Linear Mixed Model Approach to Protein Significance Analysis

JONGSOO JUN, TAESUNG PARK*
Department of Statistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826
KOREA

Abstract: Discovering protein biomarkers is one of the important issues in biomedical researches. The enzyme-linked immunosorbent assay (ELISA) is one of the traditional techniques for protein quantitation. Recently, the multiple reaction monitoring (MRM) mass spectrometry has been proposed as a new method for protein quantification and has been popular as an alternative to ELISA. However, not many analysis methods are available yet to analyse MRM data. Linear mixed models (LMMs) are effective in analysing MRM data. MSstats is one of the most widely used tools for MRM data analysis which is based on the LMMs. MSstats is well implemented on Skyline program and R programming language. However, LMMs often provide various significance results depending on model specification. Thus, sometimes it would be difficult to specify a right LMM for the analysis of MRM data. In this paper, we systematically investigated the effect of model specification on significance of proteins through simulation studies. Our results provide a practical guideline of using LMMs for MRM data analysis.

Key-Words: Linear mixed model, MRM, MSstats, Power, Protein significance analysis, Skyline

1 Introduction

Discovering protein biomarkers is one of the primary interest in biomedical researches [1]. The enzyme-linked immunosorbent assay (ELISA) is one of the traditional protein quantitation techniques that provide high sensitivity [2]. The result of ELISA is treated as the “gold standard” for targeted protein quantification [3]. However, recent studies discovered many novel proteins and the availability of highly qualified ELISAs for those proteins is limited [4], which created a need for a different technique of protein quantitation. The multiple reaction monitoring (MRM) mass spectrometry is a new method used in tandem mass spectrometry for a systematic development of targeted protein assays [1]. As an alternative to ELISA, MRM assays become gradually used in systems biology and in clinical investigations [6]. For the MRM data analysis, two sample t-test or paired t-test was applied to identify proteins that would change in abundance between two groups [8]. To test for multiple groups, one-way analysis of variance (ANOVA) was employed [9]. Recently, a linear mixed model (LMM) approach was proposed for MRM data analysis and implemented in MSstats [10] and has been popularly used [11]. The proposed LMM approach treats either or both subject and run effect as random or fixed. However, we observed that the proposed LMM approach often provides various *p*-values for the same data depending on which effects are treated as random or fixed. Moreover, the data structure also affects the performance of LMM approach. If intensity patterns of peptides from a protein are heterogeneous, there is a loss of power in LMM approach. If intensity patterns are homogeneous, there is a power gain in LMM approach.

In this paper, we systematically investigated the effect of model specification on significance of proteins through simulation studies. Our results provide a practical guideline of using LMMs for MRM data analysis.

2 Methods

For a given protein, the LMM used in MSstats is given as follows:

$$y_{i,j(i),k,l} = \mu + G_i + S(G)_{j(i)} + F_k + R_l + (G \times F)_{i,k} + (F \times R)_{k,l} + \epsilon_{i,j(i),k,l} \quad (1)$$

where $y_{i,j(i),k,l}$ denotes $\log_2(intensity)$ value of the *j*-th subject nested in the *i*-th group of the *k*-th peptide and the *l*-th run; μ is a global mean, G_i stands for the *i*-th group effect; $S(G)_{j(i)}$ stands for the *j*-th subject effect nested in *i*-th group; F_k stands for the *k*-th peptide effect; R_l stands for the *l*-th run effect, $(G \times F)_{ik}$ stands for interaction effect of the *i*-th group and the *k*-th peptide; $(F \times R)_{kl}$ stands for interaction effect of the *k*-th peptide and the *l*-th run. When all effects are treated as fixed, these parameters have the following restrictions: $\sum_{i=0}^{2} G_i = 0$, $\sum_{j=0}^{J} S(G)_{j(i)} = 0$, $\sum_{k=1}^{K} F_k = 0$, $\sum_{l=1}^{L} R_l = 0$, $\sum_{i=0}^{2}(G \times F)_{i,k} = 0$, $\sum_{k=1}^{K}(G \times F)_{i,k} = 0$, $\sum_{k=1}^{K}(F \times R)_{k,l} = 0$ and $\sum_{l=1}^{L}(F \times R)_{k,l} = 0$, and $\epsilon_{i,j(i),k,l} \sim N(0, \sigma_\epsilon^2)$. Here, G_0 stands for the effect of reference group of MRM data. When the subject effects and the run effects are treated as random, the restrictions of $S(G)_{j(i)}$, R_l and $(F \times R)_{kl}$ are replaced by $S(G)_{j(i)} \sim N(0, \sigma_S^2)$, $R_l \sim N(0, \sigma_R^2)$ and $(F \times R)_{kl} \sim N(0, \sigma_{F\times R}^2)$, respectively.

The Model (1) can be equivalently written as follows:

$$y_i = \beta_0 + g_i'\beta_G + s_i'\beta_S + f_i'\beta_F + r_i'\beta_R + (g \times f)_i'\beta_{G\times F} + (r \times f)_i'\beta_{R\times F} + \epsilon_i$$

Here, y_i is a $\log_2(intensity)$ value of the *i*-th sample; g_i is a $(G \times 1)$ group indicator variable; G stands for the number of groups except the reference group; s_i is a $(N \times 1)$ subject indicator variable, where N stands for the number of subjects except the reference sample; f_i is a $(K - 1 \times 1)$ peptide indicator variable, where K stands for the number of peptides; r_i is a $(R - 1 \times 1)$ run indicator variable, where R stands for the number of MS runs; $(g \times f)_i$ is a interaction of group and peptide indicator variable; $(r \times f)_i$ is a interaction of run and peptide indicator variable; ϵ_i is an error term that follows normal distribution with mean 0 and variance σ_ϵ^2. β_S, β_R and $\beta_{R\times F}$ are coefficients of subject, run and interaction of run and peptide, respectively. These coefficients can be treated either as fixed or random.

In most MRM data analyses, the interest lies in determining proteins that differ from groups. Thus, the hypothesis of interest is given below for comparing two groups:

$$H_0: K(\beta_{G(1)} - \beta_{G(2)}) + \sum_{k=2}^{K}(\beta_{G(1)\times F(k)} - \beta_{G(2)\times F(k)}) = 0 \quad (2)$$

where $\beta_{G(a)}$ is the coefficient of the group a and $\beta_{G(a)\times F(b)}$ is the interaction coefficient of group a and peptide b. Here, $\beta_{G(a)}$ is equal to $G_a - G_0 + (G \times F)_{(a,1)} - (G \times F)_{(0,1)}$ and $\beta_{G(a)\times F(b)}$ to $(G \times F)_{(a,b)} - (G \times F)_{(a,1)}$. Thus, the hypothesis (2) is equivalent to $H_0: G_1 = G_2$

3 Simulations

3.1 Simulation Settings

We performed simulation studies to investigate the performance of LMMs. There are four LMMs depending on how to specify the random or fixed effect: (i) LMM(FF) with fixed subject effect and fixed run effect, (ii) LMM(FR) with fixed subject effect and random run effect, (iii) LMM(RF) with random subject effect and fixed run effect, and (iv) LMM(RR) with random subject effect and random run effect. For each simulated data set, the best LMM, LMM(best), was selected among four LMMs which had the smallest Akaike Information Criterion (AIC) value. We generated parameters of model (1) either as random or fixed effects. We generated random effects from the identical normal distribution independently with mean 0 and a specific variance.

On the other hand, for the fixed effect, we generated equally spaced sequence such that the average of the sequence is 0 and its squared average is the same with the value of the variance that we specified to generate random effect. In model (1), the global mean, μ, was set to 15 and σ_ϵ^2 , the variance of $\epsilon_{i,j(i),k,l}$, was set to 0.5 throughout all simulations. $S(G)_{j(i)}$, F_k, R_l and $(F \times R)_{kl}$ are nuisance parameters to test hypothesis (2). Therefore, we fixed their effects throughout simulations; their variances for random effect or squared averages for fixed effect were set to 0.25, 0.1, 0.25 and 0.1 for $S(G)_{j(i)}$, F_k, R_l and $(F \times R)_{kl}$, respectively. The number of peptides was assumed to vary from 2 to 5 in order to investigate the effect of the number of peptides on type I error and power. Various sample sizes, 20, 50 and 100, were considered with the ratio of case and control fixed to 1:1.

3.1.1 Settings for Generating Random Effects

We considered four scenarios for generating the group effect and the interaction effect as follows.

Scenario 1: $G_2 - G_1 = 0$ and $(G \times F)_{ik} = 0$ for all *i*, *k*

Scenario 2: $G_2 - G_1 = 1$ and $(G \times F)_{ik} = 0$ for all i, k
Scenario 3: $G_2 - G_1 = 0$ and $\mathrm{Var}\{(G \times F)_{ik}\} = 0.2$
Scenario 4: $G_2 - G_1 = 0.5$ and $\mathrm{Var}\{(G \times F)_{ik}\} = 0.1$

The first scenario was to investigate the type I error rate of LMMs for testing group difference; the second to the fourth scenarios were considered to evaluate empirical power of LMMs when only group effect was present (Scenario 2), only the interaction effect was present (Scenario 3), and both group and interaction effects were present (Scenario 4). All simulation data sets were generated 1,000 times from model (1) and the significance level was set to 0.05 through simulations.

3.1.2 Settings for Generating Fixed Effects

Basically, the settings for generating fixed effects were identical to those of random effects explained in section 3.1.1 except that the fixed sequences were used as previously described. That is as follows.

Scenario 5: $G_2 - G_1 = 0$ and $(G \times F)_{ik} = 0$ for all i, k
Scenario 6: $G_2 - G_1 = 1$ and $(G \times F)_{ik} = 0$ for all i, k
Scenario 7: $G_2 - G_1 = 0$ and $\mathrm{Var}\{(G \times F)_{ik}\} = 0.2$
Scenario 8: $G_2 - G_1 = 0.5$ and $\mathrm{Var}\{(G \times F)_{ik}\} = 0.1$

Here, $\mathrm{Var}\{(G \times F)_{ik}\}$ stands for the squared average of a fixed effect. Since the effects were fixed, we could control the interaction effect of groups and peptides more easily. We considered three types of interaction model (IM). (i) The first model IM1 was when the peptide effects between two groups are the same, that is, $(G \times F)_{2,1} - (G \times F)_{1,1} = \cdots = (G \times F)_{2,K} - (G \times F)_{1,K} > 0$. The number of peptides was varied from 2 to 5. (ii) The second model IM2 was when the number of peptides was assumed to be four among which two peptides have positive effects andthe other two have negative effects. The effect sizes of peptides between two groups were assumed to be the same. Thus, the interaction effect becomes $(G \times F)_{2,1} - (G \times F)_{1,1} = (G \times F)_{2,2} - (G \times F)_{1,2} = (G \times F)_{1,3} - (G \times F)_{2,3} = (G \times F)_{1,4} - (G \times F)_{2,4}$. (iii) The third model IM3 was the case when the number of positive $(G \times F)_{2,k} - (G \times F)_{1,k}$ was assumed to vary from 1 to 4. Here, the number of peptides was assumed to be five.

3.2 Simulation Results

3.2.1 Results for Random Effects

Figure 1 shows the type I error rate of LMMs when the effects were random. Among four models, LMMs with random subject effect, LMM(RF) and LMM(RR), controlled type I error well, while LMMs with fixed subject effect, LMM(FF) and LMM(FR), did not. LMM(FR) showed the highest type I error rate. The AIC value of LMM(FF) tended to be the smallest among four LMMs. Thus, LMM(FF) was most frequently selected as the best LMM and accordingly its behavior was very similar to that of LMM(FF).

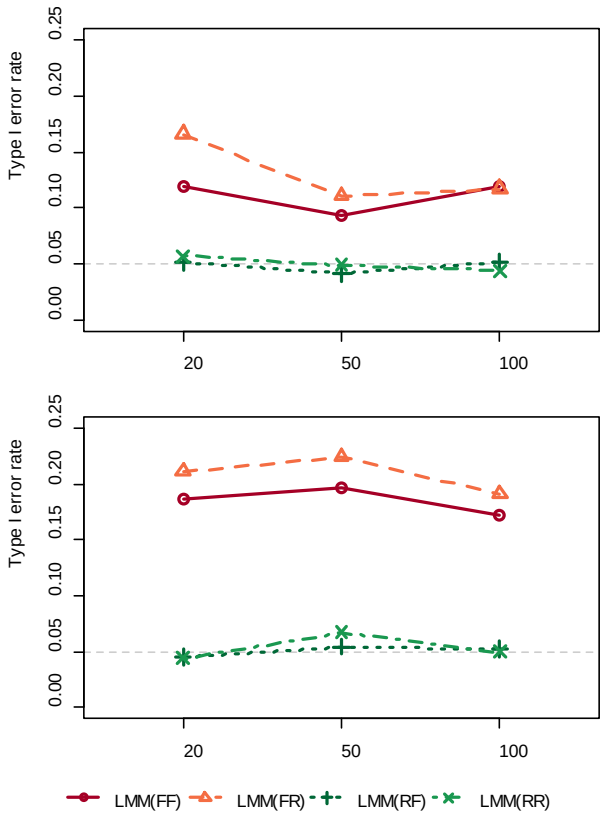

Figure 1. Type I error rate of LMMs when effects were randomly generated.
The number of peptides was 2 and 5 for the top panel and the bottom panel, respectively. The x-axis and y-axis represent the number of samples and type I error rate, respectively. Grey dotted horizontal line represents the significant level.

Regarding the effect of the number of peptides on the type I error, there are some differences among the models. For LMM(RF) and LMM(RR), there is no effect of the number of peptides, while LMM(FF) and LMM(FR) showed inflated type I error rates. LMM(best) was not affected by the number of peptides either. Since LMM(FF) and LMM(FR) did not control type I error well, their power was higher than those of LMM(RF) and LMM(RR) for scenarios 2 to 4, as shown in Figure 2. LMM(FR) showed higher power than LMM(FF), as similarly observed in Scenario 1. LMM(RR) showed relatively higher power than LMM(RF). The behaviors of the LMM(best) for scenarios 2 to 4 were similar to what we observed in

Scenario 1. Four LMMs showed consistent power patterns under all simulated cases for scenarios 2 to 4. That is, the power of LMMs has the following ordering: LMM(FR) > LMM(FF) > LMM(RR) > LMM(RF).

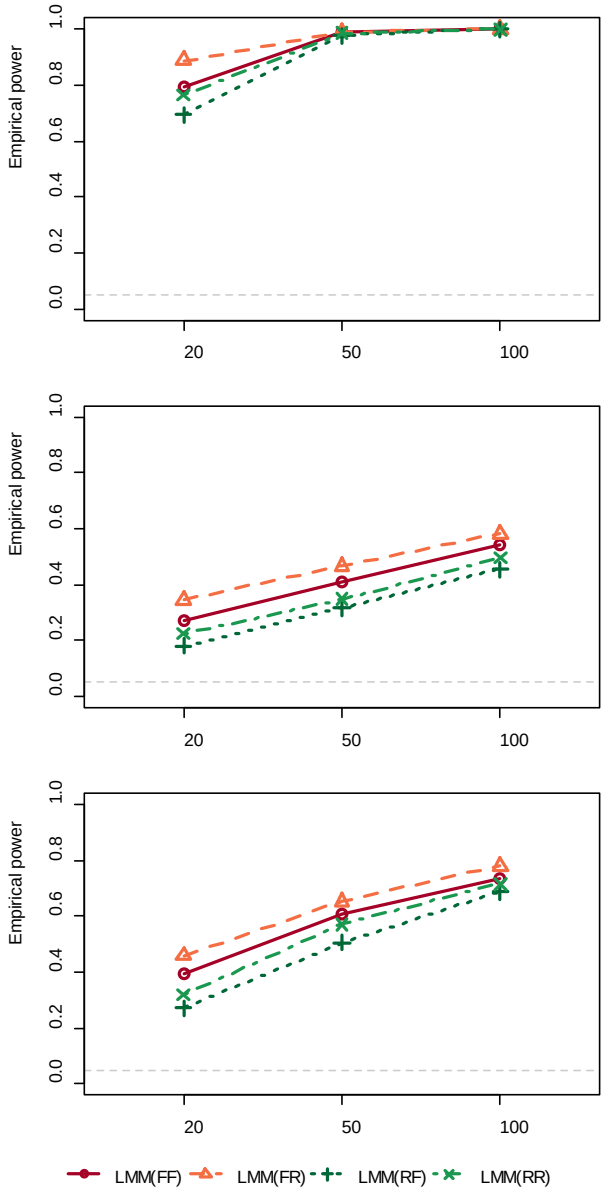

Figure 2. Estimated empirical power of LMMs when effects were randomly generated.
Top panel (Scenario 2): $\boldsymbol{G_2 - G_1 = 1}$ and $\mathbf{Var}\{(\boldsymbol{G \times F})_{ik}\} = \mathbf{0}$. Middle panel (Scenario 3): $\boldsymbol{G_2 - G_1 = 0}$ and $\mathbf{Var}\{(\boldsymbol{G \times F})_{ik}\} = \mathbf{0.2}$. Bottom panel (Scenario 4): $\boldsymbol{G_2 - G_1 = 0.5}$ and $\mathbf{Var}\{(\boldsymbol{G \times F})_{ik}\} = \mathbf{0.1}$. The number of peptides was 2. The x-axis and y-axis represent the number of samples and estimated empirical power, respectively. Grey dotted horizontal line represents the significant level.

The effect of sample sizes on power depended on the group and interaction effects, but showed very consistent patterns for all LMMs. When the group effect $G_2 - G_1 = 1$ and the interaction effect $\mathrm{Var}\{(G \times F)_{ik}\} = 0$, the sample size 20 yielded power of 0.6, while the sample size 50 yielded higher than 0.8. When the group effect $G_2 - G_1 = 0.5$ and the interaction effect $\mathrm{Var}\{(G \times F)_{ik}\} = 0.1$, the sample size of 100 yielded power higher than 0.6. On the other hand, when the group effect $G_2 - G_1 = 0$ and the interaction effect $\mathrm{Var}\{(G \times F)_{ik}\} = 0.2$, the sample size of 100 produced power lower than 0.6.

3.2.2 Results for Fixed Effects

Figure 3 shows the type I error rate of LMMs when the effects were fixed (Scenario 5). The type I error rates of LMMs with fixed effects showed different patterns from those of LMMs with random effects. All four LMMs controlled type I error well, as shown in Figure 3. LMM(RF) and LMM(RR) tend to control type I error rate more strongly than LMM(FF) and LMM(FR). Regarding the effect of the number of peptides on the type I error, there is no strong effect of the number of peptides.

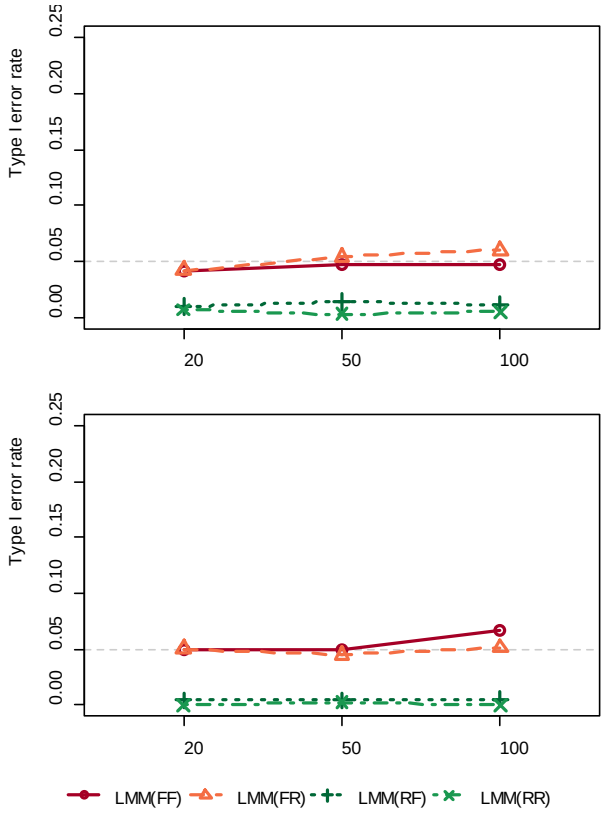

Figure 3. Type I error rate of LMMs when effects were fixedly generated.
The number of peptides was 2 and 5 for the top panel and the bottom panel, respectively. The x-axis and y-axis represent the number of samples and type I error rate, respectively. Grey dotted horizontal line represents the significant level.

The power comparison results are summarized in Figures 4 and 5. The effect of sample sizes on power depended on the group and interaction effects. Figure 4 shows the results for the interaction model IM1, when the number of peptides was two. Four LMMs showed consistent power patterns under all simulated cases for scenario 6 to 8. The power of LMM(FF) and LMM(FR) was relatively higher than those of LMM(RF) and LMM(RR) for scenarios 6 to 8. LMM(FR) showed the highest power among four LMMs. The power of LMM(RF) was slightly higher than that of LMM(RR).

The effect of sample sizes on power depended on the group and interaction effects, but showed very consistent patterns for all LMMs. Sample size 50 yielded power of 0.8 for scenarios 6 to 8.

Figure 5 shows the results for the interaction model IM2. Here, the number of peptides was assumed to be four; two peptides have positive effects and the other two have negative effects. As expected, the opposite direction of effects dramatically decreased power of all four LMMs.

Figure 6 shows the results for the interaction model IM3. The number of peptides was fixed to be five. Here, $G_2 - G_1 = 0$ and the squared average of $(G \times F)_{ik}$ was set to 0.2. As the number of positive $(G \times F)_{2,k} - (G \times F)_{1,k}$ increases, the power of all LMMs increases. The increase patterns of LMM(FF) and LMM(FR) are different from those of LMM(RF) and LMM(RR). The power of LMMs has the following ordering: LMM(FR) > LMM(FF) > LMM(RF) > LMM(RR).

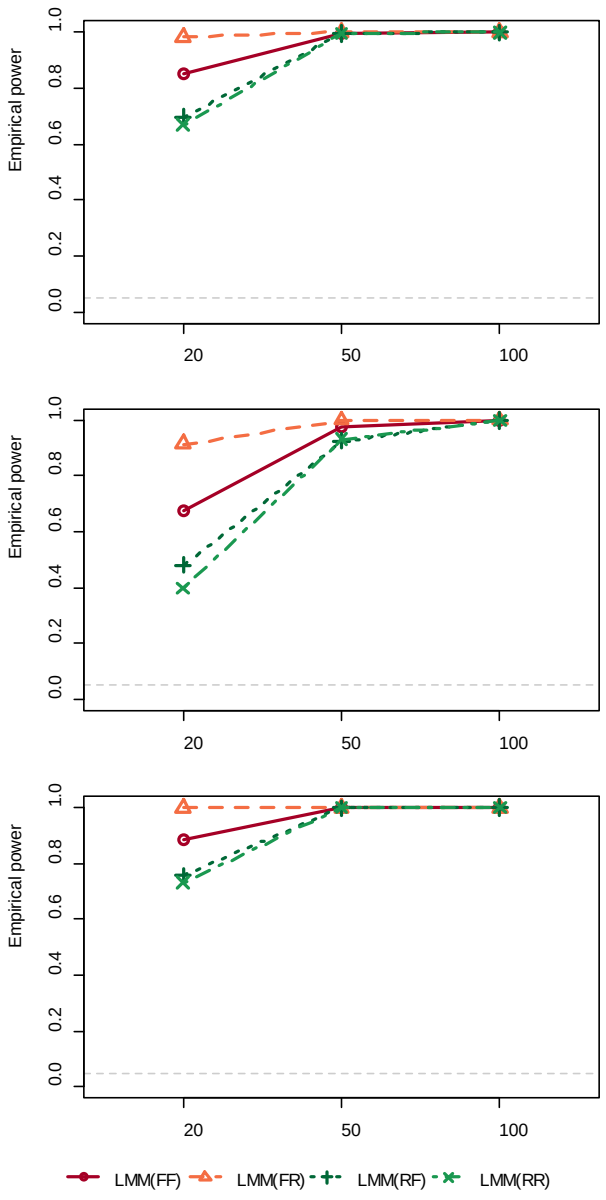

Figure 4. Estimated empirical power of LMMs when the effects are fixed and the interaction model is IM1

Top panel (Scenario 6): $\boldsymbol{G_2 - G_1 = 1}$ and $\boldsymbol{\mathrm{Var}\{(G \times F)_{ik}\} = 0}$. Middle panel (Scenario 7): $\boldsymbol{G_2 - G_1 = 0}$ and $\boldsymbol{\mathrm{Var}\{(G \times F)_{ik}\} = 0.2}$. Bottom panel (Scenario 8): $\boldsymbol{G_2 - G_1 = 0.5}$ and $\boldsymbol{\mathrm{Var}\{(G \times F)_{ik}\} = 0.1}$. The number of peptides was 2. The x-axis and y-axis represent the number of samples and estimated empirical power, respectively. Grey dotted horizontal line represents the significant level.

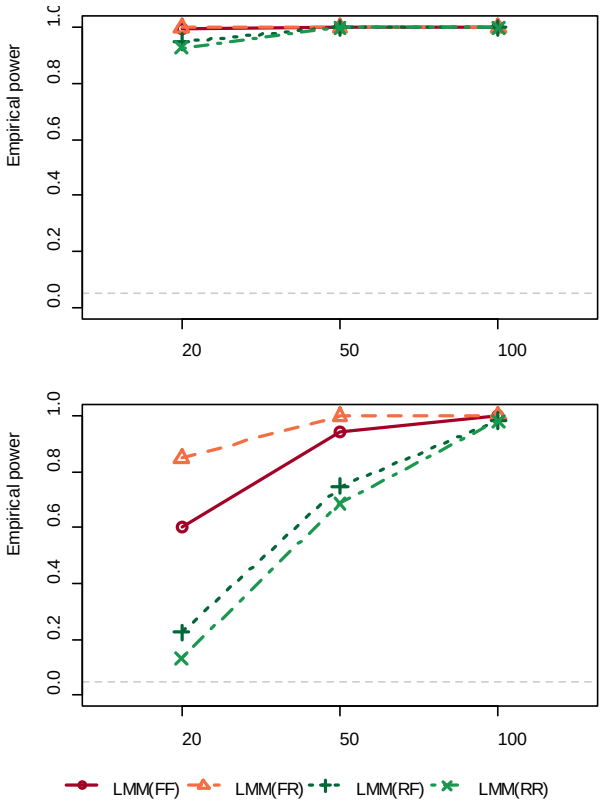

Figure 5. Estimated empirical power of LMMs when effects are fixed and the interaction models are IM1 and IM2
The x-axis represents the sample size and the y-axis represents estimated empirical power. The top panel: The number of peptides is four with the interaction model IM1. The bottom panel: The total number of peptides is four with the interaction model IM2.

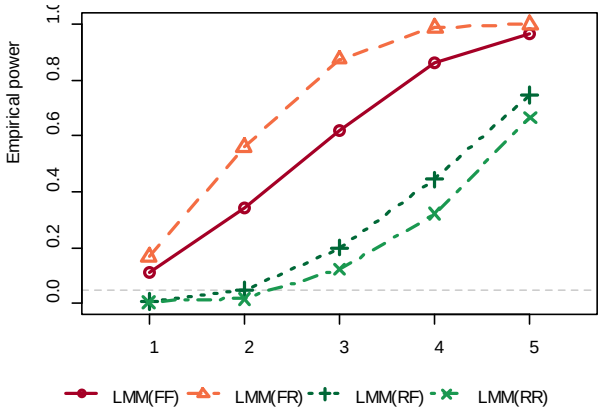

Figure 6. Estimated empirical power of LMMs when effects are fixed and the interaction model is IM3
The x-axis represents the number of positive $(G \times F)_{2,k} - (G \times F)_{1,k}$ when the number of peptides was five. The y-axis respresents estimated empirical power when the sample size was 20. Here, $G_2 - G_1 = 0$ and the squared average of $(G \times F)_{ik}$ was set to 0.2.

4 Conclusion

LMMs have been widely used to identify significant protein from MRM assay. However, LMM approach provides various significance results for the same MRM data depending on which effects are treated as random or fixed. It is well known properties of LMMs that the variance of model parameters are underestimated when the fixed effect model is fitted when the true effects are random and vice versa [14]. As a result, the significance result of LMMs may vary depending on whether the true effect is random or fixed. Thus, it is important to specify correctly the effect as random or fixed.

We examined the performance of LMMs through extensive simulation studies. We utilized AIC for the model selection. However, our empirical study showed that LMM(FF) has the smallest AIC for all simulation settings, which made it difficult to use AIC as a model selection criterion.

Based on our simulation results, we suggest the following practical guideline to use LMMs for MRM data analysis. First, if there is a strong evidence that effects are fixed, then use LMM(FR), because it controlled type I error well and provided the highest power among four LMMs. Second, if there is no evidence that effects are fixed, then use LMM(RR), because it controlled type I error and showed higher power than LMM(RF). Third, when some of peptides in a protein behaved oppositely from the others, we found out that LMMs did not perform well. Thus, the nonsignificant results should be more carefully examined.

Acknowledgement:
This work was supported by the Industrial Strategic Technology Development Program (#10045352), funded by the Ministry of Knowledge Economy (MKE, Korea).

References:
[1] FRANTZI, Maria; BHAT, Akshay; LATOSINSKA, Agnieszka. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med*, 2014, 3.1: 7.
[2] WHITEAKER, Jeffrey R.; PAULOVICH, Amanda G. Peptide immunoaffinity enrichment coupled with mass spectrometry for peptide and protein quantification. *Clinics in laboratory medicine*, 2011, 31.3: 385-396.
[3] PAN, Sheng, et al. Mass spectrometry based targeted protein quantification: methods and applications. *Journal of proteome research*, 2008, 8.2: 787-797.
[4] SHI, Tujin, et al. Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. *Proteomics*, 2012, 12.8: 1074-1092.
[5] LIN, De, et al. Comparison of protein immunoprecipitation-multiple reaction monitoring with ELISA for assay of biomarker candidates in plasma. *Journal of proteome research*, 2013, 12.12: 5996-6003.
[6] MESRI, Mehdi. Advances in Proteomic Technologies and Its Contribution to the Field of Cancer. *Advances in Medicine*, 2014, 2014.
[7] HUANG, Susan M., et al. An endogenous capsaicin-like substance with high potency at recombinant and native vanilloid VR1 receptors. *Proceedings of the National Academy of Sciences*, 2002, 99.12: 8400-8405.
[8] ZHANG, Haixia, et al. Methods for peptide and protein quantitation by liquid chromatography-multiple reaction monitoring mass spectrometry. *Molecular & Cellular Proteomics*, 2011, 10.6: M110. 006593.
[9] ZHANG, Pingbo, et al. Multiple reaction monitoring to identify site-specific troponin I phosphorylated residues in the failing human heart. *Circulation*, 2012, CIRCULATIONAHA. 112.096388.
[10] CHANG, Ching-Yun, et al. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Molecular & Cellular Proteomics*, 2012, 11.4: M111. 014662.
[11] PURSIHEIMO, Anna, et al. Optimization of statistical methods impact on quantitative proteomics data. *Journal of proteome research*, 2015, 14.10: 4118-4126.
[12] MCDONALD, John H. *Handbook of biological statistics*. Baltimore, MD: Sparky House Publishing, 2009.
[13] HEDGES, Larry V.; OLKIN, Ingram. *Statistical method for meta-analysis*. Academic press, 2014.
[14] CHUNG, Yeojin; RABE-HESKETH, Sophia; CHOI, In-Hee. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in medicine*, 2013, 32.23: 4071-4089.