BIOS 736: HMWK 1

Thomas Hsiao

Question 1

Let Y be the response vector, W the observed covariate matrix, and X the true covariate matrix. The two definitions of non-differential measurement error (ME) are

1. Y and W are conditionally independent given X. $f(Y|X,W) = f(Y|X)$.
2. $f(W|X,Y) = f(W|X)$

Then 1 implies 2 because

$$f(W|Y,X) = \frac{f(Y,X,W)}{f(Y,X)} = \frac{f(Y|X,W)f(X,W)}{f(Y|X)f(X)} = \frac{f(Y|X)f(X,W)}{f(Y|X)f(X)} = f(W|X)$$

Question 2

We repeat the simulation study from Lecture 1 pg. 8 and pg. 10 using $n = 1000$ and setting the true value of $\beta = 0$.

For the first simulation, since the Y vector is no longer dependent on the X vector, the non-differential measurement error does not bias the β anymore and both models estimate an effect of around 0. The standard error is higher due to W being noisier than X, but the test result for $\beta = 0$ is still valid.

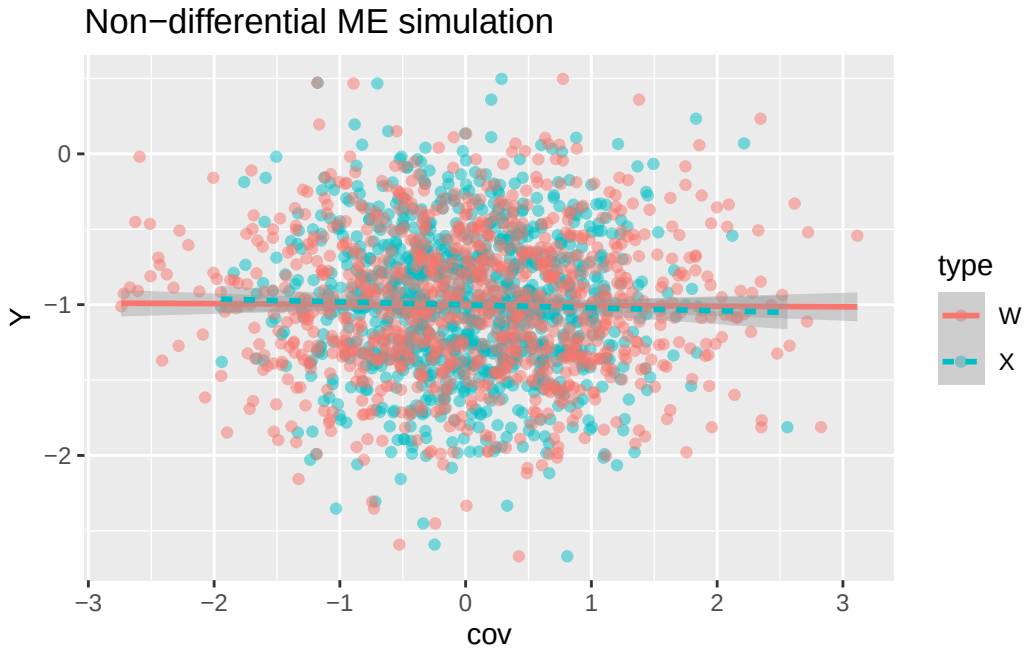
Table 1: Simulation 1: Non-differential ME

term	estimate	std.error	statistic	p.value	model
(Intercept)	-1.0014	0.0153	-65.6486	0.0000	true
X	-0.0195	0.0217	-0.9002	0.3682	true
term	estimate	std.error	statistic	p.value	model
(Intercept)	-1.0016	0.0153	-65.6009	0.0000	naive
W	-0.0041	0.0151	-0.2737	0.7844	naive

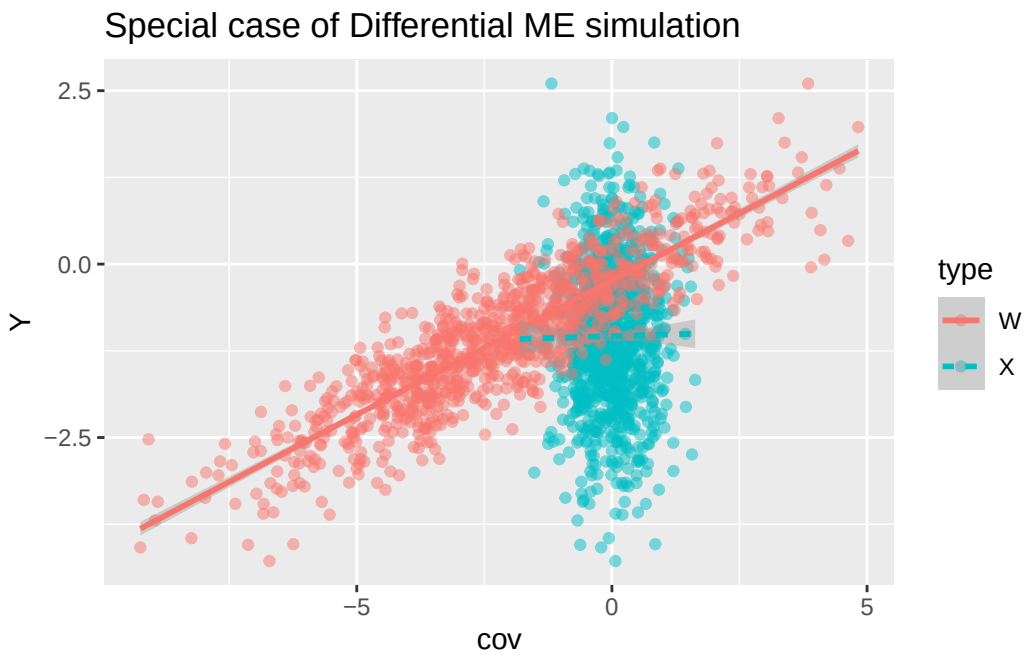
Table 2: Simulation 2: Special case of Differential ME

term	estimate	std.error	statistic	p.value	model
(Intercept)	-1.040	0.0328	-31.6856	0.0000	true
X	0.022	0.0611	0.3595	0.7193	true
term	estimate	std.error	statistic	p.value	model
(Intercept)	-0.2361	0.0210	-11.26	0	naive
W	0.3871	0.0067	57.91	0	naive

```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```



For the second simulation, the β is biased and results in an invalid hypothesis test for $\beta = 0$.

Question 3

In the original paper, Morrissey and Spiegelman (1999) compare several methods for exposure misclassification methods in case control studies with an internal validation dataset. The methods include the inverse matrix and matrix methods for 2 x 2 analyses, as well as maximum likelihood through a numerical procedure. The main tension to solving the ME problem was that while the MLE appears to be more efficient compared to the other methods, it is difficult to implement due to requiring numerical procedures that were not widely available to applied epidemiologists, the chief audience, at the time. The authors concluded MLE was the most efficient and recommended it to be seriously considered despite its practical issues. However, in the reader's response Lyles revealed through a clever reparametrization that the MLE and inverse matrix methods are equivalent under the assumption of differential misclassification. In other words - a closed form estimator of the MLE exists under the guise of a previous method. This result opened the door for epidemiologists to conduct inference under differential ME with full confidence in the efficiency of a fast estimator, without requiring complicated numerical routines.

Question 4

Part A

The naive OR is

$$OR_{naive} = \frac{n_{11}n_{02}}{n_{12}n_{01}} = \frac{50 * 87}{104 * 55} = 0.7605$$

95% CI is (0.4719, 1.226)

$$L = \exp\left\{\log(OR) - 1.96 * \sqrt{n_{11}^{-1} + n_{12}^{-1} + n_{01}^{-1} + n_{02}^{-1}}\right\}$$

$$U = \exp\left\{\log(OR) + 1.96 * \sqrt{n_{11}^{-1} + n_{12}^{-1} + n_{01}^{-1} + n_{02}^{-1}}\right\}$$

```
n11 <- 50
n02 <- 87
n12 <- 104
n01 <- 55

n13 <- 14
n14 <- 14
n15 <- 10
n16 <- 58
n03 <- 29
n04 <- 3
n05 <- 8
n06 <- 68

OR <- n11*n02 / (n12 * n01)
SE <- sqrt(1/n11 + 1/n02 + 1/n12 + 1/n01)
CI <- exp(log(OR) + 1.96*c(-SE, SE))
```

Part B

```
SE1 <- n13 / (n13+n15)
SE0 <- n03 / (n03+n05)

SP1 <- n16 / (n16 + n14)
SP0 <- n06 / (n06 + n04)
```

SE is sensitivity and SP is specificity. Since we are asked to verify whether the measurement error is non-differential, we can compute the four parameters in the internal validation dataset.

The two sensitivities are 0.583 for $D = 1$ and 0.784 for $D = 0$. The two specificities are 0.806 for $D = 1$ and 0.958 for $D = 0$. Since the sensitivities and specificities differ as a function of the response, the measurement error is most likely differential.