8-5-2015 12:00 AM

Healthy And Unhealthy Statistics: Examining The Impact Of Erroneous Statistical Analyses In Health-Related Research

Britney Allen, *The University of Western Ontario*

Supervisor: Dr. Bethany White, *The University of Western Ontario*
Joint Supervisor: Dr. John Braun, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Statistics and Actuarial Sciences

HEALTHY AND UNHEALTHY STATISTICS: EXAMINING THE IMPACT OF ERRONEOUS STATISTICAL ANALYSES IN HEALTH-RELATED RESEARCH

(Thesis format: Monograph)

by

Britney Allen

Graduate Program in Statistics and Actuarial Science

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Sound statistical analyses are essential to the advancement of medicine. Although certainly not always the case, far too many publications are based on weak or inappropriate statistical methodology, leading to questionable results. Statistical reporting guidelines and standards for research are being introduced which should help curb this problem. Wide recognition of the need for statistical methodologies aligned with research questions and study designs, and the impact when this is not the case, would help prevent this problem. In this thesis, I illustrate the consequences of erroneous statistical analyses on data from an observational study on Multiple Sclerosis and I investigate the impact of inappropriate survival analyses through a simulation study.

Keywords: Survival analysis, p-value, statistical errors, model misspecification, interval censoring

Acknowledgments

Foremost, I must take this opportunity to thank my supervisors, Dr. Bethany J.G. White and Dr. W. John Braun. Thank you to Dr. Bethany White for her thorough and significant research guidance and for her inspiring enthusiasm towards my thesis. Thank you to Dr. W. John Braun for his constructive direction and consistent oversight throughout the progression of this work, and throughout the course of my graduate studies. I am exceedingly grateful to have had the privilege to learn from both of them.

I would also like to acknowledge Meyada Widaatalla for her contribution and mining of the Multiple Sclerosis dataset, which was investigated in her previous study and has greatly motivated my research.

Finally, I would like to express gratitude for the unwavering encouragement and invaluable support I received from my family, friends and Ricky Le.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

High quality research is fundamental to the advancement of medicine but the quality of medical research depends on the valid application, interpretation and reporting of statistics [84, 100, 115]. Poor quality research has serious ethical implications, as it can produce misleading analysis and waste valuable resources [3, 4, 6, 12, 109]. The misuse of statistics has been acknowledged in health-related literature since the first notable systematic review in 1966 [7, 70, 104]. Since then, many studies have reviewed the use of statistics, and subsequent statistical reporting, in various areas of health-related research and have identified substantial errors [5, 7, 13, 15, 47, 63, 64, 70, 73, 75, 89, 90, 92, 97, 101, 107]. Notably, numerous reviews have estimated the prevalence of such errors to be quite high, observing error rates of 38% or higher [47, 63, 64, 89, 90, 92, 101, 119]. A Clinical trial is a special type of health-related study that is strictly monitored and conducted by a large research team, often directly involving those with statistical expertise. Therefore, the prevalence of errors in this research are much lower. Guidelines have been developed and adjustments have been made to journal review processes to help address the issue, particularly for other types of health-related studies [5, 7, 10, 11, 12, 22, 44, 46, 79, 83, 100, 118]. As a result, there have also been some improvements more broadly [60]. However, despite the increased awareness and preventative measures, statistical errors and poor reporting continue to be a real problem in health-related research [44, 100, 106, 109]. Furthermore, few guidelines exist that address the broad set of issues that arise in analyses specific to observational studies [100]. In this chapter, I will describe the errors in statistical analyses and statistical reporting that are commonly found in health and medical

journal publications. Although statistical errors also exist in study design, this thesis only investigates analytical errors.

1.1 Errors in Statistical Analysis

1.1.1 Inference Procedures

A common error that is discovered in systematic reviews of medical research is the application of inappropriate statistical methods [13, 94, 101, 102]. Inappropriate tests are often applied because researchers neglect to check if their data satisfy the test assumptions [57]. A consistent mistake found through many published literature reviews is the violation of independence or the failure to account for paired data [13, 69, 101, 123]. Another common mistake is the inappropriate choice of parametric versus nonparametric statistical tests [39, 67, 69, 73, 101]. The Student's t test, χ^2 test and Analysis of Variance are three types of inference procedures that are frequently applied even when test assumptions are violated [47].

Student's t Test

Many medical research papers have incorrectly applied the Student's t tests [63, 67, 89, 119]. These tests and assumptions are summarized in Table 1.1. The one-sample Student's t test is applied to quantitative data to evaluate the null hypothesis $H_0 : \mu = \mu_0$ against one of the alternative hypotheses $H_A : \mu \neq \mu_0$, $H_A : \mu > \mu_0$ or $H_A : \mu < \mu_0$ [19]. The conditions for the test are that the observations are independent, identically distributed and come from a normally distributed population [19].

The independent samples t test is applied to quantitative data to compare two populations, where μ_1 and μ_2 denote the respective population means [19]. The null hypothesis of the test is usually $H_0 : \mu_1 - \mu_2 = 0$ and is often evaluated against one of the alternative hypotheses $H_A : \mu_1 - \mu_2 \neq 0$, $H_A : \mu_1 - \mu_2 > 0$ or $H_A : \mu_1 - \mu_2 < 0$ [19]. The conditions for the independent samples t test are that the samples are independent, the observations within the samples are independent from each other, and they come from normally distributed populations

[19]. There is a slightly different version of the independent samples t test, called the pooled t test that imposes an additional strong assumption that $\sigma_1 = \sigma_2$ (i.e. it assumes the variances are equal in both populations) [34]. The homogeneity of variances assumption for the pooled Student's t test is commonly violated in medical research [47, 119]. As well, violations of the independent samples assumption have been regularly found in literature [67, 89, 101].

The sample means should follow a normal distribution in order to use the one- and two-sample Student's t tests, but most data used in medical research are not normally distributed [47, 81, 86, 92]. Therefore, it is perhaps not surprising that violations of the normality assumption are found often in systematic reviews of published medical papers [47, 64, 119].

The independent samples t test is not robust to violations of independence as correlations among the errors severely bias test results [19, 29, 45]. In the presence of matched samples or repeated measurements, the Wilcoxon Signed-Rank test is a nonparametric alternative to the paired t test when normality is violated [34]. The independent samples t test can be fairly robust against nonnormality and heteroscedasticity in a variety of situations [21]. However, when both sample sizes and variances are unequal, the true type I error rate may deviate from the nominal significance level [21, 45]. Particularly, the t test is only robust against heteroscedasticity when both populations are normally distributed, and when sample sizes are equal and greater than 15 [48]. Therefore, there is concern regarding the application of parametric tests to observational data, where researchers have less control over sample sizes. Parametric tests are also unreliable in the presence of outliers or heavy skewness [97]. When samples sizes are small and distributions differ in both shape and skew, the sampling distribution of t statistics will typically be skewed and result in biased type I error rates [21]. Nonnormality tends to result in a loss of power [29]. It is of most importance that the condition of normality is satisfied in the case of small sample sizes and such samples are used often in medical research [87, 89, 92, 97, 101]. Therefore, assumption violations of the Student's t test remain problematic, even after consideration of robustness. The Mann-Whitney U test is a nonparametric alternative to the independent samples t test when normality is not satisfied, especially when sample sizes are small [34, 67]. The Welch test is another alternative test when population

variances are unequal [54].

Table 1.1: Comparison of Hypotheses and Assumptions of Student's t Tests

	One-sample t test	Independent Samples t test
H_0	$\mu = \mu_0$	$\mu_1 - \mu_2 = \mu_{d_0}$
H_A	$\mu \neq \mu_0$, $\mu > \mu_0$, $\mu < \mu_0$	$\mu_1 - \mu_2 \neq \mu_{d_0}$, $\mu_1 - \mu_2 > \mu_{d_0}$, $\mu_1 - \mu_2 < \mu_{d_0}$
Assumptions	independence, normality	independence, normality, homogeneous variances

The erroneous technique of examining the overlap and widths of confidence intervals to test for significant differences between groups has been applied in medical research [92]. Relative to the standard method of hypothesis testing, the overlap method is deficient because it has low, inconsistent power and is more conservative [103]. Based on the plot created by Schenker and Gentleman (2001) [103], Figure 1.1 displays the asymptotic probabilities of Type I error for the standard method versus the overlap method for the null hypothesis of no difference between two population quantities. The plot illustrates how the probability for the overlap method depends on the ratio of standard errors for the two populations and never reaches the nominal significance level [103].

χ^2 Test

Systematic reviews of medical literature have identified many papers that display incorrect analyses of contingency tables and application of χ^2 tests [63, 73, 97, 119]. The χ^2 test of independence is performed to evaluate evidence against the null hypotheses of independence

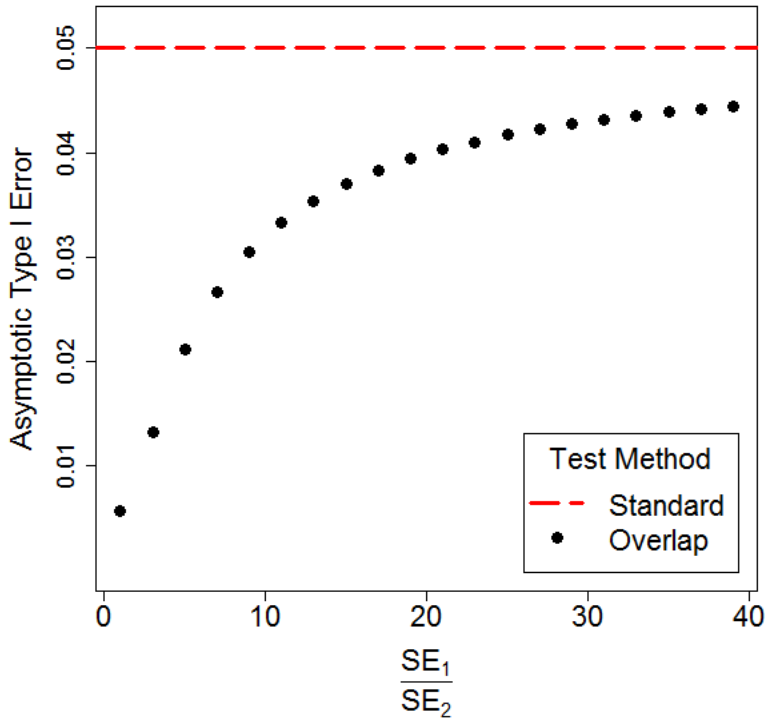

Figure 1.1: Asymptotic comparison of Type I error probabilities for the standard and overlap methods of testing differences between two populations.

between two categorical variables [78]. The χ^2 test can also be applied to multinomial data for the goodness-of-fit test which evaluates evidence against the null hypothesis $H_0 : \pi_1 = \pi_2 = \cdots = \pi_k$, where π_i is the probability of being in category i [34]. Independent observations and large sample sizes are required for valid application of the χ^2 test [78]. Specifically, the 20% rule states that 80% of the expected cell frequencies in the contingency table should be at least five and no expected cell frequency should be less than one [19]. However, the 20% rule is a controversial one, as it is found by some researchers to be conservative [25, 38]. Nevertheless, small expected cell frequencies can lead to poor chi-square approximations of multinomial probabilities [25, 72].

The χ^2 test is not robust to violations of independence, producing meaningless results when both this condition is not satisfied and the P-value is small [19, 72]. However, the application

of a χ^2 test on paired data has been found in medical literature [58]. When the expected cell counts in a 2×2 contingency table are too small, the χ^2 approximation is unsound and Fisher's exact test may be applied instead. However, many authors fail to consider the impact of sample size [13, 67, 105].

ANOVA

The one-way ANOVA is applied to multi-sample, quantitative data to assess the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against the alternative hypothesis that at least one population mean differs from the others [19]. It is essentially an extension of the independent samples pooled t test. The assumptions of ANOVA are that the samples are independent, normally distributed, simple random samples taken from k normally distributed populations with the same variance [19]. Although the previously stated misuses of parametric versus nonparametric tests involve misuses of ANOVA, specific errors concerning the application of ANOVA have also been highlighted in medical research [13, 63, 64]. The ANOVA test is not robust against violations of independence and correlations among the errors severely bias test results [19, 29, 45]. Moreover, consequences of heteroscedasticity are more serious for ANOVA than for t tests [48]. Inflated type I error rates are a consequence of heteroscedasticity, regardless of equality of sample sizes [54]. It is not unusual for heteroscedasticity to exist in medical data; therefore, violations of ANOVA assumptions are concerning [48]. The Brown and Forsythe F^* test is an alternative to one-way ANOVA when the assumption of homogeneous variances is violated [48]. Similar to t tests, nonnormality will likely result in a loss of power in the ANOVA test, especially in the case of extreme skewness [29, 45]. The Kruskal-Wallis Test is an alternative test for comparing the means of k independent groups when the assumption of normality is violated [34, 54].

P-values

In addition to a failure to satisfy test assumptions, another misuse of statistics in health and medical research is the inappropriate use of tests to make multiple comparisons without con-

trolling the overall type I error rate [13, 51, 63, 67, 92, 94, 101, 102]. Failure to adjust the *P*-value when performing multiple applications of a statistical test inflates the overall probability of making a type I error [59, 102]. Reviews of medical literature have discovered a variety of errors surrounding the interpretation and calculation of *P*-values [63, 102].

One prevalent, interpretative issue includes the comparison of *P*-values from separate analyses to evaluate differences between groups [7, 8, 63, 90]. Another prevalent error is the rigid interpretation that *P*-values greater than 0.05 reveal similarities or no effect [18, 31, 39, 123]. The probability of the null hypothesis being true is also a common misinterpretation of the *P*-value [66]. Discrepancies between reported test statistics and degrees of freedom have revealed many *P*-value calculation errors in medical publications [43, 97]. Again, the actual frequency of errors surrounding *P*-values may be underestimated in systematic reviews of medical literature because authors do not always report sufficient information to determine *P*-value accuracy [94, 97].

The *P*-value is only an indicator of how unusual the observed test statistic is under a specific null hypothesis [16]. It was not originally intended to have the central, definitive and objective position in decision making that it has in biomedical research [91, 93]. The dependence on *P*-values has lead to the deceptive practice of *P*-hacking: the fishing for statistically significant results throughout the research process [91]. However, if multiple tests are conducted at a 5% level of significance, 5% of the *P*-values are expected to be no more than 5% even when null hypotheses are true. Many of the "found" statistically significant results may just be type I errors. As well, the definitive interpretation of *P*-values is problematic because *P*-values are random variables that can be very different from sample to sample and are highly dependent on sample size [49, 111]. Moreover, they are unstable and unreliable even when study designs have considerable power [49]. The low test-retest reliability of *P*-values means that they provide poor measures of evidence against the null hypotheses [49]. Alternatively, researchers should use a variety of measures to assess evidence including graphical presentations and confidence intervals [49]. Bayes factors, such as the likelihood ratio statistic, offer another alternative to *P*-values in inference [93]. The Bayesian approach treats unknown pa-

rameters as random variables and the observed data as fixed [71]. In contrast to the frequentist approach (i.e., *P*-values), the probability of the parameters given the data is determined [71]. The Bayesian approach might be preferred over the frequentist approach because it provides straightforward and clear evaluation results [71].

Literature reviews have provided substantial evidence of the misuse of statistical inference procedures. Moreover, these studies likely understate the actual prevalence of test misuse because it is often the case that authors do not report enough information to allow readers to validate the analyses performed [39, 63, 64, 92, 94, 101].

1.1.2 Quantifying Relationships between Variables

Correlation

The Spearman correlation coefficient is a more appropriate measure of correlation than the Pearson correlation coefficient when variables have a nonlinear relationship, are not approximately normal or when data are ordinal [96]. However, the Pearson correlation coefficient is commonly used in medical journals despite assumption violations [63, 64, 96].

Regression

Multivariate linear regression models assume a linear relationship between the response and the associated predictor variables of the form $Y|X = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ [2]. Along with linearity, the regression model assumes that the n responses are independent, normally distributed random variables with constant variance $\text{Var}(Y_i \mid X_1, X_2, ..., X_p) = \sigma^2$ and mean $E(Y_i \mid X_1, X_2, ..., X_p) = \mu_{Y|X} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$ [2].

The logistic regression model also requires independent observations but assumes that the outcomes are binary with a probability of success $P(Y_i = 1 \mid X_1, X_2, ..., X_p) = \pi(X_1, X_2, ..., X_p)$ [2]. As well, continuous predictors must have a linear relationship with their logit outcomes; that is, $\text{logit}\left(\pi(X_1, X_2, ..., X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ [108].

Errors in regression analyses that have occurred often in medical literature include over-

fitting, violations of model assumptions, inappropriate evaluation of predictor variables and improper choice between linear and logistic models [64, 67, 96, 101]. A major issue is that medical researchers neglect clustering and fail to use hierarchical or mixed linear models when appropriate [39]. As well, the flawed approach of using univariate analysis as a variable selection tool in multivariate models is a prevalent error in medical research [75, 101].

Missing data is another source of error in regression analysis, as inappropriate data handling techniques can lead to biased parameter estimates and impact model conclusions [27]. A systematic review of cancer prognostic studies established missing covariate data as a common issue in multivariate survival analysis and found complete case analysis to be the most common approach used by researchers [27].

Additionally, the modeling of time-to-event data is a source of error in medical research [63]. The Cox model estimates multiplicative effects of predictors on the hazard function at time t [56]. The main assumption of the Cox model is proportional hazards, which is the condition that the ratio of hazard functions does not vary across time [56]. Time-dependent bias, which results when time-dependent variables are treated as fixed variables in survival analysis, was frequently found in leading clinical journals [116]. As well, analyzing time-to-event data as fixed time observations, without accounting for censoring or differing follow-up times, is a common issue [123]. A systematic review of research articles in clinical oncology journals revealed an absence in reporting goodness of fit assessments and a lack of reporting proportional hazards assessments for the Cox model [9].

1.1.3 Data Handling

The statistical error of using incorrect methods to analyze categorical versus quantitative data has commonly been identified in medical research [67, 101]. Specifically, the application of inappropriate analysis or descriptives statistics to ordinal data, such as the use of a parametric test, has been observed in literature [13]. Another example of inappropriate analysis is the use of a t test on categorical data [101].

The consequences of categorizing continuous data have been well documented and include

power reduction, information loss and increased type I error rates [74, 99, 110, 117, 122]. Despite the established consequences, the technique of categorizing continuous variables is still used frequently in medical research [8, 18, 36, 39, 75, 94, 95, 101].

Similar to the previous discussion concerning regression analysis, the inadequate handling of missing data is a common issue for other analyses in medical research [18, 39, 51, 53]. Popular strategies include complete case analysis and single imputation methods; both strategies are disadvantageous and lead to biased results [18, 39, 42, 51, 120].

1.2 Errors in Statistical Reporting

The reporting of statistical methods and results in medical research is another area subject to a variety of errors [89, 92].

1.2.1 Omissions

Reproducibility is an important research objective [49]. The International Committee of Medical Journal Editors recommends that authors of biomedical journal manuscripts report enough detail regarding the statistical methods used to allow readers to confirm study results [35]. Yet, the omission of important information continues to be a widespread error in statistical reporting [9, 14, 92, 97, 101]. A substantial problem in statistical reporting is the failure to provide confirmation regarding the validation of assumptions for the statistical methods used [14, 52]. Importantly, model diagnostics are also often omitted in reporting regression analysis [14, 101]. As well, insufficient descriptions of methods used for analyses of survival data are common in medical literature [9, 52].

Sensitivity analysis is recommended to assess the impact of missing data on study conclusions, but is often not included in medical research [18, 51, 53, 120]. Similarly, power analyses are largely omitted in medical literature and tend to only be used after a test fails to generate statistically significant results [39, 41, 49, 52, 89, 92, 119]. In consideration of the important research objective of reproducibility, power analysis is a valuable research tool that quantifies

the reliability of the observed *P*-value in the context of study replication [49].

Although there is some evidence that statistical reporting of methodological details has improved, the level and quality of reporting remains unsatisfactory [60].

1.2.2 Descriptive Statistics

Another statistical reporting issue in medical literature is the presentation of unsuitable numerical summaries of data. The standard deviation is a preferred descriptive statistic for symmetric data, otherwise, the range or interquartile range is preferred [87]. Conversely, the standard error of the mean is an estimate of precision in the sample mean and is an inappropriate description of the variability in a sample [88]. Nevertheless, unsuitable presentations of the standard deviation and the standard error of the mean are detected frequently in medical publications [13, 63, 88, 92, 94].

1.2.3 P-Values

Authors of medical research articles have demonstrated an undesirable reliance on *P*-values to present test results [9, 18, 39, 63, 76, 97]. Valuable information regarding effect size and precision are lost when authors report *P*-values alone without confidence intervals [76]. Authors must report effect size to provide readers with a greater understanding of the strength and relevance of study results, and to allow readers to draw their own study conclusions [31, 111]. The focus on *P*-values has diverted attention away from the practical importance of test results [91]. Moreover, it is common for authors who report confidence intervals to do so in terms of significance testing, rather than interpretation of data [41]. It is important to note that some journals have taken action to address this issue. The journal of *Epidemiology* have discouraged the presentation of *P*-values [68]. As well, the journal of *Basic and Applied Social Psychology* has entirely banned the reporting of *P*-values [121].

1.2.4 Graphical Summaries

Misleading, ambiguous and flawed graphs or figures are often sources of statistical errors in medical journals [9, 32, 63, 96]. An example of a poor graph found in medical publications is the summarization of correlation with the insertion of a regression line in a scatter plot [96]. As illustrated in Figure 1.2, using a regression line to describe correlation can misguide readers into deducing a strong linear relationship where one does not exist [96]. Erroneous graphs of survival curves include unsuitable time scales, absences of distinguishing features between multiple survival curves and displays of survival estimates inconsistent with text content [9].

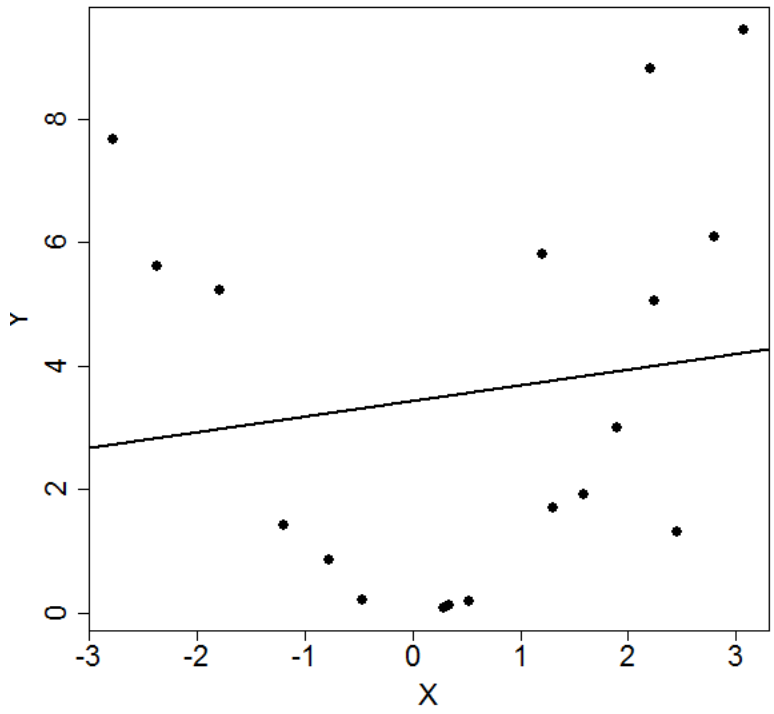

Figure 1.2: The insertion of an irrelevant regression line through the scatter of points alludes to the existence a linear relationship.

Although, errors occur mainly in basic statistical analysis, they are frequent, their consequences can be serious and they result in published papers that draw faulty or misleading

conclusions [8, 47, 70, 119]. Errors in the reporting of statistics can be equally damaging as they cause readers to lose trust in the competence of the author [37].

1.3 Remainder of the Thesis

In the following chapters, I explore the application of erroneous statistical analyses on data from an observational study on Multiple Sclerosis and investigate the impact of inappropriate survival analyses through a simulation study. Contrasting the application of inappropriate and appropriate analyses to the Multiple Sclerosis data in Chapter 2 will result in disagreement between conclusions. Similarly, disparities among the performance of the misspecified models will be observed in the simulation study in Chapter 3.

Chapter 2

Analysis of Multiple Sclerosis Data

The review of literature has established the prevalence and seriousness of inappropriate statistical methodology in medical research. Now, I provide context and relevance to the issue. In this chapter, I illustrate the consequences of erroneous statistical analysis on data from an observational study on Multiple Sclerosis. The data for the application explored here were extracted from the records of 54 patients at the London Health Sciences Centre Multiple Sclerosis Clinic in 2012 [80]. Multiple Sclerosis is an inflammatory disorder where neurological episodes affect the brain and spinal cord [30]. Patients with Multiple Sclerosis experience dysfunctions of their motor, sensory, visual, and autonomic systems [30]. Early diagnosis is important because these symptoms may worsen over time and can result in tissue damage that restricts the performance of daily activities [30]. Due to the importance of early detection of MS, the medical sciences researcher was interested in using these data to study the factors associated with delay time from onset of first symptoms to physician visit.

2.1 Contingency Tables

2.1.1 Example 1 - Analyses of Categorical Outcomes

The dataset contains information on the onset and trigger symptoms experienced by each patient. Onset symptoms are the first Multiple Sclerosis symptoms experienced by the patient, whereas trigger symptoms are defined as the symptoms that lead the patient to seek medical attention. The recorded onset and trigger symptoms were classified as either non-severe or

severe.

Is there a change in severity between onset and trigger symptoms? To address this research question, one may inappropriately construct the 2×2 contingency table displayed in Table 2.1. Then, applying the χ^2 test of homogeneity with Yates' continuity correction, $H_0 : \pi_1 = \pi_2$ is evaluated against the alternative hypothesis $H_A : \pi_1 \neq \pi_2$ and the computed test statistic is $\chi^2 = 14.14$ with a P-value $P < 0.001$. Conforming with the unhealthy reliance on P-values in published medical research, a conclusion of a statistically significant difference in severity is made and the null hypothesis is rejected. However, the independence assumption of the χ^2 test

Table 2.1: Inappropriate Contingency Table: Symptoms by Severity

	Severity		
Symptoms	Non-Severe	Severe	Total by Symptom
Onset	26	28	54
Trigger	7	47	54
Total by Severity	33	75	108

is clearly violated because each observation contributes to two cells in the contingency table.

As a step in the right direction, the contingency table should be constructed as in Table 2.2, where the total counts in the table equal the total sample size. Now, applying the χ^2 Test to Table 2.2 produces a test statistic of 0.84 with $P = 0.36$. The large P-value results in a failure to reject H_0. However, the application of the χ^2 test is still problematic for two reasons. Further calculations show that two expected cell frequencies are less than five. As well, the rows and columns of the contingency table are still not independent since the variables represent repeated measurements of severity at different stages of the disease for the same patient.

Alternatively, the McNemar test is the appropriate test to analyze these paired data in the 2×2 contingency table [58]. The McNemar test is similar to the χ^2 test of homogeneity, but the McNemar test evaluates evidence against the null hypothesis of equal proportions of subjects within the contingency table's discordant cells [58]. Discordant cells represent the pairs

Table 2.2: Contingency Table: Severities of Onset and Trigger Symptoms

	Trigger		
Onset	Non-Severe	Severe	Total
Non-Severe	5	21	26
Severe	2	26	28
Total	7	47	54

which have experienced a change in severity between onset and trigger symptoms and they are identified as counts b and c in the contingency table in Figure 2.3 [58]. The test statistic from the McNemar test statistic is $\chi^2 = 14.09$ with $P < 0.001$. Therefore, inconsistent with the last χ^2 test, the results of the McNemar test indicate that the sample provides sufficient evidence to reject the claim that there is no change in severity between onset and trigger symptoms. The results are summarized for comparison in Table 2.4. This scenario exemplifies the inappropriate analysis of contingency tables that is commonly found in medical literature, demonstrating how the failure to satisfy test assumptions leads to seemingly conflicting and erroneous conclusions.

Table 2.3: Contingency Table: Severities of Onset and Trigger Symptoms

	Trigger	
Onset	Non-Severe	Severe
Non-Severe	a	b
Severe	c	d

Table 2.4: Comparison of tests against the null hypothesis $H_0 : \pi_1 = \pi_2$: Two inappropriate χ^2 tests and the McNemar test.

Test	χ^2	***P*-value**	**Conclusion**
Test of Homogeneity on Table 2.1	14.14	< 0.001	Reject H_0
Test of Homogeneity on Table 2.2	0.84	0.36	Fail to reject H_0
McNemar Test on Table 2.2	14.09	< 0.001	Reject H_0

2.2 Two Sample Inference

2.2.1 Example 2 - Inference on Population Means

The dates of onset symptoms and first clinical visits are also recorded in the dataset. Delay time to clinical visit is defined as the difference in weeks between these two dates. The dataset also records the clinical disease course of each patient: clinically isolated syndrome (CIS) or relapsing-remitting multiple sclerosis (RRMS). CIS and RRMS are two different forms of Multiple Sclerosis. A possible research question of interest may be the following: Is the mean delay time for patients with RRMS equal to the mean delay time for those with CIS? It is important to note that some delay times are unknown, since the self-reported nature of the data led to imprecise onset and first clinical visit dates. Since only completely observed records are used to answer this question, results may be biased. The Student's t test is repeatedly misused in published medical research and could be incorrectly applied in this context. Investigating the research question in the same manner, the independent samples t test is applied and the test statistic is $t = -1.735$ with 23 degrees of freedom and P-value = 0.096. Imitating the conclusive interpretations of P-values found in published medical research, CIS patients are pronounced to not have delay times that are statistically different from RRMS patients. However, the presence of outliers and skew in Figure 2.1 show that the assumption of normality is not satisfied.

Therefore, the nonparametric Mann-Whitney-Wilcoxon test should be applied to reason-

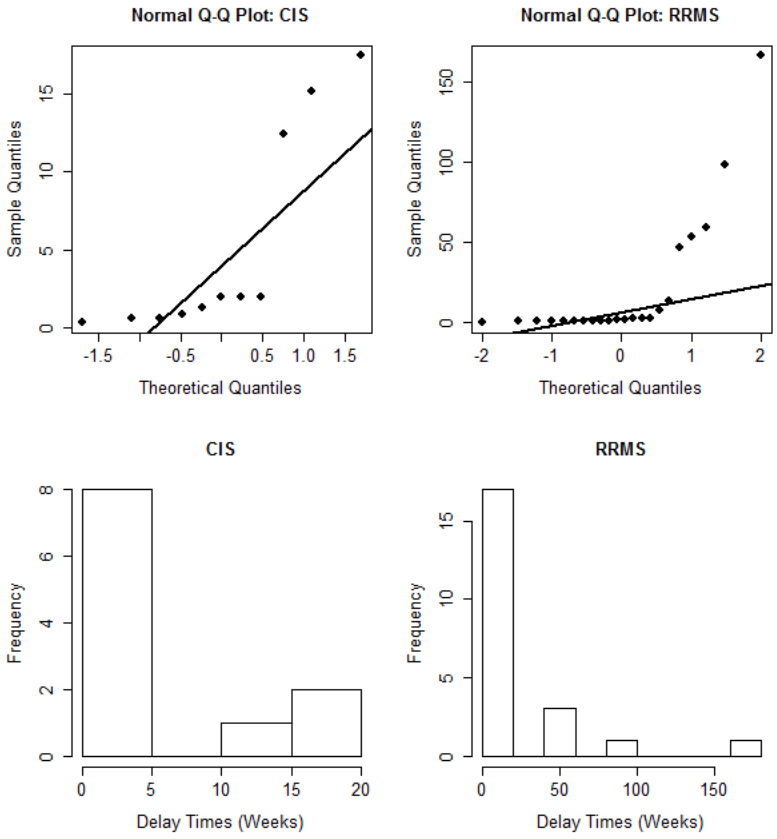

Figure 2.1: Normal quantile-quantile plots and histograms for Multiple Sclerosis patients with Clinically Isolated Syndrome (CIS) and those with Relapsing-Remitting Multiple Sclerosis (RRMS).

ably address this research question. While the assumption of normality is not required for the test, the populations must have equal spread. Table 2.5 displays measures of variability for each group. Although the standard deviation and interquartile range for the two patient groups show that the samples have quite different spread, these measures may be unreliable due to the small sizes of the samples. Therefore, the nonparametric Fligner-Killeen test is used to assess the equality of variances. The test obtains a large P-value equal to 0.63, showing a lack of evidence against unequal variances.

Moving forward with the Mann-Whitney-Wilcoxon test generates test statistic, W = 129 with P-value = 0.77. Thus, much weaker evidence against the null hypothesis is obtained when

Table 2.5: The standard deviation and interquartile range of delay times for patients with Clinically Isolated Syndrome (CIS) and patients with Relapsing-Remitting Multiple Sclerosis (RRMS).

Sample	**Sample Size** **(Complete Data Only)**	**Standard Deviation (Weeks)**	**Interquartile Range (Weeks)**
CIS	11	6.54	6.45
RRMS	22	41.71	11.13

conducting the appropriate analysis.

I have demonstrated how ignoring the assumptions of a test produces adverse results in both Example 1 and Example 2. Noting how often this occurs in published medical literature, this consequence is troublesome.

2.3 Regression Analysis

2.3.1 Survival Models

Survival Analysis refers to the methods applied to analyze time-to-event data, where the response variable is the time until an event of interest occurs [77]. Regression analyses may be applied to model the relationship between survival time and one or more explanatory variables. The essential characteristic of regression models for time-to-event data is that they take censoring into consideration [82]. Censoring is the condition where the survival time is not known exactly [82]. Observations may be right-, left- or interval-censored [77]. Right-censored observations exist when the event has not occurred before the last time point, left-censored observations exist when the event has already occurred before the first time point and interval-censored observations exist when the event is known to have occurred between two time points [82]. Two common types of parametric regression models are the accelerated failure time (AFT) regression model and the proportional hazards (PH) regression model.

Accelerated Failure Time Regression Model

An accelerated failure time (AFT) regression model is of the form

$$\log t_j = \boldsymbol{x_j}\boldsymbol{\beta}_{\text{AFT}} + w_j \tag{2.1}$$

where t_j represents survival time, $\boldsymbol{x_j}$ is the vector of covariates, $\boldsymbol{\beta}_{\text{AFT}}$ is the vector of coefficients and w_j is the random error term [82]. As displayed by Equation 2.1, the systematic effects in the model are assumed to be additive on the natural logarithm of time. Additionally, the accelerated failure time assumption assumes that

$$T_1 = \gamma T_2 \tag{2.2}$$

where T_1 and T_2 are random variables representing survival times for two groups, and γ is a constant acceleration factor equal to $\exp(-\boldsymbol{\beta}_{\text{AFT}})$ [65]. The distribution of the error term decides the regression model [82]. Namely, if w follows the standard extreme value distribution, extreme value distribution, normal distribution, or logistic distribution, then t has the exponential distribution, Weibull distribution, log-normal distribution, or log-logistic distribution respectively [82]. $\boldsymbol{\beta}_{\text{AFT}_i}$ is interpreted as the change in the logarithm of time associated with a unit change in $\boldsymbol{x_i}$, holding all other covariates constant.

Proportional Hazards Regression Model

In proportional hazards (PH) regression, covariate effects are modeled to predict the hazard ratio. The hazard rate is the instantaneous risk associated with the occurrence of an event [82]. The regression model is commonly of the form

$$h\left(t_j\right) = h_0\left(t_j\right)\exp\left(\boldsymbol{x_j}\boldsymbol{\beta}_{\text{PH}}\right) \tag{2.3}$$

where the systematic effects are assumed to act multiplicatively on the baseline hazard function, $h_0(t)$. These models are referred to proportional hazards because it assumes the ratio of hazards between two cases, $h_i(t)/h_j(t)$, is fixed at a proportional amount. That is, proportional differences in the hazard ratio are constant over time. $\boldsymbol{\beta}_{\text{PH}_i}$ is interpreted as the increase

in the log hazard rate associated with a unit increase in $\boldsymbol{x_i}$, holding all other covariates constant. However, some models may be expressed in either PH or AFT form. In fully parametric models, the baseline hazard function is assumed to follow a specific distribution such as exponential or Weibull [77]. When $h_0(t)$ is unspecified, Equation 2.3 fits the semi-parametric Cox proportional hazards model [82].

2.3.2 Example 3 - Analyses of Time-to-Event Data

Early treatment is considered to be advantageous in treating Multiple Sclerosis [40]. Notably, delay time until first clinical visit has been found to be the greatest contributor to delay in treatment [40]. Accordingly, a research question of interest may be: Which factors impact the delay time of Multiple Sclerosis patients? As mentioned in Example 2, some delay times are imprecise since the data are self-reported. These delay times are only known to be within an interval of values. To address this research question, one may inadequately handle the presence of interval censoring by imputing the intervals' midpoint values, and then use multiple linear regression to model delay time against covariates. The dataset includes variables describing patients' symptom severity, family history, fear or concern, presence of multiple symptoms or infected areas, influence of daily activities, age and gender. Applying the automated selection techniques of backwards elimination or forward selection, age of onset is found to be the only statistically significant covariate. The results from the inappropriately fitted linear regression model on the 54 patients are displayed in Table 2.6.

Table 2.6: Results from the inappropriately fitted linear regression model

Covariate	**Coefficient**	**95% Confidence Interval**	***P*-value**
Age at Onset (Years)	-8.70	$(-13.36, -4.04)$	< 0.001
Intercept	328.18	$(182.37, 473.98)$	< 0.001

Furthermore, residual-based diagnostics support inadequacies of the model. Figure 2.2 shows that the Normality assumption is not reasonable since the distribution of residuals appear

heavily right skewed, due to the presence of outliers. As well, the residuals in Figure 2.3 are not randomly scattered about zero and violate the constant variance assumption.

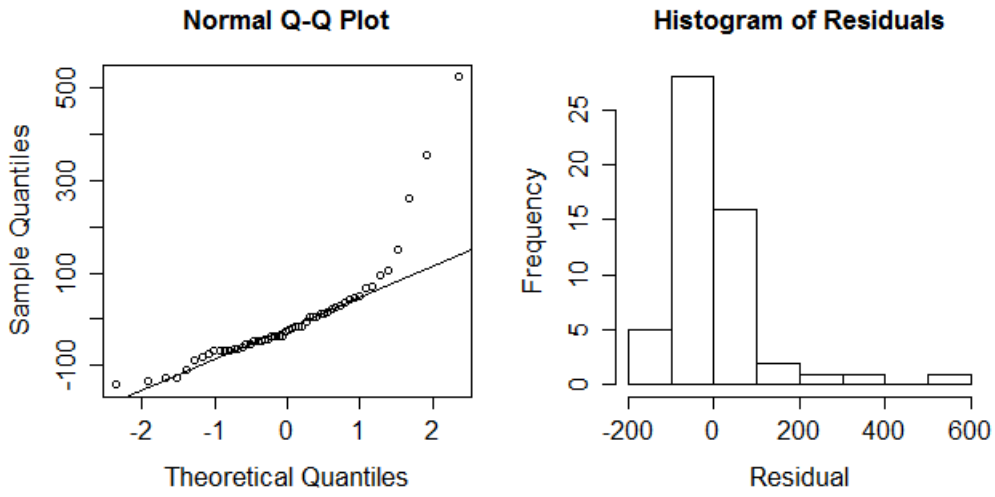

Figure 2.2: Normal quantile-quantile plot and histogram of residuals from the linear regression of imputed delay time (midpoint) on Age at Onset for evaluation of the normality assumption.

Clearly this linear model poorly represents the data. To appropriately model this interval censored, time-to-event data, survival analysis should be applied. As well, covariate selection that is based solely on statistical significance may produce results that are not very useful [24]. Therefore, a combination of previous knowledge and stepwise regression are used to select model covariates. It is well established that age and gender influence Multiple Sclerosis symptoms [20, 33, 50]. Accordingly, the effects of symptom severity, family history, patient fear or concern, presence of multiple symptoms or infected areas, and influence of daily activities are investigated after adjusting for age and gender. It is important to note that a small constant has been added to each delay time to address the presence of times equal to zero. Weibull, Exponential, Log-Normal and Log-Logistic AFT survival models are fit using the `survreg()` function in the `survival` package, where the type of censoring for the `Surv` object is specified as "interval2" [112, 113]. Only severity of onset symptoms is found to be a statistically significant covariate. These results differ considerably from the previous linear regression model.

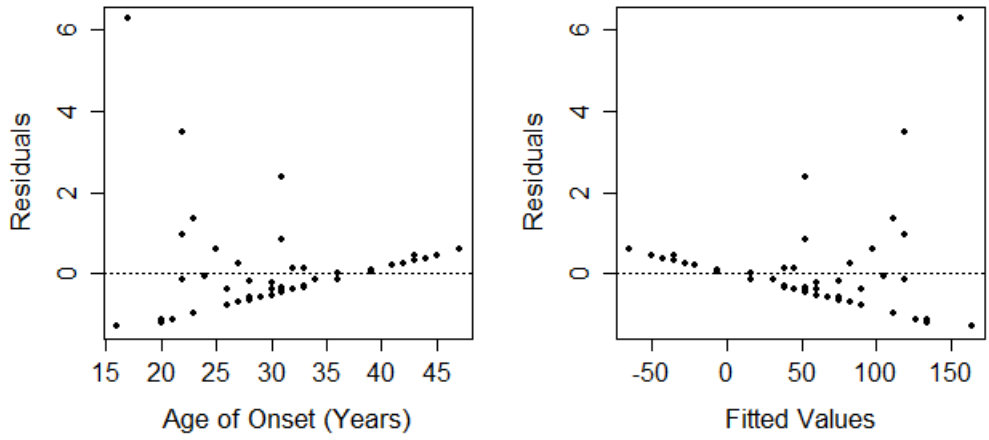

Figure 2.3: Studentized residuals from the linear regression of imputed delay time (midpoint) on Age at Onset plotted against Age of Onset and fitted values for evaluation of the constant variance assumption.

Akaikes information criterion (AIC) and the log-likelihood are then used to compare the different parametric models. Table 2.7 displays the AIC statistics and log-likelihoods for the four competing models. The Weibull model has the smallest AIC and the highest log-likelihood, suggesting a better fit, relative to the other models.

Table 2.7: Akaike's Information Criterion for the four competing models

Model	**AIC**	**Log-Likelihood**
Exponential	333.41	-162.70
Weibull	308.72	-149.36
Log-Normal	314.50	-150.59
Log-Logistic	311.17	-152.25

The results of the fitted Weibull regression model are displayed in Table 2.8.

The coefficient estimates are provided in AFT form. However, the Weibull regression

Table 2.8: Results from the fitted Weibull accelerated failure time regression model

Covariate	**Coefficient β_{AFT}**	**95% Confidence Interval**	***P*-value**
Age at Onset (Years)	−0.167	(−0.232, −0.101)	< 0.001
Gender (Male)	−0.172	(−1.293, 0.949)	0.76
Severity of Onset (Severe)	−1.030	(−1.994, −0.066)	0.04
Intercept	8.631	(6.529, 10.732)	< 0.001
Scale	1.70	(1.359, 2.123)	< 0.001

model can be estimated in both AFT and PH form through the transformation

$$\beta_{\text{PH}} = -\beta_{\text{AFT}} * a \tag{2.4}$$

where a is the Weibull shape parameter [82]. The Weibull regression model also has the feature that if the proportional hazards assumption holds, then the accelerated failure time assumption holds, as well [65]. The proportional hazards assumption is assessed by examining plots of the logarithm of the estimated cumulative hazard function against the logarithm of time. The cumulative hazard function is estimated by using the Kaplan-Meier method to estimate survival probabilities, $\widehat{\text{S}}(t)$, and applying the formula $\log(-\log \widehat{\text{S}}(t))$ [24]. Figure 2.4 displays plots of the logarithm of the cumulative hazard function against the logarithm of time for each covariate in the model. Note that age of onset has been categorized into percentiles to facilitate the plot. The assumption of proportional hazards is plausible if the plotted lines are parallel [24]. Since the lines intersect in several areas, proportional hazards is clearly violated by the gender and age covariates. Conversely, the lines for severity of onset appear reasonably parallel, except for the earlier survival times. Furthermore, the lines in the plots should be approximately straight if the assumed Weibull distribution is appropriate [65]. Although gender is questionable, the plots in Figure 2.4 roughly support this assumption.

Next the accelerated failure time assumption is assessed by examining a quantile-quantile plot of survival times. Figure 2.5 displays the quantile-quantile plot only for the severity of onset covariate, since its survival estimation is of primary concern. A straight line through

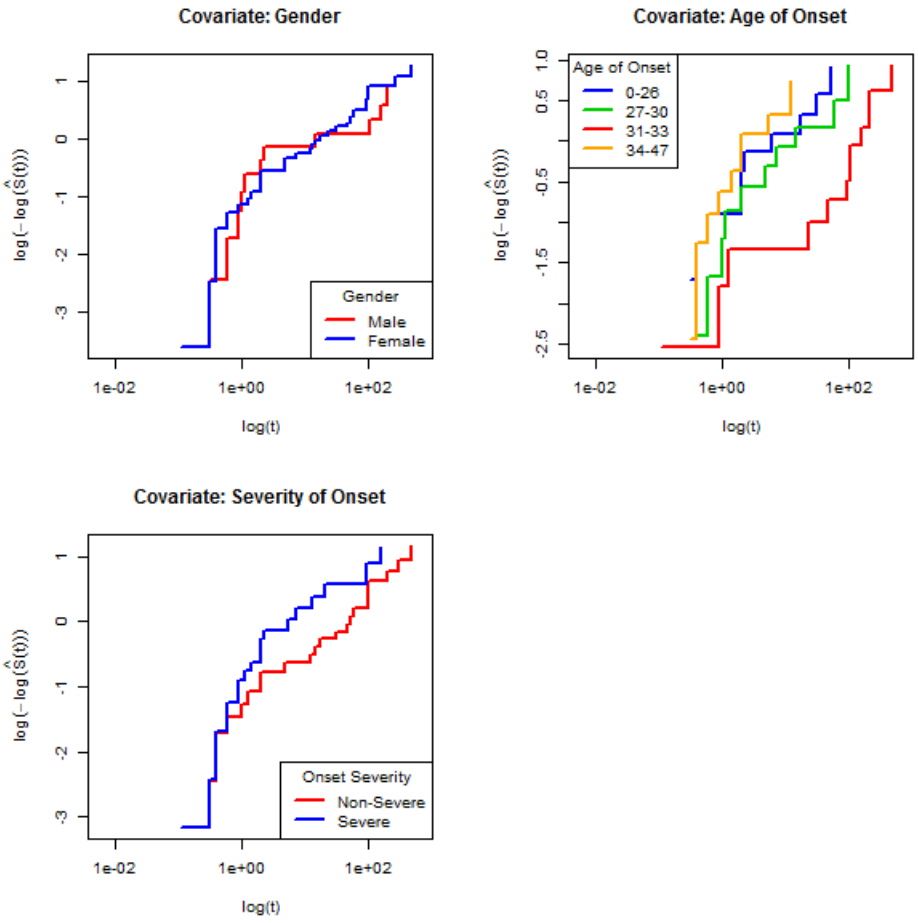

Figure 2.4: Logarithm of the cumulative hazard function for each group plotted against the logarithm of time.

the origin with slope $\exp(\widehat{\beta})$ has been added to the plot. If AFT is appropriate, the percentiles should lie along this line [24]. The plot suggests the AFT is reasonable.

To further assess model adequacy, Figure 2.6 displays predicted survival curves by the Weibull model along with the survival curves estimated by the Kaplan-Meier method for each severity group. To facilitate prediction, age was fixed at its mean value and gender was set to female. Although the probabilities are not very close, survival appears to be reasonably estimated by the model. Results are similar for males.

Next, the deviance residuals for the model are examined in Figure 2.7. The residuals are

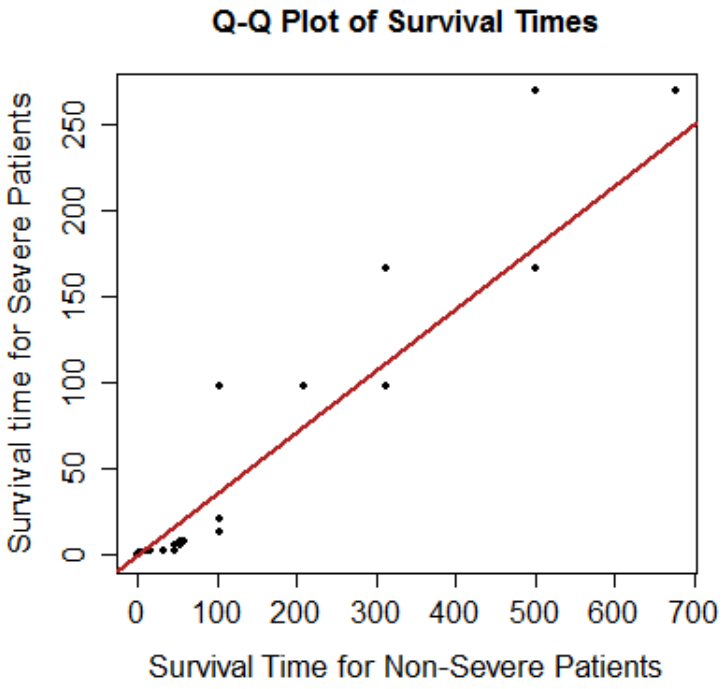

Figure 2.5: Quantile-Quantile plot of survival distribution for patients with non-severe onset symptoms against those with severe onset symptoms.

randomly scattered about zero and there are no concerning outlying points.

This example has made it evident that poor regression modeling and inadequate data handling will produce a defective model, as is often the case in published medical literature. Despite the prevalence of these statistical errors, regression analyses are frequently applied in medical research to make decisions. Hence, this is an issue with the potential to have serious implications.

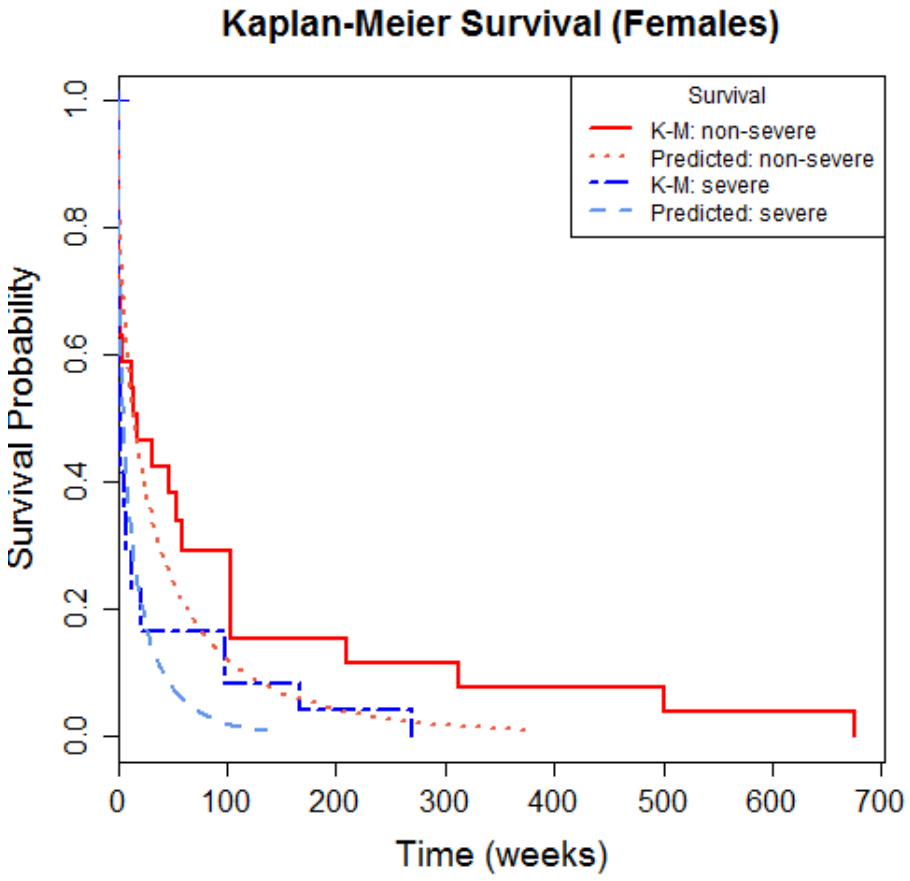

Figure 2.6: Comparison of Kaplan-Meier estimates for survival probabilities with predicted survival from the Weibull regression model

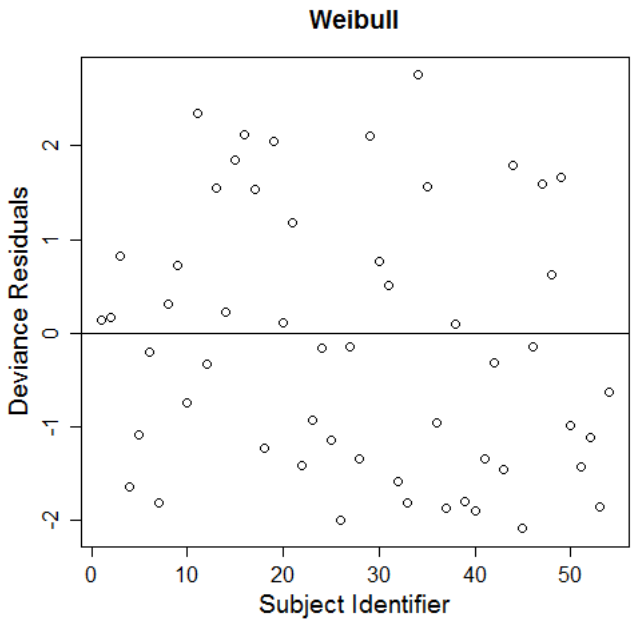

Figure 2.7: Deviance residuals from the Weibull regression model plotted against the subject identifier.

Chapter 3

Simulation Study

3.1 Objective

The previous chapter illustrated how quite different conclusions can be reached depending on whether or not appropriate analyses are used in the context of an MS application. The review of medical literature has established that the inappropriate modeling of time-to-event data, the inadequate handling of data and the severe reliance on *P*-values are all common sources of error. Since *P*-values have highly trusted and central roles in published research, it is important to investigate their behaviour. Therefore, in consideration of the prevalence of errors, the objective of the simulation study is to assess the performance of *P*-values through the application of inappropriate analyses on time-to-event data under different parameter settings.

3.2 Methods

3.2.1 Simulation Procedures

Two independent simulation studies were conducted using the statistical software R [98]. Each study consisted of 5000 simulations. The studies each used a different data generation procedure, explored different parameter scenarios and applied distinct methods of inappropriate modeling. For each simulated dataset, several statistical models were fit to allow for comparison of their performance across all of the 5000 simulated samples.

3.2.2 Data Generation

Study A - Comparison of Models without Censoring

The goal of the first simulation study, Study A, was to investigate the performance of the *P*-value for a hypothesis test associated with one covariate effect, subject to inappropriate model fitting. Time-to-event data were randomly generated of the form

$$Y = \log T = \frac{1}{a}W - \log\frac{1}{b} - \frac{1}{a}\beta_{\mathrm{PH}}X' \tag{3.1}$$

where W follows a standard extreme value (type I) distribution and β_{PH} denotes the coefficient in proportional hazards form. That is, uncensored time-to-event data were generated under the assumption of proportional hazards from a Weibull model with one covariate, as described by the hazard function

$$\mathrm{h}(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1}\exp(x'\beta_{\mathrm{PH}}) \tag{3.2}$$

where the Weibull shape and scale parameters were a and b, respectively. The Weibull distribution was selected for the simulation studies because it is highly versatile and is applied often in survival analysis [23, 61]. The flexibility of the Weibull distribution is illustrated in Figure 3.1, where the scale parameter varies from 0.5 to 2.

The inverse transform method was used to generate w from the extreme value distribution as recommended by Aaserud [1]. First, u was randomly sampled from the uniform distribution over the interval [0,1]. Then, the inverse cumulative distribution function

$$F^{-1}(u) = -\ln(-\ln u) = w \tag{3.3}$$

was applied to transform the values from the uniform distribution to those of the extreme value distribution.

Study B - Comparison of Models with Censoring

The goal of the second simulation study, Study B, was to investigate the performance of the *P*-value for a hypothesis test associated with a covariate effect, subject to the inadequate handling of interval censored time-to-event data. As in Study A, Equation 3.1 was used to generate

time-to-event data from a Weibull model with one covariate, under the assumption of proportional hazards. Next, Study B incorporated interval censoring through two separate approaches. The first approach implemented random, non-informative censoring and the second approach implemented dependent, informative censoring, both of which are described below.

The random censoring mechanism began by determining which observations were complete, s according to

$$s_i \sim \text{Bernoulli}\,(1-c) \tag{3.4}$$

where c is the specified probability of obtaining a censored time. Then, the upper and lower bounds for each interval censored time were constructed as

$$t_{a_i} = t_i - \epsilon_{a_i}$$

$$t_{b_i} = t_i + \epsilon_{b_i}$$

where ϵ_{a_i} and ϵ_{b_i} are independent $\text{Exp}\,(1)$ random variables. Thus, (t_{a_i}, t_{b_i}) was the time input for the response variable in the model. The exponential distribution with rate parameter $\lambda = 1$ was chosen to construct the intervals in order to generate intervals with widths that were similar to the interval censored delay times in the Multiple Sclerosis dataset.

For informative censoring, the distribution of censoring was dependent on time to first physician visit. It seems reasonable that the longer the delay time, the less precise the patient's recollection of the timing of his/her first symptoms. The probability of censoring, c_i was determined for each t_i generated by Equation 3.1 according to

$$c_i = \begin{cases} 0.2, & \text{if } t < 24 \\ 0.4, & \text{if } 24 \leq t \leq 48 \\ 0.6, & \text{if } t > 48 \end{cases}$$

where the cutoff points of 24 weeks and 48 weeks relate to the context of the Multiple Sclerosis dataset. Particularly, it is reasonable to suppose that censoring rates would increase the farther patients were from the date of their onset symptoms. Next, a censoring status was assigned to each time by randomly sampling a number from $\text{Bernoulli}\,(1-c_i)$ where a success represented

an uncensored observation. As in the random censoring scenario, the interval bounds were determined using observations of two independent Exp(1) random variables.

In the case where a censoring mechanism computed a non-positive lower bound for an interval censored time, the lower bound was set equal to `NA`. This ensured valid time-to-event data, but these cases represented a left censored survival event in the simulation.

3.2.3 Scenarios Investigated

The numerous scenarios investigated in the simulation study were motivated from the Multiple Sclerosis dataset. In the studies, parameters values were varied as presented in Table 3.1 and Table 3.2. A single covariate was considered in each module: a continuous one (i.e., $X \sim \text{N}(0, \sigma^2)$) and a discrete one (i.e., $X \sim \text{Bin}(1, p)$). Next, I describe the parametric scenarios explored in each study.

Study A - Comparison of Models without Censoring

The various parameters investigated in Study A are presented in Table 3.1. The time was simulated to follow a Weibull distribution with a shape parameter that ranged from 0.5 to 2.0. The four baseline Weibull hazard functions from which data were generated are displayed in Figure 3.1. The range of shape values was chosen to achieve a variety of Weibull models with monotone decreasing, constant and monotone decreasing hazards. As well, these shapes were selected to include a shape similar to the parameter estimate obtained from the Multiple Sclerosis data. Similarly, the decision to hold the Weibull scale parameter constant at $b = 100$ was influenced by the high scale estimated in the Weibull model for the Multiple Sclerosis data. The AFT coefficient for the covariate in the simulation ranges from -2 to 2. The values for the coefficient were selected in order to allow exploration of a variety of cases that vary from a model with no effect to a model with a relatively strong one. It is important to note that in one covariate scenario, X was normally distributed with a centered mean and a standard deviation of either 1 or 5. These standard deviations were chosen to include the standardized case and a value similar to that of the continuous covariate in the fitted Weibull model for the

Multiple Sclerosis data (i.e., age of onset). In the other covariate scenario, X is Bernoulli with a a probability of 0.5 or 0.8. Lastly, the sample sizes of 50 and 250 were simulated to compare a small sample size that was similar to the size of the Multiple Sclerosis dataset, to a relatively large sample size. The probabilities for the binomial covariate were motivated by dichotomous variables in the Multiple Sclerosis dataset (e.g., severity of onset symptom). A fully factorial arrangement of the parameter values in Table 3.1 form the 416 scenarios considered in the simulation.

Table 3.1: Parameters for the Weibull and covariate distributions investigated in Simulation Study A

Parameter	**Values**
Weibull Shape (a)	0.5, 1, 1.5, 2
Coefficient (β_{AFT})	± 2, ± 1.5, ± 1, ± 0.75, ± 0.5, ± 0.25, 0
Normal Covariate Scale (σ)	1, 5
Bernoulli Covariate Probability (p)	0.5, 0.8
Sample Size (n)	50, 250

Study B - Comparison of Models with Censoring

The various parameters investigated in Study B are presented in Table 3.2. Study B focused on a subset of scenarios from Study A. The time was simulated to follow a Weibull distribution with a shape of 0.5 and 2. These two shapes were chosen to target shapes which produced considerably dissimilar results in Study A. The AFT coefficient for the model varied from -1 to 1, where the extreme values of the coefficient were omitted since they generally produced uninteresting results in Study A. The standard deviation for the normally distributed covariate was fixed at 1 because the standardized case produced the most diverse results in Study A. The remaining parameters: Weibull scale, Bernoulli probability and sample size were held constant at 100, 0.5 and 50, respectively. Additionally, the random censoring mechanism explored probabilities of censoring equal to 0.2 and 0.4. The censoring probabilities were chosen to foster

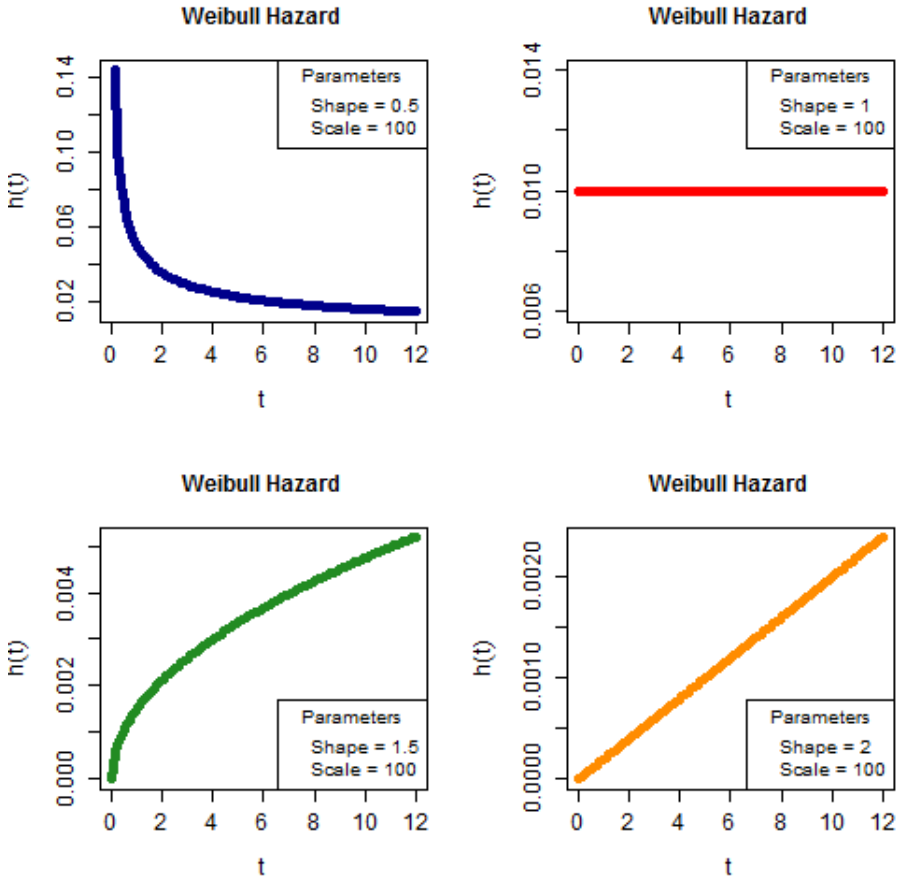

Figure 3.1: The four baseline Weibull hazard scenarios from which Study A generates data. The scale is fixed at 100. The shape assumes the values 0.5, 1, 1.5 and 2.

the comparison of low and high censor levels. However, the small sample size of 50 made it impractical to consider censoring probabilities closer to 1. Again, a fully factorial arrangement of the parameter values in Table 3.2 form the 36 scenarios considered in the simulation.

3.2.4 Methods Evaluated

After time-to-event data were generated under each scenario, the studies apply various models to estimate the coefficient for the covariate.

Table 3.2: Parameters for the Weibull, covariate and censoring distribution investigated in Simulation Study B

Parameter	**Values**
Weibull Shape (a)	0.5, 2
Coefficient (β_{PH})	± 1, ± 0.75, ± 0.5, ± 0.25, 0
Random Censoring probability (c)	0.2, 0.4

Study A - Comparison of Models without Censoring

Study A used the `survreg` function in the `survival` package to fit Weibull, log-logistic, log-normal and exponential parametric survival regression models on log transformed time, Y. As well, `coxph` was used to fit a Cox proportional hazards regression model to the data, where the Breslow method was used to handle tied event times. Furthermore, `lm` was used to fit two simple linear regression models to the data; one on time and one on log transformed time. Given the way the data were generated, the Weibull model was the correct model and the other six models were considered inappropriate (i.e., were misspecified). The four inappropriate survival models were included because they represented the frequent error in medical literature of fitting a model to data that violate assumptions. Similarly, the two simple linear regression models were fit to serve as typical applications of inappropriate methodology in medical literature.

It is important to note that the Weibull and Cox model fitting procedures tracked failures to converge within the simulation. In the rare occurrence of lack of convergence, the sample was discarded and new data were generated to evaluate all methods.

Study B - Comparison of Models with Censoring

Study B used `survreg` to fit three Weibull survival regression models. The first Weibull model classified the type of censoring for the survival object, `Surv` as "interval2". Given the way the data were generated, the first Weibull model was the correct model. Censoring was ignored

in the second Weibull model where the midpoint of the time interval was used as the response variable. In the third Weibull model, censored observations were completely disregarded; only times when events were observed were included in the model. The latter two models were evaluated because they represented inadequate approaches of handling incomplete data.

In the same manner as Study A, the three Weibull models were monitored for lack of convergence.

3.2.5 Quantities Stored

The goal of both simulation studies was to assess the performance of inference on β by looking at the results through P-values. Accordingly, the simulations stored the two-tailed P-value associated with the estimated coefficient for the covariate in each model, based on the Wald test. It is also possible to assess the performance of inference on β by inspecting the corresponding confidence intervals. For that reason, each study stored the proportion of estimated confidence intervals that fell entirely above (i.e., the lower limit was positive) or below 0 (i.e., the upper limit was negative). The confidence intervals were constructed as

$$\widehat{\beta} \pm z_{\alpha/2} se(\widehat{\beta}) \tag{3.5}$$

where α was set at 5%. This allowed evaluation of which methods incorrectly predicted the direction of association between survival time and the covariate.

Where applicable, the number of datasets that experienced convergence failure during model fit in each study were recorded to allow judgment of the methods' reliability.

3.3 Results

The simulations have revealed some marked differences between the performances of the various misspecified models. In this section, I describe the findings from Study A and Study B in detail.

3.3.1 Study A - Comparison of Models without Censoring

First, plots of the median P-values against β for each model were examined. Each model was found to have a maximum median P-value at $\beta = 0$. This was anticipated since when $\beta = 0$, the null hypothesis is true. One notable finding was that scenarios with the Bernoulli covariate tended to produce more discernible differences between the models than scenarios with the normal covariate. Figure 3.2 displays an example of this finding. Although, this is for the specific scenario where the sample size was set at 50, similar results were observed across other scenarios. This shows that inference on β for the discrete covariate was more affected by inappropriate modeling than inference on β for the continuous covariate, based on the results from this study. As well, the plot shows that the exponential model and the linear model regressed on the logarithm of time tended to have higher P-values than the other models, which indirectly motivate power and Type I error rate interpretations. Namely, higher median P-values suggested a tendency to reject H_0 less often. Conversely, the correct Weibull model tended to have the lowest median P-values across various scenarios, suggesting higher power and higher (i.e., closer to nominal) Type I error rate.

Another notable observation was that, apart from the scenario where Weibull shape = 1, the exponential model consistently behaved very differently from all the other models. This observation is seen in both of the plots in Figure 3.2. Additionally, Figure 3.3 illustrates this result for the scenario with a Bernoulli covariate and Weibull shape of 0.5. In contrast with Figure 3.2, the median P-values in Figure 3.3 are much lower for the exponential model than for the other models.

The disparity detected between the models in the median P-value plots have exposed a distinct pattern regarding the behaviour of the exponential model. Specifically, the behaviour of the P-value for β in the exponential model was linked to the value of the Weibull shape parameter. As seen in Figure 3.3, when the shape parameter was 0.5, the median P-values for the exponential model fell below the medians associated with the other six models. Conversely, Figure 3.4 shows that when the shape parameter was 1.5 or 2, the median P-value fell above

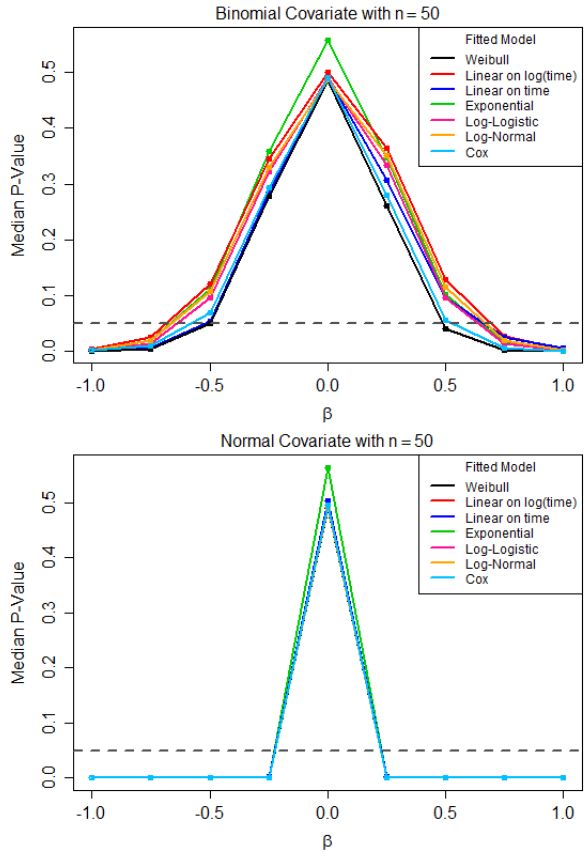

Figure 3.2: Median P-values plotted against β values for each model under the scenarios of Binomial and Normal covariates with sample size = 50.

those for the other models. When the shape parameter was equal to 1, the median P-values were very similar to those of the other models. Figure 3.3 and 3.4 illustrate this pattern under specific scenarios, but the result is consistent across other scenarios. It was clear that the behaviour of the P-value for the exponential model fluctuated depending on Weibull's shape, indicating that the performance of a test of $H_0 : \beta = 0$ is very sensitive to model misspecification involving the use of exponential models when a Weibull should be used.

While the exponential model was noticeably different from the other models, the other models also demonstrated apparent patterns in relation to each other. The median P-values corresponding to the linear model regressed on the logarithm of time tended to perform very closely to the log-normal survival model. This is a expected result since both models have

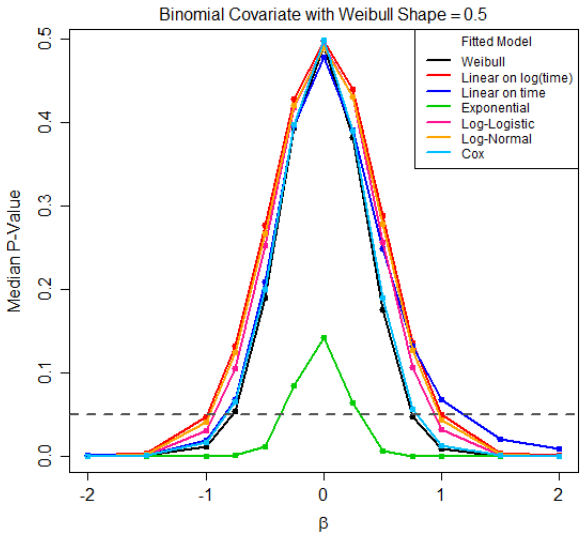

Figure 3.3: Median P-values plotted against β values for each model under the scenario of a Binomial covariate with Weibull shape = 0.5.

the same response variable and impose an assumption of normality. As well, the linear model regressed on time performed similarly to the log-logistic survival model. Likewise, the median P-values for the Weibull and Cox survival regression models performed similar to each other. Since the Cox model did not impose a distributional assumption for the baseline hazard function, its close performance to the correct Weibull model was sensible.

A more straightforward comparison of model performances occurred by examining the proportions of P-values that were below the significance level, $\alpha = 0.05$. Assessment of these proportions against values of β for each model allowed direct, empirical estimations of Type I error and power. The empirical Type I error rates were evaluated when $\beta = 0$ and power estimates were evaluated when $\beta \neq 0$. Again, the performance of exponential model related to the Weibull shape parameter. An instance of this is displayed in Figure 3.5, where the Weibull shape parameter was 2. Aside from the drop in power experienced by the Cox model, the plot shows that the exponential model obtains a lower Type I error rate and lower power, relative to the other models. This tended to be the case when the Weibull shape parameter is 1.5 or 2. Conversely, when the Weibull shape parameter was 0.5, the exponential model obtained a

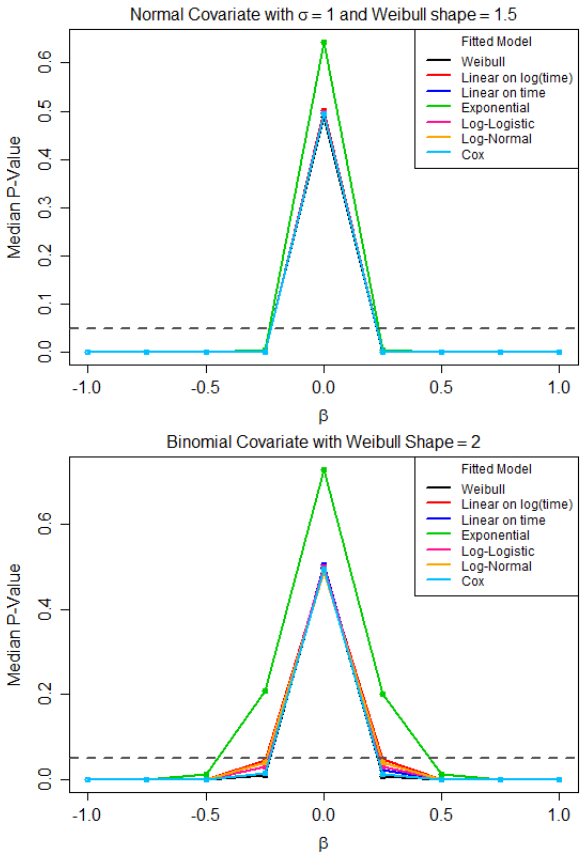

Figure 3.4: Median P-values plotted against β values for each model under the scenarios: Standard Normal covariate with Weibull shape = 1.5 and Binomial covariate with Weibull shape = 2.

higher Type I error rate and higher power relative to the other models. The decrease in power observed for the Cox model in Figure 3.5 may be due to the combined use of the Wald test to assess the significance of a binary covariate and the Breslow method to handle tied event times [62].

As well, the patterns displayed between the models with respect to their median P-values were consistent with the patterns observed between the models with respect to their P-value proportions. Apart from the erratic performance of the exponential model, the correct Weibull model was found to consistently have the highest power across all scenarios. The linear model regressed on the logarithm of time tended to have the lowest proportion of P-values less than α,

until the values of $|\beta|$ are greater than 0.75, where the linear model regressed on time then had the lowest proportion. As well, the proportions corresponding to the linear model regressed on time were found to be higher for negative coefficients and were not symmetric about $\beta = 0$, especially in the case of the Bernoulli covariate. Consequently, both linear models conduct inference on β with weak power, but the linear model regressed on time also demonstrated inconsistent performance. Excluding the exponential model, the patterns among the models and their relative order of performance did not change across the various values of the Weibull shape parameter. However, the performances of the models did become more similar.

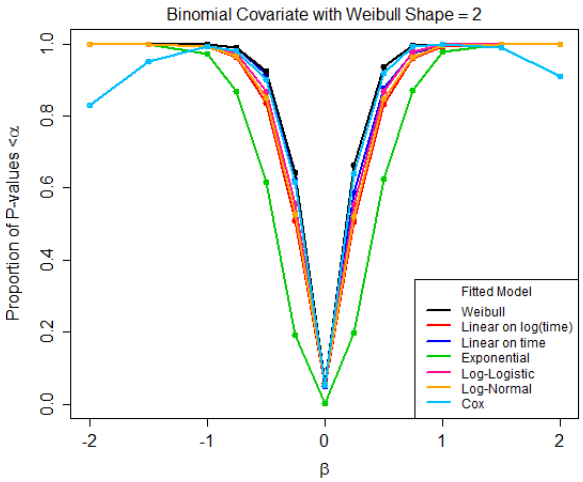

Figure 3.5: Proportion of P-values < 0.05 plotted against β for the scenario where Weibull shape = 2.

Under the null hypothesis of $\beta = 0$, P-values should look like independent and identically distributed uniform random variables [26]. Histograms were constructed to assess if simulation results were consistent with this property. It was found that every model, except for exponential, roughly satisfied this uniformity condition. Hence, inference on β did not appear to be as problematic in these inappropriate models. Figure 3.6 shows estimated densities for each model under the null hypothesis where the sample size was 250 and the Weibull shape parameter was 1.5. The histogram corresponding to the exponential model in this case is asymmetric. Clearly, testing the null hypothesis, $H_0 : \beta = 0$ was not valid in the exponential model. Histogram find-

ings were generally the same under the null hypothesis for other simulated scenarios; however, the approximation to uniform was best when the sample size was large.

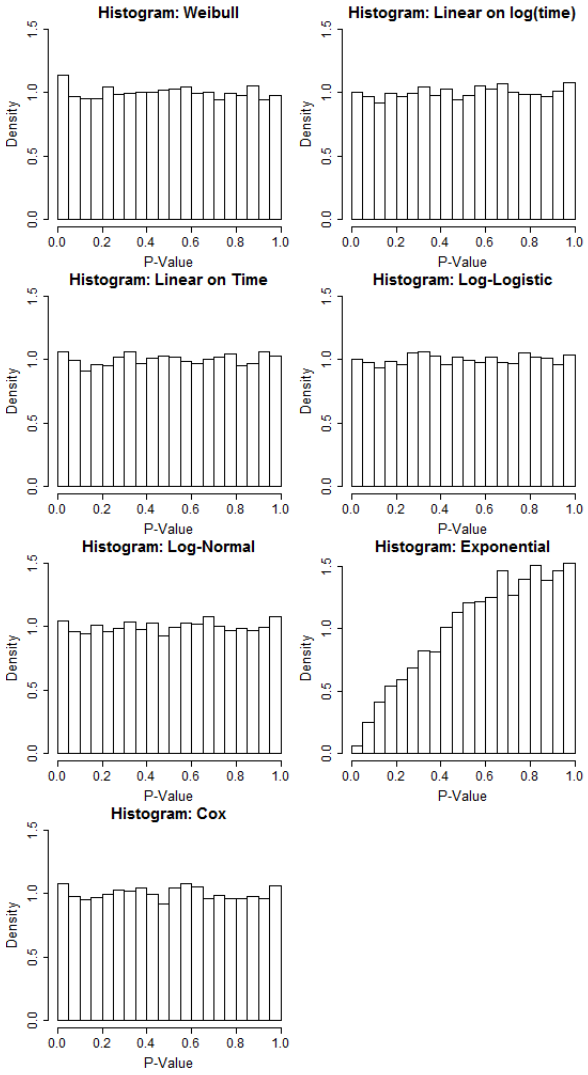

Figure 3.6: Histograms for each model under the scenario of a Normal covariate with $\beta = 0$, sample size = 250 and Weibull shape = 1.5.

Consistent with results published in Murdoch et al. (2008) which explored the one-sample t test for μ with Normal data, the distributions of P-values under the alternative hypothesis were not uniformly distributed [85]. This result is shown in Figure 3.7 for the case of a small sample size, a Weibull shape parameter 0.5 and $\beta = 0.5$. Findings were similar across other scenarios but the distributions start to show increasing right skewness as $|\beta|$ increases and as the Weibull

Shape parameter increases. Another notable finding in Figure 3.7 is that the histogram for the linear model on time is shaped differently than the others, as it appears almost bell-shaped with less density for small P-values. This corresponds to the previous finding for this linear model, where positive coefficients obtained lower power than negative coefficients.

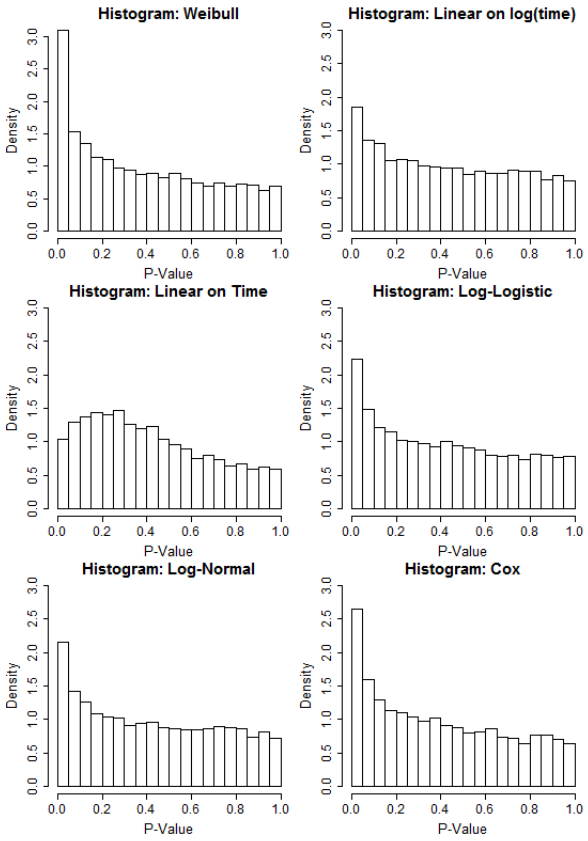

Figure 3.7: Histograms for each model under the scenario of a Binomial covariate with $\beta = 0.5$, sample size = 50 and Weibull shape = 0.5.

Next, inspection of the confidence intervals revealed that all methods produced a very small proportion of intervals that incorrectly fell entirely above or below zero, but the exponential model bore the largest proportion. Table 3.3 displays these results.

Lastly, only a small proportion of simulated datasets produced convergence failures. Specifically, most of the convergence failures occurred for the Weibull model under the scenario of a normal covariate and large values of $|\beta|$.

The differences that I have identified among the seven models in Study A are not substan-

Table 3.3: Proportion of confidence intervals incorrectly entirely above or below zero, suggesting the opposite direction of association, in Study A.

Model	**Proportion**
Weibull	0.003
Linear on time	0.003
Linear on log(time)	0.002
Exponential	0.018
Log-Logisitc	0.003
Log-Normal	0.003
Cox	0.002

tial. That is, the misspecified models have performed reasonably similar to the correct Weibull model. Specifically, they experienced a small reduction in power and a departure of the actual Type I error rate from the nominal. Although the inappropriate models were expected to perform worse than they did, the results were not too surprising since completely observed data were generated. Next, I describe the results from fitting misspecified models to data that were not as well behaved.

3.3.2 Study B - Comparison of Models with Censoring

Random Censoring

There was more disparity between the inappropriate models applied in Study B than there was between the models applied in Study A. Unlike Study A where the median *P*-values among the models with the Normal covariate were almost indistinguishable (see Figure 3.2), results were quite distinctive across the various scenarios in Study B. This suggests that inappropriate modeling in the situation of censored data has a greater impact on the behaviour of *P*-values. As well, the Weibull shape parameter continued to have a large impact on the distribution of median *P*-values. Figure 3.8 illustrates these findings for the scenario of a normal covariate

and Weibull shape equal to 0.5. In this plot, median P-values are highest for the linear model regressed on the logarithm of completely observed data and the linear model regressed on untransformed, completely observed data. P-values were the lowest for the correct Weibull model and the Weibull model applied to midpoint times. Again, median P-values motivate power and Type I error rate interpretations, where lower median P-values suggested higher power and Type I error rates.

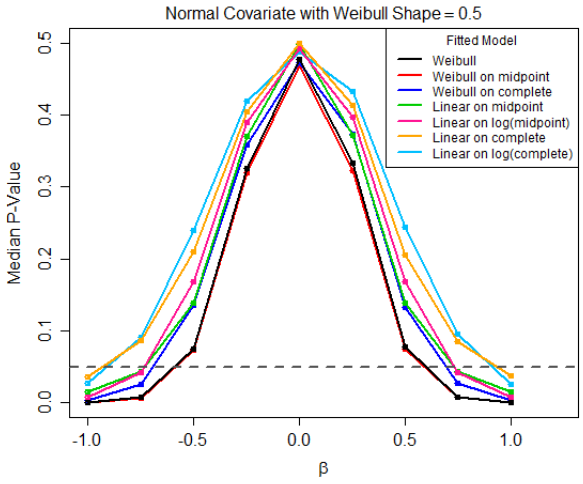

Figure 3.8: Random censoring: Median P-values plotted against β for the case of a Normal covariate and Weibull shape=0.5.

The influence of the Weibull shape parameter on the performance of P-values is explored in Figure 3.9, where the proportions of P-values less than α are plotted against β for each model. As anticipated from examining Figure 3.8, the correct Weibull model and the Weibull model regressed on midpoint times performed very similarly and had the highest power relative to the other models. Conversely, the two linear models which utilized only completely observed times consistently had the lowest power. As well, it was found that the linear model regressed on midpoint times, the linear model regressed on the logarithm of midpoint times and the Weibull model regressed on completely observed times behaved similarly, but their relative performance changed depending on the Weibull shape parameter. Specifically, the linear model regressed on midpoint times tended to have the highest power of the three models when the

Weibull shape parameter was 2.

In terms of Type I error, the Weibull model regressed on uncensored times typically had the highest Type I error rate, whereas the linear models regressed on completely observed times typically had the lowest Type I error rate. Findings were similar across other scenarios, although power was generally higher for the case of the normal covariate.

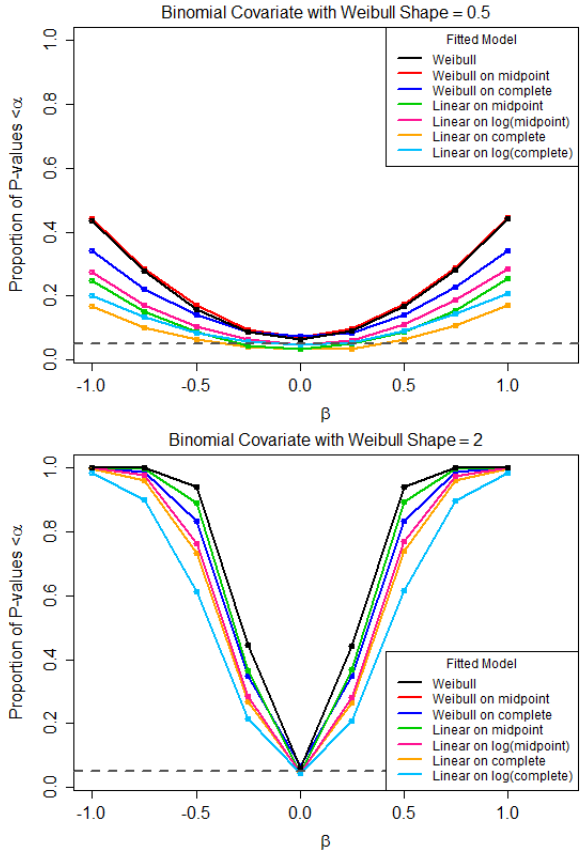

Figure 3.9: Random censoring: Proportion of P-values < 0.05 plotted against β for the case of a Binomial covariate and Weibull shapes equal to 0.5 and 2.

Next, Figure 3.10 shows the proportions of P-values less than α under the scenario of the normal covariate, grouped by censoring probability. The separation between the model performances was more apparent when the probability of censoring was higher, making the patterns in performance that were observed in Figure 3.9 more distinct in Figure 3.10. It was clear that the linear model regressed on complete data performed similarly to the linear model

regressed on the logarithm of complete data. As well, the linear model regressed on midpoint times, the linear model regressed on the logarithm of midpoint times, and the Weibull model regressed on completely observed times obtained similar proportions. Lastly, the performances of the correct Weibull model and the Weibull model regressed on midpoint times were very much alike.

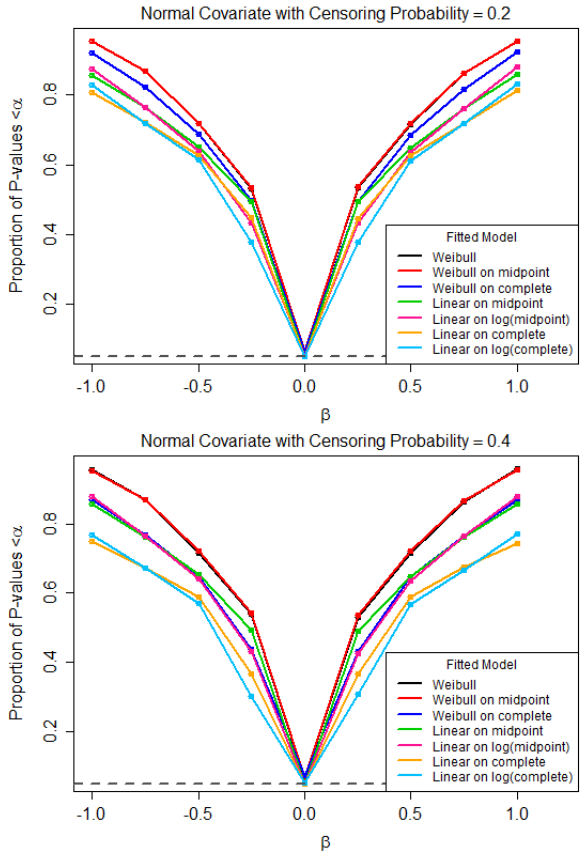

Figure 3.10: Random censoring: Proportion of P-values < 0.05 plotted against β for the case of a Normal covariate and censoring probabilities equal to 0.2 and 0.4.

The distribution of P-values appeared approximately uniform under the null hypothesis, with some exceptions. For the specific scenario of a Bernoulli covariate and a Weibull shape parameter of 0.5 displayed in Figure 3.11, the linear model regressed on midpoint times and the linear model regressed on completely observed times were not uniformly distributed. They were asymmetrical with higher densities in the lower P-value range. Therefore, inference on

β was not valid in this case. Further examination of histograms have shown that the approximation to uniform under the null hypothesis improved for all models when the Weibull shape parameter was 2. Additionally, the probability of censoring did not appear to have a substantial effect on the distribution of P-values under the null or alternative hypothesis.

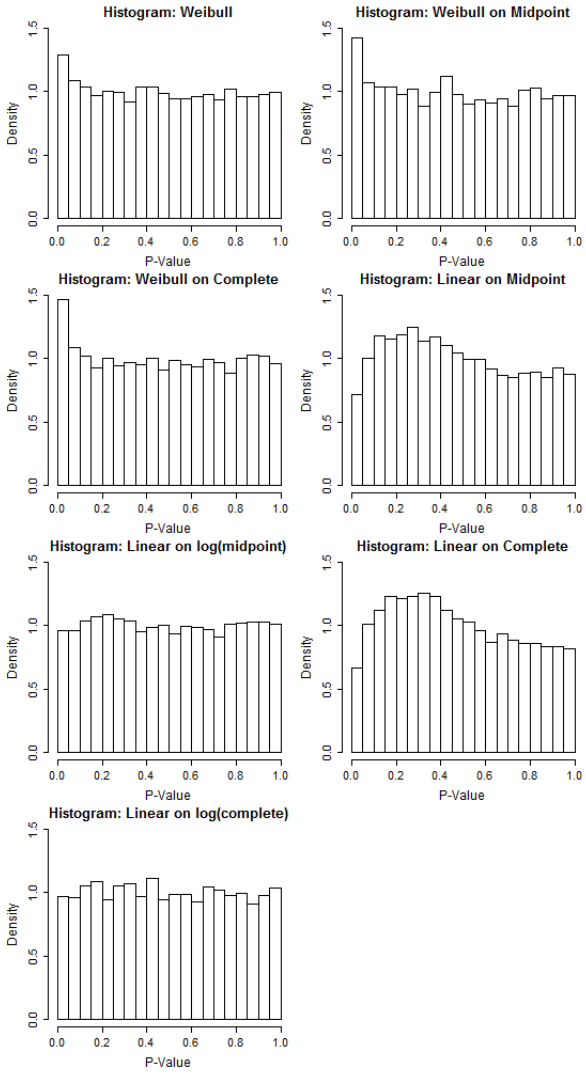

Figure 3.11: Random censoring: Histograms for each model under the scenario of a Binomial covariate with $\beta = 0$ and Weibull shape=0.5.

Next, it was observed that the scenario with the Bernoulli covariate produced a higher proportion of confidence intervals for β that incorrectly fell entirely above or below zero than the scenario with the normal covariate. These results are displayed in Table 3.4.

Lastly, it was found that only a negligible number of datasets resulted in convergence failures.

Table 3.4: Proportion of confidence intervals incorrectly entirely above or below zero, suggesting the opposite direction of association, under random censoring in Study B.

Model	**Normal Proportion**	**Binomial Proportion**
Weibull	0.001	0.004
Weibull on complete	0.002	0.007
Weibull on midpoint	0.001	0.005
Linear on midpoint	<0.001	0.003
Linear on log(midpoint)	0.001	0.005
Linear on complete	0.001	0.005
Linear on log(complete)	0.002	0.007

Informative Censoring

The main observation gained from the incorporation of informative censoring rather than random censoring was that the performances of the various models differed from each other to a greater extent. Figure 3.12 provides an example of this for the scenario of a normal covariate and a Weibull shape of 0.5. The previously observed patterns were unmistakable in the presence of informative censoring. It is clear the the Weibull model regressed on midpoint times continued to perform the closest to the correct Weibull model. It was also evident that the two linear models that only utilized completely observed times consistently obtained the lowest proportions of P-values less than α. Another important observation from Figure 3.12 is that the correct Weibull model and the Weibull model regressed on midpoint times were the most resistant to changes in the censoring mechanism, whereas the performances of the remaining five models fluctuated with changes in censoring probability or changes in censoring type.

Relative to the Bernoulli covariate, the normal covariate tended to have higher power across all values of $\beta \neq 0$. This was also a finding for random censoring, but it was more apparent

in the case of informative censoring. As well, power tended to be higher when the shape was equal to 2, rather than 0.5.

Regarding Type I error, the Weibull shape parameter of 0.5 produced greater variation in the Type I error rates among the models than the Weibull shape parameter of 2. The three Weibull survival models typically had higher Type I error rates than the four linear models. These observed patterns were similar to what was found in the case of random censoring.

Notably, inference on β was valid for all models and both values of the Weibull shape parameter, since the distributions of P-values appeared approximately uniform under the null hypothesis.

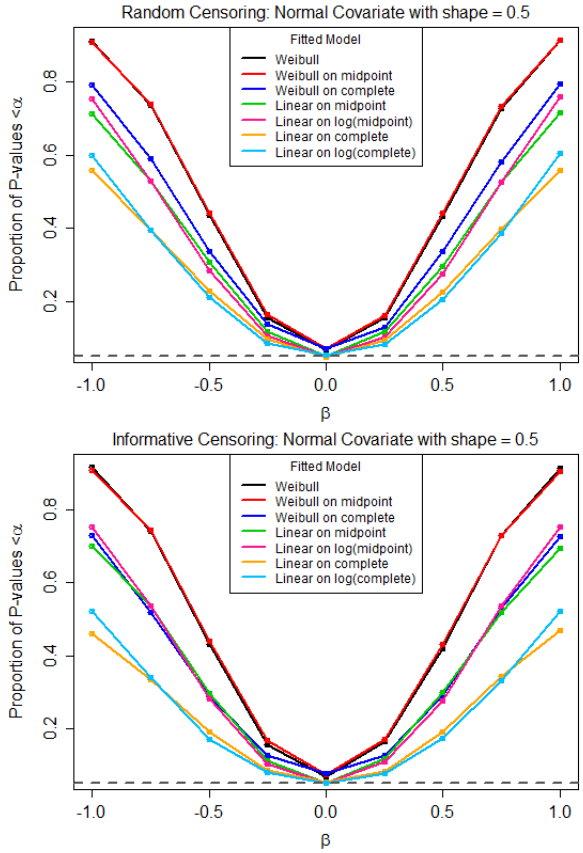

Figure 3.12: Comparison of random and informative censoring: Proportion of P-values < 0.05 plotted against β for the case of a Normal covariate with Weibull shape=0.5.

As with random censoring, the scenario with the Bernoulli covariate produced a higher

proportion of confidence intervals for β that incorrectly fell entirely above or below zero than the case of the Normal covariate. These results are displayed in Table 3.5. As well, only a negligible number of datasets resulted in convergence failures.

Table 3.5: Proportion of confidence intervals incorrectly entirely above or below zero, suggesting the opposite direction of association, under informative censoring in Study B.

Model	**Normal Proportion**	**Binomial Proportion**
Weibull	<0.001	0.004
Weibull on complete	0.002	0.011
Weibull on midpoint	0.001	0.006
Linear on midpoint	0.001	0.004
Linear on log(midpoint)	0.001	0.005
Linear on complete	0.002	0.005
Linear on log(complete)	0.003	0.008

3.4 Discussion

Based on the simulation studies, I conclude that the regression approaches investigation in Study A are acceptable when the Weibull time-to-event data are not censored. That is, inferences on β through simple linear regression t tests are reasonably adequate, but not as powerful as the correct model. However, one must be cautious when inappropriately specifying an exponential model to Weibull data. Since the exponential model is a special case of the Weibull model, I recommend applying the Weibull model instead.

In consideration of Study B, I conclude that inappropriate models and techniques for handling censored, Weibull data are also acceptable, in some cases, as they result in valid tests on β. However, the loss in power is large relative to the correct model. The loss of power is an important consequence since authors experience strong incentives to publish statistically significant results [55].

An earlier study also investigated the application of simple linear regression analyses to the midpoint of interval censored responses [114]. In contrast to the results in this simulation, the earlier study found linear regression analyses to obtain reasonable power [114]. However, the errors were simulated to follow a Gaussian distribution and little violation of the distributional assumption were investigated.

3.5 Limitations and Future Directions

One surprising result was that the Weibull model regressed on midpoint values performed very closely to the correct Weibull model. This may have occurred because the use of two independent Exp(1) random variables to create the interval censored times may have resulted in narrow intervals. I expect that differences between these two models would be more apparent if the widths of the intervals were larger. Even though the presence of censoring demonstrated an effect on the power of inference on β, perhaps the impact of censoring is more substantial when the probability of censoring is larger. Since the level of censoring was limited by the small sample size, a next strategy may be to simulate larger samples, with larger probabilities of censoring. As well, it would be interesting to increase the degree of dependence within the informative censoring mechanism to further study its impact on the performance of P-values.

Chapter 4

Conclusion

I have provided an extensive review of the flawed use of statistics in published medical research, showing that the existence of statistical errors in medical literature is widespread and not new. Statistical errors are common within many areas of research and reporting. The continued prevalence of the issue is concerning. Research informs the public, research informs subsequent research and research informs medical practice [84, 115]. This transmission of knowledge is damaged when research is published containing faulty or misleading conclusions, since readers of medical journals are generally uncritical towards authors' statistical analyses [17]. Along with the impact on literature and the research process, there are significant ethical implications and implications for science as a whole [3, 6].

I have investigated the impacts of the statistical errors identified in the literature review through application of erroneous analyses to a dataset with medical relevance. The investigation effectively demonstrated that conflicting conclusions can be reached depending on whether or not appropriate analyses are applied.

Furthermore, I have implemented a simulation study to investigate the impact of statistical errors in the framework of regression modeling. Although, the results from the simulation study have shown that inappropriate regression modeling may still be acceptable, tests on β will suffer a considerable loss of power. This is a destructive consequence because researchers face strong incentives to publish statistically significant results and a loss of power makes significant results less attainable [55]. Moreover, low power can encourage deceptive practices that lead to more statistical errors in published medical literature (e.g., *P*-hacking) [55].

Steps have been taken to improve the use of statistics in journals. For example, an experiment has shown that the inclusion of a statistical reviewer in the editorial process improves the quality of manuscripts for the *Medicina Clínica* biomedical journal [28]. Not all articles reviewed in Chapter 1 were recent and errors in published research may have changed over time, but it was not the goal of this thesis to assess these changes. Although there have been improvements in statistical practice, problems remain. Together, the consequences of statistical errors and the transmission of knowledge aggravate the issue of erroneous statistical analyses in published medical literature. A solution to this perpetuating problem may be found through comprehension and recognition of its impact.

References

[1] Aaserud, S. Residuals and Functional Form in Accelerated Life Regression Models. Master's thesis, Norwegian University of Science and Technology, 2011.

[2] Abraham, B., and Ledolter, J. *Introduction to Regression Modeling*. Thompson, Brooks/Cole, Belmont, CA, 2006.

[3] Altman, D. Statistics and ethics in medical research. VIII. Improving the quality of statistics in medical journals. *British Medical Journal 282*, 6257 (1981), 44–47.

[4] Altman, D. Statistics in medical journals. *Statistics in Medicine 1*, 1 (1982), 59–71.

[5] Altman, D. Statistics in medical journals: Developments in the 1980s. *Statistics in Medicine 10*, 12 (1991), 1897–1913.

[6] Altman, D. The scandal of poor medical-research. *British Medical Journal 308*, 6924 (1994), 283–284.

[7] Altman, D. Statistical reviewing for medical journals. *Statistics in Medicine 17*, 23 (1998), 2661–2674.

[8] Altman, D. Statistics in medical journals: Some recent trends. *Statistics in Medicine 19*, 23 (2000), 3275–3289.

[9] Altman, D., De Stavola, B., Love, S., and Stepniewska, K. Review of survival analyses published in cancer journals. *British Journal of Cancer 72*, 2 (1995), 511–518.

[10] Altman, D., Gore, S., Gardner, M., and Pocock, S. Statistical guidelines for contributors to medical journals. *British Medical Journal (Clinical research ed.) 286*, 6376 (1983), 1489–1493.

[11] American Society for Microbiology. Infection and Immunity: 2011 Instructions to Authors. *Infection and Immunity 79*, 1 (2011), 1–20.

[12] American Statistical Association. Ethical Guidelines for Statistical Practice, 2015 (accessed February 2, 2015). `http://www.amstat.org/about/ethicalguidelines.cfm`.

[13] Avram, M., Shanks, C., Dykes, M., Ronai, A., and Stiers, W. Statistical methods in anesthesia articles: An evaluation of two American journals during two six-month periods. *Anesthesia and Analgesia 64*, 6 (1985), 607–611.

[14] Bagley, S., White, H., and Golomb, B. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology 54*, 10 (2001), 979–985.

[15] Bakker, M., and Wicherts, J. The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods 43*, 3 (2011), 666–678. cited By 46.

[16] Baldi, B., and Moore, D. S. *The Practice of Statistics in the Life Sciences*, third ed. W. H. Freeman, New York, 2013.

[17] Begg, C. B., and Berlin, J. A. Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society.Series A (Statistics in Society) 151*, 3 (1988), 419–463.

[18] Bell, M., Olivier, J., and King, M. Scientific rigour in psycho-oncology trials: Why and how to avoid common statistical errors. *Psycho-Oncology 22*, 3 (2013), 499–505.

[19] Blair, R. C., and Taylor, R. A. *Biostatistics for the Health Sciences*. Pearson Prentice Hall, Upper Saddle River, N.J, 2008.

[20] Boiko, A., Vorobeychik, G., Paty, D., Devonshire, V., Sadovnick, D., and of British Columbia MS Clinic Neurologists, U. Early onset multiple sclerosis: A longitudinal study. *Neurology 59*, 7 (2002), 1006–1010.

[21] Boneau, C. A. The effects of violations of assumptions underlying the t test. *Psychological Bulletin 57*, 1 (1960), 49–64.

[22] Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., Lijmer, J., Moher, D., Rennie, D., and De Vet, H. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clinical Chemistry 49*, 1 (2003), 1–6.

[23] Braarud, I. H. Analysis of Life Regression Models for Censored Data using Pseudo Observations. Master's thesis, Norwegian University of Science and Technology, 2013.

[24] Bradburn, M., Clark, T., and Altman, D. Survival Analysis Part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. *British Journal of Cancer 89*, 4 (2003), 605–611.

[25] Bradley, D. R. Type I Error Rate of the Chi-Square Test of Independence in "R x C" Tables That Have Small Expected Frequencies. *Psychological Bulletin 86*, 6 (1979), 1290–1297.

[26] Brilhante, M. F., Brilhante, M. F., Pestana, D., Pestana, D., Sequeira, F., and Sequeira, F. Combining p-values and random p-values. In *Information Technology Interfaces, 2010 32nd International Conferenceon Information Technology Interfaces* (2010), IEEE, pp. 515–520.

[27] Burton, A., and Altman, D. Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer 91*, 1 (2004), 4–8.

[28] Cobo, E., Selva-O'Callagham, A., Ribera, J.-M., Cardellach, F., Dominguez, R., and Vilardell, M. Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PloS one 2*, 3 (2007), e332.

[29] Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics 3*, 1 (1947), 22–38.

[30] Compston, A., and Coles, A. Multiple sclerosis. *Lancet (London, England) 372*, 9648 (2008), 1502–1517.

[31] Connor, J. The value of a p-valueless paper. *American Journal of Gastroenterology 99*, 9 (2004), 1638–1640.

[32] Cooper, R., Schriger, D., and Close, R. Graphical literacy: The quality of graphs in a large-circulation journal. *Annals of Emergency Medicine 40*, 3 (2002), 317–322.

[33] Cossburn, M., Ingram, G., Hirst, C., Ben-Shlomo, Y., Pickersgill, T., and Robertson, N. Age at onset as a determinant of presenting phenotype and initial relapse recovery in multiple sclerosis. *Multiple Sclerosis Journal 18*, 1 (2012), 45–54.

[34] D'Agostino, R. B., Sullivan, L. M., and Beiser, A. S. *Introductory Applied Biostatistics*. Thompson, Brooks/Cole, United States; Australia, 2006.

[35] Davidoff, F., Godlee, F., Hoey, J., Glass, R., Overbeke, J., Utiger, R., Nicholls, M., Horton, R., Nylenna, M., Hojgaard, L., and Kotzin, S. Uniform requirements for manuscripts submitted to biomedical journals. *The Journal of the American Osteopathic Association 103*, 3 (2003), 137–149.

[36] Del Priore, G., Zandieh, P., and Lee, M.-J. Treatment of continuous data as categoric variables in Obstetrics and Gynecology. *Obstetrics and Gynecology 89*, 3 (1997), 351–354.

[37] Evans, M. Presentation of manuscripts for publication in the British Journal of Surgery. *British Journal of Surgery 76*, 12 (1989), 1311–1314.

[38] Faraone, S. V. Chi-square in small samples. *American Psychologist 37*, 1 (1982), 107–107.

[39] Fernandes-Taylor, S., Hyun, J., Reeder, R., and Harris, A. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Research Notes 4* (2011).

[40] Fernndez, O., Fernndez, V., Arbizu, T., Izquierdo, G., Bosca, I., Arroyo, R., Garca Merino, J. A., de Ramn, E., Group, N., and Group, T. N. Characteristics of multiple sclerosis at onset and delay of diagnosis and treatment in spain (the novo study). *Journal of neurology 257*, 9 (2010), 1500–1507.

[41] Fidler, F., Thomason, N., Cumming, G., Finch, S., and Leeman, J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychological Science 15*, 2 (2004), 119–126.

[42] Fielding, S., Maclennan, G., Cook, J., and Ramsay, C. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials 9* (2008).

[43] Garca-Berthou, E., and Alcaraz, C. Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology 4* (2004).

[44] Gardner, M., Altman, D., Jones, D., and Machin, D. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? *British Medical Journal (Clinical research ed.) 286*, 6376 (1983), 1485–1488.

[45] Glass, G. V., Peckham, P. D., and Sanders, J. R. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research 42*, 3 (1972), 237–288.

[46] Goodman, S., Altman, D., and George, S. Statistical reviewing policies of medical journals caveat lector? *Journal of General Internal Medicine 13*, 11 (1998), 753–756.

[47] Gore, S., Jones, I., and Rytter, E. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *British Medical Journal 1*, 6053 (1977), 85–87.

[48] Grissom, R. Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology 68*, 1 (2000), 155–165.

[49] Halsey, L., Curran-Everett, D., Vowler, S., and Drummond, G. The fickle P value generates irreproducible results. *Nature Methods 12*, 3 (2015), 179–185.

[50] Harbo, H. F., Gold, R., and Tintor, M. Sex and gender issues in multiple sclerosis. *Therapeutic Advances in Neurological Disorders 6*, 4 (2013), 237–248.

[51] Harris, A., Reeder, R., and Hyun, J. Common statistical and research design problems in manuscripts submitted to high-impact psychiatry journals: What editors and reviewers want authors to know. *Journal of Psychiatric Research 43*, 15 (2009), 1231–1234.

[52] Harris, A., Reeder, R., and Hyun, J. Common statistical and research design problems in manuscripts Submitted to High-Impact Public Health Journals. *The Open Public Health Journal 2*, 1 (2009), 44–48.

[53] Harris, A., Reeder, R., and Hyun, J. Survey of editors and reviewers of high-impact psychology journals: Statistical and research design problems in submitted manuscripts. *Journal of Psychology: Interdisciplinary and Applied 145*, 3 (2011), 195–209.

[54] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. Summarizing Monte Carlo Results in Methodological Research: The One- and Two-factor fixed effects ANOVA cases. *Journal of Educational Statistics 17*, 4 (1992), 315–339.

[55] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS biology 13*, 3 (2015), e1002106.

[56] Hess, K. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine 14*, 15 (1995), 1707–1723.

[57] Hoekstra, R., Kiers, H., and Johnson, A. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology 3*, MAY (2012).

[58] Hoffman, J. I. The incorrect use of Chi-square analysis for paired data. *Clinical and Experimental Immunology 24*, 1 (1976), 227–229.

[59] Holmes, T. Ten categories of statistical errors: A guide for research in endocrinology and metabolism. *American Journal of Physiology - Endocrinology and Metabolism 286*, 4 49-4 (2004), E495–E501.

[60] Hopewell, S., Dutton, S., Yu, L., Chan, A., and Altman, D. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in pubmed. *British Medical Journal 340* (2010), c723.

[61] Horst, R. *The Weibull Distriubtion: A Handbook*. CRC Press, Boca Raton, FL, 2008.

[62] Hsieh, F. Y. A cautionary note on the analysis of extreme data with cox regression. *The American Statistician 49*, 2 (1995), 226–228.

[63] Jin, Z., Yu, D., Zhang, L., Meng, H., Lu, J., Gao, Q., Cao, Y., Ma, X., Wu, C., He, Q., Wang, R., and He, J. A retrospective survey of research design and statistical analyses in selected chinese medical journals in 1998 and 2008. *PLoS ONE 5*, 5 (2010).

[64] Kim, J., Kim, D., and Hong, S. Assessment of errors and misused statistics in dental research. *International Dental Journal 61*, 3 (2011), 163–167.

[65] Kleinbaum, D. G., and Klein, M. *Survival Analysis: A Self-Learning Text*. Springer Scince and Business Media, New York, NY, 2005.

[66] Krzywinski, M., and Altman, N. Points of significance: Significance, P values and t-tests. *Nature Methods 506*, 11 (2013), 1041–1042.

[67] Kurichi, J., and Sonnad, S. Statistical Methods in the Surgical Literature. *Journal of the American College of Surgeons 202*, 3 (2006), 476–484.

[68] Lang, J. M., Rothman, K. J., and Cann, C. I. That confounded p-value. *Epidemiology 9*, 1 (1998), 7–8.

[69] Lang, T. Twenty statistical errors even YOU can find in biomedical research articles. *Croatian Medical Journal 45*, 4 (2004), 361–370.

[70] Lang, T. A., and Altman, D. G. Basic statistical reporting for articles published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *International Journal of Nursing Studies 52*, 1 (JAN 2015), 5–9.

[71] Lee, J. Demystify statistical significance-time to move on from the p value to Bayesian analysis. *Journal of the National Cancer Institute 103*, 1 (2011), 2–3.

[72] Lewis, D., and Burke, C. J. The use and misuse of the chi-square test. *Psychological Bulletin 46*, 6 (1949), 433–489.

[73] Lucena, C., Lpez, J., Abalos, C., Robles, V., and Pulgar, R. Statistical errors in microleakage studies in operative dentistry. A survey of the literature 2001-2009. *European Journal of Oral Sciences 119*, 6 (2011), 504–510.

[74] MacCallum, R., Zhang, S., Preacher, K., and Rucker, D. On the practice of dichotomization of quantitative variables. *Psychological Methods 7*, 1 (2002), 19–40.

[75] Mallett, S., Royston, P., Dutton, S., Waters, R., and Altman, D. Reporting methods in studies developing prognostic models in cancer: A review. *BMC Medicine 8* (2010).

[76] Marshall, S. Testing with confidence: The use (and misuse) of confidence intervals in biomedical research. *Journal of Science and Medicine in Sport 7*, 2 (2004), 135–137.

[77] Matthews, D. E. Stat 935: Analyzing time-to-event data, Winter 2002.

[78] McHugh, M. The Chi-square test of independence. *Biochemia Medica 23*, 2 (2013), 143–149.

[79] McShane, L., Altman, D., Sauerbrei, W., Taube, S., Gion, M., Clark, G., Costa, J., Di Leo, A., and Mayer, R. Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute 97*, 16 (2005), 1180–1184.

[80] Merck Serono. LHSC iMed 6.1 Database, 2012. Geneva, S.A.

[81] Micceri, T. The Unicorn,The Normal Curve, and Other Improbable Creatures. *Psychological Bulletin 105*, 1 (1989), 156–166.

[82] Mills, M. *Introducing Survival and Event History Analysis*. SAGE Publications Ltd, Thousand Oaks, CA, 2011.

[83] Moher, D., Hopewell, S., Schulz, K., Montori, V., Gtzsche, P., Devereaux, P., Elbourne, D., Egger, M., and Altman, D. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery 10*, 1 (2012), 28–55.

[84] Moses III, H., and Martin, J. Biomedical research and health advances. *New England Journal of Medicine 364*, 6 (2011), 567–571.

[85] Murdoch, D. J., Tsai, Y.-L., and Adcock, J. P-values are random variables. *The American Statistician 62*, 3 (2008), 242–245.

[86] Murphy, J. Statistical errors in immunologic research. *Journal of Allergy and Clinical Immunology 114*, 6 (2004), 1259–1263.

[87] Murray, G. The task of a statistical referee. *British Journal of Surgery 75*, 7 (1988), 664–667.

[88] Nagele, P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia 90*, 4 (2003), 514–516.

[89] Neville, J., Lang, W., and Fleischer Jr., A. Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology From January Through December 2003. *Archives of Dermatology 142*, 6 (2006), 737–740.

[90] Nieuwenhuis, S., Forstmann, B., and Wagenmakers, E.-J. Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience 14*, 9 (2011), 1105–1107.

[91] Nuzzo, R. Scientific method: Statistical errors. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature 506*, 7487 (2014), 150–152.

[92] Olsen, C. Review of the use of statistics in Infection and Immunity. *Infection and Immunity 71*, 12 (2003), 6689–6692.

[93] Panagiotakos, D. The value of p-value in Biomedical Research. *The Open Cardiovascular Medicine Journal 2* (2008), 97–99.

[94] Pilk, T. Statistics in three biomedical journals. *Physiological Research 52*, 1 (2003), 39–43.

[95] Pocock, S., Collier, T., Dandreo, K., De Stavola, B., Goldman, M., Kalish, L., Kasten, L., and McCormack, V. Issues in the reporting of epidemiological studies: A survey of recent practice. *British Medical Journal 329*, 7471 (2004), 883–887.

[96] Porter, A. Misuse of correlation and regression in three medical journals. *Journal of the Royal Society of Medicine 92*, 3 (1999), 123–128.

[97] Prescott, R., and Civil, I. Lies, damn lies and statistics: Errors and omission in papers submitted to INJURY 2010-2012. *Injury 44*, 1 (2013), 6–11.

[98] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[99] Royston, P., Altman, D., and Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine 25*, 1 (2006), 127–141.

[100] Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S., Carpenter, J., and Initiative, S. STRengthening Analytical Thinking for Observational studies: the STRATOS initiative. *Statistics in Medicine 33*, 30, SI (2014), 5413–5432.

[101] Scales Jr., C., Norris, R., Peterson, B., Preminger, G., and Dahm, P. Clinical research and statistical methods in the urology literature. *Journal of Urology 174*, 4 (2005), 1374–1379.

[102] Schatz, P., Jay, K., McComb, J., and McLaughlin, J. Misuse of statistical tests in Archives of Clinical Neuropsychology publications. *Archives of Clinical Neuropsychology 20*, 8 (2005), 1053–1059.

[103] Schenker, N., and Gentleman, J. On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician 55*, 3 (2001), 182–186.

[104] Schor, S., and Karten, I. Statistical evaluation of medical journal manuscripts. *Journal of the American Medical Association 195*, 13 (1966), 1123–1128.

[105] Scott, M., Flaherty, D., and Currall, J. Statistics: Dealing with categorical data. *Journal of Small Animal Practice 54*, 1 (2013), 3–8.

[106] Selman, T., Morris, R., Zamora, J., and Khan, K. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: Application of the STARD criteria. *BMC Women's Health 11* (2011).

[107] Smidt, N., Rutjes, A., Van Der Windt, D., Ostelo, R., Reitsma, J., Bossuyt, P., Bouter, L., and De Vet, H. Quality of reporting of diagnostic accuracy studies. *Radiology 235*, 2 (2005), 347–353.

[108] Stoltzfus, J. Logistic regression: A brief primer. *Academic Emergency Medicine 18*, 10 (2011), 1099–1104.

[109] Strasak, A., Zaman, Q., Pfeiffer, K., Gbel, G., and Ulmer, H. Statistical errors in medical research - A review of common pitfalls. *Swiss Medical Weekly 137*, 3-4 (2007), 44–49.

[110] Streiner, D. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. *Canadian Journal of Psychiatry 47*, 3 (2002), 262–266.

[111] Sullivan, G., and Feinn, R. Using Effect Size - or Why the P Value Is Not Enough. *Journal of Graduate Medical Education 4*, 3 (2012), 279–282.

[112] Terry M. Therneau, and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[113] Therneau, T. M. *A Package for Survival Analysis in S*, 2015. version 2.38.

[114] Thompson, M. L., and Nelson, K. P. Linear regression with type i interval- and left-censored response data. *Environmental and Ecological Statistics 10*, 2 (2003), 221–230.

[115] Toledo-Pereyra, L. H. Importance of medical and surgical research. *Journal of Investigative Surgery 22*, 5 (2009), 325–326.

[116] Van Walraven, C., Davis, D., Forster, A., and Wells, G. Time-dependent bias was common in survival analyses published in leading clinical journals. *Journal of Clinical Epidemiology 57*, 7 (2004), 672–682.

[117] Van Walraven, C., and Hart, R. Leave 'em alone - Why continuous variables should be analyzed as such. *Neuroepidemiology 30*, 3 (2008), 138–139.

[118] von Elm, E., Altman, D., Egger, M., Pocock, S., Gtzsche, P., and Vandenbroucke, J. The strengthening the reporting of observational studies in epidemiology (STROBE)

statement: Guidelines for reporting observational studies. *International Journal of Surgery 12*, 12 (2014), 1495–1499.

[119] White, S. Statistical errors in papers in the British Journal of Psychiatry. *British Journal of Psychiatry 135*, 10 (1979), 336–342.

[120] Wood, A., White, I., and Thompson, S. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials 1*, 4 (2004), 368–376.

[121] Woolston, C. Psychology journal bans P values. *Nature 519*, 9 (2015).

[122] Zhao, L. P., and Kolonel, L. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology 136*, 4 (1992), 464–474.

[123] Zinsmeister, A., and Connor, J. Ten common statistical errors and how to avoid them. *American Journal of Gastroenterology 103*, 2 (2008), 262–266.

Curriculum Vitae

Name:	Britney Allen
Post-Secondary Education and Degrees:	University of Western Ontario London, ON 2013 - 2015 M.Sc. Biostatistics
	University of Western Ontario London, ON 2008 - 2013 B.Sc. Statistics, Economics (Honors)
Honours and Awards:	Faculty of Science Graduate Student Teaching Award 2015
	The University of Western Ontario Gold Medal 2013
	The Northern Life Gold Medal 2013
Related Work Experience:	Graduate Teaching Assistant Department of Statistics and Actuarial Science The University of Western Ontario 2013 - 2015
	STAT 1024 Developer The University of Western Ontario London, ON Summer 2014
	Research Assistant (Internship) Human Resources and Skills Development Canada/Canada Student Loans Program Gatineau, QC 2010 - 2011