Accounting for Measurement Error in Covariates in the Context of ANCOVA Using Maximum Likelihood Estimation

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistics

in the

College of Graduate Studies

University of Idaho

by

Ana Wilson

Approved by:

Major Professor: Timothy Johnson, Ph.D.

Committee Members: Audrey Fu, Ph.D.; Chris Williams, Ph.D.

Department Administrator: Hirotachi Abo, Ph.D.

May 2022

Abstract

Analysis of covariance (ANCOVA) is a common statistical model. An implicit assumption of ANCOVA is that the covariate is measured without error. However, in many applications, there is covariate measurement error. In this case, the estimates produced by classic ANCOVA methods can include bias, causing predictions and inferences to be inaccurate. This thesis uses monte carlo simulation to examine the effectiveness of an alternative model in estimating the parameters associated with ANCOVA. This model is shown to be effective in accounting for covariate measurement error in the case where there are two treatment groups.

Acknowledgements

There are many people whose support and guidance have contributed to this thesis. I would like to express my deepest appreciation to my adviser Dr. Timothy Johnson. Thank you for helping me find a project that was interesting and that I enjoyed from start to finish. I would also like to thank Dr. Audrey Fu for being a member of my committee and offering encouragement and feedback.

I would also like to extend my deepest gratitude to Dr. Chris Williams, without whom, I would have never considered continuing my education past my bachelor's degree. Thank you for seeing something in me that I couldn't see in myself and helping me achieve more than I could have ever imagined.

I would like to extend my sincere thanks to Jaclyn Gotch, Melissa Gottschalk, and Jana Joyce. Thank you for all of your help in navigating my master's program and all my dates and deadlines. Thank you for always being a smiling, friendly face.

Dedication

This thesis is dedicated to those whose unfailing love and support helped me make it through the hard times. To my best friend, Victoria, who always had an ear to listen, a shoulder to cry on, and some commentary on trashy reality TV to take my mind off things. To Ann, who always assured me that I can do it, not matter what it was. To Tristan, with whom I shared hours of work along with endless commiseration and optimism. And to my cat, Spot, who always reminded me to take breaks to pet him.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Analysis of covariance (ANCOVA) is a commonly used statistical model. ANCOVA is used in cases where researchers are interested in modeling group differences while accounting for a continuous covariate. One of the assumptions of ANCOVA is that the covariate is measured without error. When the covariate is measured with error, the parameter estimates produced by ANCOVA may be biased [8]. The bias in the parameter estimates causes incorrect predictions and incorrect test statistics that may cause incorrect inferences [5].

Consider the CrabShip data in the Stat2Data package in R [2] [9]. This data set came from an experiment where researchers sought to investigate how stress affects the rate of oxygen intake of crabs. The researchers randomly assigned the crabs to either listen to ambient noise or noise produced by ships. They then measured the mass of the crabs and the rate of oxygen intake of each crab. In a design such as this one, if the mass is measured without error, then ANCOVA will produce unbiased estimates that are useful for prediction and inference. However, if the mass of the crabs was estimated with error, the parameter estimates produced by ANCOVA will likely produce incorrect predictions and may cause inaccurate inferences. In this case, researchers are forced to choose between using classic ANCOVA methods knowing that their parameter estimates are biased, using some method to correct for the problem, or to not use ANCOVA.

Many methods have been developed to account for covariate measurement error in the context of regression [4] [1]. One limitation of these approaches is that they often rely on knowing or assuming information about the variance associated with the covariate measurement error in order to correct for the bias. Culpepper and Aguinis applied many of these methods in the context of ANCOVA and compared their relative effectiveness in accounting for covariate measurement error [3]. Lockwood and McCaffrey performed a similar comparison in the context of pre and post test data in an educational context [7]. Both studies conclude that these methods are generally effective for accounting for covariate measurement error. The authors also remark that some of these methods are difficult or impossible to implement in practice.

The goal of this thesis is to present a method for accounting for covariate measurement error that does not require any knowledge or assumptions about the variance associated with the covariate measurement error. We will explain the model and its assumptions as well as the parameter estimation. We will then use Monte Carlo methods to demonstrate the effectiveness of the model in accounting for covariate measurement error. We will then demonstrate the model on the CrabShip data and compare it to a classic ANCOVA approach. We will also discuss some of the limitations of the proposed model.

1.1 The Model

Consider a randomized experiment with g treatment groups. Observed data include a continuous response variable and a continuous covariate. Let Y_{ig} denote the i-th response in the g-th group and X_{ig} be the covariate measurement of the i-th unit in the g-th group. Then we can assume the following model

$$Y_{ig} = \alpha_g + \beta_g X_{ig} + \epsilon_{ig}.$$

Now, suppose that the covariate X_{ig} is measured with error and we observe Z_{ig} where $Z_{ig} = X_{ig} + \zeta_{ig}$. In this case, we only observe Z_{ig}, and not X_{ig}. Additionally, we will assume that X_{ig}, ϵ_{ig}, and ζ_{ig} are all mutually independent random variables and σ_x^2, σ_ϵ^2, and σ_ζ^2 are the variances of X_{ig}, ϵ_{ig}, and ζ_{ig}, respectively. We also assume that $E(X_{ig}) = \delta$, $E(\epsilon_{ig}) = 0$, and $E(\zeta_{ig}) = 0$ for all i and g. Finally, we assume that the distribution of the covariate does not depend on the group. This can be achieved if there is random assignment of treatment groups and the treatment does not affect the covariate. In this model, if X_{ig} is measured without error, then $\sigma_\zeta^2 = 0$, $X_{ig} = Z_{ig}$, and the model corresponds to the classic ANCOVA model. Though, in ANCOVA, we typically include the additional assumption that all β_g are equal.

In the case where we assume the covariate X_{ig} is measured with error, $\sigma_\zeta > 0$. We will also allow the β_g to differ. We will call this the Covariate Error Model (CEM).

1.1 Parameterization

The parameters of primary interest are the α_g and β_g parameters for each group g and functions thereof. These parameters allow us to make inferences about expected differences between groups. For example, researchers are often interested in the difference in the expected response between two treatment groups, g and g'. We can estimate this difference either for a specific $X_{ig} = x$ or in the case where we do not condition on the covariate. In the first case, it can be shown that

$$E(Y_{ig}|X_{ig} = x) - E(Y_{ig'}|X_{ig} = x) = \alpha_g - \alpha_{g'} + (\beta_g - \beta_{g'})x.$$

In the case where we do not condition on the covariate, we have

$$E(Y_{ig}) - E(Y_{ig'}) = \alpha_g - \alpha_{g'} + (\beta_g - \beta_{g'})\delta.$$

In order to estimate these expected differences and have meaningful inferences, we need to be able to accurately estimate α_g, β_g, and δ for all g.

In order to estimate these parameters, we can use maximum likelihood estimation. First, we consider a model for the joint distribution of $(Y_{ig}, Z_{ig})'$. Then for the g-th group, we have

$$\begin{pmatrix} Y_{ig} \\ Z_{ig} \end{pmatrix} = \begin{pmatrix} \alpha_g \\ 0 \end{pmatrix} + \begin{pmatrix} \beta_g & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_{ig} \\ \epsilon_{ig} \\ \zeta_{ig} \end{pmatrix}.$$

It can be shown that

$$E \begin{pmatrix} Y_{ig} \\ Z_{ig} \end{pmatrix} = \begin{pmatrix} \alpha_g + \beta_g \delta \\ \delta \end{pmatrix}$$

and

$$Cov \begin{pmatrix} Y_{ig} \\ Z_{ig} \end{pmatrix} = \mathbf{LPL}',$$

where

$$\mathbf{L} = \begin{pmatrix} \beta_g & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and

$$\mathbf{P} = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_\zeta^2 \end{pmatrix}.$$

So, we can write the covariance matrix as

$$Cov \begin{pmatrix} Y_{ig} \\ Z_{ig} \end{pmatrix} = \begin{pmatrix} \beta_g^2 \sigma_x^2 + \sigma_\epsilon^2 & \beta_g \sigma_x^2 \\ \beta_g \sigma_x^2 & \sigma_x^2 + \sigma_\zeta^2 \end{pmatrix}.$$

We will assume that the joint distribution of X_{ig}, ϵ_{ig}, and ζ_{ig} is multivariate normal. Then, the joint distribution of Y_{ig} and Z_{ig} is also multivariate normal with the mean vector and covariance matrix given above. If we have G groups and the g-th group has n_g observations, then the likelihood function can be written as

$$L(\boldsymbol{\theta}) = \prod_{g=1}^{G} \prod_{i=1}^{n_g} f(y_{ig}, z_{ig}|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\alpha_1, \alpha_2, ..., \alpha_G, \beta_1, \beta_2, ..., \beta_G, \delta, \sigma_x^2, \sigma_\epsilon^2, \sigma_\zeta^2)$ is the vector of all parameters, and f is the probability density function for a multivariate normal distribution. $(Y_{ig}, Z_{ig})'$ is independent and identically distributed within each group g. Thus, we can write the log-likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{g=1}^{G} \log L_g(\boldsymbol{\theta})$$

where

$$\log L_g(\boldsymbol{\theta}) = (\bar{\mathbf{y}}_g - \boldsymbol{\mu}_g)' \frac{n_g}{2} (\bar{\mathbf{y}}_g - \boldsymbol{\mu}_g) - \frac{n_g}{2} \log |\boldsymbol{\Sigma}_g^{-1}| + \frac{n_g}{2} tr(\mathbf{S}_g \boldsymbol{\Sigma}_g^{-1}),$$

omitting a constant that does not rely on $\boldsymbol{\theta}$ [6]. Here we define $\bar{\mathbf{y}}_g$ to be the mean vector for the sample of n_g observations of $(Y_{ig}, Z_{ig})'$ and similarly, $\mathbf{S}_g$ is the covariance matrix for the sample. Additionally,

$$\boldsymbol{\mu}_g = \begin{pmatrix} \alpha_g + \beta_g \delta \\ \delta \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \beta_g^2 \sigma_x^2 + \sigma_\epsilon^2 & \beta_g \sigma_x^2 \\ \beta_g \sigma_x^2 & \sigma_x^2 + \sigma_\zeta^2 \end{pmatrix}.$$

By maximizing the log-likelihood function, we will be able to obtain estimates for the parameters. In particular, we will obtain estimates for our parameters of interest, α_g and β_g.

1.1 Model Identification

When we say a model is identified, we mean that there is a unique set of parameters that produce the mean vector and variance-covariance matrix which maximize the likelihood function. If we look at the case where $G = 2$ and $\beta_1 = \beta_2 = \beta$, it can be shown that the model is not identified. In this case, we have

$$\boldsymbol{\mu}_g = \begin{pmatrix} \alpha_g + \beta \delta \\ \delta \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \beta^2\sigma_x^2 + \sigma_\epsilon^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma_\zeta^2 \end{pmatrix}.$$

We can define the following set of parameters

$$\begin{aligned} \tilde{\beta} &= \beta/c, \\ \tilde{\sigma}_x^2 &= c\sigma_x^2, \\ \tilde{\sigma}_\zeta^2 &= \sigma_x^2 + \sigma_\zeta^2 - c\sigma_x^2, \\ \tilde{\sigma}_\epsilon^2 &= \beta^2\sigma_x^2 + \sigma_\epsilon^2 - \beta^2\sigma_x^2/c, \\ \tilde{\alpha}_g &= \alpha_g + \beta\delta(1 - 1/c), \end{aligned}$$

where $c > 0$ is any constant such that $\tilde{\sigma}_x^2 > 0$, $\tilde{\sigma}_\zeta^2 > 0$, and $\tilde{\sigma}_\epsilon^2 > 0$. Substituting these values into $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, give the same mean vector and variance-covariance matrix. Thus, the model is not identified in this case.

In general, it is difficult to prove that a model is identified. In the case that $G = 2$ and $\beta_1 \neq \beta_2$, the CEM seems to be identified. For there to not be any other constraints on model identification in this case is surprising. The lack of required constraints on the model help provide a versatile model to help account for covariate measurement error in the context of ANCOVA.

1.2 Method

To investigate the effectiveness of the CEM, we used simulated data for the case where $G = 2$. Data were simulated based on the assumptions of the model for chosen parameter values. Then, we estimated the parameters of the model two ways. First, using the CEM method described above and assuming covariate measurement error. Second, we used classic ANCOVA methods where no measurement error was assumed and allowing the slopes for the groups to differ. The ANCOVA estimates were obtained using the lm() function in R. The CEM estimates were obtained using maximum likelihood estimation with the optim() function in R.

To see how different parameters affected the estimates, we varied some parameters. For each set of parameters, we simulated 1000 samples and estimated the model parameters for each sample. In the case that optim() failed, the estimates were recorded as NA. The parameters we varied were the sample size n, the covariate measurement error variance σ_ζ^2, and β_2. We have discussed that when $\beta_1 = \beta_2$, the model is not identified, so we wanted to investigate how the model performed as the difference between β_1 and

β_2 approached 0. Additionally, in the case where $\sigma_\zeta^2 = 0$, we have $X_{ig} = Z_{ig}$ and there is no covariate measurement error. So we were interested to see how the CEM performed in this case and how the two models compared as σ_ζ^2 increased.

Table 1.1 gives all the parameter values used in simulations. In the first set of simulations, all values of β_2 were used with all values of n while σ_ζ^2 was kept constant at $\sqrt{2}$. In the second set of simulations, all values of β_2 were used with all values of σ_ζ^2 while n was kept constant at 500.

Table 1.1: Parameter values used for simulations

Parameter	Values
α_1	0
α_2	1
β_1	1
β_2	2, 1.5, 1.1, 1.08, 1.06, 1.04, 1.02, 1
δ	10
σ_x^2	2
σ_ϵ^2	1
σ_ζ^2	2, $\sqrt{2}$, 1, 0.5, 0
n	30, 100, 500, 1000

1.2 Maximization Failure

While obtaining the CEM estimates, we used maximum likelihood estimation which relies on being able to maximize the likelihood function. Since the CEM assumes a normal distribution for the data, that means the variance-covariance matrix, $\mathbf{\Sigma}$, must have an inverse. In some cases, the simulated data caused the variance-covariance matrix to become singular or nearly singular during the optimization algorithm and the optimization failed. Table 1.2 shows the how many times the optimization algorithm failed for each set of parameters. For each set of parameters, the total number of iterations is 1000.

Table 1.2: CEM maximization failures

β_2	n=30	n=100	n=500	n=1000	$\sigma_\zeta^2 = 0$	$\sigma_\zeta^2 = 0.5$	$\sigma_\zeta^2 = 1$	$\sigma_\zeta^2 = 2$
2	4	0	0	1	48	3	0	0
1.50	1	1	0	0	58	0	0	1
1.10	8	1	0	0	152	7	0	0
1.08	3	0	0	0	136	7	0	0
1.06	2	2	0	0	140	13	4	0
1.04	8	0	0	0	91	5	1	1
1.02	4	1	0	0	97	3	0	0
1	9	0	0	0	83	0	0	0

We can see from this table that we failed to obtain estimates commonly for a sample size of 30, especially as β_2 gets closer to 1. The failures were most common in the case where $\sigma_\zeta^2 = 0$, when there is

no measurement error. In this case, the classic ANCOVA approach and the CEM approach are equivalent, so either model can be used.

Chapter 2: Comparing Models

2.1 Parameter Estimates and Bias

For each simulation, we computed the bias for the parameter estimates for both the CEM and ANCOVA. Table 2.1 gives the bias for the parameter estimates for different sample sizes and different values of β_2. The bias for the estimates using the CEM are presented along side the bias for estimates calculated using ANCOVA. For all of these simulations, the values of the other parameters are $\alpha_1 = 0$, $\alpha_2 = 1$, $\beta_1 = 1$, $\delta = 10$, $\sigma_\epsilon^2 = 1$, $\sigma_x^2 = 2$, and $\sigma_\zeta^2 = \sqrt{2}$. Note that while the CEM produces estimates for all parameters, only the bias for the estimates of $\alpha_1, \alpha_2, \beta_1$, and β_2 are included in these tables.

Looking at table 2.1, it is evident that the CEM provides estimates with less bias than the classic ANCOVA approach when the covariate is measured with error. For every set of parameters, the CEM estimates have lower bias than the ANCOVA estimates by at least a factor of 10. This shows that the CEM does help account for covariate measurement error and provide less biased estimates.

When we consider how the estimates and bias are affected by the changing parameters, there are some interesting patterns that emerge. For the CEM, when β_1 and β_2 are farther apart, the parameter estimates are less biased as the sample size increases. Interestingly, when β_1 and β_2 are equal or very close to equal, this pattern reverses and the parameter estimates tend to become less biased with smaller sample sizes. Meanwhile, the bias in the ANCOVA estimates is mostly unchanged by the sample size. The bias in mainly affected by the change in the value of β_2, with the lowest bias when $\beta_1 = \beta_2$.

Table 2.2 gives the bias for both the CEM and ANCOVA estimates for simulations where we changed the parameter values for β_2 and σ_ζ^2. For all of these simulations, the values of the other parameters are $\alpha_1 = 0$, $\alpha_2 = 1$, $\beta_1 = 1$, $\delta = 10$, $\sigma_\epsilon^2 = 1$, $\sigma_x^2 = 2$, and $n = 500$. The sample size $n = 500$ was chosen because it had the fewest number of failed optimization attempts among the various sample sizes.

For the CEM we see two emerging patterns as we change β_2 and σ_ζ^2 together. First, we see that when as β_2 gets closer to β_1, the estimates for all parameters become more biased. In the case where $\sigma_\zeta^2 = 0$, the estimates become significantly more biased quite quickly, with the bias increasing sharply when $\beta_2 = 1.1$. For other values of σ_ζ^2, the increase is more gradual. The other interesting pattern we see is that as σ_ζ^2 increases, the parameter estimates first become less biased, then start to become more biased. In general, an increase in the amount of variance in the data can often result in less accurate parameter estimation. Of most interest here is the sharp decrease in bias between estimates when $\sigma_\zeta^2 = 0$ and when $\sigma_\zeta^2 = 0.5$. This suggests that the CEM does not perform as well estimating parameters in the case where there is no covariate measurement error in the model.

Table 2.1: Bias in parameter estimates using CEM and ANCOVA for different sample sizes with $\sigma_\zeta^2 = \sqrt{2}$

		CEM Bias				ANCOVA Bias			
β_2	Parameter	n=30	n=100	n=500	n=1000	n=30	n=100	n=500	n=1000
2	α_1	-0.137	-0.109	-0.043	-0.068	4.162	4.134	4.140	4.136
	α_2	-0.504	-0.134	-0.049	-0.049	8.234	8.229	8.272	8.311
	β_1	0.013	0.010	0.004	0.007	-0.416	-0.414	-0.414	-0.414
	β_2	0.049	0.012	0.005	0.005	-0.824	-0.824	-0.827	-0.831
1.50	α_1	-0.075	-0.426	-0.099	-0.101	4.096	4.121	4.151	4.130
	α_2	-0.349	-0.463	-0.107	-0.086	6.146	6.244	6.216	6.214
	β_1	0.009	0.043	0.010	0.010	-0.409	-0.412	-0.415	-0.413
	β_2	0.034	0.046	0.011	0.009	-0.616	-0.625	-0.621	-0.621
1.10	α_1	-0.118	0.009	0.049	0.269	4.100	4.149	4.145	4.149
	α_2	-0.172	-0.132	0.025	0.252	4.545	4.551	4.568	4.550
	β_1	0.012	-0.002	-0.005	-0.027	-0.411	-0.416	-0.414	-0.415
	β_2	0.019	0.012	-0.002	-0.025	-0.453	-0.456	-0.457	-0.455
1.08	α_1	-0.217	0.169	0.225	0.327	4.175	4.110	4.142	4.136
	α_2	-0.451	0.060	0.187	0.331	4.461	4.442	4.471	4.481
	β_1	0.022	-0.017	-0.023	-0.033	-0.417	-0.411	-0.414	-0.413
	β_2	0.045	-0.006	-0.019	-0.033	-0.446	-0.444	-0.447	-0.448
1.06	α_1	-0.128	0.110	0.444	0.393	4.164	4.145	4.150	4.139
	α_2	-0.326	0.005	0.393	0.365	4.439	4.359	4.373	4.378
	β_1	0.014	-0.011	-0.044	-0.039	-0.416	-0.414	-0.415	-0.414
	β_2	0.034	-0.001	-0.039	-0.036	-0.444	-0.436	-0.438	-0.438
1.04	α_1	-0.021	0.056	0.465	0.564	4.023	4.136	4.149	4.133
	α_2	-0.160	0.016	0.430	0.569	4.215	4.316	4.301	4.307
	β_1	0.000	-0.005	-0.046	-0.057	-0.404	-0.413	-0.415	-0.413
	β_2	0.016	-0.001	-0.043	-0.057	-0.422	-0.431	-0.430	-0.431
1.02	α_1	-0.220	-0.120	0.557	0.712	4.190	4.099	4.150	4.140
	α_2	-0.391	-0.110	0.532	0.725	4.260	4.191	4.220	4.230
	β_1	0.022	0.012	-0.056	-0.071	-0.420	-0.410	-0.415	-0.414
	β_2	0.041	0.011	-0.053	-0.072	-0.425	-0.419	-0.422	-0.423
1	α_1	0.085	0.018	0.331	0.597	4.277	4.156	4.132	4.139
	α_2	-0.014	-0.062	0.318	0.597	4.141	4.107	4.140	4.151
	β_1	-0.007	-0.002	-0.033	-0.060	-0.427	-0.416	-0.413	-0.414
	β_2	0.003	0.007	-0.032	-0.060	-0.413	-0.410	-0.414	-0.415

Table 2.2: Bias in parameter estimates using CEM and ANCOVA different covariate error variance with $n = 500$

		CEM Bias				ANCOVA Bias			
β_2	Parameter	$\sigma^2_\zeta = 0$	$\sigma^2_\zeta = 0.5$	$\sigma^2_\zeta = 1$	$\sigma^2_\zeta = 2$	$\sigma^2_\zeta = 0$	$\sigma^2_\zeta = 0.5$	$\sigma^2_\zeta = 1$	$\sigma^2_\zeta = 2$
2	α_1	-0.057	-0.012	-0.015	-0.099	-0.003	1.995	3.337	4.995
	α_2	-0.114	-0.006	-0.024	-0.127	-0.001	3.996	6.659	9.986
	β_1	0.006	0.001	0.001	0.010	0.000	-0.200	-0.334	-0.500
	β_2	0.011	0.001	0.002	0.013	0.000	-0.399	-0.666	-0.999
1.50	α_1	-0.135	-0.018	-0.085	-0.145	0.009	2.013	3.333	5.015
	α_2	-0.193	-0.029	-0.104	-0.153	0.021	3.002	4.986	7.510
	β_1	0.014	0.002	0.009	0.015	-0.001	-0.201	-0.333	-0.501
	β_2	0.019	0.003	0.010	0.015	-0.002	-0.300	-0.499	-0.751
1.10	α_1	-0.727	-0.067	0.021	0.283	0.015	1.996	3.334	5.008
	α_2	-0.816	-0.130	-0.005	0.245	-0.001	2.172	3.676	5.504
	β_1	0.073	0.007	-0.002	-0.028	-0.002	-0.199	-0.333	-0.501
	β_2	0.081	0.013	0.000	-0.025	0.000	-0.217	-0.368	-0.551
1.08	α_1	-0.716	0.168	0.298	0.505	0.008	1.995	3.348	4.993
	α_2	-0.780	0.148	0.272	0.492	0.008	2.152	3.609	5.382
	β_1	0.071	-0.017	-0.029	-0.050	-0.001	-0.200	-0.335	-0.499
	β_2	0.078	-0.015	-0.027	-0.049	-0.001	-0.215	-0.361	-0.538
1.06	α_1	-0.806	0.073	0.355	0.616	0.004	1.994	3.325	5.008
	α_2	-0.870	0.071	0.347	0.570	0.000	2.151	3.533	5.272
	β_1	0.081	-0.007	-0.036	-0.062	0.000	-0.200	-0.333	-0.501
	β_2	0.087	-0.007	-0.035	-0.057	0.000	-0.215	-0.353	-0.527
1.04	α_1	-0.743	0.280	0.313	0.601	-0.002	2.007	3.338	4.981
	α_2	-0.790	0.258	0.306	0.609	-0.003	2.087	3.487	5.191
	β_1	0.074	-0.028	-0.031	-0.060	0.000	-0.201	-0.334	-0.498
	β_2	0.079	-0.026	-0.030	-0.061	0.000	-0.209	-0.348	-0.519
1.02	α_1	-0.766	0.318	0.422	0.636	-0.003	1.994	3.346	5.006
	α_2	-0.789	0.308	0.413	0.643	-0.011	2.034	3.417	5.101
	β_1	0.077	-0.032	-0.042	-0.063	0.000	-0.199	-0.335	-0.500
	β_2	0.079	-0.031	-0.041	-0.064	0.001	-0.204	-0.341	-0.510
1	α_1	-0.649	0.323	0.318	0.722	0.023	2.006	3.338	5.002
	α_2	-0.682	0.314	0.305	0.724	-0.011	1.997	3.342	5.009
	β_1	0.065	-0.032	-0.032	-0.072	-0.002	-0.201	-0.334	-0.500
	β_2	0.068	-0.031	-0.031	-0.072	0.001	-0.200	-0.334	-0.501

Comparing the CEM estimates to the ANCOVA estimates, we can see that if there is covariate measurement error in the data, then the CEM produces less biased parameter estimates. In the case where $\sigma_\zeta^2 = 0$, ANCOVA produces estimates are are essentially unbiased. This suggests that in a situation where the covariate is measured without error, classic ANCOVA estimates have an advantage over the CEM.

For all estimates in all cases, the estimates are better for β_1 and β_2 than for α_1 and α_2. This is due in part to the fact that a small change in the slope of a line can have a large affect on the intercept of the line. Which means that the bias of the β parameters has an effect on the bias of the α parameters.

2.2 Parameter Estimate Error

Unbiasedness is a desirable property for parameter estimates, but it is not the only property to be considered. It is also important to consider the variation in the parameter estimates as well. There are a few ways to look at this. In this context, we calculated the root mean squared error (RMSE), the square root of the mean squared distance of each estimate from the true parameter value.

Table 2.3 shows the RMSE for the parameter estimates while changing the sample size and the value of β_2. The RMSE for the estimates obtained using the CEM are presented next to the RMSE for the estimates obtained using ANCOVA. In general, we can see that the RMSE for the CEM estimates is less than or equal to the RMSE for the ANCOVA estimates. For n=30, at values of β_2 close to 1, the RMSE for both estimation methods is similar. We can also see that the RMSE for the CEM shows some similar patterns to the bias for the CEM estimats. In general, the RMSE decreases as the sample size increases. For sample sizes larger than n=30, the RMSE for the CEM increases and then decreases as β_2 gets closer to 1. For all sample sizes, the CEM performs best when $\beta_2 = 2$. This is likely due to the issue of model identification when $\beta_1 = \beta_2$. For the ANCOVA estimates, the RMSE decreases slightly between n=30 and n=100 and then stays mostly constant as the sample size continues to increase. So while increasing the sample size seems to greatly improve the CEM estimates, it does not have as strong an effect on the ANCOVA estimates.

Table 2.4 shows the RMSE for the parameter estimates for different values of σ_ζ^2. The RMSE for the CEM estimates is presented alongside the RMSE for the ANCOVA estimates. We can see that for nearly every case, the RMSE for the parameter estimates is lower for the CEM estimates than for the ANCOVA estimates. The only exceptions are for α_1 and α_2 when σ_ζ^2 is low and $\beta_2 < 1.1$. Recall that when $\sigma_\zeta^2 = 0$, the ANCOVA estimates for β_1 and β_2 were unbiased while the CEM estimates showed some bias, particularly for values of β_2 that were close to 1. This suggests that in this case, there is a trade-off between unbiasedness and the variation in the estimates.

Table 2.3: Parameter RMSE for different sample sizes with $\sigma_\zeta^2 = \sqrt{2}$ using CEM and ANCOVA

		CEM RMSE				ANCOVA RMSE			
β_2	Parameter	n=30	n=100	n=500	n=1000	n=30	n=100	n=500	n=1000
2	α_1	3.554	2.088	1.012	0.723	3.545	3.293	3.238	3.224
	α_2	4.417	2.503	1.120	0.795	8.550	8.340	8.317	8.349
	β_1	0.354	0.209	0.101	0.072	0.831	0.823	0.820	0.820
	β_2	0.441	0.249	0.112	0.079	0.758	0.737	0.731	0.728
1.50	α_1	4.189	2.906	1.430	1.057	3.552	3.377	3.338	3.308
	α_2	4.817	3.077	1.509	1.085	6.528	6.458	6.378	6.369
	β_1	0.418	0.290	0.143	0.106	0.631	0.620	0.618	0.617
	β_2	0.479	0.307	0.151	0.108	0.575	0.553	0.547	0.546
1.10	α_1	4.381	3.390	2.493	1.986	3.630	3.481	3.417	3.412
	α_2	4.551	3.612	2.633	2.095	5.023	4.858	4.826	4.803
	β_1	0.437	0.339	0.249	0.199	0.512	0.494	0.487	0.489
	β_2	0.455	0.361	0.263	0.210	0.498	0.476	0.469	0.469
1.08	α_1	4.474	3.180	2.304	1.855	3.710	3.451	3.419	3.402
	α_2	4.746	3.449	2.454	1.964	4.945	4.764	4.736	4.739
	β_1	0.446	0.317	0.230	0.185	0.513	0.486	0.482	0.484
	β_2	0.474	0.345	0.245	0.196	0.496	0.468	0.467	0.467
1.06	α_1	4.358	3.300	2.039	1.724	3.726	3.487	3.429	3.412
	α_2	4.682	3.395	2.157	1.846	4.913	4.678	4.641	4.640
	β_1	0.437	0.330	0.204	0.172	0.494	0.486	0.480	0.477
	β_2	0.470	0.339	0.216	0.184	0.493	0.471	0.466	0.466
1.04	α_1	4.143	3.359	1.901	1.280	3.589	3.491	3.433	3.411
	α_2	4.476	3.444	2.026	1.375	4.716	4.638	4.573	4.575
	β_1	0.413	0.336	0.190	0.128	0.486	0.476	0.474	0.471
	β_2	0.446	0.344	0.203	0.137	0.479	0.473	0.467	0.464
1.02	α_1	4.354	3.372	1.713	1.088	3.741	3.458	3.439	3.422
	α_2	4.523	3.372	1.773	1.134	4.739	4.523	4.500	4.503
	β_1	0.434	0.338	0.171	0.109	0.490	0.472	0.469	0.468
	β_2	0.452	0.338	0.177	0.114	0.481	0.468	0.464	0.464
1	α_1	4.569	3.500	1.857	1.105	3.833	3.519	3.426	3.426
	α_2	4.477	3.496	1.911	1.150	4.653	4.445	4.424	4.428
	β_1	0.458	0.350	0.186	0.110	0.496	0.467	0.463	0.463
	β_2	0.446	0.350	0.191	0.115	0.482	0.468	0.464	0.464

Table 2.4: Parameter RMSE for different covariate error variance with $n = 500$ using CEM and ANCOVA

		CEM RMSE				ANCOVA RMSE			
β_2	Parameter	$\sigma_\zeta^2 = 0$	$\sigma_\zeta^2 = 0.5$	$\sigma_\zeta^2 = 1$	$\sigma_\zeta^2 = 2$	$\sigma_\zeta^2 = 0$	$\sigma_\zeta^2 = 0.5$	$\sigma_\zeta^2 = 1$	$\sigma_\zeta^2 = 2$
2	α_1	0.326	0.596	0.797	1.202	1.269	1.263	2.464	4.072
	α_2	0.386	0.705	0.927	1.379	0.782	4.088	6.712	10.023
	β_1	0.033	0.059	0.080	0.120	0.708	0.736	0.783	0.866
	β_2	0.038	0.070	0.092	0.137	1.225	0.928	0.786	0.709
1.50	α_1	0.413	0.885	1.148	1.745	1.077	1.309	2.540	4.190
	α_2	0.489	0.961	1.227	1.829	0.645	3.196	5.155	7.667
	β_1	0.041	0.088	0.115	0.175	0.559	0.551	0.583	0.662
	β_2	0.048	0.096	0.122	0.183	0.828	0.637	0.561	0.560
1.10	α_1	1.479	2.170	2.410	2.639	0.937	1.340	2.619	4.271
	α_2	1.603	2.348	2.581	2.784	0.598	2.465	3.944	5.756
	β_1	0.148	0.217	0.241	0.264	0.503	0.453	0.463	0.527
	β_2	0.160	0.234	0.258	0.278	0.555	0.463	0.453	0.503
1.08	α_1	1.563	1.973	2.241	2.383	0.946	1.348	2.636	4.258
	α_2	1.702	2.165	2.406	2.522	0.591	2.446	3.881	5.640
	β_1	0.156	0.197	0.224	0.238	0.501	0.448	0.460	0.524
	β_2	0.170	0.217	0.240	0.252	0.544	0.459	0.450	0.502
1.06	α_1	1.703	1.968	2.091	2.290	0.936	1.344	2.620	4.280
	α_2	1.868	2.215	2.243	2.392	0.588	2.453	3.808	5.534
	β_1	0.170	0.197	0.209	0.229	0.502	0.446	0.455	0.516
	β_2	0.187	0.221	0.224	0.239	0.533	0.450	0.449	0.501
1.04	α_1	1.730	1.840	1.881	2.039	0.937	1.366	2.638	4.256
	α_2	1.887	2.010	1.999	2.174	0.602	2.396	3.769	5.458
	β_1	0.173	0.184	0.188	0.204	0.501	0.441	0.449	0.511
	β_2	0.189	0.201	0.200	0.217	0.521	0.445	0.445	0.502
1.02	α_1	1.821	1.634	1.665	1.817	0.941	1.357	2.644	4.286
	α_2	1.878	1.729	1.769	1.850	0.569	2.346	3.702	5.376
	β_1	0.182	0.163	0.167	0.181	0.500	0.439	0.449	0.506
	β_2	0.188	0.173	0.177	0.185	0.513	0.441	0.446	0.500
1	α_1	1.733	1.617	1.780	1.786	0.907	1.380	2.645	4.285
	α_2	1.747	1.616	1.829	1.794	0.580	2.316	3.634	5.287
	β_1	0.173	0.162	0.178	0.179	0.500	0.435	0.443	0.502
	β_2	0.175	0.162	0.183	0.180	0.502	0.437	0.442	0.501

The case where $\sigma_\zeta^2 = 0$ is also the case where the CEM was most likely to fail. In the case where we are able to attain CEM estimates for a model without covariate measurement error, it may be worth comparing the model estimated by the CEM and the model estimated with ANCOVA to determine which is a better fit for the data.

Chapter 3: Using CEM and ANCOVA on the CrabShip Data

Now, we will demonstrate the use of the CEM and classic ANCOVA methods on a real data set. When using the ANCOVA approach, we will allow β_1 and β_2 to differ. We will use the CrabShip data in the Stat2Data package in R. Recall this data comes from an experiment to examine the rate of oxygen intake of crabs when exposed to different types of noise. In this experiment, the covariate measurement is the mass of the crabs. Crabs are randomly assigned to listen to ambient noise or noise from ships. Then both the mass of the crabs and the rate of oxygen intake are measured for each crab.

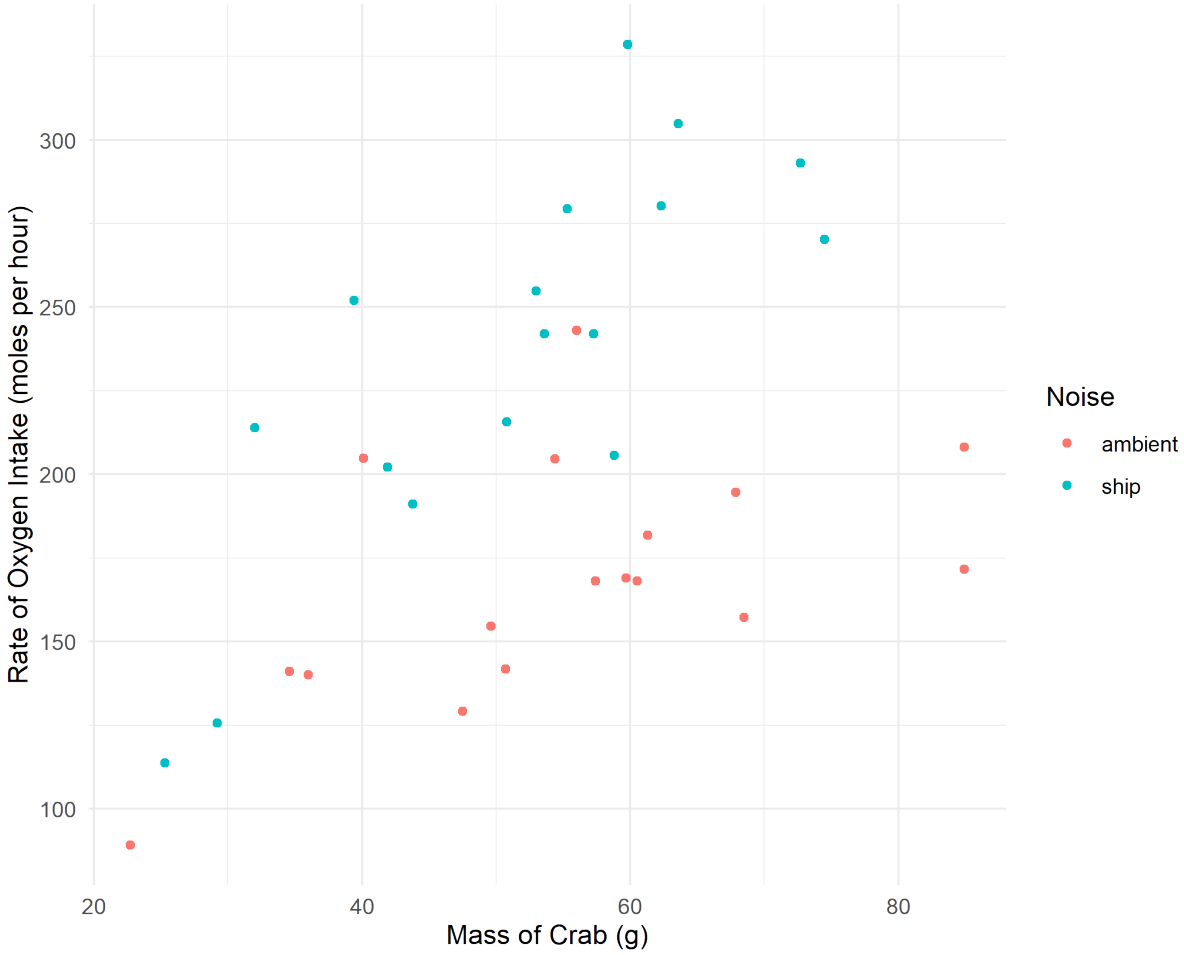

Figure 3.1: Plot of CrabShip data from the Stat2Data package in R

Using the CEM to estimate the parameters we obtain the following model:

$$E(Y_i) = \begin{cases} -23.254 + 3.548m_i, & \text{if the i-th crab is exposed to ambient noise} \\ -49.822 + 5.456m_i, & \text{if the i-th crab is exposed to ship noises} \end{cases}$$

where Y_i is the rate of oxygen intake of the i-th crab and m_i is the mass of the i-th crab. The CEM also provides estimates for the three variance parameters, $\sigma_x^2 = 100.25$, $\sigma_\epsilon^2 = 52.19$, and $\sigma_\zeta^2 = 124.97$.

Similarly, we can obtain the model using ANCOVA:

$$E(Y_i) = \begin{cases} 103.27 + 1.187m_i, & \text{if the i-th crab is exposed to ambient noise} \\ 68.88 + 3.257m_i, & \text{if the i-th crab is exposed to ship noises} \end{cases}$$

We can also visualize the models by plotting the lines over the raw data.

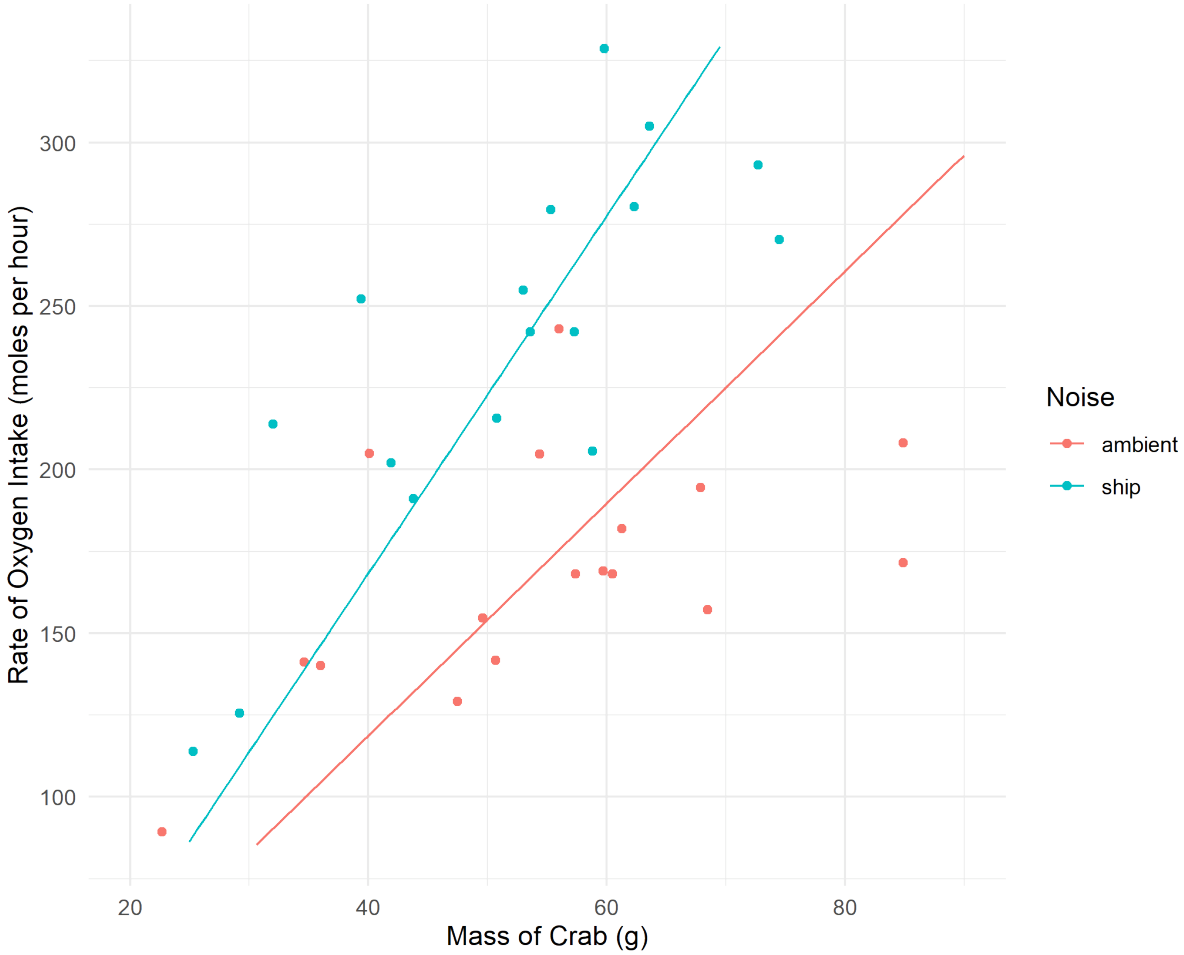

Figure 3.2: Plot of CrabShip data with the model estimated using CEM

We can see that the two procedures give very different estimates for the model parameters. Since in this

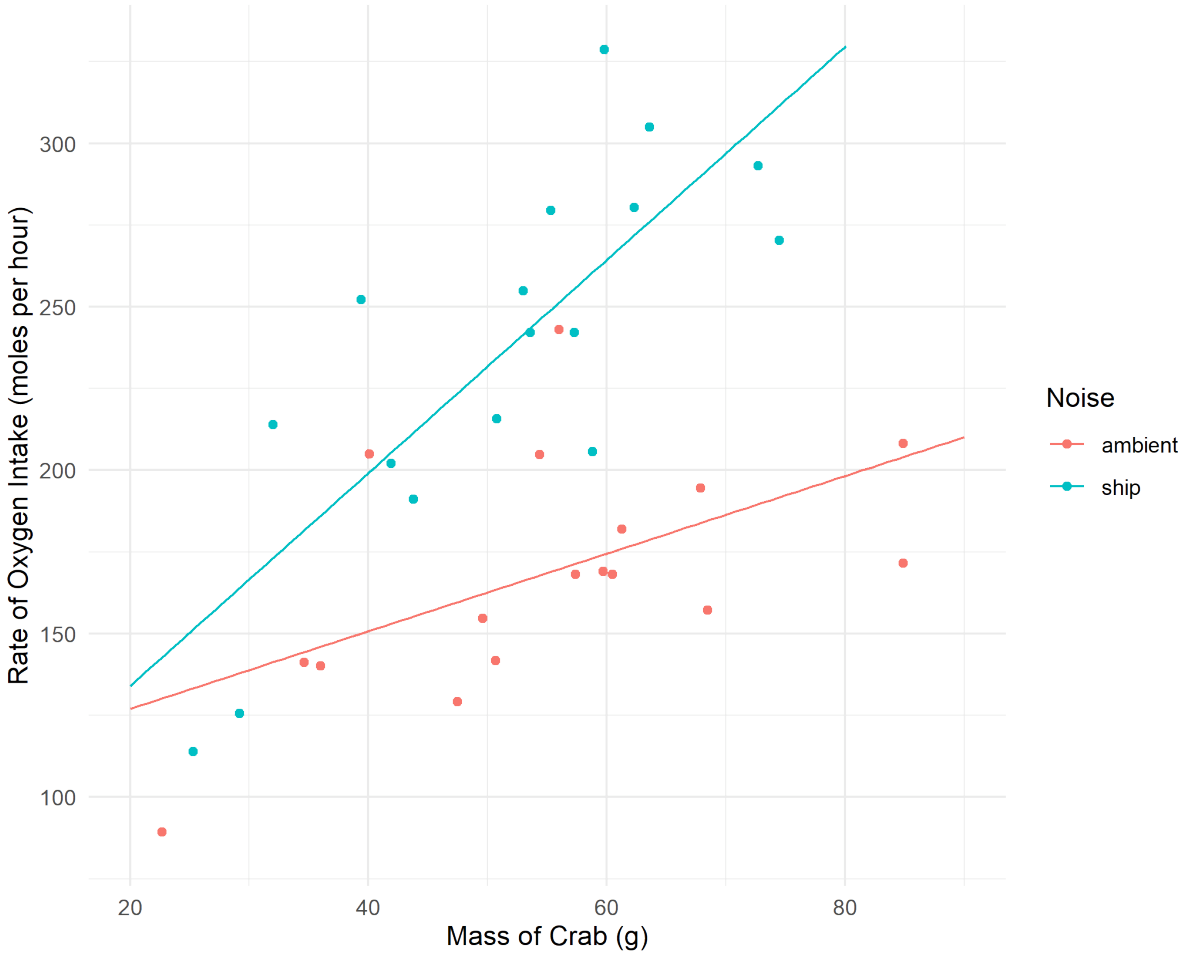

Figure 3.3: Plot of CrabShip data with the model estimated using ANCOVA

case we do not know the true value of the parameters, we cannot use that as a way to compare the models. Instead, we will use the AIC to compare the models. We can calculate the AIC $= -2 * \text{log-likelihood} + 2k$ where k is the number of parameters being estimated.

In order to accomplish this, we will reframe the classical ANCOVA model. Typically, we think of it modeling Y_{ig} given a value for Z_{ig} under the assumption that $Z_{ig} = X_{ig}$. Equivalently, we can think of ANCOVA as the joint distribution of Y_{ig} and Z_{ig} as defined in the CEM but with $\sigma^2_\zeta = 0$. So, we can use the same likelihood function that we used for the CEM, but instead of estimating σ^2_ζ, we set $\sigma^2_\zeta = 0$. We can then optimize the log-likelihood function. This produces estimates for α_1, α_2, β_1, and β_2 which are nearly identical to those produced by lm().

Since the optim() function returns the value of the log-likelihood function, it then becomes easy to calculate the AIC. Note that for the CEM, $k = 8$, since in addition to the α and β parameters, we are

also estimating the covariate mean and the three variance parameters. For ANCOVA, we are estimating one fewer parameter, so $k = 7$. The AIC for the CEM is 507.32 and the AIC for the ANCOVA model is 502.06. The ANCOVA model has a lower AIC than the CEM which would suggest that the model estimated by ANCOVA is a better fit for the data. However, the difference in AIC's is 5.26 which is small enough for there to be some ambiguity.

We know that in the case that there is no covariate error, ANCOVA produces unbiased estimates and would be the better option for estimating the model. However, if there is covariate error, the CEM produces less biased estimates. In this case, if the researchers feel confident that the measurement of crab mass has little to no error, then they might decide to use the ANCOVA estimates. However, if they think that there is likely some measurement error, then they should choose the CEM estimates.

Chapter 4: Summary and Conclusions

4.1 Limitations

There are a few limitations to the work done with the CEM in this thesis. When estimating the CEM estimates, the optimization of the log-likelihood function was not always successful. It is unclear what causes this issue to occur. Due to the error messages in R, it is clear that the problem is the variance-covariance matrix becomes singular at some point in the optimization algorithm. One likely cause, particularly in the case where $\sigma_\zeta^2 = 0$, is that the estimate for one or more of the variance parameters approaches 0.

We showed that the CEM is not identified in the case that $\beta_1 = \beta_2$. Though the estimates obtained in the case where $\beta_1 = \beta_2$ were generally less biased than the ANCOVA estimates except when $\sigma_\zeta^2 = 0$, accurate parameter estimates are not guaranteed in this case.

4.2 Future Research

There are several opportunities for future research related to the CEM. One area for possible future research is expanding the CEM to situations where $G > 2$. Similarly, there is potential for investigating the case where there are two or more covariates measured with error. These types of questions occur frequently and being able to apply the CEM in these cases could be beneficial.

Additionally, we saw that when $\sigma_\zeta^2 = 0$, the classic ANCOVA estimates were unbiased, while the CEM estimates had some bias. By the time $\sigma_\zeta^2 = 0.5$, however, classic ANCOVA was already showing more bias than the CEM. Investigating more values of σ_ζ^2 could show just how sensitive to covariate error classic ANCOVA is and where the CEM estimates become less biased than classic ANCOVA estimates. This could help give some guidance in the case where there is very little covariate error but the error is likely still present.

4.3 Conclusion

Overall, the CEM seems to be a good way to account for covariate measurement error in ANCOVA. In most cases, when error is present the CEM shows less bias and lower RMSE. However, the method has some significant drawbacks in that the optimization algorithm sometimes fails and the CEM is not identified when $\beta_1 = \beta_2$. Interestingly, when the algorithm works, it still produces parameter estimates with less bias than classic ANCOVA estimates in the case the $\beta_1 = \beta_2$ so long as there is covariate measurement error. When there is no covariate measurement error, the CEM produces slightly biased

estimates while the ANCOVA estimates are unbiased, as we would expect. In the case that there is little or no covariate error, we should continue to use classic ANCOVA.

References

[1] John P Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

[2] Ann Cannon, George Cobb, Bradley Hartlaub, Julie Legler, Robin Lock, Thomas Moore, Allan Rossman, and Jeffrey Witmer. *Stat2Data: Datasets for Stat2*, 2019. R package version 2.0.0.

[3] Steven Andrew Culpepper and Herman Aguinis. Using analysis of covariance (ancova) with fallible covariates. *Psychological Methods*, 16(2):166, 2011.

[4] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

[5] SD Hodges and PG Moore. Data uncertainties and least squares regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):185–195, 1972.

[6] JT Kent, John Bibby, and KV Mardia. *Multivariate analysis*. Academic Press Amsterdam, 1979.

[7] JR Lockwood and Daniel F McCaffrey. Correcting for test score measurement error in ancova models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1):22–52, 2014.

[8] Malcolm James Ree and Thomas R Carretta. The role of measurement error in familiar statistics. *Organizational Research Methods*, 9(1):99–112, 2006.

[9] Matthew A Wale, Stephen D Simpson, and Andrew N Radford. Size-dependent physiological responses of shore crabs to single and repeated playback of ship noise. *Biology letters*, 9(2):20121194, 2013.

Appendix A: R Code

```
#The log of the multivariate normal pdf for use in the
#log-likelihood function
nlogpdf <- function(n,y,S,m,C){
  R <- solve(C)
  n/2*(t(y-m) %*% R %*% (y-m) - log(det(R)) + sum(diag(S %*% R)))
}

#The negative log-likelihood function
nloglik <- function(theta){
  alph1 <- theta[1]
  alph2 <- theta[2]
  beta1 <- theta[3]
  beta2 <- theta[4]
  delta <- theta[5]
  sigmx <- theta[6]
  sigme <- theta[7]
  sigmz <- theta[8]

  m1 <- c(alph1 + beta1*delta, delta)
  m2 <- c(alph2 + beta2*delta, delta)

  C1 <- matrix(NA,2,2)
  C1[1,1] <- beta1^2*sigmx + sigme
  C1[1,2] <- beta1*sigmx
  C1[2,1] <- beta1*sigmx
  C1[2,2] <- sigmx+sigmz

  C2 <- matrix(NA,2,2)
  C2[1,1] <- beta2^2*sigmx + sigme
  C2[1,2] <- beta2*sigmx
```

```
  C2[2,1] <- beta2*sigmx
  C2[2,2] <- sigmx+sigmz

  return(nlogpdf(n1,y1,S1,m1,C1)+nlogpdf(n2,y2,S2,m2,C2))
}

#Simulate data, estimate parameters, and store the estimates
for(i in 1:iter){
  mydata <- expand.grid(x=rnorm(n,delta,sqrt(sigmax)),group=c("a","b")) %>%
    mutate(epsilon=rnorm(n(),0,sqrt(sigmae))) %>%
    mutate(y=case_when(
      group=="a"~alpha1+beta1*x+epsilon,
      group=="b"~alpha2+beta2*x+epsilon
    )) %>%
    mutate(z=x+rnorm(n(),0,sqrt(sigmaz))) %>% select(-epsilon)

  y1 <- with(mydata, cbind(y[group=="a"],z[group=="a"]))
  n1 <- nrow(y1)
  S1 <- cov(y1) * (n1-1)/n1
  y1 <- apply(y1,2,mean)

  y2 <- with(mydata, cbind(y[group=="b"],z[group=="b"]))
  n2 <- nrow(y2)
  S2 <- cov(y2) * (n2-1)/n2
  y2 <- apply(y2,2,mean)

  theta.hat <- rep(1,8)
  names(theta.hat) <- c("alpha1","alpha2","beta1","beta2",
  "delta","sigmax","sigmae","sigmaz")

  tmp <- try(optim(theta.hat, nloglik, method="L-BFGS-B",
                   lower= c(rep(-Inf, 5), rep(0,3)),
```

```
                    control = list(maxit = 10000)))

  if(class(tmp)!="try-error"){
    x <- as.vector(tmp$par)
  }else{x<-c(NA,NA,NA,NA,NA,NA,NA,NA)}
  par.est[i,] = x

  mz <- lm(y ~ -1+group+group:z,data=mydata)
  lmz.est[i,]=mz$coefficients

  par.diff[i,] = x-theta
}
```