Mersin Üniversitesi Eğitim Fakültesi Dergisi - Mersin University Journal of the Faculty of Education

"Mersin Üniversitesi Eğitim Fakültesi Dergisi, 2018; 14(3): 1020-1037"
Research Paper / DOI: 10.17860/mersinefd.443584

Bounded-Influence Regression Estimation for Mixture Experiments

Karma Denemelerde Sınırlı Etkili Regresyon Tahmin Edicileri

Orkun COŞKUNTUNCEL*

Abstract: Ordinary Least Squares (OLS) estimator is widely used technique for estimating the regression coefficient in mixture experiments. But this estimator is very sensitive to outliers and/or multicollinearity problems. The aim of this paper is to propose estimators for the regression parameters of a mixture model that can combat with the above problems. For this purpose, Generalized M (GM) estimation, which is more resistant to outliers in the y and / or x directions and regression estimators such as ridge and Liu, which is effective against the multicollinearity, were used together. The Mean Square Error (MSE) properties of proposed estimator has been examined and shown to be smaller than biased and GM estimates. Also performance of the combined estimator is illustrated by examples.
Keywords: Bounded-influence regression estimation, generalized M estimation, ridge regression estimator, liu estimator

Öz: Karma denemelerde regresyon katsayılarını tahmin etmek amacıyla en sık kullanılan tahmin edici En küçük kareler tahmin edicisidir. Fakat bu tahmin edici çoklu bağlantı ve/veya sapan değer problemlerine karşı çok hassastır. Bu çalışmanın amacı karma modellerin regresyon parametrelerinin tahminlerini bahsi geçen problemlere karşı daha dirençli olacak biçimde tahmin edebilecek bir tahmin edici önermektir. Bunun için y ve/veya x yönündeki aykırı değerlere kaşı daha dirençli olan Genelleştirilmiş M (GM) tahmini ile çoklu bağlantı problemine karşı etkili olan ridge ve liu gibi yanlı regresyon tahmin edicileri birlikte kullanılmıştır. Önerilen tahmin edicinin hata kareler ortalaması (MSE) incelenerek bunun yanlı ve GM tahminlerinden daha küçük olduğu gösterilmiş ve performansı örneklerle gösterilmiştir.
Anahtar Kelimeler: Sınırlı etkili regresyon tahmini, genelleştirilmiş M tahmin edicisi, ridge regresyon tahmin edicisi, liu tahmin edicisi

Introduction

Experiment with mixtures or mixture problem is a special type of experimental designs in which the response depends only on the relative proportions of the design factors or component. If x_i denote the proportion of the i^{th} component for $i = 1, 2, \ldots, q$ then a mixture problem with q components must have the following constraints

$$x_i \geq 0, \quad \sum_{i=1}^{q} x_i = 1 \ . \tag{1}$$

These are called natural constraints of experiments with mixtures. In many mixture problems there are additional component constraints of the form $L_i \leq x_i \leq U_i$ for some or all of the components (Cornell, 1990; Myers & Montgomery, 2002). These additional constraints on the components often cause multicollinearity or referred as multicollinearity problem in mixture data (John, 1984). In addition to multicollinearity, in many mixture data there may be a single design point or a small subset of points that exerts disproportionate and influence the classical estimation procedure.

* Dr. Öğretim Üyesi, Mersin Üniversitesi, Eğitim Fakültesi, Mersin-Türkiye, ORCID: 0000-0002-0599-1881, e-posta: orkunct@mersin.edu.tr

Gönderi Tarihi: 13.07.2018 - Kabul Tarihi: 15.11.2018

The mixture constraints given by (1) produce a simplex experimental region. Scheffé (1958, 1963) showed that the standard response surface polynomial

$$\eta = \beta_0 + \sum_{i=1}^{q}\beta_i x_i + \sum\sum_{i\le j}^{q}\beta_{ij}x_i x_j + \sum\sum\sum_{i\le j\le k}^{q}\beta_{ijk}x_i x_j x_k + \ldots \quad (2)$$

must be modified with respect to these constraints. Since, this polynomial has meaning for us only subject to the restriction $x_1 + \ldots + x_q = 1$ and substitute

$$x_q = 1 - \sum_{i=1}^{1-q} x_i \quad \text{and} \quad x_i^2 = x_i\left(1 - \sum_{\substack{j=1\\ j\neq i}}^{q} x_j\right)$$

in (2) and obtain the linear and second-order mixture models

$$E(Y) = \sum_{i=1}^{q}\beta_i x_i \quad (3)$$

$$E(Y) = \sum_{i=1}^{q}\beta_i x_i + \sum\sum_{i\le j}^{q}\beta_{ij}x_i x_j\,. \quad (4)$$

In this paper we will use the second-order mixture model given in (4). Note that in product optimization if number of component is small second-order mixture model is most preferable. The polynomial mixture model can be represented in matrix form as following:

$$y = X\beta + \varepsilon, \quad (5)$$

where y is a $n \times 1$ vector of observation on the response variable, X is a $n \times p$ matrix with i^{th} row containing the values of the p predictor variables (linear terms for model (3) in which case $p = q$ and linear and cross product terms for model (4)), β is a $p \times 1$ vector of unknown parameters and ε is an $n \times 1$ column vector of random errors. It is seen that we do not have the intercepts (Cornell, 1990).

Multicollinearity and influential observation

Multicollinearity arises when there are near linear dependencies in the column of X, and it destroys the precision of estimation for regression coefficient and can result in coefficient estimates that are far from the true values (Montgomery, Peck & Vining, 2001). Such linear dependence particularly arises in mixture models when the constraints on the components are active (John, 1984). Various techniques for detecting and identifying multicollinearity have been purposed. In this paper, we will use the condition number of X matrix, which is the ratio of maximum and minimum eigenvalues of $(X'X)$, and variance inflation factors (VIF), which are the diagonal element of $(X'X)^{-1}$. Marquardt (1970) indicates that if any of the VIF's is greater than 10, the corresponding OLS coefficient estimate should be considered as a poor estimate of the corresponding parameter. On the other hand, Gorman (1970) suggests that the ill-conditioning problem should be considered if the VIFs are greater than 100. Also, Belsley, Kuh and Welsch (1980) show that if the condition number is greater than 25, then there is a serious multicollinearity problem which needs to be investigated.

A point or points may be influential because of its location in *x*-space, its observed *y*-values, or both (Montgomery et al., 2001). For mixture data Montgomery and Voth (1994) used diagonal elements h_{ii} of the hat matrix $H = X(X'X)^{-1}X'$ as a measure of leverage and used this measure as a measure of location of the point in *x*-space and suggest that if the experimenter has some flexibility with respect to number of runs, then leverage and multicollinearity can be taken into account in design selection. However, in general, experimenters do not have flexibility with number of runs because of physical, chemical or economical reasons. Montgomery and Voth (1994) also suggest using a robust regression method instead of the OLS estimate to over come with the outliers in *x* direction. To measure the outlyingness of an observation one can use the Mahalanobis distance based on robust location and covariance estimators. For example, Rousseeuw and Zomeren (1990) suggest that high breakdown point estimators to obtain Mahalanobis distances (for example, the minimum volume ellipsoid (MVE) was introduced by Rousseeuw (1985)).

Coşkuntuncel (2005) consider the estimation problem of the regression coefficients in mixture experiments in the case of the combined problem of outliers in y direction and multicollinearity. In this study, it will be investigated the problem of multicollinearity and outliers in x and/or y-directions and propose some hybrid estimators to combat with these problems simultaneously. It will be mainly combine the generalized M (GM) estimators, which are effective to the outlier in *x*- and *y*-direction, with the biased estimators ridge and Liu, which are effective to the multicollinearity, to obtain estimators that can simultaneously deal with multicollinearity and outlier problems. The newly proposed hybrid estimators compared with the ridge, Liu and GM according to their MSE properties and showed that the hybrid estimators have smaller MSE. In the examples, we observed that these estimators significantly reduce the MSE.

Estimators

Ridge and Liu Estimators

The classical estimator for β (the OLS estimator) is $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. This estimator is very sensitive to multicollinearity and outliers. To deal with the multicollinearity problem, the ridge regression estimator was proposed by Hoerl and Kennard (1970a, b). The ridge regression estimator $\hat{\beta}_{R}(k)$ of β is defined as

$$\hat{\beta}_{R}(k) = (X'X + kI)^{-1}X'X\,\hat{\beta}_{OLS}, \tag{6}$$

where $k > 0$ is called the biasing or shrinkage parameter and $I = diag(1, \ldots, 1)$. There are various procedures for choosing the shrinkage parameter k and these methods depend on data. One estimator for k is $k = p\,\hat{\sigma}^2 / \hat{\beta}'_{OLS}\,\hat{\beta}_{OLS}$, where $\hat{\sigma}^2$ is the OLS estimates of σ^2.

Another technique which combats with multicollinearity is Liu estimator defined by Liu (1993). The Liu estimator is defined as

$$\hat{\beta}_{L}(d) = (X'X+I)^{-1}(X'X + dI)\,\hat{\beta}_{OLS}, \tag{7}$$

where $0 < d < 1$ is a parameter. The advantage of $\hat{\beta}_{L}(d)$ over $\hat{\beta}_{R}(k)$ is that $\hat{\beta}_{L}(d)$ is a linear function of d. Therefore it is easier to choose d in Liu estimator than k in ridge estimator.

Robust Ridge and Robust Liu Estimators Based On M Estimator

Since OLS estimator is sensitive to outliers, Ridge and Liu estimators based on the OLS will also be sensitive to outliers in *x*- and/or *y*- directions. To overcome this problem, Silvapulle (1991) proposed ridge type M estimator by combining M and ridge estimators, and showed that this estimate is better in term of MSE criteria. Ridge type M estimator (RM estimator) is defined as

$$\hat{\beta}_{RM}(k^*) = (X'X+k^*I)^{-1}X'X\,\hat{\beta}_M, \tag{8}$$

where k^* is the robust choice of shrinkage parameter and $\hat{\beta}_M$ is the M estimator of the regression coefficient that is used instead of OLS estimates. One of the robust choices of k is

$$k^* = \frac{p\hat{A}^2}{\hat{\beta}_M'\hat{\beta}_M}, \tag{9}$$

where $\hat{A}^2 = \hat{\sigma}^2\{(n-p)^{-1}\Sigma\psi^2(e_i/\hat{\sigma}^2)\}/\{n^{-1}\Sigma\psi'(e_i/\hat{\sigma}^2)\}^2$. Silvapulle (1991) also proved that $MSE\{\hat{\beta}_{RM}(k^*)\} < MSE\{\hat{\beta}_M\}$ and $MSE\{\hat{\beta}_{RM}(k^*)\} < MSE\{\hat{\beta}_R(k)\}$.

Arslan and Billor (2000) proposed Liu type M estimator to combat with combined problem of outlier in y-direction and multicollinearity. Their alternative class of Liu type M estimator (LM estimator) is defined as

$$\hat{\beta}_{LM}(d^*) = (X'X+I)^{-1}(X'X+d^*I)\;\hat{\beta}_M, \tag{10}$$

where d^* is the robust choice of the parameter d. There are many different ways to choose a robust estimate for d. In their paper they proposed the following robust estimate of d.

$$d^* = 1 - \hat{A}^2\left[\sum\frac{1}{\lambda_i(\lambda_i+1)}\Big/\sum\frac{\hat{\beta}_M'\hat{\beta}_M}{(\lambda_i+1)^2}\right], \tag{11}$$

where $\lambda_1 \geq \ldots \geq \lambda_p$ are the eigenvalues of $X'X$ and $\hat{A}^2$ is given above. Arslan and Billor (2000) also proved that $MSE\{\hat{\beta}_{LM}(d^*)\} < MSE\{\hat{\beta}_M\}$ and $MSE\{\hat{\beta}_{LM}(d^*)\} < MSE\{\hat{\beta}_L(d)\}$.

Combined estimators given above are concerned with the multicollinearity and outliers in *y*-direction. However, Huber type M estimator is not robust against outliers in *x*-direction. So, the combined estimators, Ridge type M and Liu type M estimators, will not be robust against outliers in *x* direction. To overcome this drawback of M estimators, Krasker and Welsch (1982) suggested Generalized M (GM) estimators which are effective to outliers in *y*- and *x*-directions. They introduced extra weight function depending only on *x* to down weight the outliers in *x*-direction. By doing this, they define bounded influence robust regression estimators for the regression parameter. We will briefly summaries their estimator in the following section.

Note that Huber type M estimator is not scale invariant. To make it a scale invariant estimator we have to introduce a scale parameter σ in the minimization problem and then find an estimate for it. Here we will estimate the scale parameter σ using the median absolute deviation (MAD) prior to the regression estimation (Maronna, Martin & Yohai, 2006).

$$\hat{\sigma} = \frac{1}{0.6745}.median|e_i - median(e_i)| = \frac{1}{0.6745}.mad(e_i) = 1.483.mad(e_i). \qquad (12)$$

Generalized M (GM) Estimators

In general the GM-estimator for the regression parameter can be obtained as the solution of the following minimization problem:

$$Q(\beta) = \hat{\sigma}^2 \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i'\beta}{\hat{\sigma} u_i^{\theta}}\right) u_i^{1+\theta},$$

(13)

where ρ is an even function that is non decreasing on $[0, \infty)$, u_i is weight determined by $x_1, \ldots, x_n$ and $\hat{\sigma}$ is the scale estimate given above.

In this form if we choose $\rho(x) = x^2$ and $u_i \equiv 1$ we get OLS estimator. Similarly if we choose $\rho(x) = x^2$ and $\theta = -1$ we get weighted OLS. The L_1 estimator uses $\rho(x) = |x|$ and $u_i \equiv 1$. Robust M estimator (or Huber M estimator) is obtained by setting $u_i \equiv 1$ (Simpson & Chang, 1997). Here, choosing θ is very important and there are two important values for θ which give two different bounded influence regression estimators widely known in literature.

Setting $\theta = 0$ gives Mallows type GM estimator (Krasker & Welsh 1982; Simpson, Ruppert and Carroll 1992). In this form of the minimization problem an extra weight function is introduced to downweight the outliers in x-direction. Widely used weight function is

$$u_i = \min\left[1, \left\{\frac{b}{(x_i - \bar{x}_c)'C^{-1}(x_i - \bar{x}_c)}\right\}^{1/2}\right], \qquad (14)$$

(Hampel, Ronchetti, Rousseeuw & Stahel, 1986) where $\bar{x}_c$ is a robust location estimator and C is a robust scatter estimator of X. Here b is equal to $(1 - \upsilon)$ quantile of the chi-squared distribution on $p - 1$ degrees of freedom and υ can be taken as 0.1 or 0.05. The Mallows type GM estimator put separate bounds on the residuals and the influence in x-direction. This method downweights the outlier observation regardless of the magnitude of the corresponding residual. However, any downweighting of x that does not consider how the response fits the remainder of the data may not produce efficient estimators.

To cope with this drawback of Mallows type GM estimator the Schweppe type GM estimator has been proposed (Hampel et. al., 1986). The Schweppe type GM estimator can be obtained when we take $\theta = 1$. In this case the weight function can be taken as:

$$u_i = \frac{1}{\sqrt{(x_i - \bar{x}_c)'C^{-1}(x_i - \bar{x}_c)}}.$$

(15)

This estimate downweights leverage points only if corresponding residual is large (Hampel et. al. 1986).

If ρ is a differentiable function we get the following estimating equation after setting derivative of (13) with respect to β to zero

$$\sum_{i=1}^{n} v\left(\frac{y_i - x_i'\beta}{\sigma u_i^{\theta}}\right) x_i' u_i^{1-\theta}(y_i - x_i'\beta) = 0\,, \tag{16}$$

where $v(t)$ is $\psi(t)/t$ for $t \neq 0$ and $\psi'(0)$ for $t = 0$. After some further rearranging this equation we get

$$\hat{\beta}_{GM} = \left(\sum_{i=1}^{n} \tilde{u}_i x_i x_i'\right)^{-1} \sum_{i=1}^{n} \tilde{u}_i x_i y_i\,, \tag{17}$$

where $\tilde{u}_i = v\left(\frac{y_i - x_i'\hat{\beta}}{\hat{\sigma} u_i^{\theta}}\right) u_i^{1-\theta}$. Or in matrix notation we can write

$$\hat{\beta}_{GM} = (X'\tilde{U}X)^{-1} X'\tilde{U}y, \tag{18}$$

where $y = (y_1, \ldots, y_n)'$, $X = [x_1, \ldots, x_n]'$ and $\tilde{U} = diag(\tilde{u}_1, \ldots, \tilde{u}_n)$.

The variance-covariance estimate of $\hat{\beta}_{GM}$ is of the form

$$Cov(\hat{\beta}_{GM}) = \frac{\hat{\sigma}^2}{n} P^{-1} Q P^{-1}, \tag{19}$$

where $P = \frac{1}{n}\sum_{i=1}^{n} \psi'\left(\frac{r_i}{u_i^{\theta}}\right) u_i^{1-\theta} x_i x_i'$, $Q = \frac{1}{n}\sum_{i=1}^{n} \psi^2\left(\frac{r_i}{u_i^{\theta}}\right) u_i^2 x_i x_i'$ and r_i is i^{th} standardized residual. Here is when $\theta = 0$ we get exchangeable estimate of $Cov(\hat{\beta}_{GM})$ (Du & Wiens, 2000; Maronna & Yohai, 1981) and it can be expressed as in matrix notation by

$$Cov(\hat{\beta}_{GM}) = \frac{\hat{\sigma}^2}{n-p} \frac{\sum_{i=1}^{n} \psi^2(r_i)}{\left[\sum_{i=1}^{n} \psi'(r_i)\right]^2} \Omega\,, \tag{20}$$

where $\Omega = (X'UX)^{-1}(X'W^2X)(X'UX)^{-1}$ and $U = diag(u_1, \ldots, u_n)$.

Robust Ridge and Liu Regression Estimations Based On GM Estimators

Robust Ridge Estimator for Regression Based On GM Estimator

Since GM estimators are robust against outliers in *x*- and *y*-directions it will be very useful in combining both ridge (and Liu) and GM estimate to deal with simultaneous presence of outliers in *x*- and *y*- direction and multicollinearity. In this sense Arslan and Billor (1996) proposed ridge type GM estimator (RGM estimator) which robust against the outlying observation in *x*-direction and outliers with large residuals.

They proposed two different forms of ridge type GM estimator. The first form is

$$\hat{\beta}_{RGM}(k^+) = (X'X + k^+I)^{-1}X'X\hat{\beta}_{GM}, \quad (21)$$

where $\hat{\beta}_{GM}$ is the GM estimator of β and k^+ is the one of the robust choice of k. The second form is

$$\hat{\beta}_{RGMU}(k^+) = (X'UX + k^+I)^{-1}X'UX\hat{\beta}_{GM}, \quad (22)$$

where $k^+ = p\hat{\sigma}^2_{GM} / \hat{\beta}'_{GM}\hat{\beta}_{GM}$, $U = diag(u_1, \ldots, u_n)$ and $\hat{\sigma}^2_{GM}$ is the estimate of variance using $\hat{\beta}_{GM}$. Arslan and Billor (1996) also proved that $MSE\{\hat{\beta}_{RGM}(k^+)\} < MSE\{\hat{\beta}_{GM}\}$, $MSE\{\hat{\beta}_{RGMU}(k^+)\} < MSE\{\hat{\beta}_{GM}\}$ and $MSE\{\hat{\beta}_{RGM}(k^+)\} < MSE\{\hat{\beta}_{R}(k)\}$.

Robust Liu Estimator for Regression Based On GM Estimator

In this study we combine the GM and the Liu estimators to combat with simultaneously occurrence of multicollinearity and outliers in the dataset. We expect that the resulting estimator is robust against the outliers and influence observations as well as deals with the multicollinearity. Similarly two different forms of the Liu type GM estimator can be defined.

In ordinary Liu estimator form we can replace the OLS estimator $\hat{\beta}_{OLS}$ with the $\hat{\beta}_{GM}$ which gives

$$\hat{\beta}_{LGM}(d^+) = (X'X + I)^{-1}(X'X + d^+I)\ \hat{\beta}_{GM}, \quad (23)$$

where d^+ is the robust choice of d.

We can also combine $\hat{\beta}_{GM}$ and the Liu estimation method as follows

$$\hat{\beta}_{LGMU}(d^+) = (X'UX + I)^{-1}(X'UX + d^+I)\ \hat{\beta}_{GM}. \quad (24)$$

In this form we want to take care the outliers in design matrix and then downweight them. Therefore, instead of using $X'X$ we want to use $X'UX$. By doing this we may get rid of some points in design which affects the whole procedure. However downweighting this type of observation may also affect the multicollinearity problem. We may downweight some points which may induce multicollinearity or mask multicollinearity.

Mean Squared Error Properties

In this subsection we will compare the MSEs of the proposed estimators. To do so let $\lambda_1 \geq \ldots \geq \lambda_p$ be the eigenvalues of $X'X$, $q_1, \ldots, q_p$ be the corresponding eigenvectors, and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ and $P = (q_1, \ldots, q_p)$ such that $X'X = P\Lambda P'$. Then, the regression model given in (5) can be rewritten in canonical form as following:

$$y = C\alpha + \varepsilon, \quad (25)$$

where $C = XP$, $\alpha = P'\beta$. It is seen that we do not have the constant term in our models. The OLS and the corresponding Liu estimators are

$$\hat{\alpha}_{OLS} = \Lambda^{-1} C' Y$$

and

$$\hat{\alpha}_{L}(d) = (\Lambda + I)^{-1}(\Lambda + dI)\ \hat{\alpha}_{OLS}. \tag{26}$$

Note that since $\hat{\beta} = P\hat{\alpha}$ we have $MSE(\hat{\alpha}_{OLS}) = MSE(\hat{\beta})$. Thus, it is sufficient to consider only the canonical form compare the MSEs.

We can also find the GM estimator of α. If $\hat{\alpha}_{GM}$ is the corresponding GM estimators of α, then the corresponding $\hat{\alpha}_{LGM}$ and $\hat{\alpha}_{LGMU}$ estimators become

$$\hat{\alpha}_{LGM}(d) = (\Lambda + I)^{-1}(\Lambda + dI)\ \hat{\alpha}_{GM} \text{ and } \hat{\alpha}_{LGMU}(d) = (B + I)^{-1}(B + dI)\ \hat{\alpha}_{GM}, \tag{27}$$

where $\mu_1 \geq \ldots \geq \mu_p$ are the eigenvalues of $X'UX$, $B = diag(\mu_1, \ldots, \mu_p)$ and U is given above.

Now we can give the following theorem on the MSE properties of the estimators $\hat{\alpha}_{LGM}(d)$ and $\hat{\alpha}_{LGMU}(d)$.

Theorem: Assume that $Cov(\hat{\alpha}_{GM}) = \Gamma$ is finite. Then

i) There always exists a d such that $MSE(\hat{\alpha}_{LGM}(d)) < MSE(\hat{\alpha}_{GM})$

ii) There always exists a d such that $MSE(\hat{\alpha}_{LGMU}(d)) < MSE(\hat{\alpha}_{GM})$

iii) Further if $\Gamma_{ii} < \dfrac{\sigma^2}{\lambda_i}$ is for every i, then $MSE(\hat{\alpha}_{LGM}(d)) < MSE(\hat{\alpha}_{L}(d))$ for every positive d, where Γ_{ii} is diagonal element of Γ.

Proof: The MSE expressions of $\hat{\alpha}_{GM}$, $\hat{\alpha}_{L}(d)$, $\hat{\alpha}_{LGM}(d)$ and $\hat{\alpha}_{LGMU}(d)$ are

$$MSE(\hat{\alpha}_{GM}) = \sum_{i=1}^{p} \Gamma_{ii} \tag{28}$$

$$MSE(\hat{\alpha}_{L}(d)) = \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{\lambda_i(\lambda_i + 1)^2} + (d-1)^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + 1)^2} \tag{29}$$

$$g(d) = MSE(\hat{\alpha}_{LGM}(d)) = \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{(\lambda_i + 1)^2} \Gamma_{ii} + (d-1)^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + 1)^2} \tag{30}$$

$$r(d) = MSE(\hat{\alpha}_{LGMU}(d)) = \sum_{i=1}^{p} \frac{(\mu_i + d)^2}{(\mu_i + 1)^2} \Gamma_{ii} + (d-1)^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\mu_i + 1)^2}. \tag{31}$$

i) The proof of part one of the theorem is similar to the proof of the Theorem 2.1 of Liu (1993). If we take the derivative of g(d) with respect to d we get

$$g'(d) = \sum_{i=1}^{p} \frac{2(\lambda_i + d)}{(\lambda_i + 1)^2} \Gamma_{ii} + 2(d-1) \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + 1)^2}$$

Note that $g(1) = MSE(\hat{\alpha}_{GM})$. Because of $g(1) > 0$ there exists $0 < d < 1$, which means that $g(d) < g(1)$ or equivalently $MSE(\hat{\alpha}_{LGM}(d)) < MSE(\hat{\alpha}_{GM})$.

From the proof of part (i) we see that $MSE(\hat{\alpha}_{LGM}(d))$ is minimized at

$$d = \sum_{i=1}^{p} \frac{\alpha_i^2 - \lambda_i \Gamma_{ii}}{(\lambda_i + 1)^2} \Big/ \sum_{i=1}^{p} \frac{\alpha_i^2 + \Gamma_{ii}}{(\lambda_i + 1)^2}. \qquad (32)$$

ii) The proof of part (ii) is similar to the Theorem 4.3 of Hoerl and Kennard (1970a). If we take the derivative of $r(d)$ we observe that the first of part of $r(d)$ is an increasing function of d and the second part is a decreasing function of d. Therefore we only need to prove that there always exists a $d > 0$ such that $r'(d) < 0$. The derivative of $r(d)$ is

$$r'(d) = \sum_{i=1}^{p} \frac{2(\mu_i + d)}{(\mu_i + 1)^2} \Gamma_{ii} + 2(d-1) \sum_{i=1}^{p} \frac{\alpha_i^2}{(\mu_i + 1)^2}.$$

From this one can easily see that $2\Gamma_{ii}\mu_i + 2\Gamma_{ii}d < -2\alpha_i^2 d - 2\alpha_i^2$. Therefore if $d < \min\left(\frac{\alpha_i^2 - \Gamma_{ii}\mu_i}{\Gamma_{ii} + \alpha_i^2} \right)$, $i = 1, \ldots p$ then we get $r'(d) < 0$.

iii) For the proof of part (iii) we denote the difference $MSE(\hat{\alpha}_{LGM}(d)) - MSE(\hat{\alpha}_{L}(d))$ by L. Some straightforward calculations result in

$$L = \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{(\lambda_i + 1)^2} \Gamma_{ii} - \sigma^2 \sum_{i=1}^{p} \frac{(\lambda_i + d)^2}{\lambda_i (\lambda_i + 1)^2} = \sum_{i=1}^{p} \frac{(\lambda_i + d)^2 [\lambda_i \Gamma_{ii} - \sigma^2]}{(\lambda_i + 1)^2 \lambda_i}.$$

In order to make L less than zero we should have $\lambda_{ii}\Gamma_{ii} - \sigma^2 < 0$. Thus we have $\Gamma_{ii} < \sigma^2/\lambda_i$ to get $L < 0$. This completes the proof. □

Next we will propose a robust estimator for d by substituting α_i^2 and Γ_{ii} by their unbiased estimates given in (32). Asymptotically it is known that $\hat{\alpha}_{GM}$ is normally distributed with covariance matrix $A^2\Omega$ for the Mallows case, where $A^2 = \sigma^2 E[\psi^2(\varepsilon/\sigma)]/[E\psi'(\varepsilon/\sigma)]^2$. Thus instead of Γ_{ii} we can take its estimate $\hat{A}^2 / \omega_i$ where ω_i are the eigenvalues of Ω given in (20), and

$$\hat{A}^2 = \hat{\sigma}^2 (n-p)^{-1} \sum_{i=1}^{n} \psi^2(r_i) \Big/ \left[\sum_{i=1}^{n} \psi'(r_i) \right]^2 \qquad (33)$$

(Du & Wiens, 2000). Also the unbiased estimator of α_i^2 is $\hat{\alpha}_{GMi}^2 - \hat{A}^2 / \omega_i$. Thus for the Mallows GM estimator we can propose a robust estimator of d as follows

$$\hat{d}_{GMi} = 1 - \hat{A}^2 \left[\sum_{i=1}^{p} \frac{1}{\omega_i (\lambda_i + 1)} \Big/ \sum_{i=1}^{p} \frac{\hat{\alpha}_{GMi}^2}{(\lambda_i + 1)^2} \right]. \qquad (34)$$

Examples

To illustrate the performance of the estimators that we give above we consider the four component lubricant blending data given by Snee (1975). The objective of the study was to determine the amount of additive (x_1) needed in three lubricant blends (x_2, x_3, x_4) so that a certain critical physical property would attain a desired level (John, 1984). The components

have following additional constraints instead of natural constraints of experiments with mixtures given by (1).

$$0.07 \le x_1 \le 0.18 \; 0 \le x_2 \le 0.30$$
$$0.37 \le x_3 \le 0.70 \; 0 \le x_4 \le 0.15$$

To emphasize the aim of this study we would like to create an outlier in y direction by changing the 5th response value from 12.93 to 3.417. The new value for 5th response value found with simulation editor of software packages Design Expert 6 (Stat-Ease, 2004). Preliminary examination of the data set for the second-order mixture model given in (4) shows that the smallest and the largest eigenvalues are 0.00008 and 7.57, respectively and the condition number is 93406.91 and also the smallest and the largest VIFs are 7039.66 and 65.74, respectively. These values point out that there is a serious ill-conditioning problem. We get following estimating equation for ordinary least square estimate of the standardized data.

$$\hat{y}_{OLS} = 13.55x_1 + 19.45x_2 + 0.58x_3 - 94.83x_4 + 3.66x_1x_2 - 6.22x_1x_3 + 35.54x_1x_4 + 0.02x_2x_3 + 4.27x_2x_4 + 82.86x_3x_4 \; (\hat{\sigma}^2 = 2.44)$$

For the OLS estimator we have the $MSE(\hat{\beta}_{OLS}) = 37468.21$. The following table shows the some diagnostic measure for the lubricant blending data. In the table, r_i are the standardized residuals, h_{ii} are diagonal elements of hat matrix, MD_i and RD_i are the Mahalanobis distances base on the sample variance-covariance estimators and the robust MVE estimators, respectively.

Table 1.
Some diagnostics for lubricant blending data.

Obs	r_i	h_{ii}	MD_i	RD_i
1	-0.481	0.442	0.308	8.039
2	-0.169	0.927	1.147	13.298
3	0.8759	0.528	0.528	8.768
4	-0.549	0.813	0.123	12.114
5	-2.828	0.784	0.301	208.379
6	-1.028	0.525	0.466	9.3687
7	2.339	0.869	0.340	416.093
8	-0.210	0.429	0.345	7.714
9	1.197	0.827	0.303	12.819
10	0.939	0.453	0.264	9.631
11	-0.687	0.411	0.065	8.971
12	0.847	0.427	0.027	9.046
13	-1.617	0.477	0.148	9.325
14	0.847	0.388	0.072	6.922
15	1.382	0.401	0.321	9.250
16	-0.345	0.466	0.147	8.608
17	-0.032	0.275	0.001	7.529
18	-0.319	0.558	0.382	8.597

From Table 1 following conclusion can be drawn: For 5th observation, standardized residual is larger than others as we expected. Because, this observation is adjusted as an outlier in *y*-direction. Also 7th observation has large standardized residual too. The cut off value for h_{ii} is $2p/n = 1.11$. This is bigger than 1, so we can use average of h_{ii} values 0.555 as the cut off value. 2nd, 4th, 5th, 7th and 9th observations exceed the cut off value. 2nd observation has larger classical mahalanobis distance than the others as we except, because there exists a monotone relationship between h_{ii} and MD_i of x_i (Rousseeuw & Leroy, 2003; Rousseeuw & Zomeren,

1990). For robust distance, 5th and 7th observations exceed the cut off value $\chi^2_{0.95,p-1} = \chi^2_{0.95,9}$ = 16.918. For Huber type M estimate, we get the following estimating equation:

$$\hat{y}_M = 70.37x_1 + 6.98x_2 + 17.49x_3 + 2.07x_4 - 11.71x_1x_2 - 32.71x_1x_3 - 1.75x_1x_4 - 3.33x_2x_3 - 0.23x_2x_4 - 2.32x_3x_4 \ (\hat{\sigma}^2 = 1.02).$$

Notice the difference between OLS and the M estimate. Examination of the residual we see that the 5th observation has very large residual and also its weight approximately zero and 7th observation has very small residual and its weight is one, which means, 7th observation has no problem in *y*-direction but in OLS it has large error and seems to be outlier in *y*-direction. This is, classical diagnostics methods and OLS estimate is mislead us about outlier analysis and we have to be careful when we use these methods. Thus based on diagnostics given above and M estimate results, we can conclude that the 5th observation is an outlier in *x*- and *y*-direction and the 7th observation is an outlier in *x*-direction.

For GM estimate we expected to get smallest weights in *x*-direction for 5th and 7th observations and weight in *y*-direction for 5th observation. Mallows and Schweppe types GM estimators are using the weights for *x*-direction given in (14) and (15), respectively. For Mallows type GM estimator we get following estimating equation:

$$\hat{y}_{GMM} = 62.81x_1 + 8.34x_2 + 14.92x_3 - 12.08x_4 - 9.51x_1x_2 - 29.79x_1x_3 + 3.82x_1x_4 - 2.25x_2x_3 + 0.25x_2x_4 + 14.27x_3x_4.$$

For Mallows type GM estimator, we have $MSE(\hat{\beta}_{GMM})$ = 208.13. 5th and 7th observations have weighs in *x*-direction 0.28 and 0.20, respectively and the others are 1. Also 5th observation has weight in *y*-direction near zero while the others are 1, and 5th observation has largest residual from the others. For Schweppe type GM estimator, we get following estimating equation:

$$\hat{y}_{GMS} = 71.58x_1 + 6.37x_2 + 17.46x_3 + 2.59x_4 - 11.84x_1x_2 - 33.97x_1x_3 - 1.80x_1x_4 - 2.65x_2x_3 - 0.46x_2x_4 + 2.10x_3x_4.$$

Notice the different signs for some coefficients from Mallows type GM estimate. High level multicollinearity may cause these differences. For Schweppe type GM estimator we have $MSE(\hat{\beta}_{GMS})$ = 200.60. 5th and 7th observations have weighs in *x*-direction 0.07 and 0.05, respectively. Also 5th observation has weight in *y*-direction near zero while the others are 1, and 7th observation has largest residual from the others as Mallows type GM estimate. But with these two estimates we get weights and residuals as we expect above.

Table 2 shows the coefficient estimates of ordinary ridge and Liu estimators and robust ridge and Liu estimators based on M estimator. In the table $\hat{\beta}_R$ is ordinary ridge estimate, $\hat{\beta}_L$ is ordinary Liu estimate, $\hat{\beta}_{RM}$ is ridge type M estimate, $\hat{\beta}_{LM}$ is Liu type M estimate.

Table 2.
Results for ridge and Liu estimators. [(1)]k = 0.001376, [(2)]d = 0.185, [(3)]k = 0.001556, [(4)]d = 0.14.

Term	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	VIF_{imax}	MSE
$\hat{\beta}_R$[1]	3.42	12.14	-0.56	-16.65	6.37	3.84	22.57	5.92	-9.76	27.46	95.53	1278.44
$\hat{\beta}_L$[2]	7.54	6.30	4.95	-13.32	3.65	4.05	11.60	2.87	1.65	20.02	241.22	12801.09
$\hat{\beta}_{RM}$[3]	15.26	4.58	11.71	4.06	1.40	11.87	5.90	-1.51	-0.74	0.63	87.40	1142.72
$\hat{\beta}_{LM}$[4]	16.15	3.73	8.25	4.22	1.75	2.16	4.44	2.62	0.64	4.70	138.22	5258.07

From Table 2 we observe that performance of the biased estimators based on M estimator are better than its ordinary forms in terms of MSE and VIF. Especially Ridge type M estimator gives smaller MSE and VIF values than the others. And also from VIF values we may say, for this data ridge type estimators combat with multicollinearity better than Liu type estimators. Table 3 shows the results of ridge type GM estimators. In the table $\hat{\beta}_{\text{RGM-M}}$ and $\hat{\beta}_{\text{RGMU-M}}$ are ridge type Mallows GM estimate and $\hat{\beta}_{\text{RGM-S}}$ and $\hat{\beta}_{\text{RGMU-S}}$ are ridge type Schweppe GM estimate given in (21) and (22), respectively and $k = p\hat{\sigma}^2_{GM} / \hat{\beta}'_{GM}\hat{\beta}_{GM}$. From Table 3 we can observe that ridge type GM estimators are very effective in terms of MSE and VIF.

Table 3.
Results for ridge type GM estimators.

Term	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	k	VIF_{imax}	MSE
$\hat{\beta}_{RGM-M}$	13.11	5.15	9.71	1.67	2.37	10.72	8.50	0.19	-2.34	4.28	0.0018	76.24	188.67
$\hat{\beta}_{RGMU-M}$	13.14	5.26	9.62	1.74	2.29	10.77	8.44	0.17	-2.33	4.27	0.0018	81.17	187.95
$\hat{\beta}_{RGM-S}$	15.28	4.29	11.69	4.25	1.50	11.54	6.08	-1.17	-0.92	0.59	0.0016	88.70	189.35
$\hat{\beta}_{RGMU-S}$	12.84	1.64	11.12	3.21	3.05	12.08	7.63	1.31	-1.60	1.75	0.0016	93.45	189.67

Table 4 shows the results of Liu type GM estimators. In the table $\hat{\beta}_{\text{LGM-M}}$ and $\hat{\beta}_{\text{LGMU-M}}$ are Liu type Mallows GM estimate and $\hat{\beta}_{\text{LGM-S}}$ and $\hat{\beta}_{\text{LGMU-S}}$ are Liu type Schweppe GM estimate given in (23) and (24), respectively.

Table 4.
Results for Liu type GM estimators.

Term	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	d	VIF_{imax}	MSE
$\hat{\beta}_{LGM-M}$	13.27	3.78	7.53	2.82	2.41	3.49	5.38	2.93	0.76	6.19	0.11	85.37	188.84
$\hat{\beta}_{LGMU-M}$	12.08	3.87	7.20	3.35	2.49	3.61	5.51	3.10	1.18	6.30	0.095	64.60	195.69
$\hat{\beta}_{LGM-S}$	14.34	3.54	7.91	4.36	2.20	3.24	4.67	2.90	0.64	4.77	0.1095	84.61	172.78
$\hat{\beta}_{LGMU-S}$	9.03	3.52	5.60	4.10	2.55	2.77	4.11	3.03	2.08	4.24	0.06	69.33	194.49

From Table 4 we can observe that, similar to ridge type GM estimator, Liu type GM estimators are very effective in terms of MSE and VIF.

Conclusions

In experiment with mixture, multicollinearity is a very important problem. Particularly, for models with some additional constraints it can be very serious. In the case of multicollinearity, pseudocomponent transformation may improve the VIFs, but they may not be reduced to an acceptable level. Biased regression estimators, such as ridge or Liu, should be used to reduce the effect of multicollinearity on OLS estimators. Further, there may be some outlying observations in the mixture data, and we may use a regression estimator that is robust against outliers in the data. In this paper we propose to combine biased regression estimation and robust regression estimation to produce a hybrid estimator to simultaneously combat with multicollinearity and outliers problems. We are particularly interested in combining generalized M (GM) regression estimators with the ridge and Liu estimators to produce an estimator to deal with the combined problem of multicollinearity and the outliers in a mixture data. We provide an example to show

that the hybrid estimators can successfully combat with the combined problem of multicollinearity and outliers, and these estimators can be safely used to obtain regression estimates for the mixture data sets that are suspected to have multicollinearity and outliers.

It should be taken into consideration that in order to claim that this estimate will be effective against outliers in the direction of y and/or x, it is necessary to see its effectiveness in other fields (educational sciences, econometrics etc.) and in data sets with more observations.

References

Arslan, O., & Billor, N. (1996). Robust ridge regression estimation based on the gm-estimators. *Jour. of Math. & Comp. Sci. (Math. Ser.)*, *9*(1), 1-9.

Arslan, O., & Billor, N. (2000). Robust liu estimator for regression based on an m-estimators. *Journal of Applied Statistics*, *27*(1), 39-47.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley, New York.

Cornell, J. A. (1990). *Experiments with mixtures - designs, models and the analysis of mixture data*. 2nd edition, John Wiley& Sons, Inc., New York, USA.

Coşkuntuncel, O. (2005). Robust estimators for the regression parameters of experiment with mixtures models. *International Jour. of Pure and App. Math.*, *24*(4), 459-469.

Du, Z., & Wiens, D. P. (2000). Jackknifing, weighting, diagnostics and variance estimation in generalized m-estimation. *Statistics and Probability Letters*, 46, 287-299.

Gorman, J. W. (1970). Fitting equations to mixture data with restraints on compositions. *Journal of Quality Technology*, 2, 186-194.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: the approach based on influential functions*. Wiley, New York.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69-82.

Huber, P. J. (1964). Robust estimation of a location parameters. *The Annals of Mathematical Statistics*, 35, 73-101.

Huber, P.J. (2003). *Robust statistics*. Wiley, New York.

John, R. C. (1984). Experiments with mixtures, ill conditionning, and Ridge regression. *Journal of Quality Technology*, 16, 81-96.

Krasker, W. S., & Welsch, R. E. (1982). Efficent bounded-influence regression estimation. *Journal of the American Statistical Association*, *77*(379), 595-604.

Liu, K. (1993). A new class of biased estimate in linear regression. *Communication in Statistics A*, 22, 105-123.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Wiley, New York.

Maronna, R. A., & Yohai, V. J. (1981). Asymptotic behaviour of general m-estimates for regression and scale with random carriers. *Z. Wahrsch. Verb. Geb.*, 58, 7-20

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis.* Wiley, New York.

Montgomery, D. C., & Voth, S. R. (1994). Multicolinearity and leverage in mixture experiments. *Journal of Quality Technology*, 26, 96-108.

Myers, R. H., & Montgomery, D. C. (2002). *Response surface methodology.* Wiley, New York.

Rousseeuw, P. J. (1985). A regression diagnostic for multiple outliers and leverage points. Abstract in *IMS Bull.*, 14, 399.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. Wiley, New York.

Rousseeuw, P. J., & Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association, 85*(411), 633-651.
Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society* - B, 20, 344-360.
Scheffé, H. (1963). The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society* - B, 25, 235-263.
Silvapulle, M.J. (1991). Robust ridge regression based on m-estimator. *Australian and New Zealand Journal of Statistics*, 33, 319-333.
Simpson, D. G., Ruppert, D., & Carroll, R. J. (1992). On one-step gm estimates and stability of inferences in linear regression. *Journal of American Statistical Association*, 87(418), 439-450.
Simpson, D. G., & Chang, Y. C. I. (1997). Reweighting approximate gm estimators: asymptotics and residual-based graphics. *Journal of Statistical Planning and Inference*, 57, 273-293.
Snee, R. D. (1975). Experimental designs for quadratic models in constrained mixture spaces. *Technometrics*, 17, 149-159.
Stat-Ease (2004). *DESIGN-EXPERT software for response surface methodology and mixture experiments.* Version 6, 45 days Trial. Stat-Ease, Inc. Minneapolis, MN.

Uzun Öz

Giriş

Karma denemeler, y amacının (veya amaçlarının), sistemi oluşturan bileşenlerin oranına bağlı olduğu deneysel tasarımların özel bir halidir. Yani eğer xi, i-inci bileşenin oranı ise q bileşenli karma problem $x_i \geq 0$, $x_1 + \ldots + x_q = 1$ doğal kısıtlamasına sahip olur. Birçok karma problem $L_i \leq x_i \leq U_i$ şeklinde alt ve üst sınırlardan oluşan ek kısıtlamalara sahiptir (Cornell, 1990; Myers ve Montgomery, 2002). Bileşenler üzerindeki bu kısıtlamalar karma problemde genellikle kötü koşulluluk veya iç ilişki probleminin çıkmasına neden olur (John, 1984). Bunların yanı sıra veri analizinde araştırmacıların çok sık karşılaştıkları sapan değer problemi de bulunabilir. Bu durumda en küçük kareler ile elde edeceğimiz sonuçlar güvenilir olmayacaktır.

Scheffé (1953, 1963) standart polinomun karma denemelerin doğal koşullarından dolayı düzenlenmesi gerektiğini ve aşağıdaki kanonik polinom modelin kullanılmasını önermiştir. Bu polinom bizim için $x_1 + \ldots + x_q = 1$ kısıtlaması altında anlamlıdır.

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i \leq j}^{q} \beta_{ij} x_i x_j + \sum\sum \sum_{i \leq j \leq k}^{q} \beta_{ijk} x_i x_j x_k + \ldots$$

Bu polinomda

$$x_q = 1 - \sum_{i=1}^{1-q} x_i \quad \text{and} \quad x_i^2 = x_i \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^{q} x_j \right)$$

dönüşümleri yapılarak aşağıdaki lineer ve ikinci dereceden karma modeller elde edilir.

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i$$

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i \leq j}^{q} \beta_{ij} x_i x_j \, .$$

Kanonik polinom model matris formunda; *y*, n×1 tipinde amaç değişken üzerinde gözlemlerin bir sütun vektörü, ε, n×1 tipinde rastgele hataların bir sütun vektörü, *X*, n×p tipinde

ve i-inci satırı, i-inci gözleme karşılık gelen sütunlarında p tane x_i değişkeninin değerlerini içeren matris ve β, p×1 tipinde tahmin edilecek parametrelerinin bir sütun vektörü olmak üzere,

$$y = X\beta + \varepsilon$$

şeklinde verilir (Cornell, 1990).

Çoklu bağlantı ve etkili gözlem

X katsayı matrisinin kolonları arasında kötü koşulluluktan dolayı yaklaşık lineer bağımlılık oluşabilir. Bu durumda $X'X$ matrisinin tersi alınamaz veya alınsa bile yanlış olabilir. Ayrıca regresyon katsayıları da kötü koşulluluktan dolayı gerçek değerlerinden uzaklaşabilir (Montgomery, Peck ve Vinning, 2001). Veride kötü koşulluluğun derecesini belirlemek için en çok kullanılan yöntemler koşul sayısı ve VIFi'dir. Varyans şişirme faktörleri (VIFi)'ler, $(X'X)^{-1}$ matrisinin köşegen elemanlarıdır. Marquardt (1970), VIF'lerden en az birinin 10'dan büyük olması durumunda tahmin edilen parametrenin güçlü bir tahmin olmayacağını belirtmiştir. Diğer yandan Gorman (1970), 100'den büyük VIF değerlerinin ciddi çoklu bağlantı problemini işaret ettiğini belirtmiştir. Belsley, Kuh ve Welsch (1980), 25'ten büyük koşul sayısı olması durumunun ciddi çoklu bağlantı probleminin göstergesi olduğunu belirtmişlerdir.

Bir gözlem veya gözlemlerin etkili gözlem olması onun x yönündeki konumuyla, gözlenen y değeriyle veya her ikisiyle ilgili olabilir. Karma verilerde etkili gözlemlerin bir ölçüsü olarak Montgomery ve Voth (1994), $H = X(X'X)^{-1}X'$ şapka matrisinin köşegen elemanları h_{ii}'leri kullanmışlardır. Montgomery ve Voth (1994) x yönündeki aykırı değerlere karşı robust regresyon yöntemlerini kullanmayı önermişlerdir. Bir gözlemin aykırı olma ölçüsü olarak robust Mahalanobis uzaklıkları kullanılabilir (Rousseeuw ve Zomeren, 1990).

Tahmin ediciler

Ridge ve Liu Tahminleri

Kötü koşulluluk probleminin tahminler üzerindeki etkilerini azaltmak için en çok kullanılan iki yöntem Ridge ve Liu tahmin yanlı edicileridir. $k > 0$ yanlılık parametresi olmak üzere alışılmış Ridge tahmin edicisi, $\hat{\beta}_{OLS}$ regresyon parametrelerinin en küçük kareler tahmini olmak üzere,

$$\hat{\beta}_R(k) = (X'X + kI)^{-1}X'X\hat{\beta}_{OLS}$$

şeklinde verilir (Hoerl ve Kennard, 1970a, b). Liu tahmin edicisi ise $0 < d < 1$ yanlılık parametresi olmak üzere,

$$\hat{\beta}_L(d) = (X'X + I)^{-1}(X'X+dI)\hat{\beta}_{OLS}$$

şeklinde verilir (Liu, 1993).

Ancak dikkat edilirse her iki tahmin edicide sapan değerlere karşı hassas olan en küçük kareler tahminlerine dayalıdır. Silvapulle (1991) en küçük karelerin bu dezavantajından dolayı Ridge tahmin edicisinde, en küçük kareler yerine y yönündeki sapan değerlere karşı daha dirençli olan M tahmin edicilerini kullanmasıyla elde edilen robust Ridge tahmin edicisinin kullanılmasını önermiştir. Robust Ridge tahmin edicisi, k^*, k'nın bir robust seçimi ve $\hat{\beta}_M$, β'nın bir robust M tahmin edicisi olmak üzere,

$$\hat{\beta}_{RM}(k^*) = (X'X + kI)^{-1}X'X\hat{\beta}_M$$

dir. Arslan ve Billor (2000) Liu tahmin edicisi yerine M tahmin edicilerine dayalı robust Liu tahmin edicisinin kullanılmasını önermiştir. d^*, d'nin bir robust seçimi ve $\hat{\beta}_M$, β'nın bir robust M tahmin edicisi olmak üzere robust Liu tahmin edicisi,

$$\hat{\beta}_{LM}(d^*) = (X'X + I)^{-1}(X'X+dI)\,\hat{\beta}_M$$

dir. Burada σ ölçek parametresinin bir tahmini medyan mutlak sapma (MAD) kullanılarak elde edilecektir (Maronna, et.al., 2006).

$$\hat{\sigma} = \frac{1}{0.6745}.median|e_i - median(e_i)| = \frac{1}{0.6745}.mad(e_i) = 1.483.mad(e_i).$$

Genelleştirilmiş M (GM) tahmin edicisi

Regresyon parametreleri için GM tahmin edicisi aşağıdaki minimizasyon probleminin çözülmesiyle elde edilir:

$$Q(\beta) = \hat{\sigma}^2 \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i'\beta}{\hat{\sigma} u_i^{\theta}}\right) u_i^{1+\theta},$$

Burada ρ, $[0, \infty)$ aralığında azalmayan fonksiyon, u_i,

$$u_i = \min\left[1, \left\{\frac{b}{(x_i - \bar{x}_c)'C^{-1}(x_i - \bar{x}_c)}\right\}^{1/2}\right],$$

şeklinde ağırlık fonksiyonu ve $\hat{\sigma}$ yukarıda verilen ölçek parametresinin tahminidir. Buna göre $y = (y_1, \ldots, y_n)'$, $X = [x_1, \ldots, x_n]'$ and $\tilde{U} = diag(\tilde{u}_1,...,\tilde{u}_n)$ olmak üzere matris formunda GM tahmini

$$\hat{\beta}_{GM} = (X'\tilde{U}X)^{-1}X'\tilde{U}y,$$

şeklinde elde edilir.

GM tahminine dayalı Robust Ridge ve Robust Liu tahmin edicileri

Arslan ve Billor (1996) hem *x* hem de *y* yönündeki aykırı değerlere karşı dayanıklı olan Ridge tipi GM tahmin edicisini önermiştir. k^+, k'nın bir robust seçimi olmak üzere bu tahmin edicinin iki formunu vermişlerdir. Bunlardan ilki,

$$\hat{\beta}_{RGM}(k^+) = (X'X + k^+I)^{-1}X'X\,\hat{\beta}_{GM},$$

ve ikincisi,

$$\hat{\beta}_{RGMU}(k^+) = (X'UX + k^+I)^{-1}X'UX\,\hat{\beta}_{GM},$$

şeklindedir. Ayrıca $MSE\{\hat{\beta}_{RGM}(k^+)\} < MSE\{\hat{\beta}_{GM}\}$, $MSE\{\hat{\beta}_{RGMU}(k^+)\} < MSE\{\hat{\beta}_{GM}\}$ and $MSE\{\hat{\beta}_{RGM}(k^+)\} < MSE\{\hat{\beta}_R(k)\}$ olduğu gösterilmiştir.

Bu çalışmada, aynı anda çoklu bağlantı ve aykırı değerler problemlerine karşı dirençli olan GM ve Liu tahminlerinin birleştirilmesiyle elde edilen GM tahminine dayalı Liu tahmin edicileri tanıtılacaktır. Bunu için d^+, d'nin bir robust seçimi olmak üzere Liu tahminindeki $\hat{\beta}_{OLS}$ en küçük kareler tahmini $\hat{\beta}_{GM}$ ile değiştirilerek özelliklerini incelenecektir. Bu tahmin edicinin ilk formu aşağıdaki gibidir:

$$\hat{\beta}_{LGM}(d^+) = (X'X + I)^{-1}(X'X + d^+I)\,\hat{\beta}_{GM},$$

Bir diğer formu ise

$$\hat{\beta}_{LGMU}(d^+) = (X'UX + I)^{-1}(X'UX + d^+I)\,\hat{\beta}_{GM}$$

şeklinde verilmiştir.

Hata Kareler Ortalaması Özellikleri
Yukarıda verilen regresyon modeli kanonik formda $C = XP$ ve $\alpha = P'\beta$ olmak üzere

$$y = C\alpha + \varepsilon,$$

şeklinde yazılabilir. Modelde sabit terim yoktur. Buna göre en küçük kareler ve karşılık gelen Liu tahmini sırasıyla

$$\hat{\alpha}_{OLS} = \Lambda^{-1}C'Y$$

ve

$$\hat{\alpha}_{L}(d) = (\Lambda + I)^{-1}(\Lambda + dI)\,\hat{\alpha}_{OLS}.$$

şeklindedir. GM ve buna karşılık gelen Liu tahminleri ise,

$$\hat{\alpha}_{LGM}(d) = (\Lambda + I)^{-1}(\Lambda + dI)\,\hat{\alpha}_{GM}$$

ve

$$\hat{\alpha}_{LGMU}(d) = (B + I)^{-1}(B + dI)\,\hat{\alpha}_{GM}$$

burada $\mu_1 \geq \ldots \geq \mu_p$, $X'UX$'in özdeğeri, $B = \text{diag}(\mu_1, \ldots, \mu_p)$ ve U yukarıda verilen formadadır. Şimdi $\hat{\alpha}_{LGM}(d)$ ve $\hat{\alpha}_{LGMU}(d)$ tahminleri için aşağıdaki teoremi verebiliriz.

Teorem: $\hat{\alpha}_{GM}$'nin kovaryans matrisi sınırlı olsun ve $Cov(\hat{\alpha}_{GM}) = \Gamma$ alalım.

i) $MSE(\hat{\alpha}_{LGM}(d)) < MSE(\hat{\alpha}_{GM})$ olacak şekilde bir d sayısı vardır.

ii) $MSE(\hat{\alpha}_{LGMU}(d)) < MSE(\hat{\alpha}_{GM})$ olacak şekilde bir d sayısı vardır.

iii) Γ_{ii}'ler Γ'nın köşegen elemanları olmak üzere; Eğer her i için $\Gamma_{ii} < \dfrac{\sigma^2}{\lambda_i}$ ise her pozitif d sayısı için $MSE(\hat{\alpha}_{LGM}(d)) < MSE(\hat{\alpha}_{L}(d))$'dir.

Sonuç
Karma denemelerde çoklu bağlantı problemi önemli bir problemdir. Özellikle ek kısıtlamalar bulunan modellerde bu problem daha ciddi boyutlara ulaşabilir. Karma verilerde çoklu bağlantı olması durumunda psudobileşen dönüşümü ile bazen VIF değerlerinde iyileşmeler olsa da bu genellikle yeterli seviyelerde olmamaktadır. Yanlı tahmin edicilerde en küçük kareler tahminine dayalı olup çoklu bağlantı probleminin etkilerini azaltabilirler. Ancak aykırı değerlerin olması durumunda en küçük kareler tahmini bundan çok etkilenmektedir. Bu çalışmada hem çoklu

bağlantı hem de aykırı değer problemlerine sahip karma verilerde daha tutarlı tahminler verebilecek robust yanlı tahmin edicilerin kullanılması ele alınmıştır. Bu amaçla hem *x* hem de *y* yönündeki aykırı değerlere karşı dirençli olan GM tahminine dayalı ridge ve Liu tahminleri incelenmiştir. GM tahminine dayalı Liu tahmin edicisinin MSE özellikleri incelenerek diğer tahmin edicilerle karşılaştırılmış ve karma verilerde etkili olduğu görülmüştür. Genel olarak, GM tahmin edicisine dayalı robust Liu yanlı tahmin edicisinin, hem çoklu bağlantı hem de etkili gözlemlere karşı aynı anda dirençli olduğunu söyleyebilmek için kuşkusuz başka alanlarda (eğitim bilimleri, ekonometri gibi) ve daha yüksek gözlem sayısına sahip veri gruplarında etkinliğinin incelenmesi gerekmektedir.