AN INVESTIGATION OF THE PERFORMANCE FOR ORDERED SUBSET ANALYSIS TO A GENE-ENVIRONMENT INTERACTION MODEL

by

William L. Welbourn, Jr.

A Thesis Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(BIOSTATISTICS)

May 2006

Table of Contents

List of Tables

Abstract

Ordered subset analysis (OSA) is a recent (~1999) addition to statistical methodology and its application has been limited to genetic linkage analysis. This study extends the OSA to a test of association in a gene-environment interaction model, where the explanatory variable is continuous. The application of the OSA to this model, entails: (a) Data Simulation, (b) Regression Analysis, and (c) Empirical p-value determination. The power of the OSA is determined, based on the Empirical p-values obtained via repetition of steps (a) - (c) above, and is used as a performance marker for comparison to "traditional" linear regression analysis.

The goal of the OSA is to determine the level of the continuous environmental factor, which provides maximum evidence of a difference in the mean levels of the outcome between the two gene levels for a given dichotomous gene factor.

CHAPTER I

INTRODUCTION

1.1: Motivation for this Study

This is an exciting time in the field of Biostatistics, as computers have evolved into phenomenal machines. In particular, the dynamic evolution of the personal computer has allowed the Field to broaden its "scope of analytical abilities", and burdensome computational algorithms of the past are now possible to investigate. Permutation dataset analyses are an example of such algorithms, due to the extraordinary number of orderings for a single dataset under investigation. For example, a dataset of only ten observations has $10!$ (> 3.5 million) possible orderings. The Ordered Subset Analysis (OSA) process is essentially a permutation analysis, entailing computationally challenging programming, and has been applied previously in genetic linkage analyses, "to identify subsets of families defined by the level of a trait related covariate that provide maximal evidence for linkage, without requiring a priori specification of the subset"[1]. Upon review of the literature, its current use appears to be limited in scope to genetic linkage analysis studies, and is a fairly recent addition to the fields of Biostatistics and Genetic Epidemiology (~ 1999).

The motivation for this study, stems from the paper by Hauser et. al.[1], and expands the use of the OSA to a test of association with a continuous explanatory

variable, in a gene-environment interaction model, where the environmental factor is a continuous trait and the gene factor is simply dichotomous. The OSA in this setting, attempts to reveal the subset of samples defined by the level of the environmental factor, which provides maximum evidence of a genetic effect, quantified by a difference in the mean levels of the outcome between the two levels of the gene factor. Such a subset is expected to exist when there is gene-environment interaction. It is hypothesized that the OSA for the test of the gene factor, will prove to be a useful and practical new statistical methodology for use in this modelling scheme.

In the subsequent document which follows, we will outline the methods by which the OSA is conducted (Chapter II), the simulation techniques and assumptions we have chosen for this investigation (Chapter III), summarize the results (Chapter IV), and provide remarks from it (Chapter V).

CHAPTER II

METHODS

2.1: The Linear Regression Model

As with any regression modeling scheme, the first step is to define the variables (independent and dependent) involved. Here, we assume that a random sample of observations have been recorded for analysis. For each observation of our association study, there are two independent variables which are of interest, one dichotomous and one continuous. In addition to these two independent variables, there is one dependent variable of interest for each observation, which is assumed to be continuous. An example of the dichotomous independent (gene trait), continuous independent (environmental trait), and continuous dependent variables of interest for our study setting could be, respectively, at least one gene allele for a given genotype being dominant versus both allele's for the genotype being recessive, Low-density lipoprotein (LDL) cholesterol levels, and Systolic blood pressure levels. Here, it is assumed that the classical linear regression model fits these independent variables (predictors) and the dependent variable (outcome), according to the expression

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 X_i + \beta_3 G_i * X_i + E_i, \quad (1)$$

where, for arbitrary dominant/recessive gene alleles A/a, G_i is the "genotype class indicator" for the gene trait, defined by

$$G_i = \begin{cases} 1, & \text{if the genotype for the gene under investigation is Aa or AA} \\ 0, & \text{if the genotype for the gene under investigation is aa} \end{cases},$$

X_i is the value for the continuous environmental trait, E_i is the random error, and Y_i is the value of the continuous outcome for the i^{th} observation recorded, and where each of the parameters, $\beta_j,\ j = 0, 1, 2, 3,$ is assumed to be unknown.

2.2: Ordered Subset Analysis

As was mentioned in the introduction of this paper, the OSA attempts to test for gene-environment interaction. It does this by revealing the level of the environmental trait, which provides maximum evidence of a difference in the mean levels of the outcome between the two levels of the gene trait. This is done by first sorting the recorded observations by some notion regarding the environmental trait. For clarity of our OSA description, hereinafter, we will assume that the observations are to be sorted in descending order, according to the value of the continuous environmental trait. Following the sorting of the observations, we collect the ten (value chosen for convenience) observations with the greatest value for the environmental trait and these then form a subset of observations. Simple linear regression (Y on G) is then performed on this subset, and the p-value from which is recorded. Hereinafter, when we refer to the simple linear regression of Y on G, the model

under the null hypothesis is assumed to be $Y = \gamma_0 + E$, and under the alternative hypothesis, $Y = \gamma_0 + \gamma_1 G + E$, where γ_i, $i = 0, 1$, are each assumed to be unknown parameters. The eleventh observation of the sorted observations is then added to the subset, the simple linear regression is repeated, and the p-value is recorded. This notion of "adding an observation to the subset and performing simple linear regression" is repeated until the simple linear regression has been performed on all of the recorded observations. We then determine the minimum value among all of the p-values recorded, which we denote by p_1. So, then the subset which yielded the value of p_1, suggests maximum evidence (statistically) for a difference between the mean levels of the outcome for the two levels of the gene trait. Hence, the final observation added to this subset, provides the level of the environmental trait for which the OSA sought out to determine. We denote this value of the continuous environmental trait by x_1.

There is a rather substantial drawback to simply performing the analysis of the previous paragraph. Here, we note that repetitious simple linear regression is performed on the same observations/data, and no adjustment to the overall Type I Error level is being made. Hence, it could be due to chance that we determined this level of the environmental trait (x_1) using the OSA. We need a basis by which to evaluate the use of the OSA effectively, so that we can maintain a fixed Type I Error level. This can be done by way of comparing the OSA to what we will refer as "Null Distributions."

2.3: The Continuous Exposure Ordering (CEO) Null Distribution for the OSA

To maintain a fixed Type I Error level for the OSA, we essentially need to evaluate the OSA algorithm against the same subset statistical analysis (simple linear regression of Y on G), for a random ordering of the observations. One random ordering scheme, which resembles that of the OSA, is to sort (order) the observations, randomly, according to the value of the continuous exposure trait. So, if n observations have been recorded, then we have $n! - 1$ possible orderings of the observations to compare the OSA to. This in turn implies that there is a minimum p-value for each of these subset statistical analyses, which we denote by p_{CEO_i}, $i = 1, 2, \ldots, n! - 1$. The value of p_1 is then compared to the values of the sequence $\{p_{\text{CEO}_i}\}_{i=1}^{n!-1}$, and an empirical p-value is determined. The value of the empirical p-value is equal to the proportion of the values of the sequence $\{p_{\text{CEO}_i}\}_{i=1}^{n!-1}$ which are less than the value of p_1. The beauty behind the comparison of the OSA to the random permuted orderings of the data, is that the empirical p-value can be compared to the fixed Type I Error level of our choosing. The sequence of p-values, $\{p_{\text{CEO}_i}\}_{i=1}^{n!-1}$, comprises what we will refer to as the Continuous Exposure Ordering (CEO) Null Distribution. We will see in Chapter IV below, the comparison of the OSA to the CEO Null Distribution (with the application of the underlying assumptions set forth in the subsequent chapter of this paper) is essentially a new way of testing for the interaction effect between G and X, for the model described by (1) of Section 2.1 above.

2.4: The Continuous Outcome Ordering (COO) Null Distribution for the OSA

As was mentioned in the previous section, to maintain a fixed Type I Error level for the OSA, we essentially need to evaluate the OSA algorithm against the same subset statistical analysis (simple linear regression of Y on G) for a random ordering of the observations. However, we are not restricted to the random sorting scheme, presented in the preceding section. We can randomly sort any variable(s) of the observations of our choosing. With this in mind, for each observation, we fix the ordering of the independent variables, while allowing the dependent variable vector to be sorted randomly. So, once again if n observations have been recorded, then we have $n! - 1$ possible orderings of the observations to compare the OSA to. Thus, there is a minimum p-value for each of these subset statistical analyses, which we denote by p_{COO_i}, $i = 1, 2, \ldots, n! - 1$. The value of p_1 is then compared to the values of the sequence $\{p_{\mathrm{COO}_i}\}_{i=1}^{n!-1}$, and an empirical p-value is determined. As with the comparison of the OSA to the CEO Null Distribution, the value of the empirical p-value is equal to the proportion of the values of the sequence $\{p_{\mathrm{COO}_i}\}_{i=1}^{n!-1}$ which are less than the value of p_1. The empirical p-value can then be compared to the fixed Type I Error level of our choosing. The sequence of p-values, $\{p_{\mathrm{COO}_i}\}_{i=1}^{n!-1}$, comprises what we will refer to as the Continuous Outcome Ordering (COO) Null Distribution. In Chapter IV below, we will see that the comparison of the OSA to the COO Null Distribution (with the application of the underlying assumptions set forth in the subsequent chapter of this paper) is

essentially a new way to test the simultaneous effects of the main effect for G and the interaction effect of G and X, for the model given by (1) of Section 2.1 above.

CHAPTER III

DATA SIMULATION

3.1: Overview of the Data Structure

The basis for the remainder of this study, lies in the assumptions set forth in this section. As the title of this chapter suggests, we will be proceeding with data simulation for this study. However, before we can begin discussing the data simulation process, we addressed several ideas in the preceding chapter, which need clarification for the way the simulation is to be conducted. The first of which is that we assumed a random sample of observations have been collected, but we did not specify a sample size. For our study, we fix the number of observations, n, to be 200. This sample size is small enough to be representative of data to be collected in the "real world," and seems large enough to conduct this OSA study.

The next thing that needs clarification is the underlying distribution assumptions for the independent variables G and X, as well as the random error, E. For the i^{th} observation of our study, we assume that $G_i \sim B(1, 0.3)$; $X_i, E_i \sim N(0, 1)$, and that each of these random variables are pairwise independent. The "dominant allele frequency" chosen (namely, 0.3), seems to be a reasonable value by which this study will have applicability in the "real world." Since many continuous traits are normally distributed in the "real world" such as LDL cholesterol, and in the linear

regression setting continuous predictors are typically standardized, the fact that the predictor X was chosen to follow the standard normal distribution should provide for applicability of this study to real data. Finally, recall that one of the assumptions in the linear regression model is that the residuals (random error, E) be normally distributed with mean zero and non-zero variance, σ^2. In our study, for convenience, we assume the value of σ^2 to be identically equal to one, so that the residuals follow the standard normal distribution.

The final thing we need to address here is that the OSA is to be applied to the simulated data (adhering to the model expression of (1) of Section 2.1), so that simple linear regression of Y on G is to be performed on this data. Thus, our simulated data needs to include an outcome vector $\{Y_i\}_{i=1}^{200}$. To obtain this outcome vector, we will need defined values for each of the parameters β_j of expression (1). A way of defining these parameters is through the linear relationships of the predictors G, X, and $G*X$, with Y. As was mentioned in Sections 2.3 and 2.4 above, since the OSA is essentially a new methodology of testing for the interaction effect of G and X, one idea is that we fix the magnitudes of the linear relationships between G and Y, and X and Y, while allowing for deviations in the magnitude of the linear relationship between Y and $G*X$. This should allow us to make more accurate conclusions for the OSA, since (in theory) we are essentially only allowing one of the three beta parameters (β_1, β_2, and β_3) to fluctuate in value. Moreover, the linear relationship between Y and $G*X$ is the focus of the OSA, and so of the three beta parameters (β_1, β_2, and β_3), the associated beta parameter for this

relationship (β_3) should be the main parameter of interest. This being said, we proceed to define the procedure for assigning values to the beta parameters, where we assume that the linear relationship between Y and G, and Y and X essentially remain fixed. Let $r^2_{\mathrm{YZ}_1|Z_2,Z_3}$ be the square of the second-order partial correlation coefficient for the linear relationship of Y and Z_1, controlling for each of the effects of Z_2 and Z_3, where $Z_i \in \{G, X, G*X\}$, such that $Z_i \neq Z_j$, $j \neq i = 1, 2, 3$. Further, let $\boldsymbol{R^2}$ be the vector of the squared second-order partial correlation coefficients, $\left(r^2_{\mathrm{YG}|X,G*X}, r^2_{\mathrm{YX}|G,G*X}, r^2_{Y(G*X)|G,X}\right)$. Essentially, for the data simulation process, we will choose values for the vector $\boldsymbol{R^2}$, and calculate the beta parameters from these choices. It can be shown, given an arbitrary vector $\boldsymbol{R^2}$, the applicable beta parameters for the respective squared second-order partial correlation coefficients of $\boldsymbol{R^2}$, $(\beta_1, \beta_2, \beta_3)$, under the distribution assumptions set forth for each of the predictors G and X, are

$$\beta_1 = \sqrt{\frac{r^2_{YG|X,G*X}(n-4)}{1-r^2_{YG|X,G*X}}}\sqrt{\frac{1}{0.21\,n}},$$

$$\beta_2 = \sqrt{\frac{r^2_{YX|G,G*X}(n-4)}{1-r^2_{YX|G,G*X}}}\sqrt{\frac{1}{0.7\,n}}, \tag{2}$$

and

$$\beta_3 = \sqrt{\frac{r^2_{Y(G*X)|X,G}(n-4)}{1-r^2_{Y(G*X)|X,G}}}\sqrt{\frac{1}{0.21\,n}}.$$

The final note that we provide here is that the beta parameter β_0 is essentially a nuisance parameter, and so for simulation purposes we assign the value of zero to this parameter.

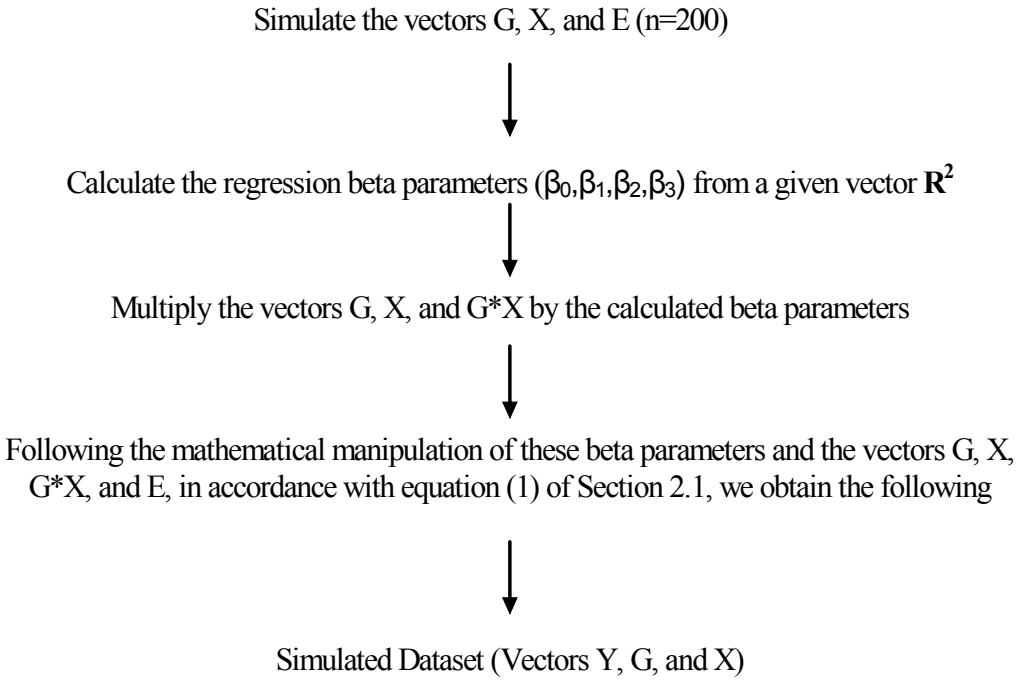
3.2: Obtaining Simulated Data for the OSA

In reviewing the underlying linear regression model, given by (1) above, we note that for each observation there are two predictors G_i and X_i, in addition to the outcome Y_i. We also note here that in the data simulation process, the outcome Y_i is derived through the mathematical manipulation of these two predictors, the beta parameters, and the random error (E_i) assigned to the respective observation. Thus, for each observation, the data simulation is essentially a three step process: (a) we need values assigned to each of the four beta parameters (β_0, β_1, β_2, β_3), (b) we require values assigned to each of the predictors G_i and X_i, as well as the random error value E_i, and (c) the values obtained in steps (a) and (b) require mathematical manipulation, in accordance to the model expression given by (1) above, so that the outcome value Y_i is obtained.

As we mentioned in the preceding section, for any given value of the vector $\boldsymbol{R^2}$ (of our choosing), applying the formulas given by (2) allows us to calculate the beta parameters β_1, β_2, and β_3. Since β_0 is fixed to the value of zero, we now have a means by which to assign values to the four beta parameters (β_0, β_1, β_2, β_3). For

each observation, the values of G_i, X_i, and E_i are derived via simulation through the R[2] software package (Version 2.1.1), in accordance to the respective underlying distribution assumptions set forth in the second paragraph of the preceding section, and that $\mathcal{X}_i$ is independent of $\mathcal{X}_j$, all $i \neq j$, $\mathcal{X} \in \{G, X, E\}$. That is, each of the elements of the sequence $\{\mathcal{X}_i\}_{i=1}^{200}$ are independently identically distributed random variables. These simulated values are then mathematically manipulated with the beta parameters, to obtain the outcome value Y_i. Hereinafter, we will refer to the collection of the simulated sets (vectors) $\{G_i\}_{i=1}^{200}$, $\{X_i\}_{i=1}^{200}$, and $\{Y_i\}_{i=1}^{200}$ as a dataset. It is noted here that the simulated vectors $\{G_i\}_{i=1}^{200}$, $\{X_i\}_{i=1}^{200}$, and $\{E_i\}_{i=1}^{200}$, are "recycled" and mathematically manipulated with the beta parameters (in accordance to (1) of Section 2.1) for each choice of the vector $\boldsymbol{R^2}$. In our study, we provide for nine choices of the vector $\boldsymbol{R^2}$, so that nine outcome vectors $\{Y_i\}_{i=1}^{200}$ are generated for the simulated vectors $\{G_i\}_{i=1}^{200}$, $\{X_i\}_{i=1}^{200}$, and $\{E_i\}_{i=1}^{200}$. In addition, to alleviate assumed simulation variation from data simulation, we simulate 500 of each of the vectors $\{G_i\}_{i=1}^{200}$, $\{X_i\}_{i=1}^{200}$, and $\{E_i\}_{i=1}^{200}$. Thus, 500 datasets are analyzed for each choice of $\boldsymbol{R^2}$. Figure 3.1 below, shows this simulation process for one dataset.

Figure 3.1: Flow Chart of the OSA Data Simulation Process



3.3: A Further Assumption and the Error Inherited in this Study

An unfortunate consequence of this permutation analysis is the magnitude of the integer $n! \equiv 200!$. As we mentioned in Sections 2.3 and 2.4 above, the respective CEO and COO Null Distributions for the OSA, each comprise $n! - 1$ p-values. Moreover, each of these p-values represents the minimum value of approximately 191 p-values which originate from the subset simple linear regression analyses of Y on G. Thus, the total number of simple linear regression analyses, required to obtain the "full" structure for each of these Null Distributions is greater than $n! \cdot (n - 10)$. The idea here is that to generate the entire Null Distribution (CEO or COO), a great many linear regression analyses would be required. So, essentially we need to sample a subset of the $n! - 1$ random orderings for each of the Null Distributions for the OSA. For each of the Null Distributions, we choose to subset 10,000 of the $n! - 1$ random orderings. The number of subset random orderings chosen (10,000) seems to be large enough to provide reliable conclusions regarding the OSA, while small enough for timely computational completion. There is an advantage to the number of subsets chosen. Namely, 10,000 subset random orderings represents less than 10^{-192} of the total random orderings for either of the Null Distributions. Hence, the probability of any one of these 10,000 subsets being chosen more than once, is nearly zero. On this same note, the mere fact that a very minute sampling of the total random orderings for these Null Distributions is a disadvantage, and provides a source of error in our study.

CHAPTER IV

RESULTS

4.1: Summary Measures for the OSA

In this section, we summarize basic summary measures from this study. Table 4.1 below, displays (a) the nine choices for the vector $\boldsymbol{R^2}$, (b) the corresponding beta parameters (β_1, β_2, β_3) for these nine choices of $\boldsymbol{R^2}$, (c) the median (of the 500 datasets simulated) subset sample size where the minimum p-value, p_1, occurs for the OSA, for each of the choices of $\boldsymbol{R^2}$, and (d) the mean value (of the 500 datasets simulated) for the minimum values of the continuous predictor (X), of the subsets which yield the value of p_1 under the OSA (or simply the sample mean of the values of x_1), for each choice of the vector $\boldsymbol{R^2}$. From this table, we see that for each of the nine choices of $\boldsymbol{R^2}$, the values of $r^2_{\mathrm{YG}|X,G*X}$ and $r^2_{\mathrm{YX}|G,G*X}$, were each fixed to the value of 0.01, while the value of $r^2_{Y(G*X)|G,X}$ was allowed to change. The choices for the vector $\boldsymbol{R^2}$, corresponds to the discussion in Section 3.1, where we mentioned that since the OSA is essentially a new methodology of testing for the interaction effect of G and X, the notion of allowing deviations in the magnitude of the linear relationship between Y and $G*X$, while maintaining fixed magnitudes of the linear relationships between G and Y, and X and Y, would be desirable for the OSA investigation. In addition, the spectrum of values chosen for

the statistic $r^2_{Y(G*X)|G,X}$, allows a broad range of measures for this OSA investigation.

Table 4.1: Summary Measures for the OSA

$\boldsymbol{R}^2$	*Beta Parameters* $(\beta_1,\beta_2,\beta_3)^*$	*Median Subset Sample Size Where* p_1 *Occurs for the OSA***	*Mean Value of* x_1^{**}
(0.01, 0.01, 0.00)	(0.217, 0.119, 0.000)	103.0	-0.15
(0.01, 0.01, 0.01)	(0.217, 0.119, 0.217)	83.0	0.15
(0.01, 0.01, 0.02)	(0.217, 0.119, 0.309)	81.0	0.22
(0.01, 0.01, 0.05)	(0.217, 0.119, 0.496)	74.5	0.33
(0.01, 0.01, 0.075)	(0.217, 0.119, 0.615)	74.0	0.34
(0.01, 0.01, 0.10)	(0.217, 0.119, 0.720)	73.0	0.36
(0.01, 0.01, 0.15)	(0.217, 0.119, 0.907)	73.0	0.36
(0.01, 0.01, 0.20)	(0.217, 0.119, 1.080)	72.5	0.38
(0.01, 0.01, 0.25)	(0.217, 0.119, 1.247)	72.0	0.38

*The beta parameters are calculated from each of the respective squared second – order correlation coefficient values given within the vector $\boldsymbol{R}^2$, applying expression (2) of Section 3.1.

**Based on 500 datasets for each value of $\boldsymbol{R}^2$ listed.

Moreover, we see that for each choice of the vector $\boldsymbol{R}^2$, the values of the calculated beta parameters β_1 and β_2 remain fixed at the respective values of 0.217 and 0.119 (column 2 of Table 4.1). This result makes sense, since the sample size for our study is fixed to $n = 200$ and since the values of $r^2_{\mathrm{YG}|X,G*X}$ and $r^2_{\mathrm{YX}|G,G*X}$, are each fixed to the value of 0.01. Then, expression (2) of Section 3.1 implies that each of the beta parameters β_1 and β_2 will remain fixed for any choice of $\boldsymbol{R}^2$. We also see an increasing trend in the parameter β_3, as the value of $r^2_{Y(G*X)\,X|G,X}$ increases for

the choice of $\boldsymbol{R^2}$. This result also makes sense, in light that each of the beta parameters, calculated from expression (2), are increasing functions for increasing values of the respective squared second-order correlation coefficients.

The third column of Table 4.1, suggests that the subset sample size, which yields the minimum p-value for the OSA, p_1, decreases as the magnitude of $r^2_{Y(G*X)\,X|G,X}$ increases. It can be shown (Appendices D and E) that under the OSA, the expected value of the maximum likelihood estimate (MLE) for the parameter γ_1 (for the simple linear regression of Y on G), is $E(\hat{\gamma}_1) = \beta_1 + \beta_3 \overline{X}$, provided that an "ample" (Appendices D and E) subset sample size is analyzed. By ample, we imply that the subset under OSA investigation should include at least 5% of the total random sample collected, although the notion of this "minimum threshold" value could be investigated further in a future study. So, for our investigation, this requires that the minimum subset sample size under investigation is ten ($= n * 5\,\%$) observations, which is precisely the case for this study. Since the datasets under the OSA are in descending order, according to the value of the continuous trait (X), then a larger subset sampled, implies a smaller expected value for the parameter γ_1. This in turn implies that the expected difference in the mean levels of the outcome between the two levels of the gene factor will decrease (since the simple linear regression of Y on G is equivalent to the independent two-sample t-test), thereby resulting in greater p-values for the null hypothesis that γ_1 equals zero. This is precisely the opposite of what the OSA methodology is all about (finding the minimum p-value). Hence, under the OSA methodology, we expect to be

analyzing smaller subset sample sizes where the minimum p-value, p_1, occurs. This is exactly what the data from this study suggest.

The final column of Table 4.1, suggests that the minimum value of the continuous trait (X), for the subset which yields the minimum p-value, p_1, is increasing for increasing values of $r^2_{Y(G*X)\,X|G,X}$. This result is expected from the discussion of the immediately preceding paragraph. That is, since the datasets are sorted in descending order (under the OSA), according to the value of the continuous trait (X), and we said that smaller subset sample sizes yield the minimum p-value (p_1) for increasing values of $r^2_{Y(G*X)\,X|G,X}$, then we would expect the minimum value of the continuous trait (X) for these subset samples to also be increasing for increasing values of $r^2_{Y(G*X)\,X|G,X}$.

4.2: The Expected Value of $\hat{\gamma}_1$

Table 4.2 below, summarizes the differentiation in the expected value of the slope parameter (γ_1), for the simple linear regression of Y on G, among (a) the OSA, (b) each of the CEO and COO Null Distributions for the OSA, and (c) Traditional Linear Regression Analysis (TLRA). The term "Traditional Linear Regression Analysis" is taken to mean the simple linear regression of Y on G for the entire sample analyzed ($n = 200$). Each of the values given in this table was rigorously derived from probability theory, as well as data simulation, under the distribution

assumptions presented in the preceding chapter, the derivations of which are provided in Appendix A.

Table 4.2: Summary of how the OSA and Traditional Linear Regression Differs*

*OSA***	*CEO Null Distribution*	*COO Null Distribution*	*Traditional Linear Regression*
$\beta_1 + \beta_3 \overline{X}$	β_1	0	β_1

* Expected value of the MLE for the slope parameter (γ_1) of the linear regression of Y on G.
** Approximate value, based on the simulation analysis of Appendices *D* and *E*.

An immediate and interesting observation from this table is that the expected value of the MLE for the regression slope parameter (γ_1) under the CEO Null Distribution, is the same as that for the TLRA. This is not surprising, since the CEO Null Distribution is derived, based on subsets of the entire dataset, such that the veracity of the data has not been compromised. A distinction though is that the subset sample size analyzed under the CEO Null Distribution is allowed to change (ranging from 10 to 200 observations), whereas the sample size analyzed under the TLRA is statically fixed to the value of $n = 200$. However, the main result that we want to take from this table is in the comparison of the OSA to each of the Null Distributions, since this is the essence of the OSA Methodology. We see that when comparing the OSA to the CEO Null Distribution, the expected value of the MLE for γ_1 for the OSA has an excess "β_3 term", so that we are essentially testing the effect of the parameter β_3. Hence, this comparison has the essence of testing the

null hypothesis $H_0 : \beta_3 = 0$ in the TLRA setting. We will investigate this notion in the subsequent section below. Moreover, we also see that when comparing the OSA to the COO Null Distribution, the expected value of the MLE for γ_1 for the OSA has excess "β_1 and β_3 terms", so that we are essentially testing the simultaneous effects of the parameters β_1 and β_3. Thus, this comparison has the essence of testing the null hypothesis $H_0 : \beta_1 = \beta_3 = 0$ in the TLRA setting.

4.3: The Comparison of the OSA to TLRA

In this section, we investigate how the results from this OSA study compares to that of TLRA. Table 4.3 below summarizes (a) The proportion of the 500 datasets (for each of the respective values of $\boldsymbol{R^2}$) which yield empirical p-values less than 0.05 (essentially, the power to reject the null distribution at the 0.05 level of significance) for the OSA, based on each of the respective CEO and COO Null Distributions, and (b) The power to reject each of the respective null hypotheses $H_0 : \beta_3 = 0$ and $H_0 : \beta_1 = \beta_3 = 0$ in the Traditional Linear Regression setting, at the Type I Error Rate of 0.05, based on the 500 datasets for each of the respective values of $\boldsymbol{R^2}$.

Table 4.3: Summary for the Power of the OSA, Versus Traditional Linear Regression Analysis†

$r^2_{Y(G*X)\|G,X} \in \boldsymbol{R^2}$	*Power of the OSA Based on the CEO Null Distribution*	*Power of the OSA Based on the COO Null Distribution*	*Power of the TLRA for the Null Hypothesis* $\beta_3 = 0^*$	*Power of the TLRA for the Null Hypothesis* $\beta_1 = \beta_3 = 0^*$
0%	0.036(a)	0.146(a)	0.058(b)	0.204(c)
1%	0.244	0.378	0.320	0.400
2%	0.422	0.516	0.534	0.580
5%	0.726	0.782	0.892	0.886
7.5%	0.866	0.888	0.980	0.970
10%	0.914	0.938	0.994	0.994
15%	0.968	0.980	1.000	0.998
20%	0.992	0.996	1.000	1.000
25%	0.996	0.998	1.000	1.000

†Based on the 500 datasets simulated for each of the respective values of $\boldsymbol{R^2}$ listed.

* Based on the alternative hypothesis $Y = \beta_0 + \beta_1 G + \beta_2 X + \beta_3 G * X + E$; Test statistic based on the Likelihood Ratio Test.

(a) Power of the OSA at the Empirical p – value level of 0.05.

(b) Proportion of the datasets which yield p – values less than 0.05. Test statistic is asymptotically chi – square distributed , with one degree of freedom.

(c) Proportion of the datasets which yield p – values less than 0.05. Test statistic is asymptotically chi – square distributed, with two degrees of freedom.

We begin our discussion here, with the former of these two summaries. The fact that each of the respective values for column three is greater than that of column two is to be expected. To justify this notion, consider an arbitrary simulated dataset. Then, under the COO Null Distribution, the outcome values (Y) no longer possess the "data structure" assumed under the model given by (1) of Section 2.1. Thus, an arbitrary outcome value, under the "sorting rule" for this Null Distribution, has essentially the probability of 0.3 (the "dominant gene" allele frequency

chosen for this study) for being randomly assigned to an observation whose simulated predictor value for G being one, and the probability of 0.7 for being randomly assigned to an observation whose simulated predictor value for G being zero. The result of this is that the expected value for the MLE of γ_1 is zero, as we noted in Table 4.2 above. This in turn implies that under the COO Null Distribution, we expect about 5% of the values for the sequence $\{p_{\text{COO}_i}\}_{i=1}^{n!-1}$ to be less than the Type I Error rate of 0.05. However, under the CEO Null Distribution, we have that the expected value for the MLE of γ_1 to be β_1. Since $\beta_1 > 0$, then under this Null Distribution, we expect more than 5% of the values for the sequence $\{p_{\text{CEO}_i}\}_{i=1}^{n!-1}$ to be less than the Type I Error rate of 0.05. Hence, when we compare the value of p_1 for the OSA to each of the Null Distributions, we expect the empirical p-value under the COO Null Distribution to be smaller than that under the CEO Null Distribution. Therefore, we would expect the OSA/COO Null Distribution comparison to yield a greater proportion of empirical p-values less 0.05, than that of the OSA/-CEO Null Distribution comparison. This is precisely what has entailed from our simulated data, and what we set out to argue to be true.

Columns four and five, summarize the power for rejecting the respective null hypotheses $H_0 : \beta_3 = 0,$ and $H_0 : \beta_1 = \beta_3 = 0,$ each against the alternative model $Y = \beta_0 + \beta_1 G + \beta_2 X + \beta_3 G * X + E,$ at the Type I Error rate of 0.05. The results of the comparison of these two columns also makes sense. Because the value of β_1 is fixed and because of the "extra" degree of freedom for the chi-square test statistic, as the value of the statistic $r^2_{Y(G*X)|G,X}$ increases, we expect a greater increase of

distinct in the way each of the respective processes are carried out, and hence is a possible explanation of why the results of columns two and four differ.

Recall also that the OSA, when compared to the COO Null Distribution appears to be essentially comparing the null hypothesis $H_0 : \beta_1 = \beta_3 = 0$, in the TLRA setting. Thus, as was the case above, we would expect the power of the OSA in this case to resemble that of the TLRA for the null hypothesis $H_0 : \beta_1 = \beta_3 = 0$. Again, the results of columns three and five do suggest this, but the power of the OSA/COO Null Distribution is somewhat less than that of TLRA. A similar argument to that provided in the preceding paragraph, suggests that the differences in these two methodologies, provides a possible explanation of why the results of columns three and five differ.

CHAPTER V

CONCLUSION

5.1: Remarks from This Study

The results of this simulation study, suggest that the OSA is a promising new statistical methodology for association testing, when applied to the gene-environment interaction model with a continuous explanatory variable. The results from the simulation analysis conducted, suggests that the power of the OSA, when compared to each of the CEO and COO Null Distributions, is on par with the respective null hypotheses $H_0 : \beta_3 = 0$ and $H_0 : \beta_1 = \beta_3 = 0$ in the Traditional Linear Regression Analysis setting. The advantage that the OSA brings over TRLA, is the ability to "pinpoint the marker" (the value of x_1) among the continuous exposure trait which provides maximum evidence of a difference in the mean levels of the outcome between the two levels of the gene factor. The implication of this is that for values of the continuous exposure trait, which are greater than x_1, the interaction of the continuous exposure trait and the gene factor is much more evident, due to greater levels of the outcome measured in the "gene carriers" ($G_i = 1$) as compared to "non-gene carriers" ($G_i = 0$). Thus, sampling only those subjects (from a given population) whose continuous exposure trait being greater than x_1, provides the implication of greater power to detect this interaction, versus sampling from the entire population. This brings about a more efficient means by

which to sample the given population, the result of which could lead to a decrease in real world study costs.

The generalizability of the results from this investigation appear very good. Here, we note that each of the values for the statistics $r^2_{\mathrm{YG}|X,G*X}$ and $r^2_{\mathrm{YX}|G,G*X}$, were fixed to the value of 0.01. Although this would appear to rule out most real data from OSA investigation, we see that the results (power) of the OSA/CEO Null Distribution comparison should be similar for other choices of these two statistics (for each of the choices of $r^2_{Y(G*X)|G,X}$). Recall, this comparison is essentially testing the interaction effect of the predictors G and X, and other choices of $r^2_{\mathrm{YG}|X,G*X}$ and $r^2_{\mathrm{YX}|G,G*X}$ should not affect the results of this OSA/Null comparison. Moreover, deviations of the values for $r^2_{\mathrm{YG}|X,G*X}$, should increase the power of the OSA/COO Null Distribution comparison. Recall, this comparison is essentially comparing the simultaneous effects of the predictor G and the interaction of the predictors G and X. Thus, increases in the parameter value for β_1 (via increases in the value for $r^2_{\mathrm{YG}|X,G*X}$) would increase the power of the OSA, when compared to this Null Distribution. In addition, deviations in the value of the statistic $r^2_{\mathrm{YX}|G,G*X}$, should not affect the results of the OSA/COO Null Distribution comparison. Therefore, deviations in the values of the statistics $r^2_{\mathrm{YG}|X,G*X}$ and $r^2_{\mathrm{YX}|G,G*X}$ (for a given choice of $r^2_{Y(G*X)|G,X}$), should not compromise the use of the OSA methodology, and are actually welcomed.

Moreover, the robust nature of the linear regression model assumptions, should allow for OSA applicability, to data which deviate from the model assumptions set forth by expression (1) of Section 2.1. Furthermore, fluctuations in the dominant gene allele frequency should not violate the use of the OSA. However, in populations where very large or very small values for the dominant gene allele frequency exists, one must exercise caution and draw a large enough random sample, so that each subsample (comprised of each gene group) is large enough for analysis. For example, suppose that a random sample of $n = 200$ observations has been drawn from a population whose dominant gene allele frequency is 5%. Then, we expect the "gene-carrier" group to be comprised of a mere ten observations (the subsample comprised of $G_i = 1$ observations). Similarly, we expect the "non gene-carrier" group to be comprised of 190 observations (the subsample comprised of $G_i = 0$ observations). Thus, it may not be feasible to suggest the use of the OSA in this case, since the gene-carrier group subsample size is relatively small when compared to the non gene-carrier group. Along these lines, there could exist a function of the population dominant allele frequency and the random sample size, to determine the minimum sample size to be drawn, so that the OSA is suggested for use. This is something that could be investigated further in a future study.

5.2: Future Studies for the OSA

Overall, we see that this OSA investigation offers tremendous optimism for future studies, and merely "scratches the surface" of its potential. There are six points that could be investigated in future OSA studies. The first of which is allowing for the values within the vector $\boldsymbol{R^2}$ to be more dynamic. This offers a potential increased generalizability of the OSA to more real world populations, due to allowing the outcome vector Y (for the model expression (1) of Section 2.1) to change for fixed values of the vectors G, X, and E. The second thing that could be addressed in a future study is, adjusting the distribution assumptions for G and X (or even simply just modifying the parameter assumptions for the distribution assumptions for this study) as well as the independence assumption between these two predictors. This again increases the potential generalizability of the OSA to more real world populations. For our investigation, the sample size was fixed to 200 observations. So, the third thing that could be investigated in the future would be to examine the power of the OSA, relative to TRLA, as sample size fluctuates. The fourth thing that could be investigated further is the idea of obtaining other random ordering methods to which to compare the OSA. Thus, additional OSA/-Null Distributions could be investigated in the future. The final two things which could be investigated in future studies, each comprise of the notion for minimization of computational time. The first of these is investigating the idea of minimizing the number of random orderings which need to be sampled from each of the Null Distributions, while allowing for valid and accurate conclusions to be drawn

from the OSA. The second idea for minimization of computational time is investigating the minimum number of datasets which should be simulated so that valid and accurate power calculations for the OSA are obtained.

BIBLIOGRAPHY

[1] Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M (2004). Ordered Subset Analysis in Genetic Linkage Mapping of Complex Traits, *Genetic Epidemiology,* **27**, 53–63.

[2] R Development Core Team (2005). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. ISBN 3- 900051-07-0. http://www.R-project.org.

APPENDIX A

THE RELATIONSHIP BETWEEN $\hat{\gamma}_1$ AND Y

For this and the subsequent sections which follow, we assume that a random sample of n observations has been simulated, according to the methods outlined in Section 3.2 of this paper, for an arbitrary choice of the vector $\boldsymbol{R^2}$. Recall, the data for this study is generated according to the modelling scheme, given by (1) of Section 2.1 above as

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 X_i + \beta_3 G_i * X_i + E_i.$$

Let $n_1 = \sum_{i=1}^{n} G_i$, and let $n_2 = n - n_1$. For each $j \in \{1, 2\}$, we define the subsequences, $\{Y_{j_i}\}_{i=1}^{n_j}$, and $\{X_{j_i}\}_{i=1}^{n_j}$ of the respective sequence $\{Y_i\}_{i=1}^{n}$, and $\{X_i\}_{i=1}^{n}$ by

$$\mathcal{X}_i \in \{\mathcal{X}_{j_k}\}_{k=1}^{n_j} \text{ iff } G_i = 2 - j,$$

for all $i = 1, 2, \ldots, n$; $\mathcal{X} \in \{X, Y\}$. Finally, for each $j = 1, 2$, let the sample mean of the subsequence $\{\mathcal{X}_{j_i}\}_{i=1}^{n_j}$, be denoted by $\overline{\mathcal{X}}_j$. Recall that the MLE for the simple linear regression of Y on G is given by the expression

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n} \left(G_i - \overline{G}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(G_i - \overline{G}\right)^2}, \tag{3}$$

where $\overline{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i$; $\mathcal{X} \in \{G, Y\}$. Here, we have

$$\sum_{i=1}^{n}\left(G_i-\overline{G}\right)\left(Y_i-\overline{Y}\right)=\sum_{i:G_i=1}\left(1-\overline{G}\right)\left(Y_i-\overline{Y}\right)+\sum_{i:G_i=0}\left(-\overline{G}\right)\left(Y_i-\overline{Y}\right)$$

$$=\sum_{i:G_i=1}\left(Y_i-\overline{Y}\right)-\overline{G}\left(\sum_{i:G_i=1}\left(Y_i-\overline{Y}\right)+\sum_{i:G_i=0}\left(Y_i-\overline{Y}\right)\right)$$

$$=\sum_{i:G_i=1}\left(Y_i-\overline{Y}\right)-\overline{G}\left(\sum_{i=1}^{n}Y_i-n_1\,\overline{Y}-n_2\,\overline{Y}\right)$$

$$=\sum_{i:G_i=1}\left(Y_i-\overline{Y}\right)-\overline{G}\left(n\,\overline{Y}-\underbrace{(n_1+n_2)}_{=n}\,\overline{Y}\right)$$

$$=\sum_{i:G_i=1}\left(Y_i-\overline{Y}\right).$$

We also have,

$$\begin{aligned}\sum_{i=1}^{n}\left(G_i-\overline{G}\right)^2&=n_1\left(1-\overline{G}\right)^2+n_2\left(-\overline{G}\right)^2\\&=n_1-2\,n_1\,\overline{G}+(n_1+n_2)\,\overline{G}^2\\&=n_1-2\,n_1\,\overline{G}+n\,\overline{G}^2.\end{aligned}\tag{5}$$

We make the observation that $\overline{G}=\frac{n_1}{n}$, so that

$$n_1-2\,n_1\,\overline{G}+n\,\overline{G}^2=n_1\left(1-2\,\frac{n_1}{n}+\frac{n_1}{n}\right)=n_1\left(1-\frac{n_1}{n}\right).$$

Thus, from (3), (4), and (5) above we have

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n}\left(G_i - \overline{G}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(G_i - \overline{G}\right)^2}$$
$$= \frac{\sum_{i:G_i=1}\left(Y_i - \overline{Y}\right)}{n_1\left(1 - \frac{n_1}{n}\right)}$$
$$= \frac{n\left(\overline{Y}_1 - \overline{Y}\right)}{n - n_1}.$$

We also have $n\,\overline{Y} = \left(n_1\,\overline{Y}_1 + n_2\,\overline{Y}_2\right)$, so that

$$\hat{\gamma}_1 = \frac{n\left(\overline{Y}_1 - \overline{Y}\right)}{n - n_1}$$
$$= \frac{(n - n_1)\,\overline{Y}_1 - n_2\,\overline{Y}_2}{n - n_1}$$
$$= \overline{Y}_1 - \overline{Y}_2.$$

Therefore, the MLE for the regression parameter γ_1 is equal to the difference of the sample means for the continuous outcomes of the two gene factors.

APPENDIX B

THE CEO NULL DISTRIBUTION

This section investigates the expected value for the MLE of the regression parameter γ_1, for a random ordering of the n observations, according to the value of the continuous predictor X. Here, we have

$$Y_{1_i} = \beta_0 + \beta_1(1) + \beta_2 X_i + \beta_3(1) * X_i + E_i = \beta_1 + (\beta_2 + \beta_3) X_i + E_i,$$

and

$$Y_{2_k} = \beta_0 + \beta_1(0) + \beta_2 X_k + \beta_3(0) * X_k + E_k = \beta_2 X_k + E_k,$$

for all $i = 1, 2, \ldots, n_1$, and $k = 1, 2, \ldots, n_2$, and where under the simulation assumptions we have noted that $\beta_0 = 0$. Then, under the CEO Null Distribution, we have

$$\begin{aligned} \overline{Y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} (\beta_1 + (\beta_2 + \beta_3) X_i + E_i) \\ &= \beta_1 + (\beta_2 + \beta_3) \overline{X}_1 + \overline{E}_1, \end{aligned}$$

and

$$\begin{aligned} \overline{Y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} (\beta_2 X_i + E_i) \\ &= \beta_2 \overline{X}_2 + \overline{E}_2. \end{aligned}$$

Thus, we have

$$\begin{aligned} E(\overline{Y}_1) &= E\left(\beta_1 + (\beta_2 + \beta_3)\,\overline{X}_1 + \overline{E}_1\right) \\ &= \beta_1 + (\beta_2 + \beta_3)\,E(\overline{X}_1) + E(\overline{E}_1) \\ &= \beta_1, \end{aligned}$$

and

$$\begin{aligned} E(\overline{Y}_2) &= E\left(\beta_2\,\overline{X}_2 + \overline{E}_2\right) \\ &= \beta_2\,E(\overline{X}_2) + E(\overline{E}_2) \\ &= 0, \end{aligned}$$

where we have made use of the fact that, asymptotically, $E(\overline{\mathcal{X}}_i) = 0$, for each $\mathcal{X} \in \{X, E\}$, and $i \in \{1, 2\}$. It then follows that $E\left(\overline{Y}_1 - \overline{Y}_2\right) = E(\hat{\gamma}_1) = \beta_1$, under the CEO Null Distribution, as we provided in Table 4.2. Here we note that the mathematical derivations for this section were based on the assumption of the analysis of the entire dataset (I.e., all n observations were assumed to be applied to the linear regression of Y on G). However, the above argument holds for arbitrary subsets analyzed from the simple linear regression of Y on G, under asymptotic conditions. Therefore, under the CEO Null Distribution, we conclude that $E(\hat{\gamma}_1) = \beta_1$.

APPENDIX C

THE ORDER STATISTICS OF THE STANDARD NORMAL DISTRIBUTION FOR THE OSA

Recall, under the OSA, we sort the observations of a given dataset in descending order, according to the value of the continuous trait (X). It was derived in Appendix A above that the MLE for γ_1 is the difference between the sample means of the outcomes for the two gene groups. So, then from the derivations of the immediately preceding section, we have that

$$\begin{aligned} E(\overline{Y}_1) &= \beta_1 + (\beta_2 + \beta_3)\, E(\overline{X}_1) + E(\overline{E}_1) \\ E(\overline{Y}_2) &= \beta_2\, E(\overline{X}_2) + E(\overline{E}_2), \end{aligned} \tag{6}$$

where $\overline{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{j_1} Y_{1_i}$ and $\overline{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{j_2} Y_{2_i}$, where $j_i \leq n_i$, for each $i = 1, 2$. Under the OSA, we still have that $E(\overline{E}_1) = E(\overline{E}_2) = 0$. However, since the observations are sorted in descending order, according to the value of the continuous predictor X, it follows that $E(\overline{X}_i) > 0$, $i = 1, 2$. This is the case, since we expect to obtain the minimum p-value, p_1, from the simple linear regression of Y on G, for a proper subset of a given dataset. Column three of Table 4.1 above supports this claim. We see that the maximum of the median subset sample sizes, which yield the minimum p-value p_1 for the OSA, is just slightly over half of the observations. That is, since the standard normal distribution is symmetric, we expect the subset which yields the minimum p-value, p_1, to be such that $X_i > 0$ for all (or nearly all)

of the observations contained within it. Thus, to obtain the expected value of $\overline{X}_1$ and $\overline{X}_2$, it deems necessary that we examine the order statistics for the standard normal distribution.

Here, let $X \sim N(0, 1)$, and let $f_X(x)$ be the density function for X at $x \in \mathbb{R}$. Let $n \in \mathbb{N}$ be arbitrary, and let $X_1, X_2, \ldots, X_n$, be a random sample from the distribution of X. Now, for each $i = 1, 2, \ldots, n$, let $X_{(i)}$ be the i^{th} order statistic (O.S.) for the random sample $\{X_i\}_{i=1}^n$. Recall, the density function for $X_{(i)}$ at $x \in \mathbb{R}$, $f_{X_{(i)}}(x)$, is given by

$$f_{X_{(i)}}(x) = i \cdot \binom{n}{i} f_X(x) \, (F_X(x))^{i-1} \, (1 - F_X(x))^{n-i},$$

where $F_X(x)$ is the cumulative distribution function for X, at x. Let $j \in \mathbb{N}$, $j < n$, be such that $X_{(j-1)} \leq 0 \leq X_{(j)}$, and consider $\overline{X} = \frac{1}{n-j+1} \sum_{i=j}^{n} X_{(i)}$. We note that the focus of the continuous predictor X under the OSA, is on the "upper half" of the normal probability density function, and as such that portion of the density is our focus here. So, then to determine $E(\overline{X})$, it deems necessary to determine $E(X_{(i)})$, for all $j \leq i \leq n$. Then, we have

$$\begin{aligned} E(X_{(i)}) &= \int_{x_{(i)} \in \mathbb{R}} x_{(i)} \, f_{X_{(i)}}(x_{(i)}) \, d\, x_{(i)} \\ &= \int_{-\infty}^{\infty} x \cdot i \cdot \binom{n}{i} f_X(x) \, (F_X(x))^{i-1} \, (1 - F_X(x))^{n-i} \, d\, x. \end{aligned}$$

We make the substitution, $u = F_X(x)$, so that $d\, u = f_X(x) \, d\, x$.

Thus, we have

$$E(X_{(i)}) = \int_{\lim_{x\to-\infty}F(x)}^{\lim_{x\to\infty}F(x)} F_X^{-1}(u)\, i\cdot\binom{n}{i}(u)^{i-1}\,(1-u)^{n-i}\,du.$$

We define the *error function,* $\mathrm{Erf}(x)$, to be

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt,$$

so that

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}}\int_0^{\frac{x}{\sqrt{2}}} e^{-t^2}\,dt \\ &= \frac{1}{2}\left(1 + \mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\right). \end{aligned}$$

Since $u = F(x)$, this implies that $F_X^{-1}(u) = \sqrt{2}\ \mathrm{Erf}^{-1}(2u-1)$. It is noted here that the function $\mathrm{Erf}(x)$ is continuous and monotone on $\mathbb{R}^+$, so that it is a one-to-one function. This implies that the inverse function $\mathrm{Erf}^{-1}(x)$ exists. Thus, we have

$$\begin{aligned} E(X_{(i)}) &= \int_0^1 F_X^{-1}(u)\, i\cdot\binom{n}{i}(u)^{i-1}\,(1-u)^{n-i}\,du \\ &= i\cdot\binom{n}{i}\int_0^1 \sqrt{2}\ \mathrm{Erf}^{-1}(2u-1)\,(u)^{i-1}\,(1-u)^{n-i}\,du, \end{aligned} \tag{7}$$

which is not integrable by any straightforward means (at least by no means that this author is aware of). A way around this, is that of the n random variables, $\{X_i\}_{i=1}^n$,

simulated for this OSA investigation, we would expect to find a value of X_i, "centered" around each of the $\left(\frac{1}{n}\right)^{\text{th}}$ percentiles of the probability density function (pdf) of X. That is, we assume that $X_{(i)} \in \mathbb{R}$, for which

$$1 - \frac{(n-i+1)}{n+1} = F_X(X_{(i)}). \tag{8}$$

This provides that

$$\begin{aligned} X_{(i)} &= F_X^{-1}\left(1 - \frac{(n-i+1)}{n+1}\right) \\ &= \sqrt{2}\ \text{Erf}^{-1}\left(2\left(1 - \frac{(n-i+1)}{n+1}\right) - 1\right) \\ &= \sqrt{2}\ \text{Erf}^{-1}\left(1 - \frac{2\,(n-i+1)}{n+1}\right). \end{aligned} \tag{9}$$

It is noted here that the value for $F_X(X_{(i)})$, given in (8) above, is the expected value for the i^{th} O.S. for a random sample (size n) of $U(0, 1)$, the proof of which is left to the reader. This implies that we can simulate the order statistics for the standard normal distribution, from the order statistics for the uniform distribution. This makes sense, as we mentioned above that we expect to see a value of X_i, around each of the $\left(\frac{1}{n}\right)^{\text{th}}$ percentiles of the pdf for X.

APPENDIX D

THE ASYMPTOTIC BEHAVIOR OF THE OSA, PART I

So, Appendix C above, provided a means by which we can now proceed to investigate the expected values of $\overline{X}_i,\ i = 1, 2$. Now, suppose that n is large (tending toward infinity), so that we may examine the limiting behaviors of each of the expected values, $E(\overline{X}_i)$. A problem which we now face is that we are essentially focused on the maximum order statistic for the X_i, $X_{(n)}$, and the order statistics which immediately follow in descending suit. Also,

$$\lim_{n\to\infty} X_{(n-j)} = \lim_{n\to\infty} \sqrt{2}\ \mathrm{Erf}^{-1}\left(1 - \frac{2\,(j+1)}{n+1}\right) = \infty,$$

for all fixed values of $j \in \mathbb{N} \bigcup \{0\}$. We claim that $E(\overline{X}_1) - E(\overline{X}_2) \to 0$, as the subset sample size under the OSA increases. Here, we note that each of the sequences $\{X_{k_i}\}_{i=1}^{n_k}$ is comprised of the order statistics $\{X_{(i)}\}_{i=n}^{n-j}$, where $j + 1 = n_1 + n_2$, and $k = 1, 2$.

We consider the subset for the OSA, such that $\{X_{k_i}\}_{i=1}^{n_k} \subset \{X_{(i)}\}_{i=n}^{n-j}$, for a fixed $j \in \mathbb{N}$. It follows that

$$\lim_{n\to\infty} \left(E(\overline{X}_1) - E(\overline{X}_2)\right) = \infty - \infty,$$

an indeterminate form. The problem here is that the $X_{(i)}$ which satisfy (8) and (9) above are upper limits of integrals, of which have no easily obtainable closed-ended form. So, to show the desired result, we simulate a large sequence of values for $\{X_{(i)}\}_{i=1}^{n}$. We choose the value of n to be 10^9 (one billion), and simulate these values using the *Mathematica* (Version 4.0) software package, applying the formula for $X_{(i)}$ given by (9) above. In doing this, for a more efficient process, we choose to calculate only the values $\{X_{(i)}\}_{i=n/2+1}^{n}$. There are two advantages to doing this. The first of which is that under the OSA methodology, we are essentially only interested in the greater (positive) values for the X_i. So, in simulating the X_i over the entire domain of the reals, the negative values would essentially be discarded, making for extraneous, unneeded simulated values. The simulation methodology we have adopted, alleviates the simulation of these negative X_i values, thus improving the efficiency of the simulation. The second advantage to the use of this simulation algorithm is that there is no need for a sorting algorithm, following the simulation of the values for the X_i. For the value of n chosen (one billion), a sorting algorithm could require a substantial amount of time to complete (if it can even be done). Thus, we have improved the efficiency of the simulation a second time, by removing the demand of a sorting algorithm for the simulated X_i values. Following the simulation of the sequence $\{X_{(i)}\}_{i=n/2+1}^{n}$, we simulate the sequence $\{G_i\}_{i=1}^{n/2}$, under the assumptions for this study (I.e., the G_i are i.i.d. Bernoulli(0.3) random variables), and "affix" a value of G_i to each of the values of $X_{(i)}$. We are then able to examine the relationship between the two sample means $\overline{X}_1$ and $\overline{X}_2$.

The next issue that we address is how to summarize the results from this rather large simulation of data. Let n_1 and n_2 be as previously defined, with the exception that $n_1 = \sum_{i=1}^{n/2} G_i$, and $n_2 = \frac{n}{2} - n_1$. As we noted above, let $\left\{X_{(i)_j}\right\}_{i=1}^{n_j}$, be the collection of the sequence of the order statistics $\{X_{(i)}\}_{i=n/2+1}^{n}$, $j = 1, 2$, for which

$$X_{(i)} \in \left\{X_{(i)_j}\right\}_{i=1}^{n_j} \text{ iff } G_i = 2 - j.$$

For an arbitrary $i \in \mathbb{N} \bigcap \left[1, \frac{n}{2}\right]$, let $n_1^* = \sum_{k=1}^{i} G_k$, and $n_2^* = i - n_1^*$, be the number of elements for each of the gene groups, for the subset under OSA analysis. Further, for each $j = 1, 2$, let $\overline{X}_{j_i} = \frac{1}{n_j^*} \sum_{k=n_j-n_j^*+1}^{n_j} X_{(k)_j}$, the sample mean of the order statistics for each of the gene groups for the i^{th} subset for the OSA. Further, we denote the sample mean for each of the sequences $\{\overline{X}_{j_k}\}_{k=1}^{i}$ by $\overline{\overline{X}}_{j_i}$, and for the sequence $\left\{\overline{X}_{1_k} - \overline{X}_{2_k}\right\}_{k=1}^{i}$ by $\overline{\overline{X}_{1_i} - \overline{X}_{2_i}}$. Finally, we denote the standard deviation for each of the sequences $\{\overline{X}_{j_k}\}_{k=1}^{i}$, by S_{j_i}, and for the sequence $\left\{\overline{X}_{1_k} - \overline{X}_{2_k}\right\}_{k=1}^{i}$ by S_i. So, for each $i = 1, 2, \ldots, \frac{n}{2}$, we could summarize the values of $\overline{X}_1$ and $\overline{X}_2$, and the absolute value of their difference, for each of the respective subsets of the OSA, in tabular form. However, this table would be extremely burdensome to read through and analyze. Instead, for fixed values of $i = 1, 2, \ldots, \frac{n}{2}$, and for each of the sets $\left\{X_{(k)_j}\right\}_{k=n_j-n_j^*+1}^{n_j}$, we summarize the means and standard deviations for the collection of all sample means for $\overline{X}_1$, $\overline{X}_2$, and $\overline{X}_1 - \overline{X}_2$, for the subsets which would be analyzed under the OSA prior to (or at) "time i," the notation for which was defined earlier in this paragraph. These summary measures provide us with measures regarding the behaviors of $\overline{X}_1$ and $\overline{X}_2$, as the subset sample size for the

OSA increases. Table D.1 below summarizes the results from this simulation analysis.

Table D.1: Summary of the Asymptotic Behavior of the Continuous Predictor X Under the OSA*

Subset Sample Size for the OSA (i)	$\overline{\overline{X}}_{1_i} \pm \frac{S_{1_i}}{\sqrt{i}}$ (n_1^*)	$\overline{\overline{X}}_{2_i} \pm \frac{S_{2_i}}{\sqrt{i}}$ (n_2^*)	$\overline{\overline{X}_{1_i} - \overline{X}_{2_i}} \pm \frac{S_i}{\sqrt{i}}$
$0.5^{(a)}$	$2.432 \pm 0.474^{(b)}$ (14,998,837)	$2.432 \pm 0.473^{(b)}$ (35,001,163)	$10.3^{(b)} \pm 12.25^{(c)}$
1.0	2.163±0.364 (29,995,338)	2.163±0.364 (70,004,662)	7.63±6.14
1.5	1.992±0.314 (44,994,097)	1.992±0.314 (105,005,903)	5.42±4.10
2.0	1.863±0.284 (59,998,202)	1.863±0.284 (140,001,798)	5.70±3.07
2.5	1.757±0.264 (75,000,353)	1.757±0.264 (174,999,647)	6.30±2.46
3.0	1.667±0.249 (89,998,031)	1.667±0.249 (210,001,969)	6.52±2.05
3.5	1.587±0.238 (104,997,680)	1.587±0.238 (245,002,320)	6.43±1.76
4.0	1.515±0.229 (119,995,272)	1.515±0.229 (280,004,728)	6.14±1.54
4.5	1.449±0.222 (135,000,664)	1.449±0.222 (314,999,336)	5.86±1.37
5.0	1.388±0.216 (150,000,685)	1.388±0.216 (349,999,315)	5.92±1.23

* Based on the simulation of *a* sample size of one billion observations from the order statistics of the Standard Normal Distribution.

(a) Times the value of 10^8.

(b) Times the value of 10^{-4}.

(c) Times the value of 10^{-8}.

Column one of this table summarizes the subset sample size for the OSA, columns two and three summarize the respective values for $\overline{\overline{X}}_{1_i} \pm \frac{S_{1_i}}{\sqrt{i}}$ and $\overline{\overline{X}}_{2_i} \pm \frac{S_{2_i}}{\sqrt{i}}$ (as well as n_1^* and n_2^*), and column four summarizes the values of $\overline{\overline{X}_{1_i} - \overline{X}_{2_i}} \pm \frac{S_i}{\sqrt{i}}$. The results from columns two and three, suggests that the distributions of the means, $\overline{X}_1$ and $\overline{X}_2$, are asymptotically the same. Column four of this table helps to support this notion, as we see that the difference in these means appears (very small sample mean for the difference in these means, and very small values for the respective standard error of the mean difference of these means) to be converging to zero.

In addition to the above simulation analysis, here we also investigate the behavior of the summary measures provided in column four of Table D.1, for different values of n. This will provide us with more accurate conclusions, regarding the behavior of $E\left(\overline{X}_1 - \overline{X}_2\right)$ as $n \to \infty$. Table D.2 below summarizes the simulation results of the values for $\overline{\overline{X}_{1_i} - \overline{X}_{2_i}} \pm \frac{S_i}{\sqrt{i}}$, where $i = \frac{j\,\%*n}{100}$, $j \in \{k \in \mathbb{N} : k \bmod 5 = 0, \text{and } k \leq 50\}$, and where $n = 10^k$ for all $k = 2, 3, \ldots, 9$. The results of this simulation analysis, suggests that as the random sample size drawn increases, the expected value for the difference, $\overline{X}_1 - \overline{X}_2$, is converging to zero.

Table D.2: Summary of the Behavior of the Continuous Predictor X Under the OSA, for Various Random Sample Sizes Drawn

$j\%^*$	$n = 10^2$	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$	$n = 10^7$	$n = 10^8$	$n = 10^9$
5%	0.580(a) 0.440(b)	0.113 0.061	0.038 0.007	0.012 $<10^{-3}$	0.003 $<10^{-4}$	0.001 $<10^{-4}$	0.0005 $<10^{-5}$	0.0001 $<10^{-6}$
10%	0.379 0.219	0.080 0.031	0.023 0.004	0.010 $<10^{-3}$	0.003 $<10^{-4}$	0.0007 $<10^{-5}$	0.0003 $<10^{-6}$	0.00008 $<10^{-7}$
15%	0.303 0.147	0.082 0.020	0.018 0.003	0.008 $<10^{-3}$	0.003 $<10^{-4}$	0.0008 $<10^{-5}$	0.0002 $<10^{-6}$	0.00005 $<10^{-7}$
20%	0.240 0.112	0.084 0.015	0.017 0.002	0.008 $<10^{-3}$	0.003 $<10^{-4}$	0.0009 $<10^{-5}$	0.0002 $<10^{-6}$	0.00006 $<10^{-7}$
25%	0.197 0.091	0.082 0.012	0.015 0.002	0.008 $<10^{-3}$	0.002 $<10^{-4}$	0.0008 $<10^{-5}$	0.0002 $<10^{-6}$	0.00006 $<10^{-7}$
30%	0.169 0.076	0.073 0.010	0.014 0.001	0.008 $<10^{-3}$	0.002 $<10^{-4}$	0.0008 $<10^{-5}$	0.0002 $<10^{-6}$	0.00007 $<10^{-7}$
35%	0.153 0.066	0.064 0.009	0.014 0.001	0.007 $<10^{-3}$	0.002 $<10^{-4}$	0.0008 $<10^{-5}$	0.0002 $<10^{-6}$	0.00006 $<10^{-7}$
40%	0.139 0.058	0.057 0.008	0.015 0.001	0.007 $<10^{-3}$	0.002 $<10^{-4}$	0.0007 $<10^{-5}$	0.0001 $<10^{-6}$	0.00006 $<10^{-7}$
45%	0.126 0.052	0.053 0.007	0.016 0.001	0.006 $<10^{-4}$	0.001 $<10^{-4}$	0.0007 $<10^{-5}$	0.0001 $<10^{-6}$	0.00006 $<10^{-7}$
50%	0.122 0.046	0.048 0.006	0.017 0.001	0.006 $<10^{-4}$	0.001 $<10^{-5}$	0.0006 $<10^{-5}$	0.0001 $<10^{-6}$	0.00006 $<10^{-7}$

* This is the percentage of the total (n) sample size under OSA investigation. Equivalently, this percentage when multiplied by n is the subset sample size analyzed under the OSA.

(a) $\overline{X_{1_i} - X_{2_i}}$, where $i = \frac{j\% * n}{100}$.

(b) $\frac{S_i}{\sqrt{i}}$, where $i = \frac{j\% * n}{100}$.

Thus, under the OSA, the simulation analyses for Tables D.1 and D.2, suggest that $E\left(\overline{X}_1 - \overline{X}_2\right) \to 0$ as $n \to \infty$ (provided at least 5% of the observations for a given random sample, are drawn for the subset under OSA investigation). Hence, as $n \to \infty$, from (6) above we have

$$\begin{aligned} E\left(\overline{Y}_1 - \overline{Y}_2\right) &= \beta_1 + (\beta_2 + \beta_3)\, E(\overline{X}_1) + E(\overline{E}_1) - \left(\beta_2\, E(\overline{X}_2) + E(\overline{E}_2)\right) \\ &= \beta_1 + \beta_3\, E(\overline{X}_1). \end{aligned}$$

Therefore, these data suggest that the asymptotic behavior of the OSA is such that $E(\hat{\gamma}_1) = \beta_1 + \beta_3\, E(\overline{X})$, where $\overline{X} \in \left\{\overline{X}_1, \overline{X}_2\right\}$, provided that at least 5% of the total random sample be drawn for the subset under OSA investigation.

APPENDIX E

THE ASYMPTOTIC BEHAVIOR OF THE OSA, PART II

In the previous section, we investigated the asymptotic behavior of the OSA, as the random sample size tends to infinity. The results from the above analysis are generalizable to a single random sample drawn of size n. In this section we will investigate the behavior of $E(\overline{X}_1 - \overline{X}_2)$, for a fixed value of $n = 200$, where we repeatedly simulate random samples, X_i of size n, to alleviate simulation variation. This is equivalent to the investigation in the previous section. To see this, suppose that $\{X_{1_i}\}_{i=1}^{200}$, $\{X_{2_i}\}_{i=1}^{200}$, ..., $\{X_{m_i}\}_{i=1}^{200}$, are a collection of random samples of size $n = 200$, each from the common distribution X, where $X \sim N(0, 1)$ and $m \in \mathbb{N}$ is arbitrary. For each $j = 1, 2, ..., m$, and $k = 1, 2$, let $\{X_{k_{j_i}}\}_{i=1}^{n_{k_j}} \subset \{X_{j_i}\}_{i=1}^{200}$, for which

$$X_{j_i} \in \{X_{k_{j_i}}\}_{i=1}^{n_{k_j}} \text{ iff } G_i = 2 - k,$$

where X_{j_i} is substituted for X_i in the model expression given by (1) of Section 2.1 above. Thus, under the assumptions for this study,

$$E(|\{X_{k_{j_i}}\}_{i=1}^{n_{k_j}}|) = n_{k_j} = \begin{cases} 0.3\,n, \text{ if } k = 1 \\ 0.7\,n, \text{ if } k = 2' \end{cases}$$

for all j. For each choice of $k = 1, 2$, consider the grand mean of the X_i for each of the gene groups, across all of the m random samples for X drawn,

$$\frac{1}{\sum_{j=1}^{m} n_{k_j}} \sum_{j=1}^{m} \sum_{i=1}^{n_{k_j}} X_{k_{j_i}}.$$

Now, one of the assumptions for this investigation is that the predictors G and X are independent of each other, for each observation drawn. We also said above that $E\left(\left| \left\{ X_{k_{j_i}} \right\}_{i=1}^{n_{k_j}} \right| \right) = n_{k_j} = n_k$, for all j, where n_k is a fixed rational number, dependent only on the value of k. Thus,

$$E\left(\frac{1}{\sum_{j=1}^{m} n_{k_j}} \sum_{j=1}^{m} \sum_{i=1}^{n_{k_j}} X_{k_{j_i}} \right) = E\left(\frac{1}{\sum_{j=1}^{m} n_k} \sum_{j=1}^{m} \sum_{i=1}^{n_k} X_{k_{j_i}} \right)$$

$$= \frac{1}{m\, n_k} E\left(\sum_{j=1}^{m} \sum_{i=1}^{n_k} X_{k_{j_i}} \right),$$

which provides that the expected value for the grand mean of X (for each of the respective gene groups), for the m samples drawn, is essentially equivalent to the sample mean of the aggregation of the m samples of X (for each of the respective gene groups). Thus, repeatedly simulating random samples for the X_i is equivalent to simulating one dataset with $m * n$ observations. Hence, this methodology provides another (in addition to Appendix D above) means by which we can examine the asymptotic behaviors of $\overline{X}_1$ and $\overline{X}_2$. For arbitrary choices of $j \in \mathbb{N}$, $j = 1, 2, \ldots, m$, and $i \in \mathbb{N} \bigcap [1, 100]$, let $n_{1_j}^* = \sum_{k=1}^{i} G_k$, and $n_{2_j}^* = i - n_{1_j}^*$, be the number of elements for each of the gene groups, for the i^{th} subset of the j^{th} random sample under OSA analysis. Further, for each $k = 1, 2$, let $\overline{X}_{k_{j_i}} = \frac{1}{n_{k_j}^*} \sum_{w=n_{k_j}-n_{k_j}^*+1}^{n_{k_j}} X_{(w)_{k_j}}$, be the sample mean of the order statistics for each of the gene groups for the i^{th} subset of the j^{th} random sample for the OSA.

Finally, we denote the standard deviation of the sequence $\left\{\overline{X}_{1_{j_i}} - \overline{X}_{2_{j_i}}\right\}_{j=1}^{m}$ by S_i. Table E.1 below summarizes the results of simulating $m = 10^k$, $k = 1, 2, ..., 5$, random samples of $\{X_i\}_{i=1}^{200}$.

Table E.1: Summary of the Behavior of the Continuous Predictor X Under the OSA, for Repeated Random Sampling of $n = 200$ Observations.

$v\%^*$	$m = 10^1$	$m = 10^2$	$m = 10^3$	$m = 10^4$	$m = 10^5$
5%	0.223(a) 0.193(b)	0.035 0.047	0.058 0.014	0.055 0.004	0.055 0.001
10%	0.068 0.071	0.020 0.020	0.003 0.007	0.0009 0.0022	0.001 0.001
15%	0.038 0.051	0.002 0.017	0.002 0.006	0.0001 0.0018	0.0002 0.0006
20%	0.079 0.056	0.003 0.016	0.003 0.005	0.0008 0.0017	0.0001 0.0005
25%	0.045 0.054	0.002 0.015	0.005 0.005	0.001 0.002	0.0002 0.0005
30%	0.025 0.065	0.007 0.014	0.005 0.005	0.002 0.001	0.0002 0.0005
35%	0.034 0.058	0.005 0.014	0.003 0.005	0.002 0.001	0.0003 0.0005
40%	0.019 0.056	0.024 0.014	0.004 0.004	0.002 0.001	0.0002 0.0004
45%	0.004 0.050	0.030 0.014	0.0001 0.0042	0.003 0.001	0.0002 0.0004
50%	0.012 0.043	0.030 0.013	0.0005 0.0042	0.003 0.001	0.0002 0.0004

* This is the percentage of the total ($n = 200$) sample size under OSA investigation, for each of the $j = 1, 2, ..., m$ random samples simulated.

(a) Sample mean of the sequence $\left\{\overline{X}_{1_{j_i}} - \overline{X}_{2_{j_i}}\right\}_{j=1}^{m}$, where $i = \frac{v\% * n}{100}$.

(b) $\frac{S_i}{\sqrt{i}}$, where $i = \frac{v\% * n}{100}$.

The results of this table confirm that of the previous section. That is, these data suggest as more and more random samples are drawn, the difference in the sample means of $\overline{X}_1$ and $\overline{X}_2$ is approaching zero. In addition, we see that as the number of random samples increases, the standard error of the mean difference is decreasing. This implies the sample mean difference between $\overline{X}_1$ and $\overline{X}_2$ is becoming more "stable," and a more accurate inference can be ascertained from the simulation results, as the number of random samples drawn increases.

Therefore, as in the previous section, we are able to conclude that asymptotically, the expected value of $\hat{\gamma}_1 = \beta_1 + \beta_3 E(\overline{X})$, where $\overline{X} \in \{\overline{X}_1, \overline{X}_2\}$, provided that at least 5% of the total random sample be drawn for the subset under OSA investigation.

APPENDIX F

THE COO NULL DISTRIBUTION

During the simulation of the data for this study, we assumed that $P(G_i = 1) = 0.3$, for all $i = 1, 2, ..., 200$. Thus, we would expect approximately 30% of the 200 observations to be simulated to the {Aa, AA} genotype group, and 70% of the 200 observations to be simulated to the {aa} genotype group. That is, $G_i = 1$ for approximately 30% of the observations, and $G_i = 0$ for approximately 70% of the observations. So, now we consider the "random shuffling" of the outcomes (Y), under the COO Null Distribution. Recall, from Section 2.4, under the COO Null Distribution, the predictors G and X for each observation remains fixed, while the outcome vector is "shuffled." Let $M_{200\times2}$ be the predictor matrix, comprised of G and X, such that column one of M contains the values of G simulated, while column two of M contains the values of X simulated for a given dataset. Thus, when the outcome vector is shuffled, each of the values of Y_i is "matched" to a predictor vector (row) of the observations (row M_j), where $j = 1, 2, ..., 200$. Next, we note that the outcome value Y_i, matched to the predictor row M_j upon the random shuffling scheme, may or may not be such that the value of G_i which contributed to the "generation" of Y_i equals that of $M_{j,1}$. Let A be the event that a randomly shuffled value Y_i is "matched" to a predictor vector M_j, such that the value of G_i which contributed to the "generation" of the value of Y_i equals that of $M_{j,1}$. Fur-

ther, let B be the compliment of the event A. Let C be the event that a randomly chosen value of Y_i was generated from an observation whose value for G_i equals one. Similarly, let D be the compliment of C. Since the random shuffling scheme of the Y vector is independent of the data simulation, it follows that

$$P(A, C) = P(A \mid C)\, P(C) = (0.3)\,(0.3) = 0.09$$
$$P(A, D) = P(A \mid D)\, P(D) = (0.7)\,(0.7) = 0.49$$
$$P(B, C) = P(B \mid C)\, P(C) = (0.7)\,(0.3) = 0.21$$
$$P(B, D) = P(B \mid D)\, P(D) = (0.3)\,(0.7) = 0.21$$

That is, asymptotically, we would expect approximately $58\%(= 0.09 + 0.49)$ of the values for Y_i to be matched to a predictor vector M_j, such that the value of $M_{j,1}$ equals that of the value of G_i which generated Y_i. Moreover, since the Y_i are *randomly* shuffled, then

$$E\left(\overline{Y} \mid A, C\right) = E\left(\overline{Y} \mid B, C\right) = E(\overline{Y}_1) = \beta_1$$
$$E\left(\overline{Y} \mid A, D\right) = E\left(\overline{Y} \mid B, D\right) = E(\overline{Y}_2) = 0,$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are as defined in Appendix C above. Now, under the random sorting of the vector Y, let $Y_1' = \{Y_i : M_{j,1} = 1\}$, and let $Y_2' = \{Y_i : M_{j,1} = 0\}$, for all $i, j = 1, 2, \ldots, 200$. That is, the set Y_1' is the collection of the Y_i which are sorted to an observation whose predictor value for G is one, and similarly the set Y_2' is the collection of the Y_i which are sorted to an observation whose predictor value for G is zero. Also, for each $i = 1, 2$, let $\overline{Y_i'} = \frac{1}{|Y_i'|} \sum_{y_k \in Y_i'} y_k$.

Then,

$$
\begin{aligned}
E(\overline{Y_1'}) &= E\left(\overline{Y} \mid A,\, C\right) P(A \mid C)\, P(C) + E\left(\overline{Y} \mid B,\, D\right) P(B \mid D)\, P(D) \\
&= \beta_1 (1)\, (0.3) + (0)\, (1)\, (0.7) \\
&= 0.3\, \beta_1,
\end{aligned}
$$

and

$$
\begin{aligned}
E(\overline{Y_2'}) &= E\left(\overline{Y} \mid A,\, D\right) P(A \mid D)\, P(D) + E\left(\overline{Y} \mid B,\, C\right) P(B \mid C)\, P(C) \\
&= (0)\, (1)\, (0.7) + \beta_1 (1)\, (0.3) \\
&= 0.3\, \beta_1,
\end{aligned}
$$

which implies that $E\left(\overline{Y_1'} - \overline{Y_2'}\right) = 0$. But, this expected value is the estimated slope parameter for the linear regression of Y regressed on G, which is executed for the generation of the COO Null Distribution, as we have noted in Table 4.2 above.