ABSTRACT

Title of thesis: AN EM ALGORITHM FOR MIXED-TYPE MULTIPLE OUTCOME REGRESSIONS WITH APPLICATIONS TO A PROSTATE CANCER STUDY

JoAnn Rudd, Master of Arts, 2008

Thesis directed by: Professor Paul J. Smith
Department of Mathematics

We propose a joint model for binary and continuous responses using a latent variable for the binary response. The observed continuous response and the latent response are treated as correlated normals obeying a bivariate regression model. We develop an EM algorithm to find maximum likelihood estimates for the parameters. We perform the E–step analytically and use an iterative algorithm for the M–step.

The algorithm is applied to a prostate cancer clinical trial whose goal was to assess therapeutic effects of diethylstilbestrol (DES) in advanced cancer patients and to assess possible excess cardiovascular mortality. Therapeutic effects were measured as prostatic acid phosphatase (PAP) levels follow–up and whether the patient progressed to stage IV or died of cancer. The treatment reduced PAP levels but not the incidence of cancer mortality within a six–month time frame. Higher doses of DES were associated with increased risk of cardiovascular–related death.

AN EM ALGORITHM FOR MIXED TYPE MULTIPLE OUTCOME REGRESSIONS WITH APPLICATIONS TO A PROSTATE CANCER STUDY

by

JoAnn Rudd

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Arts
2008

Advisory Committee:
Professor Paul J. Smith, Chair/Advisor
Professor Benjamin Kedem
Professor Eric Slud

Acknowledgments

I would like to thank Dr. Paul Smith, my thesis advisor, for guiding me through the thesis process and for his expertise. I would like to thank Don Corle of NIH NCI Division of Cancer Prevention, Biometry Research Group for providing me with data from the Veteran Administration Cooperative Urological Research Group (VACURG) studies and answering my questions about the data.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Biomedical studies typically have several endpoints which may be somehow related. If these are all continuous, then multivariate regression is a natural way to estimate the parameters jointly. Similarly, several categorical endpoints can be jointly modeled using the multinomial distribution. If the endpoints are a mixture of continuous and binary, they can be modeled separately using linear regression for the continuous variables and probit or logistic regression for the binary ones. The EM (expectation-maximization) algorithm can be used to link these two models, thereby simultaneously estimating the regression parameters and the correlation between the responses. We develop an EM algorithm specifically for this problem using a latent normal variable model for the binary variable, and we apply it to a prostate cancer dataset.

Our data comes from one of several prostate cancer studies from the Veterans Administration Cooperative Urological Research Group (VACURG) in the 1960s and 1970s. The studies tested the effects of various hormone therapies on the cancer, and raised concerns about mortality due to cardiovascular disease associated with one of the treatments, which was a daily dose of diethylstilbestrol (DES).

In Chapter 2 we give background information on the VACURG study, including details about biomarkers that we used as variables in the analysis. We also describe

some of the methods of analyzing multiple endpoints of mixed type that have been presented in the literature, and we outline the expectation maximization algorithm and some of its extensions. In Chapter 3, we delineate our exploratory analyses of the VACURG data, culminating with univariate regression models. We describe our model and present our algorithm for estimating the maximum likelihood estimates in Chapter 4, where we also present results of a simulation to examine the performance of our algorithm. In Chapter 5, we apply the algorithm to the prostate data for two key models. The first jointly models the log of PAP at follow–up and cancer death or advance to stage IV cancer as the two response variables, and the second jointly models log of PAP with cardiovascular death. Chapter 6 provides conclusions and suggestions for future research.

Chapter 2

Literature Review

2.1 Overview

In this chapter we present background information on the VACURG study of prostate cancer, treatment of multiple mixed-type endpoints, and the EM algorithm.

2.2 The VACURG Study

The Veterans Administration Cooperative Urological Research Group (VACURG) was formed in 1960 and conducted three large-scale randomized prostate cancer treatment studies (Byar 1973). The Group recruited United States military veterans from 15 Veterans Administration Hospitals across the country. Each study had two main parts: one on stage I and II patients and the other on stage III and IV patients. The studies evaluated the effects of various hormone therapies and surgeries that affect hormone levels. For stage I and II patients, Study 1 compared radical prostatectomy and placebo to radical prostatectomy and 5 mg diethylstilbestrol (DES), which is a synthetic estrogen. Stage III and IV patients were randomized to receive a placebo, 5 mg DES daily, orchiectomy and placebo, or orchiectomy and DES. The first arm of Study 2, which began in 1967, on stage I and II patients, received either placebo or prostatectomy and placebo. In the sec-

ond arm, each patient was randomized to one of four treatment groups; one group received a placebo, and the other three each took one of the following daily doses of diethylstilbestrol (DES): 0.2 mg, 1 mg, or 5 mg. Study 3, which began in 1969, employed the same treatment groups for stage I and II patients as did Study 2. The stage III and IV patients received one of the following treatments: 1 mg DES, Premarin, or Provera.

Patients were excluded if they had already begun some kind of treatment, and the stage of their disease was determined by the physician before randomization. Patients were classified as stage III if they had a local metastasis, and as stage IV if they had either a distant metastasis or elevated levels of prostatic acid phosphatase (PAP). PAP is an enzyme that is produced by the prostate that is used in staging because 65% to 85% of patients with bone metastases have elevated levels of PAP (Byar, Corle, and Brown 1988). Sixty percent of the patients classified as stage IV in the VACURG studies were classified on the basis of elevated PAP levels in the absence of other evidence of metastases (Byar, Corle, and Brown 1988).

The patients were followed for periods of up to nine years and assessed at intervals by the physicians. During each follow-up visit many measurements were taken from the patients.

Here we focus on the part of Study 2 which involved 506 patients who had advanced to stages III and IV cancer. They were assessed at baseline and again at a six-month follow-up. The men ranged from 48 to 99 years of age, with a mean age of 71.5. At the physician visits, the physicians recorded the extent of any metastases the patient had, the degree to which the patient was experiencing pain, measures of

activity level, whether the patient's breast had enlarged or become tender as a result of the therapy, whether the patient had changed treatments, estimated surface area of the tumor, PAP, and Gleason histology, which was measured only at baseline. The Gleason score is a tumor grade based on the microscopic character of the prostate tissue.

Both at baseline and at the follow-up visit, the data collected included measurements on the cardiovascular system because of concerns that were raised during Study 1. Study 1 showed that patients on diethylstilbestrol did experience fewer deaths due to prostate cancer, but the benefit was lessened because of increased risk of death from diseases of the cardiovascular system (Bailar and Byar, 1970). These measurements included the patient's disease history, the blood pressure, heart rate, and results of an electrocardiogram.

Variables of prognostic value would include age, since prostate cancer is a disease that is associated with long-term exposure to the male hormones. Baseline metastases and PAP would both be good predictors of a patient's outcome because they are associated with staging.

2.2.1 Endpoints

Death is a primary endpoint. Patients were coded as alive if they were still alive at the end of the study, or the cause of death was classified as one of the following: cancer of the prostate, heart or vascular diseases, stroke, pulmonary embolus, other cancers, respiratory diseases, other specific non-cancer cause, unspecified non-cancer

cause, or unknown. One expects the treatment to lower the risk of dying from prostate cancer, but it turns out that the drug had toxic effects which increase the incidence of cardiovascular- related deaths. An optimal treatment is one that increases overall survival.

Aside from its value as a prognostic indicator, PAP can be used to measure a patient's tumor progression. Elevated levels of PAP are associated with progress of the tumor. PAP values within an individual normally vary to within $\pm 50\%$ of the mean for that individual (Byar, Corle, and Brown, 1988). For this reason, only changes greater than 50% from baseline are considered significant during treatment. These significant fluctuations in PAP levels reflect progress of the tumor or improvement, which can be attributed to the treatment the patient is undergoing.

Another indicator of cancer progression would be the development of metastasis.

The investigators in this study used intent-to-treat analysis, meaning that patients were analyzed in the groups to which they were randomized, whether or not they stayed on that particular treatment (Bailar and Byar, 1970). The patients' urologists were able to change the patient's treatment according to the patient's best interest. If the physician does decide to change the treatment or perform surgery, that would indicate that the patient is not doing well.

2.2.2 Summary of Findings

All three VACURG studies indicated that prostate cancer is responsive to hormones, and they provided information about the risks of cardiovascular incidents associated with estrogen therapies. These studies continue to be cited in discussions of the complex issues of when to start hormone therapy and which patients should be treated.

2.3 Regression Analysis of Multiple Mixed-Type Response Variables

A difficulty in modelling mixed-type outcomes is specifying their joint distribution. Several authors have considered the binary outcome as the manifestation of a latent continuous random variable. Catalano and Ryan (1992) and Fitzmaurice and Laird (1995) represented the joint distributions as products of a marginal and conditional component. The parameters of the joint model have been estimated using generalized estimating equations or stochastic EM algorithms, avoiding specifying the likelihoods in closed form.

Catalano and Ryan (1992) developed a joint model for a continuous endpoint, $\mathbf{Y}_1$, and a binary endpoint, $\mathbf{Y}_2^*$, by assuming that the binary event occurs only if a latent normal random variable, $\mathbf{Y}_2$, exceeds a certain level. The joint distribution of $\mathbf{Y}_1$ and $\mathbf{Y}_2^*$ was formed as the product of the marginal distribution of $\mathbf{Y}_1$ and the conditional distribution of $\mathbf{Y}_2^*|\mathbf{Y}_1$. Their model was complicated because they considered the case of correlation between individuals in clusters. They used quasi-likelihood to estimate the regression parameters.

Fitzmaurice and Laird (1995) also jointly considered a continuous and binary response variable, but their interest was in estimating the marginal regression parameters, and they were uninterested in estimating the correlation between the two responses. They also formed the joint distribution as a product of the marginal and conditional distributions, but they used the binary endpoint as the conditioning variable, in contrast to Catalano and Ryan's model. They also used quasi-likelihood to estimate the parameters.

Sammel, Ryan, and Legler (1997) modeled discrete and continuous outcomes using latent variables, but their model accommodates the general case of the continuous variable from any regular exponential family and the latent variable from any regular exponential family. The coefficients of the covariates can then be estimated using MCEM, assuming the distributions of the latent variable and observed continuous variable are known.

Geys et al. (2001) introduced two latent variable models for continuous and discrete outcomes. One models the underlying latent variable using a Plackett distribution. The other uses a normal distribution to model the underlying latent variable. They also use pseudolikelihood methods to estimate the parameters.

Gueorguieva and Agresti (2001) also used a latent variable approach, using a correlated probit model. They used Monte Carlo ECM to find the maximum likelihood estimates of the parameters.

2.4 The EM Algorithm

The Expectation Maximization (EM) algorithm is a procedure to compute maximum likelihood estimates in cases of incomplete data. It is most useful in situations in which, had the all the data been observed, the maximum likelihood estimates would be easy to compute. The EM algorithm is an iterative procedure in which each iteration consists of two steps: the expectation step, or E-step, and the maximization step, or M-step.

Suppose one observes some incomplete data, $\mathbf{X}$, and $\mathbf{X}^*$ is the complete, unobservable data which comes from a distribution $f(\mathbf{x};\boldsymbol{\theta})$, which depends on a parameter $\boldsymbol{\theta}$, and suppose that one is interested in finding the maximum likelihood estimate for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$. The E-step consists of computing the expectation of the complete-data log likelihood, conditional on the observed data and the current guess of the parameters, $\hat{\boldsymbol{\theta}}_k$:

$$E_{\hat{\boldsymbol{\theta}}_k}(\ln f(\mathbf{X};\boldsymbol{\theta})|\mathbf{X})$$

The M-step then maximizes the conditional expectation found in the E-step with respect to $\boldsymbol{\theta}$. The value that maximizes the conditional expectation then becomes the current guess, and the E-step begins again.

$$\hat{\boldsymbol{\theta}}_{k+1} = \underset{\theta}{\operatorname{argmax}}\, E_{\hat{\boldsymbol{\theta}}_k}(\ln f(\mathbf{X};\boldsymbol{\theta})|\mathbf{X})$$

The process repeats until the values converge.

The method was popularized by Dempster, Laird and Rubin (1977), who formally defined the algorithm and provided various examples of applications, demonstrating its versatility. They showed that each successive iteration of the algorithm

increases the likelihood. Wu (1983) proved important properties of the algorithm. He showed that the algorithm produces estimates that do converge to a value which is a stationary point of the incomplete data likelihood, as long as the conditional expectation of the complete data likelihood is continuous.

Although the EM algorithm does not have a built-in procedure for calculating estimates of the variance-covariance matrix of the parameter estimates, Louis (1982) described a procedure to find the observed information matrix which requires taking derivatives only of the complete-data likelihood, which is usually more tractable than that of the observed data. Several other schemes to estimate the variance have been proposed since that of Louis. Some are described in detail in the book by McLachlan and Krishnan (2008).

Various extensions and modifications of the EM algorithm have been presented to simplify calculations or decrease the the computation time. One is the expectation-conditional maximization (ECM) algorithm, proposed by Meng and Rubin (1993). The ECM algorithm is useful when the complete-data likelihood is difficult to maximize. Instead of maximizing the conditional complete-data likelihood in the M-step, a series of conditional maximization (CM) steps maximize the conditional complete-data likelihood conditional on some of the parameters being estimated. Since each maximization is over fewer parameters, the entire algorithm can converge faster.

The ECM algorithm was extented to the ECME algorithm by Liu and Rubin (1994). ECME represents 'expectation-conditional maximization either,' and in this algorithm, some of the CM-steps maximize the conditional expectation of the actual,

or observed-data likelihood, conditional on some of the parameters, rather than the conditional expectation of the complete-data likelihood, conditional on some of the parameters. Although the ECME can be difficult to program because it involves the observed-data likelihood, the total computation time can be considerably reduced.

In some situations, the E-step can be difficult because of the integration involved in computing the expectation of the likelihood. In this case, one may simulate the missing data from the conditional distribution of the missing data, given the observed data and the current estimate of the parameters. This approach, called Monte Carlo EM (MCEM), was suggested be Wei and Tanner (1990). It uses the principle of Monte Carlo integration to estimate conditional expectation of the complete-data loglikelihood by taking the empirical loglikelihood based on the simulated sample. Then this approximated expectation is maximized in the M-step. For the other EM extensions mentioned here, ECM and ECEM, the attractive property of monotonically increasing sequence of incomplete likelihoods continues to hold as it does with EM itself. This is not true of MCEM because of the Monte Carlo error associated with the introduction of sampling within the E-step. This property is recovered with high probability using Caffo et al.'s (2005) ascent-based MCEM, which determines the Monte Carlo sample size at each E-step according to the data.

Chapter 3

Exploratory Analysis of Prostate Data

3.1 Overview

We did a number of exploratory analyses on the data. Here we present those analyses that motivate our model. In Section 3.2 we describe the contents of the data and how we handled discrepancies we found. Then we provide summary statistics in Section 3.3. Important endpoints are described in Section 3.4, and we develop regression models for each of those endpoints in Section 3.5.

3.2 Structure of the Data

The data consists of two records each for 506 patients enrolled in the trial. The first observation is labeled as the "initial history," or baseline set of measurements. The next record was taken at follow-up, which occurred about six months after the initial appointment. This study began in 1967 and followed some patients for more than ten years.

There were 292 stage III patients and 214 stage IV patients. They were randomized to four groups as shown in Table 3.1. Some of the measurements recorded were age, height, weight, the date of the exam, basic disease history, whether the patient had any matastases, whether the patient was experiencing dilation of the

upper urinary tract, the level of prostatic acid phosphatase, presence and extent of pain, level of hemoglobin, activity level, presence of breast tenderness or enlargement, which are side effects of DES, blood pressure, heart rate, electrocardiogram results, and cause of death if died on study. Cholesterol and level of alkaline phosphatase were recorded in some patients, but there were so many missing values of these variables that they were not used in any of our analyses. This was a multicenter study; one of the variables recorded for each patient was at which of 15 Veterans Administration Hospitals the patient was seen.

Table 3.1: Treatment Group Randomization by Stage

		Dose in Milligrams			
		Placebo	0.2	1.0	5.0
Stage	III	75	73	73	71
	IV	53	52	55	54

There were fairly detailed records of whether the patient was on or off study and why; the patients were categorized into on study, off study or died on study, off study because the patient received no treatment or not the assigned treatment, or off study because the patient was found to not have met the requirements for enrollment in the study. For each patient who was not recorded as on study, an explanation was provided, such as "patient too ill to continue study treatment," or "treated before placed on study." The frequencies for each reason are shown in Table 3.2. We excluded from the analysis 20 patients who were removed from treatment

or who should never have been enrolled in the study, and from this point on, our analyses reflect the group excluding those 20 patients. For the patients that we analyzed, we regarded their six–month follow–up record as occurring at six months after beginning treatment, even if they had expired while on–study prior to the six–month follow–up appointment.

Table 3.2: Patients Excluded From Analyses

	Frequency
Patient psychotic or otherwise refuses treatment	1
Patient left area with no contact information	1
Previous hormone treatment	1
Malignant prostate lesion other than carcinoma	1
Treated before placed on study	2
Stage misclassified	10
Protocol violated; no pre-treatment workup	2
Drug violation or treatment violation	2

Nine patients had missing values of PAP measured at follow-up. PAP values within an individual vary to within ±50% of the mean for that individual (Byar, Corle, and Brown 1988). Seventy patients had identical values recorded for the level of PAP at baseline and at follow-up. This was a surprising finding since the probability that such a variable quantity would equal the same amount on two occasions six months apart is zero. This might possibly be due to a missing value

at the follow-up that was imputed from the baseline measurement. However, for those patients who had the same values coded for both baseline and follow-up PAP, 54% died on study, while only 14% of the total group died on study. We conduct analyses both with and without these patients.

3.3 Basic Characteristics of the Patients

Summary statistics for the patients' ages are shown in Table 3.3. They are given by stage.

Table 3.3: Age by Stage

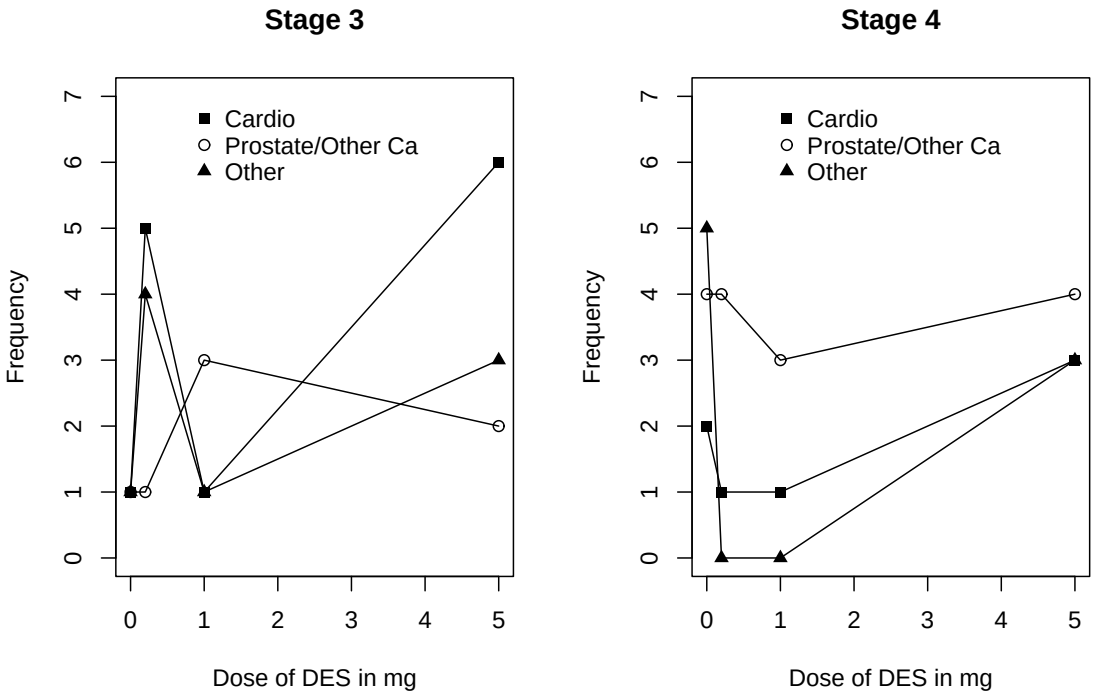
		n	$\bar{x}$	$s.d.$	min.	max.
Stage	III	243	71.82	6.69	49	89
	IV	172	71.40	7.11	48	85

The number of patients at each Veterans Administration hospital varied. The hospital with the most patients in the study was the one in Minneapolis, which had 119 patients. The one with the fewest was Omaha, which had five patients. The median number of patients was 27.

In all, 59 patients died during the time frame our data covers. Figure 3.1 shows the number of deaths due to prostate or other cancers, cardiovascular events, and other causes, by each dose group and by stage. There were not many deaths, but we can see that there were more cancer deaths for stage IV than for stage III,

which is to be expected.

Figure 3.1: Frequency of Different Causes of Death by Dose and by Stage



3.4 Important Endpoints

We would like to quantify how well the study drug affects the progression of the prostate cancer as well as quantify any toxicity.

Dying of prostate cancer would certainly be the primary endpoint of interest. Death due to other cancers would be strongly related. Developing new metastases would indicate disease progression, as would a change in treatment. We constructed

a binary endpoint to reflect progression of the prostate cancer which is equal to one if the patient died of prostate or some other cancer, developed a new metastasis, had a change in treatment, or had surgery while on study.

As discussed in 2.2.1, prostatic acid phosphatase (PAP) is an excellent tumor marker. The follow-up PAP levels had a skewed distribution, but they looked more symmetric on a log scale. Byar (1973) studied the effects of the different doses of DES on PAP, and he also looked at the effects on testosterone levels and on the size of the prostate as determined by rectal examination. We do not have access to the testosterone levels of the patients, but we did look at surface area of the prostate as a response.

The first VACURG study revealed that DES had toxic effects which caused cardiovascular disease. To quantify the effect of the toxicity and how it relates to dose, we looked at another binary endpoint which indicates death due to cardiovascular causes. We included death from heart or vascular diseases, death from stroke, and death from pulmonary embolus, following Byar (1973).

3.5 Model Selection

We fitted models for three endpoints discussed in Section 3.4: the log of the PAP level at follow-up, the binary indicator of cancer progression, and cardiovascular-related death.

We fitted models to the data with and without the 70 patients for which the baseline PAP was equal to the follow-up PAP. Including these observations in the

model for the response of log of PAP at follow-up will tend to have a flattening effect of the regression curve. One expects that the higher the dose, the lower the log of PAP, yet for these patients, their follow-up PAP is the same as their baseline PAP. So the effect of including these patients will be to underestimate the effect of dose on log of PAP at follow-up. However, as we noted in Section 3.2, these patients were more likely to have died than the other patients in the study. Not including these patients could bias the results because it would lower the estimated probability of death.

For the log of PAP at follow-up, to which we will refer here as 'logpap2,' we fitted a linear regression model using complete-case analysis. We looked at various predictors starting with the dose level, stage, and the baseline PAP level. However, staging in this study was performed using the baseline level of PAP and presence of distant metastasis (Byar, Corle, and Brown 1988), so that stage and baseline PAP were highly correlated. For this reason, stage was not a significant predictor in the presence of PAP, so we excluded it from the model. We looked at Gleason histology, which was an ordinal variable, but it was correlated with the log of initial PAP. The sample correlation coefficient was 0.24 with p-value less than 0.0001 for the test of the null hypothesis that the correlation is zero. Age was a significant predictor with a negative coefficient, so that the older a patient is, the less his PAP was at follow-up. This may be because cancer generally progresses more slowly in older patients. Dose was highly significant with a negative coefficient, showing that higher doses do reduce levels of PAP, which would suggest that the drug is having an effect on the cancer. Dose squared was also significant, so we retained it along with dose,

age, and logpap. Tables 3.4 and 3.5 show the coefficient estimates and estimated standard errors along with test statistics p-values for tests of difference from zero for the individual coefficients, both with and without those patients with identical values for baseline and follow-up PAP. The value of R^2 was 0.56, or 0.53 excluding the patients with imputed levels of PAP. We did not find that hospital location was a good predictor; the F-test of a better model including location versus the model shown in Table 3.5 had a p-value of 0.1563. A Q-Q plot of the residuals is given in Figure 3.2. There is some departure from normality because of the heavy tails of the distribution, as evidenced by the increased slope at the edges of the curve.

Table 3.4: Model: $\text{logpap2} = \beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{dose} + \beta_3 \,\text{dose}^2 + \beta_4 \,\text{age} + \text{error}$, including patients with imputed values of PAP

Variable	Estimate	*s.e.*	T-test statistic	p-value
intercept	1.04	0.53	1.98	0.0484
logpap	0.69	0.03	21.08	$< 2e-16$
dose	-0.99	0.16	-6.00	$4.49e-9$
dose^2	0.15	0.03	4.90	$1.41e-6$
age	-0.01	0.01	-1.65	0.0997

Surface area was also good outcome variable, but we did not find covariates to fit a model that was superior to logpap2. We display the model in Table 3.6.

In fitting the binary prostate cancer progression variable, which we refer to as 'binend,' we used probit regression, again using complete-case analysis. We started

Table 3.5: Model: $\text{logpap2} = \beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{dose} + \beta_3 \,\text{dose}^2 + \beta_4 \,\text{age} + \text{error}$, excluding patients with imputed values of PAP

Variable	Estimate	*s.e.*	T-test statistic	*p*-value
intercept	1.54	0.59	2.61	0.0095
logpap	0.67	0.04	17.99	$< 2e-16$
dose	-1.07	0.19	-5.71	$2.44e-8$
dose^2	0.16	0.04	4.52	$8.88e-6$
age	-0.02	0.01	-2.28	0.0232

out with logpap and dose as predictors. Dose did not turn out to be significant, which may be because this model is based on the number of events within six months, which is a relatively short period of time. Logpap was a good predictor, as was surface area of the prostate. Hemoglobin was significant, with a negative coefficient, meaning that patients who started out with a higher level of hemoglobin were less likely to have binend equal to one. Results for this model are given in Table 3.7.

For the cardiovascular-related death outcome, which we label 'carddeath,' for brevity, we investigated age, history of cardiovascular disease, dose, electrocardiogram, and hemoglobin as predictors in a probit regression model. Age did turn out to be significant, as expected, since this type of disease is associated with aging. Also, a history of cardiovascular disease was a good predictor of carddeath. The coefficient for dose was significant and had a positive sign, showing that the dose

Table 3.6: Model: surface area 2 $= \beta_0 + \beta_1$ logpap $+ \beta_2$ dose $+ \beta_3$ dose^2 $+ \beta_4$ surface area + error

Variable	Estimate	*s.e.*	T-test statistic	*p*-value
intercept	5.34	0.85	6.31	$7.25e-10$
logpap	0.83	0.29	2.84	0.0047
dose	-7.41	1.42	-5.23	$2.67e-7$
dose^2	1.22	0.27	4.55	$6.98e-6$
surface area	0.61	0.04	17.46	$2e-16$

increases, so does the incidence of carddeath. No other variables entered the model. Table 3.8 shows the results of the model.

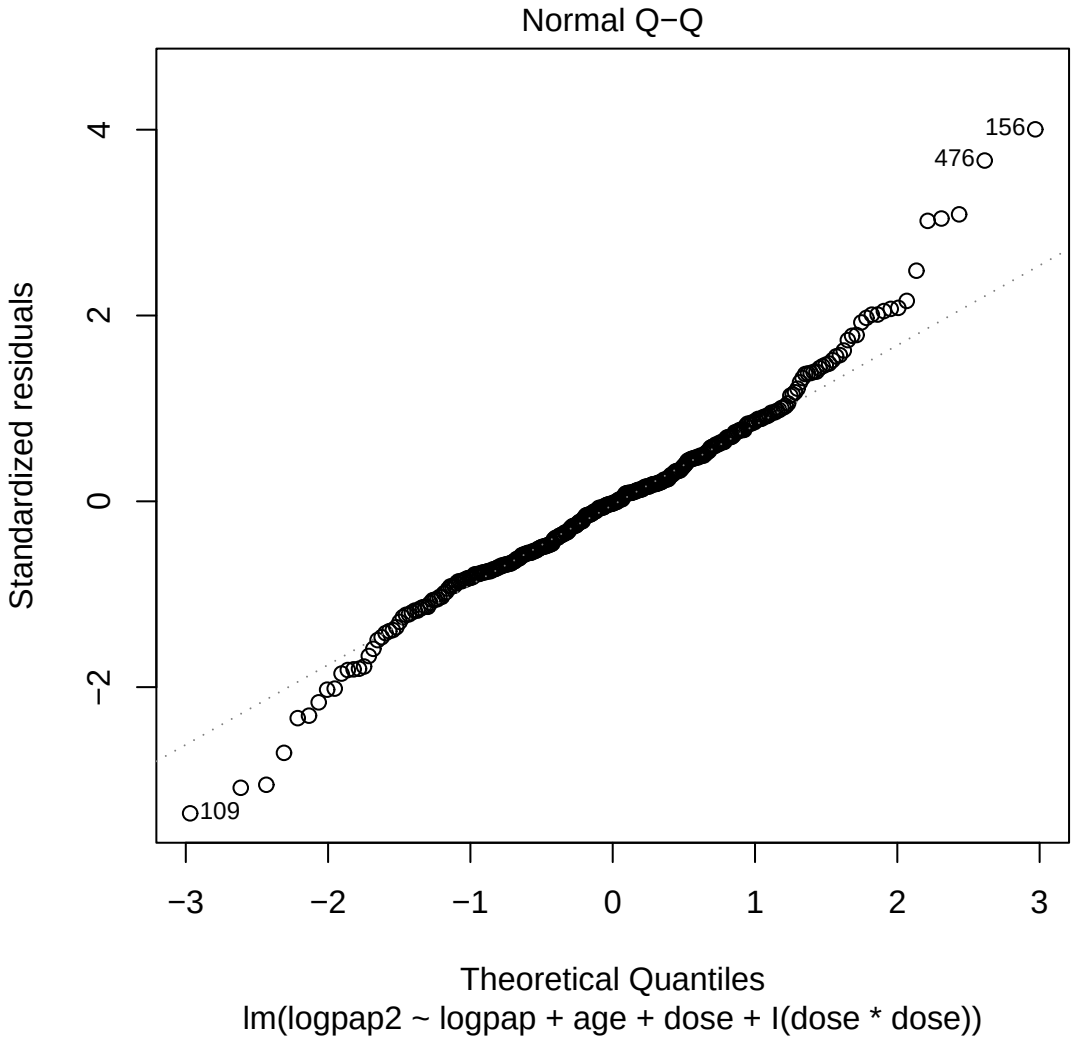
Table 3.7: Model: $P[\text{binend} = 1] = \Phi(\beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{surface area} + \beta_3 \,\text{dose} + \beta_4 \,\text{hemoglobin})$

Variable	Estimate	*s.e.*	T-test statistic	*p*-value
intercept	-0.07	0.52	-0.14	0.88633
logpap	0.13	0.05	2.82	0.00480
surface area	0.02	0.01	3.01	0.00265
dose	0.02	0.04	0.45	0.65227
hemoglobin	-0.09	0.04	-2.29	0.02210

Table 3.8: Model: $P[\text{carddeath} = 1] = \Phi(\beta_0 + \beta_1 \,\text{cardiodisease} + \beta_2 \,\text{age} + \beta_3 \,\text{dose})$

Variable	Estimate	*s.e.*	Z-test statistic	*p*-value
intercept	-5.93	1.54	-3.85	0.000116
cardio disease	0.88	0.24	3.64	0.000274
age	0.05	0.02	2.41	0.015872
dose	0.11	0.05	2.39	0.016737

Figure 3.2: QQ Plot of Residuals for Logpap2 Model



Chapter 4

Parameter Estimation Using the EM Algorithm

4.1 Overview

From Chapter 3, the most important endpoints that emerged from our exploratory analyses were prostatic acid phosphatase, occurrence of cardiovascular-related death, and progression of the prostate cancer as evidenced by new metastases, death from cancer, or a change in treatment. The first is continuous and, after a log transformation, easily lends itself to analyses using ordinary least squares. The latter two are binary, and can be modeled by logistic or probit regression. However, our interest lies in the joint distribution of the pair to allow for dependence between the two response variables. To this end, we can view the cardiovascular event as the observable product of some underlying disease process which follows a continuous distribution. If we condition on the value of the variables that we actually observe, we can use the expectation-maximization (EM) algorithm to find the maximum likelihood estimates of regression coefficients and variances.

The EM algorithm alternates between the E-step and the M-step. The E-step consists of taking the conditional expectation of the log likelihood of the complete model, given the value of the observed data. We compute this expectation analytically in Section 4.4. The M-step finds the values of the parameters that maximize the expectation found in the E-step. In Section 4.5, we describe our maximization

process. The maximum likelihood estimates are then substituted into the expectation, beginning a new iteration. We investigate the performance of the algorithm in Section 4.6. We outline a procedure to estimate standard errors of the estimates using bootstrap sampling in Section 4.7.

4.2 Model

For each individual, we have two endpoints. One is continuous, and one is binary. The data are $(Y_{1i}, Y_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}),\ i = 1, \ldots, n$, where (Y_{1i}, Y_{2i}) is the the row vector of endpoints for the i^{th} observation, and $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ are vectors of covariates consisting of the i^{th} row of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Let $\mathbf{Y}_1$ be the n-dimensional vector of the continuous endpoints and let $\mathbf{Y}_2$ be the vector of binary responses. We view the binary endpoint as partial information about an unobservable continuous random variable, $\mathbf{Y}_2^*$. Specifically, $Y_{2i} = I_{\{Y_{2i}^* > 0\}}$. Let $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]$ be the matrix of observable data, and let $\mathbf{Y}^*$ be the matrix of the complete data. We assume a joint multivariate normal distribution of $\mathbf{Y}_1$ and $\mathbf{Y}_2^*$ having within-observation correlation coefficient ρ. We assume that $E[\mathbf{Y}_1]$ and $E[\mathbf{Y}_2^*]$ are linear functions of the covariate matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, so that we have the the following linear models:

$$\begin{aligned} Y_{1i} &= \mathbf{x}_{1i}\boldsymbol{\beta}_1 + \epsilon_{1i} \\ Y_{2i}^* &= \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \epsilon_{2i} \end{aligned}$$

with the pairs $(\epsilon_{1i}, \epsilon_{2i})$ iid, and $(\epsilon_{1i}, \epsilon_{2i}) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

We want to estimate the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, ρ, and σ_1^2. Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho, \sigma_1^2\}$.

4.3 Constructing the Likelihood Function

We start with the joint bivariate normal density function:

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Since $E(\mathbf{Y}_1) = \mathbf{X}_1\boldsymbol{\beta}_1$ and $E(\mathbf{Y}_2^*) = \mathbf{X}_2\boldsymbol{\beta}_2$, the complete data likelihood for n individuals is

$$\begin{aligned} L(\boldsymbol{\theta}) \;=\; & \left(\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\right)^n \prod_{i=1}^{n} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_{1i}-\mathbf{x}_{1i}\boldsymbol{\beta}_1}{\sigma_1}\right)^2\right.\right. \\ & \left.\left. -\frac{2\rho(y_{1i}-\mathbf{x}_{1i}\boldsymbol{\beta}_1)(y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2)}{\sigma_1\sigma_2} + \left(\frac{y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right)^2\right]\right\}, \end{aligned}$$

where $\mathbf{x}_{ki}$ is a row vector consisting of the i^{th} row of $\mathbf{X}_k$.

We use the EM algorithm to find the maximum likelihood estimate of $\boldsymbol{\theta}$.

4.4 The E-Step: Expectation of L Given the Observable Data

We take the conditional expectation of the complete data log likelihood analytically. We condition on the observed data, $\mathbf{Y}$. Recall that Y_{2i} is 1 when Y_{2i}^* is positive, and zero otherwise. Observe that

$$Y_{2i} = 1 \Leftrightarrow Y_{2i}^* > 0 \Leftrightarrow \frac{Y_{2i}^*}{\sigma_2} > 0,$$

so that once the value of $\mathbf{Y}_2$ is observed, the next two relations follow, and we see that there is not enough information in the data to estimate σ_2 separately. For this reason we impose the constraint $\sigma_2 = 1$, which gives

$$\rho = \frac{\sigma_{12}}{\sigma_1}.$$

The conditional expectation of the complete data log likelihood is

$$\begin{aligned}
&E_{\hat{\boldsymbol{\theta}}}(\ln L(\boldsymbol{\theta})|\mathbf{Y}_1, \mathbf{Y}_2) = && (4.1)\\
&\quad -n\log(2\pi\sigma_1\sqrt{1-\rho^2}) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}\left(\frac{y_{1i}-\mathbf{x}_{1i}\boldsymbol{\beta}_1}{\sigma_1}\right)^2 \\
&\quad -\frac{1}{2(1-\rho^2)}\sum_{i:y_{2i}=1} E_{\hat{\boldsymbol{\theta}}}((y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1,\mathbf{Y}_2) && (4.2)\\
&\quad +\frac{\rho}{1-\rho^2}\sum_{i:y_{2i}=1}\left(\frac{y_{1i}-\mathbf{x}_{1i}\boldsymbol{\beta}_1}{\sigma_1}\right) E_{\hat{\boldsymbol{\theta}}}(y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2|\mathbf{Y}_1,\mathbf{Y}_2) && (4.3)\\
&\quad -\frac{1}{2(1-\rho^2)}\sum_{i:y_{2i}=0} E_{\hat{\boldsymbol{\theta}}}((y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1,\mathbf{Y}_2) && (4.4)\\
&\quad +\frac{\rho}{1-\rho^2}\sum_{i:y_{2i}=0}\left(\frac{y_{1i}-\mathbf{x}_{1i}\boldsymbol{\beta}_1}{\sigma_1}\right) E_{\hat{\boldsymbol{\theta}}}(y_{2i}^*-\mathbf{x}_{2i}\boldsymbol{\beta}_2|\mathbf{Y}_1,\mathbf{Y}_2). && (4.5)
\end{aligned}$$

In order to find the expectations in (4.2)–(4.5), we present the following lemma:

Lemma 1. *Let Z_1, Z_2 be standard normal random variables with correlation coefficient ρ. Then*

(i)

$$E(Z_2|Z_1 = z_1, Z_2 > c) = \rho z_1 + \frac{\sqrt{1-\rho^2}\,\phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)}{1-\Phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)},$$

(ii)

$$\begin{aligned}
&E(Z_2^2|Z_1 = z_1, Z_2 > c) = \\
&\quad \rho^2 z_1^2 + 1 - \rho^2 + \frac{\sqrt{1-\rho^2}\,\phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)(c+\rho z_1)}{1-\Phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)},
\end{aligned}$$

(iii)

$$E(Z_2|Z_1 = z_1, Z_2 < c) = \rho z_1 - \frac{\sqrt{1-\rho^2}\phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)}{\Phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)},$$

(iv)

$$E(Z_2^2|Z_1 = z_1, Z_2 < c) = \\ \rho^2 z_1^2 + 1 - \rho^2 - \frac{\sqrt{1-\rho^2}\phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)(c+\rho z_1)}{\Phi\left((c-\rho z_1)/\sqrt{1-\rho^2}\right)}.$$

Proof. For (i), we need to evaluate the following limit:

$$\lim_{h\to 0} \frac{(1/h)\, I_1}{(1/h)\, I_0},$$

where

$$I_1 = \int_{z_1}^{z_1+h} \int_{z_1=c}^{\infty} \frac{z_2}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)}{2(1-\rho^2)} \right) dz_2\, dz_1$$

and

$$I_0 = \int_{z_1}^{z_1+h} \int_{z_1=c}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)}{2(1-\rho^2)} \right) dz_2\, dz_1.$$

By completing the square in the exponential term, the integral in the numerator becomes:

$$I_1 = \int_{z_1}^{z_1+h} \phi(z_1) \int_{z_1=c}^{\infty} \frac{z_2}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-\left(z_2 - \rho z_1\right)^2}{2(1-\rho^2)} \right) dz_2\, dz_1.$$

Substituting

$$w = \frac{z_2 - \rho z_1}{\sqrt{(1-\rho^2)}},$$

and recognizing the standard normal density function, we have

$$I_1 = \int_{z_1}^{z_1+h} \phi(z_1) \int_{\frac{c-\rho z_1}{1-\rho^2}}^{\infty} (\rho z_1 + w\sqrt{1-\rho^2})\phi(w)\, dw\, dz_1.$$

Multiplying the terms and noting that

$$\int w\phi(w)\, dw = -\phi(w),$$

we have

$$I_1 = \int_{z_1}^{z_1+h} \phi(z_1) \left\{ \rho z_1 \left[1 - \Phi\left(\frac{c-\rho z_1}{\sqrt{(1-\rho^2)}} \right) \right] + \sqrt{(1-\rho^2)}\phi\left(\frac{c-\rho z_1}{\sqrt{1-\rho^2}} \right) \right\} dz_1.$$

Using the same technique of completing the square, I_0 is evaluated as

$$I_0 = \int_{z_1}^{z_1+h} \phi(z_1) \left[1 - \Phi\left(\frac{c-\rho z_1}{\sqrt{1-\rho^2}} \right) \right] dz_1.$$

After using the Mean Value Theorem and taking limits, we obtain (i).

For (ii), we must calculate

$$\lim_{h\to 0} \frac{(1/h)\, I_2}{(1/h)\, I_0},$$

with

$$I_2 = \int_{z_1}^{z_1+h} \int_{z_1=c}^{\infty} \frac{z_2^2}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-(z_1^2 - 2\rho z_1 z_2 + z_2^2)}{2(1-\rho^2)} \right) dz_2\, dz_1.$$

As before, we complete the square in the exponential and make the change of variables $w = (z_2 - \rho z_1)/\sqrt{1-\rho^2}$. Then I_2 becomes

$$I_2 = \int_{z_1}^{z_1+h} \phi(z_1) \int_{\frac{c-\rho z_1}{1-\rho^2}}^{\infty} (\rho z_1 + w\sqrt{1-\rho^2})^2 \phi(w)\, dw\, dz_1.$$

We expand the square. The integral of the linear term is evaluated as in (i), and the square term is evaluated using integration by parts. We obtain

$$I_2 = \int_{z_1}^{z_1+h} \phi(z_1) \left\{ \left[1 - \Phi\left(\frac{c - \rho z_1}{\sqrt{1-\rho^2}} \right) \right] (\rho^2 z_1^2 + (1-\rho^2)) + \sqrt{1-\rho^2}\, \phi\left(\frac{c - \rho z_1}{\sqrt{1-\rho^2}} \right) (c + \rho z_1) \right\} dz_1,$$

giving (ii).

The calculations for (iii) and (iv) are similar to those of (i) and (ii), except the limits of the inner integrals in an expression such as I_1 are from $z_2 = -\infty$ to $z_2 = c$. □

Taking the expectation in (4.3), we add and subtract $\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2$ to get

$$\begin{aligned} & E_{\hat{\boldsymbol{\theta}}}(y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2 | \mathbf{Y}_1, \mathbf{Y}_2) \\ &= E_{\hat{\boldsymbol{\theta}}}\left(y_{2i}^* - \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 + (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2) \,|\mathbf{y}_{1i}\,, \mathbf{y}_{2i}^* > 0 \right) \\ &= \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2 + E_{\hat{\boldsymbol{\theta}}}\left(y_{2i}^* - \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 \,\middle|\, \frac{\mathbf{y}_{1i} - \mathbf{x}_{1i}\hat{\boldsymbol{\beta}}_1}{\hat{\sigma}_1}\,, \mathbf{y}_{2i}^* - \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 > -\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 \right). \end{aligned}$$

Taking z_{1i} to be $(y_{1i} - \mathbf{x}_{1i}\hat{\boldsymbol{\beta}}_1)/\hat{\sigma}_1$, z_{2i} to be $y_{2i}^* - \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2$, and letting $c = -\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2$, we have

$$E_{\hat{\boldsymbol{\theta}}}(y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2 | \mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2 + E(z_{2i} | z_{1i}, z_{2i} > c).$$

Using the results of Lemma 1(i), we have that the expectation in (4.3) is

$$E_{\hat{\boldsymbol{\theta}}}(y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2 | \mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \hat{\rho} z_{1i} + \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c - \hat{\rho} z_{1i}) / \sqrt{1-\hat{\rho}^2} \right)}{1 - \Phi\left((c - \hat{\rho} z_{1i}) / \sqrt{1-\hat{\rho}^2} \right)},$$

since under the condition that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, the correlation between z_{1i} and z_{2i} is $\hat{\rho}$.

Looking now at the expectation in (4.2), we add and subtract $\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2$ again to get

$$E_{\hat{\boldsymbol{\theta}}}\left((y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1, \mathbf{Y}_2\right) = E_{\hat{\boldsymbol{\theta}}}\left(\left[y_{2i}^* - \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 + (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)\right]^2 |\mathbf{y}_{1i}\ , \mathbf{y}_{2i}^* > 0\right).$$

Multiplying and making the same substitutions, we have

$$\begin{aligned}
&E_{\hat{\boldsymbol{\theta}}}\left((y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1, \mathbf{Y}_2\right) \\
&= E_{\hat{\boldsymbol{\theta}}}\left(z_{2i}^2 + 2z_{2i}(\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2) + (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|z_{1i}, z_{2i} > c\right) \\
&= (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2 + 2(\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)E_{\hat{\boldsymbol{\theta}}}(z_{2i}|z_{1i}, z_{2i} > c) + E_{\hat{\boldsymbol{\theta}}}(z_{2i}^2|z_{1i}, z_{2i} > c).
\end{aligned}$$

Applying Lemma 1(i) and (ii), we find that the expectation in (4.2) is

$$\begin{aligned}
&E_{\hat{\boldsymbol{\theta}}}((y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1, \mathbf{Y}_2) \\
&= (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2 + 2(\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)\left[\hat{\rho}z_{1i} + \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}{1-\Phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}\right] \\
&\quad +\hat{\rho}^2 z_{1i}^2 + 1 - \hat{\rho}^2 + \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)(c+\hat{\rho}z_{1i})}{1-\Phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}.
\end{aligned}$$

Similar calculations yield the following for the expectation in (4.5):

$$E_{\hat{\boldsymbol{\theta}}}(y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2|\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \hat{\rho}z_{1i} - \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}{\Phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)},$$

and similarly for the expectation in (4.4):

$$\begin{aligned}
&E_{\hat{\boldsymbol{\theta}}}((y_{2i}^* - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2|\mathbf{Y}_1, \mathbf{Y}_2) \\
&= (\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)^2 + 2(\mathbf{x}_{2i}\hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{2i}\boldsymbol{\beta}_2)\left[\hat{\rho}z_{1i} - \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}{\Phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}\right] \\
&\quad +\hat{\rho}^2 z_{1i}^2 + 1 - \hat{\rho}^2 - \frac{\sqrt{1-\hat{\rho}^2}\phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)(c+\hat{\rho}z_{1i})}{\Phi\left((c-\hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}.
\end{aligned}$$

4.5 The M-Step: Maximizing with Respect to the Parameters

The conditional expectation of the complete data likelihood (4.1) is a function of the parameters, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, ρ, and σ_1, their estimated values from the previous EM iteration, $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\rho}$, and $\hat{\sigma_1}$, and the observed data, $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{X}_1$, and $\mathbf{X}_2$. The M-step involves finding the parameter vector $\boldsymbol{\theta}$ that maximizes this function. We devised an iterative two-part algorithm in R to find the estimates that maximize the conditional expectation (4.1).

We first maximize the profile likelihood with respect to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ analytically, holding ρ and σ_1 as constants using their estimates from the previous EM iteration. We take the partial derivatives of the conditional expectation (4.1) with respect to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and set each equal to zero, creating two matrix equations that are linear in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\begin{bmatrix} \mathbf{X}_1^t\mathbf{X}_1 & -\rho\sigma_1\mathbf{X}_1^t\mathbf{X}_2 \\ -\rho\mathbf{X}_2^t\mathbf{X}_1 & \sigma_1\mathbf{X}_2^t\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^t\mathbf{Y}_1 - \rho\sigma_1\mathbf{X}_1^t\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 - \rho\sigma_1\mathbf{X}_1^t f(\mathbf{Z}_1, \mathbf{C}) \\ -\rho\mathbf{X}_2^t\mathbf{Y}_1 + \sigma_1\mathbf{X}_2^t\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \sigma_1\mathbf{X}_2^t f(\mathbf{Z}_1, \mathbf{C}) \end{bmatrix}, \tag{4.6}$$

where $\mathbf{C} = -\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$, $\mathbf{Z}_1 = (\mathbf{Y}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1)/\hat{\sigma}_1$ and

$$f(\mathbf{Z}_1, \mathbf{C}) = \hat{\rho}z_{1i} + (-1)^{I\{y_{2i}=0\}} \frac{\sqrt{1-\hat{\rho}^2}\,\phi\left((c_i - \hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}{1 - \Phi\left((-1)^{I\{y_{2i}=0\}}(c_i - \hat{\rho}z_{1i})/\sqrt{1-\hat{\rho}^2}\right)}.$$

In the function f, the negative ones are multiplied only by the rows of the argument vectors for which $y_{2i} = 0$. We solved this system of equations using R.

The next step is to take the solutions, which we call $\hat{\hat{\boldsymbol{\beta}}}$, and substitute them for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in (4.1). We now optimize with respect to ρ and σ_1, using the optim() function in R. The L-BFGS-B method of optim() is a quasi-Newton algorithm which

allows one to specify an upper and lower bound for each variable (R 2007). We take the values returned by optim(), $\hat{\hat{\rho}}$ and $\hat{\hat{\sigma_1}}$, and substitute them in (4.6) for ρ and σ_1, respectively. This begins the next cycle of the M-step. We wrote this two-step process as a loop in R (R 2007) which repeats until all the parameter estimates converge. Let $[\hat{\hat{\boldsymbol{\beta}}}_{1,i}, \hat{\hat{\boldsymbol{\beta}}}_{2,i}]^t$ be the values of the estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ computed on the i^{th} iteration of the M-step algorithm, and $[\hat{\hat{\rho}}_i, \hat{\hat{\sigma}}_{1,i}]^t$ be the values of the estimates of ρ and σ_1 computed on the i^{th} iteration. We required that

$$\left(\begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_{1,i+1} \\ \hat{\hat{\boldsymbol{\beta}}}_{2,i+1} \end{bmatrix} - \begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_{1,i} \\ \hat{\hat{\boldsymbol{\beta}}}_{2,i} \end{bmatrix} \right)^t \left(\begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_{1,i+1} \\ \hat{\hat{\boldsymbol{\beta}}}_{2,i+1} \end{bmatrix} - \begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_{1,i} \\ \hat{\hat{\boldsymbol{\beta}}}_{2,i} \end{bmatrix} \right) < 10^{-5}$$

and

$$\left(\begin{bmatrix} \hat{\hat{\rho}}_{i+1} \\ \hat{\hat{\sigma}}_{1,i+1} \end{bmatrix} - \begin{bmatrix} \hat{\hat{\rho}}_i \\ \hat{\hat{\sigma}}_{1,i} \end{bmatrix} \right)^t \left(\begin{bmatrix} \hat{\hat{\rho}}_{i+1} \\ \hat{\hat{\sigma}}_{1,i+1} \end{bmatrix} - \begin{bmatrix} \hat{\hat{\rho}}_i \\ \hat{\hat{\sigma}}_{1,i} \end{bmatrix} \right) < 10^{-5}.$$

After the M-step has converged to a set of parameter estimates, $\hat{\boldsymbol{\theta}}$ is updated with those values, completing one iteration of the EM cycle. Next a new E-Step begins. The conditional expectation (4.1) and $f(\mathbf{Z}_1, \mathbf{C})$ are reevaluated using the updated value of $\hat{\boldsymbol{\theta}}$, and then a new M-Step begins. The EM process is written in a loop which contains the M-Step loop and repeats until the estimated values converge. Let $\hat{\hat{\boldsymbol{\theta}}}_i$ be the value of the estimate of $\boldsymbol{\theta}$ computed on the j^{th} iteration of the EM algorithm. At each iteration of the EM algorithm, the previous guess was $\hat{\boldsymbol{\theta}}$, and the current guess is denoted by $\hat{\hat{\boldsymbol{\theta}}}$. Again, we required an absolute convergence:

$$\left(\begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_1 \\ \hat{\hat{\boldsymbol{\beta}}}_2 \end{bmatrix} - \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} \right)^t \left(\begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}}_1 \\ \hat{\hat{\boldsymbol{\beta}}}_2 \end{bmatrix} - \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} \right) < 10^{-5}$$

and

$$\left(\begin{bmatrix} \hat{\hat{\rho}} \\ \hat{\hat{\sigma}}_1 \end{bmatrix} - \begin{bmatrix} \hat{\rho} \\ \hat{\sigma}_1 \end{bmatrix} \right)^t \left(\begin{bmatrix} \hat{\hat{\rho}} \\ \hat{\hat{\sigma}}_1 \end{bmatrix} - \begin{bmatrix} \hat{\rho} \\ \hat{\sigma}_1 \end{bmatrix} \right) < 10^{-5}.$$

4.6 Performance of the Algorithm on Simulated Data

We performed simulations to test the algorithm under varying configurations of the parameters and correlations between the covariate matrices. We first generated three independent random samples of size 100 from the standard normal distribution: $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$. For these covariates, we generated 1000 correlated $\mathbf{Y}_1$ and $\mathbf{Y}_2^*$ vectors from the following distribution:

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{Y}_2^* &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \end{aligned}$$

where $\mathbf{X}_1 = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_3 \end{bmatrix}$ and $\mathbf{X}_2 = \begin{bmatrix} \mathbf{U}_2 & \mathbf{U}_3 \end{bmatrix}$ with the pairs $(\epsilon_{1i}, \epsilon_{2i})$ iid, and $(\epsilon_{1i}, \epsilon_{2i}) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix},$$

so that the correlation between Y_{1i} and Y_{2i}^* is 0.35. We used $\mathbf{Y}_2^*$ to construct a vector of binary random variables, $\mathbf{Y}_2$, whose i^{th} row is $I\{y_{2i}^* > 0\}$. For each of the $\mathbf{Y}_1$, $\mathbf{Y}_2$ vector pairs, we executed our program which calculated maximum likelihood estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and σ_1. So, from the same set of covariate matrices, we have 1000 sets of response vectors and 1000 sets of the corresponding parameter estimates resulting from our algorithm. For each parameter, we took the sample

mean of the 1000 estimates and subtracted the true value to compute the bias. We also computed the sample standard deviations to estimate the standard error of the estimators, and we calculated an estimate of $\sqrt{MSE}$. The results are displayed in Table 4.1.

Table 4.1: Results of 1000 MC samples of size 100 with uncorrelated covariates and $\rho = 0.35$

Parameter	True Value	Bias	$\widehat{s.e.}$	Bias/s.e.	$\sqrt{MSE}$
$\beta_{1,0}$	-0.10	-0.005	0.100	-0.051	0.100
$\beta_{1,1}$	2.00	-0.002	0.113	-0.020	0.113
$\beta_{1,2}$	1.00	0.001	0.089	0.015	0.089
$\beta_{2,0}$	0.40	0.050	0.188	0.265	0.194
$\beta_{2,1}$	0.70	0.091	0.212	0.429	0.230
$\beta_{2,2}$	1.00	0.134	0.229	0.587	0.265
σ_1	1.00	-0.017	0.072	-0.230	0.074
ρ	0.35	0.019	0.155	0.122	0.156

Each EM algorithm converged after around 8 iterations. From Table 4.1, we see that the algorithm gives nearly unbiased estimates of the parameters $\beta_{1,0}$, $\beta_{1,1}$, $\beta_{1,2}$, σ_1, and ρ with relatively small standard error. The estimates of $\beta_{2,0}$, $\beta_{2,1}$, and $\beta_{2,2}$ show moderate bias. For this simulation, we initialized each $\hat{\boldsymbol{\theta}}$ using the marginal regression estimates. We also tried starting the algorithm with bad initial values, but the algorithm continued to produce good estimates, although it used

more iterations before converging.

We also wanted to compare our estimators with the marginal regression estimators. For each of the $\mathbf{Y}_1$ vectors in the simulation, we also computed the least squares estimates of $\boldsymbol{\beta}_1$ and σ_1, and we ran a probit regression for each $\mathbf{Y}_2$ to estimate $\boldsymbol{\beta}_2$. To compare these estimates with those obtained by our program, we took the ratio of the sample variance of the estimates obtained by the marginal regression and the sample variance of the estimates obtained by our program. We also calculated the ratio of the estimated mean squared error for for both estimates. These quantities are displayed in Table 4.2.

Table 4.2: Comparison of the results of the EM algorithm with the marginal regression estimates from simulation with $\rho = 0.35$

Parameter	$Var(\hat{\boldsymbol{\theta}}_{init})/Var(\hat{\boldsymbol{\theta}}_{EM})$	$MSE(\hat{\boldsymbol{\theta}}_{init})/MSE(\hat{\boldsymbol{\theta}}_{EM})$
$\beta_{1,0}$	0.998	0.999
$\beta_{1,1}$	1.050	1.050
$\beta_{1,2}$	1.002	1.001
$\beta_{2,0}$	1.011	1.006
$\beta_{2,1}$	1.044	1.036
$\beta_{2,2}$	1.026	1.017
σ_1	1.030	0.979

From Table 4.2, we see that the MC variances of the initial estimates and final estimates are similar. We perform the simulation again, with higher correlation

between $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Results are presented for uncorrelated, mildly correlated, and highly correlated covariate matrices in Tables 4.3–4.8. From Table 4.4, we see that the ratios of the variances for the two estimates are higher now, for the case when $\rho = 0.71$.

Table 4.3: Results of 1000 MC samples of size 100 using the same uncorrelated covariates from Table 4.1 and $\rho = 0.71$

Parameter	True Value	Bias	$\widehat{s.e.}$	Bias/s.e.	$\sqrt{MSE}$
$\beta_{1,0}$	-0.10	-0.002	0.101	-0.018	0.101
$\beta_{1,1}$	2.00	0.006	0.091	0.061	0.091
$\beta_{1,2}$	1.00	0.004	0.090	0.039	0.090
$\beta_{2,0}$	0.40	0.096	0.174	0.551	0.199
$\beta_{2,1}$	0.70	0.179	0.194	0.924	0.264
$\beta_{2,2}$	1.00	0.250	0.235	1.061	0.343
σ_1	1.00	-0.015	0.071	-0.215	0.072
ρ	0.71	0.099	0.069	1.437	0.121

In each simulation study, we see that generally $Var(\hat{\boldsymbol{\theta}}_{init})/Var(\hat{\boldsymbol{\theta}}_{EM}) \geq 1$, and when $\rho = 0.71$, the gain in efficiency can be substantial. Similarly, we see that $MSE(\hat{\boldsymbol{\theta}}_{init})/MSE(\hat{\boldsymbol{\theta}}_{EM}) \geq 1$, with substantial gain when $\rho = 0.71$. This suggests that even after accounting for bias, the EM estimates are more accurate than the marginal regression estimates.

Table 4.4: Comparison of the results of the EM algorithm with the marginal regression estimates from a simulation with $\rho = 0.71$ and uncorrelated covariates

Parameter	$Var(\hat{\boldsymbol{\theta}}_{init})/Var(\hat{\boldsymbol{\theta}}_{EM})$	$MSE(\hat{\boldsymbol{\theta}}_{init})/MSE(\hat{\boldsymbol{\theta}}_{EM})$
$\beta_{1,0}$	1.002	1.002
$\beta_{1,1}$	1.385	1.387
$\beta_{1,2}$	1.000	1.000
$\beta_{2,0}$	1.230	1.207
$\beta_{2,1}$	1.478	1.259
$\beta_{2,2}$	1.211	1.113
σ_1	1.044	1.000

4.7 Estimating Standard Errors Using Bootstrap

The EM algorithm does not have a built-in facility to estimate the covariance matrix of the parameters. We can use bootstrap resampling to estimate the standard errors of the estimates.

Suppose one wants to estimate the standard error of a statistic. The statistic is computed based on a given dataset; its value is denoted as $\tilde{\boldsymbol{\theta}}$. Bootstrapping is a resampling procedure in which rows of the $n \times k$ data matrix are sampled with replacement to form a new data matrix of the same dimension as the original. The statistic is computed for this new dataset. This is repeated r times to obtain r values of the statistic, denoted $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_r^*$. The sample mean and sample variance of

Table 4.5: Results of 1000 MC samples of size 100 using mildly correlated covariates ($\mathbf{U}_3 = 0.3\mathbf{U}_1 - 0.4\mathbf{U}_2 + \boldsymbol{\epsilon}, \epsilon_i \sim N(0,1)$) and $\rho = 0.71$

Parameter	True Value	Bias	$\widehat{s.e.}$	Bias/s.e.	$\sqrt{MSE}$
$\beta_{1,0}$	-0.10	-0.001	0.101	-0.007	0.101
$\beta_{1,1}$	2.00	-0.003	0.086	-0.033	0.087
$\beta_{1,2}$	1.00	-0.002	0.115	-0.018	0.115
$\beta_{2,0}$	0.40	0.092	0.168	0.549	0.191
$\beta_{2,1}$	0.70	0.153	0.187	0.818	0.241
$\beta_{2,2}$	1.00	0.222	0.241	0.921	0.327
σ_1	1.00	-0.020	0.069	-0.287	0.072
ρ	0.71	0.100	0.063	1.587	0.118

these r values are computed. The bias of the statistic is estimated as the sample mean of the $\boldsymbol{\theta}^*$ values minus $\tilde{\boldsymbol{\theta}}$, and the standard error of the statistic is estimated by the sample standard deviation of the $\boldsymbol{\theta}^*$ values.

We used simulated data to check the performance of the bootstrap standard error estimates against the the estimates from our simulation. We simulated $\mathbf{X}_1$ and $\mathbf{X}_2$ and 1000 observations of $\mathbf{Y}_1$ and $\mathbf{Y}_2^*$ in manner identical to that described Section 4.6, and we computed the EM estimates for each simulated dataset, which we used to compute standard error estimates. On each simulated dataset, we also drew 200 bootstrap samples, which we used to compute standard error estimates. The calculations were performed using the R function boot(). We present the results

Table 4.6: Comparison of the results from a simulation with $\rho = 0.71$ and mildly correlated covariates

Parameter	$Var(\hat{\boldsymbol{\theta}}_{init})/Var(\hat{\boldsymbol{\theta}}_{EM})$	$MSE(\hat{\boldsymbol{\theta}}_{init})/MSE(\hat{\boldsymbol{\theta}}_{EM})$
$\beta_{1,0}$	1.005	1.005
$\beta_{1,1}$	1.505	1.509
$\beta_{1,2}$	1.007	1.007
$\beta_{2,0}$	1.155	1.111
$\beta_{2,1}$	1.584	1.357
$\beta_{2,2}$	1.282	1.181
σ_1	1.049	0.971

in Table 4.9.

We assume the Monte Carlo estimates are accurate, since they result from 1000 datasets, and we compare the bootstrap estimates to the MC estimates. We see from the table that there are some differences. The bootsrap grossly overestimated the standard error of the estimate for $\beta_{2,0}$, $\beta_{2,1}$, and $\beta_{2,2}$, and underestimated their biases.

Table 4.7: Results of 1000 MC samples of size 100 using highly correlated covariates ($\mathbf{U}_3 = -0.7\mathbf{U}_1 - 0.8\mathbf{U}_2 + \boldsymbol{\epsilon}, \epsilon_i \sim N(0,1)$) and $\rho = 0.71$

Parameter	True Value	Bias	$\widehat{s.e.}$	Bias/s.e.	$\sqrt{MSE}$
$\beta_{1,0}$	-0.10	0.004	0.106	0.035	0.106
$\beta_{1,1}$	2.00	0.002	0.106	0.022	0.106
$\beta_{1,2}$	1.00	0.003	0.080	0.043	0.080
$\beta_{2,0}$	0.40	0.093	0.177	0.527	0.200
$\beta_{2,1}$	0.70	0.178	0.233	0.766	0.293
$\beta_{2,2}$	1.00	0.261	0.271	0.962	0.377
σ_1	1.00	-0.021	0.069	-0.301	0.072
ρ	0.71	0.101	0.068	1.476	0.122

Table 4.8: Comparison of the results from a simulation with $\rho = 0.71$ and highly correlated covariates

Parameter	$Var(\hat{\boldsymbol{\theta}}_{init})/Var(\hat{\boldsymbol{\theta}}_{EM})$	$MSE(\hat{\boldsymbol{\theta}}_{init})/MSE(\hat{\boldsymbol{\theta}}_{EM})$
$\beta_{1,0}$	1.001	1.001
$\beta_{1,1}$	1.296	1.297
$\beta_{1,2}$	1.070	1.071
$\beta_{2,0}$	1.163	1.110
$\beta_{2,1}$	1.525	1.327
$\beta_{2,2}$	1.179	1.095
σ_1	1.045	0.959

Table 4.9: Comparison of Standard Error and Bias Estimates from Monte Carlo Simulation and Bootstrap Using Slightly Correlated Covariates: ($\mathbf{U}_3 = 0.3\mathbf{U}_1 - 0.4\mathbf{U}_2 + \boldsymbol{\epsilon}, \epsilon_i \sim N(0,1)$) and $\rho = 0.35$

Parameter	True Value	MC Bias	BS Bias	MC $\widehat{s.e.}$	BS $\widehat{s.e.}$
$\beta_{1,0}$	-0.1	-0.001	0.000	0.100	0.099
$\beta_{1,1}$	2	0.001	0.000	0.114	0.114
$\beta_{1,2}$	1	-0.002	0.000	0.113	0.113
$\beta_{2,0}$	-0.4	0.043	0.027	0.175	0.242
$\beta_{2,1}$	0.7	0.078	0.056	0.199	0.323
$\beta_{2,2}$	1	0.127	0.089	0.257	0.448
σ_1	1	-0.018	-0.017	0.069	0.068
ρ	0.35	0.038	0.008	0.143	0.139

Chapter 5

Results of Multiple Endpoint Analysis of Prostate Data

From Chapter 3, we have three endpoints of interest. One is the log of prostatic acid phosphatase (PAP) as measured at follow-up, or logpap2. The second is an indicator of death due to cancer, metastases, or a change in treatment, which we call 'binend.' The third is an indicator of death due to cardiovascular disease, called 'carddeath.' In Section 4.6 we saw that our EM program estimates the regression parameters with reasonable accuracy when the data are normally distributed. Here we give the results of our program when applied to the prostate cancer data.

5.1 The Joint Model for Log of PAP and Binend

In Section 3.5 we developed a linear model for the response logpap2 as a function of logpap at baseline, dose, dose squared, and age. We used probit regression to model binend as a function of logpap at baseline, surface area of the prostate, dose, and hemoglobin, so that the two models are

$$\begin{aligned} \text{logpap2} &= \beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{dose} + \beta_3 \,\text{dose}^2 + \beta_4 \,\text{age} + \text{error} \\ P[\text{binend} = 1] &= \Phi(\beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{surface area} + \beta_3 \,\text{dose} + \beta_4 \,\text{hemoglobin}). \end{aligned}$$

Tables 5.1 and 5.2 give the results of the joint model with logpap2 and binend including and excluding, respectively, the 70 patients with questionable values of PAP recorded. The estimates from the separate regressions, which were given in

Chapter 3, are given, along with the EM estimates and standard errors as estimated by bootstrap resampling. The EM estimates are close to those estimated using regression, with the exception of the dose coefficient for the binend model in Table 5.2. Excluding the 70 patients had the effect of changing both intercept terms. The estimates of ρ are different in the two cases also, but in both cases positive and significantly different from zero. For the predictors of logpap2, the dose coefficient is negative, and even after one accounts for the dose squared term, which is positive, the higher the dose of DES, the lower the logpap2, but the effect lessens as the dose increases, because the squared term has a positive coefficient. Even over the short period of six months, the estrogen is reducing the PAP levels. The estimated standard error for the dose coefficient in the logpap2 model is fairly high, but we saw in Section 4.6 that the bootstrap tends to overestimate standard errors. Age had a negative coefficient in both cases, which could be explained by the slower progression of cancer in older patients. Its effect was only significant in the model excluding the 70 patients.

As a predictor of binend, dose is not significant regardless of whether we include or exclude the 70 patients with questionable values of PAP at follow-up, but this may be because we only have the statuses of patients as of about 6 months after beginning treatment. The immediate effect of dose on logpap2 does not carry over to the endpoint binend. Hemoglobin was a significant predictor of binend in the model including those 70 patients, but not the one excluding them, but surface area of the prostate was significant for both cases.

Table 5.1: EM results for joint model of logpap2 and binend, including the 70 patients with imputed values of logpap2

Parameter	GLM Est.	EM Est.	BS Bias	BS $\widehat{s.e.}$	p-val*
Coefficients for logpap2 regression					
logpap2 int.	0.975	1.102	0.073	0.601	0.0668
logpap	0.692	0.691	-0.002	0.046	0.0000
age	-0.011	-0.013	-0.001	0.008	0.1130
dose	-0.987	-0.962	-0.001	0.174	0.0000
dose squared	0.152	0.148	0.000	0.033	0.0000
Coefficients for binend regression					
binend int.	-0.162	-0.102	-0.047	0.506	0.8407
logpap	0.123	0.118	-0.005	0.048	0.0138
dose	0.024	0.030	-0.001	0.036	0.4010
surface area	0.018	0.015	0.001	0.006	0.0135
hemoglobin	-0.082	-0.084	0.002	0.038	0.0273
σ_1	0.989	0.983	-0.006	0.047	0.0000
ρ	?	0.270	0.007	0.064	0.0000

* Two-sided p-value based on a Wald test

Table 5.2: EM results for joint model of logpap2 and binend, excluding the 70 patients with imputed values of logpap2

Parameter	GLM Est.	EM Est.	BS Bias	BS $\widehat{s.e.}$
Coefficients for logpap2 regression				
logpap2 int.	1.461	1.412	0.061	0.675
logpap	0.672	0.672	0.004	0.053
age	-0.017	-0.017	-0.001	0.009
dose	-1.068	-1.061	0.000	0.184
dose squared	0.159	0.157	0.000	0.036
Coefficients for binend regression				
binend int.	-0.785	-0.755	-0.027	0.740
logpap	0.142	0.137	0.000	0.056
dose	-0.011	0.002	-0.005	0.049
surface area	0.022	0.021	0.001	0.007
hemoglobin	-0.063	-0.064	0.000	0.054
σ_1	1.039	1.032	-0.010	0.056
ρ	?	0.174	-0.002	0.085

5.2 The Joint Model for Log of PAP and Cardiovascular-Related Death

Our other binary endpoint was cardiovascular-related death, or carddeath, which we modeled using a probit regression with age, dose, and history of cardiovascular disease as covariates. We used our EM program to jointly model the two endpoints carddeath and logpap2:

$$\begin{aligned} \text{logpap2} &= \beta_0 + \beta_1 \,\text{logpap} + \beta_2 \,\text{dose} + \beta_3 \,\text{dose}^2 + \beta_4 \,\text{age} + \text{error} \\ P[\text{carddeath} = 1] &= \Phi(\beta_0 + \beta_1 \,\text{cardio disease} + \beta_2 \,\text{age} + \beta_3 \,\text{dose}). \end{aligned}$$

The results are displayed in Tables 5.3 and 5.4. Again, the estimates from the EM program are not very different from those obtained by fitting separate regression models. Including or excluding the 70 patients with imputed values of PAP makes a big difference. The estimate of ρ without those patients is not significantly different from zero, while the estimate computed with those patients is significant, with a z-score of about 3.8. The coefficient estimates for the predictors of logpap2 were similar to those obtained in the joint model for binend and logpap2.

As a predictor for cardiac death, dose has a significant effect if we include the 70 patients in the analysis, but not if we exclude them. This may be because of the comparatively high number of deaths among those 70 patients that we noted in Section 3.2, so that excluding them suppresses evidence of the effect of dose on cardiac-related death. The coefficient of the term indicating a history of cardiovascular disease had a very high standard error, causing it to not be significant,

whether or not those patients were included in the analyses. This may be because patients who already had a history of cardiovascular disease were being monitored and possibly medicated because of the concerns raised in Study 1 of the VACURG studies about cardiovascular event risks associated with DES.

5.3 Summary

We can see that in this study it makes a difference whether or not the 70 patients with questionable values of logpap2 are included in the analyses. The one observation of theirs that is in question is the value of PAP measured at follow-up, and we have seen that whether or not we include them does not make a big difference on the estimates of the predictors of logpap2. For this reason, we are inclined to trust the estimates based on including them in the analysis, because they include more observations. Having said that, we conclude that the drug DES significantly reduces PAP, increases the probability of cardiovascular-related death, but does not influence binend, at least over the initial six months of treatment. After observing the patients over the course of the entire study, Byar (1973) recommended that treatment be withheld until the patient's symptoms require relief, and at that time taking the 1 mg dose.

Table 5.3: EM results for joint model of logpap2 and carddeath, including the 70 patients with imputed values of logpap2

Parameter	GLM Est.	EM Est.	BS Bias	BS $\widehat{s.e.}$	p-val*
Coefficients for logpap2 regression					
logpap2 int.	0.975	0.966	-0.055	0.608	0.1118
logpap	0.692	0.694	0.002	0.046	0.0000
age	-0.011	-0.011	0.001	0.008	0.1842
dose	-0.987	-0.967	-0.021	0.165	0.0000
dose squared	0.152	0.148	0.005	0.031	0.0000
Coefficients for carddeath regression					
carddeath int.	-6.230	-6.163	-0.244	1.590	0.0001
cardio dis. hist.	1.018	0.999	0.235	0.987	0.3112
dose	0.119	0.101	0.001	0.050	0.0454
age	0.052	0.052	0.000	0.019	0.0057
σ_1	0.989	0.982	-0.013	0.051	0.0000
ρ	?	0.297	-0.001	0.078	0.0001

* Two-sided p-value based on a Wald test

Table 5.4: EM results for joint model of logpap2 and carddeath, excluding the 70 patients with imputed values of logpap2

Parameter	GLM Estimates	EM Estimates	BS Bias	$\widehat{BS s.e.}$
Coefficients for logpap2 regression				
logpap2 intercept	1.461	1.461	-0.033	0.696
logpap	0.672	0.674	-0.003	0.052
age	-0.017	-0.017	0.000	0.009
dose	-1.068	-1.068	-0.007	0.191
dose squared	0.159	0.159	0.002	0.036
Coefficients for caddeath regression				
carddeath intercept	-7.811	-7.697	-1.800	4.422
cardio disease hist.	0.349	0.342	0.533	1.545
dose	0.051	0.043	-0.011	0.222
age	0.074	0.073	0.014	0.046
σ_1	1.039	1.032	-0.009	0.055
ρ	?	0.173	-0.007	0.168

Chapter 6

Conclusions

6.1 Summary

We have developed an EM algorithm with an analytically–performed E–step which gives maximum likelihood estimates of regression coefficients, variances, and correlation for a binary and a continuous response variable. We did a limited simulation study in which we showed that the algorithm can increase the efficiency considerably when the two responses are highly correlated. The EM algorithm does not include estimates of the standard errors, so we used the bootstrap to estimate them.

We examined a prostate cancer study, but first did some data screening in which we raised questions about the accuracy of the records for some of the patients. We found preliminary regression models based on complete–case records. We fitted the mean log of PAP as a linear function of the predictors, and we used probit regression to model the incidence of cardiovascular death as a response and a composite indicator of cancer progression as a response. We found that dose reduced PAP and increased probability of cardiovascular–related death, but did not significantly affect binend, which indicates cancer mortality and progression to stage IV cancer. Our EM algorithm also showed these results, and it estimated moderate positive correlation between both logpap2 and binend and logpap2 and carddeath.

6.2 Possible Topics for Future Research

We used the ordinary bootstrap to estimate the standard errors of the estimates computed by our algorithm, but we saw in our simulation that the bootstrap tended to overestimate the standard errors of the coefficients of the binary random variable's predictors. A logical next step would be to find improved methods of estimating the standard errors. Improved methods could involve a variation of the bootstrap by using parametric bootstrap or bootstrapping the residuals. A different direction would be to calculate the standard errors analytically by using the methods of Louis (1982) or any of a number presented in the book by McLachlan and Krishnan (2008).

We have also seen that the estimates of the coefficients of the continuous variable's predictors from our algorithm are very close to those computed from the linear regression. This suggests that we could reduce the dimension of the problem by only estimating the correlation, ρ, and the coefficients of the binary endpoint's predictors, thereby reducing the computation time. Some analytic study is needed to assess the performance of this modified algorithm, followed by more comprehensive simulation studies.

Bibliography

[1] Bailar, J. C., and Byar, D. P. (1970) "Estrogen for Cancer of the Prostate: Early Results With 3 Doses of Diethylstilbestrol and Placebo," *Cancer* 26, 257-261.

[2] Byar, D. P. (1973) "The Veterans' Administration Cooperative Urological Research Group's Studies of Cancer of the Prostate," *Cancer* 32, 1126-1130.

[3] Byar D. P., Corle D. K., and Brown, C. C. (1988), "Observations on Acid Phosphatase in the VACURG Studies of Prostatic Cancer," in *Clinical Aspects of Prostate Cancer* (1989 ed.), eds. W. J. Catalona, D. S. Coffey, and J. P. Karr. New York: Elsevier Science, 29-36.

[4] Caffo, B. S., Jank, W., and Jones, G. L. (2005) "Ascent-Based Monte Carlo Expectation-Maximization," *Journal of the Royal Statistical Society*, Ser. B, 67, 235-251.

[5] Catalano, P. J., and Ryan, L. M. (1992) "Bivariate Latent Variable Models for Clustered Discrete and Continuous Outcomes," *Journal of the American Statistical Association* 87, 651-658.

[6] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

[7] Fitzmaurice, G. M., and Laird, N. M. (1995) "Regression Models for a Bivariate Discrete and Continuous Outcome With Clustering," *Journal of the American Statistical Association* 90, 845-852.

[8] Geys, H., Regan, M. M., Catalano, P. J., and Molenberghs, G. (2001) "Two Latent Variable Risk Assessment Approaches for Mixed Continuous and Discrete Outcomes From Developmental Toxicity Data," *Journal of Agricultural, Biological and Environmental Statistics* 6, 340-355.

[9] Gueorguieva, R. V., and Agresti, A. (2001) "A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous responses," *Journal of the American Statistical Association* 96, 1102-1112.

[10] Liu, C., and Rubin, D. B. (1994) "The ECME Algorithm: a Simple Extension of EM and ECM With Faster Monotone Convergence. *Biometrika* 81, 633-648.

[11] Louis, T. A. (1982) "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226-233.

[12] McLachlan, G. J., Krishnan, T. (2008) *The EM Algorithm and Extensions* (3rd ed.), Hoboken, NJ: John Wiley.

[13] Meng, X.-L., and Rubin, D. (1993) "Maximum Likelihood Estimation via the ECM Algorithm: a General Framework," *Biometrika* 80, 267-278.

[14] Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997) "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society*, Ser. B, 59, 667-678.

[15] Wei, G. C. G., and Tanner, M. A. (1990) "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association* 85, 699-704.

[16] Wu, C. F. J. (1983) "On the Convergence Properties of the EM Algorithm," *Annals of Statistics* 11, 95-103.