3-28-2005

Constructing Confidence Intervals of the Summary Statistics in the Least-Squares SROC Model

Ming-Yu Fan
University of Washington, myfan@u.washington.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Key words: meta-analysis, receiver operating characteristic curve, least squares estimates

1 Introduction

To evaluate the performance of a binary-scale diagnostic test, whether its binary nature comes from a true binary outcome or dichotomization of a continuous outcome, the result is often described as a two-by-two table. From the two-by-two table, we can estimate the sensitivity (true positive rate) and the specificity (true negative rate), which measure how accurate a binary-scale diagnostic test is to detect the disease status. **R2C1 Since a single large size study is not easy to conduct, methods to combine the results from several independent studies are desired. Comparing to a single study, a careful structural review with rigorous meta-analysis can provide more reliable information for power analysis or sample size estimation for future studies. Some models can also be used explore the heterogeneity across studies [4].**

When the response of a diagnostic test is continuous, its sensitivity and specificity are derived by dichotomizing the outcome at a certain threshold. Different thresholds result in different pairs of sensitivity and specificity, and there is a trade-off between these two rates. This joint dependence of sensitivity and specificity is fully captured by the receiver operating characteristic (ROC) curve. When combining results from different independent studies, it is assumed that there exists an underlying probability distribution, and each study result corresponds to a specific threshold that determines the sensitivity and the false positive rate. These sensitivities and false positive rates are assumed to be on one common ROC curve, which is called the summary ROC (SROC) curve. If the underlying probability distribution is known, then we might only need to estimate a few parameters in order to fit a smooth SROC curve. Common methods such as maximum likelihood could be used to estimate the parameters and then the distribution. However, the underlying probability law is seldom known. **Moses, Shapiro, and Littenberg (1993) [7] proposed a least-**

squares approach to fit the smooth SROC curve for combining different studies, and estimated the variances of the coefficients using the standard method for least-squares estimators. That is, the randomness of the "independent variable" is ignored in the estimating process. This method soon became popular and has been used frequently in meta-analysis literature in the last decade. Our search shows that their work has been referred in approximately 200 papers, of which around 90 percent are meta-analysis applications in various fields. Several alternative approaches have been proposed in the rest papers to either fit the smooth SROC curve (e.g. hierarchical SROC model [9] [1] [5]) or derive summary statistics of a diagnostic test from multiple studies [10] [11]. Hierarchical regression approach is a more sophisticated method that takes into account the correlation between sensitivity and false positive rate and incorporates the intra-study and inter-study variation simultaneously. The method is, however, not widely used, which is possibly attributed to the complexity of the model and the estimating procedure. Despite of its popularity in medical research application, the validity of Moses et al.'s method is seldom evaluated in the literature. As pointed out by Rutter and Gatsonis in their paper [9], the effect of ignoring the randomness of the independent variable on the point and interval estimation is unknown and needs to be studied. Mitchell (2003) [6] conducted a Monte Carlo simulation to examine the method and concluded its validation. In the simulation the true sensitivities, true false positive rates, and true prevalences were considered known parameters. The simulated samples were created by adding random variation to the transformed as well as non-transformed true sensitivities and false positive rates. The biases were then computed as the differences between the "observed" and "true" values in sensitivities and false positive rates. In this paper we generated our simulated data using another approach. The biases and confidence coverage probabilities were computed on the regression parameters. We provided another estimation for the variance which was derived by using

the delta method, and compared our method to Moses et al.'s method in the simulation.

The paper is organized as follows. In Section 2 we introduce the notations as well as the method proposed by Moses et al. for fitting the smooth SROC curve. This is followed by the new estimation of the variances of the parameters in Section 3. The specifics of our simulation are provided in Section 4, while the results are illustrated in Section 5. A real example is shown in Section 6. We then conclude the paper with a discussion in Section 7.

2 Smooth SROC curve

Let Sen and FPR be the sensitivity and false positive rate of a diagnostic test in a particular study. Kardaun and Kardaun (1990) [3] suggested an empirical transformation that mapped (Sen, FPR), the ROC space, onto (V, U) plane, where $V = \text{logit}(Sen)$ and $U = \text{logit}(FPR)$. Under the assumption that the response of the test follows a logistic distribution, we can show that U and V are linearly related. More specifically, let X be the test's response for a diseased patient and Y be the test's response for a non-diseased patient. If we assume that X and Y follow logistic distributions and have the distribution functions,

$$F_X(x) = \left[1 + e^{-\left(\frac{x-r_1}{t_1}\right)}\right]^{-1} \quad \text{and} \quad F_Y(y) = \left[1 + e^{-\left(\frac{y-r_2}{t_2}\right)}\right]^{-1},$$

respectively, then it can be shown that at a particular threshold c, the test has the corresponding sensitivity (Sen),

$$Sen = \text{Prob}(X > c) = 1 - F_X(c) = \left[1 + \exp\left(\frac{c - r_1}{t_1}\right)\right]^{-1}, \tag{1}$$

and the false positive rate (FPR),

$$FPR = \text{Prob}(Y > c) = 1 - F_Y(c) = \left[1 + \exp\left(\frac{c - r_2}{t_2}\right)\right]^{-1}. \tag{2}$$

Therefore, $V(= (r_1 - c)/t_1)$ and $U(= (r_2 - c)/t_2)$ are linearly related. If X and Y do not follow logistic distributions exactly, the linear relationship might not be observed but may

approximately hold. The closer the true distributions are to logistic distributions, the closer the relationship between V and U is to linear. Some transformation can help in reaching better linearity. Moses, Shapiro, and Littenberg (1993) [7] suggested to further transform (V, U) into (D, S) where $D = V - U$ and $S = V + U$, and postulated a linear relationship between D and S for all possible thresholds:

$$D \;=\; \phi_0 + \phi_1 S\,. \tag{3}$$

The above linear association between D and S implies a relationship between FPR and Sen, and it can be transformed back into the ROC space. The resulting ROC curve is called the summary ROC (SROC) curve. For $\phi_1 \neq 0$,

$$\mathrm{SROC}(FPR) \;=\; \left[1 + e^{-\phi_0/(1-\phi_1)} \left(\frac{1-FPR}{FPR}\right)^{(1+\phi_1)/(1-\phi_1)}\right]^{-1}. \tag{4}$$

Equation (3) can be re-arranged as $V = \frac{\phi_0}{(1-\phi_1)} + U\,\frac{(1+\phi_1)}{(1-\phi_1)}$. Substituting V and U with the notations used in equations (1) and (2), we have the following equation:

$$\left(\frac{r_1 - c}{t_1}\right) \;=\; \frac{\phi_0}{1-\phi_1} \;+\; \left(\frac{r_2 - c}{t_2}\right)\frac{1+\phi_1}{1-\phi_1}\,. \tag{5}$$

Equation (5) demonstrates the relation of the parameters between the two logistic distributions through the common threshold "c" and the true regression parameters. When ϕ_0, ϕ_1, and $(r_2,\, t_2)$ are known, the parameters $(r_1,\, t_1)$ can be obtained according to the above equation. Let $\hat{S}_g$ and $\hat{D}_g$ be estimates of S and D in study g, where $g = 1, \ldots, G$. Fitting $(\hat{S}_g, \hat{D}_g)'$s to the linear model (3), we can derive the resulting least-squares estimators for ϕ_0 and ϕ_1 as follows:

$$\hat{\phi_1} \;=\; \frac{\boldsymbol{s}^T \boldsymbol{D}_1 \boldsymbol{d}}{\boldsymbol{s}^T \boldsymbol{D}_1 \boldsymbol{s}}, \quad \text{and} \quad \hat{\phi_0} \;=\; \frac{1}{G}\,\mathbf{1}^T \boldsymbol{d} - \frac{1}{G}\hat{\phi}_1 \cdot \mathbf{1}^T \boldsymbol{s}\,,$$

where $\boldsymbol{s} = (\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_G)^T$, $\boldsymbol{d} = (\hat{D}_1, \hat{D}_2, \cdots, \hat{D}_G)^T$, and $\boldsymbol{D}_1 = \boldsymbol{I}_{G\times G} - (1/G)\mathbf{1}\mathbf{1}^T$. The vector $\mathbf{1}$ is the G by 1 vector of 1's, and the matrix $\boldsymbol{I}_{G\times G}$ is the G by G identity matrix.

Note that $\boldsymbol{s}$ and $\boldsymbol{d}$ are functions of observed sensitivities and false positive rates. Therefore, they are both random variables. Moses et al. (1993) suggested to estimate the variance

of $\hat{\phi}_0$ and $\hat{\phi}_1$ by ignoring the variance of $\boldsymbol{s}$, as in regression models, and thus the variance of $\hat{\phi}_1$ equaled to $\mathrm{Var}(\hat{D}_g)/(\boldsymbol{s}^T\boldsymbol{D}_1\boldsymbol{s})$.

In addition, Moses et al. suggested yet another summary measure Q, which was defined as the point where $Q = Sen = (1 - FPR)$. Explicitly, $Q = (1+e^{-\phi_0/2})^{-1}$, and was estimated with $Q^* = (1 + e^{-\hat{\phi}_0/2})^{-1}$. Moses et al. also showed that the standard error of Q^* could be estimated as

$$\frac{SE(\hat{\phi}_0)}{8\,[\mathbf{cosh}(\phi_0/4)]^2}\,,$$

where $\mathbf{cosh}(\cdot)$ is a hyperbolic cosine function such that $\mathbf{cosh}(x) = \big[\mathbf{exp}(x) + \mathbf{exp}(-x)\,\big]/2$.

One problem with Moses' variance estimation of $\hat{\phi}_0$ and $\hat{\phi}_1$ is that the vector of independent variables, $\boldsymbol{s}$, is random. Hence their method may underestimate the true variances.

3 Estimation of the variances

In this section we will derive the new variance and covariance formulae for $\hat{\phi}_0$ and $\hat{\phi}_1$ by taking into account the additional variation due to the randomness of $\boldsymbol{s}$. Since both $\hat{\phi}_0$ and $\hat{\phi}_1$ are functions of $\boldsymbol{d}$ and $\boldsymbol{s}$, using the delta method we can approximate the variance-covariance matrix of $\sum\left(\hat{\phi}_0(\boldsymbol{d},\boldsymbol{s}),\,\hat{\phi}_1(\boldsymbol{d},\boldsymbol{s})\right)$ as the following:

$$\Sigma(\hat{\phi}_0,\hat{\phi}_1) \approx \left(\begin{array}{cc} \partial\hat{\phi}_0/\partial\boldsymbol{d} & \partial\hat{\phi}_1/\partial\boldsymbol{d} \\ \partial\hat{\phi}_0/\partial\boldsymbol{s} & \partial\hat{\phi}_1/\partial\boldsymbol{s} \end{array}\right)^T\Bigg|_{\substack{\boldsymbol{d}=\boldsymbol{\mu}_{\boldsymbol{d}} \\ \boldsymbol{s}=\boldsymbol{\mu}_{\boldsymbol{s}}}} \left(\begin{array}{cc} \mathrm{Var}(\boldsymbol{d}) & \mathrm{Cov}(\boldsymbol{d},\boldsymbol{s}) \\ \mathrm{Cov}(\boldsymbol{d},\boldsymbol{s}) & \mathrm{Var}(\boldsymbol{s}) \end{array}\right) \left(\begin{array}{cc} \partial\hat{\phi}_0/\partial\boldsymbol{d} & \partial\hat{\phi}_1/\partial\boldsymbol{d} \\ \partial\hat{\phi}_0/\partial\boldsymbol{s} & \partial\hat{\phi}_1/\partial\boldsymbol{s} \end{array}\right)\Bigg|_{\substack{\boldsymbol{d}=\boldsymbol{\mu}_{\boldsymbol{d}} \\ \boldsymbol{s}=\boldsymbol{\mu}_{\boldsymbol{s}}}},$$

where $\boldsymbol{\mu_d} = \text{Exp}(\boldsymbol{d})$, $\boldsymbol{\mu_s} = \text{Exp}(\boldsymbol{s})$, and

$$\begin{aligned}
\left.\frac{\partial\hat{\phi}_0}{\partial \boldsymbol{d}}\right|_{\substack{\boldsymbol{d}=\boldsymbol{\mu_d}\\ \boldsymbol{s}=\boldsymbol{\mu_s}}} &= \frac{1}{G}\mathbf{1} - \left(\frac{\mathbf{1}^T\boldsymbol{\mu_s}/G}{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s}}\right)\boldsymbol{D}_1\boldsymbol{\mu_s} \\
\left.\frac{\partial\hat{\phi}_0}{\partial \boldsymbol{s}}\right|_{\substack{\boldsymbol{d}=\boldsymbol{\mu_d}\\ \boldsymbol{s}=\boldsymbol{\mu_s}}} &= \left(-\frac{\mathbf{1}^T\boldsymbol{\mu_s}/G}{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s}}\right)\boldsymbol{D}_1\boldsymbol{\mu_d} + \left(-\frac{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_d}/G}{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s}}\right)\mathbf{1} + \frac{(2\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_d})(\mathbf{1}^T\boldsymbol{\mu_s}/G)}{(\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s})^2}\boldsymbol{D}_1\boldsymbol{\mu_s} \\
\left.\frac{\partial\hat{\phi}_1}{\partial \boldsymbol{d}}\right|_{\substack{\boldsymbol{d}=\boldsymbol{\mu_d}\\ \boldsymbol{s}=\boldsymbol{\mu_s}}} &= \left(\frac{1}{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s}}\right)\boldsymbol{D}_1\boldsymbol{\mu_s} \\
\left.\frac{\partial\hat{\phi}_1}{\partial \boldsymbol{s}}\right|_{\substack{\boldsymbol{d}=\boldsymbol{\mu_d}\\ \boldsymbol{s}=\boldsymbol{\mu_s}}} &= \left(\frac{1}{\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s}}\right)\boldsymbol{D}_1\boldsymbol{\mu_d} - \frac{2(\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_d})}{(\boldsymbol{\mu_s}^T\boldsymbol{D}_1\boldsymbol{\mu_s})^2}\boldsymbol{D}_1\boldsymbol{\mu_s}\,.
\end{aligned}$$

When the sample size is large enough, the observed sensitivity, $\widehat{Sen}_g$ $(g = 1, 2, \cdots, G)$ and the observed false positive rate, $\widehat{FPR}_g$, are asymptotically normally distributed following Central Limit Theorem. That is,

$$\begin{aligned}
\widehat{Sen}_g &\longrightarrow N\Big[Sen_g,\ Sen_g(1-Sen_g)/n_{1g}\Big] \\
\widehat{FPR}_g &\longrightarrow N\Big[FPR_g,\ FPR_g(1-FPR_g)/n_{2g}\Big],
\end{aligned}$$

where n_{1g} and n_{2g} are the sample sizes of the diseased and non-diseased groups, respectively. It follows that the transformed variables $\hat{V}$ and $\hat{U}$ are also asymptotically normally distributed:

$$\begin{aligned}
\hat{V}_g &= \text{logit}(\widehat{Sen}_g) \longrightarrow N(\mu_{1g},\ \sigma_{1g}) \\
\hat{U}_g &= \text{logit}(\widehat{FPR}_g) \longrightarrow N(\mu_{2g},\ \sigma_{2g}),
\end{aligned}$$

where $\mu_{1g} = \text{logit}(Sen_g)$, $\sigma_{1g} = [n_{1g}Sen_g(1-Sen_g)]^{-1}$, $\mu_{2g} = \text{logit}(FPR_g)$, and $\sigma_{2g} = [n_{2g}FPR_g(1-FPR_g)]^{-1}$. Denote $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iG})^T$ and $\boldsymbol{\Sigma}_i = diag(\sigma_{i1}, \sigma_{i2}, \cdots, \sigma_{iG})$ $(i = 1, 2)$, then we can show that the vectors of $\hat{V}_i's$ and $\hat{U}_i's$ have asymptotically normal distributions:

$$\begin{aligned}
\boldsymbol{v} &= (\hat{V}_1, \hat{V}_2, ..., \hat{V}_G)^T \longrightarrow N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \\
\boldsymbol{u} &= (\hat{U}_1, \hat{U}_2, ..., \hat{U}_G)^T \longrightarrow N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2).
\end{aligned}$$

Assume that $\widehat{Sen}$ and $\widehat{FPR}$ are independent, then $\boldsymbol{v}$ and $\boldsymbol{u}$ are independent as well. Since $\boldsymbol{s} = \boldsymbol{u} + \boldsymbol{v}$, its expected value ($\boldsymbol{\mu}_s$) and variance ($\boldsymbol{\Sigma}_s$) are $\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$, respectively. Similarly, for $\boldsymbol{d} = \boldsymbol{v} - \boldsymbol{u}$, the expected value is $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and $\mathrm{Var}(\boldsymbol{d}) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$. The covariance between $\boldsymbol{d}$ and $\boldsymbol{s}$ equals $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ under the assumption of independence between $\widehat{Sen}$ and $\widehat{FPR}$.

The predicted SROC value at a given FPR_0 can be derived from equation (4) with the estimated coefficients, $\hat{\phi}_0$ and $\hat{\phi}_1$. Since equation (4) is a function of ϕ_0 and ϕ_1, the variance of the fitted SROC curve can be estimated with the delta method again. With any fixed FPR_0, the variance of the corresponding predicted SROC value, denoted $\widehat{SROC}_0 = \widehat{SROC}(FPR_0)$, can be approximated as the following:

$$\mathrm{Var}(\widehat{SROC}_0) \approx \left(d_0 , \quad d_1 \right)\Big|_{\substack{\hat{\phi}_0=\phi_0 \\ \hat{\phi}_1=\phi_1}} \Sigma\left(\hat{\phi}_0, \hat{\phi}_1\right) \left(d_0 , \quad d_1 \right)^T\Big|_{\substack{\hat{\phi}_0=\phi_0 \\ \hat{\phi}_1=\phi_1}},$$

where $d_0 = \partial(\widehat{SROC}_0)/\partial\hat{\phi}_0$ and $d_1 = \partial(\widehat{SROC}_0)/\partial\hat{\phi}_1$.

The programs to derive the variances and confidence intervals based on the method described above are written in R and S-plus, and are available upon request. With an average modern computer, the result of a meta-analysis with 60 data points can be returned within several seconds.

4 Simulation

We conducted a simulation study to compare the confidence intervals constructed with two different variance estimation methods. The number of studies (G) to be included in the meta-analysis, as well as the true values of ϕ_0 and ϕ_1, were determined before each simulating process. Therefore, every simulated data set consisted of G observations. Each observation represented a study and included several elements of the study: the sensitivity (Sen_g), the false positive rate (FPR_g), the sample size for the diseased group (n_{1g}), and the sample size for the non-diseased group (n_{2g}). **Because we wanted to allow the sample sizes of different studies to vary, we generated G variables from a log-normal distribution**

with mean equal to 5 and standard deviation equal to 0.25, and rounded up these variables to represent the sample sizes for the G studies. The reason we chose log-normal distribution was to generate a skewed distribution of the sample size. The mean was around 153; the first and third quartiles were 125 and 176, respectively; and the probability of generating a sample size greater than 300 was less than 0.2%. This was done separately for disease and non-disease groups. Thus the numbers of diseased and non-diseased subjects could be different within each study, as well as across studies.

Next we generated G false positive rates. With the pre-determined ϕ_0 and ϕ_1, the corresponding sensitivities could be derived. In order to create a perfect linear relation between D and S, we generated the G false positive rates from a logistic distribution. Without loss of generality, we assumed that $r_2 = 0$ and $t_2 = 1$ in (2). Then G uniform variables were generated as the thresholds "c", and plugging these variables into (2) we obtained the G false positive rates. For each false positive rate, the corresponding sensitivity was derived from a logistic distribution with the same "c" value. The values of r_1 and t_1 in equation (1) could be computed according to (5), hence the logistic distribution for X and consequently the sensitivity could be determined.

The sensitivities and false positive rates described in the last paragraph were related in perfect linearity, and thus the sum of the squared residuals from the least squares model (3) would be zero. They represented the scenario when there were no random errors in measuring the true sensitivities and false positive rates, and that they both followed logistic distributions. In order to approximate real samples, we added random noises to the sensitivities and false positive rates. For each study g, the previously generated sensitivity was considered as the "true" sensitivity. We then generated a variable from a binomial distribution with study-specific sample size of the diseased group (from the aforementioned log-normal distribution) and the "true" sensitivity. That is, an W was generated from $Binomial(n_{1g}, Sen_g)$. The "observed" sensitivity was computed as the ratio of that variable over the sample size, W/n_{1g}. Moses et al. suggested to add 0.5 adjustment to the numerator and 1 to the denom-

inator to avoid probabilities that were zero or one. We followed this advice and performed simulations with and without this correction to explore its effect on both point and interval estimations. The "observed" false positive rate was generated in a similar matter. This adjustment was also added to the example analyzed in Section 6.

After the "observed" sensitivities and false positive rates were generated, we computed the estimates for ϕ_0, ϕ_1, $SROC(0.1)$, and Q, and constructed confidence intervals using two different methods. The simulation was repeated 1000 times with each given set of the parameters (ϕ_0, ϕ_1, G). We tried two different values for ϕ_0, two different values for ϕ_1, two different values for G, and two options with the 0.5 correction. Hence there were in total 16 simulation outcomes. **The true SROC curves corresponding to the four combinations of ϕ_0 and ϕ_1 are shown in Figure 1. We also controlled the range of the uniform distribution from which the thresholds "'c" were generated to change the intra-study variation. The broadest range allowed for the uniform distribution was determined by the true regression coefficients, the parameters r_i and $t_i(i = 1, 2)$ in (1) and (2), and the constraints which required the true sensitivities and specificities to be greater than or equal to 0.5. When this range was used for the uniform distribution, the intra-study variation dominated the inter-study variation by about 10 folds. We then tried a smaller range for the uniform distribution to examine the result when intra-study and inter-study variation were similar.** We chose $(\phi_0, \phi_1) = (2.6, -0.5)$ for the first simulation because they were the estimates from the data analysis in Section 6. Later we increased ϕ_0 to 4 and/or increased ϕ_1 to 0 to examine how different linear relations affected the coverage probabilities of the confidence intervals. Two different numbers of studies, 10 and 50, were used in the simulation.

5 Result

The simulation result can be found in Table 1 and Table 2. The overall coverage

probability of the intercept is better than the coverage probability of the slope, and is relatively unsensitive to the changes of other parameters. **The intercept is usually of better interest than the slope because the summary statistic, Q, is a function of the intercept.** In the result the coverage probability of Q does not vary as much as that of the slope, either. The biases of the regression coefficients decrease as the number of studies increase. The effect of the 0.5 correction on the biases is small comparing to the effect of the number of studies. **Although our method incorporates the variance of the "independent variable" (s) and Moses et al.'s method does not, it is still possible to observe a wider confidence interval for Q or A using the latter method. It is because the delta method accounts for both variances and the covariance between $\hat{S}$ and $\hat{D}$, which is equal to $\mathbf{Var}(logit(\widehat{Sen})) - \mathbf{Var}(logit(\widehat{FPR}))$. The covariance between $\hat{S}$ and $\hat{D}$ can be negative if the variance of $logit(\widehat{Sen})$ is smaller than the variance of $logit(\widehat{FPR})$. When the covariance is negative with large absolute value, the variance derived by using the delta method can be smaller than the variance derived by ignoring the randomness of the independent variable.**

We summarize the overall coverage probability as well as the stratified coverage probability in order to examine the marginal effect of the number of studies, 0.5 correction, and the choice of true regression coefficients. The result is shown in Table 3. The upper half and the lower half of Table 3 are derived from Table 1 and Table 2, respectively. When the intra-study variation dominates the inter-study variation, the overall coverage probability, which is the mean of all the 16 coverage probabilities in Table 1, is slightly higher using our method, although both methods appear to underestimate the interval. The averaged coverage probability stratified by the number of studies suggests that the differences caused by the number of studies is moderate (ranges from 0% to 5%), except for the slope (8% and 11%). Higher number of studies is associated with lower coverage, and our method is more sensitive to the number than Moses et

al.'s method. Adding the 0.5 correction does not show substantial effect on the coverage probability (the difference for the slope: 4% and 7%; for others: from 0% to 3%). Lower value of true ϕ_0 is associated with higher coverage probability, and its effect is stronger on the slope than other parameters. Similar relation is observed on the true ϕ_1. When the true ϕ_1 is -0.5, the coverage probability is higher than it is when the true ϕ_1 is 0, except for the slope. There seems to be an interactive effect between the values of true ϕ_0 and ϕ_1 on the coverage probability of the slope (Table 1). The coverage of the slope is the lowest when the true ϕ_0 is 4 and the true ϕ_1 is -0.5, but this pattern is not observed on the coverage probability of the other 3 parameters.

When the intra-study variation and the inter-study variation are similar, the overall coverage probability of the confidence interval increases if constructed using our method, but it says almost unchanged when the confidence interval is constructed based on Moses et al.'s method. As a result, the coverage probability of our method is noticeably higher than the one based on Moses et al.'s method. Lower number of studies, no 0.5 correction, lower ϕ_0, and lower ϕ_1 are associated with higher coverage probability, and the effects are all considered small to moderate (range: 0% to 7%).

Generally speaking, the confidence interval constructed using our method has higher coverage probability than the confidence interval constructed according to the method proposed by Moses et al., and the difference is larger when the inter-study variation and the intra-study variation are similar. Although in this scenario our method tends to overestimate the confidence interval, the coverage probability of the two important summary parameters Q and A is close to the nominal level. Moses et al.'s method is relatively unsensitive to the ratio of intra-study and inter-study variation, and consistently underestimates the confidence interval in both scenarios. When the intra-study variation dominates the inter-study variation (ratio$\approx$10), the coverage probability of the confidence

interval falls below 95%, regardless of the method. The number of studies and the 0.5 correction also affect the coverage probability, although the effect does not appear to be substantial. Our result indicates that ϕ_0 and ϕ_1 might have joint influence on the coverage probability. Further exploration is needed to determine why the confidence interval of the slope is severely underestimated when specific combination of ϕ_0 and ϕ_1.

6 Example

In 1999, Congress required polygraph screening examinations of over 1,300 employees at Department of Energy weapons laboratories because of the concerns about security violations. However, the accuracy of polygraph examinations was still of question, and consequently the results were rarely admissible as evidence in court. In January 2001 the National Research Council's Committee on National Statistics convened a committee to review relevant research and assess the scientific validity of polygraph examinations from a structured literature review. The resulting report, "The Polygraph and Lie Detection" was published in year 2002 [8]. We analyzed the polygraph data, which consisted of 59 studies of polygraph and lie detection conducted by various authors from year 1959 to 2000. The sample size ranged from 12 to 252. The true behavior of each study subject was recorded in a binary format, while the result of polygraph could be binary or ordinal with three categories: non-deceptive, intermediate, and deceptive. We merged the non-deceptive and intermediate groups to form a binary testing result. The choice was arbitrary, and there was no drastic change, e.g. changes of the signs of the regression coefficients, when we combined the intermediate and deceptive groups. One study recorded the result in 10 categories, and thus was excluded from our analysis. After re-categorizing the ordinal outcome into a binary variable, the smallest and largest sensitivities observed among the studies were 0.53 and 0.96, respectively. For false positive rate, the observed values ranged from 0.008 to 0.643.

The analysis is shown in Table 4. The least-squares estimates of ϕ_0 and ϕ_1 are 2.63 and

-0.5, respectively. The 95% confidence interval for ϕ_0 based on Moses' method is (2.20, 3.06), which is not much different from the confidence interval constructed with our estimation of the variance, (2.28, 2.99). If $\boldsymbol{s}$ is treated as a fixed vector, the confidence interval for ϕ_1 constructed according to this assumption is narrower than the one based on the random $\boldsymbol{s}$ assumption. The fitted sensitivity corresponding to the case when the false positive rate equals to 0.1 is 0.74, and the confidence interval based on Moses' method is slightly wider than ours. The Q^* statistic is 0.79, and the confidence intervals generated from the two methods are similar. **Figure 2 shows the plot of the data points and the smooth SROC curve. The analysis was repeated by excluding the three data points which had false positive rate greater than 0.5, and the change in the result was small. Therefore, we performed the analysis with all 58 data points.**

Figure 3 (a) shows a scatterplot of $B = \mathbf{log}[Sen/(1-Sen)] - \mathbf{log}[FPR(1-FPR)]$ **versus** $S = \mathbf{log}[Sen/(1-Sen)] + \mathbf{log}[FPR/(1-FPR)]$**, and Figure 3 (b) illustrates the residuals from fitting the data with the linear model (3). From Figure 3 (a) we can see that the data points probably do not follow a linear trend. The zero mean assumption appears to be violated since in Figure 3 (b) the spread of the residuals is not symmetric to the zero line. The sensitivity and specificity of the two data points associated with the highest residuals are** $(0.96, 0.98)$ **and** $(0.96, 0.99)$**, respectively. The residuals of these two data points remain the highest when the three data points with low specificity are excluded. The variance varies substantially across different** S **and hence the homogeneity assumption is likely to be violated, too.**

7 Discussion

Moses, Shapiro, and Littenberg (1993) [7] proposed to fit the smooth SROC curve using least-squares estimates, and ignored randomness of the independent variable when estimating the variances. From our simulation it showed that

these estimates were biased. Moses et al. (1993) [7] and Mitchell (2003) [6] both examined the effect of the 0.5 correction on the biases and concluded that the effect was substantial when the sample sizes were small. In our simulation we observe that when the sample sizes are larger (around 150), the effect of the 0.5 correction on the biases is negligible, which is consistent with the aforementioned findings. However, the biases remain large when the sample sizes of the studies increase. This is probably due to the "attenuation effect" of the measurement error [2]. Assuming that the observed independent variable Z is in fact a linear combination of the real independent variable X and a random error u. The expected value of the regression coefficient estimated without accounting for the randomness of Z equals to the true coefficient times a factor k, where $k = \mathbf{Var}(X)/(\mathbf{Var}(X) + \mathbf{Var}(u))$. Because k is less than or equal to one, the regression coefficient is underestimated. This phenomenon is also called "attenuation". The magnitude of the bias depends on the ratio of the variance of u over the variance of X. The larger the ratio is, the larger the bias will be. In the model proposed by Moses et al., the measurement error might not be of the linear form, but the issue of measurement error is still existent. The estimation without correcting the measurement error may subject to biases.

When the research interest focus on comparing two or more ROC curves, the issue of biases is likely to remain present. Since the bias (attenuation effect) is a function of the variances of the random error and the independent variable, taking the difference of the two ROC curves is unlikely to cancel out the biases, unless the two underlying distributions have the same variance components. The variance of the difference between the two ROC curves can be estimated as the sum of the two individual variances if the underlying distributions are assumed to be independent, or the estimation can incorporate the covariance between the two ROC curves if the underlying distributions are assumed to be correlated. For example, the covariance between two estimates of Q can be derived as suggested

by Zhou, Obuchowski, and McClish (2002) [12]. Without properly correcting the biases, however, the poor coverage is likely to be observed, no matter how the variance is estimated.

The coverage probability of the confidence intervals constructed based on our method is generally higher than the ones constructed using the method proposed by Moses et al. When the intra-study variation is about 10-times higher than the inter-study variation, both methods appear to underestimate the confidence intervals, although the coverage probability using our method is closer to the nominal level. When the intra-study variation is similar to the inter-study variation, the coverage probability of the 95% confidence interval based on Moses et al.'s method is still around 90%, while the coverage probability of our method is a little over 95%. Number of studies, 0.5 correction, and the choice of true regression coefficients have relatively small impact on the confidence coverage comparing to the ratio of intra-study variation over inter-study variation.

Although our method shows slightly better coverage probability than Moses et al.'s method, there is yet space to improve. Our method is sensitive to the ratio of intra-study variation over inter-study variation, which can result in both over-estimation and under-estimation. However, our method is still recommended since its overall coverage probability is closer to the nominal level than the confidence coverage by Moses et al.'s method, and it is easy to be implemented. Although several other advanced models have been proposed to fit the smooth SROC curve, the method proposed by Moses et al. remains to be extremely popular in meta-analysis literature because of its elegancy and simplicity. Methods that allow more flexible and robust estimation, e.g. hierarchical regression model [9], usually also require higher computing power and advanced statistical knowledge, and hence are appealing to a smaller group of investigators. We believe that, with a better estimation of the variance, the confidence coverage using Moses et al.'s method will improve substantially if the bias of the point es-

timate can be corrected. Approaches such as accounting for measurement error might lead to a more satisfactory estimation since it incorporates the random variation of the independent variable. This will be one of the directions of our future research.

Acknowledgement

We would like to thank Peter B. Imrey for providing us with the example data set.

References

[1] Dukic, V., and Gatsonis, C., *Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds.* Biometrics, **59**(4), 936-946, (2003).

[2] Fuller, W. A., *Measurement error models.* Wiley & Sons, New York, (1987).

[3] Kardaun, J. W. P. F. and Kardaun, O. J. W. F., *Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation.* Methods of Information in Medicine, **29**, 12-22, (1990).

[4] Lijmer, J. G., Bossuyt P. M. M. and Heisterkamp S. H., *Exploring sources of heterogeneity in systematic reviews of diagnostic tests.* Statistics in Medicine, **21**, 1525-1537, (2002).

[5] Macaskill, P., *Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis.* Journal of Clinical Epidemiology, **57**, 925-932, (2004).

[6] Mitchell, M. D., *Validation of the summary ROC for diagnostic test meta-analysis: a Monte Carlo simulation.* Academic Radiology, **10**(1), 25-31, (2003).

[7] Moses, L. E., Shapiro, D. and Littenberg, B., *Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations.* Statistics in Medicine, **12**, 1293-1316, (1993).

[8] National Research Countil, *The Polygraph and Lie Detection.* Committee to Review the Scientific Evidence on the Polygraph. Devision of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, (2002).

[9] Rutter, C. M. and Gatsonis, C. A., *A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations.* Statistics in Medicine, **20**, 2865-2884, (2001).

[10] Suzuki, S., Moro-oka, T., and Choudhry, N. K., *The conditional relative odds ratio provided less biased results for comparing diagnostic test accuracy in meta-analysis.* Journal of Clinical Epidemiology, **57**(5), 461-469, (2004).

[11] Zhou, X-H, Brizendine, E. J., and Pritz, M. B., *Methods for combining rates from several studies.* Statistics in Medicine, **18**, 557-566, (1999).

[12] Zhou, X-H, Obuchowski N. A., and McClish, D. K., *Statistical methods in diagnostic medicine.* Wiley, John & Sons, Incorporated, (2002).

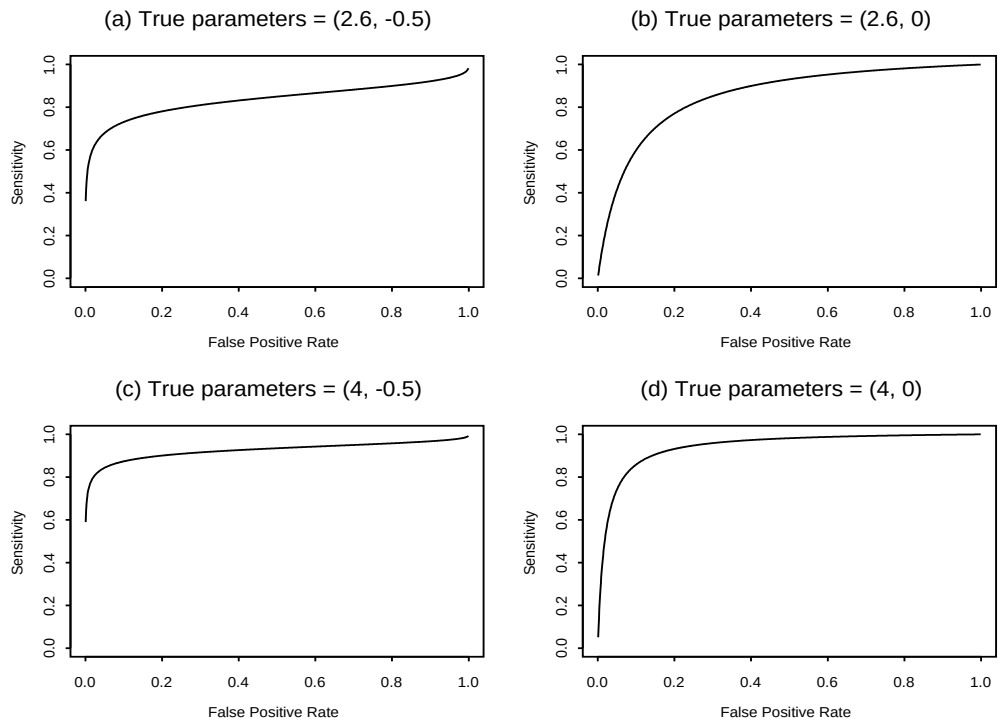


Figure 1: True ROC curves of the simulated data.

Table 1: Simulation result: mean biases and coverage probability, when intra-study variation dominates the inter-study variation

# of studies	with 0.5 correction	Biases		Coverage % - Moses et al.				Coverage % - Our method			
		ϕ_0	ϕ_1	ϕ_0	ϕ_1	Q	A	ϕ_0	ϕ_1	Q	A
True (ϕ_0, ϕ_1)=(2.6, -0.5)											
10	yes	0.12	0.06	92	88	92	92	95	97	95	96
10	no	0.12	0.05	92	91	92	91	96	98	96	96
50	yes	0.06	0.04	89	69	88	95	89	85	89	96
50	no	0.06	0.03	90	83	90	94	90	94	90	96
True (ϕ_0, ϕ_1)=(2.6, 0)											
10	yes	0.09	0.06	91	88	91	91	95	96	95	95
10	no	0.09	0.06	91	89	91	89	94	96	94	93
50	yes	0.04	0.03	91	87	90	93	90	93	90	96
50	no	0.05	0.03	87	89	87	87	87	94	87	91
True (ϕ_0, ϕ_1)=(4, -0.5)											
10	yes	0.20	0.21	91	25	89	91	91	24	89	91
10	no	0.21	0.19	90	35	89	90	90	32	88	89
50	yes	0.08	0.19	92	0	92	94	89	0	89	92
50	no	0.08	0.18	95	0	94	94	92	0	92	88
True (ϕ_0, ϕ_1)=(4, 0)											
10	yes	0.14	0.07	90	82	88	88	95	96	94	94
10	no	0.13	0.06	90	88	88	89	96	98	95	94
50	yes	0.08	0.04	82	69	80	85	86	88	84	88
50	no	0.08	0.03	81	90	80	79	86	99	84	83

$(*) A = \widehat{SROC}(0.1)$

Table 2: Simulation result: mean biases and coverage probability, when the intra-study variation and the inter-study variation are similar

# of studies	with 0.5 correction	Biases		Coverage % - Moses et al.				Coverage % - Our method			
		ϕ_0	ϕ_1	ϕ_0	ϕ_1	Q	A	ϕ_0	ϕ_1	Q	A
True (ϕ_0, ϕ_1)=(2.6, -0.5)											
10	yes	0.36	0.21	91	90	90	93	99	99	98	99
10	no	0.37	0.21	88	86	89	90	99	99	99	98
50	yes	0.15	0.09	91	89	91	94	98	98	99	98
50	no	0.17	0.10	90	83	90	95	97	96	98	98
True (ϕ_0, ϕ_1)=(2.6, 0)											
10	yes	0.08	0.23	91	91	91	87	96	99	96	92
10	no	0.08	0.23	91	91	91	85	96	99	96	92
50	yes	0.04	0.10	92	92	92	93	93	99	92	97
50	no	0.04	0.10	91	92	90	89	92	99	92	96
True (ϕ_0, ϕ_1)=(4, -0.5)											
10	yes	0.49	0.18	90	91	93	87	99	99	99	99
10	no	0.50	0.18	91	91	92	89	99	99	99	99
50	yes	0.23	0.08	89	91	92	89	98	99	99	98
50	no	0.23	0.09	89	89	92	89	92	99	99	99
True (ϕ_0, ϕ_1)=(4, 0)											
10	yes	0.11	0.24	91	88	89	89	96	98	95	92
10	no	0.10	0.24	91	89	91	87	97	98	96	90
50	yes	0.06	0.11	86	88	86	89	87	97	87	92
50	no	0.06	0.11	84	87	83	80	85	98	84	83

$(*) A = \widehat{SROC}(0.1)$

Table 3: Simulation result: summary over the 16 simulations

	Averaged coverage % of the 95% confidence interval							
	Moses et al.				Our method			
	ϕ_0	ϕ_1	Q	A	ϕ_0	ϕ_1	Q	A
When the intra-study variation dominates the inter-study variation:								
Overall	90	67	89	90	91	74	91	92
Number of studies: 10	91	73	90	90	94	80	93	94
50	88	61	88	90	89	69	88	91
With correction: Yes	90	64	89	91	91	72	91	94
No	90	71	89	89	91	76	91	91
$\phi_0 = 2.6$	90	86	90	92	92	94	92	95
4	89	49	88	89	91	55	89	90
$\phi_1 = -0.5$	91	49	91	93	92	54	91	93
0	88	85	87	88	91	95	90	92
When the intra-study variation and the inter-study variation are similar:								
Overall	90	89	90	89	95	98	96	95
Number of studies: 10	91	90	91	88	98	99	97	95
50	89	89	90	90	93	98	94	95
With correction: Yes	90	90	91	90	96	99	96	96
No	89	89	90	88	95	98	95	94
$\phi_0 = 2.6$	91	89	91	91	96	99	96	96
4	89	89	90	87	94	98	95	94
$\phi_1 = -0.5$	90	89	91	91	98	99	99	99
0	90	90	89	87	93	98	92	92

$(*) A = \widehat{SROC}(0.1)$

Table 4: Meta-analysis on the Polygraph Data

Parameter	Estimate	95% Confidence Interval	
		Moses' method	Our method
ϕ_0	2.63	(2.20, 3.06)	(2.28, 2.99)
ϕ_1	-0.50	(-0.80, -0.19)	(-0.89, -0.11)
SROC(0.1)	0.74	(0.68, 0.79)	(0.70, 0.77)
Q	0.79	(0.75, 0.82)	(0.76, 0.82)

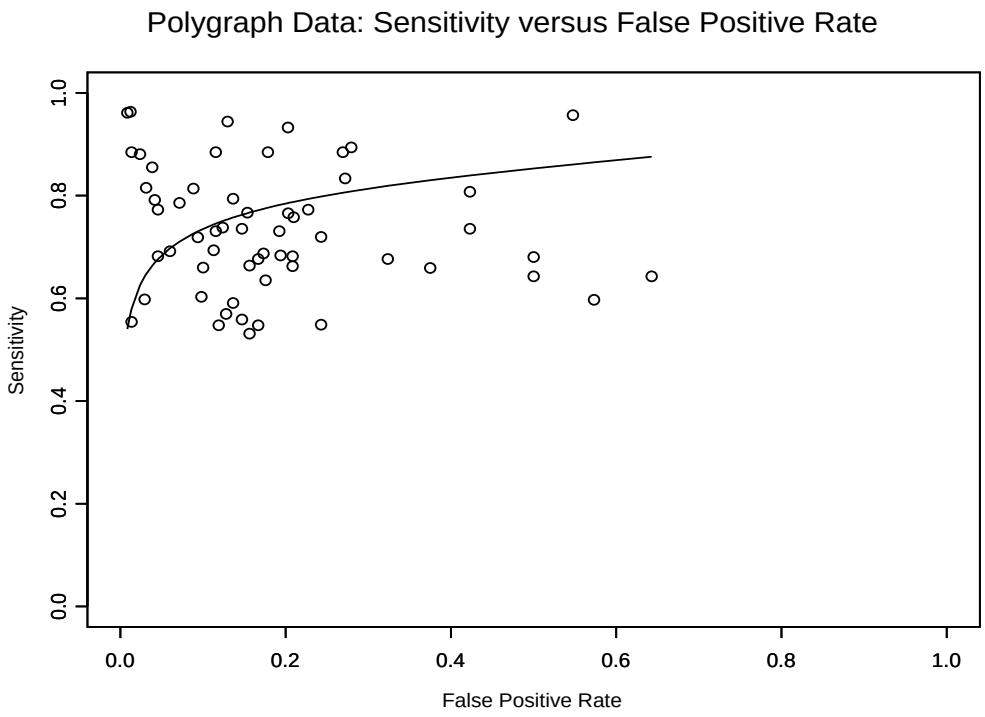
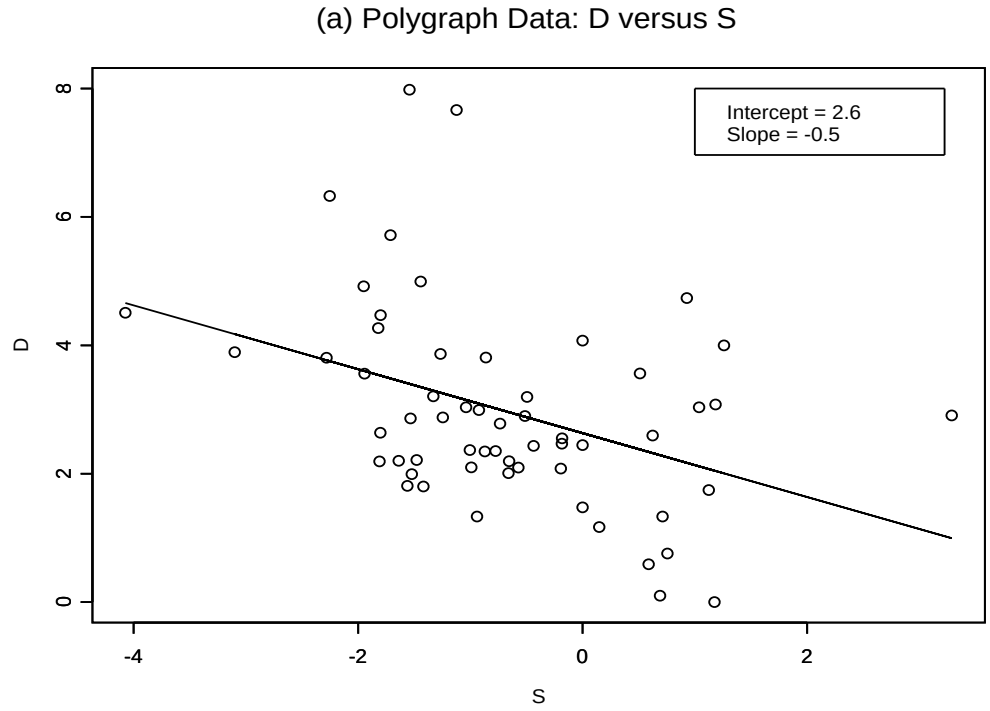


Figure 2: Descriptive plot of the Polygraph Data.



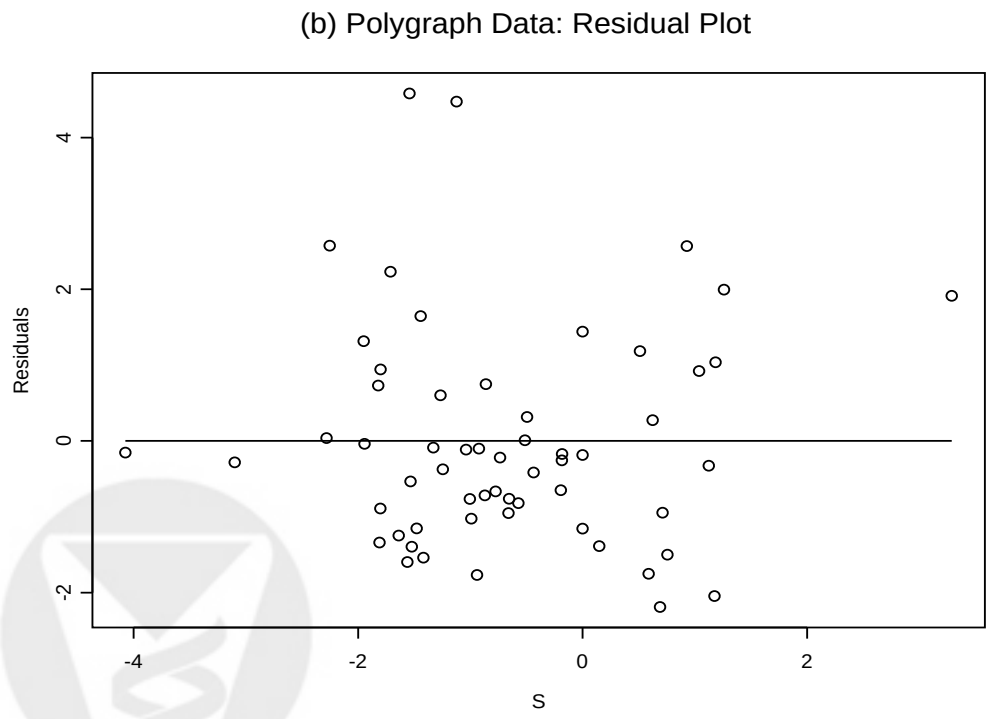


Figure 3: (a) Descriptive plot and (b) residual plot from the linear model (3).