이 학 박 사 학 위 논 문

Regression with Partially Observed Ranks on a Covariate: Distribution-Guided Scores for Ranks

부분적으로 관측된 순위 공변량을 이용한 회귀분석: 순위에 대한 분포-유도 스코어 함수

2016년 8월

서울대학교 대학원

통계학과

김 윤 응

Regression with Partially Observed Ranks

on a Covariate:

Distribution-Guided Scores for Ranks

부분적으로 관측된 순위 공변량을 이용한 회귀분석:

순위에 대한 분포-유도 스코어 함수

지도교수 임 요 한

이 논문을 이학박사 학위논문으로 제출함

2015년 10월

서울대학교 대학원

통계학과

김 윤 응

김윤응의 이학박사 학위논문을 인준함

2016년 6월

위 원 장	조 신 섭	(인)
부위원장	임 요 한	(인)
위 원	김 용 대	(인)
위 원	오 희 석	(인)
위 원	최 수 정	(인)

Regression with Partially Observed Ranks on a Covariate: Distribution-Guided Scores for Ranks

by

Yun-Eung Kim

A Thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
August, 2016

Abstract

Yun-Eung Kim
The Department of Statistics
The Graduate School
Seoul National University

This work is motivated by a hand-collected data set from one of the largest Internet portals in Korea. This data set records the top 30 most frequently discussed stocks on its on-line message board. The frequencies are considered to measure the attention paid by investors to individual stocks. The empirical goal of the data analysis is to investigate the effect of this attention on trading behavior. For this purpose, we regress the (next day) returns and the (partially) observed ranks of frequencies. In the regression, the ranks are transformed into scores, for which purpose the identity or linear scores are commonly used. In this thesis, we propose a new class of scores (a score function) that is based on the moments of order statistics of a random variable Z. The new scores are shown to be flexible in modeling the desired features (e.g., monotonicity or convexity) of the scores. In addition, if the true covariate X is drawn from a location-scale family and Z is its standardized distribution, then the least-squares estimator calculated using the proposed scores consistently estimates the true correlation between the response and the covariate and asymptotically approaches the normal distribution. We also propose a procedure for diagnosing a given score function and selecting one that is better suited to the data. We numerically demonstrate the advantage of using a correctly specified score function over that of the identity scores (or other misspecified scores) in estimating the correlation coefficient. Finally, we

apply our proposal to test the effects of investors' attention on their returns using the motivating data set.

keywords : *Concomitant variable; investor attention; linear regression model; moments of order statistics; partially observed ranks; scores of ranks.*

Student Number : 2009-20242

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is motivated by a hand-collected data set from `Daum.net`, the 2nd largest Internet portal in Korea. The `Daum.net` portal offers an online stock message board where investors can freely discuss specific stocks in which they might be interested. This portal also reports a ranked list of the top 30 stocks that are most frequently discussed by users on a daily basis. The data set was collected by the authors during the 537 trading days from October 4th, 2010, to November 23rd, 2012. Along with the rank data, we also collected financial data regarding individual companies from FnGuide (`http://www.fnguide.com`). These additional data include stock-day trading volumes classified in terms of different types of investors, stock prices, stock returns, and so on.

The purpose of analyzing the collected data is to investigate the shifts in stock returns caused by variations in investor attention. In finance, researchers are often interested in determining the motivations that drive buy-

ing and selling decisions in stock markets. It is commonly assumed that investors efficiently process relevant information in a timely manner, but in reality, it is nearly impossible to be efficient because of information overload. In particular, individual investors are often less sophisticated than are institutional investors and have a limited ability to process all relevant information. For this reason, individual investors may pay attention only to a limited amount information, perhaps that which is relatively easy to access. The phenomenon of limited attention is a well-documented cognitive bias in the psychological literature (Kahneman, 1973; Camerer, 2003). This phenomenon affects the information-processing capacities of investors and thus affects asset prices on the financial market. To empirically prove the effect of investor attention on stock returns, we regress the observed stock returns with respect to the partially observed ranks.

Regression on a (partially observed) rank covariate has not previously been extensively studied in the literature. A procedure that is commonly used in practice to address rank covariates is to (i) regroup the ranks into only a few groups (if the number of ranks is high) and (ii) treat the regrouped ranks as an ordinal categorical variable. Ordered categorical variables frequently arise in various applications and have been studied extensively in the literature. Score-based analysis is commonly used for this purpose; see Hájek (1968), Hora and Conover (1984), Kimeldorf et al. (1992), Zheng (2008), Gertheiss (2014) and the references therein. Two-step procedure for addressing a rank covariate is equivalent to defining a score function for the ranks. However, as in the case of ordinal categorical variables, the score-based approach has difficulties in chosing the score function; different choices of scores

may lead to conflicting conclusions in the analysis (Graubard and Korn, 1987; Ivanova and Berger, 2001; Senn, 2007). The recommendation for selecting the score function according to the literature is (i) to choose meaningful scores for the ordinal categorical variable based on domain knowledge of the data or (ii) to use equally spaced scores if scientifically plausible scores are not available (Graubard and Korn, 1987).

In this thesis, we seek to provide an efficient tool for approach (i) described above, for the case in which some qualitative knowledge is available regarding the ranks or the ranking variable (the variable that is ranked). More specifically, we propose a new set of score functions and study their use in linear regression. The proposed score function is based on the moments of order statistics (MOS) of a random variable Z. This score function has several interesting properties, as listed below. First, in defining the score function, we can simply choose a sequence of numbers (the scores of the ranks) that exhibits the desired features. The chosen scores become a set of MOS of a random variable Z if they satisfy a certain set of conditions, which are reviewed in Chapter 3. Second, in the linear regression model, if the true covariate X is drawn from a location-scale family and Z is its standardized distribution, then the least-squares estimator that is calculated using the proposed scores consistently estimates the true correlation between the response and the covariate and asymptotically approaches the normal distribution. In addition, the residuals of the fitted regression allow us to diagnose the given score function and to provide a tool for selecting a score function that is better suited to the data.

The remainder of this thesis is organized as follows. We review the study

for score-based analysis and choice of score function in Chapter 2. We briefly summarize the use of score function and recommendations for selecting the score function in their study. In Chapter 3, we study the properties of the MOS and the proposed score function. In this Chapter, we review various moment problems, including the moment problem of the order statistics and the Hausdorff and Markov moment problems. We also demonstrate the asymptotic equivalence between the proposed score function and the quantile function; the quantile function may provide a better illustration of the qualitative features of the score function. In Chapter 4, we apply the score function to estimate the regression coefficient of the linear model or, more precisely, to estimate the correlation coefficient between the response and the scoring variable X. We prove that the least-squares estimator that is calculated using the proposed score function consistently estimates the correlation coefficient and is asymptotically normally distributed. In addition, we discuss the procedure for selecting an appropriate score function. In Chapter 5, we numerically demonstrate that using the correctly specified score function significantly reduces the mean square error on the estimation of the correlation coefficient. In Chapter 6, we analyze the motivating data set to investigate the existence of the attention effect. Finally, in Chapter 7, we briefly summarize the paper and discuss the application of the proposed scores to regression using other auxiliary covariates.

Chapter 2

Literature Review

Regression on a partially observed rank covariate has not previously been widely studied in the literature. In statistics, a common approach for dealing with rank covariates is to (i) regroup the ranks with a lot of categories into only a few groups and (ii) treat the regrouped ranks as an ordinal categorical variable. Unlike the study for regression with a partial rank covarite, the study for ordered categorical variables have been studied extensively in many literatures. In this thesis, we briefly review the method of score-based analysis presented by them and refer to choice of score function. From the review of score-based analysis, we study for the use of score function and their role in analysis. We also learn the precautions when we choose the optimal score function. Especially, Their recommendation for selecting the score function according to the literature is (i) to choose meaningful scores for the ordinal categorical variable based on domain knowledge of the data or (ii) to use equally spaced scores if scientifically plausible scores are not available.

2.1. Score-based Analysis

2.1.1. Simple linear rank statistics

Simple linear rank statistics provide a key to solving of general theory such as asymptotically most powerful tests for some problems. Under the alternative hypothesis, the distribution of a simple linear rank statistic is determined by the following three entities: (i) regression constants (ii) distribution functions of individual observations (iii) scores. The scores are usually assumed to be generated by a function ϕ, and the regularity conditions of scores are expressed in terms of smoothness and boundedness of ϕ. Among the authors who have studied for simple linear rank statistics, Hájek (1968) has improved the asymptotic normality of simple linear rank statistics. We briefly examine the use of score funtion in their study.

Let $X_1, \ldots, X_N$ be independent random variables with continuous distribution functions $F_1, \ldots, F_N$, and let $R_1, \ldots, R_N$ denote the corresponding ranks. The simple linear rank statistics is defined as

$$\mathcal{S} = \sum_{i=1}^{N} c_i a_N(R_i), \quad i = 1, \ldots, N,$$

where $R_i = \sum_{j=1}^{N} u(X_i - X_j)$ with $u(x) = \mathcal{I}(x \geq 0)$ and $c_1, \ldots, c_N$ are arbitrary regression constants. The score function $a_N(1), \ldots, a_N(N)$ are generated by a function $\phi(t)$, $0 < t < 1$, in the following two ways:

$$a_N(i) = \phi(i/(N+1)), \quad i = 1, \ldots, N,$$
$$a_N(i) = \mathrm{E}_\phi(U_N^{(i)}), \quad i = 1, \ldots, N,$$

where $U_N^{(i)}$ denotes the ith order statistic in a sample of size N from the $\mathbf{U}(0,1)$. The first scores given by above occur in statistics yielding locally most powerful rank tests and the second scores are distinguished by simplicity.

There are several studies for relaxing the regularity conditions of three entities: (i) regression constants (ii) distribution functions of individual observations (iii) scores. The main finding of Hájek (1968) are to extend the results of Chernoff and Savage (1958) and Govindarajulu et al. (1966) from two-sample case to the general regression case by relaxing the conditions on the scores-generating function.

2.1.2. Two-way ANOVA model

The general test for main effects in the two-way ANOVA assumes independence of the observations, constant variance, and normality. When these assumptions cannot be verified, there is an alternative procedure with less assumptions. Hora and Conover (1984) consider a procedure in which all observations are ranked simultaneously without regard to block membership or levels of treatments. In their procedure, they recommend to use rank transform testing procedure that has several advantages: (i) The distribution assumptions are less strict. (ii) The α-level for the rank transform testing procedure is more robust to distibutional assumptions than the classical procedure and has greater power when the distibutions are heavy-tailed. (iii) The test is simple to apply. Their aim is to provide the asymptotic theory for the fixed effects two-way ANOVA for scores based on the ranks of the data. We simply introduce their procedure with general setups.

Let X_{ijn} be independent random variables with continous distribution function F_{ij}. The hypothese of interest are

$$\mathrm{H}_0 : F_{ij} = F_i \qquad \text{for } j = 1, \ldots, J$$

and

$$\mathrm{H}_a : F_{ij} \neq F_i \qquad \text{for at least one } j = 1, \ldots, J$$

The rank of $X_{i'j'n'}$ among the set of random variables $\{X_{ijn}\}$ is denoted by $R_{i'j'n'}$. The original observations are replaced with scores generated by a function $\phi(t), 0 < t < 1$. The score of X_{ijn} is the random variable $a_M(R_{ijn})$, where $M = IJN$ and either

$$a_M(i) = \phi(i/(M+1)), \quad i = 1, \ldots, M,$$
$$a_M(i) = \mathrm{E}_\phi(U_M^{(i)}), \quad i = 1, \ldots, M.$$

$U_M^{(i)}$ is the ith order statistic in a random sample of size M taken on U$[0, 1]$ and ϕ has a bounded second derivative, $\int_0^1 [\phi(t)]^2 dt < \infty$. Then, the sum of scores for all X_{ijn} will be denoted by $S_{i.}, S_{.j}, S_{ij}$, repectively, and the sum of all scores will be denoted by S. The usual F statistic for testing main effects in a two-way ANOVA, then given by

$$F_N = \frac{\sum_{j=1}^{J} (S_{.j} - \bar{S})^2/(J-1)N}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N} [a_M(R_{ijn}) - N^{-1}S_{ij}]^2/J(N-1)},$$

where $\bar{S} = J^{-1}S$.

The main finding of their work is to show that F_N converges in law to χ^2_{J-1}

with the assumption in which the scores $a_M(R_{ijn})$ are generated from above two scores and ϕ has a bounded second derivative and is square-integrable.

2.2. Choice of Score Function

2.2.1. $2 \times K$ contingency table

The analysis of a $2 \times K$ contingency tables with ordered categories occurs frequently in the biomedical research literature. According to Graubard and Korn (1987), researchers have been advised not to apply a chi-square test for such data, but to use an appropriate analysis that incorporates the order of the columns. Proposed the analyses may be divided into those that require a priori assignment of quantitative scores for each column, and those that do not.

A major drawback with using tests that require preassigned fixed scores is the need to specify the values for the scores. This is one of the considerations that has led some authors to recommend against the use of tests that require scores in favor of the rank tests. Additionally, allowing the investigator to select the scores leaves open the potential abuse of choosing scores that will produce a desired result. However, even though analyses based on rank statistics are apparently objective, the investigator could easily choose from among many possible linear rank statistics to obtain a desired result.

The purpose of study in Graubard and Korn (1987) is to demonstrate that the rank statistics can be poor choices for testing independence when the column margin is far from uniformly distributed. This is because of the well-known correspondence between the rank tests and tests using scores

with midranks as the preassigned scores. Therefore, the notion that rank tests avoid the arbitrary choice of column scores is misleading. Their recommendations for testing independence in an ordered $2 \times K$ contingency table are as follows:

(i) If possible, develop reasonable column scores based on the substantive meaning of the column categoried, and use them in the analysis.

(ii) If no natural column scores are available, then consider using equally spaced column scores in the analysis.

(iii) Always examine the midranks as scores to make sure they are reasonable before using a rank test.

For the appropriate choice of scores they consider exact permutation tests that are conditional on both the row and column margins in order to focus attention on the importance of the choice of column scores. A general procedure for generating such a test based on a test statistic S is as follows:

(i) Construct every $2 \times K$ table with the same margins as the observed table;

(ii) Calculate the probability of observing each of the tables under the null hypothesis of independence;

(iii) Calculate S for each of the tables; and

(iv) Add up the probabilities of the tables for which S is greater than or equal to S based on the observed data to yield the one-sided P-value.

2.2.2. Drawbacks of integer scoring

Linear rank tests are widely used when testing for independence against stochastic order in a $2 \times K$ contingency table with two treatments and K ordered outcome levels. For this purpose, numerical scores are assigned to the outcome levels. When the choice of scores is not apparent, integer (equally-spaced) scores are often considered. Ivanova and Berger (2001) begin with an example of a 2×3 table with ordered categories and claim the test is generally conservative when equally-spaced scores are chosen. They also discuss the Wilcoxon rank-sum test, and other tests whose reliance on assigning scores is rarely made explicit. For this reason they propose a new test that is an improvement of the Smirnov test in the same sense that the linear rank test with slightly perturbed scores is an improvement of the linear rank test with integer scores. The notion of integer scores and perturbed scores can be comprehensible the following example. If the three categories are assumed to be equally-spaced, then the test with equally-spaced (integer) scores (1,2,3) or (0,0.5,1) is considered. In their situation, the pertubed scores is not to be equally-spaced,that is, (0,0.49,1).

Ivanova and Berger (2001) mention the only reason to choose integer scores is that they "look good". They argue that, though somewhat less attractive, slightly perturbed integer scores will lead to better results. They also argue that if a nonlinear rank test is to be used, then the Smirnov test with slightly perturbed scores might be considered, on the basis that is more powerful than the Smirnov test, but easier to compute than the uniformly improved Smirnov test. Eventually, they show that slightly perturbed scores often lead to a uniformly more powerful test in their work.

Chapter 3

Distribution-Guided Scores for Ranks

The score function that we propose is a set of the MOS of a random variable Z. Specifically, suppose that $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed (IID) copies of the random variable Z and that $Z_{(r:n)}$ is the corresponding rth-order statistic for $r = 1, 2, \ldots, n$. The score that we propose for rank r is $S_n(r) := \mathrm{E}\big(Z_{(r:n)}\big)$.

3.1. Relationship between score and quantile function

The proposed score is closely related to the quantile of the underlying distribution of Z. Let $F_Z(z)$ for $z \in \mathcal{R}$ and $Q_Z(q)$ for $q \in [0, 1]$ be the cumulative distribution function (CDF) and the quantile function (QF), respectively, of

Z. In the estimation of $F_Z(z)$ for $\{Z_i, i = 1, 2, \ldots, n\}$, the rth-order statistic $Z_{(r:n)}$ is the $(r/n) \times 100$-th percentile point of the empirical CDF, and thus, its expectation value is approximately equal to $Q_Z(r/n)$. More specifically, given $p_r = \frac{r}{n+1}$, $q_r = 1 - p_r$, and $Q_r = Q_Z(p_r)$, we can write

$$\alpha_{(r:n)} = Q_r + \frac{p_r q_r}{2(n+2)} Q_r^{(2)} + O\big(\frac{1}{n^2}\big),$$

where $Q_r^{(2)} = -f_Z'(Q_r)/\{f_Z(Q_r)\}^3$ and $f_Z(z)$ is the probability density function of Z, which is differentiable. We refer the reader to David (2003, Section 4.6) for the details of the relationship between the MOS and the quantiles.

Consideration of the QF will provide a better understanding of the qualitative features of the proposed score function. Suppose that we expect the score function $S_n(r)$ to be convex in the sense that $S_n(r+1) - S_n(r) \geq S_n(r) - S_n(r-1)$ for $r = 2, \ldots, n-1$. From the asymptotic equivalence between the MOS and the quantiles, it is known that the convexity of the scores $S_n(r)$ is approximately equal to that of the quantile function $Q_Z(p)$. Furthermore, the convexity of $Q_Z(p)$ implies the following equivalent statements: (i) $F(z)$ is concave in z, (ii) $f'(z) \leq 0$ or (iii) $\log f(z)$ is decreasing in z.

3.2. The moment problem of the order statistics

In defining the scores, we could simply choose a sequence of n numbers, for example, $a_1, a_2, \ldots, a_n$, with the desired properties and set $S_n(r) = a_r$. The

moment problem of the order statistics involves studying the conditions for the existence of a random variable Z (equivalently, a probability measure) that satisfies $a_r = \mathrm{E}\big(Z_{(r:n)}\big)$, $r = 1, \ldots, n$, which have been well characterized by Kadane (1971). One such condition is defined by Theorem 2 of Kadane (1971), which states the following: Given n and $\{a_r\}_{1 \le r \le n}$, there exists a non-negative random variable Z such that $a_r = \mathrm{E}\big(Z_{(r:n)}\big)$, $r = 1, \ldots, n$, if and only if

$$m_r := \big(a_{r+1} - a_r\big) \Big/ \left\{ a_1 \binom{n}{r} \right\}$$

is the rth moment ($1 \le r \le n-1$) of a probability distribution on $[0, \infty)$. This condition for $m_0 = 1$ and $n' = n - 1$ can be converted into that of the Stieltjes moment problem, which concerns the existence of a positive Borel measure ν that satisfies

$$m_k = \int_0^\infty x^k d\nu(x), \quad k = 0, 1, 2, \ldots, n' \tag{3.1}$$

The necessary and sufficient condition for (3.1) is characterized in terms of the Henkel matrices

$$H_{2k} := \{m_{i+j}\}_{0 \le i,j \le k}, \; H_{2k+1} := \{m_{i+j+1}\}_{0 \le i,j \le k}, \quad k = 0, 1, \ldots. \tag{3.2}$$

The condition is that

$$\begin{aligned} &\det\big(H_0\big) > 0, \; \det\big(H_1\big) > 0, \; \cdots, \; \det\big(H_k\big) > 0 \\ &\det\big(H_{k+1}\big) = 0, \; \det\big(H_{k+2}\big) = 0, \; \cdots, \; \det\big(H_{n'}\big) = 0. \end{aligned} \tag{3.3}$$

Detailed reviews of the moment problem have been provided by Huang (1989) and Diaconis and Freedman (2003).

3.3. Features of location-scale family assumption

In the linear regression model discussed in the next chapter, we assume that the ranking variable X is drawn from the location-scale family generated by Z. This assumption provides at least two attractive features in the analysis. The first is that the use of the standardized distribution Z (or its MOS) provides a normalization procedure for data sets that are drawn from different sources and are heterogeneously distributed. For example, in the rank data from `Daum.net`, the distributions of word frequencies vary from day to day, although their qualitative features, such as heavy-tailedness, do not. Here, using the standardized distribution Z, we can adequately explain this heterogeneity on a daily time scale. The other attractive feature is related to the estimation of the parameters of the linear regression. In the linear regression, this assumption allows us to consistently estimate the correlation coefficient between Y and X using only the ranks of X. Details are provided in the following chapter.

Chapter 4

Simple Linear Regression

In this chapter, we consider a simple regression model in which only partial ranks of a covariate are observed. Specifically, suppose that $\big\{\big(Y_i, X_i\big),\ i = 1, 2, \ldots, n\big\}$ is the complete set of observations, where Y_i is the variable of primary interest and X_i is the covariate related to Y_i. For example, in our rank data from `Daum.net`, for $i = 1, 2, \ldots, n$, Y_i is a relevant outcome such as earning rate or trading volume, X_i is the frequency of on-line discussions of the ith company, and R_i is the rank of X_i among $X_1, X_2, \ldots, X_n$. In this thesis, we consider the case in which the ranks R_i are partially observed in the sense that we observe only that $U_i = R_i \mathrm{I}\big(R_i \leq m\big) + m^{+}\mathrm{I}\big(R_i > m\big)$ rather than R_i, where m^{+} is an arbitrary constant that is greater than m. This observation can also be written, in brief, as

$$Y_{[1:n]}, Y_{[2:n]}, \ldots, Y_{[m:n]}, \big\{Y_{[r:n]}, r > m\big\},$$

where $Y_{[r:n]} = Y_i \mathrm{I}\big(R_i = r\big)$ for $r = 1, 2, \ldots, n$. We denote the above partial data by $\mathbf{Y}_{[m]}$ for notational simplicity. The objective of this chapter is to identify a good estimator of $\rho = \mathrm{corr}\big(Y, X\big)$ (or the regression coefficient between Y and X) and to test $\mathcal{H}_0 : \rho = 0$ versus $\mathcal{H}_1 : \rho \neq 0$ or $\rho > 0$ using the observed data $\mathbf{Y}_{[m]}$.

4.1. Least-Squares Estimator

To estimate ρ, we make certain assumptions regarding the distributions of X and Y. We assume that the linear model of the relationship between X_i and Y_i is

$$Y_i = \mu_Y + \rho\sigma_Y \frac{X_i - \mu_X}{\sigma_X} + \epsilon_i,$$

where the ϵ_is are IID values from a distribution of mean 0 and variance σ_ϵ^2. By ordering on the X_is, we have for $r = 1, \ldots, n$

$$Y_{[r:n]} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\big(X_{(r:n)} - \mu_X\big) + \epsilon_{[r:n]}, \tag{4.1}$$

where $\rho = \mathrm{corr}\big(Y, X\big)$ and

$$\begin{aligned} \mathrm{E}\big(Y_{[r:n]}\big) &= \mu_Y + \rho\sigma_Y\alpha_{(r:n)} \\ \mathrm{var}\big(Y_{[r:n]}\big) &= \sigma_Y^2\big(\rho^2\beta_{(rr:n)} + 1 - \rho^2\big) \\ \mathrm{cov}\big(Y_{[r:n]}, Y_{[s:n]}\big) &= \rho^2\sigma_Y^2\beta_{(rs:n)},\ r \neq s \end{aligned} \tag{4.2}$$

with

$$\alpha_{(r:n)} = \mathrm{E}\left\{\frac{X_{(r:n)} - \mu_X}{\sigma_X}\right\} \quad \text{and} \quad \beta_{(rs:n)} = \mathrm{Cov}\left(\frac{X_{(r:n)} - \mu_X}{\sigma_X}, \frac{X_{(s:n)} - \mu_X}{\sigma_X}\right)$$

for $r, s = 1, 2, \ldots, n$ (David and Galambos, 1974; David, 2003).

By motivating the identities (4.1) and (4.2) given above, we propose the least-squares estimator

$$\widehat{\rho}\big(s\big) \equiv \frac{1}{\widehat{\sigma}_Y} \cdot \frac{\sum_{r=1}^{[ns]} \alpha_{(r:n)} \big\{Y_{[r:n]} - \widehat{\mu}_Y\big\}}{\sum_{r=1}^{[ns]} \alpha_{(r:n)}^2} \tag{4.3}$$

as an estimator of ρ with $s = m/n$, where, $\widehat{\mu}_Y = \sum_{i=1}^{n} Y_i/n$ and $\widehat{\sigma}_Y^2 = \sum_{i=1}^{n}(Y_i - \widehat{\mu}_Y)^2/n$ are the empirical estimators of the mean and variance, respectively, of Y.

We claim that if X is drawn from a location-scale family generated by Z, then the least-squares estimator that is calculated based on the partial observations $\mathbf{Y}_{[m]}$ with s, which is defined in (4.3), is consistent and asymptotically normally distributed with an appropriate scale, as shown in Theorem 4.1. Let the empirical variance estimators are

$$\begin{aligned}
\Psi_n^{\mathrm{I}}(s) &:= \frac{1}{n} \sum_{r=1}^{[ns]} \alpha_{(r:n)}^2 \sigma_{(r:n)}^2 \\
\Psi_n^{\mathrm{II}}(s) &:= \frac{1}{n} \sum_{r1=1}^{[ns]} \sum_{r2=1}^{[ns]} \alpha_{(r1:n)} \alpha_{(r2:n)} \beta_{(r1,r2:n)}^2 \quad \text{and} \\
\Phi_n(s) &:= \frac{1}{n} \sum_{r=1}^{[ns]} \alpha_{(r:n)}^2,
\end{aligned}$$

where $\sigma_{(r:n)}^2 = \sigma^2\big(X_{(r:n)}\big)$, and let $\Psi_\infty^{\mathrm{I}}(s)$, $\Psi_\infty^{\mathrm{II}}(s)$ and $\Phi_\infty(s)$ be the limits of

$\Psi_n^{\mathrm{I}}(s)$, $\Psi_n^{\mathrm{II}}(s)$ and $\Phi_n(s)$, respectively (under the assumption that they exist).

Theorem 4.1. *Under the assumption that X is drawn from a distribution of a location-scale family, the distribution of $\sqrt{n}\big(\widehat{\rho}(s)-\rho\big)$ converges to the normal distribution of mean 0 and variance $\big\{\Psi_\infty^{\mathrm{I}}(s)/\sigma_Y^2+\rho^2\Psi_\infty^{\mathrm{II}}(s)\big\}/\Phi_\infty^2(s)$.*

Proof. Note that $\sigma_Y^2/\widehat{\sigma}_Y^2$ converges in probability to 1 as $n\to\infty$ and $\widehat{\rho}(s)$ has same asymptotic distribution with

$$\widetilde{\rho}(s)=\frac{1}{\sigma_Y}\cdot\frac{\sum_{r=1}^{[ns]}\alpha_{(r:n)}\big\{Y_{[r:n]}-\widehat{\mu}_Y\big\}}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}$$

and hereafter we let $\widehat{\rho}(s)=\widetilde{\rho}(s)$. Then,

$$\begin{aligned}\sqrt{n}\big(\widehat{\rho}(s)-\rho\big)&=\sqrt{n}\left\{\frac{1}{\sigma_Y}\frac{\sum_{r=1}^{[ns]}\alpha_{(r:n)}\big(Y_{[r:n]}-\widehat{\mu}_Y\big)}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}-\rho\right\}\\&=\sqrt{n}\Bigg\{\frac{1}{\sigma_Y}\frac{\sum_{r=1}^{[ns]}\alpha_{(r:n)}\big(Y_{[r:n]}-m(X_{(r:n)})\big)}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}+\\&\qquad\rho\frac{\sum_{r=1}^{[ns]}\alpha_{(r:n)}\left(\left(\frac{X_{(r:n)}-\mu_X}{\sigma_X}\right)-\alpha_{(r:n)}\right)}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}+\\&\qquad\frac{1}{\sigma_Y}\frac{\sum_{r=1}^{[ns]}\alpha_{(r:n)}(\mu_Y-\widehat{\mu}_Y)}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}\Bigg\}.\end{aligned}\tag{4.4}$$

Equation (4.4) can be written as

$$\frac{1}{\sigma_Y}\frac{\sqrt{n}\sqrt{n\Psi_n^{\mathrm{I}}(1)}}{\sum_{r=1}^{[ns]}\alpha_{(r:n)}^2}\mathrm{U}(s)+\rho\sqrt{n}\mathrm{V}(s)+\frac{\sqrt{n}}{\sigma_Y}\mathrm{R}(s),$$

where $\Psi_n^{\mathrm{I}}(1) = \sum_{r=1}^{n} \alpha_{(r:n)}^2 \sigma_{(r:n)}^2/n$ and

$$\begin{aligned} \mathrm{U}(s) &= \frac{1}{\sqrt{n\Psi_n^{\mathrm{I}}(1)}} \sum_{r=1}^{[ns]} \alpha_{(r:n)} \big(Y_{[r:n]} - m(X_{(r:n)})\big) \\ &= \frac{1}{\sqrt{n\Psi_n^{\mathrm{I}}(1)}} \sum_{r=1}^{[ns]} \left(\frac{\mathrm{E}(X_{(r:n)}) - \mu_X}{\sigma_X}\right) \Big(Y_{[r:n]} - m(X_{(r:n)})\Big), \end{aligned} \tag{4.5}$$

with $m(X_{(r:n)}) = \mathrm{E}\big(Y \big| X_{(r:n)}\big) = \mu_Y + \rho\sigma_Y(X_{(r:n)} - \mu_X)/\sigma_X$ and

$$\begin{aligned} \mathrm{V}(s) &= \frac{\sum_{r=1}^{[ns]} \alpha_{(r:n)} \big\{(X_{(r:n)} - \mu_X)/\sigma_X - \alpha_{(r:n)}\big\}}{\sum_{r=1}^{[ns]} \alpha_{(r:n)}^2}, \\ \mathrm{R}(s) &= \frac{\sum_{r=1}^{[ns]} \alpha_{(r:n)} (\mu_Y - \widehat{\mu}_Y)}{\sum_{r=1}^{[ns]} \alpha_{(r:n)}^2}. \end{aligned}$$

Since $\mathrm{R}(s)$ converges in probability to 0, we only consider the $\mathrm{U}(s)$ and $\mathrm{V}(s)$. Thus, the proof of the theorem is based on the functional central limit theorem for two partial sums of rank statistics, $\mathrm{U}(s)$ and $\mathrm{V}(s)$.

We first consider the asymptotic distribution of the process of taking the weighted partial sum of the induced rank statistic, which is

$$\begin{aligned} \mathrm{U}(s) &= \frac{1}{\sqrt{n\Psi_n^{\mathrm{I}}(1)}} \sum_{r=1}^{[ns]} \alpha_{(r:n)} \big(Y_{[r:n]} - m(X_{(r:n)})\big) \\ &= \frac{1}{\sqrt{n\Psi_n^{\mathrm{I}}(1)}} \sum_{r=1}^{[ns]} \left(\frac{\mathrm{E}(X_{(r:n)}) - \mu_X}{\sigma_X}\right) \Big(Y_{[r:n]} - m(X_{(r:n)})\Big). \end{aligned} \tag{4.6}$$

The main finding of Bhattacharya (1974) is the conditional independence of $Y_{[1:n]}, \ldots, Y_{[n:n]}$ given $X_1, X_2, \ldots, X_n$ (or equivalently, $X_{(1:n)}, X_{(2:n)}, \ldots, X_{(n:n)}$).

Thus, given $\mathcal{A} = \sigma\big(X_1, X_2, \ldots, X_n, \ldots\big)$, (4.6) can be read as

$$S_{nk} = \frac{1}{\sqrt{n\Psi_n^{\mathrm{I}}(1)}} \sum_{r=1}^{k} \alpha_{(r:n)}\sigma_{(r:n)}u_r, \quad k = 1, 2, \ldots, n, \tag{4.7}$$

where the u_r are independent, with mean 0 and variance $\sigma^2_{(r:n)}$. By applying the basic concept of Skorokhod embedding (Shorack and Wellner, 2009), we obtain a sequence of stopping times $\tau_{n1}, \tau_{n2}, \ldots, \tau_{nn}$ such that

- these stopping times are conditionally independent given $\mathcal{A}$,
- $\mathrm{E}\big(\tau_{nk}\big|\mathcal{A}\big) = \sum_{r=1}^{k} \alpha^2_{(r:n)}\sigma^2_{(r:n)}\big/\{n\Psi_n^{\mathrm{I}}(1)\}$,
- $\mathrm{var}\big(\tau_{nk}\big|\mathcal{A}\big) = \sum_{r=1}^{k} \alpha^4_{(r:n)}\mathrm{E}\big\{\big(Y_{[r:n]} - m(X_{(r:n)})\big)^4\big|\mathcal{A}\big\}\big/\{n\Psi_n^{\mathrm{I}}(1)\}^2 < \infty$, and
- $\big(S_{n1}, S_{n2}, \ldots, S_{nn}\big)$ has the same distribution as $\big(W(\tau_{n1}), W(\tau_{n1}+\tau_{n2}), \ldots, W(\tau_{n1}+\tau_{n2}+\cdots+\tau_{nn})\big)$, where $\big\{W(s),\ s \in [0,\infty)\big\}$ is conventional Brownian motion.

We now consider the embedded partial-sum process $\big\{W_n(s) : 0 \le s \le 1\big\}$ that is defined by $W_n(s) = S_{n[ns]}$. As in Bhattacharya (1974), it suffices to show that

$$\sup_{0\le s\le 1}\left|\frac{1}{n}\sum_{r=1}^{[ns]}\tau_{nr} - \frac{\Psi_n^{\mathrm{I}}(s)}{\Psi_n^{\mathrm{I}}(1)}\right| \tag{4.8}$$

converges to 0 probability.

For each $s \in [0,1]$, the strong law of large numbers states that $\big(1/n\big)\sum_{r=1}^{[ns]}\tau_{nr}$ almost certainly converges to $\Psi_\infty^{\mathrm{I}}(s)\big/\Psi_\infty^{\mathrm{I}}(1)$. Both $\big(1/n\big)\sum_{r=1}^{[ns]}\tau_{nr}$ and $\Psi_n^{\mathrm{I}}(s)\big/\Psi_n^{\mathrm{I}}(1)$ are increasing functions of s. Thus, using the same arguments

(Shorack and Wellner, 2009, pp. 62), we find that their sup difference also converges to 0.

Second,

$$\sqrt{n}\mathrm{V}(s) = \frac{1}{\Phi_n(s)}\frac{1}{\sqrt{n}}\left\{\sum_{r=1}^{[ns]}\alpha_{(r:n)}\left(\frac{X_{(r:n)}-\mu_X}{\sigma_X}-\alpha_{(r:n)}\right)\right\} \tag{4.9}$$

is a linear statistic of order statistics and converges to the normal distribution with mean 0 and variance $\Psi^{\mathrm{II}}_{\infty}(s)/\Phi^2_{\infty}(s)$ (David, 2003, Theorem 11.4). Here, we remark that both $\Psi^{\mathrm{II}}_{\infty}(s)$ and $\Phi_{\infty}(s)$ can also be written as functionals of the distribution of X, as shown in (David, 2003).

Finally, summing the asymptotic results of $\mathrm{U}_n(s)$ and $\mathrm{V}_n(s)$, we find that $\sqrt{n}\big(\widehat{\rho}(s)-\rho\big)$ converges to the normal distribution with mean 0 and variance

$$\frac{\Psi^{\mathrm{I}}_{\infty}(s)/\sigma^2_Y + \rho^2\Psi^{\mathrm{II}}_{\infty}(s)}{\Phi^2_{\infty}(s)}$$

This concludes the proof. □

We conclude this chapter with two remarks regarding Theorem 4.1. First, in Theorem 4.1, from the tower property of the conditional expectation,

$$\mathrm{var}\big(\sqrt{n}\widehat{\rho}\big) > \frac{1}{\big(1/n\big)\sum_{r=1}^{[ns]}\alpha^2_{(r:n)}} \geq \frac{1}{\mathrm{var}\big(X\big)} = 1,$$

and when $\rho = 0$, the asymptotic variance of $\sqrt{n}\widehat{\rho}$ is larger than 1, which is the variance of the least-squares estimator in the case where X is completely observed. Second, it is possible to test the hypothesis $\mathcal{H}_0 : \rho = 0$ using the statistic $\mathrm{T} = \sqrt{n}\widehat{\rho}$, which has an asymptotically normal distribution of mean

0 and variance $1/\Phi_\infty(s)$.

4.2. Residual Analysis

As in the classical linear model, the residuals can provide guidance for identifying a better model and score function. The residuals are defined as $e_{[r:n]} = \big(Y_{[r:n]} - \mu_Y\big)/\sigma_Y - \widehat{\rho}\alpha_{(r:n)}$ for $r = 1, 2, \ldots, [ns]$ and estimated residuals are defined as $\widehat{e}_{[r:n]}$ by replacing μ_Y and σ_Y with their empirical estimator $\widehat{\mu}_Y$ and $\widehat{\sigma}_Y$. Since the empirical mean estimator $\widehat{\mu}_Y$ is unbiased estimator of μ_Y and the empirical variance estimator $\widehat{\sigma}_Y^2$ is asymptotically unbiased when n is large, statistical properties of $\widehat{e}_{[r:n]}$ are equivalent to $e_{[r:n]}$. Several statistical properties of the residuals, which are analogous to those in the classical linear model, are summarized as follows.

Theorem 4.2. *Under the assumptions of Theorem 4.1, the following statements are true for the residuals:*

(i) $\mathrm{E}\big(e_{[r:n]}\big) = 0$;

(ii) $\mathrm{var}\big(e_{[r:n]}\big) = \Big\{\rho^2\beta_{(rr:n)} + \big(1-\rho^2\big)\Big\} +$
$\alpha^2_{(r:n)}\frac{1}{n\sigma_Y^2}\frac{\Psi^{\mathrm{I}}_n(s)}{\Phi^2_n(s)} - 2\frac{1}{\sum_{r=1}^{[ns]}\alpha^2_{(r:n)}} \times \Big\{\rho^2\sum_{k=1}^{[ns]}\alpha_{(k:n)}\alpha_{(r:n)}\beta_{(rk:n)} + \alpha^2_{(r:n)}(1-\rho^2)\Big\}$;

(iii) $\mathrm{E}\big(e_{[r:n]}\alpha_{(r:n)}\big) = 0$; *and*

(iv) $\mathrm{E}\big(e_{[r:n]}\widehat{Y}_{[r:n]}\big) = 0$, *where* $\widehat{Y}_{[r:n]} = \widehat{\rho}\alpha_{(r:n)}$.

The proof of Theorem 4.2 requires only simple algebra and is thus omitted here. The theorem states that the residuals have mean 0 and finite variance and also states that they are uncorrelated with the scores $\alpha_{(r:n)}$ and the predicted values $\widehat{Y}_{[r:n]}$. Thus, the residual plots, which are the plots of (i) r

versus $e_{[r:n]}$, (ii) $\alpha_{(r:n)}$ versus $e_{[r:n]}$, and (iii) $\widehat{Y}_{[r:n]}$ versus $e_{[r:n]}$, have the same interpretations as those of the classical linear model.

The residual sum of squares may be another useful tool for measuring the goodness of fit of the proposed model, as in the classical linear model. The residual sum of squares in our model is defined as

$$\text{RSS} = \sum_{r=1}^{[ns]} \Big(\frac{Y_{[r:n]} - \widehat{\mu}_Y}{\widehat{\sigma}_Y} - \widehat{Y}_{[r:n]} \Big)^2$$

and will be used along with the residual plots as a guide for selecting a better score function.

Finally, the proposed least-squares estimator (4.3) assume that the regression line between $\alpha_{(r:n)}$ and $\big(Y_{[r:n]} - \widehat{\mu}_Y\big)$ has an intercept (at the y axis) of 0. Thus, if the model (or the score function) is correctly specified, then the intercept estimated by the regression (with intercept) should be close to 0, and the estimated intercept therefore serves as a measure for checking the correctness of the score function. Note that the regression (without intercept) performed in this thesis is based on observations of the top $[ns]$ ranks and assumes that the function passes through the origin (see Figure 6.4).

4.3. An Estimator with Unranked Observations

The least-squares estimator presented in Chapter 4.2 does not fully use the information contained in $\big\{Y_{[r:n]}, r > m\big\}$; it is used only to estimate μ_Y and σ_Y, not to estimate ρ itself. In this chapter, we briefly demonstrate how $\widehat{\rho}$

can be modified to incorporate these unranked observations.

We consider the following modified estimator:

$$\widehat{\rho}_{\mathrm{m}}(s) \equiv \frac{1}{\widehat{\sigma}_Y} \cdot \frac{\sum_{r=1}^{[ns]} \alpha_{(r:n)} \{Y_{[r:n]} - \widehat{\mu}_Y\} + (n - [ns]) \overline{\alpha}_{[ns]+} (\overline{Y}_{[ns]+} - \widehat{\mu})}{\sum_{r=1}^{[ns]} \alpha_{(r:n)}^2 + (n - [ns]) \overline{\alpha}_{[ns]+}^2},$$

where $\overline{\alpha}_{[ns]+} = \sum_{r=[ns]+1}^{n} \alpha_{(r:n)} / (n - [ns])$ and $\overline{Y}_{[ns]+} = \sum_{r=[ns]+1}^{n} Y_{[r:n]} / (n - [ns])$. This modified estimator also asymptotically approaches the normal distribution. Specifically, suppose that

$$\widetilde{\alpha}_{(r:n)} = \begin{cases} \alpha_{(r:n)} & r = 1, 2, \ldots, [ns], \\ \\ \overline{\alpha}_{[ns]+} & r = [ns] + 1, [ns] + 2, \ldots, n. \end{cases}$$

We also suppose that

$$\begin{aligned} \widetilde{\Psi}_n^{\mathrm{I}}(s) &= (1/n) \left\{ \sum_{r=1}^{n} \widetilde{\alpha}_{(r:n)}^2 \sigma_{(r:n)}^2 \right\} \\ \widetilde{\Psi}_n^{\mathrm{II}}(s) &= (1/n) \left\{ \sum_{r1=1}^{n} \sum_{r2=1}^{n} \widetilde{\alpha}_{(r1:n)} \widetilde{\alpha}_{(r2:n)} \beta_{(r1,r2:n)}^2 \right\} \quad \text{and} \\ \widetilde{\Phi}_n(s) &= (1/n) \sum_{r=1}^{n} \widetilde{\alpha}_{(r:n)}^2. \end{aligned}$$

As in the previous chapter, $(1/n)$-scaled limits of $\widetilde{\Psi}_n^{\mathrm{I}}(s)$, $\widetilde{\Psi}_n^{\mathrm{II}}(s)$ and $\widetilde{\Phi}_n(s)$ exist; let these limits be $\widetilde{\Psi}_\infty^{\mathrm{I}}(s) = \lim_{n\to\infty} \widetilde{\Psi}_n^{\mathrm{I}}(s)/n$, $\widetilde{\Psi}_\infty^{\mathrm{II}}(s) = \lim_{n\to\infty} \tilde{\Psi}_n^{\mathrm{II}}(s)/n$ and $\tilde{\Phi}_\infty(s) = \lim_{n\to\infty} \tilde{\Phi}_n(s)/n$, respectively. Then, we can write the following theorem.

Theorem 4.3. *Under the same assumptions as those of Theorem 4.1, the*

distribution of $\sqrt{n}\big(\widehat{\rho}_{\rm m}(s)-\rho\big)$ *converges to the normal distribution with mean* 0 *and variance* $\big\{\widetilde{\Psi}^{\rm I}_{\infty}(s)/\sigma_Y^2+\rho^2\widetilde{\Psi}^{\rm II}_{\infty}(s)\big\}/\widetilde{\Phi}^2_{\infty}(s)$

Proof.

$$
\begin{aligned}
&\sqrt{n}\ \big(\widehat{\rho}_{\rm m}-\rho\big)\\
&\qquad = \quad \sqrt{n}\Bigg\{\frac{1}{\sum_{r=1}^{[ns]}\alpha^2_{(r:n)}+\big(n-[ns]\big)\overline{\alpha}^2_{[ns]+}}\times\\
&\qquad\qquad \Bigg(\sum_{r=1}^{[ns]}\alpha_{(r:n)}\Big(Y_{[r:n]}-\widehat{\mu}_Y\Big)+\big(n-[ns]\big)\overline{\alpha}_{[ns]+}\big(\overline{Y}_{[ns]+}-\widehat{\mu}_Y\big)\Bigg)-\rho\Bigg\}\\
&\qquad = \quad \sqrt{n}\Bigg(\frac{1}{\widehat{\sigma}_Y}\frac{\sum_{r=1}^{n}\widetilde{\alpha}_{(r:n)}\big(Y_{[r:n]}-\widehat{\mu}_Y\big)}{\sum_{r=1}^{n}\widetilde{\alpha}^2_{(r:n)}}-\rho\Bigg),
\end{aligned}
$$

the distribution of which converges to the normal distribution with mean 0 and variance
$\big\{\widetilde{\Psi}^{\rm I}_{\infty}(s)/\sigma_Y^2+\rho^2\widetilde{\Psi}^{\rm II}_{\infty}(s)\big\}/\widetilde{\Phi}^2_{\infty}(s)$ following the same arguments presented in the proof of Theorem 4.1. □

Chapter 5

Numerical Study

In this chapter, we numerically investigate the advantage we can gain by choosing the correct score function to estimate $\rho = \text{corr}(Y, X)$. We view the importance of choosing score function given rare informations in Chapter 2. In numerical study, the performance of an estimator is measured in terms of its bias and its mean square error (MSE), which we numerically estimate based on 1000 simulated data sets and the estimators obtained therefrom.

5.1. Study setup

The data sets are generated from the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where β_0 and β_1 are the regression coefficient to be estimated and the ϵ_i are independently drawn from $N(0,1)$. We consider three distributions for X which has different dispersion : the uniform distribution on $[0,1]$, the standard normal distribution, and the gamma distribution with mean 1 and variance 1/3. As stated in Chapter 3, the score function of the uniform distribution is almost equivalent to the identity score function $S_n(r) = r$. However, the normal distribution and the gamma distribution have heavier tails than does the uniform distribution, and their score functions are convex in the right tail. We set the parameters δ to ensure that $\rho = 0$, 0.3, 0.5 and 0.7, where $\rho = \delta/\sigma_Y$. Finally, in each considered case, the sample size n and the number of partially observed ranks m are set to all possible combinations of $n = 500$ or 2000 and $r = 20$, 50, or 100. When estimating ρ, we apply four different scores, including the proposed MOS-based score functions obtained from the three distributions listed above and the identity score function, which is commonly used in practice. The approximated bias and MSE values are reported in Tables 5.1 and 5.2.

5.2. Interpretations and results

We can observe several interesting findings from these tables. First, the correctly specified score function performs better than do others when there exists a strong correlation between X and Y (when ρ is large). However, when $\rho = 0$, there is almost no difference among the four considered scores. Second, as the number of observations increases, in the sense that either r or n increases, the superiority of the correctly specified scores with respect to

the others becomes apparent even when ρ is not large. Third, as conjectured in Chapter 3, the scores based on the uniform distribution perform almost identically to the identity scores. This result is not surprising since the both uniform scores and identity scores are equally spaced scores. Finally, the differences between the correctly specified scores and the others are significant regardless of ρ or the sample size (r or n) when the distribution of X has a heavier right tail (the gamma distribution).

			$U(0,1)$		$N(0,1)$		$G(3,3)$	
	ρ	Dist	Bias	MSE	Bias	MSE	Bias	MSE
r=20	0.0	1:N	0.0009	0.0167	-0.0009	0.0165	-0.0025	0.0184
		U	-0.0009	0.0175	-0.0027	0.0174	0.0051	0.0170
		N	-0.0032	0.0103	0.0009	0.0104	0.0008	0.0097
		G	0.0001	0.0059	-0.0044	0.0059	0.0021	0.0057
	0.3	1:N	-0.0003	0.0161	0.0923	0.0243	0.2090	0.0600
		U	0.0026	0.0158	0.0900	0.0256	0.2122	0.0603
		N	-0.0715	**0.0141**	-0.0010	**0.0098**	0.0990	0.0194
		G	-0.1287	0.0218	-0.0794	0.0118	-0.0024	**0.0057**
	0.5	1:N	0.0012	0.0127	0.1437	0.0336	0.3430	0.1327
		U	0.0035	**0.0121**	0.1474	0.0351	0.3476	0.1354
		N	-0.1217	0.0222	0.0009	**0.0078**	0.1566	0.0330
		G	-0.2180	0.0518	-0.1241	0.0197	-0.0022	**0.0052**
	0.7	1:N	-0.0025	**0.0087**	0.2057	0.0522	0.4916	0.2535
		U	-0.0003	0.0090	0.2053	0.0517	0.4886	0.2505
		N	-0.1693	0.0334	-0.0009	**0.0057**	0.2266	0.0584
		G	-0.3051	0.0958	-0.1750	0.0338	-0.0057	**0.0040**
r=50	0.0	1:N	0.0011	0.0077	0.0007	0.0076	-0.0020	0.0069
		U	-0.0019	0.0075	-0.0005	0.0071	0.0034	0.0076
		N	-0.0038	0.0056	0.0008	0.0054	0.0008	0.0053
		G	-0.0016	0.0035	0.0004	0.0039	0.0004	0.0034
	0.3	1:N	-0.0033	**0.0064**	0.0405	0.0083	0.1116	0.0196
		U	-0.0002	0.0066	0.0427	0.0082	0.1112	0.0187
		N	-0.0418	0.0067	-0.0025	**0.0051**	0.0735	0.0110
		G	-0.0988	0.0131	-0.0604	0.0066	-0.0029	**0.0037**
	0.5	1:N	0.0022	0.0050	0.0657	0.0096	0.1878	0.0413
		U	-0.0011	**0.0049**	0.0687	0.0101	0.1882	0.0408
		N	-0.0717	0.0093	0.0022	**0.0040**	0.1217	0.0192
		G	-0.1652	0.0298	-0.0982	0.0123	-0.0023	**0.0031**
	0.7	1:N	-0.0006	0.0035	0.0897	0.0116	0.2655	0.0744
		U	0.0006	**0.0035**	0.0975	0.0133	0.2641	0.0736
		N	-0.0998	0.0127	0.0044	**0.0026**	0.1655	0.0307
		G	-0.2293	0.0544	-0.1401	0.0216	0.0040	**0.0021**
r=100	0.0	1:N	0.0005	0.0042	0.0005	0.0042	0.0040	0.0042
		U	-0.0017	0.0040	0.0000	0.0043	-0.0021	0.0039
		N	0.0003	0.0035	-0.0029	0.0036	-0.0012	0.0040
		G	0.0003	0.0028	0.0005	0.0027	0.0007	0.0027
	0.3	1:N	0.0013	0.0035	0.0107	0.0034	0.0490	0.0061
		U	0.0008	0.0037	0.0099	0.0036	0.0498	0.0062
		N	-0.0210	**0.0033**	0.0005	0.0034	0.0517	0.0063
		G	-0.0714	0.0074	-0.0479	**0.0049**	-0.0038	**0.0026**
	0.5	1:N	0.0003	0.0028	0.0161	0.0031	0.0896	0.0110
		U	-0.0002	**0.0028**	0.0152	0.0031	0.0890	0.0108
		N	-0.0352	0.0036	-0.0011	**0.0025**	0.0836	0.0096
		G	-0.1187	0.0159	-0.0823	0.0088	-0.0035	**0.0020**
	0.7	1:N	0.0008	0.0017	0.0230	0.0022	0.1210	0.0166
		U	0.0004	**0.0017**	0.0225	0.0023	0.1247	0.0175
		N	-0.0468	0.0038	-0.0002	**0.0017**	0.1186	0.0158
		G	-0.1689	0.0297	-0.1140	0.0143	-0.0016	**0.0014**

Table 5.1: $n = 500$: In the MSE columns, the numbers in bold-faced are the smallest among the evaluated score functions. In both the bias and MSE columns, the underlined numbers are the values from the correctly specified score functions.

			$U(0,1)$		$N(0,1)$		$G(3,3)$	
	ρ	Dist	Bias	MSE	Bias	MSE	Bias	MSE
r=20	0.0	1:N	-0.0007	0.0173	0.0000	0.0168	0.0006	0.0165
		U	0.0077	0.0168	0.0029	0.0166	0.0066	0.0170
		N	-0.0042	0.0072	-0.0023	0.0069	0.0008	0.0068
		G	-0.0014	0.0032	0.0017	0.0034	0.0005	0.0031
	0.3	1:N	-0.0020	**0.0157**	0.1650	0.0433	0.3691	0.1519
		U	-0.0033	0.0158	0.1630	0.0427	0.3707	0.1537
		N	-0.1068	0.0174	0.0028	**0.0062**	0.1319	0.0249
		G	-0.1680	0.0312	-0.0925	0.0118	0.0041	**0.0031**
	0.5	1:N	-0.0053	0.0123	0.2821	0.0931	0.6109	0.3891
		U	0.0009	**0.0122**	0.2714	0.0870	0.6163	0.3952
		N	-0.1822	0.0387	-0.0010	**0.0056**	0.2238	0.0566
		G	-0.2821	0.0821	-0.1537	0.0263	0.0020	**0.0031**
	0.7	1:N	-0.0013	**0.0083**	0.3825	0.1558	0.8650	0.7618
		U	0.0030	0.0091	0.3859	0.1586	0.8623	0.7575
		N	-0.2528	0.0671	0.0014	**0.0039**	0.3113	0.1026
		G	-0.3936	0.1567	-0.2182	0.0496	0.0037	**0.0029**
r=50	0.0	1:N	-0.0013	0.0071	0.0012	0.0070	0.0009	0.0071
		U	0.0024	0.0071	0.0020	0.0066	-0.0008	0.0075
		N	0.0015	0.0036	0.0040	0.0038	-0.0007	0.0033
		G	0.0002	0.0018	-0.0009	0.0018	0.0011	0.0019
	0.3	1:N	0.0016	0.0062	0.1147	0.0196	0.2589	0.0737
		U	-0.0028	**0.0062**	0.1136	0.0192	0.2667	0.0776
		N	-0.0851	0.0102	0.0012	**0.0035**	0.1121	0.0159
		G	-0.1456	0.0229	-0.0818	0.0083	0.0002	**0.0018**
	0.5	1:N	0.0025	0.0051	0.1953	0.0436	0.4396	0.1988
		U	-0.0008	**0.0048**	0.1928	0.0426	0.4389	0.1986
		N	-0.1486	0.0246	-0.0006	**0.0025**	0.1854	0.0376
		G	-0.2457	0.0617	-0.1371	0.0202	-0.0022	**0.0017**
	0.7	1:N	0.0006	0.0035	0.2698	0.0767	0.6124	0.3806
		U	0.0011	**0.0034**	0.2691	0.0761	0.6195	0.3882
		N	-0.2019	0.0426	0.0015	**0.0020**	0.2578	0.0690
		G	-0.3429	0.1184	-0.1901	0.0372	-0.0018	**0.0014**
r=100	0.0	1:N	0.0002	0.0036	0.0015	0.0038	0.0039	0.0034
		U	-0.0008	0.0035	0.0013	0.0036	-0.0002	0.0037
		N	0.0012	0.0021	0.0039	0.0022	-0.0029	0.0021
		G	-0.0003	0.0013	-0.0014	0.0013	0.0001	0.0013
	0.3	1:N	0.0031	0.0032	0.0791	0.0095	0.1863	0.0380
		U	-0.0006	**0.0029**	0.0764	0.0089	0.1847	0.0374
		N	-0.0698	0.0067	0.0000	**0.0021**	0.0920	0.0105
		G	-0.1247	0.0167	-0.0711	0.0062	-0.0006	**0.0012**
	0.5	1:N	-0.0008	**0.0024**	0.1302	0.0196	0.3124	0.1006
		U	0.0001	0.0029	0.1284	0.0191	0.3114	0.1000
		N	-0.1107	0.0139	0.0009	**0.0016**	0.1530	0.0252
		G	-0.2105	0.0453	-0.1204	0.0155	-0.0001	**0.0011**
	0.7	1:N	-0.0012	0.0018	0.1793	0.0339	0.4344	0.1910
		U	0.0001	**0.0017**	0.1793	0.0340	0.4365	0.1927
		N	-0.1551	0.0251	0.0010	**0.0011**	0.2156	0.0479
		G	-0.2923	0.0861	-0.1691	0.0292	-0.0009	**0.0009**

Table 5.2: $n = 2000$: In the MSE columns, the numbers in bold-faced are the smallest among the evaluated score functions. In both the bias and MSE columns, the underlined numbers are the values from the correctly specified score functions.

Chapter 6

Data Examples

6.1. Data Description

To investigate how the attention of investors affects stock returns, we merge the hand-collected `Daum` rank data set and the financial data from `FnGuide`. The structure of merged data sets are presented in Table 6.1 as example. We illustrate how the returns of attention-grabbing stocks fluctuate around the event dates when investors pay attention to these stocks. The variables to be used in the analysis are as follows. (1) "R": The rank of an individual stock on day t; if the rank value is 1, then the stock is the most frequently discussed stock on the `Daum` stock message board on that day. This is the key variable that measures the degree of investor attention. (2) "RN": Raw returns on day $t + 1$ (the next day) (%), which is of primary interest and is the quantity that we wish to predict. (3)" R0": Raw returns on day t

(%). (4) "R1": Raw returns on day $t-1$ (%). (5) "R2": Raw returns on day $t-2$ (%). (6) "R3": Raw returns on day $t-3$ (%). (7) "R4": Raw returns on day $t-4$ (%). (8) "R5": Raw returns on day $t-5$ (%). (9) "ME": Market capitalization (1 trillion Korean won). (10) "T": Turnover ratio defined as the trading volume divided by the number of outstanding shares. (11) "TA": Turnover ratio defined as the trading volume divided by market capitalization.

Date(yyyy.mm.dd)	R	R0	RN	R1	R2	R3	R4	R5	T	TA
2010.10.04	1	-0.62	0.33	-	-	-	-	-	0.1992	0.1977
2010.10.04	2	0.23	0.42	-	-	-	-	-	0.1238	0.1382
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010.10.04	29	0.58	0.75	-	-	-	-	-	0.1342	0.1894
2010.10.04	30	-0.25	0.12	-	-	-	-	-	0.2134	0.2423
2010.10.04	+	0.54	0.56	-	-	-	-	-	0.1849	0.1912
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010.10.04	+	0.12	0.17	-	-	-	-	-	0.2341	0.2467
2010.10.05	1	0.58	0.68	0.43	-	-	-	-	0.1385	0.1482
2010.10.05	2	-0.34	-0.31	0.12	-	-	-	-	0.2391	0.2149
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010.10.05	29	0.78	0.57	0.56	-	-	-	-	0.1451	0.1305
2010.10.05	30	0.15	0.24	0.44	-	-	-	-	0.1128	0.1592
2010.10.05	+	0.39	0.48	0.22	-	-	-	-	0.2193	0.2841
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010.10.05	+	0.59	0.12	0.48	-	-	-	-	0.1837	0.1731
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012.11.23	1	0.41	-	0.24	0.43	0.23	0.43	0.23	0.1184	0.1498
2012.11.23	2	0.76	-	0.48	0.37	0.21	0.11	-0.15	0.1038	0.1293
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012.11.23	29	-0.39	-	0.38	0.36	0.36	0.31	0.11	0.2931	0.2837
2012.11.23	30	0.31	-	0.12	0.38	0.12	-0.11	-0.38	0.1983	0.1723
2012.11.23	+	0.19	-	0.19	0.16	0.08	0.17	0.23	0.2031	0.2231
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012.11.23	+	-0.39	-	-0.22	-0.14	0.12	0.23	0.45	0.2312	0.2414

Table 6.1: Example of data set: "+" means that rank is over 30.

6.2. Attention and Predictive Stock Returns

As stated previously, the primary goal of our analysis is to determine how the returns of attention-grabbing stocks will fluctuate around the event dates when investors pay attention to these stocks. The next-day return can also be influenced by several other factors in addition to investor attention. To account for the effects of these other factors, we consider the residuals obtained after regressing the next-day return against all other covariates except the rank, "R". These residuals are obtained from the multiple linear regression model, which is defined as follows:

$$\mathrm{RN}_i = \beta_0 + \sum_{l=0}^{5} \beta_{l+1} \mathrm{R}l_i + \beta_7 \mathrm{ME}_i + \beta_8 \mathrm{T}_i + \beta_9 \mathrm{TA}_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{6.1}$$

where $n(= 1,771)$ is the total number of companies on the market. Let Y_i^t be the absolute (value of the) residual of company i obtained from the regression (6.1). We then select the absolute residuals whose ranks are reported to be within the top 30 for the primary analysis. Below, $Y_{[r:n]}^t$ is the absolute residual corresponding to rank r on day t for $t = 1, 2, \ldots, T(= 537)$.

In Figure 6.1, we plot the quantiles of $\left\{Y_{[r:n]}^t, t = 1, 2, \ldots, T\right\}$ for each $r = 1, 2, \ldots, 30$. This figure reveals non-linearity of $Y_{[r:n]}^t$ at $r = 1$ and 2, which we hypothesize reflects the heterogeneity of investor expectations with regard to highly attention-grabbing stocks. In other words, the ranking of the `Daum` board is purely determined by the attention of individual investors, and stocks related to news that is difficult to characterize as either good or bad often receive the greatest attention and the highest ranks. We introduce an additional term to explain this apparent non-linearity and consider the

model

$$Y^t_{[r:n]} = \mu^t_Y + \rho^t \sigma^t_Y \frac{X^t_{(r:n)} - \mu^t_X}{\sigma^t_X} + \gamma^t \mathrm{I}(r \leq 2) + \epsilon^t_{[r:n]}, \quad r = 1, 2., \ldots, 30, \quad (6.2)$$

for $t = 1, 2, \ldots, T$ with $T = 537$ and $n = 1,771$.

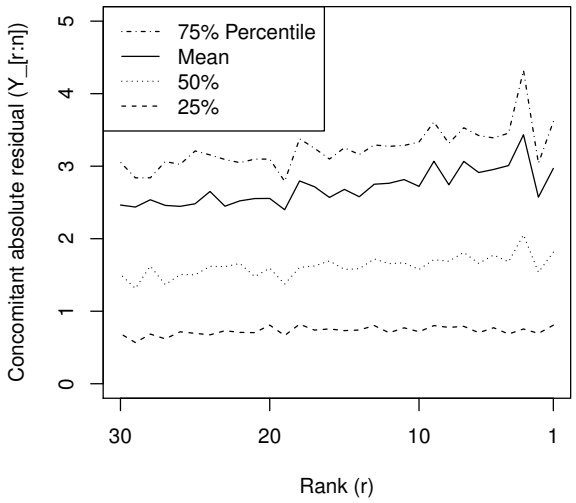

Figure 6.1: Plot of the means and quantiles of $\left\{Y^t_{[r:n]}, t = 1, 2, \ldots, T\right\}$ for each $r = 1, 2, \ldots, 30$.

6.3. Regression with Ranks

In the regression model, we consider the scores from the standardized distributions of the location-scale families generated by the following three distributions: (i) a uniform distribution on $(0, 1)$ (called the uniform score), (ii) a positive normal distribution $X = |Z|$, $Z \sim N(0, 1)$ (called the half-normal

score), and (iii) a power-law distribution X whose CDF is $F(x) = 1 - x^{-\alpha}$ with $\alpha = 2.3$ (called the power-law score). The scores are illustrated on different scales in Figure 6.2.

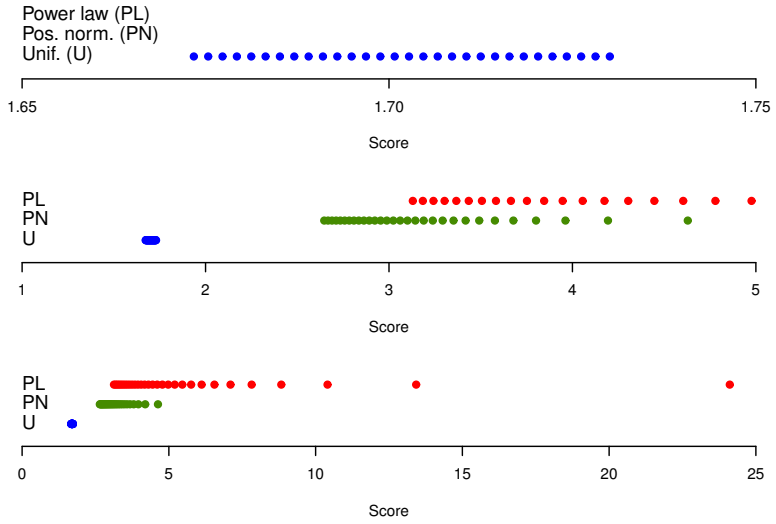

Figure 6.2: $\{\alpha_{(r:n)}\}_{r=1,\ldots,30}$ for each distribution on different scales, where $n = 1,771$.

We estimate ρ^t and γ^t to minimize the empirical squared-error loss of the model (6.2) by iterating the following steps:

1. Given the least-squares estimator of ρ, denoted by $\widehat{\rho}^t_{(0)}$, update the estimate of γ as follows:

$$\widehat{\gamma}^t = \frac{1}{2}\left[\left(Y^t_{[1:n]} - \mu^t_Y - \sigma^t_Y \widehat{\rho}^t_{(0)} \alpha_{(1:n)}\right) + \left(Y^t_{[2:n]} - \mu^t_Y - \sigma^t_Y \widehat{\rho}^t_{(0)} \alpha_{(2:n)}\right)\right].$$

2. Given the estimate of γ, denoted by $\widehat{\gamma}^t_{(0)}$, update the estimate of ρ using

the LSE proposed in the previous chapter as follows:

$$\widehat{\rho}^t = \frac{1}{\sigma_Y^t}\left\{\frac{\sum_{r=1}^{30}\alpha_{(r:n)}\big(Y_{[r:n]}^t - \mu_Y^t - \widehat{\gamma}_{(0)}^t \mathrm{I}(r\le 2)\big)}{\sum_{r=1}^{30}\alpha_{(r:n)}^2}\right\}.$$

In the analysis, the initial value $\widehat{\rho}_{(0)}^t$ is obtained from the preliminary linear regression on $\left\{\big(\alpha_{(r:n)}, (Y_{[r:n]}^t - \mu_Y^t)/\sigma_Y^t\big)\right\}_{r=3,\cdots,30}$, $t = 1,\ldots,T$, in which the data corresponding to $r = 1, 2$ are excluded. By contrast, μ_Y^t and σ_Y^t are estimated based on their empirical values as follows: $\widehat{\mu}_Y^t = \big(\sum_{r=1}^n Y_{[r:n]}^t\big)/n$ and $\big(\widehat{\sigma}_Y^t\big)^2 = \sum_{r=1}^n \big(Y_{[r:n]}^t - \widehat{\mu}_Y^t\big)^2/n$.

To choose the most appropriate score function among the three considered, we follow the guidelines presented in Chapter 4.2 and perform a residual analysis. First, we plot $\alpha_{(r:n)}$ and the quantiles of the corresponding residuals to identify any remaining trend not explained by the model (see Figure 6.3). This figure shows that the uniform score and the half-normal score exhibit additional linear trends not explained by the linear model (6.2), whereas the power-law score performs well. Second, we plot

$$\left(\alpha_{(r:n)}, \frac{Y_{[r:n]}^t - \widehat{\mu}_Y^t}{\widehat{\sigma}_Y^t}\right), \quad r = 3, 4, \ldots, 30, t = 1, 2, \ldots, T,$$

and apply the least-squares fits with/without intercept. The estimated regression line with intercept should cross the origin if the scores are correctly specified. Figure 6.4 reveals that the (estimate of) the intercept of the power-law score is closest to zero among the intercepts of the three considered scores. Finally, the residual sums of squares of the three scores are found to be 152186.7, 150706.3, and 150288.9, respectively. This finding also supports

the superiority of the power-law score function, and in the following analysis, we focus on the power-law score function.

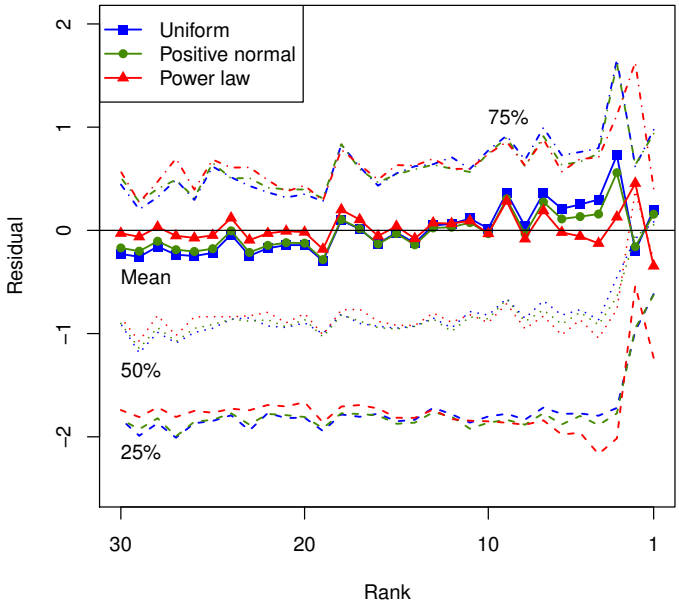

Figure 6.3: The averages and quantiles of the residuals for each rank.

6.4. Test of the Effect of Investor Attention on the Next-day Returns

The primary goal of the analysis is to investigate whether the attention of investors affects the returns of a stock on the following day. Specifically, we are interested in testing $\mathcal{H}_0 : \rho = 0$ under the assumption that $\rho^t = \rho$ for every t. To test this hypothesis, we consider a combined statistic of

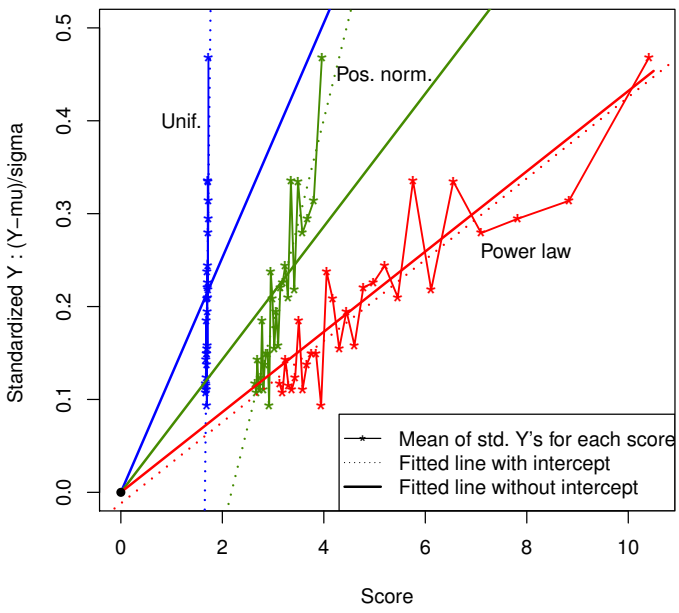

Figure 6.4: Check of proportionality between the standardized residuals and the scores. Points that are marked by '*' represent the average standardized residuals for each score (rank), the dotted line represents the fitted model for a naïve simple regression with intercept, and the solid line represents our model. Refer to Chapters 4.5 and 6.3 for details.

$\{\widehat{\rho}^t, t = 1, 2, \ldots, T\}$, that is,

$$\mathbf{t}_\rho = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} U_t, \tag{6.3}$$

where $U_t = \sqrt{n}\widehat{\rho}^t$. Here, the estimates of ρ for each day t, denoted by $\widehat{\rho}^t$, are serially dependent on each other, as are the U_ts. Thus, to obtain the reference distribution of $\mathbf{t}_\rho$, we further assume that $\{U_t,\ t = 1, 2, \ldots, T\}$ is stationary and that $\mathrm{E}|U_t|^{2+\kappa} < \infty$ for $\kappa > 0$. Under these assumptions, the

null distribution of $\mathbf{t}_\rho$ is asymptotically normal with mean 0 and variance

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T} \big(T-k\big)\mathrm{cov}(U_t, U_{t+k}).$$

The variance can be empirically estimated from the observed values of $\{U_t,\ t = 1, 2, \ldots, T\}$ as

$$\frac{1}{T} \sum_{k=0}^{m} \big(T-k\big)\widehat{\mathrm{cov}}(U_t, U_{t+k})$$

for sufficiently large m, where $\widehat{\mathrm{cov}}(U_t, U_{t+k})$ denotes the empirical covariance of the observed statistics $(U_1, U_{1+k}), (U_2, U_{2+k}), \ldots, (U_{T-k+1}, U_T)$. An additional interesting feature of the combined procedure is that the test statistic $\mathbf{t}_\rho$ is a rough estimator of ρ for all T trading days (after the scaling). It is calculated as

$$\begin{aligned}
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} U_t &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{n}\widehat{\rho}^t = \sqrt{nT} \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{\rho}^t \right) \\
&= \sqrt{nT} \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{r=1}^{30} \alpha_{(r:n)} \big\{ Y^t_{[r:n]} - \widehat{\mu}^t_Y - \widehat{\gamma}^t \mathrm{I}(r \le 2) \big\}}{\widehat{\sigma}^t_Y \sum_{r=1}^{30} \alpha^2_{(r:n)}} \\
&\approx \sqrt{nT} \frac{1}{\widehat{\sigma}_Y} \frac{\sum_{t=1}^{T} \sum_{r=1}^{30} \alpha_{(r:n)} \big\{ Y^t_{[r:n]} - \widehat{\mu}_Y - \widehat{\gamma}^t \mathrm{I}(r \le 2) \big\}}{T \sum_{r=1}^{30} \alpha^2_{(r:n)}} \\
&\approx \sqrt{nT} \widehat{\rho}^{\mathrm{lse}}, \qquad (6.4)
\end{aligned}$$

where $\widehat{\rho}^{\mathrm{lse}}$ is the least-squares estimator under the assumption that $\rho^t = \rho$ for all t. The difference between the right- and left-hand sides of (6.4) lies in the definition of $\widehat{\gamma}^t$, which is defined using $\widehat{\rho}^t$ rather than $\widehat{\rho}^{\mathrm{lse}}$.

The results of the test indicate that the average value of $\widehat{\rho}^t$, which is an estimator of ρ, is 0.043. The p-value obtained when testing $\mathcal{H}_0 : \rho = 0$

is less than 10^{-5} and statistically supports the association between investor attention and the next-day returns of the stocks.

6.5. Test on overall Correlation

In the previous chapter we used the average value of $\widehat{\rho}^t$ as an estimator of ρ. Since the estimators $\widehat{\rho}^t$ are serially dependent on each other, we considered the combined procedures for testing hypothesis. We now analyze the same data set to obtain the overall estimator $\widehat{\rho}$ instead of $\widehat{\rho}^t$ and use it for testing $\mathcal{H}_0 : \rho = 0$. This can be simpler and more intuitive analysis of data since we directly estimate ρ rather than ρ^t. The process of data analysis is the same in previous chapter. The absolute residuals $Y_{[r:n]}$ without t are obtained from (6.1) in Chapter 6.2 and we consider the regression model as

$$Y_{[r:n]} = \mu_Y + \rho\sigma_Y \frac{X_{(r:n)} - \mu_X}{\sigma_X} + \gamma \mathrm{I}(r \leq 2) + \epsilon_{[r:n]}, \quad r = 1, 2. \ldots, 30, \quad n = 1771. \tag{6.5}$$

At first, we consider the scores from the standardized distributions of the location-scale families: (i) a uniform score (ii) the half-normal score, and (iii) a power-law score with $\alpha = 2.23$. The scores are illustrated on different scales in Figure 6.5.

Then, we suggest to minimize the empirical squared-error loss of the model (6.5) for estimating ρ and γ by iterating the following steps:

1. Given the least-squares estimator $\widehat{\rho}_{(0)}$, update the estimate of γ as

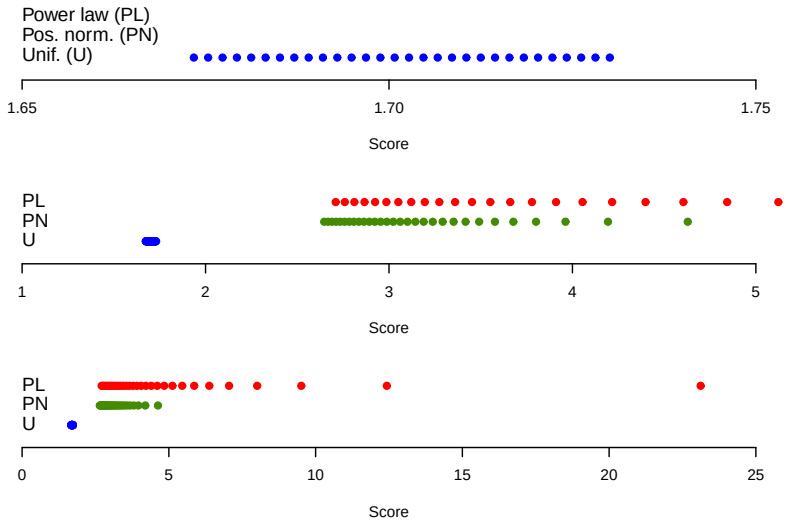

Figure 6.5: $\{\alpha_{(r:n)}\}_{r=1,\ldots,30}$ for each distribution on different scales, where $n = 1,771$.

follows:

$$\widehat{\gamma} = \frac{1}{2}\left[\left(Y_{[1:n]} - \mu_Y - \sigma_Y\widehat{\rho}_{(0)}\alpha_{(1:n)}\right) + \left(Y_{[2:n]} - \mu_Y - \sigma_Y\widehat{\rho}_{(0)}\alpha_{(2:n)}\right)\right].$$

2. Given the estimate $\widehat{\gamma}_{(0)}$, update the estimate of ρ using the LSE as follows:

$$\widehat{\rho} = \frac{1}{\sigma_Y}\left\{\frac{\sum_{r=1}^{30}\alpha_{(r:n)}\left(Y_{[r:n]} - \mu_Y - \widehat{\gamma}_{(0)}\mathrm{I}(r \le 2)\right)}{\sum_{r=1}^{30}\alpha_{(r:n)}^2}\right\}.$$

The initial value $\widehat{\rho}_{(0)}$ is obtained from the linear regression on $\left\{\left(\alpha_{(r:n)}, (Y_{[r:n]} - \mu_Y)/\sigma_Y\right)\right\}_{r=3,\cdots,30}$ and the empirical value $\widehat{\mu}_Y$ and $\widehat{\sigma}_Y^2$ are used for the estimator of μ_Y and σ_Y.

The guidelines presented in Chapter 4.2 are tools for selecting the most

appropriate score function among the three considered. Thus, in residual analysis, we first plot $\alpha_{(r:n)}$ and the quantiles of the corresponding residuals to check out any trend not explained by the model (see Figure 6.6). This figure shows that the uniform score and the half-normal score still exhibit additional linear trends and although the power-law score does not reveal linear trends, their expectation of residuals seem not to be 0. Second, we plot

$$\left(\alpha_{(r:n)}, \frac{Y_{[r:n]} - \widehat{\mu}_Y}{\widehat{\sigma}_Y}\right), \quad r = 3, 4, \ldots, 30,$$

and apply the least-squares fits with/without intercept. Figure 6.7 reveals that the (estimate of) the intercept of the power-law score is closest to zero

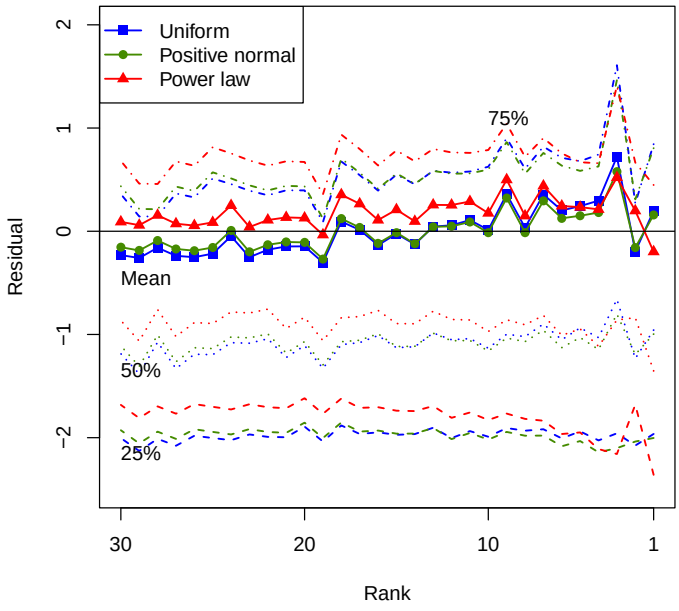

Figure 6.6: The averages and quantiles of the residuals for each rank.

among the intercepts of the three considered scores. Finally, the residual sums of squares of the three scores are calculated as 166531.7, 166194.7, and 166666.7, respectively. The power-law score has biggest residual sums of squares among others, but we choose the power-law score in testing since their estimate of intercept is closed to zero.

To test the hypothesis, $\mathcal{H}_0 : \rho = 0$, we consider a test statistic under null

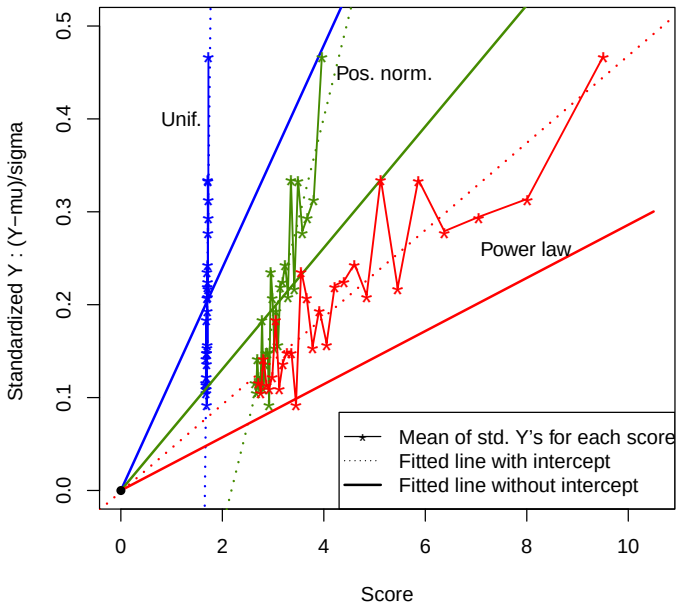

Figure 6.7: Check of proportionality between the standardized residuals and the scores. Points that are marked by '*' represent the average standardized residuals for each score (rank), the dotted line represents the fitted model for a naïve simple regression with intercept, and the solid line represents our model. Refer to Chapters 4.5 and 6.3 for details.

as

$$\mathbf{T}_{\rho} = \frac{\sqrt{n}\widehat{\rho}}{\sqrt{\mathrm{Var}(\widehat{\rho})}}, \tag{6.6}$$

where $\mathrm{Var}(\widehat{\rho})$ is asymptotic variance of estimator $\widehat{\rho}$ and it can be obtained from samples. The results of $\widehat{\rho}$ is 0.0407 and their p-value obtains when testing $\mathcal{H}_0 : \rho = 0$ is less than 10^{-5}. This result statistically provides for the association between investor attention and the next-day returns of the stocks.

Chapter 7

Concluding remarks

In this thesis, we study a regression problem based on a partially observed rank covariate. We propose a new set of score functions and study their application in simple linear regression. We demonstrate that the least-squares estimator that is calculated based on the newly proposed score consistently estimates the correlation coefficient between the response and the unobserved true covariate if the score function is correctly specified. We also define procedures based on the obtained residuals to identify the correct score function for the given data. The proposed estimator and procedures are applied to rank data collected from `Daum.net`, and we empirically verify the association between investor attention and next-day stock returns.

We now conclude the paper with a few remarks regarding our work. First, we wish to highlight that our work represents the first attempt in the literature to improve the selection of score functions for rank data. Second, we can extend the results of this thesis to multiple regression. Specifically, suppose

that we wish to consider the multiple regression model

$$Y_i = \mu_Y + \frac{X_i - \mu_X}{\sigma_X}\delta + (\mathbf{Z}_i - \mu_{\mathbf{Z}})^{\mathrm{T}}\eta + \epsilon_i$$

for an additional covariate vector $\mathbf{Z} = \big(Z_1, Z_2, \ldots, Z_q\big)^{\mathrm{T}}$. The least-squares estimates of δ is obtained from the solution of equations

$$\begin{pmatrix} \sum_{r=1}^{[ns]} \alpha_{(r:n)}^2 & \sum_{r=1}^{[ns]} \alpha_{(r:n)}\big(\mathbf{Z}_{[r:n]} - \overline{\mathbf{Z}}\big) \\ \sum_{r=1}^{[ns]} \big(\mathbf{Z}_{[r:n]} - \overline{\mathbf{Z}}\big)^{\mathrm{T}}\alpha_{(r:n)} & \sum_{r=1}^{[ns]} \big(\mathbf{Z}_{[r:n]} - \overline{\mathbf{Z}}\big)^{\mathrm{T}}\big(\mathbf{Z}_{[r:n]} - \overline{\mathbf{Z}}\big) \end{pmatrix} \begin{pmatrix} \widehat{\delta} \\ \widehat{\eta} \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{r=1}^{[ns]} \alpha_{(r:n)}\big(Y_{[r:n]} - \overline{Y}\big) \\ \sum_{r=1}^{[ns]} \big(\mathbf{Z}_{[r:n]} - \overline{\mathbf{Z}}\big)^{\mathrm{T}}\big(Y_{[r:n]} - \overline{Y}\big) \end{pmatrix}.$$

For notational simplicity, let the above equation denote

$$\begin{pmatrix} S_{\alpha\alpha} & S_{\alpha\mathbf{Z}} \\ S_{\mathbf{Z}\alpha} & S_{\mathbf{ZZ}} \end{pmatrix} \begin{pmatrix} \widehat{\delta} \\ \widehat{\eta} \end{pmatrix} = \begin{pmatrix} S_{\alpha Y} \\ S_{\mathbf{Z}Y} \end{pmatrix}.$$

Then, by the inversion formula of the partitioned matrix, we have

$$\widehat{\delta} = \Big(S_{\alpha\alpha} - S_{\alpha\mathbf{Z}}S_{\mathbf{ZZ}}^{-1}S_{\mathbf{Z}\alpha}\Big)^{-1}S_{\alpha Y} - S_{\alpha\alpha}^{-1}S_{\alpha\mathbf{Z}}\Big(S_{\mathbf{ZZ}} - S_{\mathbf{Z}\alpha}S_{\alpha\alpha}^{-1}S_{\alpha\mathbf{Z}}\Big)^{-1}S_{\mathbf{Z}Y}.$$

We hereafter need to show that the consistency and asymptotic normality of estimator of $\widehat{\delta}$ like simple regression case. Then, we will perform numerical study and data analysis.

Finally, the application of the proposed score function is not restricted to linear regression but may also be appropriate for other statistical procedures based on rank, including the well-known rank aggregation problem (Breitling

et al., 2004; Eisinga et al., 2013; Lin, 2010).

Bibliography

Barnett, V., Green, P. and Robinson, A. (1976). Concomitants and correlation estimates. *Biometrika*, **63(2)**, 323-328.

Bhattacharya, P. (1974). Convergence of sample paths of normalized sums of induced order statistics. *The Annals of Statistics*, **2(5)**, 1034–1039.

Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, **573**, 83-92.

Camerer, C. (2003). The behavioral challenge to economics: Understanding normal people. Conference Series, Proceedings, 48. Paper presented at Federal Bank of Boston 48th Conference on 'How humans behave: Implications for economics and policy'.

Chernoff, H. and Savage, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics*, **29**, 972-994.

Da, Z., Engelberg, J. and Gao, P. (2011). In search of attention. *Journal of Finance*, **66**, 1461-1499.

David, H. and Galambos, J. (1974). The asymptotic theory of concomitants of order statistics. *Journal of Applied Probability*, **11(4)**, 762-770.

David, H. (2003). *Order Statistics*. John Wiley and Sons, New Jersey.

Diaconis, P. and Freedman, D. (2003). The Markov moment problem and de Finetti's theorem: Part I and Part II. Technical Report:2003-01, Department of Statistics, Stanford University.

Eisinga, R., Breitling, R. and Heskes, T. (2013). The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Letters*, **587**, 677-682.

Gertheiss, J. (2014). ANOVA for factors with ordered levels. *Journal of Agricultural, Biological, and Environmental Statistics*, **19(2)**, 258–277.

Govindarajulu, Z., LeCam, L. and Raghavachari, M. (1966). Generalizations of theorems of Chernoff and Savage on the asymptotic normality of test statistics. *Proc, Fifth Berkely Symp. Math. Statistical Probability*, **1**,609–638.

Graubard, B. and Korn, E. (1987). Choice of column scores for testing independence in ordered 2 $\times$ K contingency tables. *Biometrics*, **43**, 471-476.

Ivanova, A. and Berger, V. (2001). Drawbacks to integer scoring for ordered categorical data. *Biometrics*, **57**, 567-570.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Annals of Mathematical Statistics,* **39,** 325-346.

Hora, S. and Conover, W. (1984). The F statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, **79**, 668-673.

Huang, J. (1989). Moment problem of order statistics: A review. *International Statistical Review*, **57**, 59-66.

Kadane, J. (1971). A moment problem for order statistics. *The Annals of Mathematical Statistics*, **42**, 745-751.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall, New Jersey.

Kimeldorf, G., Sampson, A. and Whitaker, L. (1992). Min and max scorings for two-sample ordinal data. *Journal of the American Statistical Association*, **87**, 241-247.

Lin, S. (2010). Space oriented rank-based data integration. *Statistical Applications in Genetics and Molecular Biology*, **9(1)**, Article 20.

Seasholes, M. and Wu, G. (2007). Predictable behavior, profits, and attention. *Journal of Empirical Finance*, **14**, 590-610.

Senn, S. (2007). Drawbacks to noninteger scoring for ordered categorical data. *Biometrics*, **63**, 296-298.

Shorack, G. and Wellner, J. (2009). *Empirical processes with applications to statistics*, The Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wang, X., Stokes, L., Lim, J. and Chen, M. (2006). Concomitants of multivariate order statistics with application to judgment poststratification. *Journal of the American Statistical Association*, **101(476)**, 1693–1704.

Yang, S. (1981). Linear function of concomitants of order statistics with application to nonparametric estimation of a regression function. *Journal of the American Statistical Association*, **76**, 658-682.

Zheng, G. (2008). Analysis of ordered categorical data: two score-independent approaches. *Biometrics*, **64**, 1276-1279.

국 문 초 록

부분적으로 관측된 순위 공변량을 이용한 회귀분석:
순위에 대한 분포-유도 스코어 함수
Regression with Partially Observed Ranks
on a Covariate:
Distribution-Guided Scores for Ranks

본 논문은 한국의 대형 포털 사이트에서 손수 수집된 자료에 동기를 얻어 시작되었다. 포털 사이트의 온라인 메시지 게시판에서 가장 빈번하게 언급이 된 상위 30개의 주식들을 기록하였고 언급이 된 횟수는 투자자들이 주식에 기인한 주목도를 측정하기 위한 도구로 고려되었다. 자료 분석의 실증적인 목표는 주식에 대한 주목도가 투자 행위에 미치는 영향을 연구하는 것인데 이를 위해, 다음날 주식의 수익률과 부분적으로 관측이 된 언급 횟수의 순위들을 회귀분석 하였다. 회귀분석 안에서 순위들은 흔히 사용되는 항등 스코어 혹은 선형 스코어의 형태로 변환이 된다. 제안하는 새로운 종류의 스코어는 순서 통계량의 적률에 기반하고 스코어의 바람직한 특성들(단조성 혹은 볼록성)을 모델링하는데 유연한 것으로 보인다. 게다가, 실제 변량 X가 위치척도집단으로부터 추출되었고 Z가 그것들의 표준화된 분포라고 할 때 제안한 스코어를 이용하여 계산한 최소자승통계량은 반응변수와 공변량 사이의 실제 상관관계를 일관성 있게 추정하고 최소자승통계량은 점근적으로 정규분포에 근사한다. 또한, 주어진 스코어 함수를 진단하고 자료에 가장 잘 맞는 스코어 함수를 선택하기 위한 절차도 제안하였다. 상관계수를 추정하는데 있어 정확하게 명시된 스코어 함수를 사용할 때의 장점을 수치적으로 입증하고 수집된 자료를 이용하여 투자자 주목이 수익율에 미치는 영향을

검정하는데 제안한 방법들을 적용하였다.

주요어 : 수반 변수; 투자자 주목; 선형 회귀 모형; 순서통계량 적률; 부분적 관측 순위; 순위 스코어
학 번 : 2009-20242