} + t_{ik})cov(b_{1i}, b_{2i}) + t_{ij}t_{ik}V(b_{2i}) + 0$.[2]

- The two-stage random effects formulation:

There are two-stages to the formulation of the b_i random effects for the model defined above in equation 62. According to Hand et. al.[5], "b_i are random effects, yielding a combined contribution to Y_i via the design matrix Z_i, and β is a vector of fixed effects." In two-stage

models, the vector of the random errors ε_i are first based on variation within individuals and then are based on variation between individuals.

- Stage 1: ε_i based on within-individual variation

The repeated measures on each of the observations in the dataset will follow some regression model, denoted as in equation 67.

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = Z_i\beta_i + \varepsilon_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix} \begin{bmatrix} \beta_{1i} \\ \beta_{2i} \end{bmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \text{ (equation 67)}$$

The set of covariates will be the same set, but each of the observations will have different regression coefficients. Each observation gets its own linear regression models, with each equation involving the same covariates from the covariant matrix Z_i.[2]

- Stage 2: ε_i based on between-individual variation

Next, the diagnostics need to be evaluated for the random individual-specific effects β_i.If A_i is a $q \times p$ matrix of the between-individual covariates, the mean of β_i is denoted $E(\beta_i) = A_i\beta$ and the covariance of β_i is denoted as $Cov(\beta_i) = G$. Then, the mean and covariances of β_i are specified with the specification of a model, like with the treatment vs. control group scenario[2].

- Choice among random effects covariance models

When determining which random effects covariance model to use, compare two nested models. One of said models has q correlated random effects (the reduced model), and the other has q+1 correlated random effects (the full model). The null hypothesis of comparing the two nested models is based on the likelihood ratio test. According to Fitzmaurice et. al[2], "when comparing two nested models, one with q correlated random effects, the other with q + 1 correlated random effects, the null distribution of the likelihood ratio test is a 50:50 mixture of chi-squared distributions with q and q + 1 degrees of freedom." If the null hypothesis is rejected, given some α, then the reduced model is the better model to fit the data.

- Fixed Effects vs. Mixed Effects Models
  - Definition: Linear Fixed Effects Models

Linear fixed effects models are useful because they overcome the limitations of the adjustments of confounding variables included in regression models. Such limitations are that there will always be confounders that should have been included in the regression model and that such confounders will likely be difficult to measure and collect. According to Fitzmaurice et. al[2], "the fundamental idea underlying fixed effects models is the control of all potential confounding variables that remain stable across repeated measurement occasions and whose effects on the response are assumed to be constant over time." There are two types of confounders in longitudinal data: observed and the time-invariant confounders. Fixed effect models aim to remove both types of confounders from the regression model, and they require

repeated measures of the response variable that has two or more levels and the set of covariates for some subset of the sample needs to be variable over the repeated measurement occasions.

The linear fixed effects model is denoted as a $p \times 1$ follows in equation 68.

$$Y_{ij} = X_{ij}^T \beta + W_i^T \gamma + \alpha_i + \varepsilon_{ij} \text{ (equation 68)}$$

The type of covariates that change with time, or the time-varying covariates are put into the $q \times 1$ vector X_{ij}^T. The type of covariates that do not change with time, or the time-invariant covariates are put into the $(p-q) \times 1$ vector W_i^T. According to Fitzmaurice et. al[2], "to accommodate unbalanced data, we assume that there are n_i repeated measurements of the response on the i[th] subject and that each $\sim j$ is observed at time t_{ij}." Additionally, there are random errors within the subjects, denoted as $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. The $p \times 1$ vector α_i is the fixed effects vector that represent the time-invariant or stable characteristics of the observations that are not included in the W_i^T vector defined. The main difference between the linear fixed effects model and the linear mixed effects model is the fact that in the latter model, the vector α_i is a random, not fixed, set of effects[2].

For example, suppose there are two measurement occasions repeated and the features that the linear fixed effects must have are all satisfied. The linear fixed effects model will be denoted as in equation 69.

$$Y_{ij} = X_{ij}^T \beta + W_i^T \gamma + \alpha_i + \varepsilon_{ij}, \text{ for } j = 1, 2 \text{ (equation 69)}$$

According to Fitzmaurice et. al[2], "cannot estimate both α_i and γ from the data at hand" in the linear fixed effects model described. The vector of time-invariant covariates $W_i^T \gamma$ cannot be estimated due to the perfect collinearity that it has with the vector α_i. But both measures will disappear as will be shown below.

However, using the Ordinary Least Squares (ORS) regression method the β can be estimated. Taking the within-subject changes, the model becomes $Y_{i2} - Y_{i1} = X_{i2}^T \beta - X_{i1}^T \beta + W_i^T \gamma - W_i^T \gamma + \alpha_i - \alpha_i + \varepsilon_{i2} - \varepsilon_{i1} = (\varepsilon_{i2} - \varepsilon_{i1}) + (X_{i2}^T - X_{i1}^T) * \beta$. So the OLS regression estimates β through the fitting of a simple linear regression model of $(Y_{i2} - Y_{i1})$ on $(X_{i2}^T - X_{i1}^T)$. Note that the error terms $E(\varepsilon_{i2} - \varepsilon_{i1}) = 0$ and $E(\varepsilon_{i2} - \varepsilon_{i1}) = 2\sigma_\varepsilon^2$. According to Fitzmaurice et. al[2], "The fixed effects model can only remove the potential confounding by those measured and unmeasured time-invariant covariates whose effects on the response remain constant over time. That is, conditional on X_{ij} and W_i, it must be assumed that the effect of any time-invariant confounder on Y_{i1} is the same as on Y_{i2}." Therefore, the fixed effects estimate of β, the effect of X_{ij} on Y_{ij} will not be biased.

- Fixed Effects versus Mixed Effects: Bias-Variance Trade-off

According to Fitzmaurice et. al[2], "fixed effects models remove the potential for bias due to certain types of confounding variables, namely measured and unmeasured time-invariant confounders whose effects on the response can be assumed to be constant across measurement occasions." Such a property is unique to fixed effect models; that is, the linear mixed effects

model does not share this property with the fixed effect model because the former model requires that the time-invariant and stable characteristics of the observations in the dataset have stronger assumptions made about them than the latter model, treating the α_i term as random and uncorrelated with X_{ij} instead of fixed and correlated with X_{ij}.

Due to these strong assumptions, the linear mixed effects model might produce biased estimated of the effect of X_{ij} on Y_{ij}, if such assumptions do not hold for the α_i term. However, the linear mixed effect model is still often used because it can estimate a mixture of time-varying and time-invariant covariates, whereas the fixed effect model can only estimate the time-varying effects. In addition, the mixed effect model, even though it might produce bias estimated, is more efficient than the fixed effects model at calculating such estimates. (Fitzmaurice 246).

- General linear models (GLMs) for Longitudinal Data
    - Definition and Feature of GLMs

General linear models, or GLMs, are very important in the modeling of regression models such as the longitudinal data analysis. It should be recalled that in longitudinal studies of a continuous response variable with MCAR missing data and a large amount of data does not require a multivariate normal assumption. According to Fitzmaurice et. al[2], "A characteristic feature of generalized linear models is that a suitable non-linear transformation of the mean response is related to a linear function of the covariates. This non-linearity raises some additional issues concerning the interpretation of the regression coefficients in models for longitudinal data." GLMs can be extended from univariate responses, such as continuous, binary, ordinal, and count responses, to ANOVA models given a normally distributed response, logistic regression models for binary response, and even using Poisson or log-linear regression models for count response data, and more. Note that the distributions mentioned here for the mean response transformations are mainly used in the biomedical analyses.

Assumptions of GLMs

GLMs have an assumption about their distribution, a systematic component to the model, and a ***link function*** that completes a transformation on the response data. The distributional assumption of the GLM model is that GLMs have a "response variable has a probability distribution belonging to the exponential family of distributions", such as the normal, Bernoulli, and Poisson distributions, according to Fitzmaurice et. al[2]. Such an assumption is defined as the random component of the GLM. For each of these distributions referred to, the variance of the response is denoted by equation 70, where $\phi > 0$ is the dispersion or scaling parameter, $v(\mu_i)$ is the variance of $\mu_i = E(Y_i|X_i)$.

$$Var(Y_i) = \phi * v(\mu_i) \text{ (equation 70)}$$

The table of the normal, Bernoulli/binomial, and Poisson distributions' variance functions and canonical links are shown below in Figure 53.

Distribution	Variance Function, $v(\mu)$	Canonical Link
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1-\mu)$	Logit: $\log\left(\frac{\mu}{1-\mu}\right) = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

Figure 53: from Fitzmaurice (294).

Systematic Component of GLMs

The systematic component of a GLM is donated as $\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$ and this notation species all of the covariates in vector X_i on the vector of mean responses Y_i, where η_i is the linear predictor, with the unknown coefficient vector β. Generally, the term $\beta_1 X_{i1}$ has $X_{i1} = 1$, and is thus the intercept term in the model. According to Fitzmaurice et. al[2], "linearity strictly applies to the regression parameters, but not necessarily to the covariates." The linearity restriction of the linear predictor requires that the covariates must be linear (that is, not in exponents). Any violation of the linearity assumption of the covariates can be fixed with transforming the mean response.

Link function transformation

Such a transformation to the mean response of at GLM is done through the use of a known link function, which is the same as the linear predictor described above and is denoted as in equation 71.

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik} = X_i^T \beta \text{ (equation 71)}$$

There are two main types of link functions: canonical, or a unique and derived for any distribution, and non-canonical, or arbitrary and have no relation to the distribution (296). The canonical link functions for the normal, Bernoulli, and Poisson distributions are shown in figure 53 above. According to Fitzmaurice et. al[2], "for example, the logit link function is the canonical link function associated with the Bernoulli and binomial distributions; the probit link function is a non-canonical link function for these distributions that is often adopted for the analysis of binary data from toxicological experiments." Note that the most widely used distributions have a canonical link function.

The Maximum Quasi-Likelihood Estimation (MQLE)

According to Hand et. al.[5], "the phrase '***maximum quasi-likelihood estimation'*** refers to the estimating equation, $q(y, \beta) = 0$, used to obtain the MQL estimator $\hat{\beta}$ of β" (130). MQLE is different from GLMs because GLMs relate to the introduction of μ_i to a link function g. If $y = (y_1, \ldots, y_n)^T$ and μ_i' is a p×1 vector with $\partial\mu_i / \partial\beta_j$ as the jth element, the ***quasi-score function*** for the β is denoted as in equation 72[5].

$$q(y, \beta) = \sum v(\mu_i)^{-1}(y_i - \mu_i)\,\mu_i' \text{ (equation 72)}$$

When y_i is in the "linear exponential family distribution with specifications [link function g for μ_i and $Var(Y_i) = \phi * v(\mu_i)$], then $q(y, \beta) \propto {\partial logL}/{\partial \beta}$. In this case, $\hat{\beta}$ is a genuine maximum likelihood estimator (MLE)," according to Hand et. al.[5]. The linear exponential family of distributions is shown above in Figure 53, in addition to the exponential loglinear model. If y_i is not in the linear exponential family distribution, then the resulting MQLE $\hat{\beta}$ will be asymptotic and have similar properties to the true MLE $\hat{\beta}$. This is due to the resulting quasi-score function having the characteristics of a true score function. Note that a simple case of the MQLE is using it with the weighted least squares.[5]

Now, if $Var(Y_i)$ is correctly specified with a link function g for μ_i and $Var(Y_i) = \phi * v(\mu_i)$], then the estimating equation, $q(y, \beta) = 0$ is asymptotically optimal. According to Hand et. al.[5], "the MQLE $\hat{\beta}$ has large-sample distribution approximately $N(\beta, V_\beta)$ with $V_\beta = \phi(X^T W X)^{-1}$, where X is the n×q matrix with ith row $x_i^T, W = diag[w_i]$, $w_i^{-1} = v(\mu_i)g'(\mu_i)^2$." The estimate of the V_β comes from estimating ϕ and using the $\hat{\mu}_i = \mu_i(\hat{\beta})$ as a substitute of w_i. The estimate of the scaling parameter ϕ is denoted in equation 73.

$$\hat{\phi} = (n-q)^{-1} \sum {(Y_i - \hat{\mu}_i)^2}/{v(\hat{\mu}_i)} \text{ (equation 73)}$$

When Y_i has an exponential family distribution with the mean μ_i and variance $Var(Y_i)$ specified correctly with a link function g for μ_i and $Var(Y_i) = \phi * v(\mu_i)$], then the likelihood can be found for Y: $L(Y; \mu, \phi)$. That is, for the n observations in Y, the likelihood "depends on the parameter β through μ, …[and the likelihood will be written] as $L(Y; \mu, \phi)$ rather than the 'correct form $L(Y; \beta, \phi)$," according to Hand et. al.[5]. The maximized likelihood is denoted $L(Y; \hat{\mu}, \phi)$. What is called the ***deviance*** is the scaled log-likelihood ratio and is denoted as in equation 74.

$$D(Y; \hat{\mu}) = 2\{logL(Y; Y, \phi) - logL(Y; \hat{\mu}, \phi)\} \text{ (equation 74)}$$

The likelihood $L(Y; Y, \phi)$ in this expression is a resulting likelihood from the saturated model, where unconstrained Y_i estimate for Y is substituted in for μ_i.

A log-likelihood ratio hypothesis test can be completed to determine any inferences on the MQLE. Such a test can be used to verify "that the score function for each of these examples has the form of $q(y, \beta)$ … and that, therefore, the MQLE $\hat{\beta}$ is a true MLE," according to Hand et. al.[5]. If hypothesis H places a constraint on β, then the parameter dimension will reduce from q to q_1. If $\hat{\beta}_H$ is the MLE of β under the null hypothesis H and if this holds the standard asymptotic likelihood theory's instructions, then equation 75 is used.

$$2\{\log(L(Y; \hat{\mu}, \phi)) - \log(L(Y; \hat{\mu}_H, \phi))\} = \phi^{-1}\{D(Y; \hat{\mu}_H) - D(Y; \hat{\mu})\} \dot{\sim} \chi^2_{q_1 - q} \text{ (equation 75)}$$

According to Hand et. al.[5], "If ϕ is unknown an estimate needs to be inserted ,…, and then one can refer $\phi^{-1}\{D(Y; \hat{\mu}_H) - D(Y; \hat{\mu})\}$ to the $F_{q_1 - q, n-q}$."

- Ordinal Regression

Ordinal regression occurs when the response is an ordinal categorical variable with 3 or more responses that are ordered in some way. It is a generalized version of logistic regression and is considered to be a multivariate GLM. The transformation to the response data will be applied to the cumulative response probabilities[2]. Suppose that the latent (or unobserved) continuous variable L_i and the ordinal response variable Y_i can be observed only when the L_i is in one of the K intervals of time. The Y_i is denoted as in equation 76.

$$Y_i = \begin{cases} 1, if -\infty < L_i \leq \alpha_1 \\ \vdots \\ K, if\ \alpha_{K-1} < L_i \leq \infty \end{cases} \text{ (equation 76)}$$

The cut-off points $\alpha_1, \dots \alpha_{K-1}$ are fixed values and are determined by the context of the situation. If the linear regression model for L_i holds then equation 77 is applied.

$$L_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i = X_i^T \beta + e_i \text{ (equation 77)}$$

The standard logistic regression model denoted as the errors I, in equation 78.

$$e_i = L_i - X_i^T \beta \sim Logistic(\mu = 0, \sigma^2 = \frac{\pi^2}{3}) \text{ (equation 78)}$$

In this distribution of L_i, therefore, is changing with the covariates X_i^T (Fitzmaurice 313-314).

The covariate effects, or the β vector must be invariant for the proportional hazards assumption to be met. For response level k, letting $F_{ik} = \Pr(Y_i = k)$, the non-proportional odds model is denoted as in equation 79.

$$logit(F_{ik}) = \alpha_k + \beta_{k1} X_{i1} + \cdots + \beta_{kp} X_{ip} \text{ (equation 79)}$$

Hence, the covariate effects β depend on response k, as does the log odds ratio for the $logit(F_{ik})$ model, denoted as in equation 80.

$$\beta_{k1} = log\left[\frac{F_{ik}(X_{i1}=c+1)/\{1-F_{ik}(X_{i1}=c+1)\}}{F_{ik}(X_{i1}=c)/\{1-F_{ik}(X_{i1}=c)\}}\right] \text{ (equation 80)}$$

The $logit(F_{ik})$ model is used to test the null hypothesis on the proportionality assumption denoted as $H_0: \beta_{1j} = \cdots = \beta_{K-1,j} = \beta_j$ for each of the p covariates, j = 1,…,p. This test of the proportionality assumption has $df = \#distinct\ \beta\ in\ H_a - \#distinct\ \beta\ in\ H_0 = (K-1) \times p - p = (K-2) \times p$. Regression parameters can be estimated via the use of Maximum Likelihood Estimation (MLE), which maximizes the multinomial likelihood of the ordinal response. The means of the K-1 cumulative random variables are jointly related to the covariates[2].

- Residual Analyses and Diagnostics

Residuals are meant to assess the fit of the model and to indicate any outliers in the dataset, if there are any. For each of the observations in the dataset, the vector of residuals of the outcome over time is denoted as in equation 81.

$$r_i = Y_i - X_i \hat{\beta} \text{ (equation 81)}$$

which are the statistics that estimate the parameters in the vector of the true errors for an observation in a dataset, denoted as $e_i = Y_i - X_i\beta$. Note that the covariance of the residuals is approximately the same as the covariance of the true errors: $Cov(r_i) \approx Cov(e_i) = \Sigma_i$. A systematic trend can be detected in a plot of the residuals of all of the j observations $r_{ij} = Y_{ij} - {X_{ij}}^T\hat{\beta}$ vs. the predicted mean responses for all of the observations $\hat{\mu}_{ij} = X_{ij}^T\hat{\beta}$ be fitting of a smooth curve through the scatterplot. If the scatterplot has the data randomly scattered around the smooth curve of zero, then there is no systematic pattern.

According to Fitzmaurice et. al[2], "because the variance is not necessarily constant, the scatterplot of the residuals against the predicted values, or against time, will not necessarily have a constant range. As a result, standard residual diagnostics for examining either the homogeneity of the residual variance or autocorrelation among the residuals should be avoided altogether." Such residuals will likely be correlated with the covariates collected for the observations in the longitudinal dataset.

- Define Transformed Residuals

Residuals might need to be transformed in longitudinal studies so that they will have constant variance and zero correlation. The Cholesky decomposition is used to create a transformed residual. First, the Cholesky decomposition of the estimated covariance matrix of the residuals should be found $\hat{\Sigma}_i = L_i L_i^T$, with the lower triangular matrix $L_i = L_i^{-1}$. The set of correlated residuals with heterogeneous variances is denoted as in equation 82.

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\hat{\beta}) \text{ (equation 82)}$$

The kth transformed residual in the row of r_i^* estimates $\frac{Y_{ik} - E(Y_{ik}|Y_{i1},\ldots,Y_{ik-1})}{\sqrt{Var(Y_{ik}|Y_{i1},\ldots,Y_{ik-1})}}$. The transformed predicted values is denoted as in equation 83, for observation i.

$$\hat{\mu}_i^* = L_i^{-1}\hat{\mu}_i = L_i^{-1}X_i\hat{\beta} \text{ (equation 83)}$$

The plot of the transformed residuals r_{ij}^* vs. the transformed predicted values $\hat{\mu}_{ij}^*$ will show if there is a systematic pattern in the data, in the same way as above. In addition, the "normal quantile plot (or so-called quantile-quantile or Q-Q plot) of the transformed residuals can be used to assess the normal distribution assumption and to identify outliers," according to Fitzmaurice et. al[2].

- Define Aggregating Residuals

The basic idea of aggregating the residuals over some of the coordinates is to check the model by discerning the signals from the noise in such scatterplots of the residuals vs. fitted response values discussed above. The main advantage of aggregate residuals is that they allow "us to determine whether any apparent pattern in the observed sum of the residuals is evidence of a systematic trend or simply due to natural variation. This way we can remove a large degree of subjectivity from the assessment of graphical displays of residuals, placing residual diagnostics on a more objective footing," according to Fitzmaurice et. al[2].

Aggregating residuals require that the residuals are summed over the individual covariates X_{ijk} and the fitted values ${X_{ij}}^T\hat{\beta}$. According to Fitzmaurice et. al[2], "the cumulative sum of the residuals over the fitted values is denoted as" in equation 84

$$W_f(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{n_i} I({X_{ij}}^T\hat{\beta} \le x)r_{ij} \text{ (equation 84)}$$

This cumulative sum is useful for testing of the linear assumption and the any systematic trend resulting from such a plot indicates the need for a transformation on the responses or of the mean response to make no pattern of the data. Also, note that there is an alternative moving sum aggregating residual which is not as influenced by any residuals that are associated with small covariate values.

- Residuals in the context of GLMs

There are two types of residuals that can be used as alternatives of the ***traditional standardized residual*** in the context of GLM. Such residuals that correspond to can also be expressed as Y_i are denoted as in equation 85.

$$(Y_i - \mu_i)/\sigma_i \text{ (equation 85)}$$

Where σ_i is the standard deviation of Y_i ($\sqrt{Var(Y_i)}$). According to Hand et. al.[5], the ***standardized Pearson residual*** takes the estimates from the quasi-likelihood, $\hat{\mu}_i = \mu_i(\hat{\beta})$ and $\hat{\sigma}_i = \hat{\phi} * v(\hat{\mu}_i)$, and inserts them into the equation for the traditional standardized residual, $(Y_i - \hat{\mu}_i)/\hat{\sigma}_i$, while also improving the residual equation through the use of the (i,i)the element h_{ii} of hat matrix denoted in equation 86 to obtain 87.

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2} \text{ (equation 86)}$$

$$r_i = (Y_i - \hat{\mu}_i)/\left\{\hat{\sigma}_i\left(1 - \hat{h}_{ii}\right)^{1/2}\right\} \text{ (equation 87)}$$

Here, the distribution ϕ is unknown, so it needs to be estimated with the $\hat{\phi}$ defined above. Also, note that the variance of the standardized Pearson residual is closer to 1 than the traditional standardized residual.

Next, the ***deviance residual*** is defined by how they contribute to the deviance, when the likelihood case is considered to be true. Furthermore, according to Hand et. al.[5], "$D(Y;\hat{\mu})$ [defined above]…is the sum of n terms, say D_i [with] i = 1,…,n." For $s_i = sign(Y_i - \hat{\mu}_i)$, $\hat{h}_{ii}$ tas defined above, the ith deviance residual is denoted as $s_i D_i^{1/2}$. A variance corrected form for the deviance residual is denoted as in equation 89.

$$s_i D_i^{1/2}\left(1 - \hat{h}_{ii}\right)^{-1/2} \text{ (equation 89)}$$

- Define: Semi-Variogram

A semi-variogram can be used to determine the fit of a model for the covariance. The semi-variogram, for longitudinal data, is denoted as in equation 90, which is the halved value of the expected squared difference between the individual residuals r_{ij} and r_{ik}.

$$\gamma(h_{ijk}) = \frac{1}{2}E(r_{ij} - r_{ik})^2 = \frac{1}{2}E\left(r_{ij}^2 - r_{ik}^2 - 2r_{ij}r_{ik}\right) = \frac{1}{2}Var(r_{ij}) + \frac{1}{2}Var(r_{ik}) - Cov(r_{ij}, r_{ik})$$

(equation 90)

Applying this formula for the semi-variogram to the transformed residuals defined above, the semi-variogram becomes $\gamma(h_{ijk}) = \frac{1}{2}E(r_{ij}^* - r_{ik}^*)^2 = \frac{1}{2}Var(r_{ij}^*) + \frac{1}{2}Var(r_{ik}^*) - Cov(r_{ij}^*, r_{ik}^*) = \frac{1}{2}(1) + \frac{1}{2}(1) - (0) = 1$. According to Fitzmaurice et. al[2], "in a correctly specified model for the covariance, the plot of the semi-variogram for the transformed residuals versus the time elapsed between the corresponding observations should fluctuate randomly around a horizontal line centered at 1." Note that the sample semi-variogram is denoted as $\hat{\gamma}(h)$, where the observations are all h units (of time) apart and when the data is unbalanced, a smooth curve, such as a lowess curve can be fit to the scatterplot of $\hat{\gamma}(h)$ to detect if the selected model correctly specifies the covariance matrix. If the transformed residuals are centered randomly around 1, ten the covariance structure is correct for the fitted model.

- Marginal Models
    - Definition (in general)

According to Fitzmaurice et. al[2], "the need to distinguish models according to the interpretation of their regression coefficients has led to the use of the terms 'marginal models' and 'mixed effects models'; the former are often referred to 'population-average models,' the latter as 'subject-specific' models." The former models will be discussed below. Marginal model indicates that the mean response model is only dependent on the covariates of interest. That is, there is no dependence on random effects or responses recorded previously.

Marginal models are similar to the other types of models, except that they do not require observations to have the same count of measurements made repeatedly and that those repeated measurements do not have to be simultaneously collected for all observations, making this model accommodating to unbalanced longitudinal datasets. Although the longitudinal data structure and notation are the same as denoted above in equation 38, the only difference is that the responses Y_i for the n_i repeated measurements are not assumed to be continuous. They can be binary, ordinal, or distributed in some other way mentioned above.

Marginal models make inferences mainly about the population mean of the longitudinal data. There are three parts to the specification of the marginal model for a longitudinal dataset. First, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, the conditional mean of each response is dependent on the covariates X_{ij} through the link function defined above for GLMs: $g(\mu_{ij}) = X_{ij}^T\beta$. Second, as denoted as above for GLMs, the variance of each of the responses Y_{ij} depends on the mean $Var(Y_{ij}|X_{ij}) = \phi * v(\mu_{ij})$. (For balanced designs, the scale parameter ϕ_j is used.) Lastly, Fitzmaurice et. al[2] "the conditional within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, α (and also depends on the means, μ_{ij})." This last specification represents the biggest extension of GLMs to longitudinal data and recognizes that longitudinal data does not have an independence assumption between the mean repeated responses in the within-individual observations.

- Generalized Estimating Equations (GEE)

One of the main advantages of using marginal models is that there is no distributional assumption, vs. models like the mixed and fixed models. According to Fitzmaurice et. al[2], "to avoid distributional assumptions for Y_i we would apply the method of estimation known as generalized estimating equations (GEE)." GEE is an alternative to the MLE approach. This method is a convenient method for modeling the mean and pairwise within-subject association structure. GEE is a good method of estimating marginal models.

Suppose a marginal model that's specified in the above section on marginal models. The GEE estimator for the regression parameter β for the marginal model or for the generalized linear models for longitudinal data is a result of minimizing the following this objective function with respect to β. Such an objective function is denoted as in equation 91.

$$\sum_{i=1}^{N}\{y_i - \mu_i(\beta)\}^T V_i^{-1}\{y_i - \mu_i(\beta)\} \text{ (equation 91)}$$

The μ_i is simply the vector of mean responses. Such a vector has elements in it denoted as "$\mu_{ij} = \mu_{ij}(\beta) = g^{-1}(X_{ij}^T\beta)$," according to Fitzmaurice et. al[2].

The symbol denoted as "V_i is known as a 'working' covariance matrix to distinguish it from the true underlying covariance among the Y_i," according to Fitzmaurice et. al[2], and is assumed to be known. According to Fitzmaurice et. al[2], "the 'working' covariance matrix [means] that V_i approximates the true underlying covariance matrix for Y_i; that is $V_i \approx Cov(Y_i)$…the diagonal entries of V_i are the variances and the off-diagonal terms are the 'working' covariances." If α represents the pairwise correlations among the responses and scale parameter ϕ is defined as it is in the marginal section above, then the V_i is a function of $\alpha, \phi,$ and β.

By taking the first partial derivative of the objective function with respect to β, the generalized estimating equation vector is denoted as

$$\sum_{i=1}^{N} D_i^T V_i^{-1}(y_i - \mu_i) \text{ (equation 92)}$$

where the deviance matrix $D_i = \partial\mu_i/\partial\beta$. The generalized estimating equations are "functions of β and α," as said by Fitzmaurice et. al[2]. With that said, a two-stage iterative estimation procedure for $\alpha, \phi,$ and β is:

1. V_i is estimated when current estimates of $\alpha \; and \; \phi$ are given
    a. The β estimate is updated and is therefore the solution to the GEE.
2. With the newest estimate of β obtained in (1.), the estimates of $\alpha \; and \; \phi$ can be found.
    a. With the defined standardized residuals $e_{ij} = \left(Y_{ij} - \hat{\mu}_{ij}\right)/\sqrt{v\left(\hat{\mu}_{ij}\right)}$:
        i. $\hat{\varphi} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} e_{ij}^2}{\sum_{i=1}^{N} n_i}$
        ii. In balanced designs, $\alpha_{jk} = \left(\frac{1}{\hat{\varphi}N}\right)\sum_{i=1}^{N} e_{ij}e_{ik}$

Note that, regarding the marginal models, this GEE method can be used for example for when there is an ordinal response (See Fitzmaurice 356-357), such as a binary outcome.

- Missing Data
    - Issues with Missing Data

There are almost always missing data that need to be accounted for in longitudinal data sets. According to Fitzmaurice et. al[2], "in longitudinal studies in the health sciences, missing data are the rule, not the exception". Such datasets are said to be unbalanced, as mentioned above. When missing data is in a dataset, "there will be a loss of information and a reduction in the precision with which changes in the mean response over time can be estimated." Such precision reduction is related to how much of the data is missing, where more missing data means more reduction in precision.

According to Fitzmaurice et. al[2], "a missing data mechanism is a probability model for distribution of a set of response indictor variables." There are three models for modeling missing data: Missing at Random (MAR), Missing Completely at Random (MCAR), and Not Missing at Random (NMAR). Each model has its own assumptions about whether the missing values are related to the response(s) or whether the response is observed or unobserved.

There are two types of random missing data. According to Fitzmaurice et. al[2], "the distinction between these two mechanisms determines the appropriateness of maximum likelihood estimation under the assumption of a multivariate normal distribution for the responses and GLS without requiring assumptions about the shape of the distribution". The two types of missing data are MCAR data and MAR data. As stated by Fitzmaurice et. al[2], "MCAR and MAR are often referred to as ignorable mechanisms." However, there is a trade-off of using the "ignorable" term. That is, when there is MAR longitudinal data, it should not always be ignored or left as is. The missing values in the dataset can be imputed from the known observed data.

MCAR data is data that has missing data with the probability that such missing responses are not related to values that should have been collected or the set of responses that were observed by the collector. In other words, "that is, longitudinal data are MCAR when missingness in Y; is simply the result of a chance mechanism that does not depend on either observed or unobserved components of Y_i," according to Fitzmaurice et. al[2]. The consequences of longitudinal data having the MCAR data type is the longitudinal inferences will be valid still, even with such missing data. Without any assumptions of the distribution of Y_i, and given estimates of the means, variances, and covariances, valid estimates of β can be obtained through the GLS estimator. Such an estimator of β "does not require assumptions about the joint distribution of the longitudinal responses" (95). The Maximum Likelihood (ML) and GLS estimators, with whatever the true distribution of Y_i is, have the same properties.

The MAR data, on the other hand, is data that has missing response data with the probability that such missing responses are not related to values that should have been collected. According to Fitzmaurice et. al[2], "put another way, if subjects are stratified on the basis of similar values for the responses that have been observed, missingness is simply the result of a chance mechanism that does not depend on the values of the unobserved responses."

But, the missingness of the data is dependent on the observed responses. The distribution of Y_i of the target population is distinct from the distribution of Y_i where the different strata that's defined by patterns of missingness in the data. When the individuals who have completed the study are only used in the analysis, the analysis is not valid, and the estimates obtained from such data are biased estimates. When the mean response and covariance among the responses are correctly specified, therefore providing the correct multivariate normal distribution, then the ML estimator of β will be considered as a valid estimator.

In addition, it is very important to distinguish between MCAR and MAR. If there are n repeated measurements collected from an individual of the response variable, then set of all responses is denoted by $Y_i = (Y_{i1}, \dots, Y_{in})^T$, an nx1 vector. However, there will likely be some missing dataset in a realistic sense. Letting $R_i = (R_{i1}, \dots, R_{in})^T$, then if the $R_{ij} = 1$ if Y_{ij} is an observed variable. But, if Y_{ij} is not observed, then $R_{ij} = 0$. Using the R_i (an nx1) vector, the vector responses Y_i is split into two subsets: the subset of the observed (Y_i^O) and the subset of the missing observations (Y_i^M). According to Fitzmaurice et. al[2], "The MAR assumption is far less restrictive on Pr(R_i) than MCAR and may be considered to be a more plausible assumption about missing data in many applications." The default missing data assumption should therefore be the MAR.

- Methods for Dealing with dropout

Drop out occurs when, at some repeated measurement time k, a patient no longer records their response. That is, some values in the subset of the response vector Y_i , $\{Y_{ik}, Y_{ik+1}, \dots, Y_{in}\}$, are missing. So, the $R_{ik} = R_{ik+1} = \cdots = R_{in} = 0$. According to Fitzmaurice et. al[2], "the key issue is whether those who 'drop out' and those who remain in the study differ in any further relevant way." If they do differ, then the there is likely to be bias in the complete case analysis. The dropout of the individual(s) can be random, completely random, or not random, as described above for the missing value types (MAR, MCAR, and NMAR). For the latter type of dropout, not at random, is also known as informative dropout because the dropout was meaningful in context of the research study.

Suppose there are 5 repeated measurements for each of the N individuals in a study. Then, according to Fitzmaurice et. al[2], "Suppose that repeated measurements Y_{it} $(i = 1, \dots, N; t = 1, \dots, 5)$ are generated from a multivariate normal distribution with mean response $E(Y_{it}) = \mu_{it} = \beta_1 + \beta_2 t$ and covariance $Cov(Y_{is}, Y_{it}) = \rho^{|s-t|}$, for $\rho \geq 0$." When a dropout occurs for an individual in the dataset, the response indicator vector R_{it}, $t = 1, \dots, 5$, is replaced with D_i, a dropout indicator variable, which is equal to k when an individual who drops out of the study in between the measurement occasions (k-1) and k. So, in other words, there are $D_i - 1$ observed responses for an individual who dropped out.

The probability that a dropout occurs at any of the repeated measurement occasions, given that this is the first time the dropout has occurred, is denoted in this context as in equation 93.

$$\log\left\{\frac{\Pr\,(D_i=k|D_i\geq k,Y_{i1},\ldots,Y_{ik})}{\Pr\,(D_i>k|D_i\geq k,Y_{i1},\ldots,Y_{ik})}\right\}=\theta_1+\theta_2(Y_{ik-1}-\mu_{ik-1})+\theta_3(Y_{ik}-\mu_{ik})$$ (equation 93)

Suppose that the first response (k = 1) in the measurement occasion has been completely observed. That is, $\Pr(D_i=1|D_i\geq 1,Y_{i1},\ldots,Y_{ik})=0$. For the MCAR dropout situation, $\theta_2=\theta_3=0$. For the MAR dropout situation, $\theta_3=0$. And last but not least, for the NMAR dropout situation, $\theta_3\neq 0$.[2]

- Using Multiple Imputation and Weighting Methods

There are several common methods to dealing with missing values in general in datasets and with the specific case of dropout. These include multiple imputation and weighting methods. There are also lesser-used such as complete-case analysis and available data analysis. The traditional methods are complete-case analysis and imputation. These traditional methods are often used in situation where there is a need for a complete and balanced dataset to analyze the data.

Imputation is the more common method of dealing with missing data. Imputation occurs when the missing values of a variable in a dataset are filled in with a value that summarizes the variable, such as a mean or median. According to Fitzmaurice et. al[2], "in multiple imputation the missing values are replaced by a set of m plausible values, thereby acknowledging the uncertainty about what values to impute for the missing responses." Such m filled-in data sets yield a total count of m distinct sets of parameter estimates. The standard errors are also therefore found for each of their parameter estimates via multiple imputation. Usually, if the number of imputations m is small, between perhaps 10 and 30, the estimates will be realistic at approximating the sampling variability.

Complete-case analysis is "performed by excluding any subjects that do not have data at all intended measurement occasions," according to Fitzmaurice et. al[2]. This approach to handling missing values is less commonly used because it is problematic and is inefficient. It leads to reduced statistical power in analysis. In addition, the available-data analysis is more efficient than the complete-case analysis. The available-data analysis "incorporates vectors of repeated measures of unequal length in the analysis," according to Fitzmaurice et. al[2]. Such analyses use partial information taken from those who dropout. However, a trade-off with the use of available-data analysis is that the estimates for the mean responses that are produced will be biased.

Lastly, weighted methods are another alternative to dealing with missing values. According to Fitzmaurice et. al[2], "in weighting methods, the under-representation of certain response profiles in the observed data is taken not account and corrected" (509). A single weight, denoted as in equation 94, is calculated for each of the observations who did not dropout of the study.

$$w_i=\{\Pr\,(D_i=n+1)\}^{-1}=\{\Pr(D_i>1|D_i\geq 1)\times\ldots\times\Pr\,(D_i>n|D_i\geq n)\}^{-1}=(\pi_{i1}\times\ldots\times\pi_{in})^{-1}$$

(equation 94)

Here, given the information on all data up to the $(k-1)^{st}$ occasion, the $(k-1)^{st}$ occasion's remaining observations are used to calculate $\pi_{ik} = \Pr(D_i > k | D_i \geq k)$.

- Longitudinal data analysis in R Framingham heart study data.

The Framingham heart study was a major epidemiological study. According to the official Framingham heart study documentation[4],

> "The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects" (1).

The Framingham heart study was completed in Framingham, Massachusetts and data that is being used in the analysis is a subset of the original data. This collected data runs from 1956 to 1968. There are three different times that repeated measurements were taken for each patient, each time 6 years apart.[4]

The data collected were meant to document any issues patients had with Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction or Cerebral Hemorrhage (Stroke) or death, in addition to the basic variables collected in medical datasets, such as age, diabetes status, glucose levels, heart rate, and time. Thus, the primary variables that will be looked at in the longitudinal data analysis are in table 2 below, along with the summary statistics or counts of the categories:

variable name	description	summary
RANDID	The random id chosen for each subject in the heart study, to protect their identities.	
Age	The age of the subject in years	AGE Min. :32.00 1st Qu.:48.00 Median :54.00 Mean :54.79 3rd Qu.:62.00 Max. :81.00
SEX	the sex of the subject F/M	SEX freq 1 1 5022 2 2 6605
Time	Number of days since baseline exam	TIME Min. : 0 1st Qu.: 0 Median :2156 Mean :1957 3rd Qu.:4252 Max. :4854
Period	one of the 3 repeated measurement times that the subject went in to get	> count(framingham$PERIOD) x freq 1 1 4434 2 2 3930 3 3 3263

	measured on a variety of variables.	
Prevchd	Prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina) (binary 0 - no/1 – yes)	PREVCHD freq 1 0 10785 2 1 842
Mi_fchd	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease (binary – yes/no)	> count(framingham$MI_FCHD) x freq 1 0 9839 2 1 1788
Timemifc	Defined as above for the first MI_FCHD event during follow-up	TIMEMI Min. : 0 1st Qu.:7212 Median :8766 Mean :7594 3rd Qu.:8766 Max. :8766
HEARTRTE	Heart rate (Ventricular rate) in beats/min	HEARTRTE Min. : 37.00 1st Qu.: 69.00 Median : 75.00 Mean : 76.78 3rd Qu.: 85.00 Max. :220.00 NA's :6
CURSMOKE	if the patient currently smokes cigarettes during the repeated measurement	CURSMOKE freq 1 0 6598 2 1 5029
DIABETES	diabetes status (Y/N)	DIABETES freq 1 0 11097 2 1 530
GLUCOSE	Casual serum glucose (mg/dL)	GLUCOSE Min. : 39.00 1st Qu.: 72.00 Median : 80.00 Mean : 84.12 3rd Qu.: 89.00 Max. :478.00 NA's :1440

Table 2: the variables of interest in the Framingham Heart Study

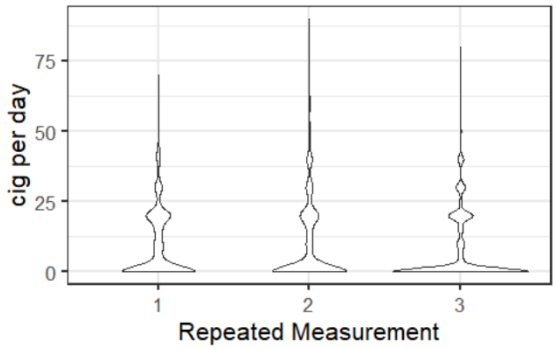


Figure 54: violin plots of cigarettes per day vs. repeated measurement

Now, violin plots of cigarettes per day, glucose, and heart rate vs. repeated measurement In Figure 54 above of the number of cigarettes per day vs. the repeated measurement time, notice how a lot of the people who smoke between 0 and 5 went up by repeated measurement three. The violin plots are less dense because some people dropped out or were lost to follow-up.

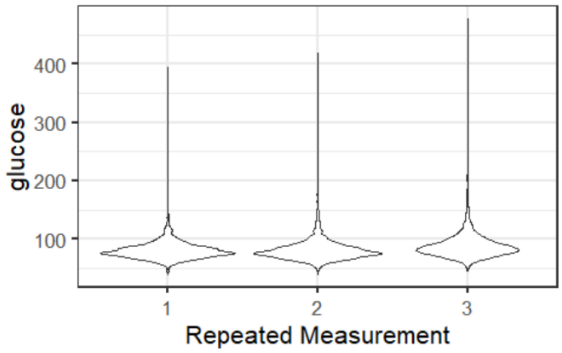


Figure 55: violin plots of glucose vs. repeated measurement

For each repeated measurement occasion in Figure 55 above of glucose level, it appears that over time, the overally glucose levels went up. For each repeated measurement occasion in Figure 56 below of heart rate, it appears that over time, the overall heart rate level stayed about the same.

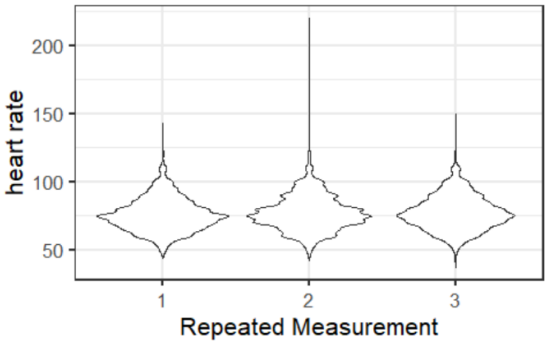


Figure 56: violin plots of heart rate vs. repeated measurement

Heart rate and glucose levels will be the primary variables of interest. The line plots of heart rate and glucose levels are shown in Figure 57 and 58 below. Notice how, in the heart rate line plot below, most of the lines that connect all of the are all between about 50 and 115 beats/min and stay consistent over time. On the contrary, in the glucose level line plot, there are ample outliers and many of the glucose levels go up over time or are very inconsistant over time.

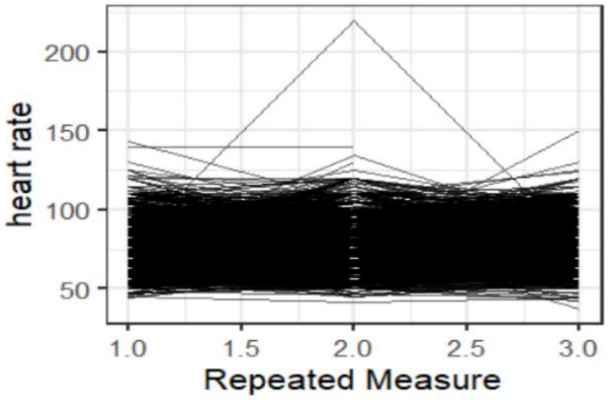


Figure 57: line plot of heart rate vs. repeated measurement occasion

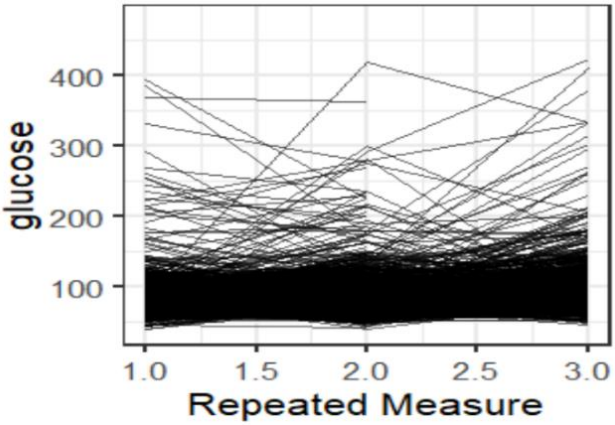


Figure 58: line plot of glucose vs. repeated measurement occasion

The random intercept term model for when the outcome is heart rate is shown in Figure 59 below. The fixed effect is period (a.k.a. the repeated measurement occasion) and the random effect intercept term by the random identification assigned to get subject. The model converges, according to the REML criterion at convergence. The scaled residual statistics are also given for the model, where the median residual is nearly zero and the max reisual is a major outlier of about 12. In addition, the random and fixed effects summaries are given.

The random effect summary shows the estimates of the amount of variability for the random intercept and the random with individual error, labeled as “residual”. There is a not much variability in the starting heart rate in beats/min for each subject in the beginning of the study which isn’t surprising given the observations made about the violin plots and line plots above. Larger variability values indicate more variability about the aggregate and the individual’s

trajectories for the random intercepts and within individual error, respectively. On the contrary, the fixed effects summary shows the average trends across all individuals in the data. The average starting place, labeled the intercept, and the average change in heart rate for each repeated measurement occasion can be thought of as a one unit change in the time metric.

```
Linear mixed model fit by REML ['lmerMod']
Formula: HEARTRTE ~ 1 + PERIOD + (1 | RANDID)
   Data: framingham

REML criterion at convergence: 89244.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7560 -0.5688 -0.0613  0.5054 12.0613

Random effects:
 Groups   Name        Variance Std.Dev.
 RANDID   (Intercept) 77.03    8.777
 Residual             78.58    8.865
Number of obs: 11621, groups:  RANDID, 4434

Fixed effects:
            Estimate Std. Error t value
(Intercept)  75.1068     0.2495 300.992
PERIOD        0.9699     0.1048   9.258

Correlation of Fixed Effects:
       (Intr)
PERIOD -0.778
```

Figure 59: random intercept term model for heart rate

The random slope term model is a more common model for longitudinal data bcause it is more realistic and allows each of the individuals in the dataset to have their own trajectory. According to DataCamp[1], the random slope model also "model the dependency due to the repeated measures more adequently, like decreasing correlation as time lag increases." Such a model for when the outcome is heart rate is shown in Figure 60 below. Just as in the random intercept model above, the fixed effect is period (a.k.a. the repeated measurement occasion) and the random effect slope and intercept term by the random identification assigned to get subject. Once again the model converges, according to the REML criterion at convergence. The scaled residual statistics are also given for the model, where the median residual is nearly zero and the max residual is a major outlier of about 12. In addition, the random and fixed effects summaries are given.

The random effect summary shows the estimates of the amount of variability for the random slope and intercept and the random within-individual error, labeled as "residual". In the random effects output, the PERIOD term indicates that there is not a lot of variation in the trajectories of the indviduals in the study because the variance is about 4.3, the std. dev. is about 2.08, and the correlation is nearly zero. Generally in longitudinal models, variation in the random slopes is much smaller than the intercepts DataCamp[1]. The analysis of the random intercept follows through the same as the above for the fixed effects results.

```
Linear mixed model fit by REML ['lmerMod']
Formula: HEARTRTE ~ 1 + PERIOD + (1 + PERIOD | RANDID)
   Data: framingham

REML criterion at convergence: 89225.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4645 -0.5526 -0.0614  0.4866 12.3100

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 RANDID   (Intercept) 82.041   9.058
          PERIOD       4.318   2.078    -0.27
 Residual             74.481   8.630
Number of obs: 11621, groups:  RANDID, 4434

Fixed effects:
            Estimate Std. Error t value
(Intercept)  75.0910     0.2478 303.056
PERIOD        0.9824     0.1078   9.113

Correlation of Fixed Effects:
       (Intr)
PERIOD -0.774
convergence code: 0
Model failed to converge with max|grad| = 0.00287297 (tol = 0.002, component 1)
```

Figure 60: random slope term model for heart rate

To answer the question of which of these models – the random intercept model and the random slope model – fits better, the ANOVA is used to produce such results to make the comparison. There are other methods, but the ANOVA method is the most commonly used. The ANOVA output will include Akaike information criterion (AIC), Bayesian information criterion (BIC), and log-likelihood values that are minimized during the estimation of the model's parameters using the maximum likelihood estimation method. In all such methods of selecting the best model, the smaller the statistic, the better. AIC and BIC are based on the log-likelihood function, but have more penalities for the amount of parameters in the model that are being estimated and are meant to prevent overfitting of the model, with the BIC being the stricter method. The AIC statistic is "recommended when the true model is not included in the comparison" according to DataCamp[1]. With empirical data, the true model is usually never known, therefore the AIC will be focused on in the selection of the best fitting model on the longitudinal data. As, note that nested model comparisons are also often compared using the ANOVA output.

The results for the ANOVA for the heartrate output models of the random intercept model and the random slope model are shown in Figure 61 below. The AIC, BIC, and log-likelihood of the latter model is slightly smaller than the former. Therefore, the random slope model is yields a better fit to the longitudinal data. Notice that the chi-square test that compares the to models is statistically signficant, which indicates that the two models are different in regards to how well they fit the data.

```
Data: framingham
Models:
HEARTRATE_ri: HEARTRTE ~ 1 + PERIOD + (1 | RANDID)
HEARTRATE_ri2: HEARTRTE ~ 1 + PERIOD + (1 + PERIOD | RANDID)
              Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
HEARTRATE_ri   4 89248 89278 -44620    89240
HEARTRATE_ri2  6 89233 89277 -44610    89221 19.774      2  5.082e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 61: AIC, BIC, and log-likelihood for heart rate models

There are some neat ways of presenting statistical results in a visual way. Such methods are helpful to understanding and checking if the interpretations of the results from the model are correct. The linear mixed effect model adjusts for dependency in the dataset by introducing random effects. By creating a custom function for looking at the correlation structure to explore the correlations implied in the model, the correlation structures of the models that have the random slopes or random intercepts can be compared. The correlation structure function is shown in Figure 62 (a) below.

In the case of the heart rate models produced above, the custom function corr_structure was used to generate the model implied correlations from the repeated measurements for the 3 occasions they were taken at for the first person in the study, as shown in Figure 62 (b) below. As you can see, the correlations across the measurements are constant and moderate (about 0.5); that is, not close to one. Notice however, that the correlations for the random slope model are slightly higher than those of the random interept model, which is consistant with the above results. The constant correlation over time is referred to as, in statistics, the compound correlation. The correlation stuctures for the random intercept model (a) and the random slope model (b), respectively can also be looked at in the plots in Figure 63 below as well. The analysis for such plots is the same as that for the corr_stuctute function output.

```
> #Compound Symmetry:
> corr_structure <- function(object, num_timepoints, intercept_only = TRUE) {
+   variance <- VarCorr(object)
+   if(intercept_only) {
+     random_matrix <- as.matrix(object@pp$X[1:num_timepoints, 1])
+     var_cor <- random_matrix %*% variance[[1]][1] %*% t(random_matrix) +
+       diag(attr(variance, "sc")^2, nrow = num_timepoints,
+            ncol = num_timepoints)
+   } else {
+     random_matrix <- as.matrix(object@pp$X[1:num_timepoints, ])
+     var_cor <- random_matrix %*% variance[[1]][1:2, 1:2] %*%
+       t(random_matrix) + diag(attr(variance, "sc")^2,
+                               nrow = num_timepoints, ncol = num_timepoints)
+   }
+   Matrix::cov2cor(var_cor)
+ }
```

```
> #the custom function corr_structure
> #is used to generate the model implied
> #correlations from the repeated
> #measurements for the 3 time points
> #for the first person in the study.
> corr_structure(HEARTRATE_ri, 3) %>%
+   round(2)
    1   2   3
1 1.0 0.5 0.5
2 0.5 1.0 0.5
3 0.5 0.5 1.0
> corr_structure(HEARTRATE_ri2, 3) %>%
+   round(2)
     1    2    3
1 1.00 0.52 0.52
2 0.52 1.00 0.52
3 0.52 0.52 1.00
```

Figure 62 (a) left, (b) right.

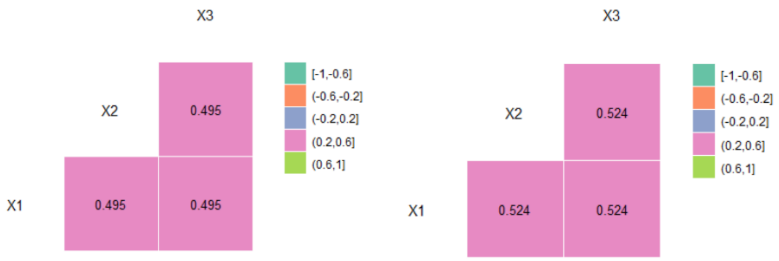


Figure 63 (a) left, (b) right.

Next, some linear mixed effect models will be created with some potiental predictors of the outcome variable heartrate. Adding predictors to the model builds up the fixed effects to help explain variation and differentiate the average trend from other models. In the model fitted in Figure 64 below, the variable SEX is added to the model as a predictor variable.

From the fixed effects output, there are one additional term that was added to the summary when the variable SEX was added. The new terms represents the average deviations from the intercept starting heart rate in beats/minute, for females. In this model, as the PERIOD variable starts at zero, the intercept represents the average heart rate for female subjects to be about 75 beats/min. To find the average heart rate at period 0 for males, the intercept and the estimated coefficient of the SEX are added to yield 74.66 + 2.52 = 77.12 or about 77 beats/min.

The random effects summary for the model also says that there's an ample amount about the model. The mentioned summary shows the variables of the intercepts and slopes after the fixed effects are adjusted for. The variation in the random intercepts in the output in Figure 64 below when compared to the output in Figure 60 above, where sex is not added to the model as a potential predictor are 80.178 and 76.33, respectively. This represents a 1% increase in the variation of the intercepts from the addtion of the introduction of sex into the model as a potential predictor. The variable sex did not explain much of the variation in the intercepts since the variance of the intercept went slightly up from the random slope model without sex.

```
Linear mixed model fit by REML ['lmerMod']
Formula: HEARTRTE ~ 1 + PERIOD + SEX_f + (1 + PERIOD | RANDID)
   Data: framingham

REML criterion at convergence: 89161.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4823 -0.5497 -0.0642  0.4877 12.2975

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 RANDID   (Intercept) 80.178   8.954
          PERIOD       4.303   2.074    -0.27
 Residual             74.519   8.632
Number of obs: 11621, groups:  RANDID, 4434

Fixed effects:
            Estimate Std. Error t value
(Intercept)  73.6820     0.3029 243.218
PERIOD        0.9738     0.1078   9.034
SEX_fSEX 2    2.5226     0.3138   8.039

Correlation of Fixed Effects:
            (Intr) PERIOD
PERIOD      -0.628
SEX_fSEX 2 -0.579 -0.008
```

Figure 64: random slope term model for heart rate with variable sex

An example of a model where an added predictor predicts the response variable well is demonstrated here, where the outcome of interest is glucose level measured in mg/dL. First, the standard random intercept and random slope models were fit without any potential predictor variables added to the model. Second, the predictor diabetes was added to the random slope model to predict glucose.

The random intercept term model for when the outcome is glucose is shown in Figure 65 below. The fixed effect is period (a.k.a. the repeated measurement occasion) and the random effect intercept term by the random identification assigned to get subject. The model converges, according to the REML criterion at convergence. The scaled residual statistics are also given for the model, where the median residual is nearly zero and the max residual is a major outlier of about 12. In addition, the random and fixed effects summaries are given.

```
Linear mixed model fit by REML ['lmerMod']
Formula: GLUCOSE ~ 1 + PERIOD + (1 | RANDID)
   Data: framingham

REML criterion at convergence: 93181.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.4842 -0.4145 -0.0991  0.2642 12.2074

Random effects:
 Groups   Name        Variance Std.Dev.
 RANDID   (Intercept) 286.7    16.93
 Residual             351.8    18.76
Number of obs: 10187, groups:  RANDID, 4366

Fixed effects:
            Estimate Std. Error t value
(Intercept)  77.3148     0.5469  141.36
PERIOD        3.7955     0.2425   15.65

Correlation of Fixed Effects:
       (Intr)
PERIOD -0.810
```

Figure 65: random intercept term model for glucose

The random effect summary shows that there is a moderate amount of variability in the starting glucose in mg/dL for each subject in the beginning of the study. This observation is not surprising given the observations made about the violin plots and line plots above of glucose vs. repeated measurement occasion. Larger variability values indicate more variability about the aggregate and the individual's trajectories for the random intercepts and within-individual error, respectively.

The random slope term model is shown in Figure 66 below. According to DataCamp[1], the random slope model also "model the dependency due to the repeated measures more adequently, like decreasing correlation as time lag increases." Just as in the random intercept model above, the fixed effect is period (a.k.a. the repeated measurement occasion) and the random effect slope and intercept term by the random identification assigned to get subject. Again, the model converges, according to the REML criterion at convergence. The scaled residual statistics are also given for the model, where the median residual is nearly zero and the max residual is a major outlier of about 10.7. In addition, the random and fixed effects summaries are given.

The random effect summary shows the estimates of the amount of variability for the random slope and intercept and the random with individual error, labeled as “residual”. In the random effects output, the PERIOD term indicates there are is a lot of variation in the trajectories of the individuals in the study because the variance is about 452.52, the std. dev. is about 21.27, and the correlation is nearly zero. The analysis of the random intercept follows through the same as the above for the fixed effects results.

```
Linear mixed model fit by REML ['lmerMod']
Formula: GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID)
   Data: framingham

REML criterion at convergence: 92991.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.5743 -0.3745 -0.0785  0.2483 10.6723

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 RANDID   (Intercept) 452.52   21.27
          PERIOD       77.98    8.83    -0.59
 Residual             279.61   16.72
Number of obs: 10187, groups:  RANDID, 4366

Fixed effects:
            Estimate Std. Error t value
(Intercept)  77.3734     0.5486  141.03
PERIOD        3.7475     0.2663   14.07

Correlation of Fixed Effects:
       (Intr)
PERIOD -0.815
```

Figure 66: random slope term model for glucose

The results for the ANOVA for the glucose output models of the random intercept model and the random slope model are shown in Figure 67 below. The AIC, BIC, and log-likelihood of the latter model is slightly smaller than the former. Therefore, the random slope model is yields a better fit to the longitudinal data. Notice that the chi-square test that compares the to models is statistically signficant, which indicates that the two models are different in regards to how well they fit the data.

```
> anova(GLUCOSE_ri, GLUCOSE_ri2)
refitting model(s) with ML (instead of REML)
Data: framingham
Models:
GLUCOSE_ri: GLUCOSE ~ 1 + PERIOD + (1 | RANDID)
GLUCOSE_ri2: GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID)
            Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
GLUCOSE_ri   4 93188 93217 -46590    93180
GLUCOSE_ri2  6 93002 93046 -46495    92990 189.89      2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 67: ANOVA comparing models from Figures 65 and 66

In the case of the glucose models produced above, the custom function corr_structure was used to generate the model implied correlations from the repeated measurements for the 3 occasions they were taken at for the first person in the study, as shown in Figure 68 below. As you can see, the correlations across the measurements are constant and moderate (about 0.45); that is, not close to one. Notice however, that the correlations for the random slope model are higher than those of the random intercept model (0.62 and 0.45, respectively), which is

consistant with the above results. The constant correlation over time is referred to as, in statistics, the compound correlation. The correlation stuctures for the random intercept model (a) and the random slope model (b), respectively can also be looked at in the plots in Figure 69 below as well.

```
> corr_structure(GLUCOSE_ri, 3) %>%
+   round(2)
     1    2    3
1 1.00 0.45 0.45
2 0.45 1.00 0.45
3 0.45 0.45 1.00
> corr_structure(GLUCOSE_ri2, 3) %>%
+   round(2)
     1    2    3
1 1.00 0.62 0.62
2 0.62 1.00 0.62
3 0.62 0.62 1.00
```

Figure 68: correlation structures for the random intercept model and the random slope model for glucose

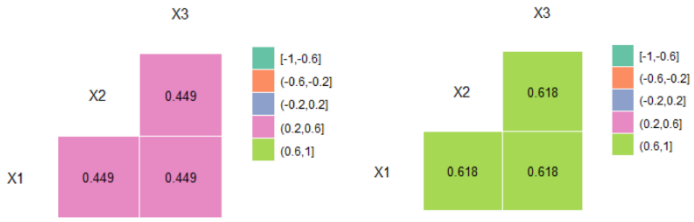


Figure 69 (a) left, (b) right.

In the model fitted in Figure 70 below, the variable DIABETES or diabetes status, was added to the model as a predictor variable. From the fixed effects output, there is one additional term that was added to the summary when this variable was added. The new terms represent the average deviations from the intercept starting glucose amount in mg/dL, for people with no diabetes. In this model, as the PERIOD variable starts at one, the intercept represents the average glucose for non-diabetic subjects to be about 77.4 mg/dL. To find the average glucose at period 1 for subjects with diabetes, the intercept and the estimated coefficient of the DIABETES are added to yield 77.4 + 61.18 = 138.58 or about 139 mg/dL, which is high.

The random effects summary also says an ample amount about the model. The mentioned summary shows the variables of the intercepts and slopes after the fixed effects are adjusted for. The variation in the random intercepts in the output in Figure 70 below when compared to the output in Figure 66 above, where diabetes is not added to the model as a potential predictor are 263.03 and 452.52, respectively. This represents a 50% decrease in the variation of the intercepts from the addition of the introduction of diabetes into the model as a potential predictor. The variable diabetes explained plenty of the variation in the intercepts since the variance of the intercept went slightly up from the random slope model without sex.

```
Linear mixed model fit by REML ['lmerMod']
Formula: GLUCOSE ~ 1 + PERIOD + DIABETES_0 + (1 + PERIOD | RANDID)
   Data: framingham

REML criterion at convergence: 90750.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.6729 -0.3951 -0.0660  0.2984 10.5438

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 RANDID   (Intercept) 263.03   16.218
          PERIOD       80.06    8.948   -0.77
 Residual             283.10   16.825
Number of obs: 10187, groups:  RANDID, 4366

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          77.3773     0.5045  153.36
PERIOD                2.2079     0.2657    8.31
DIABETES_0DIABETES 1  61.1893     1.1354   53.89

Correlation of Fixed Effects:
            (Intr) PERIOD
PERIOD      -0.873
DIABETES_01 -0.006 -0.096
```

Figure 70: random slope term model for glucose with diabetes

BINARY ANALYSIS

When the outcome being looked at is categorical, such as when the outcome variable is binary (has two posible outcomes: 0 or 1, for example for a yes/no question), the analysis is binary. There are specific methods to analyzing such results from the generalized linear mixed effect model that's used. Perhaps, with the given variables in the framingham dataset, the variable MI_FCHD, which indicates whether a subject was Hospitalized for Myocardial Infarction or Fatal Coronary Heart Disease during the study period (see table 2 above) is the outcome variable to be studied. A potential predictor variable for such a variable is Prevchd which indicates whether there is a prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina) in the subject being studied.

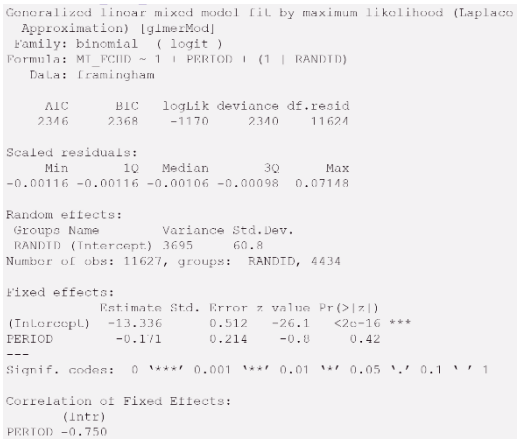
The generalized linear mixed model (GLMM) is used to model the outcome vs. the potenital predictor. It is a more general version of the linear mixed effect model that was used above for situatiosn whrre there is a continuous outcome variable being predicted. That is, the GLMM can handle any type of outcome: categorical or continunous. Such a model explores the log-odds of the success, which is typically coded as outcome = 1. The log-odds represents the probablilty ratio of the outcome equalling 1 vs. the outcome equalling 0, with such a ratio of probabilities then transformed by a log. The family argument n the glmm function is set to binomial since there is only two outcomes for the outcome variable. The output from the GLMM is very similar to that of the outcome from the continuous outcome fitted lmer output from all output Figures above. The primary difference is that the parameter estimates for the fixed effects are in the logistic metric.

Notice that, in the output in Figure 71 below of the GLMM output, as expected, the PREVCHD variable is statistically significant, as is the repeated measurement occasion PERIOD as the random slope and the random intercept term. There are no outliers according to the scaled residual output. In addition, there is some variance in the Hospitalized for Myocardial Infarction or Fatal Coronary Heart Disease status for the subjects in the study because the random effect variance from this model is 2751 and has a standard deviation of 52.45.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: MI_FCHD ~ 1 + PERIOD + PREVCHD + (1 | RANDID)
   Data: framingham

     AIC      BIC   logLik deviance df.resid
  2649.5   2678.9  -1320.7   2641.5    11623

Scaled residuals:
      Min        1Q    Median        3Q       Max
-0.075519 -0.001499 -0.001139 -0.000865  0.081191

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 2751     52.45
Number of obs: 11627, groups:  RANDID, 4434

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.4446     0.4733 -26.291   <2e-16 ***
PERIOD       -0.5494     0.2217  -2.478   0.0132 *
PREVCHD      23.4277     0.7285  32.160   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
        (Intr) PERIOD
PERIOD  -0.741
PREVCHD -0.414 -0.038
```

Figure 71: GLMM output for the MI_FCHD with PREVCHD variable

When the GLMM model with no predictor (PREVCHD) is computed (i.e. the baseline function, as shown in Figure 72 below. Notice that, in the output of the GLMM output, the repeated measurement occasion PERIOD is not sitatistically significant. The random intercept term is statistially significant, however. There are no outliers according to the scaled residual output. In addition, there is some variance in the Hospitalized for Myocardial Infarction or Fatal Coronary Heart Disease status for the subjects in the study because the random effect variance from this model is 3965 and has a standard deviation of 60.8, which is higher than the model with PREVCHD.



```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: MI_FCHD ~ 1 + PERIOD + (1 | RANDID)
   Data: framingham

     AIC      BIC   logLik deviance df.resid
    2346     2368    -1170     2340    11624

Scaled residuals:
     Min       1Q   Median       3Q      Max
-0.00116 -0.00116 -0.00106 -0.00098  0.07148

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 3695     60.8
Number of obs: 11627, groups:  RANDID, 4434

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -13.336      0.512   -26.1   <2e-16 ***
PERIOD        -0.171      0.214    -0.8     0.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       (Intr)
PERIOD -0.750
```

Figure 72: GLMM output for the MI_FCHD without PREVCHD variable

The results of the two GLMM outputs can be compared using the AICc statistics. The output for comparing the statistics is in Figure 73 below. The framework for AICc is similar to AIC used in the continuous outcome LMM models, where the smaller the statistic, the better fit the model is to the data. In the case in Figure 73 below, the results suggest that the presence of PREVCHD may not be that helpful in predicting the MI_FCHD because the model with no PREVCHD in it has a better fit. However, the variance in the model with PREVCHD is smaller and all values are statistically significant. This model is still a good model to fit the data ad predict the outcome MI_FCHD and is very interpretable.

```
> aictab(list(MI_FCHD_baseline, MI_FCHD_output),
+        c("no PREVCHD", "PREVCHD"))

Model selection based on AICc:

            K AICc Delta_AICc AICcWt Cum.Wt    LL
no PREVCHD  3 2346          0      1      1 -1170
PREVCHD     4 2649        304      0      1 -1321
```

Figure 73: AICc of the models in Figures 71 and 72

GEE MODEL FITTING

The GLMM has an alternative model called the generalized estimating equations (GEE) model output. Such models are useful only when the aggregate trend is of interest, instead of subject-specific interpretations. A couple of GEE models were fit to the same forumla with the same outcome variable and predictor variables as in the GLMM output above. The GEE model is a marginal model, as discussed in the GEE section above in more detail, because only the average trend is directly estimated. The output from the GEE has a correlation structure attached which is also known as a working correlation matrix. Such a matrix will help account for any repeated measurement dependency. It is assumed that the repeated measurements are independent however, but that is usually not the case for repeated measurements.

In Figure 74 (b), the augument corstr is equal to the value “exchangeable,” which indicates that a different working correlation matrix will be used. Such a correlation matrix will have equal correlation between all timepoints. This output differs from the original output in Figure 74 (a) because the estimated correlation parameters at the very bottom of the output for both models are different. With the exchangeable working correlation stucture in part (b), the

observation within an individual are not treated as independent, rather they are assumed to have a constant correlation across all of the time points. Both parameter estimates and correlation structures have about the same output though, there are small differences, although nothing too large to cause significant concern.

It often is useful to compare several different correlation structures in the model to explore if there is any impact on the results and if the consistancy is therefore broken. Other specifications for the correlation structure are “ar1”, first order autoregressive structure (see Figure 74 (c)) and “unstructured” (see Figure 74 (d)). The former has decreasing correlation as the time lag between observations increases, which is common in longitudinal datasets. The latter will estimate a unique correlation for each time lag, which can be data intensive with many time points. In all of the results in Figure 74, the parameter estimates and correlation structures have about the same output though, there are small differences, although nothing too large to cause significant concern.

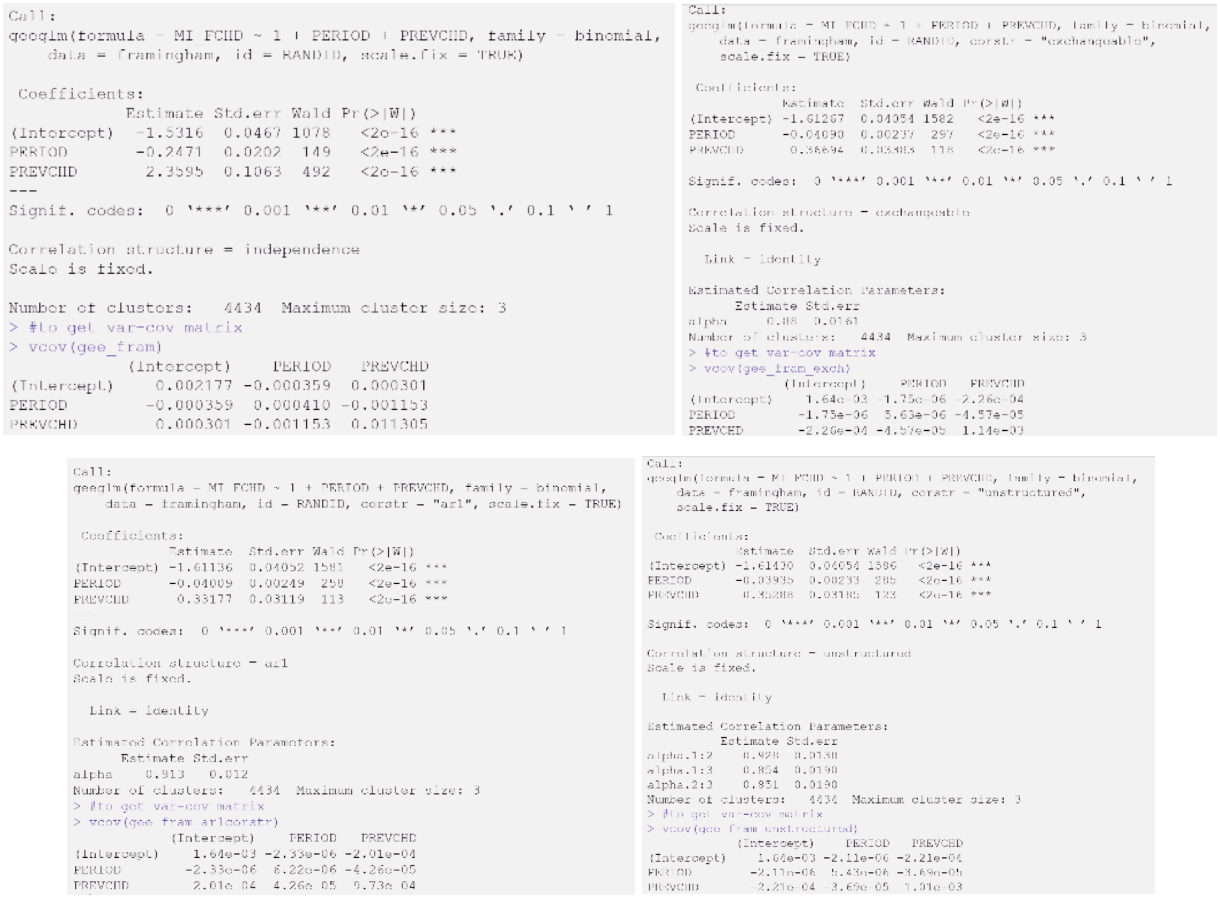


Figure 74 (a) top left, (b) top right, (c) bottom left, (d) bottom right.

Which GEE model is the best? There are several ways to select the best model. The QIC statistic stands for quasi-likelihood under the independence model criterion. The QIC statistic is needed for GEE models since GEE does not use maximum likelihood estimation like the GLMM model. Similar to the AIC, QIC is scaled so that smaller values indicate better model fit and it is useful when selecting working correlation. The model with the smallest value of the QIC statistic

is considered to be the best model. Taking a look at Figure 75 below, the four models' QIC scores from Figure 75 (a), (b), (c), and (d) are shown below. The models with exchangeable, ar1, and unstructured are all tied on the value of QIC as the smallest value. Although relatively subjective when picking the best model, the gee_ar1corstr model was selected because the estimated correlation parameter is the highest of all of the models

	QIC
gee_fram	9078
gee_fram_exch	9073
gee_fram_ar1corstr	9073
gee_fram_unstructured	9073

Figure 75: QIC scores for the four GEE models

GLMM and GEE models are not able to be directly compared. Instead, the type of model (GLMM and GEE) should be fit on the data based on the purpose of the model and the ultimate research goal. Although sibjective and debatable, in the case of the predication of outcome MI_FCHD, the best GLMM that would be the best model to go with because the random effects are meaningful to compare individuals with each other. The GLMM model with PREVCHD would probably be the overall best model in terms of fit and interpretability.

Conclusions on Longitdinal Data Analysis

So, longitudinal data analysis is very common in clinical trials and medical research to help look at re[eated measurement data of individuals responses over time. Next, the two powerful time-dependent modelling techniques discussed will be put into one common model. Such a model works with data with both time-to-event and repeated measurement data.

Joint Modeling of Survival and Longitudinal Models

Definition

The time to event data used in Survival analysis models and repeated measurements that longitudinal data analysis models use can be combined in a joint model to tell a more detailed story of the data, for example, cancer data with biomarker variables recorded for some subjects in a study looking at comparing cancer treatments. According to Ibrahim et. al.[6], "Joint models for longitudinal and time-to-event data are model that bring these two data types together (simultaneously) into a single model so that one can infer the dependence and association between the longitudinal biomarker and time to event to better assess the effect of a treatment." Joint modeling is common in studies quality-of-life (QOL) that study quality of life while recording the time-to-event end point and repeated responses. Using the cancer example stated above, Ibrahim et. al.[6] says "in cancer vaccine trials, immunologic measures such as immunoglobulin G or immunoglobulin M response are often measured longitudinally, and it is of interest to examine their association with time to event."

Why they are used.

According to Ibrahim et. al.[6], "joint models are increasingly used in clinical trials because (1) they provide more efficient estimates of the treatment effects on the time to event, (2) they provide more efficient estimates of the treatment effects of the longitudinal marker, and (3)

they reduce bias in the estimates of the overall treatment effect, that is, the treatment effect on survival and the longitudinal marker." When the bias is minimized on an estimate or statistic of a parameter, such as the treatment effect, the estimate becomes a better estimate for the parameter. In addition, the incorporation of the longitudinal data into the design of a study, such as a clinical trial, could possibly yield the study to have smaller sample sizes and higher power. Joint models produce more accurate and precise treatment effect estimates when compared to the results of, for example, Cox PH model and longitudinal models alone.

Notation

If T_i is the observed failure time for subject i and C_i is the censoring time, then $T_i = \min(T_i^*, C_i)$. That is, the observed failure time of subject i is the minimum of the true event time (T_i^*) and the censoring time. Next, the indicator function denoted as in equation 95, which is the event indicator.

$$\delta_i = I(T_i^* \leq C_i) \text{ (equation 95)}$$

When $T_i^* \leq C_i$, $\delta_i = 1$ and when not $\delta_i = 0$. Therefore, according to Rizopoules, D.[11], "the observed data for the time-to-event outcome consist of the pairs {(T_i, δ_i), I = 1, …, n}." Let equation 96 be the trajectory function, or the function of time that contains the true unobserved longitudinal process.

$$M_i(t) = \{m_i(u), 0 \leq u \leq 1\} \text{ (equation 96)}$$

The longitudinal model part of the joint model usually used is the linear mixed effects model, as discussed above in the linear mixed effects section. Recall that the linear mixed effects equation is denoted as in equation 97, with the parameters and variables defined as above.[11]

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i = m_i(t) + \varepsilon_i \text{ (equation 97)}$$

That is, β is the fixed effect vector, b_i is the random effects vector (with a MVN distribution), and ε_i is the random error vector (N(0, σ^2) distributed). The X_i is the matrix of covariates and Z_i is the design matrix that links the vector of random effects b_i to Y_i, the vector of the responses. The latter vector is also called the trajectory function.

The survival portion of the joint model is typically a parametric distribution such as a Weibull model. The survival model's hazard function at time t is denoted as in equation 98, with the υ being the direct treatment effect on the time to event.

$$h(t) = h_i(t|M_i(t), Z_i) = lim_{dt\to 0} \Pr\{t \leq T_i^* < t + dt | t \geq T_i^*, M_i(t), Z_i\}/dt = h_0(t)e^{\omega(X_i\beta + Z_i b_i) + \upsilon Z_i} = h_0(t)e^{\omega(m_i(t)) + \upsilon Z_i} \text{ (equation 98)}$$

The ω is the measurement of "the association between the longitudinal marker and the time to event," according to Ibrahim et. al.[6]. Notice that the trajectory function Y_i connects the longitudinal to the survival model to make the joint model. So that there is no confusion with the number of parameters in this equation, it is important to clarify the parameters. The three different treatment effects in this equation of $h(t)$. (1) β, fixed effect vector and the "treatment effect on the longitudinal marker"; (2) υ, the time to event treatment effect; and (3) $\omega\beta + \upsilon$, the

overall treatment effect.[6] A visualization of how this can be shown is in Figure 76 below (see Appendix for the translated notation explanations of the joint model from the different author's models).

If $\omega = 0$ (if the estimate of ω is nearly zero) then the longitudinal marker and the event time are not associated. This implies that "information from the longitudinal marker does not improve on the estimate for the survival treatment effect υ compared with an analysis based on the time-to-event data alone." If this occurs, then a joint model is not needed. When $\omega < 0$, the hazard defined above is decreasing. This implies that when the longitudinal marker is increased, the time to event also increases.[6]

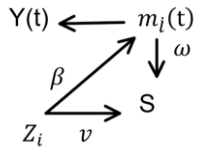


Figure 76: Causal diagram - Y(t) is the observed longitudinal data; $m_i(t)$ is the trajectory function; Z_i is the treatment, S is survival, β is treatment effect on the longitudinal process; υ is the treatment effect on the survival; ω is the effect of longitudinal process on survival. (Visual Analog of Figure 1 in Ibrahin et. Al. translated into Schveder's notation)

To estimate the joint model defined above in equation 98, maximum likelihood estimation can be used, or the MCMC technique with Bayes theorem can be applied. The former is one of the traditional approaches. According to Rizopoules, D.[11], "Maximum likelihood estimation for joint models is based on the maximization of the loglikelihood corresponding to the joint distribution of the time-to-event and longitudinal outcomes $\{T_i, \delta_i, Y_i\}$." It is assumed that, for both longitudinal and survival processes, the vector of time-independent random effects (i.e. b_i) is the underlying effect. Such an assumption, in other words, means the random effects in vector b_i "account for both the association between the longitudinal and event outcomes", in addition to the longitudinal process's repeated measurement correlations, as stated by Rizopoules, D.[11] The conditional independence of the longitudinal outcomes and the parameter vector $\theta = (\theta_t^T, \theta_y^T, \theta_b^T)$ with b_i is therefore present. The first item in the vector θ is the event time outcome vector; the second item is a vector for the longitudinal outcomes; the third item is a vector of the distinct parameters of the random-effects covariance matrix. The conditionally independent probability is denoted as follows in equation 99.

$$p(T_i, \delta_i, Y_i|b_i; \theta) = p(T_i, \delta_i|b_i; \theta)p(Y_i|b_i; \theta), \text{ with } p(Y_i|b_i; \theta) = \prod_j p\{Y_i(t_{ij})|b_i; \theta\}$$

(equation 99)

Furthermore, the joint log-likelihood is denoted as in equation 100.

$$\log\left(p(T_i, \delta_i, Y_i|b_i; \theta)\right) = log \int p(T_i, \delta_i|b_i; \theta_t, \beta)[\prod_j p\{Y_i(t_{ij})|b_i; \theta\}]\, p(b_i; \theta_b)db_i$$

(equation 100)

The likelihood of the survival portion of the joint model is denoted as follows in equation 101 and 102.

$$p(T_i, \delta_i | b_i; \theta_t, \beta) = \{h_i(T_i | M_i(T_i);\ \theta_t, \beta)\}^{\delta_i} S_i(T_i | M_i(T_i); \theta_t, \beta) \quad \text{(equation 101)}$$

where

$$S_i(T_i | M_i(T_i), Z_i; \theta_t, \beta) = P(t < T_i^* | M_i(T_i), Z_i; \theta_t, \beta) = e^{-\int_0^t h_i(s | M_i(s);\ \theta_t, \beta) ds} \text{ (equation 102)}$$

The longitudinal portion of the log-likelihood equation is denoted as in equation 103, and "is the univariate normal density for the longitudinal responses."

$$p\{Y_i(t_{ij}) | b_i; \theta\} \text{ (equation 103)}$$

Lastly, the portion denoted as in equation 104 is called the multivariate normal density of the random effects in the b_i vector.

$$p(b_i; \theta_b) \text{ (equation 104)}$$

To estimate the integral that has no analytical solution in the log-likelihood function mentioned above, some methods can be applied based on the situation. Numerical integration techniques, including Gaussian quadrature and Monte Carlo methods, as well as joint model Laplace approximations for higher dimension b_i vectors (more than 3 entries, for example) can be successfully applied. The Expectation-Maximization (EM) algorithm can be used to maximize the log-likelihood that was approximated. This EM algorithm is typically used when the entries in b_i are treated as missing data, despite the issues with slow convergence near the maximum. (Please see the Appendix section on the EM algorithm for more detail.)

Residuals

Each subject i has a defined observed part, which contains all of the observed measurements of the ith individual, and missing part, which contains within the longitudinal response vector. The observed part is denoted as in equation 105

$$Y_i^O = \{Y_i(t_{ij}) : t_{ij} < T_i, j = 1, \dots, n_i\} \text{ (equation 105)}$$

It has the missing part is similarly denoted as in equation 106.

$$Y_i^m = \{Y_i(t_{ij}) : t_{ij} \geq T_i, j = 1, \dots, n_i'\} \text{ (equation 106)}$$

Finding the residuals of the fitted joint model can be very problematic due to the missing values and any nonrandom dropout of the subjects that might occur in the data. A dropout mechanism is defined as "the conditional distribution of the time-to-dropout given the complete vector of longitudinal responses (Y_i^O, Y_i^m)" and can be derived using the assumptions of joint modeling discussed above, according to Rizopoules, D.[11]. The dropout mechanism is denoted as in equation 107, with $p(b_i | Y_i^O, Y_i^m; \theta)$ representing the posterior distribution of random effects vector b_i.

$$p(T_i^* | Y_i^O, Y_i^m; \theta) = \int p(T_i^* | b_i; \theta) p(b_i | Y_i^O, Y_i^m; \theta) db_i \text{ (equation 107)}$$

(It would be remiss not to discuss what a posterior distribution is and where it comes from. Please see the appendix section on Bayes theorem and prior and posterior distributions for more detail.) The time-to-dropout depends on Y_i^m, which "implies that the observed data, upon which the residuals are calculated, are not a random sample of the target population, and therefore

should not be expected to exhibit the standard properties of zero mean, constant variance, etc," as talked about in Rizopoules, D.[11] Therefore, the traditional approaches of evaluating the model assumptions with the residual vs. fitted and/or the observed data is typically problematic for a joint model produced. Such an issue can be overcome by "augmenting the observed data with randomly imputed longitudinal responses under the complete data model, corresponding to the longitudinal outcomes that would have been observed had the patients not dropped out," according to Rizopoules, D.[11]. This can be done with multiple imputation, or the repeated sampling, given the observed data is known, from the posterior distribution of the vector Y_i^m which is collected from a simulation described in Rizopoules, D.[11]. The residuals of the missing data (such as in the dropout case(s)) have the perk of having the same properties as the full data model Rizopoules, D.[11].

Acknowledgements

Thanks to Dr. Sujay Datta for helping me with this paper, as well as the professors and staff at The University of Akron department of statistics for helping me through this project and with obtaining my masters degree.

References

1. DataCamp. Interactive Course: Longitudinal Analysis in R. 2019.

   Available at: https://learn.datacamp.com/courses/longitudinal-analysis-in-r.

   Accessed April 10, 2020.

2. Fitzmaurice, G, Laird, N, Ware, J. In: Balding, D, Cressie, N, Fitzmaurice, G, Goldstein, H,

   Johnstone, I, Molenberghs, G, Scott, D, Smith, A, Tsay, R, Weisberg, S.

   Applied Longitudinal Analysis. Second Edition. Hoboken, New Jersey, USA. John Wiley & Sons, Inc.; 2011.

3. Framingham Heart Study Data

   https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation.pdf?link_time=2020-04-26_14:45:55.648172

4. Framingham Heart Study Teaching Dataset. National Heart, Blood, and Lung Institute.

   Requested data at https://biolincc.nhlbi.nih.gov/requests/teaching-dataset-request/ . Accessed March 31st, 2020.

5. Hand, D, Crowder, M. In: Chatfield, C, Zidek, J. Practical Longitudinal Data Analysis. First edition.

   London, U.K.: Chapman & Hall; 1996.

6. Ibrahim, J, Chu, Haitao, Chen, L. Basic Concepts and Methods for Joint Models of Longitudinal

and Survival Data. *Journal of Clinical Oncology*. 2010; 28: 2796-2801.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4503792/

Accessed: April 1, 2020.

7. Jenkins. Survival Analysis. 2005.

https://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf

8. Kleinbaum, D, Klein, M. In: Gail, M, Krickeberg, K, Samet, J, Tsiatis, A, Wong, W.

Survival Analysis A Self-Learning Text. Second edition. New York, NY, USA: New York, NY; 2005.

9. Moore, D. In: Gentleman, R, Hornik, K, Parmigiani, G. Applied Survival Analysis Using R.

Switzerland: Springer International Publishing Switzerland; 2016.

https://www.springer.com/gp/book/9783319312439#otherversion=9783319312453

Accessed November 9, 2019.

10. National Heart, Blood, and Lung Institute. Framingham Heart Study Longitudinal Data

Documentation.
https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation.pdf?link_time=2020-04-26_14:45:55.648172 Accessed March 31st, 2020.

11. Rizopoules, D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data.

Journal of Statistical Software. 2010; 35(9): 1-33.

https://www.jstatsoft.org/article/view/v035i09/v35i09.pdf

Accessed: April 1, 2020.

12. Stevenson, M. An Introduction to Survival Analysis. EpiCentre at Massey University, New Zealand.

Stevenson, M.; 2007.

http://www.biecek.pl/statystykaMedyczna/Stevenson_survival_analysis_195.721.pdf

Accessed September 25, 2019.

13. Survival data. http://www.stat.rice.edu/~sneeley/STAT553/Datasets/survivaldata.txt.

Updated March 17, 2005. Accessed October 14, 2020.

Kalbfleisch JD. Heart transplant data. 1980.

Kalbfleisch JD. VA lung cancer data. 1980.

Used R version 3.6.3 (2020-02-29) with RStudio .

Appendix

The translation of the notation of the joint model from the different authors of the articles used for the joint modeling research.

Author	Linear Mixed effect model	Trajectory function	Treatment effect longitudinal marker	Treatment effect on time to event	Overall treatment effect	Notes
Schveder	$Y_i = \beta X_i + b_i Z_i + \varepsilon_i$	$m_i(t) = \beta X_i + b_i Z_i$	β	v	$\omega\beta + v$	Follows notation from Fitzmaurice. $\varepsilon_i \sim MVN$
Ibrahim et. al.	$Y_{ij} = X_{ij} + \varepsilon_{ij}$	$X_{ij} = \theta_{0i} + \theta_{1i} * t_{ij} + \gamma Z_i$	γ	α	$\beta\gamma + \alpha$	$\varepsilon_{ij} \sim N(0, \sigma^2)$ $h(t) = h_0(t)e^{\beta X_{ij} + \alpha Z_i}$,
Rizopoules	$y_i(t) = X_i^T(t)\beta + Z_i^T(t)b_i + \varepsilon_i(t)$	$m_i(t) = X_i^T(t)\beta + Z_i^T(t)b_i$	β	γ^T	$\alpha\beta + \gamma^T$	$h(t)$ ω_i is a vector of covariates $M_i(t) = \{m_i(u), 0 \le u < t\}$ is a history of true unobserved longitudinal process up to time t.

Compare:
$e^{\beta X_{ij} + \alpha Z_i}$ to $e^{\gamma^T \omega_i + \alpha m_i(t)}$
(from Ibrahim et. al.) to (Rizopoules)

$\beta_{(I)} = \alpha_{(R)}$: measures the association between the longitudinal marker and time to event or, in other words, quanitifes the effect of the underlying longitudinal outcome to time to event
$\alpha_{(I)} = \gamma^T{}_{(R)}$: direct treatment effect on time to event, or in other words, vector of regression coefficients
$Z_{(I)} = \omega_{i(R)}$: treatment indictor of history of disease
$X_{ij(I)} = m_i(t)_{(R)}$: trajectory function

Trajectory function comparisons of Ibrahin et. al. and Rizopoules

$m_i(t)_{(R)} = X_{ij(I)}$

Rizopoules		Ibrahin et. al
$b_{i_{(R)}}$	=	$\theta_{1i_{(I)}}$
$Z_i^T(t)_{(R)}$	=	$t_{ij_{(I)}}$
$X_i^T(t)_{(R)}$	=	$Z_{i_{(I)}}$
$\beta_{(R)}$	=	$\gamma_{(I)}$

Finally, to translate all of this notation so that quotes and facts can be used from both papers Ibrahim et. al. and Rizopoules.

Schveder MLEM notation has the same notation as Rizopoules.

Bayes' theorem and the prior and posterior distributions

According to Ellinor et. al.[14], when given evidence E (such as given a set of parameters, i.e. the observed and missing measurements of the ith individual, as well as the event time outcome vector), the probability of the evidence E happening, the probabilities of a given hypothesis H (such as the dropout or event occurring at some particular true time T*), the

probability of hypothesis H happening (ak.a. the prior probability), and the probability pf the evidence happening given the hypothesis can be updated using a conditional probability known as the posterior probability:

$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$

The $\frac{P(E|H)}{P(E)}$ is the likelihood ratio.

14. Ellinor, A, Williams, C, Strandberg, A.
Bayes' Theorem and Conditional Probability.
Brilliant.org. Accessed May 11, 2020.
https://brilliant.org/wiki/bayes-theorem/

The Expectation-Maximization Algorithm

Expectation-Maximization (EM) algorithm is very good with dealing with missing data and is a powerful technique needed to infer the parameters. When there are latent values or features in datasets, which is often the case in datasets with missing values, the EM algorithm can be used. Models with latent or unobserved features include the Hidden Markov model. According to Misra[15], "the question is: does having latent features makes any difference in the estimation of parameters? It turns out, yes. Estimating model parameters does get a little tricky if latent features (or missing data) are involved." The EM algorithm is an alternative to the maximum likelihood estimation approach in that the MLE method does not accommodate for missing or latent values and is an efficient approach to estimating the a model when there are missing or latent values.

If the maximum likelihood approach is considered for parameter estimation with missing or latent values, where V = set of observed variables, Z = set of latent variables, and θ = set of model parameters, then the following objective function is "intractable" because the "parameters are coupled because of summation inside the log. This makes the optimization using Gradient Ascent (or any iterative optimization technique in general)," according to Misra[15]. Thus, a more efficient and practical method of estimation should be used, such as the EM algorithm.

$$L(w) = logP(\text{data}) = log\Pi_{i=1}^{N} P(Y_i|V_i)$$

$$L(w) = \sum_{i=1}^{N} logP(Y_i|V_i) = \sum_{i=1}^{N} log \sum_{h \in Z_i} P(Y_i|V_i, h)$$

Expectation Maximum (EM) algorithm is described as:

- When the Z is known, the optimization of complete log-likelihood P(V, Z|θ)* is easy
  - To know Z must use the posterior probability P(Z|V, θ)

- So, therefore, "consider the expected value of complete data log likelihood under the posterior distribution of latent variables. This step of finding the expectation is called the E-step. In the subsequent M-step, we maximize this expectation to optimize θ."

Expectation Maximum (EM) algorithm steps are as follows:

1. Choose an initial setting for the parameters and denote these initial values set as θ^{old}
2. E step: Evaluate the posterior P(Z|V, θ^{old})
3. Mstep: Evaluate the new parameter θ^{new} denoted as

$$\theta^{\text{new}} = \text{argmax}_\theta \sum_z P(Z|V, \theta^{\text{old}}) log P(V, Z|\theta)$$

4. Check for convergence of the log-likelihood or parameter values
   - If algorithm didn't convergence, set θ^{old} = θ^{new} and return to the E step.

Also, note that Gibbs Sampling is a special type of EM algorithm and that the EM algorithm uses the KL divergence.

15. Misra, R. Inference using EM algorithm. Towards Data Science. 2019. https://towardsdatascience.com/inference-using-em-algorithm-d71cccb647bc Accessed April 28, 2020.

R code

```
################################################################

################################################################

################################################################

#############Rice University Survival Data Set Analysis##########

################################################################

################################################################

################################################################

######data preparation######

#install.packages("readxl")

library(readxl)


#C:/Users/kasch/Dropbox/Statistics Career Stuff/FALL_2019/Masters Paper/Survival
Analysis

setwd("C:/Users/kasch/Dropbox/Statistics Career Stuff/FALL_2019/Masters Paper/Survival
Analysis")

excel_sheets("Rice University Survival Data.xlsx")

# [1] "VA lung cancer data "      "Primary Biliary Cirrhosis" "Lupus nephritis data"

# [4] "Bladder cancer data "      "Brain cancer data"         "Cervical cancer data "

# [7] "Ovarian cancer data"       "Heart transplant data "
```

```
#Use some of these as toy-data examples and one or two as main examples

va_lung <- read_excel("Rice University Survival Data.xlsx", sheet = "VA lung cancer
data ")

bladder_cancer <- read_excel("Rice University Survival Data.xlsx", sheet = "Bladder
cancer data ")

ovarian_cancer <- read_excel("Rice University Survival Data.xlsx", sheet = "Ovarian
cancer data")

pb_cirrhosis <- read_excel("Rice University Survival Data.xlsx", sheet = "Primary
Biliary Cirrhosis")

brain_cancer <- read_excel("Rice University Survival Data.xlsx", sheet = "Brain cancer
data")

heart_transplant <- read_excel("Rice University Survival Data.xlsx", sheet = "Heart
transplant data ")

Lupus_nephritis <- read_excel("Rice University Survival Data.xlsx", sheet = "Lupus
nephritis data")

cervical_cancer <- read_excel("Rice University Survival Data.xlsx", sheet = "Cervical
cancer data ")



#install.packages("survival")

library(survival)



#####Analysis of the brain cancer data#####

brain_cancer <- read_excel("Rice University Survival Data.xlsx", sheet = "Brain cancer
data")

library(readxl)



#a.    Model selection and interpretation

#i.    Covariance adjustment

coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age)

coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age +
strata(brain_cancer$group))

coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ brain_cancer$age +
brain_cancer$group)



#ii.   Categorical and continuous covariates

# brain_cancer$group[brain_cancer$group == 1] = "group 1"

# brain_cancer$group[brain_cancer$group == 0] = "group 0"
```

```
group <- factor(brain_cancer$group)
age <- brain_cancer$age

model.matrix(~ group + age)[,-1]

group <- relevel(group, ref="1")
model.matrix(~ group + age)[,-1]
model.matrix(~ group + age + group:age)[,-1]



result.cox <- coxph(Surv(brain_cancer$"survival time", brain_cancer$status) ~ group +
age)
summary(result.cox)

# iii. Hypothesis Testing for Nested Models
modelA.coxph <- coxph(Surv(brain_cancer$"survival time", brain_cancer$status)
                      ~ brain_cancer$age)
modelB.coxph <- coxph(Surv(brain_cancer$"survival time", brain_cancer$status)
                      ~ brain_cancer$group)
modelC.coxph <- coxph(Surv(brain_cancer$"survival time", brain_cancer$status)
                      ~ brain_cancer$age + brain_cancer$group)

llA <- logLik(modelA.coxph)
llB <- logLik(modelB.coxph)
llC <- logLik(modelC.coxph)

likelihoodratioteststat <- 2*(llC-llA)
pchisq(likelihoodratioteststat, df = 1, lower.tail = F)

#iv.   The Akaike Information Criterion for Comparing Non-nested models
AIC(modelA.coxph)
AIC(modelB.coxph)
AIC(modelC.coxph)
```

```
# Model assessment and diagnostics
# a.   Assessing Goodness of Fit Using Residuals
# i.   Martingale and Deviance residuals
library(survival)
result.0.coxph <- coxph(Surv(brain_cancer$"survival time",
                             brain_cancer$status) ~ 1)
rr.0 <- residuals(result.0.coxph, type = "martingale")
par(mar=c(1,1, 1, 1)) # par(mfrow=c(3,2))
plot(rr.0 ~ brain_cancer$age,
     xlab = "age", ylab = "Martingale Residuals",
)
title("martingale residuals vs. age")
lines(1:100,
      rep(0, 100),
      type = "l",
      lty = 5)

#ii.   Case Deletion Residuals
result.coxph <- coxph(Surv(brain_cancer$"survival time",
                           brain_cancer$status) ~
                        brain_cancer$age)
coef.all <- result.coxph$coef[1]
coef.all

#to get the jackknife residuals
age <- brain_cancer$age
n.obs <- length(brain_cancer$"survival time")
jkbeta.vec <- rep(NA, n.obs)
#i <- 1 #for testing
for (i in 1:n.obs) {
  tt.i <- brain_cancer$"survival time"[-i]
  delta.i <- brain_cancer$status[-i]
  age.i <- age[-i]
```

```
  result.coxph.i <- coxph(Surv(tt.i, delta.i) ~ age.i)
  coef.i <- result.coxph.i$coef[1]
  jkbeta.vec[i] <- (coef.all - coef.i) #the jackknife residuals
}

index.obs <- 1:n.obs
plot(jkbeta.vec ~ index.obs, type="h",
     xlab="Observation", ylab="Change in coefficient for age",
     cex.axis=1.3, cex.lab=1.3)
abline(h=0)
#identify(jkbeta.vec ~ index.obs)

resid.dfbeta <- residuals(result.coxph, type="dfbeta")
n.obs <- length(ttr)
index.obs <- 1:n.obs
plot(resid.dfbeta[,4] ~ index.obs, type="h",
     xlab="Observation", ylab="Change in coefficient")
abline(h=0)
#identify(resid.dfbeta[,4] ~ index.obs)



#Checking PH assumption
# i.   Log cumulative hazard plots
result.surv0 <- survfit(Surv(brain_cancer$"survival time",
                             brain_cancer$status) ~ brain_cancer$group,
                        subset={brain_cancer$group == "0"})
time0 <- result.surv0$time
surv0 <- result.surv0$surv
cloglog0 <- log(-log(surv0))
logtime0 <- log(time0)

result.surv1 <- survfit(Surv(brain_cancer$"survival time",
                             brain_cancer$status) ~ brain_cancer$group,
```

```
                        subset={brain_cancer$group == "1"})
time1 <- result.surv1$time
surv1 <- result.surv1$surv
cloglog1 <- log(-log(surv1))
logtime1 <- log(time1)

plot(cloglog0 ~ logtime0, type="s", col="blue", lwd=2)
lines(cloglog1 ~ logtime1, col="red", lwd=2, type="s")
legend("bottomright",
       legend=c("control group","treatment gorup"),
       col=c("blue","red"),
       lwd=2)

# ii.  Schoenfeld Residuals
result.coxph <- coxph(Surv(brain_cancer$"survival time",
                           brain_cancer$status) ~
                        brain_cancer$age)

result.coxph$coef
residuals(result.coxph, type="schoenfeld")
resid.unscaled <- residuals(result.coxph, type="schoenfeld")
resid.scaled <- resid.unscaled*result.coxph$var[1,1]*sum(brain_cancer$status)

resid.scaled + result.coxph$coef
resid.sch <- cox.zph(result.coxph)
resid.sch$y
result.sch.resid <- cox.zph(result.coxph, transform="km")
plot(result.sch.resid)
result.sch.resid




#####Jasa data set analysis#####
```

```
#Working with Time Dependent Covariates
#Moore's data analysis
library(survival)
result.heart <- coxph(Surv(futime, fustat) ~ transplant + age +
                        + surgery, data=jasa)
summary(result.heart)

ind30 <- jasa$futime >= 30
transplant30 <- {{jasa$transplant == 1} & {jasa$wait.time < 30}}
summary(coxph(Surv(futime, fustat) ~ transplant30 + age + surgery,
              data=jasa, subset=ind30))

id <- 1:nrow(jasa)
jasaT <- data.frame(id, jasa)
id.simple <- c(2, 5, 10, 12, 28, 95)
id.simple
heart.simple <- jasaT[id.simple, c(1, 10, 9, 6, 11)]
heart.simple
summary(coxph(Surv(futime, fustat)~transplant,
              data = heart.simple))

sdata <- tmerge(heart.simple, heart.simple, id=id,
                death=event(futime, fustat),
                transpl=tdc(wait.time))
heart.simple.counting <- sdata[,-(2:5)]
# drop columns 2 thorugh 5
heart.simple.counting

summary(coxph(Surv(tstart, tstop, death) ~ transpl,
              data=heart.simple.counting))

#full dataset:
tdata <- jasa[, -c(1:4, 11:14)]
```

```
tdata$futime <- pmax(.5, tdata$futime)
indx <- {{tdata$wait.time == tdata$futime} &
    !is.na(tdata$wait.time)}
tdata$wait.time[indx] <- tdata$wait.time[indx] - .5
id <- 1:nrow(tdata)
tdata$id <- id
sdata <- tmerge(tdata, tdata, id=id,
                death = event(futime, fustat),
                trans = tdc(wait.time))
jasa.counting <- sdata[,c(7:11, 2:3)]
head(jasa.counting)
summary(coxph(Surv(tstart, tstop, death) ~
                trans + surgery +
                age, data=jasa.counting))
#####Working with multiple survival outcomes and competing risks#####
library(survival)
#install.packages("asaur")
library(asaur)

ashkenazi[ashkenazi$famID %in% c(1, 11, 93), ]



###### The exponential distribution#######
#to make a generic figure for the paper
t <- seq(from = 1, to = 90, by = 1)
lambda = 0.25
ht <- lambda
Ht <- lambda * t
St <- exp(-lambda * t)
par(mfrow = c(2,3), pty = "s")
plot(t, rep(ht, times = length(t)), ylim = c(0, 1),
     lwd = 2, type = "s", xlab = "Time", ylab = "h(t)",
     main = "Instantaneous Hazard of Exponential")
plot(t, Ht, ylim = c(0, 25), lwd = 2,
```

```
     type = "s", xlab = "Time", ylab = "H(t)",
     main = "Cumulative Hazard of Exponential")
plot(t, St, ylim = c(0, 1), lwd = 2,
     type = "s", xlab = "Time", ylab = "S(t)",
     main = "Survival of Exponential")

###### The Weibull distribution #######
#to make a generic figure for the paper
t <- seq(from = 1, to = 90, by = 1)   #lambda = tau and p = c
lambda = 0.25; p = 0.5
ht <- lambda * p * t^(p - 1)
Ht <- lambda * t^p
St <- exp(-lambda * t^p)
par(mfrow = c(2,3), pty = "s")
plot(t, ht, ylim = c(0, 0.5), lwd = 2,
     type = "s", xlab = "Time", ylab = "h(t)",
     main = paste("Instantaneous Hazard of Weibull, tau =", lambda))
plot(t, Ht, ylim = c(0, 5), lwd = 2,
     type = "s", xlab = "Time", ylab = "H(t)",
     main = paste("Cumulative Hazard of Weibull, tau =", lambda))
plot(t, St, ylim = c(0, 1), lwd = 2,
     type = "s", xlab = "Time", ylab = "S(t)",
     main = paste("Survival of Weibull, tau =", lambda))

###### Plots of hazard using different values of lambda and p (Weibull): ######
t <- seq(from = 0, to = 10, by = 0.1)
tau <- 1; c05 <- 0.5; c10 <- 1.0; c15 <- 1.5; c30 <- 3.0
h05 <- tau * c05 * (tau * t)^(c05 - 1)
h10 <- tau * c10 * (tau * t)^(c10 - 1)
h15 <- tau * c15 * (tau * t)^(c15 - 1)
h30 <- tau * c30 * (tau * t)^(c30 - 1)
plot(t, h05, type = "l", ylim = c(0, 6), xlab = "Time",
     ylab = "h(t)", lty = 1, lwd = 2)
lines(t, h10, lty = 2, lwd = 2)
```

```
lines(t, h15, lty = 3, lwd = 2)

lines(t, h30, lty = 4, lwd = 2)

legend(4, 6, legend = c("tau = 1, c = 0.5", "tau = 1, c = 1.0", "tau = 1, c = 1.5",
"tau = 1, c = 3.0"),

       lty = c(1,2,3,4), lwd = c(2,2,2,2),

       bty = "n", cex = 0.75)

###### The lognormal distribuion ######

library(survival)

# x = rlnorm(500,1,.6)

# grid = seq(0,25,.1)

#

# plot(grid,dlnorm(grid,1,.6),type="l",xlab="x",ylab="f(x)")

# lines(density(x),col="red")

#

# legend("topright",c("True Density","Estimate"),lty=1,col=1:2)

par(mfrow = c(2,3), pty = "s")

x   <- seq(.1, 3, length=30) #quantiles

haz <- dsurvreg(x, 0, 0.5, distribution = "lognormal")/ (1-psurvreg(x, 0, 0.5,
distribution = "lognormal"))

plot(x, haz, ylab="Hazard", xlab = "t",

     main = "Lognormal Instantaneous Hazard with sigma = 0.5")


x   <- seq(.1, 3, length=30) #quantiles

haz <- dsurvreg(x, 0, 1, distribution = "lognormal")/ (1-psurvreg(x, 0, 1,
distribution = "lognormal"))

## Not run:

plot(x, haz, ylab="Hazard", xlab = "t",

     main = "Lognormal Instantaneous Hazard with sigma = 1")


x   <- seq(.1, 3, length=30) #quantiles

haz <- dsurvreg(x, 0, 2, distribution = "lognormal")/ (1-psurvreg(x, 0, 2,
distribution = "lognormal"))

## Not run:

plot(x, haz, ylab="Hazard", xlab = "t",

     main = "Lognormal Instantaneous Hazard with sigma = 2")
```

```
###### The loglogistic distribution ######

par(mfrow = c(2,2), pty = "s")

x   <- seq(.1, 3, length=30) #quantiles

haz <- dsurvreg(x, 0, 0.5, distribution = "loglogistic")/ (1-psurvreg(x, 0, 0.5,
distribution = "loggaussian"))

plot(x, haz, ylab="Hazard", xlab = "t",

     main = "Loglogistic Instantaneous Hazard with l = 0.5")


# x   <- seq(.1, 3, length=30) #quantiles

# haz <- dsurvreg(x, 0, 1, distribution = "loglogistic")/ (1-psurvreg(x, 0, 1,
distribution = "loggaussian"))

# plot(x, haz, ylab="Hazard",

#      main = "Loglogistic Instantaneous Hazard with sigma = 1")


x   <- seq(.1, 3, length=30) #quantiles

haz <- dsurvreg(x, 0, 2, distribution = "loglogistic")/ (1-psurvreg(x, 0, 2,
distribution = "loggaussian"))

plot(x, haz, ylab="Hazard", xlab = "t",

     main = "Loglogistic Instantaneous Hazard with l = 2")







####### Comparison of Kaplan-Meier and Weibull estimates of survival:


######Analysis of Heart transplant dataset#######

hearttrans.km <- survfit(Surv(Time, Status) ~ 1,

                         conf.type = "none",

                         type = "kaplan-meier",

                         data = heart_transplant)

# Kaplan-Meier survival function with confidence intervals:

plot(hearttrans.km,

     xlab = "Time",
```

```
     ylab = "S(t)",
     main = "Heart Transplant Survival, Kaplan-Meier",
     conf.int = TRUE)

# Flemington-Harrington estimator
hearttrans.fh <- survfit(Surv(Time, Status) ~ 1,
                         type = "fleming-harrington",
                         data = heart_transplant)
plot(hearttrans.fh,
     xlab = "Time",
     ylab = "S(t)",
     main = "Heart Transplant Survival, fleming-harrington",
     conf.int = TRUE)

# With this data set the difference between the Kaplan-Meier
# and the Flemington-Harrington estimate of survival is not obvious.
# A closer comparison of the two functions:
tmp <- as.data.frame(cbind(km = hearttrans.km$surv, fh = hearttrans.fh$surv))
head(tmp)
tail(tmp)




### Weibull accuracy estimates the Kaplan-Meier.
library(survival)
# Fit parametric (Weibull) and non-parametric (Kaplan-Meier)
# survival functions to the observed data:
hearttrans.we <- survreg(Surv(Time, Status) ~ 1, dist = "weib",
                         data = heart_transplant)
hearttrans.km <- survfit(Surv(Time, Status) ~ 1,
                         conf.type = "none", type = "kaplan-meier",
                         data = heart_transplant)
```

```
# Using the Weibull distribution (the intercept) = -log() and  (scale) = 1 / p.
# Thus the scale parameter  = exp(-) and p
# = 1 / . See Venables and Ripley p 360 and
# Tableman and Kim p 78 for further details.

p <- 1 / hearttrans.we$scale
lambda <- exp(-hearttrans.we$coeff[1])
t <- 1:1000
St <- exp(-(lambda * t)^p)
hearttrans.we <- as.data.frame(cbind(t = t, St = St))
# Compare the two estimates of survival:
plot(hearttrans.km, xlab = "Time",
     ylab = "Cumulative proportion to experience event")
lines(hearttrans.we$t, hearttrans.we$St, lty = 2)
legend(x = "topright", legend = c("Kaplan-Meier", "Weibull"),
       lty = c(1,2), bty = "n")
# The Weibull distribution provides an adequate fit to the observed data up to day
1000,
# then appears to underestimate survivorship.



# Cumulative hazard plots can provide an alternative method for assessing the
# appropriateness of a parametric approach to describe survivorship.
# Here we plot cumulative hazard as a function of time (to check for
# consistency with the exponential distribution)
# and log cumulative hazard as a function of log time
# (to check for consistency with the Weibull distribution).
# Both plots show departure from linearity.
hearttrans.km <- survfit(Surv(Time, Status) ~ 1,
                         conf.type = "none", type = "kaplan-meier", data =
heart_transplant)
Ht <- -log(hearttrans.km$surv)
t <- hearttrans.km$time
par(pty = "s", mfrow = c(1,2))
plot(t, Ht, type = "s", xlim = c(0, 1000),
```

```
     xlab = "Time", ylab = "Cumulative hazard", lwd = 2)
plot(log(t), log(Ht), type = "s", xlim = c(0, 7),
     xlab = "Time (log)", ylab = "Cumulative hazard (log)", lwd = 2)


######Analysis of VA lung cancer dataset######
# VA lung cancer data
#
# Patients with advanced, inoperable lung cancer were treated with
# chemotherapy.
# N = 137
# Veteran's Administration Lung Cancer Trial
# Taken from Kalbfleisch and Prentice, pages 223-224
#
# Variables
# Treatment 1=standard, 2=test
# Cell type 1=squamous, 2=small cell, 3=adeno, 4=large
# Survival in days
# Status 1=dead, 0=censored
# Karnofsky score (measure of general performance, 100=best)
# Months from Diagnosis
# Age in years
# Prior therapy 0=no, 10=yes


# setwd("C:/Users/kasch/Dropbox/Statistics Career Stuff/FALL_2019/Masters
Paper/Survival Analysis")
library(readxl)
va_lung <- read_excel("Rice University Survival Data.xlsx", sheet = "VA lung cancer
data ")


va_lung.km <- survfit(Surv(Survival, Status) ~ 1,
                      conf.type = "none",
                      type = "kaplan-meier",
                      data = va_lung)
# Kaplan-Meier survival function with confidence intervals:
plot(va_lung.km,
```

```
     xlab = "Time",
     ylab = "S(t)",
     main = "VA Ling Cancer Survival, Kaplan-Meier",
     conf.int = TRUE)

# Flemington-Harrington estimator
va_lung.fh <- survfit(Surv(Survival, Status) ~ 1,
                      type = "fleming-harrington",
                      data = va_lung)
plot(va_lung.fh,
     xlab = "Time",
     ylab = "S(t)",
     main = "VA Ling Cancer Survival, fleming-harrington",
     conf.int = TRUE)

# With this data set the difference between the Kaplan-Meier
# and the Flemington-Harrington estimate of survival is not obvious.
# A closer comparison of the two functions:
tmp <- as.data.frame(cbind(km = va_lung.km$surv, fh = va_lung.fh$surv))
head(tmp)
tail(tmp)



# Instantaneous hazard
va_lung.km <- survfit(Surv(Survival, Status) ~ 1,
                      conf.type = "none",
                      type = "kaplan-meier",
                      data = va_lung)

# Work out the proportion that fail at each evaluated time period:
prop.fail <- va_lung.km$n.event/va_lung.km$n.risk
#number of events per number of risks
```

```
# Work out the length of time over which these failure occur:
time <- va_lung.km$time
time
time0 <- c(0, time[-length(time)])
#add a zero to the vector trick, remove the last obs and bind with 0.
time0

# Divide prop.fail by the time interval over which those failures occur
# (that is, time - time0) to get the probability of failing
# per unit time, i.e. the instantaneous hazard:
haz <- prop.fail/(time - time0)

# # Plot the result:
# plot(time, haz, ylim = c(0,0.06), type = "s", xlab = "Days to relapse", ylab =
"h(t)")
# lines(lowess(time[-1], haz[-1], f = 0.10))

# Tidier plot:
plot(time, haz, type = "n", xlab = "Days to relapse",
     ylab = "h(t)", ylim = c(0,0.06))
lines(lowess(time[-1], haz[-1], f = 0.10))

### Weibull accuracy estimates the Kaplan-Meier.
library(survival)
# Fit parametric (Weibull) and non-parametric (Kaplan-Meier)
# survival functions to the observed data:
va_lung.we <- survreg(Surv(Survival, Status) ~ 1, dist = "weib",
                      data = va_lung)
va_lung.km <- survfit(Surv(Survival, Status) ~ 1,
                      conf.type = "none", type = "kaplan-meier",
                      data = va_lung)
# Using the Weibull distribution (the intercept) = -log()
# and  (scale) = 1 / p.
# Thus the scale parameter  = exp(-) and p
# = 1 / . See Venables and Ripley p 360 and
```

```
# Tableman and Kim p 78 for further details.

p <- 1 / va_lung.we$scale
lambda <- exp(-va_lung.we$coeff[1])
t <- 1:1000
St <- exp(-(lambda * t)^p)
va_lung.we <- as.data.frame(cbind(t = t, St = St))
# Compare the two estimates of survival:
plot(va_lung.km, xlab = "Time",
     ylab = "Cumulative proportion to experience event",
     main = "Compare the two estimates of survival")
lines(va_lung.we$t, va_lung.km$St, lty = 2)
legend(x = "topright", legend = c("Kaplan-Meier", "Weibull"),
       lty = c(1,2), bty = "n")
# The Weibull distribution provides an adequate fit
# to the observed data up to day 1000,
# then appears to underestimate survivorship.



# Cumulative hazard plots can provide an
# alternative method for assessing the
# appropriateness of a parametric approach
# to describe survivorship.
# Here we plot cumulative hazard as a
# function of time (to check for
# consistency with the exponential distribution)
# and log cumulative hazard as a function of log time
# (to check for consistency with the Weibull distribution).
# Both plots show departure from linearity.
va_lung.km <- survfit(Surv(Survival, Status) ~ 1,
                      conf.type = "none", type = "kaplan-meier",
                      data = va_lung)
Ht <- -log(va_lung.km$surv)
t <- va_lung.km$time
```

```
par(pty = "s", mfrow = c(1,2))
plot(t, Ht, type = "s", xlim = c(0, 1000),
     xlab = "Time", ylab = "Cumulative hazard", lwd = 2)
plot(log(t), log(Ht), type = "s", xlim = c(0, 7),
     xlab = "Time (log)", ylab = "Cumulative hazard (log)",
     lwd = 2)


#####STRATIFICATION

va_lung$Treatment <- as.character(va_lung$Treatment)
# Kaplan-Meier survival function of days to lung cancer,
# stratifying by treatment:
va_lung.km <- survfit(Surv(Survival, Status) ~ Treatment,
                      type = "kaplan-meier", data = va_lung)
# plot(va_lung.km, xlab = "Time",
#      ylab = "Cumulative proportion to experience event",
#      lty = c(1,2),
#      legend.text = c("Treatment 1", "Treatment 2"),
#      legend.pos = 1,
#      legend.bty = "n"
#      )

plot(va_lung.km,
     xlab = "Time",
     ylab = "Cumulative proportion to experience event",
     lty = c(1,2))

######Residuals
# In the survdiff function the argument rho = 0
# returns the log-rank or Mantel-Haenszel test, rho = 1
# returns the Peto and Peto modification of the
# Gehan-Wilcoxon test. Mantel-Haenszel test:
survdiff(Surv(Survival, Status) ~ Treatment,
```

```
         data = va_lung, na.action = na.omit, rho = 0)
#https://www.graphpad.com/support/faq/how-do-the-three-methods-compare-to-survival-curves-log-rank-mantel-haenszel-gehan-wilcoxon-differ/

# Peto and Peto modification of the Gehan-Wilcoxon test:
survdiff(Surv(Survival, Status) ~ Treatment,
         data = va_lung, na.action = na.omit, rho = 1)

# SELECTION OF COVARIATES
# Set contrasts for cell type and prison. Set the
# reference category for cell type, making cell type 1 (base = 1)
# the reference category:
va_lung$"Cell type" <- factor(va_lung$"Cell type",
                              levels = c(1, 2, 3, 4), labels = c("1", "2", "3", "4"))
contrasts(va_lung$"Cell type") <- contr.treatment(4, base = 1, contrasts = TRUE)
levels(va_lung$"Cell type")

# Same for prior therapy, making absence of a prior therapy the reference category:
va_lung$"Prior therapy" <- as.factor(va_lung$"Prior therapy")
# va_lung$"Prior therapy" <- factor(va_lung$"Prior therapy",
#                                   levels = c(0, 1), labels = c("0", "1"))
contrasts(va_lung$"Prior therapy") <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(va_lung$"Prior therapy")

# Assess the influence of Cell type, Prior therapy, Treatment, and
# Karnofsky score on days to relapse.

# First of all categorise Karnofsky score into four classes based on quartiles:
quantile(va_lung$"Karnofsky score", probs = c(0.25, 0.50, 0.75))
hist(va_lung$"Karnofsky score",
     xlab = "Karnofsky score", main = "Histogram of Karnofsky score")

# Quartiles for meth_dose are 50, 60 and 70.
# Create a categorical variable based on Karnofsky:
Karnofsky.cat <- rep(0, length(va_lung[,5]))
```

```
Karnofsky.cat[va_lung$"Karnofsky score" < 40] <- 1
Karnofsky.cat[va_lung$"Karnofsky score" >= 40 & va_lung$"Karnofsky score" < 60] <- 2
Karnofsky.cat[va_lung$"Karnofsky score" >= 60 & va_lung$"Karnofsky score" < 75] <- 3
Karnofsky.cat[va_lung$"Karnofsky score" >= 75] <- 4
va_lung <- cbind(va_lung, Karnofsky.cat)

# Assess the effect of Cell type, Prior therapy, and Karnofsky score on days to
relapse:
va_lung.km01 <- survfit(Surv(Survival, Status) ~ va_lung$Karnofsky.cat,
                        type = "kaplan-meier", data = va_lung)
va_lung.km02 <- survfit(Surv(Survival, Status) ~ va_lung$"Prior therapy",
                        type = "kaplan-meier", data = va_lung)
va_lung.km03 <- survfit(Surv(Survival, Status) ~ va_lung$"Cell type",
                        type = "kaplan-meier", data = va_lung)

# Plot all Kaplan-Meier curves on one page. The mark.time = FALSE argument disables
the censor marks:
par(pty = "s", mfrow = c(2,2))
plot(va_lung.km03, xlab = "Time to death", ylab = "Cumulative proportion to
       experience event", main = "Cell type", lty = c(1,4), mark.time = FALSE,
     legend.text = c("Cell type 1","Cell type 2", "Cell type 3", "Cell type 4"),
     legend.pos = 0, legend.bty = "n", cex = 0.80)
plot(va_lung.km02, xlab = "Time to death", ylab = "Cumulative proportion to
       experience event", main = "Prior Therapy", lty = c(1,2), mark.time = FALSE,
     legend.text = c("Yes","No"), legend.pos = 0, legend.bty = "n", cex = 0.80)
plot(va_lung.km01, xlab = "Time to death", ylab = "Cumulative proportion to
       experience event", main = "Karnofsky Score", lty = c(1,2,3,4), mark.time =
       FALSE, legend.text = c("Karnofsky 1","Karnofsky 2", "Karnofsky 3", "Karnofsky
4"), legend.pos = 0,
     legend.bty = "n", cex = 0.80)

# Log-rank tests:
survdiff(Surv(Survival, Status) ~ va_lung$Karnofsky.cat,
         data = va_lung, na.action = na.omit, rho = 0)
survdiff(Surv(Survival, Status) ~ va_lung$"Prior therapy",
         data = va_lung, na.action = na.omit, rho = 0)
```

```
survdiff(Surv(Survival, Status) ~ va_lung$"Cell type",
         data = va_lung, na.action = na.omit, rho = 0)

# Fit the cox proportional hazards multivariable model
# survival tim4 depends on prior therapy, karnofsky score, and cell type:
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Prior therapy" + va_lung$"Cell
type",
                       method = "breslow", data = va_lung)
summary(va_lung.cph01)
# Variables karnofsky score and cell type significantly
# influence time to death
# Variable Prior therapy
# is not signficant, or in the model significance (P = 0.06).
# Drop variable Prior therapy (using the update function):

va_lung.cph02 <- update(va_lung.cph01, ~. - va_lung$"Prior therapy")
summary(va_lung.cph02)
# Does va_lung.cph02 provide a better fit to the data than va_lung.cph01?
x2 <- 2 * (va_lung.cph02$loglik[2] - va_lung.cph01$loglik[2])
1 - pchisq(x2,1)
# Removing Prior therapy has no effect on model fit (P = 1.0).

##### Check scale of continuous covariates (method 1):
# Replace the continuous covariate Karnofsky score with design (dummy) variables.
# Plot the estimated coefficients versus the midpoint of each group:
va_lung$Karnofsky.cat <- factor(va_lung$Karnofsky.cat, labels=c("1", "2", "3", "4"))
contrasts(va_lung$Karnofsky.cat) <- contr.treatment(4, base = 1, contrasts = TRUE)

va_lung.cph03 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Prior therapy" + va_lung$"Cell
type",
                       method = "breslow", data = va_lung)
summary(va_lung.cph03)
va_lung.cph03$coefficients
```

```
va_lung.cph03$coefficients[3:5]

x <- c(((50 + min(as.numeric(va_lung$Karnofsky.cat)))/2), 55,
65,((max(as.numeric(va_lung$Karnofsky.cat)) + 70)/2))

y <- c(0, va_lung.cph03$coefficients[3:5])

plot(x, y, xlim = c(0, 100), type = "l", xlab = "Karnofsky Score", ylab = "Regression
coefficient")


# Scale of continuous covariates linear | no transformations

# required for variable Karnofsky score

# Check scale of continuous covariates (method 2)

va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~

                         Karnofsky.cat + va_lung$"Prior therapy" + va_lung$"Cell
type",

                       method = "breslow", data = va_lung)

va_lung.mi <- residuals(va_lung.cph01, type = "martingale") #Martingale residuals

va_lung.hi <- va_lung$Status - va_lung.mi

va_lung.clsm <- lowess(as.numeric(va_lung$Karnofsky.cat), va_lung$Status)

va_lung.hlsm <- lowess(as.numeric(va_lung$Karnofsky.cat), va_lung.hi)

va_lung.yi <- log(va_lung.clsm$y / va_lung.hlsm$y) +

  (va_lung.cph01$coefficients[3] * as.numeric(va_lung$Karnofsky.cat))

# Plot covariate values versus Martingale residuals

# and va_lung.yi versus covariate values:

par(pty = "s", mfrow = c(1,2))

plot(as.numeric(va_lung$Karnofsky.cat), va_lung.mi,

     xlab = "Karnofsky Score", ylab = "Martingale Residual",

     main = "Martingale Residuals of Karnofsky Score")

lines(lowess(as.numeric(va_lung$Karnofsky.cat), va_lung.mi))

plot(va_lung.yi, as.numeric(va_lung$Karnofsky.cat),

     xlab = "calculated parameters", ylab = "Karnofsky Score")

# A linear relationship is evident between the covariate

# values and each of the calculated parameters,

# indicating that described the continuous variable

# Karnofsky score is linear in its log hazard.


#Interactions:

#Check for significance of the interaction between the categorical variables:
```

```
va_lung.cph04 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + #va_lung$"Prior therapy" +
                         va_lung$"Cell type" +
                         (Karnofsky.cat * va_lung$"Cell type"),
                       method = "breslow", data = va_lung)
summary(va_lung.cph04)
# The P-value of the Wald test for the interaction term
# Prior therapy * Cell type is significant (p=4e-10).

##### Testing the proportional hazards assumption
# Plot scaled Schoenfeld residual plots:
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Cell type", #+ va_lung$"Prior
therapy"
                       method = "breslow", data = va_lung)
va_lung.zph <- cox.zph(va_lung.cph01)
par(pty = "s", mfrow = c(2,2))
plot(va_lung.zph[1], main = "Karnofsky Score",
     ylab = "Karnofsky score"); abline(h = 0, lty = 2)
plot(va_lung.zph[2], main = "Cell type",
     ylab = "Cell type"); abline(h = 0, lty = 2)
# plot(va_lung.zph[2], main = "Prior therapy",
#      ylab = "Prior therapy"); abline(h = 0, lty = 2)

# The variability band for clinic displays a negative slope over time,
# suggesting non-proportionality of hazards.

# Formally test of the proportional hazards assumption
# for all variables in va_lung.cph01:
cox.zph(va_lung.cph01, global = TRUE)
# Using the cox.zph function, rho is the Pearson product-moment
# correlation between the scaled Schoenfeld residuals and time.
# The hypothesis of no correlation is tested using test statistic chisq.
# In the above example, the significant cox.zph test for
# Karnofsky.cat (P < 0.01) implies that the proportional
```

```
# hazards assumption as been violated for the Karnofsky.cat variable.
# This notion is supported by the Schoenfeld residual plots.

# An alternative (and less sensitive) means of
# testing the proportional hazards
# assumption is to plot log[-log S(t)] vs time:
library(survival)
celltype.km <- survfit(Surv(Survival, Status) ~
                         va_lung$"Cell type",
                       type = "kaplan-meier",
                       data = va_lung)
celltype.km$strata[1]
celltype <- c(rep(1, times = celltype.km$strata[1]),
              rep(2, times = celltype.km$strata[2]),
              rep(3, times = celltype.km$strata[3]),
              rep(4, times = celltype.km$strata[4])
)
celltype.km.haz <- as.data.frame(cbind(celltype,
                                       time = celltype.km$time,
                                       surv = celltype.km$surv))
celltype1 <- log(-log(celltype.km.haz$surv[celltype.km.haz$celltype == 1])) #plot
log[-log S(t)] vs time
celltype2 <- log(-log(celltype.km.haz$surv[celltype.km.haz$celltype == 2])) #plot
log[-log S(t)] vs time
celltype3 <- log(-log(celltype.km.haz$surv[celltype.km.haz$celltype == 3])) #plot
log[-log S(t)] vs time
celltype4 <- log(-log(celltype.km.haz$surv[celltype.km.haz$celltype == 4])) #plot
log[-log S(t)] vs time
plot(c(celltype.km.haz$time[celltype.km.haz$celltype == 1],
       celltype.km.haz$time[celltype.km.haz$celltype == 2],
       celltype.km.haz$time[celltype.km.haz$celltype == 3],
       celltype.km.haz$time[celltype.km.haz$celltype == 4]),
     c(celltype1, celltype2, celltype3, celltype4),
     type = "n",
     ylim = c(-5, 2), xlab = "Survival Time",
     ylab = "Log cumulative hazard",
```

```
     main = "cell type")
lines(celltype.km.haz$time[celltype.km.haz$celltype == 1],
      celltype1, type = "s", lty = 1)
lines(celltype.km.haz$time[celltype.km.haz$celltype == 2],
      celltype2, type = "s", lty = 2)
lines(celltype.km.haz$time[celltype.km.haz$celltype == 3],
      celltype3, type = "s", lty = 3)
lines(celltype.km.haz$time[celltype.km.haz$celltype == 4],
      celltype4, type = "s", lty = 4)
legend(x = "topleft",
       legend = c("cell type 1",
                  "cell type 2",
                  "cell type 3",
                  "cell type 4"),
       lty = c(4, 1), bty = "n")
# We could be talked into concluding that the
# -log[-log S(t)] vs time plots for Cell type are not parallel,
# not conflicting with the endings of the cox.zph test
# and the Schoenfeld residual plots.

Karnofsky.cat.km <- survfit(Surv(Survival, Status) ~
                              Karnofsky.cat,
                            type = "kaplan-meier",
                            data = va_lung)
Karnofsky.cat.km$strata[1]
Karnofsky.cat <- c(rep(1, times = Karnofsky.cat.km$strata[1]),
                   rep(2, times = Karnofsky.cat.km$strata[2]),
                   rep(3, times = Karnofsky.cat.km$strata[2]),
                   rep(4, times = Karnofsky.cat.km$strata[2]))
Karnofsky.cat.haz <- as.data.frame(cbind(Karnofsky.cat,
                                         time = Karnofsky.cat.km$time,
                                         surv = Karnofsky.cat.km$surv))
Karnofsky.cat1 <- log(-log(Karnofsky.cat.haz$surv[Karnofsky.cat.haz$Karnofsky.cat ==
1])) #plot log[-log S(t)] vs time
```

```
Karnofsky.cat2 <- log(-log(Karnofsky.cat.haz$surv[Karnofsky.cat.haz$Karnofsky.cat ==
2])) #plot log[-log S(t)] vs time

Karnofsky.cat3 <- log(-log(Karnofsky.cat.haz$surv[Karnofsky.cat.haz$Karnofsky.cat ==
3])) #plot log[-log S(t)] vs time

Karnofsky.cat4 <- log(-log(Karnofsky.cat.haz$surv[Karnofsky.cat.haz$Karnofsky.cat ==
4])) #plot log[-log S(t)] vs time


plot(c(Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 1],

       Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 2],

       Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 3],

       Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 4]),

     c(Karnofsky.cat1, Karnofsky.cat2, Karnofsky.cat3, Karnofsky.cat4),

     type = "n",

     ylim = c(-5, 2), xlab = "Survival Time",

     ylab = "Log cumulative hazard",

     main = "Karnofsky.cat")

lines(Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 1],

      Karnofsky.cat1, type = "s", lty = 1)

lines(Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 2],

      Karnofsky.cat2, type = "s", lty = 2)

lines(Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 3],

      Karnofsky.cat3, type = "s", lty = 3)

lines(Karnofsky.cat.haz$time[Karnofsky.cat.haz$Karnofsky.cat == 4],

      Karnofsky.cat4, type = "s", lty = 4)

legend(x = "topleft",

       legend = c("Karnofsky.cat 1", "Karnofsky.cat 2",

                  "Karnofsky.cat 3", "Karnofsky.cat 4"),

       lty = c(1, 1), bty = "n")

# We could be talked into concluding that the

# -log[-log S(t)] vs time plots for Karnofsky.cat are not parallel,

# not conflicting with the endings of the cox.zph test

# and the Schoenfeld residual plots.


##### Residuals

# Deviance residuals:

va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
```

```
                          Karnofsky.cat + va_lung$"Cell type",
                        method = "breslow", data = va_lung)
va_lung.res <- residuals(va_lung.cph01, type = "deviance")
par(pty = "s", mfrow = c(2, 2))
#boxplot(va_lung.res ~ va_lung$"Prior therapy", main = "Prior Therapy"); abline(h = 0,
lty = 2)
boxplot(va_lung.res ~ va_lung$"Cell type", main = "Cell type"); abline(h = 0, lty = 2)
plot(as.numeric(va_lung$Karnofsky.cat), va_lung.res,
     xlab = "Karnofsky Score", ylab = "Deviance residual",
     main = "Karnofsky Score"); abline(h = 0, lty = 2)

# The following plots show the change in each regression
# coeffcient when each
# observation is removed from the data influence statistics
# (using a common scale for the vertical axis: -0.1 to +0.1):
va_lung.res <- resid(va_lung.cph01, type = "dfbeta")
par(mfrow = c(2, 2))
main <- c("Karnofsky.cat", "Prior therapy", "Cell type")
for (i in 1:3){
  plot(1:238, va_lung.res[,i], type = "h", ylim = c(-0.1,0.1), xlab =
         "Observation", ylab = "Change in coefficient")
  title(main[i])
}
# The above plots give an idea of the influence
# individual observations have
# on the estimated regression coefficients for each
# covariate. Data sets where the influence plot
# is tightly clustered around zero
# indicate an absence of influential observations.

# Now plot the Martingale residuals:
res <- residuals(va_lung.cph01, type = "martingale")
X <- as.matrix(va_lung[,c("Prior therapy", "Cell type", "Karnofsky.cat")])
par(mfrow = c(2,2))
for(j in 1:3){
```

```
  plot(X[,j], res, xlab =
         c("Prior therapy", "Cell type", "Karnofsky.cat")[j],
       ylab = "Martingale residuals")
  abline(h = 0, lty = 2)
  lines(lowess(X[,j], res))
}

# par(mfrow = c(2,2))
# b <- coef(va_lung.cph01[1:3])
# for(j in 1:3){
#          plot(X[,j], b[j] * X[,j] + res, xlab
#               = c("Prior therapy", "Cell type", "Karnofsky.cat")[j],
#               ylab = "Component + residual")
#          abline(lm(b[j] * X[,j] + res ~ X[,j]), lty = 2)
#          lines(lowess(X[,j], b[j] * X[,j] + res, iter = 0))
# }
# Error in b[j] * X[, j] : non-numeric argument to binary operator

# Overall goodness of fit
# Cox model:
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Prior therapy" + va_lung$"Cell
type",
                       method = "breslow", data = va_lung)
summary(va_lung.cph01)

# Log partial likelihood for the [intercept-only] model and for the fitted model:
#   va_lung.cph01$loglik[1]; va_lung.cph01$loglik[2]
# Compute Schemper and Stare (1996) R2 manually:
r.square <- 1 - exp((2/length(va_lung[,1])) *
                      (va_lung.cph01$loglik[1] - va_lung.cph01$loglik[2]))
r.square


##### Dealing with violation of the proportional hazards assumption:
```

```
# From the analyses conducted so far,
# we conclude that the proportional hazards
# assumption has been violated for the variable Karnofsky score
# One method of dealing with this is to stratify
# the model by Karnofsky score
# This means that we produce a separate baseline hazard
# function for each level of Karnofsky score
# Note that by stratifying we cannot obtain a hazard
# ratio for Karnofsky score since the `Karnofsky score effect'
# is absorbed into the baseline hazard.
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Cell type",
                       method = "breslow", data = va_lung)
va_lung.cph04 <- coxph(Surv(Survival, Status, type = "right") ~
                         strata(Karnofsky.cat) + va_lung$"Cell type",
                       method = "breslow", data = va_lung)
summary(va_lung.cph04)

# Compare the original model with the stratfied model:
x2 <- 2 * (va_lung.cph04$loglik[2] - va_lung.cph01$loglik[2])
1 - pchisq(x2, 1)
#[1] 0
# The stratified model provides a significantly better fit.

# Parameterising Karnofsky score as
# a time dependent covariate would be one
# option for dealing with non-proportionality of hazards
# and retaining the ability to quantify
# the effect of Karnofsky score.
# Plot Kaplan-Meier survival curves for each Karnofsky score,
# adjusting for the effect of cell type:
plot(survfit(va_lung.cph04), lty = c(1,3),
     xlab = "Survival Time",
     ylab = "Cumulative proportion to experience event",
```

```
     main = "Survival Fit of Karnofsky Score adjusting for Cell type",
     legend.text = c("Karnofsky 1","Karnofsky 2",
                     "Karnofsky 3", "Karnofsky 4"),
     legend.pos = 0, legend.bty = "n")



#Exponential and Weibull models:
library(survival)
library(readxl)


#va_lung <- read_excel("Rice University Survival Data.xlsx", sheet = "VA lung cancer
data ")
# setwd("C:/Users/kasch/Dropbox/Statistics Career Stuff/FALL_2019/Masters
Paper/Survival Analysis")


#Cox proportional hazards model (for comparison):
va_lung.cph01 <- coxph(Surv(Survival, Status, type = "right") ~
                         Karnofsky.cat + va_lung$"Cell type",
                       method = "breslow", data = va_lung)
summary(va_lung.cph01)


#Exponential model:
va_lung.exp01 <- survreg(Surv(Survival, Status, type = "right") ~
                           Karnofsky.cat + va_lung$"Cell type",
                         dist = "exp", data = va_lung)
summary(va_lung.exp01)
shape.exp = 1/ va_lung.exp01$scale
shape.exp


# Weibull model:
va_lung.wei01 = survreg(Surv(Survival, Status, type = "right") ~
                          Karnofsky.cat + va_lung$"Cell type",
                        dist = "weib", data = va_lung)
summary(va_lung.wei01)
shape.wei = 1 / va_lung.wei01$scale
```

```
shape.wei

# Lognormal model:
va_lung.lognorm01 = survreg(Surv(Survival, Status, type = "right") ~
                              Karnofsky.cat + va_lung$"Cell type",
                            dist = "lognormal", data = va_lung)
summary(va_lung.lognorm01)
shape.lognorm = 1 / va_lung.lognorm01$scale
shape.lognorm

# Loglogistic model
va_lung.loglogistic01 = survreg(Surv(Survival, Status, type = "right") ~
                                  Karnofsky.cat + va_lung$"Cell type",
                                dist = "loglogistic", data = va_lung)
summary(va_lung.loglogistic01)
shape.loglogistic = 1 / va_lung.loglogistic01$scale
shape.loglogistic

#Compare the three models using AIC:
extractAIC(va_lung.cph04)
extractAIC(va_lung.exp01)
extractAIC(va_lung.wei01)
extractAIC(va_lung.lognorm01)
extractAIC(va_lung.loglogistic01)



# The AIC for the Cox model is the smallest, indicating that this model provides the
best
# fit with the data (this is consistent with
# the diagnostics we ran earlier to assess how consistent the
# data was with the exponential and Weibull distributions).

# Additional plotting options are available using the Design package.
# Re-run Weibull model using the psm function:
#install.packages("rms") #previously Design
```

```
library(rms)
####package 'Design' is not available (for R version 3.6.1)

va_lung.wei = psm(Surv(Survival, Status, type = "right") ~
                    Karnofsky.cat + va_lung$"Cell type",
                  dist = "weibull", data = va_lung)

# Plot survivorship for each Cell type for each patient
# and receving a maximum karnofsky score of 40:
# survplot(va_lung.wei, what = c("survival","hazard"), va_lung$'Cell type' =
c("1","2","3","4"), va_lung$'Karnofsky score' = 50)
#
###See Moore page 146 to plot weibull:
va_lung.wei.survreg = survreg(Surv(Survival, Status) ~
                                va_lung$"Cell type",
                              dist = "weibull", data = va_lung)
summary(va_lung.wei.survreg)

va_lung.coxph.survreg = coxph(Surv(Survival, Status) ~
                                va_lung$"Cell type",
                              data = va_lung)
summary(va_lung.coxph.survreg)

mu0.hat <- va_lung.coxph.survreg$coef[1]
sigma.hat <- va_lung.wei.survreg$scale
alpha.hat <- 1/sigma.hat
lambda0.hat <- exp(-mu0.hat)

tt.vec <- 0:16
surv0.vec <- 1 - pweibull(tt.vec,
                          shape=alpha.hat,
                          scale=1/lambda0.hat)

gamma.hat <- va_lung.wei.survreg$coef[2]
```

```
surv1.vec <- surv0.vec^(exp(-gamma.hat/sigma.hat))


coxph.surv.est <- survfit(va_lung.coxph.survreg,
                          newdata=data.frame(list(grp=c("2","3"))))


plot(coxph.surv.est, col=c("red", "black"))
lines(surv0.vec ~ tt.vec, col="red")
lines(surv1.vec ~ tt.vec)


# Here we use the psm function in the rms library to develop an AFT model.
# The psm function is a modification of survreg
# and is used for fitting the accelerated failure time family of parametric survival
models.


library(rms)
va_lung.aft01 <- psm(Surv(Survival, Status, type = "right") ~
                       Karnofsky.cat + va_lung$"Cell type",
                     dist = "weibull", data = va_lung)
va_lung.aft01


####
# What is the effect of  Cell type of 2
# on survival time
# (after adjusting for the effect of presence of Karnofsky.cat)?
####


exp(va_lung.aft01$coefficients[3])


####
# A patient with a Cell type of 2
# cuts the patient survival time in about half (by a scale of 0.44).
# What does this mean in terms of calendar time?
####
```

```
log.t <- as.numeric(va_lung.aft01$coefficients[1] +
                      (va_lung.aft01$coefficients[3]* 1))
exp(log.t)

####
# Patients with a Cell type of 2 remained
# on the cancer treatment program for an additional 19 days
# (compared with those treated in the other
# three levels of the cell type.

#lognormal?
# va_lung.aft02 <- psm(Surv(Survival, Status, type = "right") ~
#                        Karnofsky.cat + va_lung$"Cell type",
#                      dist = "lognormal", data = va_lung)
# va_lung.aft02
#####################################################################
#####################################################################
#####################################################################
#############Framingham Dataset Analysis in LDA###################
#####################################################################
#####################################################################
#####################################################################
setwd("C:/Users/kasch/OneDrive/Desktop")
framingham <- read.csv("frmgham2.csv", stringsAsFactors = F)

#####look at the data#####
View(framingham)

library(tibble)
# glimpse(framingham)
summary(framingham)
library(nlme)
```

```
# head(framingham)

names(framingham)

# install.packages("dplyr")

library(dplyr)

# hospmi <- as.factor(framingham$hospmi)

# timeap <- framingham$timeap

# death <- framingham$death

#count(framingham, death) #not working.
https://stackoverflow.com/questions/45986155/r-error-in-usemethodgroups-no-applicable-method-for-groups-applied-to




#####clean the data#####

#Make categorical variables into 0 ans 1

#in general:

#No: 0

#Yes: 1

#variables are

#death, angina, hospmi, mi_fchd, anychd, stroke,

#cvd, hyperten, cursmoke1, diabetes1, bpmeds1,

#prevchd1, prevap1, prevmi1, prevstrk1, prehyp1, hdlc1, ldlc1,

#cursmoke2, diabetes2, bpmeds2,

# prevchd2, prevap2, prevmi2, prevstrk2, prevhyp2,

#cursmoke3, bmi3, diabetes3, bpmeds3, prevchd3, prevap3, prevmi3

#prevstrk3, prevhyp3


#DONT NEED: #framingham["death"] <- as.factor(framingham["death"])

#code that built for loop below: for reference if needed

# framingham["death"][framingham["death"] == 'No'] <- "0"

# framingham["death"][framingham["death"] == 'Yes'] <- "1"

#convert to numeric if needed:

#framingham["death"] <- as.numeric(framingham["death"][, 1])


#this doesn't work: (CAN DELETE)
```

```
#as.numeric(factor(framingham["death"][framingham["death"]=="No"])) <- 0


#framingham[cat_var[1]][framingham[cat_var[1]] == "No"]
#
# library(plyr)
# cat_var <- c("death", "angina", "hospmi", "mi_fchd", "anychd", "stroke",
#              "cvd", "hyperten", "cursmoke1", "diabetes1", "bpmeds1",
#              "prevchd1", "prevap1", "prevmi1", "prevstrk1",
#              "prevhyp1",
#              "cursmoke2", "diabetes2", "bpmeds2", "prevchd2",
#              "prevap2", "prevmi2", "prevstrk2", "prevhyp2",
#              "cursmoke3", "diabetes3", "bpmeds3",
#              "prevchd3", "prevap3", "prevmi3",
#              "prevstrk3", "prevhyp3")
# #c = 1
# for (c in 1:length(cat_var)){
#   framingham[cat_var[c]][framingham[cat_var[c]] == 'No'] <- "0"
#   framingham[cat_var[c]][framingham[cat_var[c]] == 'Yes'] <- "1"
#   #convert to numeric if needed:
#   # framingham[cat_var[c]] <- as.numeric(framingham[cat_var[c]][, 1])
#
#   #library(plyr)
#   print(count(framingham[cat_var[c]]))
# }



#impute missing values in numeric variables with the median?


#variables: totchol1, cigpday1, bmi1, heartrte1, glucose1, totchol2,
#age2, sysbp2, diabp2, cigpday2, bmi2, heartrte2, glucose2,
#totchol3, age3, sysbp3, diabp3, cigpday3, bmi3, heartrte3, glucose3,
#hdlc3, ldlc3, bmidiff
```

```
#####make counts of the data: #####
library(plyr)
cat_var <- c("SEX", "CURSMOKE", "CIGPDAY", "DIABETES", "BPMEDS",
             "educ", "PREVCHD", "PREVAP", "PREVMI", "PREVSTRK",
             "PREHYP", "PERIOD", "DEATH", "ANGINA", "HOSPMI",
             "MI_FCHD", "ANYCHD", "STROKE", "CVD", "HYPERTEN")
for (c in 1:length(cat_var)){
  print(count(framingham[cat_var[c]]))
}
count(framingham$PERIOD)
count(framingham$MI_FCHD)

#####framingham correlations#####
#data in "long" format

#can convert to wide format
#with functions in dplyr package.
library(dplyr)
library(tidyr)
#install.packages("corrr")
library(corrr)
library(tidyverse)

# which(is.na(framingham$HEARTRTE))
# framingham$HEARTRTE[which(is.na(framingham$HEARTRTE))] <- 0

# Doesn't work!!!!!!
framingham %>%
  mutate(PERIOD = paste0('Repeated_Measure_', PERIOD)) %>%
  #mutate(TIME = paste0('TIME_', TIME)) %>%
  #spread(TIME, HEARTRTE) %>%
  spread(PERIOD, HEARTRTE) %>%
  select(Repeated_Measure_1, Repeated_Measure_2:Repeated_Measure_3) %>%
  # select(#RANDID, DIABETES,
```

```
  #        TIME_1, TIME_2:TIME_3) %>%
  cor(use = "pairwise.complete.obs") %>%
  shave(upper = FALSE) %>%
  fashion(decimals = 3)

#DATACAMP EXAMPLE:
#correlate(framingham, )
#EXAMPLE (Datacamp)
# BodyWeight %>%
#   mutate(Time = paste0('Time_', Time)) %>%
#   spread(Time, weight) %>%
#   select(Rat, Diet, Time_1, Time_8, everything())



cor(framingham$TIME, framingham$TIME)


#####Summary statistics#####
#install.packages("tidyverse")
library(tidyverse)
framingham %>%
  group_by(PERIOD) %>%
  summarize(mean_HEARTRTE = mean(HEARTRTE, na.rm = TRUE),
            med_HEARTRTE = median(HEARTRTE, na.rm = TRUE),
            min_HEARTRTE = min(HEARTRTE, na.rm = TRUE),
            max_HEARTRTE = max(HEARTRTE, na.rm = TRUE),
            sd_HEARTRTE = sd(HEARTRTE, na.rm = TRUE),
            num_miss = sum(is.na(HEARTRTE)),
            n = nrow(framingham))


framingham %>%
  group_by(PERIOD) %>%
  summarize(mean_AGE = mean(AGE, na.rm = TRUE),
            med_AGE = median(AGE, na.rm = TRUE),
```

```
            min_AGE = min(AGE, na.rm = TRUE),
            max_AGE = max(AGE, na.rm = TRUE),
            sd_AGE = sd(AGE, na.rm = TRUE),
            num_miss = sum(is.na(AGE)),
            n = nrow(framingham))



framingham %>%
  group_by(PERIOD) %>%
  summarize(mean_GLUCOSE = mean(GLUCOSE, na.rm = TRUE),
            med_GLUCOSE = median(GLUCOSE, na.rm = TRUE),
            min_GLUCOSE = min(GLUCOSE, na.rm = TRUE),
            max_GLUCOSE = max(GLUCOSE, na.rm = TRUE),
            sd_GLUCOSE = sd(GLUCOSE, na.rm = TRUE),
            num_miss = sum(is.na(GLUCOSE)),
            n = nrow(framingham))

#####violin plot: ######
library(ggplot2)
ggplot(framingham, aes(x = factor(PERIOD), y = CIGPDAY)) +
  geom_violin() +
  xlab("Repeated Measurement") +
  ylab("cig per day") +
  theme_bw(base_size = 16)

ggplot(framingham, aes(x = factor(PERIOD), y = GLUCOSE)) +
  geom_violin() +
  xlab("Repeated Measurement") +
  ylab("glucose") +
  theme_bw(base_size = 16)

ggplot(framingham, aes(x = factor(PERIOD), y = HEARTRTE)) +
  geom_violin() +
```

```
  xlab("Repeated Measurement") +
  ylab("heart rate") +
  theme_bw(base_size = 16)

ggplot(framingham, aes(x = factor(PERIOD), y = TIMEMIFC)) +
  geom_violin() +
  xlab("Repeated Measurement") +
  ylab("time hospitalized for MI or FCH") +
  theme_bw(base_size = 16)

#GLUCOSE LOOKS LIKE IT COULD BE DIFFERENT
#AS WELL AS CIGPDAY

#####LINE PLOT(S):#####
#install.packages("nlme")
library(nlme)
library(ggplot2)
# framingham$randid <- as.factor(framingham$randid)
# framingham$randid <- as.character(framingham$randid)
set.seed(123)

#framinghamrandidsample <- sample(framingham$RANDID, 5)



library(data.table)
framinghamtable <- data.table(framingham)

framinghamsample <- sample_n(x = framinghamtable, size = 10)

library(ggplot2)
ggplot(framingham, aes(x = PERIOD,
                       y = HEARTRTE)) +
  geom_line(aes(group = RANDID), #aes(group = RAND2),
```

```
            alpha = 0.6) +
  geom_smooth(se = FALSE, size = 2) +
  theme_bw(base_size = 16) +
  xlab("Repeated Measure") +
  ylab("heart rate")


ggplot(framingham, aes(x = PERIOD,
                       y = GLUCOSE)) +
  geom_line(aes(group = RANDID), #aes(group = RAND2),
            alpha = 0.6) +
  geom_smooth(se = FALSE, size = 2) +
  theme_bw(base_size = 16) +
  xlab("Repeated Measure") +
  ylab("glucose")


ggplot(framingham, aes(x = PERIOD,
                       y = TIMEMIFC)) +
  geom_line(aes(group = RANDID), #aes(group = RAND2),
            alpha = 0.6) +
  geom_smooth(se = FALSE, size = 2) +
  theme_bw(base_size = 16) +
  xlab("Repeated Measure") +
  ylab("time hosptialization for MI or FC")



# length(unique(framingham$RANDID)) #[1] 4434
# length(unique(framinghamsample$RANDID))


#####LMER fitting and function intro: #####
# lmer stands for Linear Mixed Effects Regression
# Used for continuous outcomes
# Other names:
# - Hierarchical linear models
```

```
# - Linear mixed models
# - Multi-level models
# - Growth models
# lmer arguments:
#   lmer(outcome ~ fixed_effects + (random_effects | individual), data = data)


#load/download these packages:
library(nlme)
library(dplyr)
library(lme4)


#EXAMPLE from Datacamp course:
# BodyWeight <- mutate(BodyWeight, Time = Time - 1)
# body_ri <- lmer(weight ~ 1 + Time + (1 | Rat), data = BodyWeight)
# summary(body_ri)


#FIT ON FRAMINGHAM data:


#random intercept term model
# framingham <- mutate(framingham, PERIOD_0 = PERIOD - 1)


HEARTRATE_ri <- lmer(HEARTRTE ~ 1 + PERIOD + (1 | RANDID), data = framingham)
summary(HEARTRATE_ri)


GLUCOSE_ri <- lmer(GLUCOSE ~ 1 + PERIOD + (1 | RANDID), data = framingham)
summary(GLUCOSE_ri)


#Addition of Random Slope terms in the model
#Allow each individual to have their own trajectory
HEARTRATE_ri2 <- lmer(HEARTRTE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data =
framingham)
summary(HEARTRATE_ri2)
```

```
GLUCOSE_ri2 <- lmer(GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data = framingham)
summary(GLUCOSE_ri2)




# **see video and documentation for help with the analysis.**
#Compare HEARTRATE_ri and HEARTRATE_ri2
#use ANOVA:
anova(HEARTRATE_ri, HEARTRATE_ri2)
anova(GLUCOSE_ri, GLUCOSE_ri2)

#Compound Symmetry:
corr_structure <- function(object, num_timepoints, intercept_only = TRUE) {
  variance <- VarCorr(object)
  if(intercept_only) {
    random_matrix <- as.matrix(object@pp$X[1:num_timepoints, 1])
    var_cor <- random_matrix %*% variance[[1]][1] %*% t(random_matrix) +
      diag(attr(variance, "sc")^2, nrow = num_timepoints,
           ncol = num_timepoints)
  } else {
    random_matrix <- as.matrix(object@pp$X[1:num_timepoints, ])
    var_cor <- random_matrix %*% variance[[1]][1:2, 1:2] %*%
      t(random_matrix) + diag(attr(variance, "sc")^2,
                              nrow = num_timepoints, ncol = num_timepoints)
  }
  Matrix::cov2cor(var_cor)
}

# the custom function corr_structure
# is used to generate the model implied
# correlations from the repeated
# measurements for the 3 time points
# for the first person in the study.
```

```
corr_structure(HEARTRATE_ri, 3) %>%
  round(2)
corr_structure(HEARTRATE_ri2, 3) %>%
  round(2)
# As you can see, the correlations
# acrros the measurements
# are constant and very high
# medicore (about 0.5); that is,
# not close to one.
# the constant correlation
# over time is referred to as,
# in statistics, the
# compound correlation.

corr_structure(GLUCOSE_ri, 3) %>%
  round(2)
corr_structure(GLUCOSE_ri2, 3) %>%
  round(2)



#How does lmer adjust for data dependency?
#-Random effects help control dependency
#-Custom function (next slide) will
# help explore model implied correlations

#Visually show dependency:
#Here, you can see in the upper triangle of the
#figure that is in color of all of the cells are the same,
#indicating the same correlation between
#measurements.
#install.packages("GGally")
library(GGally)
```

```
#heart rate
ggcorr(data = NULL, cor_matrix = corr_structure(HEARTRATE_ri, 3),
       label = TRUE, label_round = 3, label_size = 3.5, palette = 'Set2',
       nbreaks = 5)

ggcorr(data = NULL, cor_matrix = corr_structure(HEARTRATE_ri2, 3),
       label = TRUE, label_round = 3, label_size = 3.5, palette = 'Set2',
       nbreaks = 5)

#glucose
ggcorr(data = NULL, cor_matrix = corr_structure(GLUCOSE_ri, 3),
       label = TRUE, label_round = 3, label_size = 3.5, palette = 'Set2',
       nbreaks = 5)

ggcorr(data = NULL, cor_matrix = corr_structure(GLUCOSE_ri2, 3),
       label = TRUE, label_round = 3, label_size = 3.5, palette = 'Set2',
       nbreaks = 5)
#***see analysis of the correlation structure
#from the online datacamp video***



#####adding predictors to the models above####

#Add a categorical predictor
#Predictors are commonly added
#to the fixed portion of the model

#load/download these packages:
library(nlme)
library(dplyr)
library(lme4)
library(plyr)
library(tidyr)
```

```
library(ggplot2)


#variable SEX
framingham <- framingham %>%
  mutate(SEX_f = paste("SEX", SEX, sep = " "))


#heart rate with sex:
HEARTRATE_ri3 <- lmer(HEARTRTE ~ 1 + PERIOD + SEX_f +
                        (1 + PERIOD | RANDID),
                      data = framingham)
summary(HEARTRATE_ri3)


#glucose with sex:
GLUCOSE_ri3 <- lmer(GLUCOSE ~ 1 + PERIOD + SEX_f +
                      (1 + PERIOD | RANDID),
                    data = framingham)
summary(GLUCOSE_ri3)


#variable DIABETES
# framingham <- framingham %>%
#   mutate(DIABETES_0 = paste("DIABETES", DIABETES, sep = " "))


######heart rate with ___: #####
#converges with CURSMOKE
#heart rate does not converge with
#PREVCHD, MI_FCHD, DIABETES, or PREV_MI
framingham <- framingham %>%
  mutate(CURSMOKE_0 = paste("CURSMOKE", CURSMOKE, sep = " "))
HEARTRATE_ri3_CURSMOKE <- lmer(HEARTRTE ~ 1 + PERIOD + CURSMOKE_0 +
                                 (1 + PERIOD | RANDID),
                               data = framingham)
summary(HEARTRATE_ri3_CURSMOKE)
```

```
model_residuals = residuals(HEARTRATE_ri3_CURSMOKE, na.rm = TRUE)
hist(model_residuals)
#distribution of the random effects:
random_effects <- ranef(HEARTRATE_ri3_CURSMOKE)$RANDID %>%
  mutate(id = 1:nrow(random_effects)) %>%
  gather("variable", "value", -id)
summary(random_effects)


ggplot(random_effects, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~variable, scales = 'free_y') +
  theme_bw(base_size = 14)


#Compare models
# random slope
HEARTRATE_rs <- lmer(HEARTRTE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data = framingham,
REML = FALSE)
#  interaction
HEARTRATE_CURSMOKE_INT <- lmer(HEARTRTE ~ 1 + PERIOD + CURSMOKE_0:PERIOD +
                                 (1 + PERIOD | RANDID), data = framingham, REML =
FALSE)


#AICc comparisons
# install.packages("AICcmodavg")
library(AICcmodavg)
aictab(list(#HEARTRATE_rs,
  HEARTRATE_ri3_CURSMOKE,
  HEARTRATE_CURSMOKE_INT),
  modnames = c(#'random slope',
    'CURSMOKE intercept',
    'CURSMOKE interaction'))


#install.packages("MuMIn")
```

```
library(MuMIn)


#no predictors:
r.squaredGLMM(HEARTRATE_rs)


#PREVMI intercept only:
r.squaredGLMM(HEARTRATE_ri3_CURSMOKE)


#PREVMI interaction:
r.squaredGLMM(HEARTRATE_CURSMOKE_INT)



#####glucose with ___: #####
#converges with PREVCHD, CURSMOKE, DIABETES
#glucose does not converge with  MI_FCHD or PREV_MI.


#GLUCOSE with PREVCHD
framingham <- framingham %>%
  mutate(PREVCHD_0 = paste("PREVCHD", PREVCHD, sep = " "))


GLUCOSE_ri3_PREVCHD <- lmer(GLUCOSE ~ 1 + PERIOD + PREVCHD_0 +
                              (1 + PERIOD | RANDID),
                            data = framingham)
summary(GLUCOSE_ri3_PREVCHD)


model_residuals = residuals(GLUCOSE_ri3_PREVCHD, na.rm = TRUE)
hist(model_residuals)
#distribution of the random effects:
random_effects <- ranef(GLUCOSE_ri3_PREVCHD)$RANDID %>%
  mutate(id = 1:nrow(random_effects)) %>%
  gather("variable", "value", -id)
summary(random_effects)
```

```
ggplot(random_effects, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~variable, scales = 'free_y') +
  theme_bw(base_size = 14)


#Compare models
# random slope
GLUCOSE_rs <- lmer(GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data = framingham,
REML = FALSE)
#  interaction
GLUCOSE_PREVCHD_INT <- lmer(GLUCOSE ~ 1 + PERIOD + PREVCHD_0:PERIOD +
                              (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
#AICc comparisons
# install.packages("AICcmodavg")
library(AICcmodavg)
aictab(list(GLUCOSE_rs,
            GLUCOSE_ri3_PREVCHD), #,
       #GLUCOSE_PREVCHD_INT),
       modnames = c('random slope',
                    'PREVCHD intercept')) #,
#'PREVCHD interaction'))


#install.packages("MuMIn")
library(MuMIn)


#no predictors:
r.squaredGLMM(GLUCOSE_rs)


#PREVMI intercept only:
r.squaredGLMM(GLUCOSE_ri3_PREVCHD)


#PREVMI interaction:
r.squaredGLMM(GLUCOSE_PREVCHD_INT)
```

```
###GLUCOSE with CURSMOKE###
framingham <- framingham %>%
  mutate(CURSMOKE_0 = paste("CURSMOKE", CURSMOKE, sep = " "))


GLUCOSE_ri3_CURSMOKE <- lmer(GLUCOSE ~ 1 + PERIOD + CURSMOKE_0 +
                               (1 + PERIOD | RANDID),
                             data = framingham)
summary(GLUCOSE_ri3_CURSMOKE)


model_residuals = residuals(GLUCOSE_ri3_CURSMOKE, na.rm = TRUE)
hist(model_residuals)
#distribution of the random effects:
random_effects <- ranef(GLUCOSE_ri3_CURSMOKE)$RANDID %>%
  mutate(id = 1:nrow(random_effects)) %>%
  gather("variable", "value", -id)
summary(random_effects)


ggplot(random_effects, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~variable, scales = 'free_y') +
  theme_bw(base_size = 14)


#Compare models
# random slope
GLUCOSE_rs_CURSMOKE <- lmer(GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data =
framingham, REML = FALSE)
#  interaction
GLUCOSE_CURSMOKE_INT <- lmer(GLUCOSE ~ 1 + PERIOD + CURSMOKE_0:PERIOD +
                               (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
#AICc comparisons
# install.packages("AICcmodavg")
library(AICcmodavg)
```

```
aictab(list(GLUCOSE_rs_CURSMOKE,
            GLUCOSE_ri3_CURSMOKE, #,
            GLUCOSE_CURSMOKE_INT),
       modnames = c('random slope',
                    'CURSMOKE intercept',
                    'CURSMOKE interaction'))

#install.packages("MuMIn")
library(MuMIn)

#no predictors:
r.squaredGLMM(GLUCOSE_rs_CURSMOKE)

#PREVMI intercept only:
r.squaredGLMM(GLUCOSE_ri3_CURSMOKE)

#PREVMI interaction:
r.squaredGLMM(GLUCOSE_CURSMOKE_INT)

#GLUCOSE with DIABETES
framingham <- framingham %>%
  mutate(DIABETES_0 = paste("DIABETES", DIABETES, sep = " "))

GLUCOSE_ri3_DIABETES <- lmer(GLUCOSE ~ 1 + PERIOD + DIABETES_0 +
                               (1 + PERIOD | RANDID),
                             data = framingham)
summary(GLUCOSE_ri3_DIABETES)

model_residuals = residuals(GLUCOSE_ri3_DIABETES, na.rm = TRUE)
hist(model_residuals)
#distribution of the random effects:
random_effects <- ranef(GLUCOSE_ri3_DIABETES)$RANDID %>%
  mutate(id = 1:nrow(random_effects)) %>%
```

```
  gather("variable", "value", -id)
summary(random_effects)


ggplot(random_effects, aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~variable, scales = 'free_y') +
  theme_bw(base_size = 14)


#Compare models
# random slope
GLUCOSE_rs_DIABETES <- lmer(GLUCOSE ~ 1 + PERIOD + (1 + PERIOD | RANDID), data =
framingham, REML = FALSE)
#  interaction
GLUCOSE_DIABETES_INT <- lmer(GLUCOSE ~ 1 + PERIOD + DIABETES_0:PERIOD +
                               (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
#AICc comparisons
# install.packages("AICcmodavg")
library(AICcmodavg)
aictab(list(GLUCOSE_rs_DIABETES,
            GLUCOSE_ri3_DIABETES, #,
            GLUCOSE_DIABETES_INT),
       modnames = c('random slope',
                    'CURSMOKE intercept',
                    'CURSMOKE interaction'))


# Explained variance
# Explained variance can help evaluate the model
# Represents the ratio of explained variance to total variance
# Larger values are better
# MuMIn package and r.squaredGLMM() function can be used for calculation
#install.packages("MuMIn")
library(MuMIn)
```

```
#no predictors:
r.squaredGLMM(GLUCOSE_rs_DIABETES)

#PREVMI intercept only:
r.squaredGLMM(GLUCOSE_ri3_DIABETES)

#PREVMI interaction:
r.squaredGLMM(GLUCOSE_DIABETES_INT)

#####Model Comparisons and Explained Variance#####
# The corrected Akaike's information
# criterion (AICc) corrects for small samples
# AICc converges to AIC with large samples

framingham <- framingham %>%
  mutate(PREVMI_f =
           paste("PREVMI", PREVMI, sep = " "))

# random slope
HEARTRATE_rs <- lmer(HEARTRTE ~ 1 + PERIOD +
                       (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
# diet intercept
HEARTRATE_PREVMI <- lmer(HEARTRTE ~ 1 + PERIOD + PREVMI_f +
                           (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
# diet interaction
HEARTRATE_PREVMI_int <- lmer(HEARTRTE ~ 1 + PERIOD + PREVMI_f + PERIOD:PREVMI_f +
                               (1 + PERIOD | RANDID), data = framingham, REML = FALSE)
#AICc comparisons
# install.packages("AICcmodavg")
library(AICcmodavg)
aictab(list(HEARTRATE_rs, HEARTRATE_PREVMI, HEARTRATE_PREVMI_int),
       modnames = c('random slope', 'PREVMI intercept', 'PREVMI interaction'))
```

```
# Explained variance
# Explained variance can help evaluate the model
# Represents the ratio of explained variance to total variance
# Larger values are better
# MuMIn package and r.squaredGLMM() function can be used for calculation
#install.packages("MuMIn")
library(MuMIn)

#no predictors:
r.squaredGLMM(HEARTRATE_rs)

#PREVMI intercept only:
r.squaredGLMM(HEARTRATE_PREVMI)

#PREVMI interaction:
r.squaredGLMM(HEARTRATE_PREVMI_int)




#####Exploring and Modeling Dichotomous Outcomes#####
#install.packages("HSAUR2")
library(HSAUR2)
# Generalized linear mixed model (GLMM)
# Explores the log-odds of success
# Success refers to the outcome coded as 1
# Continuous models are not appropriate
# due to predictions often being out of bounds
# due to mean and variance being related

#ANGINA ~ PERIOD + HEARTRTE
#HOSPMI ~ PERIOD + PREVCHD
#MI_FCHD ~ PERIOD + PREVCHD + BMI + AGE +
```

```
#Changes in the outcome variable over time
# framingham <- framingham %>%
#   mutate(#outcome_dich = ifelse(outcome == "yes or no", 1, 0),
#          PERIOD_0 = PERIOD - 1)
# framingham %>%
#   group_by(PERIOD_0) #%>%
# summarise(prop_outcome = mean(outcome_dich),
#           num = nrow(framingham))


#HOSPMI ~ PERIOD + PREVCHD
HOSPMI_baseline <- glmer(HOSPMI ~ 1 + PERIOD +
                           (1 | RANDID),
                         data = framingham, family = binomial)
summary(HOSPMI_baseline)


HOSPMI_output <- glmer(HOSPMI ~ 1 + PERIOD + PREVCHD +
                         ( 1 | RANDID),
                       data = framingham, family = binomial)
summary(HOSPMI_output)


#MI_FCHD ~ PERIOD + PREVCHD + BMI + AGE +
MI_FCHD_baseline <- glmer(MI_FCHD ~ 1 + PERIOD #+ AGE +
                          + ( 1 | RANDID),
                          data = framingham, family = binomial)
summary(MI_FCHD_baseline)


MI_FCHD_output <- glmer(MI_FCHD ~ 1 + PERIOD + PREVCHD #+ BMI #+ AGE +
                        + ( 1 | RANDID),
                        data = framingham, family = binomial)
summary(MI_FCHD_output)


# Model selection GLMM
# aictab() function from AICcmodavg package can be used for GLMM
```

```
library(AICcmodavg)
aictab(list(MI_FCHD_baseline, MI_FCHD_output),
       c("no PREVCHD", "PREVCHD"))



##### Fit GEE model #####
#install.packages("geepack")
library(geepack)
#https://cran.r-project.org/web/packages/geepack/geepack.pdf

gee_fram <- geeglm(MI_FCHD ~ 1 + PERIOD + PREVCHD, data = framingham,
                   id = RANDID, family = binomial, scale.fix = TRUE)
# Extract model summary
summary(gee_fram)
#to get var-cov matrix
vcov(gee_fram)

gee_fram_exch <- geeglm(MI_FCHD ~ 1 + PERIOD + PREVCHD, data = framingham,
                        id = RANDID, family = binomial,
                        corstr = 'exchangeable', scale.fix = TRUE)
# Extract model summary
summary(gee_fram_exch)
#to get var-cov matrix
vcov(gee_fram_exch)

# Specifying working correlations
# An optional argument, corstr is used to control the working correlation matrix
# Accounts for the dependency due to repeated measures
# The default is independence

gee_fram_ar1corstr <- geeglm(MI_FCHD ~ 1 + PERIOD + PREVCHD,
                             data = framingham,
                             id = RANDID, family = binomial,
```

```
                              corstr = 'ar1', scale.fix = TRUE)
# Extract model summary
summary(gee_fram_ar1corstr)
#to get var-cov matrix
vcov(gee_fram_ar1corstr)



gee_fram_unstructured <- geeglm(MI_FCHD ~ 1 + PERIOD + PREVCHD, data = framingham,
                                id = RANDID, family = binomial,
                                corstr = 'unstructured', scale.fix = TRUE)
# Extract model summary
summary(gee_fram_unstructured)
#to get var-cov matrix
vcov(gee_fram_unstructured)

#Model Selection
# QIC
# QIC = quasi-likelihood under the independence model criterion
# GEE does not use maximum likelihood estimation like GLMM
# QIC needed for GEE
# MuMIn package calculates this statistic
#install.packages("MuMIn")
library(MuMIn)

QIC(gee_fram, gee_fram_exch, gee_fram_ar1corstr, gee_fram_unstructured)
# Evaluating working correlation
# QIC can help select working correlation matrix
QIC(gee_fram_ar1corstr)
QIC(gee_fram_unstructured)

#QIC(gee_fram, )
```