Asymptotic Expansion of Risk for a Regression Model with respect to α-Divergence with an Application to the Sample Size Problem

Yo Sheena[*]

April 2017

For a regression model, we consider the risk of the maximum likelihood estimator with respect to α-divergence, which includes the special cases of Kullback-Leibler divergence, Hellinger distance and χ^2 divergence. The asymptotic expansion of the risk with respect to the sample size n is given up to the order n^{-2}. We observed how the risk convergence speed (to zero) is affected by the error term distributions and the magnitude of the joint moments of the standardized explanatory variables under three concrete error term distributions: a normal distribution, a t-distribution and a skew-normal distribution. We try to use the (approximated) risk of m.l.e. as a measure of the difficulty of estimation for the regression model. Especially comparing the value of the (approximated) risk with that of a binomial distribution, we can give a certain standard for the sample size required to estimate the regression model.

MSC(2010) *Subject Classification*: Primary 60F99; Secondary 62F12
Key words and phrases: alpha divergence, asymptotic expansion, regression model, asymptotic risk

1 Introduction

We consider the following regression model;

$$y = \beta' \tilde{x} + \sigma\epsilon, \tag{1}$$

[*] Faculty of Economics and Law, Shinshu University. Faculty of Data Science, Shiga University

where

$$\beta' = (\beta_0, \beta_1, \dots, \beta_p)$$

is the $p+1$-dimensional parameter vector, while

$$\tilde{x}' = (x_0, x'), \quad x_0 \equiv 1, \qquad x' = (x_1, \dots, x_p),$$

and $x' = (x_1, \dots, x_p)$ is a p-dimensional explanatory random vector. ϵ is the error term. We assume that the distributions of ϵ is known, but the distribution of x is unknown. The unknown parameters to be estimated are $\beta \in R^{p+1}$ and $\sigma (> 0)$. Without loss of generality, we can assume that $x' = (x_1, \dots, x_p)$ is standardized, i.e.

$$E[x_i] = 0, \quad i = 1, \dots, p, \qquad E[x_i x_j] = \begin{cases} 1, & \text{if } 1 \le i = j \le p, \\ 0, & \text{if } 1 \le i \neq j \le p. \end{cases} \tag{2}$$

Let $f(\epsilon)$ and $h(x)$ respectively be the p.d.f.'s of ϵ and x, then the p.d.f. of (y, x) is given by

$$f(y, x \,|\, \beta, \sigma) \triangleq f_x(y|\, \beta, \sigma)\, h(x), \tag{3}$$

where

$$f_x(y|\, \beta, \sigma) \triangleq \frac{1}{\sigma} f\left(\frac{y - \beta'\tilde{x}}{\sigma}\right).$$

We assume that $f(\epsilon)$ is positive and differentiable three times over the real line.

Let's consider the maximum likelihood estimators (say $\hat{\beta}, \hat{\sigma}$) of β, σ. One way to evaluate the performance of m.l.e. is the closeness of the predictive distribution designated by the p.d.f

$$f(y, x \,|\, \hat{\beta}, \hat{\sigma}) = f_x(y|\, \hat{\beta}, \hat{\sigma})\, h(x) = \frac{1}{\hat{\sigma}} f\left(\frac{y - \hat{\beta}'\tilde{x}}{\hat{\sigma}}\right) h(x) \tag{4}$$

to the true distribution given by (3).

We adopt divergences as the measure of closeness between two given distributions. A divergence is a premetric. Namely a divergence function $D[d_1 : d_2]$ evaluated at two distributions d_1 and d_2 on a same sigma field $\mathcal{X}$ satisfies

$$D[d_1 : d_2] \ge 0 \text{ for any distributions } d_1 \text{ and } d_2$$

with equality iff $d_1 = d_2$, but it is asymmetric, and the "triangular inequality" does not always hold.

Among possible divergences, *f-divergence* is natural in dealing with probability distributions. (See Amari and Nagaoka [3], Vajda [9].) First f-divergence is parameter-free. If we change the way of parametrization of a parametric model, f-divergence is invariant in the following sense. Suppose a distribution d on $\mathcal{X}$ can be designated by a parameter θ in a parametric model $P_\theta = \{(d|\theta) \,|\, \theta \in \Theta\}$, while it is expressed in another parametrization as $(d|\eta)$ in $P_\eta = \{(d|\eta) \mid \eta \in H\}$. If $(d|\theta_i)$ and $(d|\eta_i)$ is the same distribution for $i = 1, 2$,

$$D[(d|\theta_1) : (d|\theta_2)] = D[(d|\eta_1) : (d|\eta_2)].$$

Second it is invariant with respect to the transformation between the random variables that retains information. Let $Y(X)$ be a sufficient statistic for the parametric model of a random object X, then f-divergence satisfies

$$D[(X|\theta_1):(X|\theta_2)] = D[(Y|\theta_1):(Y|\theta_2)],$$

where $(X|\theta_i)$ is the distribution of X given by a parameter θ_i $(i=1,2)$.

In order to proceed a practical investigation of regression models, we need a more specific form of f-divergence. In this paper we focus on an α-divergence. It is an important subclass of f-divergence. Generally a divergence gives a geometrical structure on the manifold of a parametric distribution model, $P_\theta = \{(d|\theta)\,|\,\theta \in \Theta\}$. (See Eguchi [5], Amari and Nagoka [3].) The possible geometrical structures given by f-divergence can be realized by α-divergences. Furthermore it is a basic divergence from the perspective of information geometry since it gives rise to a "dual" structure between α and $-\alpha$ for the manifold of the given parametric model (see Eguchi [5], Amari [1], and Amari and Cichocki [2]). Specifically α-divergence $(-\infty < \alpha < \infty)$ between the two distributions, each of which is given respectively by the p.d.f. $f(x;\theta_1)$ and $f(x;\theta_2)$, is defined as

$$\overset{\alpha}{D}[\theta_1:\theta_2] = \begin{cases} \frac{4}{1-\alpha^2}\Big\{1-\int_{\mathfrak{X}} f^{(1-\alpha)/2}(x;\theta_1)f^{(1+\alpha)/2}(x;\theta_2)d\mu\Big\}, & \text{if } \alpha \neq \pm 1, \\ \int_{\mathfrak{X}} f(x;\theta_2)\log\Big(f(x;\theta_2)/f(x;\theta_1)\Big)d\mu, & \text{if } \alpha = 1, \\ \int_{\mathfrak{X}} f(x;\theta_1)\log\Big(f(x;\theta_1)/f(x;\theta_2)\Big)d\mu, & \text{if } \alpha = -1. \end{cases} \tag{5}$$

α-divergence is a broad class of divergences. Actually it includes Kullback–Leibler divergence $(\alpha=-1)$, the Hellinger distance $(\alpha=0)$ and χ^2 divergence $(\alpha=3)$.

We will measure the performance of m.l.e. $\hat{\beta}$, $\hat{\sigma}$ by the expected α-divergence between two distributions (3) and (4);

$$\overset{\alpha}{ED}(\beta,\sigma) \triangleq E\big[\overset{\alpha}{D}[(\hat{\beta}(\boldsymbol{y},\boldsymbol{x}),\hat{\sigma}(\boldsymbol{y},\boldsymbol{x})):(\beta,\sigma)]\big], \tag{6}$$

where $(\boldsymbol{y},\boldsymbol{x}) = \Big((y_1,x_1),\ldots,(y_n,x_n)\Big)$ are n independent random samples from the true distribution (3). In other words, we evaluate the performance of m.l.e. using the risk of m.l.e. with respect to an α-divergence. However, this risk of m.l.e. can not be gained explicitly in many (most in a practical sense) cases, hence its asymptotic expansion with n is useful since it gives a good approximation under a large size of samples. Sheena [7] gave the asymptotic expansion of $\overset{\alpha}{ED}$ up to the n^{-2} order for a general parametric model. (Henceforth, we will call the truncated $\overset{\alpha}{ED}$ up to the n^{-2} order by the name of "the approximated $\overset{\alpha}{ED}$".) In this paper, we focused ourselves on the regression model (1), and derived the approximated $\overset{\alpha}{ED}$ for it.

The result for a general regression model (1) is still too lengthy to be out of use for a practical purpose. So we narrowed our scope further to some specific error distributions. (See Mathematica program in Appendix of Sheena [8] which enables us to calculate the approximated $\overset{\alpha}{ED}$, once the p.d.f. (and its derivatives) of an arbitrary error distribution is given.) This paper is constructed as follows; In Section 2, we explained how the general result of [7] is applied to the regression model. In Section 3, we considered three specific error term distributions and observed an explicit form of the expansion of $\overset{\alpha}{ED}$: a normal distribution (Section 3.1), a t-distribution (Section 3.2), a skew-normal distribution (Section 3.3). In Section 3.4, we made a comparison among these three error distributions. Throughout Section 3, we considered the case where the explanatory variable x has a homogeneous distribution (i.e. invariant w.r.t. the permutations of the $x_i, i = 1, \ldots, p$). We combined the above error term distributions with various types of joint moments of x to gain a concrete form of the approximated $\overset{\alpha}{ED}$ as the function of n, p, α. We observed how n, p, α affect $\overset{\alpha}{ED}$. In Section 4, we treated two real datasets, which give us examples for non-homogeneous distribution of x.

As one of the possible applications of $\overset{\alpha}{ED}$, we considered the sample size problem, that is, "how large sample size is required to estimate the parameters of the regression model (1) ?". When a parametric distribution model is given, the difficulty of estimation (specification) of the parameter for that model could be measured in various ways. Sheena [7] proposed to measure it by the approximated $\overset{\alpha}{ED}$. In the paper, the author tried to use the approximated $\overset{\alpha}{ED}$ of a binomial distribution model $B(n,p)$ as a benchmark since it gives us an intuitive interpretation. For example, if a parametric distribution model has a similar value of $\overset{\alpha}{ED}(\theta)$ (at a given θ) to $B(10, 0.01)$, we can understand that the task of the estimation is hard, since the value 0.01 is too small to be estimated from as little as 10 samples. On the contrary, $\overset{\alpha}{ED}(\theta)$ of the model is close to that of $B(10, 0.5)$, it is a relatively easy task to estimate the parameter.

In this paper, we formalized this idea and proposed two indicators (*I.D.E.* and *R.S.S.*) that could be used for a sample size problem. In Section 2, we gave the definition of the both indicators. In Section 3 and 4, we calculated their concrete values under the given error distributions and the moments of x, and tried to give a solution to the sample size problem.

2 Asymptotic risk of m.l.e. w.r.t. α-divergence

First we introduce a general result of Sheena [7] on the asymptotic risk of m.l.e. with respect to α-divergence. In order to improve readability, we use Einstein's summation convention, that is, the summation carried out as every pair of upper and lower index moves from 1 to p.

Let $\mathcal{P}$ be a parametric family of probability distributions on a space $\mathfrak{X}$, which is given by a family of positive-valued densities $f(x;\theta)$ on $\mathfrak{X}$ with respect to a measure μ:

$$\mathcal{P} = \{f(x;\theta) \mid \theta = (\theta^1, \ldots, \theta^p) \in \Theta\}, \tag{7}$$

where Θ is an open set in R^p.

Consider the maximum likelihood estimator $\hat{\theta}(\boldsymbol{X})$ of θ based on n samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ independently chosen from the distribution $f(x;\theta)$. Closeness $\hat{\theta}$ and the true parameter θ is measured by (5), namely $\overset{\alpha}{D}[\hat{\theta}(\boldsymbol{X}) : \theta]$. The risk is defined as the expectation of this random variable;

$$\overset{\alpha}{ED}(\theta) \triangleq E_\theta\big[\overset{\alpha}{D}[\hat{\theta}(\boldsymbol{X}) : \theta]\big]. \tag{8}$$

The asymptotic expansion of $\overset{\alpha}{ED}$ w.r.t. n is given by

$$\begin{aligned}
&\overset{\alpha}{ED} \\
&= \frac{p}{2n} + \frac{1}{24n^2} \\
&\times \Big[(\alpha')^2\big\{3\overset{e}{F} + 3T^{ijk}T_{ijk} - 6\langle \overset{e}{A}{}_i^j, (\overset{m}{A}{}_j^i - \overset{e}{A}{}_j^i)\rangle - 3\langle \overset{e}{A}{}_i^i, (\overset{m}{A}{}_j^j - \overset{e}{A}{}_j^j)\rangle + 3p^2 + 6p\big\} \\
&\qquad + \alpha'\big\{3\overset{e}{F} - 5T^{ijk}T_{ijk} - 6T^i_{is}T^{js}_j + 6\langle \overset{e}{A}{}_i^j, (\overset{m}{A}{}_j^i - \overset{e}{A}{}_j^i)\rangle + 3\langle \overset{e}{A}{}_i^i, (\overset{m}{A}{}_j^j - \overset{e}{A}{}_j^j)\rangle \\
&\qquad\qquad - 3p^2 - 6p\big\} \\
&\qquad + 12\langle \overset{e}{A}{}_j^i, \overset{e}{A}{}_i^j\rangle - 2\langle \overset{e}{A}{}_j^i, \overset{m}{A}{}_i^j\rangle - \langle \overset{e}{A}{}_i^i, \overset{m}{A}{}_j^j\rangle + T_{ijk}T^{ijk} + 9T^i_{is}T^{js}_j + 8\overset{e}{R}{}_{ij}{}^{ij} - 9\overset{e}{F}\Big] \\
&+ o(n^{-2}),
\end{aligned} \tag{9}$$

where $\alpha' = (1-\alpha)/2$. The main term equals $p/2n$. p/n is the ratio of the number of the parameters to the sample size. (We will call this quantity "$p-n$ ratio" hereafter.) The coefficient of n^{-2}, i.e. the terms inside the bracket have a geometrical meaning if we view $\mathcal{P}$ as a Riemannian manifold. We omit the geometrical explanation (see Sheena [7]), and just describe their formal definitions.

Define the following notations; for $1 \le i, j, k, l \le p$,

$$\begin{aligned}
&L_{(ij)} \triangleq E_\theta[l_{ij}], \quad L_{ij} \triangleq E_\theta[l_i l_j], \\
&L_{(ij)k} \triangleq E_\theta[l_{ij}l_k], \quad L_{ijk} \triangleq E_\theta[l_i l_j l_k] \\
&L_{(ij)(kl)} \triangleq E_\theta[l_{ij}l_{kl}], \quad L_{(ijk)l} \triangleq E_\theta[l_{ijk}l_l], \quad L_{(ij)kl} \triangleq E_\theta[l_{ij}l_k l_l], \quad L_{ijkl} \triangleq E_\theta[l_i l_j l_k l_l],
\end{aligned} \tag{10}$$

$$\begin{aligned}
&L11 \triangleq g^{ij}g^{kl}L_{(il)jk}, \quad L12 \triangleq g^{ij}g^{kl}L_{(ij)kl}, \quad L13 \triangleq g^{ij}g^{kl}L_{ijkl}, \\
&L14 \triangleq g^{ij}g^{kl}L_{(ik)(jl)}, \quad L15 \triangleq g^{ij}g^{kl}L_{(ij)(kl)}, \\
&L21 \triangleq g^{ij}g^{kl}g^{su}L_{(ik)s}L_{jlu}, \quad L22 \triangleq g^{ij}g^{kl}g^{su}L_{(ij)k}L_{lsu}, \\
&L23 \triangleq g^{ij}g^{kl}g^{su}L_{iks}L_{jlu}, \quad L24 \triangleq g^{ij}g^{kl}g^{su}L_{ijk}L_{lsu}, \\
&L25 \triangleq g^{ij}g^{kl}g^{su}L_{(ik)s}L_{(jl)u}, \quad L26 \triangleq g^{ij}g^{kl}g^{su}L_{(ij)k}L_{(su)l},
\end{aligned} \tag{11}$$

where (g^{ij}) is the inverse matrix of (g_{ij}) given by

$$g_{ij} \triangleq L_{ij} (\equiv -L_{(ij)}),$$

and

$$l_i \triangleq l_i(x;\theta) \triangleq \frac{\partial}{\partial\theta_i}\log f(x;\theta), \quad l_{ij} \triangleq l_{ij}(x;\theta) \triangleq \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\theta), \quad \cdots,$$

$$E_\theta[h(x;\theta)] \triangleq \int_{\mathfrak{X}} h(x;\theta) f(x;\theta) d\mu.$$

Then each term of (9) is defined as follows.

$$\begin{aligned}
\overset{e}{F} &= g^{ij} g^{ks} \big(2L_{(is)jk} + L_{(ks)ij} + L_{ijks}\big) \\
&\quad - g^{ks} g^{uj} g^{li} L_{ijk} \big(2L_{(su)l} + L_{sul}\big) \\
&\quad - g^{ti} g^{uj} g^{ks} L_{(it)s} L_{juk} \\
&= 2L11 + L12 + L13 - 2L21 - L23 - L22, && (12) \\
T_{ijk} T^{ijk} &= L_{ijk} L_{stu} g^{is} g^{jt} g^{ku} = L23, && (13) \\
T^i_{is} T^{js}_j &= L_{ijk} L_{stu} g^{ij} g^{st} g^{uk} = L24, && (14) \\
\overset{e}{R}_{ij}{}^{ij} &= g^{ij} g^{sk} \big(L_{(ki)(js)} - L_{(ij)(ks)} + L_{(ki)js} - L_{(ij)ks}\big) \\
&\quad + g^{sk} g^{ti} g^{uj} \big(-L_{(ki)j} L_{(st)u} + L_{(it)s} L_{(uj)k} + L_{sit} L_{(uj)k} - L_{stu} L_{(ij)k}\big) \\
&= L14 - L15 + L11 - L12 - L25 + L26 + L22 - L21, && (15) \\
\langle \overset{e}{A}{}^j_i, \overset{e}{A}{}^i_j \rangle &= g^{jk} g^{li} L_{(ik)(jl)} - g^{jk} g^{li} g^{st} L_{(ik)s} L_{(jl)t} - p \\
&= L14 - L25 - p, && (16) \\
\langle \overset{e}{A}{}^i_i, \overset{e}{A}{}^j_j \rangle &= g^{ik} g^{jl} L_{(ik)(jl)} - g^{ik} g^{jl} g^{st} L_{(ik)s} L_{(jl)t} - p^2 \\
&= L15 - L26 - p^2, && (17) \\
\langle \overset{e}{A}{}^j_i, \overset{m}{A}{}^i_j \rangle &= g^{jk} g^{li} L_{(ik)jl} + g^{jk} g^{li} L_{(ik)(jl)} \\
&\qquad\qquad - g^{jk} g^{li} g^{st} L_{(ik)s} L_{(jl)t} - g^{jk} g^{li} g^{st} L_{(ik)s} L_{jlt} \\
&= L11 + L14 - L25 - L21, && (18) \\
\langle \overset{e}{A}{}^i_i, \overset{m}{A}{}^j_j \rangle &= g^{ik} g^{jl} L_{(ik)jl} + g^{ik} g^{jl} L_{(ik)(jl)} \\
&\qquad\qquad - g^{ik} g^{jl} g^{st} L_{(ik)s} L_{(jl)t} - g^{ik} g^{jl} g^{st} L_{(ik)s} L_{jlt} \\
&= L12 + L15 - L26 - L22. && (19)
\end{aligned}$$

Now we apply (9) to the case where $\mathcal{P}$ is given by

$$\mathcal{P} = \{f(y, x \,|\, \beta, \sigma) \,|\, \beta \in R^{p+1},\ \sigma > 0\},$$

where $f(y, x \,|\, \beta, \sigma)$ is given by (3).

Accordingly we define the following notations; for $i, j, k, l = 0, 1, \ldots, p, \sigma$

$L_{(ij)} \triangleq E_{\beta,\sigma}[l_{ij}], \quad L_{ij} \triangleq E_{\beta,\sigma}[l_i l_j],$
$L_{(ij)k} \triangleq E_{\beta,\sigma}[l_{ij} l_k], \quad L_{ijk} \triangleq E_{\beta,\sigma}[l_i l_j l_k]$
$L_{(ij)(kl)} \triangleq E_{\beta,\sigma}[l_{ij} l_{kl}], \quad L_{(ijk)l} \triangleq E_{\beta,\sigma}[l_{ijk} l_l], \quad L_{(ij)kl} \triangleq E_{\beta,\sigma}[l_{ij} l_k l_l], \quad L_{ijkl} \triangleq E_{\beta,\sigma}[l_i l_j l_k l_l],$

where

$$l_i \triangleq \partial_i \log f(y, x|\beta, \sigma), \quad l_{ij} \triangleq \partial_i \partial_j \log f(y, x|\beta, \sigma), \quad l_{ijk} \triangleq \partial_i \partial_j \partial_k \log f(y, x|\beta, \sigma)$$

with $\partial_i (i = 0, 1, \ldots, p, \sigma)$ defined by

$$\partial_i = \begin{cases} \frac{\partial}{\partial \beta_i} & \text{if } 0 \leq i \leq p, \\ \frac{\partial}{\partial \sigma} & \text{if } i = \sigma, \end{cases}$$

and

$$E_{\beta,\sigma}[h(y, x; \beta, \sigma)] = \int_{R^p} \int_R h(y, x; \beta, \sigma) f(y, x|\beta, \sigma) dy dx.$$

We also define other notations.
For $0 \leq i, j \leq p$,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

For $0 \leq i, j, k, l \leq 4$,

$$\eta[i, j, k, l] \triangleq \int_{-\infty}^{\infty} \Big(\frac{d^3 \log f(y)}{dy^3}\Big)^i \Big(\frac{d^2 \log f(y)}{dy^2}\Big)^j \Big(\frac{d \log f(y)}{dy}\Big)^k y^l f(y) dy. \tag{20}$$

For $i, j, k, l \in \{0, 1, \ldots, p, \sigma\}$

$$\begin{aligned} m[i, j, k] &\triangleq E[\dot{x}_i \dot{x}_j \dot{x}_k] = \int_{R^p} \dot{x}_i \dot{x}_j \dot{x}_k h(x) dx, \\ m[i, j, k, l] &\triangleq E[\dot{x}_i \dot{x}_j \dot{x}_k \dot{x}_l] = \int_{R^p} \dot{x}_i \dot{x}_j \dot{x}_k \dot{x}_l h(x) dx, \end{aligned} \tag{21}$$

where

$$\dot{x}_i = \begin{cases} x_i & \text{if } i \in \mathcal{I} \triangleq \{1, 2, \ldots, p\}, \\ 1 & \text{if } i \in \mathcal{S} \triangleq \{0, \sigma\}. \end{cases}$$

Straightforward calculation leads to the following results (see Appendix A of [8] for the detailed calculation).

$$g_{ij} = \delta_{ij} \sigma^{-2} \eta[0, 0, 2, 0] = -\delta_{ij} \sigma^{-2} \eta[0, 1, 0, 0], \quad 0 \leq i, j \leq p. \tag{22}$$

$$g_{i\sigma} = \begin{cases} \sigma^{-2} \eta[0, 0, 2, 1] = -\sigma^{-2} \eta[0, 1, 0, 1] & \text{if } i = 0, \\ 0 & \text{if } 1 \leq i \leq p. \end{cases} \tag{23}$$

$$\begin{aligned} g_{\sigma\sigma} &= \sigma^{-2}(1 + 2\eta[0, 0, 1, 1] + \eta[0, 0, 2, 2]). \\ &= -\sigma^{-2}(1 + \eta[0, 1, 0, 2] + 2\eta[0, 0, 1, 1]) \end{aligned} \tag{24}$$

$$g^{ij} = \delta_{ij} \sigma^2 \eta^{-1}[0, 0, 2, 0], \quad 1 \leq i, j \leq p. \tag{25}$$

$$g^{0i} = g^{\sigma i} = 0, \quad 1 \leq i \leq p. \tag{26}$$

$$g^{00} = \sigma^2 \Delta^{-1}(1 + 2\eta[0, 0, 1, 1] + \eta[0, 0, 2, 2]). \tag{27}$$

$$g^{0\sigma} = \sigma^2 \Delta^{-1} \eta[0, 1, 0, 1]. \tag{28}$$

$$g^{\sigma\sigma} = \sigma^2 \Delta^{-1} \eta[0, 0, 2, 0]. \tag{29}$$

$$(\Delta = \eta[0, 0, 2, 0](1 + 2\eta[0, 0, 1, 1] + \eta[0, 0, 2, 2]) - \eta^2[0, 1, 0, 1])$$

For $i, j, k, l = 0, 1, \ldots, p, \sigma$,

$$\begin{aligned} L_{(ij)k} &= \sigma^{-3} m[i,j,k] \eta_{(ij)k} && (30) \\ L_{ijk} &= \sigma^{-3} m[i,j,k] \eta_{ijk} && (31) \\ L_{(ij)(kl)} &= \sigma^{-4} m[i,j,k,l] \eta_{(ij)(kl)} && (32) \\ L_{(ijk)l} &= \sigma^{-4} m[i,j,k,l] \eta_{(ijk)l} && (33) \\ L_{(ij)kl} &= \sigma^{-4} m[i,j,k,l] \eta_{(ij)kl} && (34) \\ L_{ijkl} &= \sigma^{-4} m[i,j,k,l] \eta_{ijkl}, && (35) \end{aligned}$$

where for $0 \leq i, j, k, l \leq p$,

$$\begin{aligned}
\eta_{(ij)k} &= -\eta[0,1,1,0] && (36) \\
\eta_{(i\sigma)k} &= -(\eta[0,1,1,1] + \eta[0,0,2,0]) && (37) \\
\eta_{(ij)\sigma} &= -(\eta[0,1,0,0] + \eta[0,1,1,1]) && (38) \\
\eta_{(i\sigma)\sigma} &= -(\eta[0,1,0,1] + \eta[0,1,1,2] + \eta[0,0,2,1]) && (39) \\
\eta_{(\sigma\sigma)i} &= -(\eta[0,1,1,2] + 2\eta[0,0,2,1]) && (40) \\
\eta_{(\sigma\sigma)\sigma} &= -(1 + 3\eta[0,0,1,1] + \eta[0,1,0,2] + 2\eta[0,0,2,2] + \eta[0,1,1,3]) && (41) \\
\eta_{ijk} &= -\eta[0,0,3,0] && (42) \\
\eta_{ij\sigma} &= -(\eta[0,0,2,0] + \eta[0,0,3,1]) && (43) \\
\eta_{i\sigma\sigma} &= -(2\eta[0,0,2,1] + \eta[0,0,3,2]) && (44) \\
\eta_{\sigma\sigma\sigma} &= -(1 + 3\eta[0,0,1,1] + 3\eta[0,0,2,2] + \eta[0,0,3,3]) && (45) \\
\eta_{(ij)(kl)} &= \eta[0,2,0,0] && (46) \\
\eta_{(i\sigma)(kl)} &= \eta[0,2,0,1] + \eta[0,1,1,0] && (47) \\
\eta_{(i\sigma)(j\sigma)} &= \eta[0,2,0,2] + 2\eta[0,1,1,1] + \eta[0,0,2,0] && (48) \\
\eta_{(ij)(\sigma\sigma)} &= \eta[0,1,0,0] + \eta[0,2,0,2] + 2\eta[0,1,1,1] && (49) \\
\eta_{(i\sigma)(\sigma\sigma)} &= \eta[0,1,0,1] + \eta[0,2,0,3] + 3\eta[0,1,1,2] + 2\eta[0,0,2,1] && (50) \\
\eta_{(\sigma\sigma)(\sigma\sigma)} &= 1 + \eta[0,2,0,4] + 4\eta[0,0,2,2] + 2\eta[0,1,0,2] \\
&\quad + 4\eta[0,0,1,1] + 4\eta[0,1,1,3] && (51) \\
\eta_{(ijk)l} &= \eta[1,0,1,0] && (52) \\
\eta_{(ijk)\sigma} &= \eta[1,0,0,0] + \eta[1,0,1,1] && (53) \\
\eta_{(ij\sigma)k} &= 2\eta[0,1,1,0] + \eta[1,0,1,1] && (54) \\
\eta_{(i\sigma\sigma)j} &= 4\eta[0,1,1,1] + 2\eta[0,0,2,0] + \eta[1,0,1,2] && (55) \\
\eta_{(ij\sigma)\sigma} &= 2\eta[0,1,0,0] + \eta[1,0,0,1] + 2\eta[0,1,1,1] + \eta[1,0,1,2] && (56) \\
\eta_{(i\sigma\sigma)\sigma} &= 4\eta[0,1,0,1] + \eta[1,0,0,2] + 4\eta[0,1,1,2] + 2\eta[0,0,2,1] + \eta[1,0,1,3] && (57) \\
\eta_{(\sigma\sigma\sigma)i} &= 6\eta[0,1,1,2] + 6\eta[0,0,2,1] + \eta[1,0,1,3] && (58) \\
\eta_{(\sigma\sigma\sigma)\sigma} &= 2 + 6\eta[0,1,0,2] + 6\eta[0,0,1,1] + \eta[1,0,0,3] \\
&\quad + 2\eta[0,0,1,1] + 6\eta[0,1,1,3] + 6\eta[0,0,2,2] + \eta[1,0,1,4] && (59) \\
\eta_{(ij)kl} &= \eta[0,1,2,0] && (60)
\end{aligned}$$

$$\eta_{(ij)k\sigma} = \eta[0,1,1,0] + \eta[0,1,2,1] \tag{61}$$

$$\eta_{(i\sigma)jk} = \eta[0,1,2,1] + \eta[0,0,3,0] \tag{62}$$

$$\eta_{(ij)\sigma\sigma} = \eta[0,1,0,0] + 2\eta[0,1,1,1] + \eta[0,1,2,2] \tag{63}$$

$$\eta_{(i\sigma)j\sigma} = \eta[0,1,1,1] + \eta[0,0,2,0] + \eta[0,1,2,2] + \eta[0,0,3,1] \tag{64}$$

$$\eta_{(\sigma\sigma)ij} = \eta[0,0,2,0] + 2\eta[0,0,3,1] + \eta[0,1,2,2] \tag{65}$$

$$\eta_{(i\sigma)\sigma\sigma} = \eta[0,1,0,1] + 2\eta[0,1,1,2] + 2\eta[0,0,2,1] + \eta[0,1,2,3] + \eta[0,0,3,2] \tag{66}$$

$$\eta_{(\sigma\sigma)i\sigma} = 2\eta[0,0,2,1] + \eta[0,1,1,2] + \eta[0,0,2,1] + 2\eta[0,0,3,2] + \eta[0,1,2,3] \tag{67}$$

$$\begin{aligned}\eta_{(\sigma\sigma)\sigma\sigma} &= 1 + 4\eta[0,0,1,1] + \eta[0,1,0,2] + 5\eta[0,0,2,2] \\ &\quad + 2\eta[0,1,1,3] + 2\eta[0,0,3,3] + \eta[0,1,2,4]\end{aligned} \tag{68}$$

$$\eta_{ijkl} = \eta[0,0,4,0] \tag{69}$$

$$\eta_{ijk\sigma} = \eta[0,0,3,0] + \eta[0,0,4,1] \tag{70}$$

$$\eta_{ij\sigma\sigma} = \eta[0,0,2,0] + 2\eta[0,0,3,1] + \eta[0,0,4,2] \tag{71}$$

$$\eta_{i\sigma\sigma\sigma} = 3\eta[0,0,2,1] + 3\eta[0,0,3,2] + \eta[0,0,4,3] \tag{72}$$

$$\eta_{\sigma\sigma\sigma\sigma} = 1 + 4\eta[0,0,1,1] + 6\eta[0,0,2,2] + 4\eta[0,0,3,3] + \eta[0,0,4,4]. \tag{73}$$

If we insert these results (25),...,(35) into (11), we can calculate the values of (12) to (19). Note that the summation (by Einstein's convention) in (11) to (19) is carried over the range $0, 1, \ldots, p, \sigma$ for each index. The calculation process is so lengthy that we used Mathematica [6]. The general result expressed with abstract notations $\eta[i,j,k,l]$ (see (20)) and $m[i,j,k], m[i,j,k,l]$ (see (21)) could be given, but it is too complicated to be out of use. Instead we put the Mathematica program in Appendix B of [8] so that we can easily calculate the approximated $\overset{\alpha}{ED}$ once the error term distribution and the moments of the explanatory variables are given, which respectively determine $\eta[i,j,k,l]$ and $m[i,j,k], m[i,j,k,l]$.

Generally $\overset{\alpha}{ED}$ for the parametric model (7) depends on θ. However $\overset{\alpha}{ED}$ for the regression model (1) is independent of β, σ. This is obvious from the fact that (25),...,(35) include only σ, but it vanishes at (11). We report that if the support of $f(\epsilon)$ is not the whole real line (e.g. $f(\epsilon) = 0$ for negative values of ϵ), $\eta[i,j,k,l]$, hence $\overset{\alpha}{ED}$ could be dependent on (β, σ).

In the next section, we give the explicit result when an error distribution and the moments of x are specified. We consider three specific cases where the error term distribution is respectively a normal distribution, a t-distribution and a skew-normal distribution. The different sets of the moments of x are combined with these error distributions to give illustrating examples.

Now we mention one of the possible applications of the approximated $\overset{\alpha}{ED}$. For a parametric distribution model (7), we naturally raise the following questions;

1. At which point θ, is the parameter most difficult to be estimated ?

2. Compared with another model, this model is easier or harder to be estimated ?

We propose to use the approximated $\overset{\alpha}{ED}$ to give an answer to these questions. Maximum likelihood is the most common estimation method and intrinsic to the model, hence it is natural to measure "the difficulty of estimating the model " by its performance such as the risk w.r.t a certain loss function. As we mentioned in Introduction, the risk w.r.t. α-divergence has favorable properties to answer to the above questions. In this paper we will use the approximated $\overset{\alpha}{ED}$ as a measure of the estimation difficulty.

In the case of the regression model (1), the answer to the first question is obvious. Since $\overset{\alpha}{ED}$ is constant (independent of β, σ^2), the difficulty of estimation is same all over the parameter space. Concerning the second question, we take the binomial distribution model $B(n, m)$ (n: the sample size, m: the probability of an event) as the benchmark for comparison.

The asymptotic expansion of $\overset{\alpha}{ED}$ for the binomial distribution $B(n, m)$ is given by

$$\begin{aligned}&\overset{\alpha}{ED}\\ &= \frac{1}{2n} + \frac{1}{24n^2}\Big[(\alpha')^2(3M-9) + \alpha'(-11M+29) + 10M - 22\Big] + o(n^{-2}),\end{aligned} \tag{74}$$

where $\alpha' = (1-\alpha)/2$ and $M \triangleq 1/m + 1/(1-m)$. (See the subsection 3.2 of Sheena [7].) For Kullback-Leibler divergence, put $\alpha = -1$, then we have

$$\overset{-1}{ED} = \frac{1}{2n} + \frac{1}{12n^2}(M-1) + o(n^{-2}). \tag{75}$$

The graph of the approximated $\overset{-1}{ED}$ for $B(10, m)$ is given in Figure 1. We notice that the approximated $\overset{-1}{ED}$ is stable around the area $0.1 \le m \le 0.9$, however it rapidly increases outside this area.

Let $\overset{\alpha}{ED}_{B(n,m)}$ denote the approximated $\overset{\alpha}{ED}$ for $B(n, m)$ and $\overset{\alpha}{ED}_R(n)$ denote that for a specific regression model where all the elements of the regression model (p, the error term distribution, the moments of x) are specified, hence $\overset{\alpha}{ED}_R$ is considered as the function of the sample size n. Here we propose an indicator of the difficulty of estimation.

–Indicator of the Difficulty of Estimation (I.D.E.)–

> Use a k times binomial experiment $B(k, m)$ as a benchmark. Solve the equation on m
>
> $$\overset{\alpha}{ED}_{B(k,m)} = \overset{\alpha}{ED}_R((p+2)k) \tag{76}$$

We easily notice the equation (76) is independent of k. Taking the sample size for the regression model as $(p+2)k$, we get the same $p-n$ ratio $1/k$ between the two models. Hence it makes sense to compare the n^{-2} order terms. The solution m tells us intuitively how difficult the parameter estimation is for the regression model. For example if $m = 0.001$, then we easily understand the estimation is difficult since it is difficult to estimate m as small as 0.001 based on just 10 samples. On the contrary, if we

Figure 1: $\overset{-1}{ED}$ of $B(10, m)$

have $m = 0.8$, then the estimation from 10 samples seems not so hard unless we require high precision.

The above equation (76) might have no real roots, that is, the left-hand side of the equation is larger than the right-hand side for any m. In this case, we can conclude that the regression model could be estimated more easily than the binomial model with the same $p - n$ ratio.

In a reverse way, we can use the approximated $\overset{\alpha}{ED}$ of the regression model for giving an answer to the sample size problem, that is, how large sample size is required to estimate the parameters of the regression model (1).

–Required Sample Size (R.S.S.)–

> Use a 10 times normal coin toss $B(10, 0.5)$ as a benchmark. Solve the equation
>
> $$\overset{\alpha}{ED}_{B(10,0.5)} = \overset{\alpha}{ED}_R(n). \tag{77}$$

The solution n indicates the sample size large enough to guarantee as easy estimation as 10 times normal coin toss.

The equation (77) could have no real roots. This means that the left-hand side of the equation is larger than the right-hand side for any n. Since the equation is based on the "approximated" $\overset{\alpha}{ED}$, we must notice that this does not necessarily mean just a small sample (e.g. $p+2$ samples) is enough for the estimation of the regression model. For the approximation to work well, the appropriate sample size is needed. If we want a concrete solution on the sample size problem, it could be gained by choosing an appropriately large k of $B(k, 0.5)$ instead of 10 on the left-hand side of (77).

Figure 2: p.d.f.'s of three error distributions.

3 Homogeneous Explanatory Variables

In this section, we consider three concrete forms of error distribution: a normal distribution, a t-distribution and a skew-normal distribution. A normal distribution is a theoretically basic error distribution. We are interested in how the fat tail property of a t-distribution or the skewness of a skew-normal distribution affects the (approximated) $\overset{\alpha}{ED}$. For contrasting these properties, we choose 3 for the d.f. of the t distribution and 3 for the shape parameter of the skew-normal distribution. Figure 2 is the graph of the p.d.f.'s of the three error distributions; the standard normal distribution ($N(0,1)$), the t-distribution with the d.f. of 3 ($t(3)$), the skew-normal distribution with the shape parameter of 3 ($SN(3)$).

$\overset{\alpha}{ED}$ also depends on the moments of x. As we can see from the definition (30)–(35), the maximum order of the joint moments of x is four that appear in the expansion of $\overset{\alpha}{ED}$ up to the n^{-2} order. In this section, we consider the homogeneous case where the distribution of $x = (x_1, \dots, x_p)$ is invariant w.r.t. any permutation of the elements. This is not practical but this case helps us observe the effect of the dimension p, so called "the curse of dimension".

Here we define the notations of the homogeneous moments of x as follows. For all distinguished i, j, k, l $(1 \leq i, j, k, l \leq p)$,

$$\begin{aligned}
& m_4 \triangleq E[x_i^4],\ m_{31} \triangleq E[x_i^3 x_j],\ m_{22} \triangleq E[x_i^2 x_j^2], \\
& m_{211} \triangleq E[x_i^2 x_j x_k],\ m_{1111} \triangleq E[x_i x_j x_k x_l], && (78) \\
& m_3 \triangleq E[x_i^3],\ m_{21} \triangleq E[x_i^2 x_j],\ m_{111} \triangleq E[x_i x_j x_k], && (79) \\
& m_2 \triangleq E[x_i^2] = 1,\ m_{11} \triangleq E[x_i x_j] = 0, && (80) \\
& m_1 \triangleq E[x_i] = 0, && (81) \\
& m_0 \triangleq E[x_0] = 1. && (82)
\end{aligned}$$

Under these homogeneous moments, we can state the approximated $\overset{\alpha}{ED}$ explicitly for each error distribution as a function of n, p, α and these moments . The result is given in the following subsections.

We used the following four distributions of x as specific examples of the moments of x when we want to analyze the approximated $\overset{\alpha}{ED}$ in a more concrete form:

1. The standard p-dimensional normal distributions, $N_p(0, I_p)$

$$\begin{aligned}&m_4 = 3,\ m_{31} = 0,\ m_{22} = 1,\ m_{211=0},\ m_{1111} = 0,\\ &m_3 = 0,\ m_{21} = 0,\ m_{111} = 0.\end{aligned}$$

2. The standard p-dimensional t-distribution, $t_p(0, I_p, \nu)$, that is, the p dimensional multivariate t-distribution with zero mean vector, the unit matrix as the scale matrix and the degree of freedom ν. Its p.d.f. is given by

$$h(x) \propto \left(1 + \nu^{-1} \sum_{i=1}^{p} x_i^2\right)^{-(\nu+p)/2}$$

Note that $E[x_i] = 0,\ i = 1, \dots, p$ and

$$Cov(x_i, x_j) = E[x_i x_j] = \begin{cases} \nu/(\nu - 2) & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

for $\nu > 2$. Therefore after the normalization (2), we have

$$E[x_i^2 x_j^2] = \begin{cases} 3(\nu - 2)/(\nu - 4) & \text{if } i = j, \\ (\nu - 2)/(\nu - 4) & \text{if } i \neq j, \end{cases}$$

under the condition $\nu > 4$. Notice that the effect of the fourth moment is enhanced by $(\nu - 2)/(\nu - 4)$ compared to the case $x \sim N_p(0, I_p)$. We want to check the effect of the fat tail property of a t-distribution. Here we put ν as 4.2, then we have

$$\begin{aligned}&m_4 = 33,\ m_{31} = 0,\ m_{22} = 11,\ m_{211} = 0,\ m_{1111} = 0,\\ &m_3 = 0,\ m_{21} = 0,\ m_{111} = 0.\end{aligned}$$

3. A completely controlled distribution, where each $x_i, i = 1, \dots, p$ is independently and identically distributed as $P(x_i = 1) = P(x_i = -1) = 1/2$.

$$\begin{aligned}&m_4 = 1,\ m_{31} = 0,\ m_{22} = 1,\ m_{211} = 0,\ m_{1111} = 0,\\ &m_3 = 0,\ m_{21} = 0,\ m_{111} = 0.\end{aligned}$$

4. Pareto distributions, where each $x_i, i = 1, \dots, p$ is independently and identically distributed as $P(b)$, Pareto distributions with Pareto index b. Its p.d.f. is given by

$$h(x) = \begin{cases} \prod_{i=1}^{p} b x_i^{-(b+1)} & \text{if } x_i > 1 \text{ for } i = 1, \dots, p, \\ 0 & \text{otherwise.} \end{cases}$$

After the normalization (2), we have

$$\begin{aligned} m_3 &= \text{Skewness of P(b)} = \frac{2(b+1)}{b-3}\sqrt{\frac{b-2}{b}}, \quad b>3,\\ m_4 &= \text{Kurtosis of P(b)} = \frac{6\left(b^3+b^2-6b-2\right)}{b(b-3)(b-4)}+3 \quad b>4. \end{aligned}$$

We are interested in the effect of the strong skewness and heavy tail of Pareto distribution. Here we put b as 4.2. Consequently

$$\begin{aligned} m_4 &= \frac{8129}{21},\ m_{31}=0,\ m_{22}=1,\ m_{211}=0,\ m_{1111}=0,\\ m_3 &= \frac{26}{63}\sqrt{231},\ m_{21}=0,\ m_{111}=0. \end{aligned}$$

In the following subsections, we will state the approximated $\overset{\alpha}{ED}$ for each error distribution. For more specific forms under a fixed α or given moments of x, see Section 3 of [8]. The derivation of $\eta[i,j,k,l]$ of (20) is also stated there.

3.1 Normal Error Term Distribution

$$\begin{aligned} &\overset{\alpha}{ED}\\ &= \frac{p+2}{2n}\\ &\quad + \frac{1}{96n^2}\Big(3p(-27-8\alpha+3\alpha^2)m_4 + 3p(p-1)(-27-8\alpha+3\alpha^2)m_{22}\\ &\qquad\qquad 4\alpha^2(12p+21)+4\alpha(6p^2+36p+50)\\ &\qquad\qquad +4(12p^2+60p+75)\Big)\\ &\quad + o(n^{-2}) \end{aligned} \tag{83}$$

The n^{-2} order term has the following properties;

1. The maximum dimension of p is two, hence $\overset{\alpha}{ED}$ is asymptotically determined by the $p-n$ ratio.

2. Other moments than m_4 and m_{22} do not appear.

3. The coefficients of m_4 and m_{22} are non-positive when $3\alpha^2-8\alpha-27\leq 0$, that is,

$$-1.95\cdots \leq \alpha \leq 4.62\cdots. \tag{84}$$

For α within this interval, the larger m_{22} or m_4 gets, the less $\overset{\alpha}{ED}$ becomes. The divergences often used in statistical literature are all included in this interval: K-L divergence ($\alpha=-1$), K-L dual divergence ($\alpha=1$), Hellinger Divergence ($\alpha=0$), χ^2 divergence ($\alpha=3$).

Figure 3: $\overset{-1}{ED}$ when $\epsilon \sim N(0,1)$

Figure 4: $\overset{-6}{ED}$ when $\epsilon \sim N(0,1)$

We made a numerical comparison to see the effect of the joint moments of x. We set $p = 10$ and $n = 12k$, which means $p-n$ ratio equals $1/k$ since the number of the parameters of the regression model (1) equals 12 when $p = 10$. Figure 3 is the graph of the approximated $\overset{-1}{ED}$'s corresponding to each distribution of x above-mentioned as k varies from 5 to 100. (The graph for the controlled distribution is always quite similar to that for the normal distribution, hence for the clarity of the figures we will omit it in every figure hereafter.) We put as the benchmark the approximated $\overset{-1}{ED}$ of the binomial model $B(k, 0.5)$, that is, the k-times normal coin toss model.

We notice that heavy tail property of Pareto distribution $P(4.2)$ or t-distribution $t(4.2)$

Table 1: I.D.E. & R.S.S. for $N(0,1)$ error distribution

	I.D.E.	R.S.S.
$x \sim N_{10}(0, I_{10})$	*	111(10)
$x \sim t_{10}(0, I_{10}, 4.2)$	*	322(40)
x is controlled	*	112(10)
x is $i.i.d.P(4.2)$	*	741(110)

decreases difficulty in estimating the parameter, especially the large m_4 value 8129/21 of $P(4.2)$ makes the estimation easier. On the contrary, if m_4 and m_{22} are as small as those of $N_{10}(0, I_{10})$ (or the controlled distribution), then the difficulty of estimation is close to the normal coin toss.

Here we refer to the question how large sample size is required for the good approximation of $\overset{\alpha}{ED}$ by the expansion up to the n^{-2} order term. It is very difficult to give a general answer to this question, but at least for a specific model, obviously we should not use the approximation unless it is positive or decreasing with respect to n. For example, in Figure 3, we see that the approximation for t_{10} should be used for $k > 10$, namely $n > 120$.

We observed that the effects of m_4 and m_{22} depends on α. If α is outside the interval (84), the large value of m_{44} or m_{22} enhances the difficulty of the estimation. For example, if $\alpha = -6$, the order of various distributions of x is completely reversed to that for the case $\alpha = -1$ as we can see from Figure 4.

Now we consider I.D.E. and R.S.S. introduced in Section 2. We take Kullback-Leibler divergence ($\alpha = -1$) as an example. Let p be 10. When x is distributed as $N_p(0, I_p)$, we have

$$\overset{-1}{ED}(n) = \frac{6}{n} - \frac{217}{12n^2}.$$

"Indicator of the difficulty of estimation" is given as the solution of m for the equation

$$\frac{1}{2k} + \frac{1}{12k^2}(M-1) = \frac{1}{2k} - \frac{217}{12 \times (12k)^2}$$

(See (75) for the left-hand side.) Actually this quadratic equation of m does not have the real roots. The left-hand side is always larger than the right-hand side. This means the estimation of the regression model is easier than the coin toss problem under the same $p-n$ ratio.

Sample size determination is solving the next equation;

$$\frac{1}{20} + \frac{1}{400} = \frac{6}{n} - \frac{217}{12n^2}.$$

where $n = 111$ is the rounded solution. For the other distributions (joint moments) of x, we can similarly calculate I.D.E. and R.S.S..The result is given in Table 1. "*" indicates that the equation has no solutions. The number in the parenthesis in R.S.S. shows the sample size of the binomial model in the left-hand side of (77) (see the last paragraph of Section 2.) With the sample size given by R.S.S., the $p-n$ ratio of the regression model equals 12/R.S.S., while that of the coin toss model is equal to the reciprocal of the number in the parenthesis. Hence R.S.S divided by the number in the parenthesis could be another indicator. It is smaller for $t_{10}(0, I_{10}, 4.2)$ or $P(4.2)$ than that for $N_{10}(0, I_{10})$ or the controlled distribution. The large joint moments m_4, m_{22} for $t_{10}(0, I_{10}, 4.2)$ or $P(4.2)$ make estimation easier. We can guess that the large oscillation of x is helpful to estimate the values of β. Nevertheless of these differences, in general, the estimation for the regression model under the normal error distribution is not so troublesome, since 10 times as large sample size as the dimension of the parameter guarantees relatively easy estimation.

3.2 t Error Term Distribution

$$\begin{aligned}
&\overset{\alpha}{ED} \\
&= \frac{p+2}{2n} \\
&\quad + \frac{1}{384n^2}\Big(6p(-45-17\alpha+3\alpha^2)m_4 + 6p(p-1)(-45-17\alpha+3\alpha^2)m_{22} \\
&\qquad\qquad + \alpha^2(78+60p) + \alpha(154+144p+30p^2) \\
&\qquad\qquad + 861+888p+195p^2\Big) \\
&\quad + o(n^{-2})
\end{aligned} \tag{85}$$

The n^{-2} order term has similar properties as in the case of $N(0,1)$.

1. The dimension of p is two, hence $\overset{\alpha}{ED}$ is asymptotically determined by the $p-n$ ratio.
2. Other moments than m_4 and m_{22} do not appear.
3. The coefficients of m_4 and m_{22} are non-positive when $3\alpha^2 - 17\alpha - 45 \leq 0$, that is,

$$-1.97\cdots \leq \alpha \leq 7.63\cdots . \tag{86}$$

   For α within this interval, the larger m_{22} or m_4 gets, the less $\overset{\alpha}{ED}$ becomes. The divergences often used in statistical literature are all included in this interval.

We noticed that if the error term distribution is the standard normal distribution (see (83)) or $t(3)$ distribution (see (85)), only m_4 and m_{22} among the moments (78) and (79) appear in the asymptotic expansion of $\overset{\alpha}{ED}$ up to the n^{-2} order. On this phenomena, we have the following general result.

Proposition 1. *If the error term distribution is quadratic, namely $f(\epsilon) = g(\epsilon^2)$ for some function $g(\cdot)$, then the asymptotic expansion of $\overset{\alpha}{ED}$ up to the order n^{-2} includes only m_4 and m_{22} among the third and forth order joint moments of x.*

<*Proof*> From (11), we notice that the third or forth order moments of x in the expansion of $\overset{\alpha}{ED}$ up to the order n^{-2} are generated from the terms $m[i,j,k,l]$ and $m[i,j,k]m[s,t,u]$.

The forth order moments arise from $m[i,j,k,l](1 \leq i,j,k,l \leq p)$ in $L11$ to $L15$. Since $m[i,j,k,l]$ is multiplied with $g^{ij}g^{kl}$ as in (11), and g^{ij} vanishes unless $i=j$, the possible moments coming from $E[x_i x_j x_k x_l]$ are only m_4 and m_{22}.

On the other hand, $m[i,j,k]m[s,t,u]$ come from either term of $L21, \ldots, L26$. We notice that if the third moments are generated from these terms, they are always multiplied with $\eta[0,1,1,0]$ or $\eta[0,0,3,0]$. (See (30), (31).) If $f(\epsilon) = g(\epsilon^2)$, then we have

$$\tfrac{d}{dy}\log f(y) = 2y\tfrac{g'(y^2)}{g(y^2)},$$

Figure 5: $\overset{-1}{ED}$ when $\epsilon \sim t(3)$

$$\tfrac{d^2}{dy^2} \log f(y) = \tfrac{2}{g^2(y^2)} \Big(g'(y^2) + 2y^2 g''(y^2) g(y^2) + 2y^2 (g'(y^2))^2 \Big).$$

Therefore $\eta[0,1,1,0]$ and $\eta[0,0,3,0]$ vanishes. *Q.E.D.*

We made a numerical comparison under the condition $p = 10$ and $n = 12k$, which means $p - n$ ratio equals $1/k$. Figure 5 is the graph of the approximated $\overset{-1}{ED}$'s corresponding to each distribution above-mentioned except for the controlled distribution as k varies from 5 to 100. We put as the benchmark the approximated $\overset{-1}{ED}$ of the binomial model $B(k, 0.5)$.

Just like the case of $N(0,1)$, heavy tail property of Pareto distribution $P(4.2)$ or t-distribution $t(4.2)$ eases difficulty in estimating the parameter. On the contrary, if m_4 and m_{22} are as small as those of $N(0,1)$ (or the controlled distribution), then the difficulty of estimation is close to the normal coin toss.

It was also observed that the effects of m_4 and m_{22} depends on α. If α is outside the interval (86), the large value of m_{44} or m_{22} enhances the difficulty of the estimation. For example, see Figure 6 for $\alpha = -6$, where the $\overset{-6}{ED}$ for the t or Pareto distribution is larger than that of the normal distribution.

We considered I.D.E. and R.S.S. w.r.t. Kullback-Leibler divergence under the condition $p = 10$ for each distribution of x. Table 2 shows the result.The same comments hold as in the case of $N(0,1)$. The large value of m_4 or m_{22} of the t-distribution or Pareto distribution makes the estimation easier compared to the normal distribution or the controlled distribution. Generally speaking, irrespective of the above difference, the estimation of the regression model under t-distribution error is not so hard. With 10 times as large sample size as the parameter dimension, we can estimate the parameter without much trouble.

Figure 6: $\overset{-6}{ED}$ when $\epsilon \sim t(3)$

Table 2: I.D.E. & R.S.S. for $t(3)$ error distribution

	I.D.E.	R.S.S.
$x \sim N_{10}(0, I_{10})$	*	117(10)
$x \sim t_{10}(0, I_{10}, 4.2)$	*	246(30)
x is controlled	*	118(10)
x is $i.i.d.P(4.2)$	*	689(90)

3.3 Skew-Normal Error Term Distribution

Since $\eta[i,j,k,l]$ for the skew-normal distribution can not be analytically gained, the following $\overset{\alpha}{ED}$ is calculated from the numerically gained $\eta[i,j,k,l]$'s.

$$\begin{aligned}
&\overset{\alpha}{ED} \\
&= \frac{p+2}{2n} \\
&\quad + \frac{1}{n^2}\Big(0.175p(-8.570 - 2.451\alpha + \alpha^2)m_4 \\
&\qquad\quad + 0.175p(p-1)(-8.570 - 2.451\alpha + \alpha^2)m_{22} \\
&\qquad\quad + 0.217p(-0.302 + \alpha)m_3^2 \\
&\qquad\quad + p(0.065p^2 + 0.130\alpha p^2 - 0.522p + 0.457 - 0.130\alpha)m_{21}^2 \\
&\qquad\quad + 0.087p(-1.504p^2 + \alpha p^2 + 4.513p - 3\alpha p - 3.001 + 2\alpha)m_{111}^2 \\
&\qquad\quad + 0.260p(p-1)(0.500 + \alpha)m_3 m_{21} \\
&\qquad\quad + (0.988 + 0.689p)\alpha^2 + (2.352 + 2.074p + 0.385p^2)\alpha \\
&\qquad\quad + 3.385 + 2.823p + 0.583p^2\Big) \\
&\quad + o(n^{-2}) \qquad (87)
\end{aligned}$$

Note that the numbers above are rounded off to three decimal place. We observe the following points for n^{-2} order term.

Figure 7: $\overset{-1}{ED}$ when $\epsilon \sim SN(3)$

Table 3: I.D.E. & R.S.S. for $SN(3)$ error distribution

	I.D.E.	R.S.S.
$x \sim N_{10}(0, I_{10})$	*	101(10)
$x \sim t_{10}(0, I_{10}, 4.2)$	*	536(70)
x is controlled	*	105(10)
x is $i.i.d.P(4.2)$	*	1499(210)

1. The dimension of p is three, hence if p increases with a constant $p - n$ ratio, n^{-2} order term could diverge for some given α and the moments of x. Then it is not enough to increase the sample size proportionally to the number of the explanatory variables in order to keep $\overset{\alpha}{ED}$ at a certain level.

2. m_3, m_{21} and m_{111} appear in the expansion that do not appear in the case of $N_p(0, I_p)$ or $t_p(0, I_p, \nu)$. The effect of these moments are rather complicated and depends on α and p, For example, when p is large enough, the larger absolute value of m_{21} decreases the approximated $\overset{\alpha}{ED}$ for $\alpha = -1$, but vice versa for $\alpha = 1, 0, 2$.

3. The larger m_4 and m_{22} decreases $\overset{\alpha}{ED}$ if $\alpha^2 - 2.451\alpha - 8.570 < 0$, namely
$$-1.95 \cdots < \alpha < 4.40 \cdots .$$
Again α's such as $-1, 0, 1, 3$ are all included in this interval.

We made a numerical comparison under the condition $p = 10$ and $n = 12k$, which means $p - n$ ratio equals $1/k$. Figure 7 is the graph of the approximated $\overset{-1}{ED}$'s corresponding to each distribution above-mentioned except for the controlled distribution as k varies from 5 to 100. We put as the benchmark the approximated $\overset{-1}{ED}$ of the binomial model $B(k, 0.5)$. The graph of the approximated $\overset{-1}{ED}$ for the case where x has Pareto distribution is still decreasing when k is around 100, hence the approximation is only

Figure 8: $\overset{-6}{ED}$ when $\epsilon \sim SN(3)$

feasible when $k > 100$. We observe again that large values of m_4 and m_{22} of Pareto distribution $P(4.2)$ or t-distribution $t(4.2)$ lead to easier estimation than the case of normal distribution $N(0,1)$ when $\alpha = -1$. However, just like the case when the error term has a normal distribution or t-distribution, the order of difficulty in the estimation is completely reversed with another α. For example, see Figure 8 for the case when $\alpha = -6$.

We considered I.D.E. and R.S.S. w.r.t. Kullback-Leibler divergence under the condition $p = 10$ for each distribution of x. Table 3 shows the result. I.D.E. tells us that with any case of the moments of x, the regression model is easier to be estimated than the binomial model with the same $p-n$ ratio. If we divide R.S.S. with the number in the parenthesis, it is always less than 12. This means the $p-n$ ratio is always larger than that of the binomial model which has the same level of estimation difficulty as the regression model. Especially when the distribution (moments) of x is given as t-distribution or Pareto distribution, it makes the estimation easier.

3.4 Comparison between different error distributions

In this subsection, under a fixed distribution (moments) of x, we compared the approximated $\overset{\alpha}{ED}$'s for the three error distributions: the standard normal distribution (say $\overset{\alpha}{ED}_n$), the t-distribution with the d.f. of 3 (say $\overset{\alpha}{ED}_t$) and the skew-normal distribution with the shape parameter of 3 (say $\overset{\alpha}{ED}_s$). All comparisons are made under the condition $p = 10, n = 12k$.

The order of the approximated $\overset{\alpha}{ED}$ among the three error distributions depends on α. We pick up two values of α, $\alpha = -1$ and $\alpha = -6$ as contrasting cases and summarized the results in Table 4. We notice that the order is completely reversed between $\alpha = -1$ and $\alpha = -6$. Under the fixed α, $\overset{\alpha}{ED}_n$, $\overset{\alpha}{ED}_t$, $\overset{\alpha}{ED}_s$ keep the same order irrespective of the distribution of x.

We also present the graphs of $\overset{-1}{ED}_n$, $\overset{-1}{ED}_t$, $\overset{-1}{ED}_s$ for each fixed distribution (moments) of x with the reference to that of the normal coin toss model $B(k, 0.5)$. We notice that

Table 4: Comparison between different error distributions

	$\alpha = -1$	$\alpha = -6$
$x \sim N_{10}(0, I_{10})$	$\overset{-1}{ED}_t > \overset{-1}{ED}_n > \overset{-1}{ED}_s$	$\overset{-6}{ED}_s > \overset{-6}{ED}_n > \overset{-6}{ED}_t$
$x \sim t_{10}(0, I_{10}, 4.2)$	$\overset{-1}{ED}_t > \overset{-1}{ED}_n > \overset{-1}{ED}_s$	$\overset{-6}{ED}_s > \overset{-6}{ED}_n > \overset{-6}{ED}_t$
x is controlled	$\overset{-1}{ED}_t > \overset{-1}{ED}_n > \overset{-1}{ED}_s$	$\overset{-6}{ED}_s > \overset{-6}{ED}_n > \overset{-6}{ED}_t$
x is $i.i.d.P(4.2)$	$\overset{-1}{ED}_t > \overset{-1}{ED}_n > \overset{-1}{ED}_s$	$\overset{-6}{ED}_s > \overset{-6}{ED}_n > \overset{-6}{ED}_t$

there is only little difference among the three error distributions and the normal coin toss model for the Kullback-Leibler divergence, especially when the distribution of x is $N_{10}(0, I_{10})$ or controlled.

As for I.D.E. and R.S.S., we can make the comparison between different error term distributions if we look through Tables 1, 2 and 3 with a fixed distribution of x. I.D.E. again indicates that the regression model can be more easily estimated than the coin toss model with any of the three error distributions.Though R.S.S. shows the sample size required do not differ so much among the error term distributions, if pressed, t(3) requires a bit larger size of samples. If we divide R.S.S with the number in the parenthesis, we notice that $t(3)$ is always larger than the other distributions.

4 Real Data –non-homogeneous explanatory variables–

In this section we deal with two real datasets. As well as examples of non-homogeneous explanatory variables, these datasets serves as concrete cases to which the general results in the previous sections can be applied. We calculate the sample moments of the explanatory variables of those datasets and use them as examples of the following moments of x (These datasets also include the dependent variables, but we do not use them here.)

$$m[i,j,k] = E[x_i x_j x_k] \qquad m[i,j,k,l] = E[x_i x_j x_k x_l] \qquad 1 \le i,j,k,l \le p. \tag{88}$$

First in order to standardize x as in (2), we transform x into its principal component scores. Then we calculate the moments of the transformed x,

$$n^{-1}\sum_{t=1}^{n} x_{ti}x_{tj}x_{tk} \qquad n^{-1}\sum_{t=1}^{n} x_{ti}x_{tj}x_{tk}x_{tl} \qquad 1 \le i,j,k,l \le 11,$$

and use them instead of (88) for the calculation of the aggregated sample moments

$$M_{2a} \triangleq \sum_{i,j,k \in \mathcal{I}} m^2[i,j,k] \tag{89}$$

Figure 9: $\overset{-1}{ED}$ when $x \sim N_{10}(0, I_{10})$

Figure 10: $\overset{-1}{ED}$ when $x \sim t_{10}(0, I_{10}, 4.2)$

Figure 11: $\overset{-1}{ED}$ when x is controlled

Figure 12: $\overset{-1}{ED}$ when x is i.i.d. as Pareto(4.2)

$$M_{2b} \triangleq \sum_{i,j,k\in\mathcal{I}} m[i,i,k]m[j,j,k] \tag{90}$$

$$M_1 \triangleq \sum_{i,k\in\mathcal{I}} m[i,i,k,k]. \tag{91}$$

Actually $\overset{\alpha}{ED}$ is affected by the moments of x only through these aggregated moments. (See the last part of Appendix A of [8].)

Since the results for those datasets are quite similar among different α's ($\alpha = -1, 0, 1, -6, 6$), we focus ourselves on the case $\alpha = -1$.

– Example 1: Wine Quality –
This is the famous dataset on wine quality used in Cortez et.al. [4]. The data file is available at U.C.I. Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets /Wine+Quality). We used the white wine dataset. The dataset is as follows;
y (dependent variable) $= (y_t)_{1\le t\le n}$: the quality score of the wine form 0 to 10.
x (explanatory variables) $= (x_{ti})_{1\le t\le n, 1\le i\le 11}$: $n \times 11$ real value data on the quantity of the chemical substances in the wines . Each column is the data for the corresponding explanatory variable. x_1: "fixed acidity", x_2: "volatile acidity", ... , x_{11}: "alcohol".
n (sample size): 4898

The values of (89) to (91) for this dataset are as follows;

$$M_{2a} = 0.000326899, \qquad M_{2b} = 0.000230836, \qquad M_1 = 0.116967. \tag{92}$$

M_{2a} or M_{2b} is the summation over 11^3 pieces of the squared 3-dimensional joint moments of x. Since their averages $M_{2a}/11^3$ and $M_{2b}/11^3$ are quite small compared to the unit variance of x_i, this indicates x are quite symmetric around the origin. $M_1/11^2$ is also much smaller than 1, hence the distribution of x has shorter tail than the normal distribution. $\overset{\alpha}{ED}$ is given as

$$\overset{\alpha}{ED} = \begin{cases} \dfrac{6.5}{n} + \dfrac{6.386\alpha^2 + 48.804\alpha + 91.026}{n^2} & \text{if } \epsilon \sim N(0,1), \\ \dfrac{6.5}{n} + \dfrac{1.927\alpha^2 + 13.948\alpha + 89.043}{n^2} & \text{if } \epsilon \sim t(3), \\ \dfrac{6.5}{n} + \dfrac{8.586\alpha^2 + 71.639\alpha + 104.856}{n^2} & \text{if } \epsilon \sim SN(3). \end{cases} \tag{93}$$

Figure 13 (k varies from 5 to 200) shows the graphs of $\overset{-1}{ED}$ for three error distributions under this moments of x and $n = 13k$. We also put the graph of $\overset{-1}{ED}$ of $B(0.5, k)$ as a reference. We see that $\overset{-1}{ED}$'s of the four cases are quite close to each other. There is almost no difference among the error distributions. Besides, the estimation difficulty of the regression model is similar to that of the normal coin toss with the same $p-n$ ratio irrespective of the error distributions.

Figure 13: $\overset{-1}{ED}$ for the wine data

Table 5: I.D.E. & R.S.S. for the wine data

	I.D.E.	R.S.S.
$N(0,1)$	0.66	130(10)
$t(3)$	0.81	135(10)
$SN(3)$	*	130(10)

I.D.E. and R.S.S. is stated in Table 5. I.D.E. shows that when the error distribution is $SN(3)$, then the estimation is the easiest and that when $\epsilon \sim t(3)$, the estimation gets slightly more difficult. As for R.S.S., we notice that there is very little difference among the three error distributions and that the estimation is relatively easy. Around 130 samples guarantee as easy estimation as the 10-times normal coin toss problem.

We can evaluate the actual sample size 4898 of this dataset by answering the following question; how large sample size n for the normal coin toss model $B(0.5, n)$ is required in order to attain the same level of easiness in the estimation as the regression model with the moments of x as in (92) and the sample size 4898 ? For example, if the error distribution is $N(0,1)$, then the answer is given as the solution of the equation

$$\frac{1}{2n} + \frac{1}{8n^2}((\alpha')^2 - 5\alpha' + 6) = \frac{6.5}{4898} + \frac{6.386\alpha^2 + 48.804\alpha + 91.026}{4898^2}, \tag{94}$$

where the left-hand side is (74) with $M = 4$.
The rounded solution when $\alpha = -1$ equals 376 or 377 for the three error distributions, which means the sample size 4898 for the regression is equivalent to the 376 (377) times normal coin toss in view of the estimation difficulty. We see that the estimation is fairly easy with this sample size.

– Example 2: Communities and Crime –
This data combines socio-economic data for each community within USA from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI U.C.R.. You can download the data file from U.C.I. Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime).

The original data contains 124 explanatory variables from “population” to “PolicBudgPerPop”. We excluded the explanatory variables that contains missing data (denoted by "?" in the original dataset) . Besides we excluded the variable "numbUrban","PctRecImmig8" and "OwnOccMedVal" because the following correlations exceed 0.99: Corr(“population”, “numUrban”), Corr(“PctRecImmig5”,”PctRecImmig8”), Corr(“PctRecImmig8”,”PctRecImmig10”), Corr(“OwnOccLowQuart”,”OwnOccMedVal”). After this process, the dataset is as follows;
y (dependent variable) $= (y_t)_{1\le t\le n}$: The candidates of y are 18 attributes from “murders” (the number of the murders committed in the community) to “nonViolPerPop”(the number per capita of non-violent crimes committed in the community). They are the numbers of the committed crimes categorized in various ways.
x (explanatory variables) $= (x_{ti})_{1\le t\le n, 1\le i\le 99}$: $n\times 99$ real value data on the socio-economic character of the community. x_1: “population”, x_2: “household”(mean people per household) ,..., x_{99}: “LemasPctOfficDrugUn”(the percent of officers assigned to drug units).
n (sample size): 2215
We used principle component sores as the standardized x. The aggregated sample moments are given by

$$M_{2a} = 1708.97, \qquad M_{2b} = 1749.28, \qquad M_1 = 2604.5.$$

$M_{2a}/99^3$, $M_{2b}/99^3$ and $M_1/99^2$ are much smaller than unit. Like the wine data, the distribution of x is symmetric and short-tailed. Using these values we calculated $\overset{\alpha}{ED}$, which is given by

$$\overset{\alpha}{ED} = \begin{cases} \dfrac{50.5}{n} + \dfrac{294.547\alpha^2 + 1949.71\alpha + 2953.58}{n^2} & \text{if } \epsilon \sim N(0,1), \\ \dfrac{50.5}{n} + \dfrac{137.758\alpha^2 + 111.409\alpha + 3376.96}{n^2} & \text{if } \epsilon \sim t(3), \\ \dfrac{50.5}{n} + \dfrac{526.088\alpha^2 + 3232.22\alpha + 1976.26}{n^2} & \text{if } \epsilon \sim SN(3). \end{cases} \tag{95}$$

Figure 14 (k varies from 5 to 200) shows the graphs of $\overset{-1}{ED}$ for the three error distributions under these moments of x and $n = 101k$. We also put the graph of $\overset{-1}{ED}$ of $B(0.5, k)$ as a reference. The comment for Example 1 holds for this data. We see that $\overset{-1}{ED}$'s for the three error distributions are almost same. Compared to the normal coin toss with the same $p - n$ ratio, the regression model is on the same level for the estimation difficulty.

You can see I.D.E. and R.S.S. in Table 6. We notice that it is slightly harder to estimate the parameters when $\epsilon \sim t(3)$, but, generally speaking, for the regression model with these moments of x, estimating the parameters is not a hard task if we have around 1000 samples. We evaluate the sample size 2215 in a similar way to (94). If the error distribution is $N(0, 10)$, then solving the equation

$$\frac{1}{2n} + \frac{1}{8n^2}((\alpha')^2 - 5\alpha' + 6) = \frac{50.5}{2215} + \frac{294.547\alpha^2 + 1949.71\alpha + 2953.58}{2215^2} \tag{96}$$

Figure 14: $\overset{-1}{ED}$ for the crime data

Table 6: I.D.E. & R.S.S. for the crime data

	I.D.E.	R.S.S.
$N(0,1)$	*	987(10)
$t(3)$	0.72	1025(10)
$SN(3)$	*	947(10)

gives us an evaluation of the actual sample size. When $\alpha = -1$, the rounded solution is 22 or 23 for the three error distributions. Though this number is much smaller than 376(377) in Example 1, the estimation is still not a hard task since 22-times normal coin toss gives us plenty of information.

5 Summary and Discussion

- $\overset{\alpha}{ED}$ is constant for the parameter β, σ.
- The main term (n^{-1} term) of the asymptotic expansion of $\overset{\alpha}{ED}$ is $(p+2)/n$, that is, $p - n$ ratio.
- For the second term (n^{-2} term) of the expansion, we observe the following points.
  1. The maximum dimension of p depends on the error term distribution. It can be more than two as in the case $\epsilon \sim SN(3)$, where it is not enough to increase the sample size proportionally to p for reliable estimation (so called "the curse of dimension").
  2. The joint moments that appear in the term is maximally of the forth order. What moments appear is different among the error term distributions. If it is a quadratic distribution (e.g. $N(0,1)$, $t(\nu)$), then the moments m_4 and m_{22} only appear.
  3. The effect of m_4 and m_{22} depends on α. When $\alpha = -1, 0, 1, 3$, the larger m_4 and m_{22} decreases the difficulty of the estimation. In a geometrical view,

   there is no preference among α's. Each α gives its own geometrical structure to Riemannian manifold formed by the parametric distribution model (see e.g. Amari and Nagaoka [3]). However there might be values for α that is "natural" in a statistical sense or "appropriate" for a purpose of the statistical analysis.

   4. The effect of the error term distributions also depends on α. For example, the order of the estimation difficulty among the three error distributions is quite different between $\alpha = -1$ and $\alpha = -6$.
   5. The difference between the three error term distributions we investigated is relatively small if we use Kullback-Leibler divergence.This might be due to the assumption that we know the error term distributions, hence are able to use m.l.e. In most applications, the actual error term distribution is unknown, and m.l.e. is not applicable. It is of much interest what would happen to the risk of the predictive distribution, if we use another estimator such as the least squares estimator.

- We proposed measuring the (asymptotic) difficulty of estimation by the approximated $\overset{\alpha}{ED}$ and tried to give a suggestion on the sample size. It is a method comparing the approximated $\overset{\alpha}{ED}$ of the regression model to that of a binomial model $B(n, m)$. I.D.E. tells the difficulty of estimation by the value of m of $B(k, m)$, which has the same $p - n$ ratio as the regression model (1) of the sample size $(p + 2)k$. R.S.S. gives the sample size n for the regression model which leads to the same difficulty of estimation as $B(10, 0.5)$ (If it is needed, a more large value than 10 will be used for the binomial model).
   1. Though there exist small difference between the error term distributions and the moments of x, in most cases we investigated, the regression model is easier to be estimated than the normal coin toss $B(k, 0.5)$ under the same $p-n$ ratio $1/k$.
   2. The sample size $n = 10(p + 2)$ guarantees the good performance of the estimation at the same level as the 10-times normal coin toss irrespective of the error term distributions and the moments of x which we investigated in this paper.

Acknowledgment

This research was supported by JSPS KAKENHI Grant Number 25380265.

References

[1] S. Amari. Alpha divergence is unique, belonging to both classes of f-divergence and Bregman divergence *IEEE Trans. Information Theory*, 55:4925-4931, 2009.

[2] S. Amari and A. Chichocki. Information geometry of divergence function. *Bulletin of the Polish Academy of Sciences : Technical Sciences*, 58:183-195, 2010.

[3] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Translations of Mathematical Monographs 191. American Mathematical Society, 2000.

[4] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547-553, 2009.

[5] S. Eguchi. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Mathematical Journal*, 15: 341-391, 1985.

[6] Wolfram Research, Inc. *Mathematica 10.4*, 2016.

[7] Y. Sheena. *Asymptotic expansion of the risk of maximum likelihood estimator with respect to α-divergence as a measure of the difficulty of specifying a parametric model*, arXiv:1510.08226, 2016.

[8] Y. Sheena. *Asymptotic Expansion of Risk for a Regression Model with respect to α-Divergence with an Application to the Sample Size Problem–Complete Version–*, arXiv:1703.10107, 2017.

[9] I. Vajda. *Theory of Statistical Inference and Information*, Kluwer Academic Publishers, 1989.