# Simulation Study of 2-Variable Measurement Error in a Linear Regression Model

Kevin Yan

Department of Statistics
University of Toronto

July 2015

# Outline

# Outline

# Introduction

- In this report, we will be looking at the effect of measurement error on a 2 variable, linear regression model.

# Introduction

- In this report, we will be looking at the effect of measurement error on a 2 variable, linear regression model.
- Linear regression was chosen because it is probably the most well-known and used model in statistics, hence it is important and useful to see what effect measurement error will have on the outcomes of linear regression models.

# Introduction

- In this report, we will be looking at the effect of measurement error on a 2 variable, linear regression model.
- Linear regression was chosen because it is probably the most well-known and used model in statistics, hence it is important and useful to see what effect measurement error will have on the outcomes of linear regression models.
- In addition, we decide to study the case with 2 linear predictors that have varying degrees of measurement error. This allows us to study the some effects that may intuitively be interesting, and may occur in real studies, where measurements are required.

# Outline

## 2-Variable Model:

- In the second half of the study, data will be generated from the following model
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \epsilon$$

# 2-Variable Model:

- In the second half of the study, data will be generated from the following model
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, C)$$
  $$C = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix}$$

# 2-Variable Model:

- In the second half of the study, data will be generated from the following model
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, C)$$
  $$C = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix}$$
- Throughout the entire study, we will set $\sigma_1^2, \sigma_2^2$=5, and the correlation between each predictor variable will be 0.2

# 2-Variable Model:

- In the second half of the study, data will be generated from the following model
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, C)$$
  $$C = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix}$$
- Throughout the entire study, we will set $\sigma_1^2, \sigma_2^2$=5, and the correlation between each predictor variable will be 0.2
- Thus we see that $\sigma_{1,2} = 0.2 * 5 = 1 = \sigma_{2,1}$

# 2-Variable Model:

- In the second half of the study, data will be generated from the following model
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, C)$$
$$C = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix}$$
- Throughout the entire study, we will set $\sigma_1^2, \sigma_2^2$=5, and the correlation between each predictor variable will be 0.2
- Thus we see that $\sigma_{1,2} = 0.2 * 5 = 1 = \sigma_{2,1}$
- 100 pairs of Xs will be generated from the following model for each sample, with different settings for $\mu_1$ and $\mu_2$(among other variables described later), and we will fit a regression model onto each sample. This process will be repeated 10,000

# 3-variable Model:

- In the first part of study, we will be looking at data that will be generated from the following model:
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \epsilon$$

# 3-variable Model:

- In the first part of study, we will be looking at data that will be generated from the following model:
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 \end{bmatrix})$$

# 3-variable Model:

- In the first part of study, we will be looking at data that will be generated from the following model:
- $$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \epsilon$$
- $$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim Normal(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 \end{bmatrix})$$
- Same as before, the correlations between all 3 predictor variables $= 0.2$ and thus all $\sigma_{i,j} = 1$

- However the problem that we consider involves measurement error, and so we will introduce measurement error using the following model:

- However the problem that we consider involves measurement error, and so we will introduce measurement error using the following model:
- $$X_{obs1} = X_1 + u_1$$
$$X_{obs_2} = X_2 + u_2$$

- However the problem that we consider involves measurement error, and so we will introduce measurement error using the following model:
- $$X_{obs1} = X_1 + u_1$$
  $$X_{obs_2} = X_2 + u_2$$
- $$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim Normal(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, U = \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix})$$

- However the problem that we consider involves measurement error, and so we will introduce measurement error using the following model:
- $$X_{obs1} = X_1 + u_1$$
  $$X_{obs_2} = X_2 + u_2$$
- $$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim Normal(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, U = \begin{bmatrix} \sigma^2_{u1} & \sigma_{u12} \\ \sigma_{u21} & \sigma^2_{u2} \end{bmatrix})$$
- Note: Correlation between measurement errors will be denoted as $u_{12}$ and thus $\sigma_{u12} = \sigma_{u1} * \sigma_{u2} * u_{12}$

# Outline

# What parameters will be changing? (Correlation of Measurement Error)

- The first parameter we will be looking at will be the correlation between the measurement errors (ie: $\sigma_{u12}$)

# What parameters will be changing? (Correlation of Measurement Error)

- The first parameter we will be looking at will be the correlation between the measurement errors (ie: $\sigma_{u12}$)
- First of all, we will look at the case where the correlation between $u_1$ and $u_2$ is zero, which in this case, will mean that the measurement errors are independent. This is a very likely scenario, because it's possible that we can be measuring these variables in isolation of each other, and with different tools, thus removing any relationship between the measurement errors.

# What parameters will be changing? (Correlation of Measurement Error)

- The first parameter we will be looking at will be the correlation between the measurement errors (ie: $\sigma_{u12}$)
- First of all, we will look at the case where the correlation between $u_1$ and $u_2$ is zero, which in this case, will mean that the measurement errors are independent. This is a very likely scenario, because it's possible that we can be measuring these variables in isolation of each other, and with different tools, thus removing any relationship between the measurement errors.
- Secondly, it's possible that we will be looking at the case where there is large positive correlation between the measurement errors. This may be the case where we measure both variables on the same subject, and with the same instrument.

- Thirdly, to relax the scenario above, we can have low/medium correlation between the measurement errors. This scenario may occur when perhaps we use different instruments, but some environmental aspects are the same, and hence will result in some sort of linear relationship between the measurement errors.

# What parameters will be changing? (Beta-Discrepancy)

- The second parameter we will take a look at will be the settings of betas. We will only be altering the level of $\beta_2$, which we can look at as increasing the difference between the levels of the betas in the model. Thus we will vary $\beta_2$ through the values 5,10,20.

# What parameters will be changing? (Beta-Discrepancy)

- The second parameter we will take a look at will be the settings of betas. We will only be altering the level of $\beta_2$, which we can look at as increasing the difference between the levels of the betas in the model. Thus we will vary $\beta_2$ through the values 5,10,20.
- The settings we will use for $\beta_1$ will be fixed to 2, this is so we can more easily see the effect of what we're doing in our ability to estimate, and perform hypothesis testing under the regression model.

# What parameters will be changing? (Beta-Discrepancy)

- The second parameter we will take a look at will be the settings of betas. We will only be altering the level of $\beta_2$, which we can look at as increasing the difference between the levels of the betas in the model. Thus we will vary $\beta_2$ through the values 5,10,20.
- The settings we will use for $\beta_1$ will be fixed to 2, this is so we can more easily see the effect of what we're doing in our ability to estimate, and perform hypothesis testing under the regression model.
- $\beta_3$ will also be set to 0, in order to be able to observe Type I Error of the t-test. (Note: Again we only have the 3rd predictor variable for the 1st part of the study)

# What parameters will be changing? (Beta-Discrepancy)

- The second parameter we will take a look at will be the settings of betas. We will only be altering the level of $\beta_2$, which we can look at as increasing the difference between the levels of the betas in the model. Thus we will vary $\beta_2$ through the values 5,10,20.
- The settings we will use for $\beta_1$ will be fixed to 2, this is so we can more easily see the effect of what we're doing in our ability to estimate, and perform hypothesis testing under the regression model.
- $\beta_3$ will also be set to 0, in order to be able to observe Type I Error of the t-test. (Note: Again we only have the 3rd predictor variable for the 1st part of the study)
- We can interpret the change in betas, as altering the fundamental relationship between the response variable and the predictor variables holding all else equal (ie: values of predictor variables and their units)

# What parameters will be changing? (Beta-Discrepancy)

- The second parameter we will take a look at will be the settings of betas. We will only be altering the level of $\beta_2$, which we can look at as increasing the difference between the levels of the betas in the model. Thus we will vary $\beta_2$ through the values 5,10,20.
- The settings we will use for $\beta_1$ will be fixed to 2, this is so we can more easily see the effect of what we're doing in our ability to estimate, and perform hypothesis testing under the regression model.
- $\beta_3$ will also be set to 0, in order to be able to observe Type I Error of the t-test. (Note: Again we only have the 3rd predictor variable for the 1st part of the study)
- We can interpret the change in betas, as altering the fundamental relationship between the response variable and the predictor variables holding all else equal (ie: values of predictor variables and their units)
- However, it just of interest just to see if becomes harder to predict the values of betas as we change them, and if it will be more difficult to perform hypothesis tests.

# What parameters will be changing? (Mean-Discrepancy)

- The third parameter we will be changing will be the mean of the second predictor variable, which we can interpret as changing the difference in size between the measurements of the 3 predictor variables

# What parameters will be changing? (Mean-Discrepancy)

- The third parameter we will be changing will be the mean of the second predictor variable, which we can interpret as changing the difference in size between the measurements of the 3 predictor variables
- The means of $X_1$ and $X_3$ will be fixed at 10, while the mean of $X_2$ will vary between 20 and 100.

# What parameters will be changing? (Mean-Discrepancy)

- The third parameter we will be changing will be the mean of the second predictor variable, which we can interpret as changing the difference in size between the measurements of the 3 predictor variables
- The means of $X_1$ and $X_3$ will be fixed at 10, while the mean of $X_2$ will vary between 20 and 100.
- Practically this can be seen as the effect of changing the units of particular variable. We will then observe how this change can affect our ability to predict betas, and our ability to test our estimates.

# Outline

# Overview

- In this section, we will be generating data from a 3-variable model We will then try to fit a linear regression model onto this data.

# Overview

- In this section, we will be generating data from a 3-variable model We will then try to fit a linear regression model onto this data.
- we will try to understand the effect of changing the size of the betas, size of the means of the predictor variables, as well as the magnitude of correlation between the measurement errors on the characteristics of interest described before. These being: power, type I error and bias

# Overview

- In this section, we will be generating data from a 3-variable model We will then try to fit a linear regression model onto this data.
- we will try to understand the effect of changing the size of the betas, size of the means of the predictor variables, as well as the magnitude of correlation between the measurement errors on the characteristics of interest described before. These being: power, type I error and bias
- In total we will have $2 * 3 * 3 = 18$ combinations, corresponding the the means, betas and correlation respectively.

# Overview

- In this section, we will be generating data from a 3-variable model We will then try to fit a linear regression model onto this data.
- we will try to understand the effect of changing the size of the betas, size of the means of the predictor variables, as well as the magnitude of correlation between the measurement errors on the characteristics of interest described before. These being: power, type I error and bias
- In total we will have $2 * 3 * 3 = 18$ combinations, corresponding the the means, betas and correlation respectively.
- An increase in the discrepancy of the means (at least not an increase of 90 units) doesn't appear to have any significant/noticeable impact on the properties we are interested in.

# Slope X1

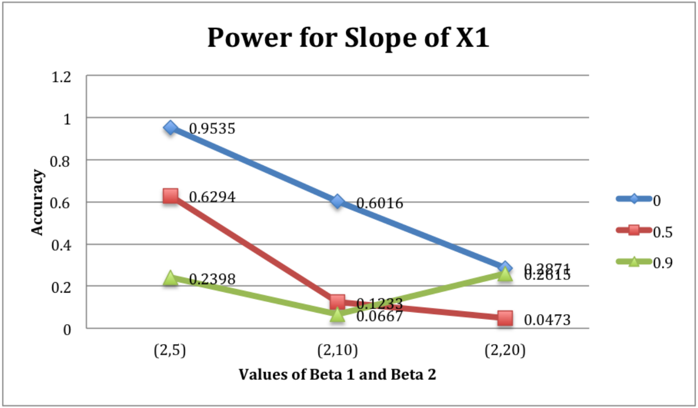
- 


Figure: Power of the test for slope of X1

# Slope X1

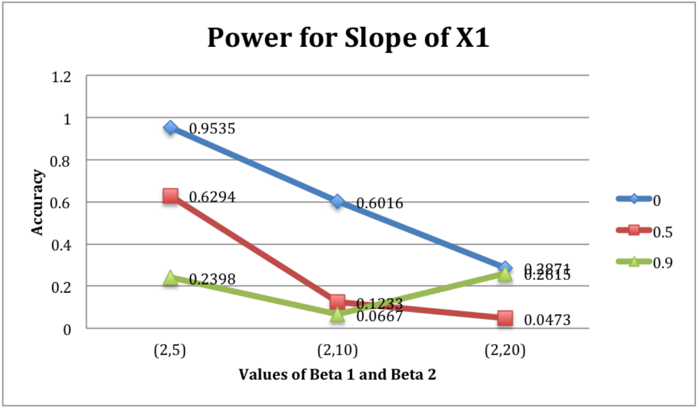
- 


Figure: Power of the test for slope of X1

- We see higher power for lower levels of measurement error correlation(in general).

# Slope X1

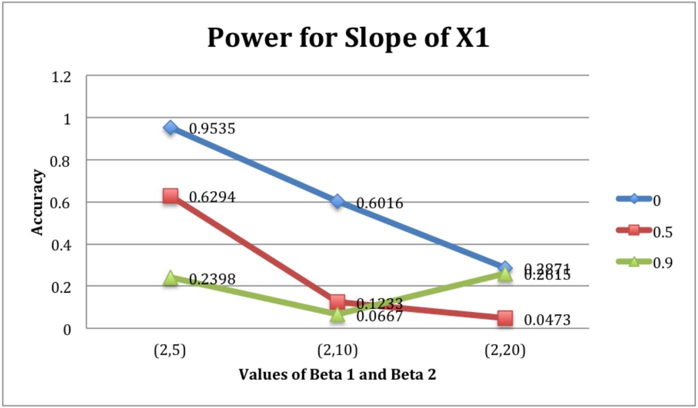
- 


Figure: Power of the test for slope of X1

- We see higher power for lower levels of measurement error correlation(in general).
- We are able to see that there is tendency for power to decrease as we increase the $\sigma_{u12}$ (with the exception of $\sigma_{u12} = 0.9$, which drops but bounces back up).

# Slope X1

- We see higher power for lower levels of measurement error correlation(in general).

# Slope X1

- We see higher power for lower levels of measurement error correlation(in general).
- We are able to see that there is tendency for power to decrease as we increase the $\sigma_{u12}$ (with the exception of $\sigma_{u12} = 0.9$, which drops but bounces back up).

# Slope X1

- We see higher power for lower levels of measurement error correlation(in general).
- We are able to see that there is tendency for power to decrease as we increase the $\sigma_{u12}$ (with the exception of $\sigma_{u12} = 0.9$, which drops but bounces back up).
- We see that slight bounce back because in general the estimates/bias(actual sign value) is decreasing with $\beta_2$. Eventually at 20, the estimate is negative enough to make the model believe that it's not that likely to be 0.

# Slope X3

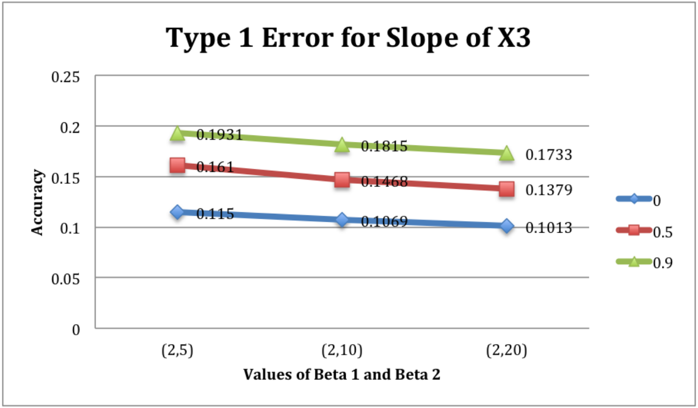



Figure: Type I Error for Slope of $X_3$

# Slope X3

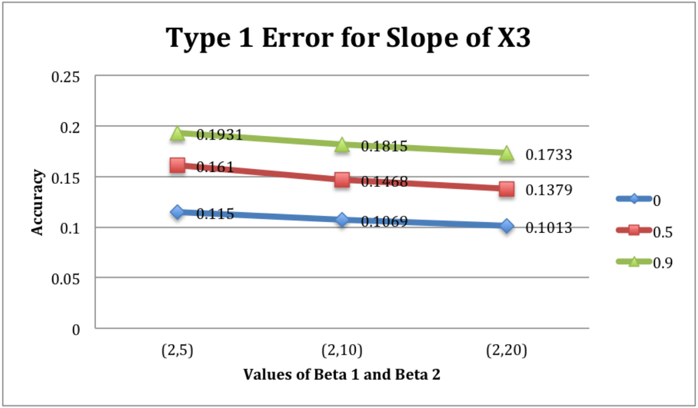
- 


Figure: Type I Error for Slope of $X_3$

- We see that as we increase correlation, that the level of Type I Error increases.

# Slope X3

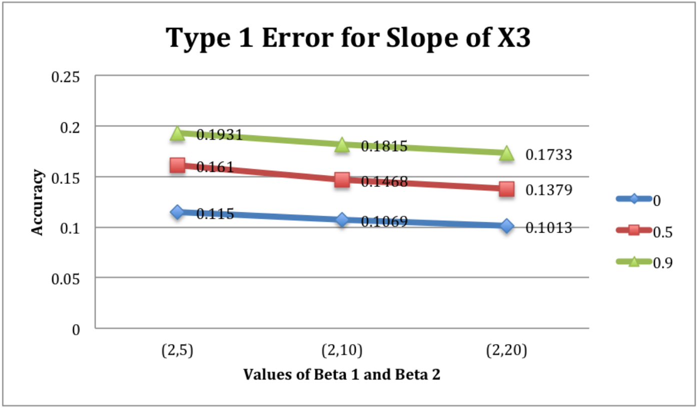
- 


Figure: Type I Error for Slope of $X_3$

- We see that as we increase correlation, that the level of Type I Error increases.
- we're able to see that Type I Error actually tends to drop as we increase the the discrepancy in betas.

# Quick Explanation

What happens here is that as we increase $\beta_2$, the variability of the bias/estimates increases for A3, thus it becomes more likely that the test believes 0 is a possible value. This is the case even though bias is increasing along with $\beta_2$

# Total Bias

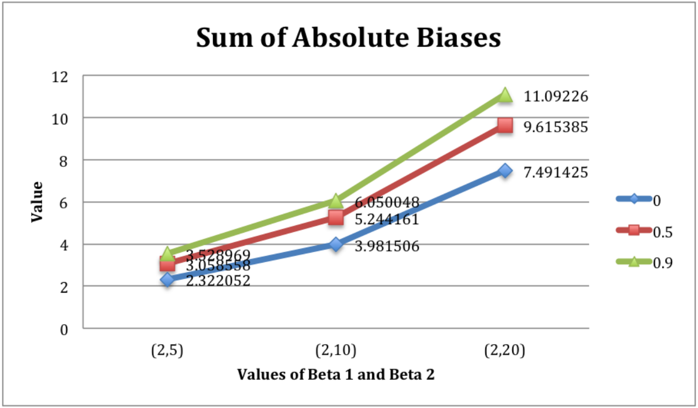
- 


Figure: Magnitude of all bias

# Total Bias

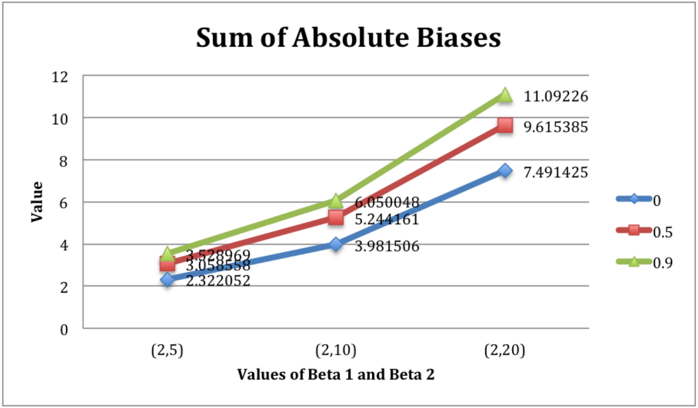
- 


Figure: Magnitude of all bias

- As we increase $\sigma_{u12}$, the total bias(sum of the absolute values of BIAS1,BIAS2,BIAS3) also increase.

## ...

- Additionally, it appears that if we increase the discrepancy of the betas, that the sum of the absolute bias of our estimates actually increases. In other words, as we increase the value of $\beta_2$, the amount of bias exhibited by our estimates increases.

- Additionally, it appears that if we increase the discrepancy of the betas, that the sum of the absolute bias of our estimates actually increases. In other words, as we increase the value of $\beta_2$, the amount of bias exhibited by our estimates increases.
- Moreover, the increase seems more significant as we move from (2,10) to (2,20), compared to the increase from (2,5) to (2,20). Thus it appears as though bigger $\beta_2$s, and hence bigger beta-deviations result in greater increases in bias.

# Summary

- In general, as we increase the correlation between the measurement errors, we observe a decrease in the power of the tests for the slope of the first predictor variable, as well as an increase in Type I Error associated with the test for the slope of third predictor variable. In addition, the biases get larger with the change in correlation as well.

# Summary

- In general, as we increase the correlation between the measurement errors, we observe a decrease in the power of the tests for the slope of the first predictor variable, as well as an increase in Type I Error associated with the test for the slope of third predictor variable. In addition, the biases get larger with the change in correlation as well.
- Increasing correlation in measurement error results in multi-colinearity problems in $X_{obs}$, and the estimates of linear regression.

# Summary

- In general, as we increase the correlation between the measurement errors, we observe a decrease in the power of the tests for the slope of the first predictor variable, as well as an increase in Type I Error associated with the test for the slope of third predictor variable. In addition, the biases get larger with the change in correlation as well.
- Increasing correlation in measurement error results in multi-colinearity problems in $X_{obs}$, and the estimates of linear regression.
- However the picture of how increasing discrepancy in betas is not as clear.
  - In terms of Type 1 Error, we actually do see that it tends to decrease as the beta discrepancy enlarges.
  - On the other hand, the power of the test of the slope of $X_{obs1}$ tends to decrease when we increase the value of $\beta_2$. Also, the sum of absolute biases is shown to increase as a result of increasing $\beta_2$

# Summary

- In general, as we increase the correlation between the measurement errors, we observe a decrease in the power of the tests for the slope of the first predictor variable, as well as an increase in Type I Error associated with the test for the slope of third predictor variable. In addition, the biases get larger with the change in correlation as well.
- Increasing correlation in measurement error results in multi-colinearity problems in $X_{obs}$, and the estimates of linear regression.
- However the picture of how increasing discrepancy in betas is not as clear.
  - In terms of Type 1 Error, we actually do see that it tends to decrease as the beta discrepancy enlarges.
  - On the other hand, the power of the test of the slope of $X_{obs1}$ tends to decrease when we increase the value of $\beta_2$. Also, the sum of absolute biases is shown to increase as a result of increasing $\beta_2$
- Thus we conclude that overall, increase the beta-discrepancy (ie: increasing $\beta_2$ results in poorer performance in power and biases, but better performance in the level of type I error.

# Outline

# intro

- In this section we will switch over to the 2-variable model.

# intro

- In this section we will switch over to the 2-variable model.
- What we want to do here is try to use the estimates made under the model with $X_{obs}$ and recover the true betas.
  Below, we will derive the equations in which to do so:

# intro

- In this section we will switch over to the 2-variable model.
- What we want to do here is try to use the estimates made under the model with $X_{obs}$ and recover the true betas.
  Below, we will derive the equations in which to do so:
- First we must assume that there exists a linear model st:
  $Y = \alpha_0 + \alpha_1 X_{obs1} + \alpha_2 X_{obs2} + \delta$

# intro

- In this section we will switch over to the 2-variable model.
- What we want to do here is try to use the estimates made under the model with $X_{obs}$ and recover the true betas.
  Below, we will derive the equations in which to do so:
- First we must assume that there exists a linear model st:
  $Y = \alpha_0 + \alpha_1 X_{obs1} + \alpha_2 X_{obs2} + \delta$
- $$\begin{aligned} cov(Y, X_{obs1}) &= cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, X_1 + u_1) \\ &= \beta_1 \sigma_1^2 + \beta_2 \sigma_{1,2} \\ cov(Y, X_{obs1}) &= cov(\alpha_0 + \alpha_1 X_{obs1} + \alpha_2 X_{obs2} + \delta, X_{obs1}) \\ &= \alpha_1 \sigma_{obs1}^2 + \alpha_2 \sigma_{obs1,obs2} \end{aligned}$$

# intro

- In this section we will switch over to the 2-variable model.
- What we want to do here is try to use the estimates made under the model with $X_{obs}$ and recover the true betas.
  Below, we will derive the equations in which to do so:
- First we must assume that there exists a linear model st:
  $Y = \alpha_0 + \alpha_1 X_{obs1} + \alpha_2 X_{obs2} + \delta$
- 
$$\begin{aligned} cov(Y, X_{obs1}) &= cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, X_1 + u_1) \\ &= \beta_1 \sigma_1^2 + \beta_2 \sigma_{1,2} \\ cov(Y, X_{obs1}) &= cov(\alpha_0 + \alpha_1 X_{obs1} + \alpha_2 X_{obs2} + \delta, X_{obs1}) \\ &= \alpha_1 \sigma_{obs1}^2 + \alpha_2 \sigma_{obs1,obs2} \end{aligned}$$

- 
$$\implies \beta_1 \sigma_1^2 + \beta_2 \sigma_{1,2} = \alpha_1 \sigma_{obs1}^2 + \alpha_2 \sigma_{obs1,obs2}$$
  by symmetry we also have the following:
$$\implies \beta_2 \sigma_2^2 + \beta_1 \sigma_{1,2} = \alpha_2 \sigma_{obs2}^2 + \alpha_1 \sigma_{obs1,obs2}$$

# Calculations Continued

- $$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
$$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
$$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 + \sigma_{u1}^2 & \sigma_{1,2} + \sigma_{u1,u2} \\ \sigma_{1,2} + \sigma_{u1,u2} & \sigma_2^2 + \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

# Calculations Continued

- $$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
  $$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
  $$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 + \sigma_{u1}^2 & \sigma_{1,2} + \sigma_{u1,u2} \\ \sigma_{1,2} + \sigma_{u1,u2} & \sigma_2^2 + \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

- $$ICC_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{u1}^2} \implies \sigma_1^2 = ICC_1(\sigma_1^2 + \sigma_{u1}^2)$$
  Equivalently:
  $$ICC_2 = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_{u2}^2} \implies \sigma_2^2 = ICC_2(\sigma_2^2 + \sigma_{u2}^2)$$

# Calculations Continued

- $$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
$$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{obs1}^2 & \sigma_{obs1,obs2} \\ \sigma_{obs1,obs2} & \sigma_{obs2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$
$$\implies \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma 1,2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 + \sigma_{u1}^2 & \sigma_{1,2} + \sigma_{u1,u2} \\ \sigma_{1,2} + \sigma_{u1,u2} & \sigma_2^2 + \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

- $$ICC_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{u1}^2} \implies \sigma_1^2 = ICC_1(\sigma_1^2 + \sigma_{u1}^2)$$

  Equivalently:
$$ICC_2 = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_{u2}^2} \implies \sigma_2^2 = ICC_2(\sigma_2^2 + \sigma_{u2}^2)$$

- $$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} ICC_1(\sigma_1^2 + \sigma_{u1}^2) & \sigma_{1,2} \\ \sigma 1,2 & ICC_2(\sigma_2^2 + \sigma_{u2}^2) \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 + \sigma_{u1}^2 & \sigma_{1,2} + \sigma_{u1,u2} \\ \sigma_{1,2} + \sigma_{u1,u2} & \sigma_2^2 + \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

# 2-variable model observations

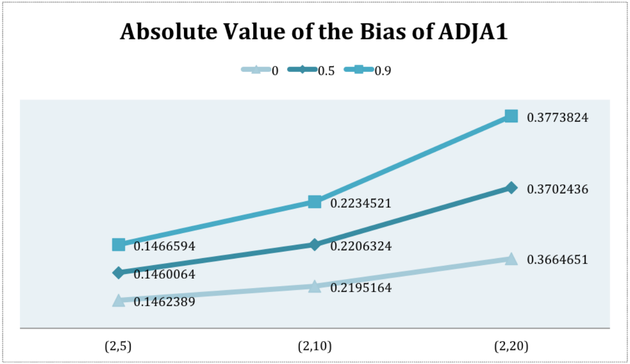
- 


Figure: ADJA1 BIAS

# 2-variable model observations

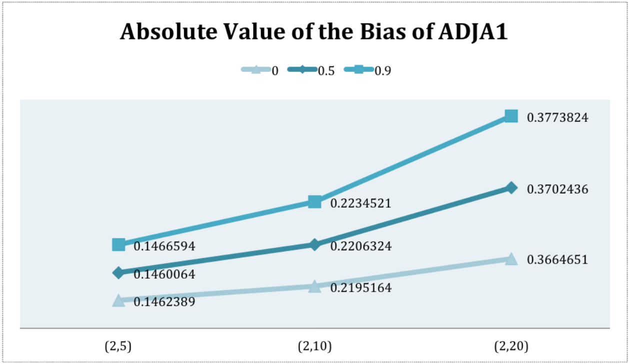
- 


Figure: ADJA1 BIAS

- In general we see that as we increase correlation, the bias tends to increase along with it (with the slight exception of (2,5)). Therefore as we increase the amount of correlation in the measurement errors, it becomes harder for us to actually estimate the true value of $\beta_1$ using the adjustment mechanism in section 4.1.

# ...

Moreover, we see that as we increase the beta-discrepancy, the level of bias will also increase. Thus as the difference between the betas enlarges, recovering the actual value of $\beta_1$ using our estimates $\hat{\alpha}$ from the model under $X_{obs}$ becomes more difficult.

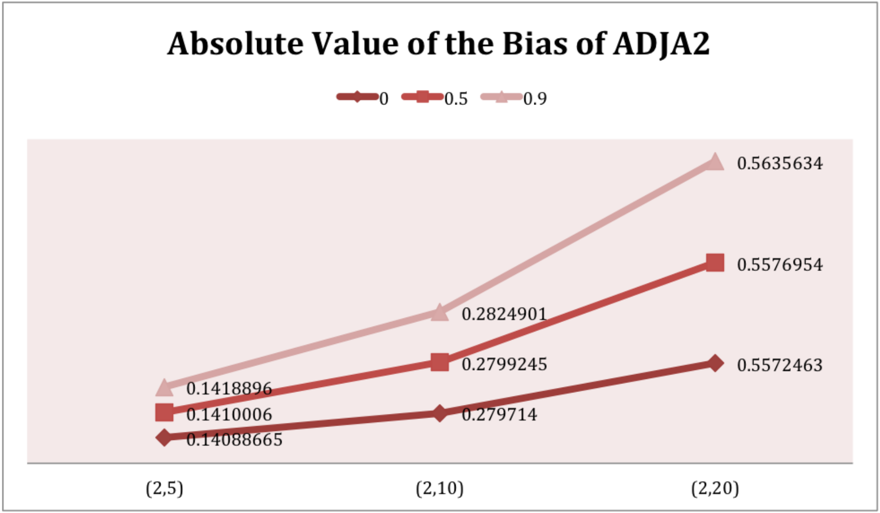



Figure: ADJA2 BIAS

## …

- Here we notice that in all settings of the betas, the bias increases as we increase the measurement error correlation. Initially in section 3, we noticed that as we increased correlation, the bias of the estimates increased as well

- Here we notice that in all settings of the betas, the bias increases as we increase the measurement error correlation. Initially in section 3, we noticed that as we increased correlation, the bias of the estimates increased as well
- It makes sense that the worse the estimates get, the adjustments we make on those estimates will also follow a similar trend.

# ...

- Here we notice that in all settings of the betas, the bias increases as we increase the measurement error correlation. Initially in section 3, we noticed that as we increased correlation, the bias of the estimates increased as well
- It makes sense that the worse the estimates get, the adjustments we make on those estimates will also follow a similar trend.
- Next, we notice that within different settings of betas, the average value of the biases increase as well. We see that for the low setting(of beta discrepancy), he values hover around 0.14, while as we increase the discrepancy we get values around 0.28 and finally around 0.55 for the highest setting.

- Here we notice that in all settings of the betas, the bias increases as we increase the measurement error correlation. Initially in section 3, we noticed that as we increased correlation, the bias of the estimates increased as well
- It makes sense that the worse the estimates get, the adjustments we make on those estimates will also follow a similar trend.
- Next, we notice that within different settings of betas, the average value of the biases increase as well. We see that for the low setting(of beta discrepancy), he values hover around 0.14, while as we increase the discrepancy we get values around 0.28 and finally around 0.55 for the highest setting.
- Because we see a similar pattern above for ADJA1, it appears as though increasing the discrepancy in betas will make it more difficult for us to properly estimate and recover the true value of betas.

# Issues that Arise

- The adjustment process results in relatively good estimates of the true betas. In fact, the BIAS of A1 can get as bad as 4.04, and the BIAS of A2 can be as bad 6.02; When compared with the max biases of ADJA1 (0.377), and ADJA2 (0.56).

# Issues that Arise

- The adjustment process results in relatively good estimates of the true betas. In fact, the BIAS of A1 can get as bad as 4.04, and the BIAS of A2 can be as bad 6.02; When compared with the max biases of ADJA1 (0.377), and ADJA2 (0.56).
- However a key issue is the fact that we need to estimate:

$$\sigma_1, \sigma_2, \sigma_{u1}, \sigma_{u2}$$

# Issues that Arise

- The adjustment process results in relatively good estimates of the true betas. In fact, the BIAS of A1 can get as bad as 4.04, and the BIAS of A2 can be as bad 6.02; When compared with the max biases of ADJA1 (0.377), and ADJA2 (0.56).
- However a key issue is the fact that we need to estimate:

$$\sigma_1, \sigma_2, \sigma_{u1}, \sigma_{u2}$$

- Estimating these values will require repeated measurements of each predictor variable for each individual in the sample.

# Issues that Arise

- The adjustment process results in relatively good estimates of the true betas. In fact, the BIAS of A1 can get as bad as 4.04, and the BIAS of A2 can be as bad 6.02; When compared with the max biases of ADJA1 (0.377), and ADJA2 (0.56).
- However a key issue is the fact that we need to estimate:

$$\sigma_1, \sigma_2, \sigma_{u1}, \sigma_{u2}$$

- Estimating these values will require repeated measurements of each predictor variable for each individual in the sample.
- In the analysis above, we simulated the scenario where to took 2 measurements of both $X_1$ and $X_2$ for each individual/rep.

- $$E(MSB) = m\sigma_x^2 + \sigma_u^2$$
  $$E(MSW) = \sigma_u^2$$
  $$\implies \sigma_x^2 = \frac{E(MSB) - E(MSW)}{m}$$

  where $m$ is the number of repeated measurements, n is the sample size and:

- $$E(MSB) = m\sigma_x^2 + \sigma_u^2$$
  $$E(MSW) = \sigma_u^2$$
  $$\implies \sigma_x^2 = \frac{E(MSB) - E(MSW)}{m}$$
  where $m$ is the number of repeated measurements, n is the sample size and:
- $$MSB = \frac{m\sum_{i=1}^{n}(\overline{X_i} - \overline{X})^2}{n-1}$$
  $$MSW = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}(X_{ij} - \overline{X_i})^2}{m-1}\right)}{n}$$

- $$E(MSB) = m\sigma_x^2 + \sigma_u^2$$
  $$E(MSW) = \sigma_u^2$$
  $$\implies \sigma_x^2 = \frac{E(MSB) - E(MSW)}{m}$$
  where $m$ is the number of repeated measurements, n is the sample size and:
- $$MSB = \frac{m\sum_{i=1}^{n}(\overline{X_i} - \overline{X})^2}{n-1}$$
  $$MSW = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{m}(X_{ij} - \overline{X_i})^2}{m-1}\right)}{n}$$
- Thus by the law of large numbers and consistency of the sample standard deviation/variance estimator, we know that if increase the sample size, we will get better estimates of these values.

# Sample size = 50

▶

| (2,5) | BIASA1ADJ(50) | BIASA1ADJ(100) | BIASA2ADJ(50) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.148 | -0.146 | 0.204 | 0.141 |
| 0.5 | -0.148 | -0.146 | 0.204 | 0.141 |
| 0.9 | -0.149 | -0.147 | 0.204 | 0.142 |

Table 1: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=5

| (2,10) | BIASA1ADJ(50) | BIASA1ADJ(100) | BIASA2ADJ(50) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.239 | -0.220 | 0.411 | 0.280 |
| 0.5 | -0.241 | -0.221 | 0.412 | 0.280 |
| 0.9 | -0.244 | -0.223 | 0.414 | 0.282 |

Table 2: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=10

| (2,20) | BIASA1ADJ(50) | BIASA1ADJ(100) | BIASA2ADJ(50) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.422 | -0.366 | 0.826 | 0.557 |
| 0.5 | -0.427 | -0.370 | 0.828 | 0.558 |
| 0.9 | -0.436 | -0.377 | 0.833 | 0.564 |

Table 3: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=20

- We notice that in all scenarios, the magnitude of the biases are larger in the case where we use a smaller sample size. This makes sense of course because as we have fewer people, the more variable our estimates will be, and through a law of large numbers argument, the greater the sample size, the closer we are to the true value of ICC, and thus the true value of betas.

- We notice that in all scenarios, the magnitude of the biases are larger in the case where we use a smaller sample size. This makes sense of course because as we have fewer people, the more variable our estimates will be, and through a law of large numbers argument, the greater the sample size, the closer we are to the true value of ICC, and thus the true value of betas.
- Moreover, we can also quite easily see that as we increase the value of the correlation between measurement errors, the magnitude of the bias tends to increase. It's also easy to see that as we increase $\beta_2$, this results in a general increase in the amount of bias in our estimates.

# Outline

# Log-Normal Measurement Error

- In this section, instead of having measurement error follow a joint-normal distribution with 0 mean, we instead let it follow a joint log-normal distribution. This will cover the case when the distribution of the measurement error is skewed(to the right).

# Log-Normal Measurement Error

- In this section, instead of having measurement error follow a joint-normal distribution with 0 mean, we instead let it follow a joint log-normal distribution. This will cover the case when the distribution of the measurement error is skewed(to the right).
- To implement this we randomly draw both $u_1$ and $u_2$ from a standard log-normal distribution, then we standardize both, and implement a correlation structure between the two (similar to what we did to define a multivariate normal distribution).

# Observations

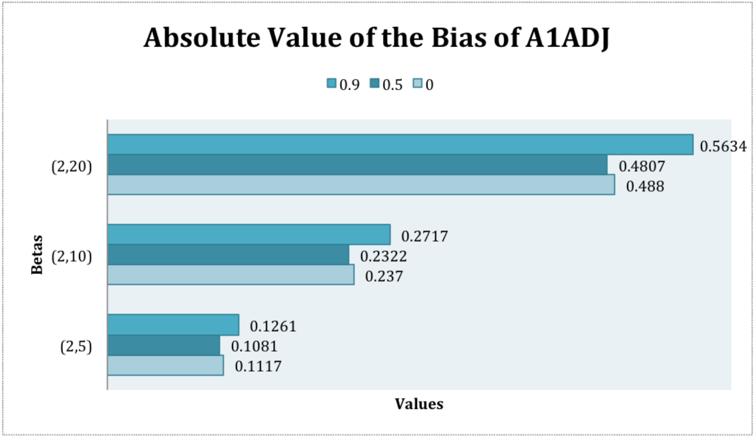



Figure:

## ...

- We observe that as we increase the discrepancy in betas, the magnitude of the bias of our adjusted estimate tends to become worse

## ...

- We observe that as we increase the discrepancy in betas, the magnitude of the bias of our adjusted estimate tends to become worse
- We also notice a V-shaped pattern in how correlation tends to affect the magnitude of the bias of ADJA1.

## ...

- We observe that as we increase the discrepancy in betas, the magnitude of the bias of our adjusted estimate tends to become worse
- We also notice a V-shaped pattern in how correlation tends to affect the magnitude of the bias of ADJA1.
- As we increase correlation from 0 to 0.5, we notice a slight decrease in bias, followed by a relatviely larger increase when $\sigma_{u12}$ becomes 0.9.

# Observations

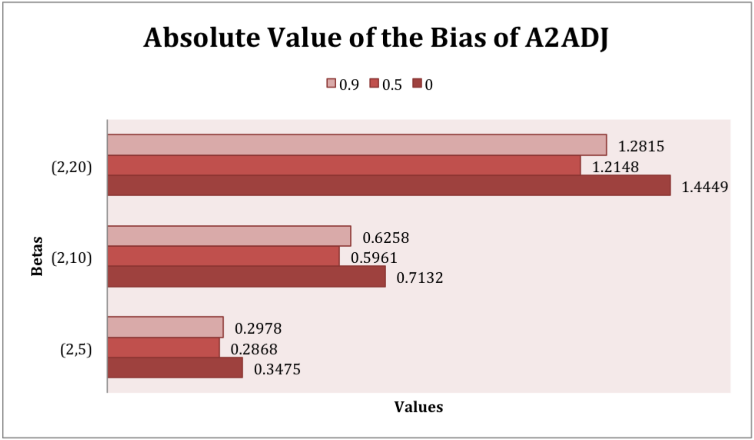



Figure:

## ...

- Just like ADJA1 we notice that as we increase the beta-discrepancy, that the bias of ADJA2 tends to increase.

# ...

- Just like ADJA1 we notice that as we increase the beta-discrepancy, that the bias of ADJA2 tends to increase.
- Moreover, we see that again there is a V-shaped pattern in how the bias is affected by correlation.

- Just like ADJA1 we notice that as we increase the beta-discrepancy, that the bias of ADJA2 tends to increase.
- Moreover, we see that again there is a V-shaped pattern in how the bias is affected by correlation.
- Similar but opposite to ADJA1, we notice the largest value of bias for 0 correlation, followed by a large(relative) drop in bias for $\sigma_{u12}$=0.5, and finally a small increase as we move from 0.5 to 0.9.

# All out comparison

| (2,5) | BIASA1ADJ(log) | BIASA1ADJ(100) | BIASA2ADJ(log) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.112 | -0.146 | 0.347 | 0.141 |
| 0.5 | -0.108 | -0.146 | 0.287 | 0.141 |
| 0.9 | -0.126 | -0.147 | 0.298 | 0.142 |

Table 4: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=5

| (2,10) | BIASA1ADJ(log) | BIASA1ADJ(100) | BIASA2ADJ(log) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.237 | -0.220 | 0.713 | 0.280 |
| 0.5 | -0.232 | -0.221 | 0.596 | 0.280 |
| 0.9 | -0.272 | -0.223 | 0.626 | 0.282 |

Table 5: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=10

| (2,20) | BIASA1ADJ(log) | BIASA1ADJ(100) | BIASA2ADJ(log) | BIASA2ADJ(100) |
|---|---|---|---|---|
| 0 | -0.488 | -0.366 | 1.445 | 0.557 |
| 0.5 | -0.481 | -0.370 | 1.215 | 0.558 |
| 0.9 | -0.563 | -0.377 | 1.282 | 0.564 |

Table 6: Comparison of adjusted estimates for $\beta_1$=2, $\beta_2$=20

Figure:

- It is only the case for $\beta_1$=2, $\beta_2$=5 that the bias of ADJA1 (under the model with lognormal measurement error) has lower bias than the original case. In all other scenarios, we see that the bias for both adjusted estimates are much higher for the model with lognormal measurement error compared to the model with just normal measurement error.

- It is only the case for $\beta_1$=2, $\beta_2$=5 that the bias of ADJA1 (under the model with lognormal measurement error) has lower bias than the original case. In all other scenarios, we see that the bias for both adjusted estimates are much higher for the model with lognormal measurement error compared to the model with just normal measurement error.
- Again, we notice however, that the amount of bias (absolute value) exhibits a V-like pattern as we increase the correlation levels;

## ...

- It is only the case for $\beta_1$=2, $\beta_2$=5 that the bias of ADJA1 (under the model with lognormal measurement error) has lower bias than the original case. In all other scenarios, we see that the bias for both adjusted estimates are much higher for the model with lognormal measurement error compared to the model with just normal measurement error.
- Again, we notice however, that the amount of bias (absolute value) exhibits a V-like pattern as we increase the correlation levels;
- Addtionaly, the level of bias tends to increase as we increase $\beta_2$ as well.

# All out comparison

| (2,5) | BIASA1ADJ(50) | BIASA1ADJ(log) | BIASA2ADJ(50) | BIASA2ADJ(log) |
|---|---|---|---|---|
| 0 | -0.148 | -0.112 | 0.204 | 0.347 |
| 0.5 | -0.148 | -0.108 | 0.204 | 0.287 |
| 0.9 | -0.149 | -0.126 | 0.204 | 0.298 |

Table 7: Comparison between n=50 and log-normal measurement errors

| (2,10) | BIASA1ADJ(50) | BIASA1ADJ(log) | BIASA2ADJ(50) | BIASA2ADJ(log) |
|---|---|---|---|---|
| 0 | -0.239 | -0.237 | 0.411 | 0.713 |
| 0.5 | -0.241 | -0.232 | 0.412 | 0.596 |
| 0.9 | -0.244 | -0.272 | 0.414 | 0.626 |

Table 8: Comparison between n=50 and log-normal measurement errors

| (2,20) | BIASA1ADJ(log) | BIASA1ADJ(50) | BIASA2ADJ(50) | BIASA2ADJ(log) |
|---|---|---|---|---|
| 0 | -0.488 | -0.422 | 0.826 | 1.445 |
| 0.5 | -0.481 | -0.427 | 0.828 | 1.215 |
| 0.9 | -0.563 | -0.436 | 0.833 | 1.282 |

Table 9: Comparison between n=50 and log-normal measurement errors

# Analysis

- In addition, we can compare the results from the case where we reduced sample size to 50 with the results we observed from introducing log-normal measurement error.

# Analysis

- In addition, we can compare the results from the case where we reduced sample size to 50 with the results we observed from introducing log-normal measurement error.
- What we notice is that aside from the case for BIASA1ADJ, when $\beta_2 = 5$, log-measurement error causes higher bias values than reducing the sample size by half.

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error
- Having skewed measurement errors, makes estimates less accurate, hence increasing the value of bias for all estimates.

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error
- Having skewed measurement errors, makes estimates less accurate, hence increasing the value of bias for all estimates.
- After adjustments, we notice that ADJA1 still under-estimate $\beta_1$, while ADJA2 over-estimate $\beta_2$

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error
- Having skewed measurement errors, makes estimates less accurate, hence increasing the value of bias for all estimates.
- After adjustments, we notice that ADJA1 still under-estimate $\beta_1$, while ADJA2 over-estimate $\beta_2$
- MA1 tends to decrease as we increase correlations, and thus making the bias of A1 higher

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error
- Having skewed measurement errors, makes estimates less accurate, hence increasing the value of bias for all estimates.
- After adjustments, we notice that ADJA1 still under-estimate $\beta_1$, while ADJA2 over-estimate $\beta_2$
- MA1 tends to decrease as we increase correlations, and thus making the bias of A1 higher
- MA2 tends to increase as increase correlation, thus making the bias lower

# Summary

- As we increase the beta-discrepancy/the size of $\beta_2$, the the level of bias of our adjusted estimates tends to increase.
- There also appears to be higher levels of bias for more extreme levels of correlation (ie: 0, 0.9 vs 0.5)
- We consistently underestimate the values of the betas, when we use the model with measurement error
- Having skewed measurement errors, makes estimates less accurate, hence increasing the value of bias for all estimates.
- After adjustments, we notice that ADJA1 still under-estimate $\beta_1$, while ADJA2 over-estimate $\beta_2$
- MA1 tends to decrease as we increase correlations, and thus making the bias of A1 higher
- MA2 tends to increase as increase correlation, thus making the bias lower
- Log-Measurement error seems to make biases worse than reducing the sample size by half.