Rank Procedures for Testing Linear Hypotheses in Repeated Measures Design

Hossein Mansouri * Danush K. Wijekularathna†

Abstract

Repeated measure data arise when for each subject a vector of observations is taken. A unique feature of the repeated measures data is the correlation structure between observations. Often it is of interest to test hypotheses concerning the parameters of a linear model for such data. The parametric models for repeated measure data have been studied extensively in the literature. Several tests based on ranks are also available.

In this expository article, we formulate a class of aligned rank tests for testing linear hypotheses for parameters of a linear model for repeated measures data.

Key Words: Dependent observations, Rank tests, Sub-hypothesis, Estimating equation, Wilcoxon score, Linear regression

1. Introduction

Repeated measures data analysis is used in a wide range of fields such as biostatistics, education, economics, psychology, agriculture, health sciences. There are several rough synonyms for repeated measures data in some fields. For instance, the term ”panel data” is more common in economics, the term ”longitudinal data” is most commonly used in biostatistics, and ”repeated measures” is more common in an agricultural framework.

Even though the term ”longitudinal data” is a rough synonym for ”repeated measures”, there are sometimes slight differences in the meaning of these terms. Longitudinal data occur when we repeatedly take the same type of measurement across time on the subjects in a study. Repeated measures are also multiple measurements on each of several individual subjects, but they are not necessarily through time. However, the terms ”repeated measures” and ”longitudinal data” are used interchangeably.

Since we repeatedly take the same type of measurement across time on the subjects, data are not independent. So the major problem of this model is covariance structure among these repeated measurements and we must account for the dependency in data using more complex or complicated statistical methods. Unfortunately, the appropriate analytical methods are not much developed yet(Hedeker, D., and Gibbons, 2006) .

The problem of sub hypothesis testing in repeated measures models based on rank statistics has received a considerable interest in statistical literature. Diggle,Heagerty,Lian and Zeger (2013) proposed a test statistic for the sub hypotheses testing on repeated model with multivariate normal error structure.Although the parametric methods are robust against misspecification of the dependence structre, it is not robust against outliers.

Barefield(2001) described two testing procedures in his dissertation, the Wald-type test based on the rank estimate of the parameters and the aligned rank test.The aligned rank method was developed in the linear model for independent data, by Chiang and Puri(1984).The aligned rank test statistic proposed by Barefield has a limiting χ^2 distribution under H_0. Since he used the general score, this method can be applied for only homogeneous error distributions. Further, he discussed the method assuming that each

*First author's affiliation, Dept. of Mathematics and Statistics, Texas Tech University, Lubbock,TX 79413
†Second author's affiliation,Dept.of Mathematics and Statistics, Texas Tech University, Lubbock,TX 79413

subjects has equal number of observations. Our proposed test statistics is a development of the Barefield's test statistics. To estimate regression parameters, he extended the results of Brunner and Denker (1994) to the case of longitudinal data that can handle quantitative regressors. He define a linear rank statistics, which use the general score to estimate regression parameters.

Jung and Ying (2003) generalized the classical Wilcoxon-Mann-Whitney method, proposing an independent working model for parameter estimation. They proposed a rank-score test for testing hypotheses about the entire parameters of a linear model.

Kloke, Mckean and Rashid(2009) extended Jaeckel's (1972) ordinary rank estimator to models with dependent block error structure. They consider the same model as Jung and Ying but homogeneous distributions and general rank score functions. Their estimators are based on the joint ranking of all residuals.They studied two test statistics to test linear hypotheses. The first test is the Wald-type test previously studied by Barefield (2001) for balanced data. A second test utilizes the reduction in dispersion.Even though proposed two test statistics can be used for the problems with unequal observation for each subjects, it can't be applied for the problems with heterogeneous error structure.

In this expository article we present a summary formulation of a test statistic for testing sub-hypotheses for the longitudinal model based on the aligned rank score following the model considered similar to Jung and Ying (2003).Proposed test statistic can be used for any repeated data design with missing observations, heterogeneous error distributions, unequal measurement times for different subjects, unequal number of observations for each subject and any covariance structure of the measurement of the same subjects.

2. Rank Regression

Consider the linear model $Y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \epsilon_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq k_i$. Y_{ij} is the response, $\mathbf{x_{ij}}$ is a $p \times 1$ vector of explanatory variables, and ϵ_{ij} is the error term for the i-th subject at the j-th time point. Note that there are a total of $N = \sum_{i=1}^{n} k_i$ observations.The vector $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters. Let $\mathbf{Y}_i = [Y_{i1}, ..., Y_{ik_i}]^{'}$, $\mathbf{X}_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{ik_i}]^{'}$, $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, ..., \epsilon_{ik_i}]^{'}$ and let $\mathbf{Y} = [\mathbf{Y}_1^{'}, ..., \mathbf{Y}_n^{'}]^{'}\mathbf{X} = [\mathbf{X}_1^{'}, ..., \mathbf{X}_n^{'}]^{'}$,and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1^{'}, ..., \boldsymbol{\epsilon}_n^{'}]^{'}$.we can write the model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

For $\mathbf{b} \in \Re^p$ Jung and Ying(2003) use Jackel's dispersion function with uniform scores

$$L_n(\mathbf{b}) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}) R_{ij}(\mathbf{b}) \tag{2}$$

where $\overline{\mathbf{x}} = N^{-1} \sum \sum \mathbf{x}_{ij}$ and $R_{ij}(\mathbf{b})$ is the rank of $(Y_{ij} - \mathbf{x}_{ij}^{'}\mathbf{b})$ among the $Y_{11} - \mathbf{x}_{11}^{'}\mathbf{b}, ..., Y_{nk_n} - \mathbf{x}_{nk_n}^{'}\mathbf{b}$, and $\mathbf{b}$ is a fixed $p \times 1$ vector.

We use the basic assumptions of Rank statistics. ε_{ij} and $\varepsilon_{ij^{'}}$ will in general be dependent for each i though ε_{ij} and $\varepsilon_{i^{'}j^{'}}$ are always independent if $i \neq i^{'}$.

2.1 The Aligned Rank Test for Subhypotheses

Without loss of generality, We can consider the following representation of the linear model of the model (1)

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{3}$$

We also partition $\mathbf{L}_{(n)}(\mathbf{b})$ as,

$$\mathbf{L}_n(\mathbf{b}) = (\mathbf{L}_{n(1)}(\mathbf{b}), \mathbf{L}_{n(2)}(\mathbf{b}))$$

Where $\mathbf{L}_{n(1)}(\mathbf{b}) = (L_{n1}(\mathbf{b}), ..., L_{nr}(\mathbf{b}))$, $\mathbf{L}_{n(2)}(\mathbf{b}) = (L_{n(r+1)}(\mathbf{b}), ..., L_{np}(\mathbf{b}))$.

Assume that we are interested in testing

$$H_0 : \boldsymbol{\beta}_2 = 0 \;\; vs \;\; H_A : \boldsymbol{\beta}_2 \neq 0$$

where $\mathbf{X}_1$, $\mathbf{X}_2$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $N \times (p-r)$, $N \times r$, $(p-r) \times 1$, and $r \times 1$, respectively and $\boldsymbol{\beta}_1'$ is unspecified under both hypotheses. Let $\hat{\mathbf{L}}_{n(2)}$ be the last $p-r$ components of $\mathbf{L}_n(\hat{\boldsymbol{\beta}}_1)$.

Then the proposed test statistic for testing H_0 against H_A is given by

$$L_W = \hat{\mathbf{L}}_{n(2)}' \hat{C}_n^{-1} \hat{\mathbf{L}}_{n(2)}$$

where;

$$\hat{C}_n = (\hat{V}_{n,22} - \hat{V}_{n,12}' \hat{A}_{n,11}^{-1} \hat{A}_{n,12} + \hat{A}_{n,12}' \hat{A}_{n,11}^{-1} (\hat{V}_{n,11} \hat{A}_{n,11}^{-1} \hat{A}_{n,12} - \hat{V}_{n,12}))$$

Where $\hat{V}_{n,ij}$'s and $\hat{A}_{n,ij}$'s are estimators of $V_{n,ij}$'s and $A_{n,ij}$'s defined as follows.

$$A_n = \int_{-\infty}^{\infty} \mathbf{A}_{n,F}(t) dG_n'(t) - \int_{-\infty}^{\infty} \mathbf{A}_{n,G}(t) dF_n'(t) = \begin{bmatrix} A_{n,11} & A_{n,12} \\ A_{n,12}' & A_{n,22} \end{bmatrix}$$

where

$\mathbf{A}_{n,F}(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \mathbf{x}_{ij} f_{ij}(t)$

$\mathbf{A}_{n,G}(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \bar{\mathbf{x}})' f_{ij}(t)$

$\hat{F}(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} I(\varepsilon_{ij} \leq t)$

$F_n(t) = E(\hat{F}(t))$

$\hat{G}(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}) I(\varepsilon_{ij} \leq t)$

$G_n(t) = E(\hat{G}(t))$,

$G_n'(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})' F_{ij}(t)$,

$F_n'(t) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} F_{ij}(t)$

and

$$V_n = Var(\mathbf{S}_n(\mathbf{0})) = N^{-1} \sum_{i=1}^{n} \boldsymbol{\xi}_i \boldsymbol{\xi}_i' = \begin{bmatrix} V_{n,11} & V_{n,12} \\ V_{n,12}' & V_{n,22} \end{bmatrix},$$

where

$$\mathbf{S}_n(\mathbf{b}) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}) F_{ij}(Y_{ij} - \mathbf{b}' \mathbf{x}_{ij})$$

$$\boldsymbol{\xi}_i = \sum_j [(\mathbf{x}_{ij} - \overline{\mathbf{x}})\{F_n(\varepsilon_{ij}) - \frac{1}{2}\} - G_n(\varepsilon_{ij})]$$

Theorem.
The proposed test statistics L_W has a limiting chi-squared distribution with $p - r$ degrees of freedom. That is,

$$L_W \to \chi^2(p - r), \quad as \;\; n \to \infty$$

REFERENCES

Barefield, E. Rank regression in longitudinal data analysis. Dis- sertation, Texas Tech University, https://repositories.tdl.org/ttu- ir/bitstream/handle/2346/15877/31295017220483.pdf?sequence=1, 2001.

Brunner, E., and Denker, M. Rank statistics under dependent observations and applications to factorial designs. Statistical Planning and Inference 42, 3 (1994), 353-378.

Chiang, C.-Y., and Puri, M. M. Rank procedures for testing subhypotheses in linear regression. Ann. Inst. Statist. Math. 36, 1 (1984), 3550.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. Analysis of Longitudinal Data. Oxford, 2013.

Hedeker, D., and Gibbons, R. D. Longitudinal data analysis. John Wiley & Sons, 2006.

Jaeckel, L. A. Estimating regression coefficients by minimizing the dispersion of the residuals. Annals of Mathematical Statistics 43, 5 (1972), 1449-1458.

Jung, S.-H., and Ying, Z. Rank-based regression with repeated measurements data. Biometrika 90, 3 (2003), 732-740.

Kloke, J. D., McKean, J. W., and Rashid, M. M. Rank-based estimation and associated inferences for linear models with cluster correlated errors. J. Amer. Statist. Assoc. 104, 485 (2009), 384-390.