DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 28

5-1-2008

Robust General Linear Models and Graphics via a User Interface (Web RGLM)

Kimberly Crimin
Wyeth Research

Asheber Abebe
Auburn University

Joseph W. McKean
Western Michigan University, joseph.mckean@wmich.edu

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 318-330

1538 – 9472/05/$95.00

Statistical Software Applications & Review

Robust General Linear Models and Graphics via a User Interface (Web RGLM)

Kimberly Crimin
Wyeth Research

Asheber Abebe
Auburn University

Joseph W. McKean
Western Michigan University

Rank-based procedures provide superior estimation and testing techniques when the data deviate from normality or contain gross outliers. However, these robust techniques are rarely incorporated in a nonparametric statistics or methods courses due to the lack of computational tools. One reason for this is the existence of certain unavoidable complexities in the numerical methods due to the absence of a closed-form solution for the rank estimation problem. This article introduces a user interface, Web RGLM, which may be used to perform rank-based analyses of linear models across the World Wide Web. These models include simple location problems to complicated ANOVA and ANCOVA designs with multiple comparison procedures. The robust and least squares analyses are presented side-by-side for immediate comparisons. Web RGLM meets many of the computational demands of the classroom as well as the computational demands of quantitative researchers. Several illustrative examples are provided.

Key words: R-estimation, RGLM, rank-based procedures, least squares, analysis of covariance

Introduction

Recent work on rank-based procedures for linear models has brought together a unified analysis of linear models analogous to the traditional analysis based on least squares. The rank-based analysis includes estimation, confidence procedures, testing of general linear hypotheses, and diagnostic methods. These rank-based analyses generalize the classical nonparametric rank procedures for one and two sample location problems and they inherit the robustness and high efficiency of these simple methods. The recent article by McKean (2004) reviews this analysis while the monograph by Hettmansperger and McKean (1998) presents a thorough discussion of these rank-based analyses. Chapter 9 of the second edition of Hollander and Wolfe (1999) also offers a recent discussion of these methods. In Section 4, we give a quick overview of the rank-based analyses that are on our web page.

Traditional least squares analyses are based on estimation by minimizing the Euclidean (squared) norm, while the rank-based procedures are based on the minimization of a different norm. The minimization of this norm is a benign numerical problem which can be handled by existing numerical methods. However, to be of practical use these procedures must be easily computed. In this article we present an easy-to-use web version of these rank-based procedures. It allows the user to 'point-and-click' to perform these analyses for simple location problems through complex experimental designs. The output offers numerical results and diagnostic plots, produced by the R language; see Ihaka and Gentleman

Kimberly Crimin is a Senior Principal Biostatistician in a Non-clinical Statistics group at Wyeth. She also works with robust procedures for discriminant analysis. Asheber Abebe is Associate Professor of Statistics at Auburn University. His research interests include robust regression and classification and exact simultaneous inference procedures. Joseph McKean is a Professor of Statistics. His research areas include nonparametric and robust statistics, along with computational algorithms for these procedures. He has co-authored three books in statistics. E-mail him at: joseph.mckean@wmich.edu

(1996). Another advantage of the output is that it offers side-by-side comparisons of the robust and least squares (LS) analyses. If the analyses disagree then the user may choose to explore the data to determine reasons for this disagreement. This side-by-side comparison also serves as a very useful teaching tool. For instance, the student can immediately see the impact that perturbations of the data have on the LS and robust analyses.

Our web-based version of these analyses is discussed. Several examples are provided. It is found http://fisher.stat.wmich.edu/slab/RGLM/.

Web-Based RGLM

RGLM, (Robust General Linear Model), is the name of the FORTRAN program that performs the robust general linear model estimation and hypotheses testing described in Section 5. It was developed by Kapenga, McKean, and Vidmar (1988), and follows algorithms listed below. For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the package RGLM returns a robust fit of this model. To make this program accessible to researchers, scientists and students, a web interface to RGLM was created. All the analyses discussed in this article were obtained using the Web based RGLM, which is available at http://fisher.stat.wmich.edu/slab/RGLM/.

The web interface to RGLM is module driven. Each module represents a different linear model that can be run using Web RGLM. Figure 1 is a screen capture of the home page for RGLM.

Note that many of the usual designs are given as options, from simple location models to complicated crossed factorial designs. When a user clicks on the name of the desired linear model (see Figure 1), a form is returned which allows the user to input the data and further customize the desired analysis. Some of the analysis options are: residuals and studentized residual plots, data plots appropriate for the model, contrasts along with type of interval, and type of scores used to estimate cell location (either Wilcoxon or signed-rank Wilcoxon). Each data input page describes the format of the input data set and contains an example data set. Data may be directly typed into the data entry window or entered as a file that resides on the user's machine.

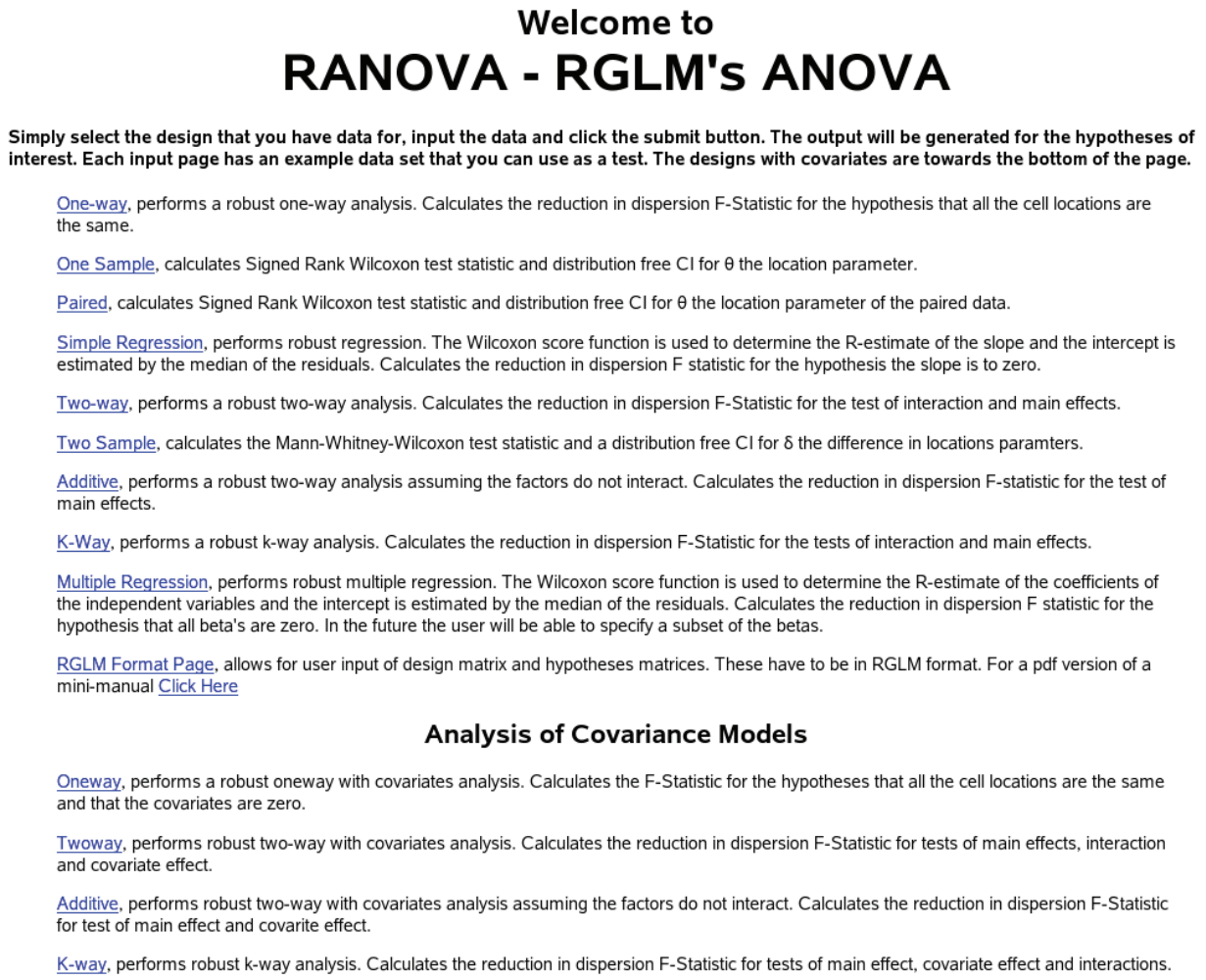
The freeware R (Ihaka & Gentleman, 1996) is used to produce the residual and data plots. Clicking the "Submit Data" button will result in a run of the desired analysis with the selected options. Clicking the "Clear Form" button will result in a default form and an empty data window.

For each module, both the traditional and rank-based analyses are provided. This summary has served as a useful teaching tool in applied nonparametric courses and methodology courses, in general. For a given data set, students can easily see if there is a difference in the analyses. In the case where the analyses differ, students can then try to determine why they differ by using residual plots and exploring the data to see if the discrepancy is caused by outliers or decidedly non-normal data, etc. It forces them to decide which analysis, if any, to use. Further, students can easily see how sensitive the robust and traditional analyses are by changing data points. For example, consider a one-sample problem. By repeatedly changing a data point, in a few seconds the student can have the data base to do comparison sensitivity plots of the Hodges-Lehmann estimator and the sample average.

The Web version of RGLM will run on any browser that is compatible with forms and, if the user selected residual or data plots, with graphics. All of the computations are done on the side of the server, reducing the hardware and software requirements of the user and ensuring uniformity of the output.

Behind the Scenes

The Web version of RGLM is a collection of CGI scripts, written in Perl (see Srinivasan (1997)), UNIX shell and FORTRAN programs. The statistical software R is used to obtain the user selected plots. The home page for Web RGLM and the input page for each linear model exist as separate HTML documents. The HTML page displaying the output is created by



Welcome to

RANOVA - RGLM's ANOVA

Simply select the design that you have data for, input the data and click the submit button. The output will be generated for the hypotheses of interest. Each input page has an example data set that you can use as a test. The designs with covariates are towards the bottom of the page.

One-way, performs a robust one-way analysis. Calculates the reduction in dispersion F-Statistic for the hypothesis that all the cell locations are the same.

One Sample, calculates Signed Rank Wilcoxon test statistic and distribution free CI for θ the location parameter.

Paired, calculates Signed Rank Wilcoxon test statistic and distribution free CI for θ the location parameter of the paired data.

Simple Regression, performs robust regression. The Wilcoxon score function is used to determine the R-estimate of the slope and the intercept is estimated by the median of the residuals. Calculates the reduction in dispersion F statistic for the hypothesis the slope is to zero.

Two-way, performs a robust two-way analysis. Calculates the reduction in dispersion F-Statistic for the test of interaction and main effects.

Two Sample, calculates the Mann-Whitney-Wilcoxon test statistic and a distribution free CI for δ the difference in locations paramters.

Additive, performs a robust two-way analysis assuming the factors do not interact. Calculates the reduction in dispersion F-statistic for the test of main effects.

K-Way, performs a robust k-way analysis. Calculates the reduction in dispersion F-Statistic for the tests of interaction and main effects.

Multiple Regression, performs robust multiple regression. The Wilcoxon score function is used to determine the R-estimate of the coefficients of the independent variables and the intercept is estimated by the median of the residuals. Calculates the reduction in dispersion F statistic for the hypothesis that all beta's are zero. In the future the user will be able to specify a subset of the betas.

RGLM Format Page, allows for user input of design matrix and hypotheses matrices. These have to be in RGLM format. For a pdf version of a mini-manual Click Here

Analysis of Covariance Models

Oneway, performs a robust oneway with covariates analysis. Calculates the F-Statistic for the hypotheses that all the cell locations are the same and that the covariates are zero.

Twoway, performs robust two-way with covariates analysis. Calculates the reduction in dispersion F-Statistic for tests of main effects, interaction and covariate effect.

Additive, performs robust two-way with covariates analysis assuming the factors do not interact. Calculates the reduction in dispersion F-Statistic for test of main effect and covarite effect.

K-way, performs robust k-way analysis. Calculates the reduction in dispersion F-Statistic for tests of main effect, covariate effect and interactions.

Figure 1: RGLM Home Page

the CGI script once RGLM has executed. This section provides a brief overview of the behind the scenes workings of Web RGLM.

RGLM is the main FORTRAN program that performs the robust analysis. RGLM requires three input files of a specific format. One file contains options for the rank-based analysis, another file contains the $\boldsymbol{X}|\boldsymbol{Y}$ augmented matrix, where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{Y}$ is the data matrix, and the third file contains the hypotheses matrices. To shorten the learning curve for the user and to make Web RGLM 'point-and-click', a FORTRAN program creates the three RGLM input files from the information provided by the user in the data input page.

The use of a FORTRAN program to create the input files also provides some security for the server, since the form only sends data and options to the CGI program and not commands. Within the CGI program, the data is checked to make sure it only contains digits. If characters other than digits are found, then an error page is returned to the user indicating an

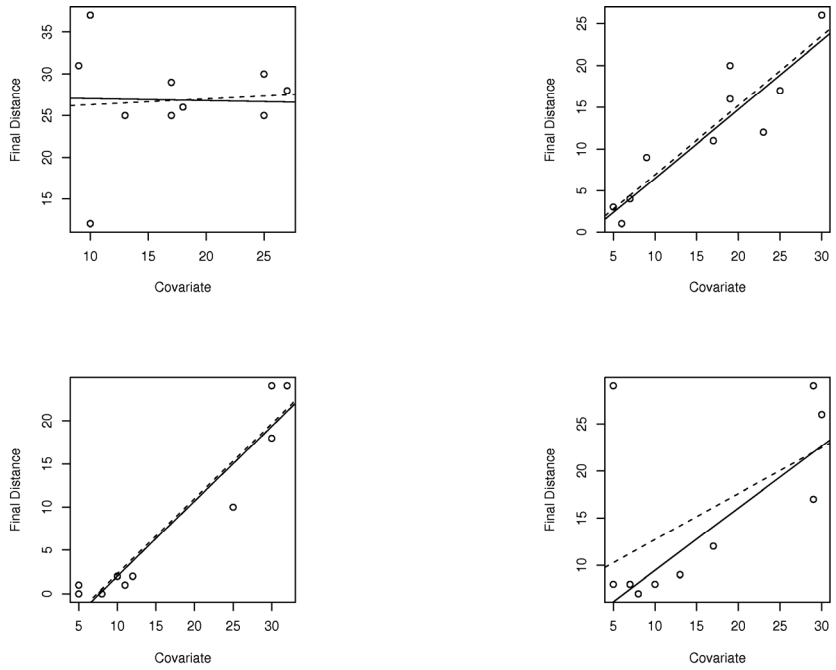


Figure 2: Covariate vs. Final Distance by Treatment. The solid line is the robust fit, while the dashed line is the LS fit. Plots are for: Control, Upper Left Panel; Treatment 1, Upper Right Panel; Treatment 2, Lower Left Panel; and Treatment 3, Lower Right Panel.

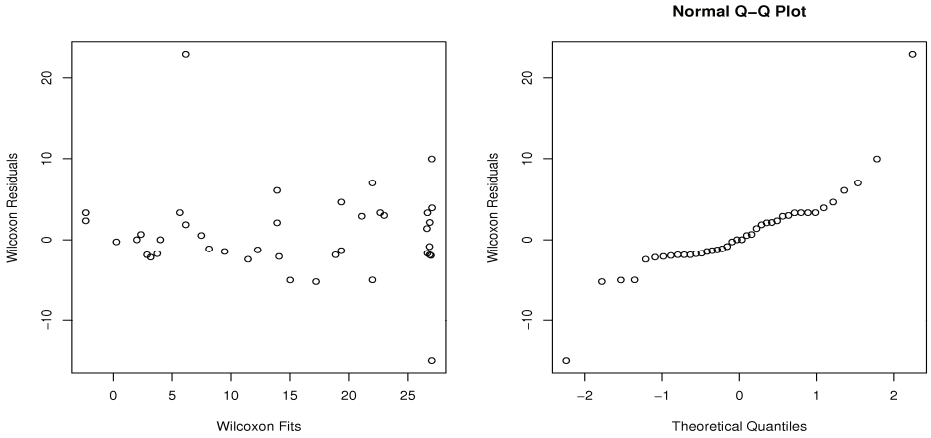


Figure 3: Residual Plots

error in the data input file. The FORTRAN program that creates the RGLM input files is the only program that is run with a system call that uses input provided by the user. All other system calls are executed on data files created by this FORTRAN program. This strategy limits the number of doors left open to the server.

The RGLM program allows the user to specify the name of the input file containing the augmented $X|Y$ matrix and the hypothesis matrix, but does not allow the user to specify the names of the output files. To allow multiple users to run Web RGLM at the same time without clobbering each other, each user is assigned a user ID. The user ID maps to a temporary directory and all files created for that run are stored in the temporary directory. After the HTML page containing the output is returned, all files in the temporary directory are removed along with the temporary directory. If a user runs multiple analyses in the same web session, a temporary directory is created and removed on each run. An earlier version of Web RGLM stored the user ID as a cookie. In this previous version, the temporary directories were removed after a prescribed length of time. This caused unwanted complications when a web session exceeded the allowable time.

When a user selects data plots or residuals plots, the CGI program writes the R code to create the plots to a file. Then R is run in batch mode, producing a postscript file containing the plots. To display the plots in an HTML page, the postscript file is converted to a gif file using Netpbm graphics utilities available at sourceforge.net/projects/netpbm.

Examples

Using the Web version of RGLM we offer three illustrative examples of the rank-based analysis, comparing it with the traditional Least Squares analysis in each case. We use the default Wilcoxon scores. These scores are based on a linear score function (see Section 5.2) and for the one and two-sample location problems these scores result in the usual Mann-Whitney-Wilcoxon analyses. They require no tuning constants. Other scores can be used, as briefly cited in Section 4.

Analysis of Covariance: Snake Data Set

The dataset used for this example was discussed in Afifi and Azen (1972). The purpose of the experiment was to compare methods of reducing human's fear of snakes. There are three methods intended to reduce ones fear of snakes and one control, or placebo. Forty subjects were randomly assigned to the four treatments. To measure ones fear of snakes, a behavior approach test was used to determine how close one could walk towards a snake without feeling uncomfortable. The behavior approach test was given to each subject before and after treatment. The score on the before treatment test was taken as a covariate.

To obtain the rank-based analysis of this data set using Web RGLM proceed as follows: from the home page, click on "Oneway" under Analysis of Covariance Models (see Figure 1) and drop the data and covariate into the data boxes. For this analysis, we included covariate by treatment interaction in the model and used cell medians as the estimates of location.

There are several options for plots available to the user. For the analysis below, we requested covariate versus response by treatment, residuals versus fitted values and a normal q-q plot of the residuals. These plots are shown in Figure 2 and Figure 3, respectively. The residual plot indicates that the data are heteroscedastic which can be eliminated by the square root transformation applied to the response variable.

Figure 4 contains the analysis part of the output from Web RGLM. It is clear from the plots of the response, final distance, by treatment, Figure 2, that the treatment slope parameters are not the same. The comparison analyses show that the robust F-test for parallelism detects this difference with a p value of 0.01, but that the LS F-test with p value is 0.09 fails to detect this difference at the 5% level. Based on the q-q plot of residuals, Figure 3, the underlying error structure appears to be heavy tailed, so the difference in the analyses is not surprising.

One-Way Analysis: Creatine Data Set

For our second example we have chosen a data set from a pharmaceutical study. The data

RGLM's ANOVA Output

One Way with Covariates Output

Estimates of Cell Location and Standard Errors

	Wilcoxon Rank Based tau-hat=3.94311				Cell Means sigma-hat=5.81833			
	Estimate	SE	t-ratio	p-value	Estimate	SE	t-ratio	p-value
FinalDistance-1	27.2819	3.62576	7.52447	1.44004e-08	25.6149	5.26431	4.86577	2.93435e-05
FinalDistance-2	-1.78198	2.77338	-0.64253	0.525106	-1.38736	3.97956	-0.34862	0.729658
FinalDistance-3	-6.67944	2.43761	-2.74015	0.00995914	-6.39363	3.46804	-1.84358	0.0745225
FinalDistance-4	2.85244	2.39913	1.18895	0.243208	7.80823	3.40912	2.2904	0.0287341
Cov*FinalDistance-1	-0.0242747	0.195477	-0.124182	0.901948	0.0693046	0.288439	0.240274	0.81165
Cov*FinalDistance-2	0.825997	0.149463	5.52643	4.29335e-06	0.83046	0.220543	3.76552	0.000673494
Cov*FinalDistance-3	0.867943	0.118588	7.319	2.5455e-08	0.868668	0.174984	4.96427	2.20457e-05
Cov*FinalDistance-4	0.660261	0.127124	5.19382	1.13076e-05	0.489658	0.187581	2.61039	0.0136462

Coefficients of Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.797625	0.76295

Hypothesis: Cell Locations are the same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
FinalDistance effect	12.662	3,32	1.26958e-05	9.62872	3,32	0.000111222

Hypothesis: Slopes are the Same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
Cov*FinalDistance effect	4.06897	3,32	0.0147894	2.33664	3,32	0.0922737

Figure 4: Screen Capture of Rank-Based Analysis of Snake Data

set contains the results of an experiment that was run on mice to determine the effects of different doses of an experimental compound on the amount of creatine cleared from the body. The mice were randomly divided into six groups. The first group formed the control which had a dose level of 0 of the compound. The other five groups each had a different dose of the compound. The data have been corrected for the body weights of the mice. Thus the appropriate design is a one-way design. Besides the test of an overall effect, it was of interest to compare the five groups to the control. On the RGLM page, "Oneway" was selected.

One of the plots we checked on the form was the comparison boxplots of the levels which is shown in Figure 5. Besides the apparent outliers, this plot indicates that all the treatment levels may be significantly different from the placebo.

For the contrast query on the RGLM one-way page, we checked *versus control* and entered the level (1) for the control. We selected the Tukey-Kramer multiple comparison procedure, (MCP). As shown in Figure 6, the Wilcoxon ANOVA detects these differences, the *F*-statistic has the value 7.24 with *p*-value 0.00001. In contrast, note that the LS *F*-statistic has *p*-value 0.056. The outliers impaired its

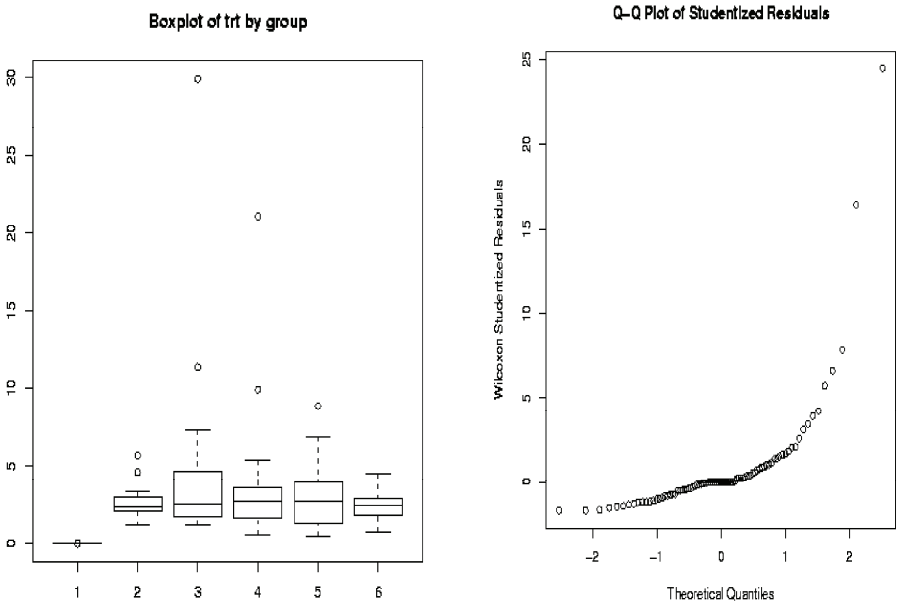


Figure 5: Comparison Boxplots of the observations by level and *q-q* plot of Wilcoxon Studentized Residuals for Creatine Data

Hypothesis: Cell Locations are the same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
trt effect	7.23668	5,80	1.23382e-05	2.26084	5,80	0.0561732

Tukey-Kramer

	Wilcoxon R Contrasts			Least Squares Contrasts		
Contrast	Estimate	Error Term	Interval	Estimate	Error Term	Interval
2 1	2.36633	1.75528	(0.61105,4.12161)	2.75157	5.16927	(-2.4177,7.92083)
3 1	2.5935	1.64005	(0.953452,4.23355)	5.17179	4.82991	(0.341872,10.0017)
4 1	2.42003	1.601	(0.819034,4.02103)	3.92346	4.7149	(-0.791448,8.63836)
5 1	2.37465	1.601	(0.773654,3.97565)	3.0671	4.7149	(-1.64781,7.782)
6 1	2.34548	1.61942	(0.72606,3.9649)	2.463	4.76916	(-2.30617,7.23216)

Figure 6: Wilcoxon and Least Squares ANOVAs for Creatine Data

power. The table in Figure 6 summarizes the MCP study. For the Wilcoxon analysis, the Tukey-Kramer procedure declares that all five levels differ significantly from the control while the LS version of the Tukey-Kramer procedure only declares that the third level differs significantly from the control.

Multiple Regression: Snow Geese Data Set

In this example, we consider the snow geese data set discussed on page 441 of Hollander and Wolfe (1999). It is a multiple regression problem with four predictors. The response is the time, minutes before (-) or after (+) sunrise, that lesser snow geese leave their overnight roost sites to fly to their feeding areas. The predictors are: x_1, the air temperature in Celsius; x_2, relative humidity; x_3, light intensity; and x_4, percent cloud cover. Data were collected for *n*=36 days. We assume the linear model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, \quad i = 1, 2, ..., 36. \qquad (1)$$

Besides estimating the regression coefficients, the following two hypotheses are of interest:

$$H_{01} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \qquad (2)$$

$$H_{02} : \beta_1 = \beta_2 = 0 \qquad (3)$$

Hollander and Wolfe used the *rregr* command of Minitab to perform this analysis. We show how it is easily performed by the RGLM web page.

On the web page (Figure 1), click on "Multiple Regression". Next, drop in the data in the form **X** *Y* into the data box. The user has a choice on the estimate of the intercept, either the median of the residuals or the Hodges-Lehmann estimate of location based on the residuals. The hypothesis H_{01} is the usual regression hypothesis that all regression coefficients are zero, except for the intercept. Web RGLM always obtains the robust and LS tests for this hypothesis. For the second hypothesis, H_{02}, the reduced model is $Y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$. To obtain the test of H_{02}, indicate that this reduced model is to be fit by entering 3 and 4 into the box labeled "Enter column ids, *1* to *p*, to include in the reduced model."

Figure 7 shows the output. The full model estimates are given in the first table. The robust and LS fits are similar, except for the estimate of β_1 in which the fits differ by about a half of a standard error. This may have been caused by the one outlier in the data set as seen in the residual and *q*-*q* plots of the robust fit as shown in Figure 8. The tests that all regression coefficients are $0, H_{01}$, are given in the third table, while the tests of H_{02} are given in the last table. This later hypothesis concerns dropping β_1. As with the estimate of β_1, the robust *F* test is more significant than the LS *F* test.

Conclusion

The statistical computation tool introduced in this article uses state-of-the-art web interfacing to provide users access to robust nonparametric methods. In addition to the traditional ASCII text output provided by RGLM, Web RGLM provides graphics for visual assessment of the data and model diagnostics. Graphics associated with rank-based procedures have customarily been produced using other statistical software after the output from RGLM is manually edited. With the web interface available, this cumbersome activity is now unnecessary. Moreover, the user is not limited to specific score functions. The RGLM Format page gives the user the option of choosing a score function, in addition to several other options, thus, retaining the flexibility of RGLM. There is an online manual describing customized analyses which the user can download.

There is an experimental companion to Web RGLM that uses high breakdown (HBR) techniques. This can be found at the URL: http://fisher.stat.wmich.edu/slab/RGLM/HBR2. As with the Web RGLM page, it offers side-by-side comparisons of the high breakdown and LS fits. These techniques, developed by Chang et al. (1999), use a stochastically weighted Wilcoxon norm to obtain estimators that are robust to outliers in both design and response space, while the Wilcoxon analysis is only robust in response space. We plan on finishing this page in the future. Also, we are planning future expansions of the page to other designs, including nested designs, generalized estimating equations, nonlinear models, and mixed models.

Multiple Regression Output

Parameter Estimates and Standard Errors

	Wilcoxon R tau-hat=8.42223				Least Squares sigma-hat=8.09161			
	Estimate	SE	t-ratio	p-values	Estimate	SE	t-ratio	p-values
Intercept	-51.4134	9.19765	-5.58984	3.95836e-06	-52.9939	8.78728	-6.03076	1.12429e-06
Beta 1	1.03877	0.275392	3.77196	0.000685556	0.912981	0.264581	3.45066	0.00163517
Beta 2	0.126197	0.118441	1.06549	0.294883	0.142532	0.113791	1.25257	0.219729
Beta 3	2.53563	0.781934	3.24277	0.00283114	2.516	0.751238	3.34914	0.00214101
Beta 4	0.089625	0.0457174	1.96041	0.058988	0.0922053	0.0439228	2.09926	0.0440324

Coefficients of Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.692526	0.758723

Hypothesis: Model does not fit

Wilcoxon R			Least Squares		
F-Statistic	df	p-value	F-Statistic	df	p-value
17.4554	4,31	1.19209e-07	24.3707	4,31	0

Reduced Model Parameter Estimates and Standard Errors

	Wilcoxon R tau-hat=9.82016				Least Squares sigma-hat=9.29604			
	Estimate	SE	t-ratio	p-values	Estimate	SE	t-ratio	p-values
Intercept	-54.931	6.89528	-7.96646	2.78399e-09	-55.1992	6.53087	-8.45205	7.18156e-10
Beta 3	4.11602	0.621827	6.61924	1.3675e-07	4.37963	0.588639	7.44026	1.24672e-08
Beta 4	0.177072	0.0446428	3.96641	0.000356513	0.176112	0.0422602	4.16734	0.000199932

Reduced Model Coefficients of Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.595624	0.661004

Hypothesis: User Specified Reduced Model

Wilcoxon R			Least Squares		
F-Statistic	df	p-value	F-Statistic	df	p-value
6.57319973451212	2,31	0.00417089	6.27759176687857	2,31	0.0051403

Figure 7: RGLM's ANOVA Output for the Snow Geese Data

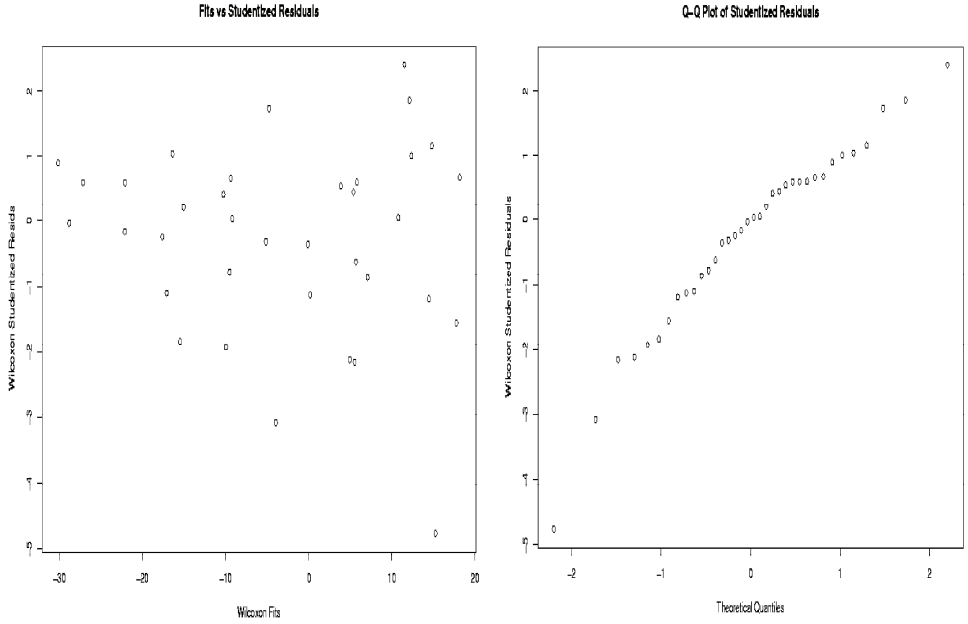


Figure 8: Wilcoxon studentized residual and *q*-*q* plots and LS ANOVAs for the Snow Geese Data

Background

Just like the traditional least squares procedures, rank-based procedures give a unified approach to testing and estimation problems. The recent monograph by Hettmansperger and McKean (1998) ('HM' hereinafter) gives a detailed treatise of rank-based procedures for handling problems of estimation and testing in situations ranging from the simple one sample location problems to the highly complicated multi-factor experimental designs. In this section we briefly review rank-based estimation and testing procedures and direct the reader to HM for further details.

Linear Models

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ denote the $n \times 1$ vector of observations which follows the linear model

$$\mathbf{Y} = \mathbf{1}\alpha + \boldsymbol{\eta} + \mathbf{e},\ \boldsymbol{\eta} \in \Omega \qquad (4)$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones, Ω is a subspace of $\mathbf{R}^n$ spanned by the columns of a centered $n \times p$ design matrix $\mathbf{X}$, and $\mathbf{e}$ is an $n \times 1$ vector of random errors.

In addition to estimating α and $\boldsymbol{\eta}$ we test general linear hypotheses such as

$$H_0 : \boldsymbol{\eta} \in \omega \text{ versus } H_A : \boldsymbol{\eta} \in \Omega \cap \omega^{\perp} \quad (5)$$

where $\omega \subset \Omega$ is $p - q$ dimensional for $0 \le q \le p$. In the following we shall refer to the model given in (4) as the full model and the same model under H_0 as the reduced model.

R-Estimation

The estimate of $\boldsymbol{\eta}$ will be obtained by minimizing the distance between $\mathbf{Y}$ and the space Ω. The distance we minimize for *R*-estimation is based on the *R* pseudonorm defined as

$$\|\mathbf{u}\|_{\varphi} = \sum_{i=1}^{n} a(R(u_i))u_i,\ \mathbf{u} \in \mathbf{R}^n \qquad (6)$$

where $R(u_i)$ denotes the rank of u_i among $u_1, \ldots, u_n$, and $a(i) = \varphi(i/(n+1))$ for some nondecreasing score function φ defined on the interval (0,1) and standardized such that

$\int \varphi = 0$ and $\int \varphi^2 = 1$. . For the proof that (6) is indeed a pseudonorm the reader is referred to McKean and Schrader (1980). The set $\{a(1),a(2),\ldots,a(n)\}$ is called the set of rank scores. The most common *R* scores used in practice are the Wilcoxon scores which are generated by $\varphi(u) = \sqrt{12}(u - 0.5)$; i.e a linear score function. In the simple location models, the rank-based analyses based on this score function are the Mann-Whitney-Wilcoxon procedures. The L_1 pseudonorm is another popular special case of (6) obtained when $\varphi(u) = \text{sgn}(u - 0.5)$. In the location cases, analyses based on the sign scores are the median (Mood) procedures.

The *R*-estimator of $\boldsymbol{\eta}$ is a vector $\hat{\mathbf{Y}}_\varphi$ such that

$$\left\|\mathbf{Y} - \hat{\mathbf{Y}}_\varphi\right\|_\varphi = \min_{\boldsymbol{\eta}\in\Omega}\left\|\mathbf{Y} - \boldsymbol{\eta}\right\|_\varphi \equiv D(\Omega) \quad (7)$$

The *R*-estimates are analogous to the least squares estimates. Suppose we use the Euclidean norm $\|\mathbf{u}\|_{LS}^2 = \sum (u_i - \bar{u})^2$. There the estimator is, of course, $\hat{\mathbf{Y}}_{LS} = \mathbf{HY}$ where $\mathbf{H} = \mathbf{X(X'X)}^{-1}\mathbf{X'}$ is the projection matrix onto the column space of the centered design matrix $\mathbf{X}$. To obtain the *R*-estimates we simply replace the Euclidean norm by the norm given in (6).

Estimation of Regression Coefficients

Rewriting (4) as $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta}$ is a $p\times 1$ vector, the *R*-estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_\varphi$, is the solution vector of the *p* normal equations

$$\mathbf{X}\hat{\boldsymbol{\beta}}_\varphi = \hat{\mathbf{Y}}_\varphi \quad (8)$$

Based on $\hat{\boldsymbol{\beta}}_\varphi$, we estimate the intercept as

$$\hat{\alpha}_S = med\left\{Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\varphi\right\} \quad (9)$$

Assume that the random errors follow a distribution *G* with density *g* and median $\theta_e = G^{-1}(1/2)$. Under some mild regularity conditions

$$\begin{pmatrix}\hat{\alpha}_S \\ \hat{\boldsymbol{\beta}}_\varphi\end{pmatrix} \text{ has an approximate}$$

$$N_{p+1}\left(\begin{bmatrix}\alpha \\ \boldsymbol{\beta}\end{bmatrix}, \begin{bmatrix} n^{-1}\tau_S^2 & \mathbf{0}' \\ \mathbf{0} & \tau_\varphi^2\left(\mathbf{X'X}\right)^{-1}\end{bmatrix}\right) \quad (10)$$

distribution, where

$$\begin{aligned}\tau_\varphi &= \left[\int \varphi(u)\varphi_g(u)du\right]^{-1} \\ \tau_S &= \left[2g(\theta_e)\right]^{-1}, \text{ and} \\ \varphi_g(u) &= -\left[g(G^{-1}(u))\right]^{-1}\left[g'(G^{-1}(u))\right]\end{aligned} \quad (11)$$

Thus we have an asymptotic $100(1-\gamma)\%$ confidence interval for the linear combination $\mathbf{l}'\boldsymbol{\beta}$ given by

$$\mathbf{l}'\hat{\boldsymbol{\beta}}_\varphi \pm t_{(\gamma/2,n-p-1)}\hat{\tau}_\varphi\sqrt{\mathbf{l}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{l}} \quad (12)$$

where $\hat{\tau}_\varphi$ is an estimate of τ_φ obtained as in Koul *et. al.* (1987), briefly discussed below.

Estimation of Scale

Let $\hat{\mathbf{e}}$ denote the vector of residuals based on the *R*-fits and let $\varsigma = \tau_\varphi^{-1}$. Then from (11) it follows upon integrating by parts that

$$\varsigma = \int_{-\infty}^{\infty} g(x)d\varphi(G(x)) \quad (13)$$

The estimate of $g(x)$ is obtained using the rectangular kernel density estimator

$$\hat{g}_n(x) = (2nt_n)^{-1}\sum_{i=1}^{n} I(|\, x - \hat{e}_i \,| \le t_n)$$

where $2t_n$ is the window width which will be decided later and $I(A)$ is the indicator function of the event A. Hence an estimate of ς is,

$$\hat{\varsigma} = (2nt_n)^{-1}\left\{\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\varphi(j/(n+1)) - \varphi(j/(n+1))\right) I\left(|e_{(j)} - e_{(i)}| \le t_n\right)\right\}$$

where $\hat{e}_{(j)}$ denotes the jth ordered residual. Using the mean value theorem, standardize the expression in braces above as

$$H_n(z) = (n^2\{(\varphi(1) - \varphi(0))c\})^{-1}$$
$$\sum_{j=1}^{n}\sum_{i=1}^{n}\varphi'(R(\hat{e}_j/(n+1)))I(|\hat{e}_j - \hat{e}_i| \le z)$$

The constant c is chosen so that H_n is an empirical distribution function of the absolute differences $|\hat{e}_j - \hat{e}_i|$. Then choose $t_{n\delta}$ so that $H_n(t_{n\delta}) = 1-\delta$ for $0 \le \delta \le 1$. Our estimate of ς is then,

$$\hat{\varsigma} = \frac{H_n(t_{n\delta}/\sqrt{n})(\varphi(1) - \varphi(0))}{2t_{n\delta}/\sqrt{n}} \qquad (14)$$

Thus our estimate of τ_φ is given by

$$\hat{\tau}_\varphi = \hat{\varsigma}^{-1}$$

Koul et al. (1987) showed that this estimate is consistent for τ_φ under both symmetric and asymmetric error distributions.

Testing

Testing the hypothesis given in (5) will be performed using an F-type test statistic given by

$$F_\varphi = \frac{[D(\omega) - D(\Omega)]/q}{\hat{\tau}_\varphi/2} \qquad (15)$$

where $D(\omega) \equiv \min_{\boldsymbol{\eta}\in\omega}\|\mathbf{Y} - \boldsymbol{\eta}\|_\varphi$ is the minimum dispersion under the restriction imposed by H_0. The quantity qF_φ has an asymptotic χ^2 distribution. Small sample studies, however, indicate that F should be compared to F distribution critical values with q and $n-p$ degrees of freedom.

Algorithm

Consider the QR-decomposition of $\mathbf{X}$

$$\mathbf{Q}'\mathbf{X} = \mathbf{R} \qquad (16)$$

where $\mathbf{R}$ is an $n \times p$ upper triangular matrix of rank p and $\mathbf{Q}$ is an $n \times n$ orthogonal matrix. We may write $\mathbf{Q}$ as $[\mathbf{Q}_1\ \mathbf{Q}_2]$ where $\mathbf{Q}_1$ is an $n \times p$ matrix whose columns form an orthonormal basis for the column space of $\mathbf{X}$. We can now write the kth Newton step as

$$\hat{\mathbf{e}}^{(k)} = \hat{\mathbf{e}}^{(k-1)} - \hat{\tau}_\varphi \mathbf{H}\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)})) \qquad (17)$$

where $\mathbf{H} = \mathbf{Q}_1\mathbf{Q}_1'$ and $\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)}))$ is a vector whose jth component is $a(R(\hat{e}_i^{(k-1)}))$. Here is the formal algorithm. Let ε_D be a given tolerance.

Step 0: Set k=1. Obtain initial residuals $\hat{\mathbf{e}}^{(k-1)}$, $\hat{\tau}_\varphi^{(k-1)}$, and the ($k$-1)th step dispersion, $D^{(k-1)}$.

Step 1: Get $\hat{\mathbf{e}}^{(k)}$ as in (17). Obtain $\hat{\tau}_\varphi^{(k)}$, and $D^{(k)}$.

- If $D^{(k)} < D^{(k-1)}$, then go to Step 2.
- Else perform a linear search (see HM pp. 186-187) along the direction $\hat{\tau}_\varphi \mathbf{H}\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)}))$ for a value which minimizes D, then go to Step 2.

Step 2: If $[D^{(k-1)} - D^{(k)}]/D^{(k-1)} < \varepsilon_D$, then go to Step 3. Otherwise set $\hat{\mathbf{e}}^{(k-1)} = \hat{\mathbf{e}}^{(k)}$ and go to Step 1.

Step 3: Obtain estimates as $\hat{\mathbf{Y}} = \mathbf{Y} - \hat{\mathbf{e}}^{(k)}$, $\hat{\tau}_\varphi = \hat{\tau}_\varphi^{(k)}$, and $\hat{\boldsymbol{\beta}}$ by solving $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}}$.

As a final note we mention that the QR-decomposition can be used to form reduced model design matrices for testing the hypotheses in (5) (see Theorem 3.7.2 of HM).

References

Abebe, A. , Crimin, K. , McKean, J. W., Haas, J. V. & Vidmar, T. J. (2001), Rank-based procedures for linear models : Applications to pharmaceutical science data. *Drug Information Journal*, *35*, 347-371.

Afifi, A. A. & Azen, S. P. (1972), *Statistical analysis: A computer oriented approach*. New York: Academic Press.

Chang, W. H., McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1999), High breakdown rank regression, Journal of the American Statistical Association, *94*, 205-219.

Hettmansperger, T. P., & McKean, J. W. (1998), *Robust nonparametric statistical methods*, London: Arnold.

Hettmansperger, T. P., McKean, J. W., & Sheather, S. J. (2000), Robust nonparametric methods, *Journal of the American Statistical Association, 95*, 1308-1312.

Hollander, M., & Wolfe, D. A. (1999), *Nonparametric statistical methods, Second edition*, New York: Wiley.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5, 229-314.

Kapenga, J. A., McKean, J. W., & Vidmar, T. J. (1995), *RGLM: Users manual*, version 2, *SCL Technical Report*, Dept. of Statistics, Western Michigan University.

Koul, H. L. Sievers, G. L. & McKean, J. W. (1987), An estimator of the scale parameter for the rank analysis of linear models under general score functions, *Scandinavian Journal of Statistics, 14*, 131-141.

McKean, J. W. (2004), Robust analyses of linear models, *Statistical Science, 19*, 562-570.

McKean, J. W., & Schrader, R. (1980), The geometry of robust procedures in linear models, *Journal of the Royal Statistical Society, Series B, 42*, 366-371.

McKean, J. W., Vidmar, T. J., & Sievers, G. L. (1989), A robust two stage multiple comparison procedure with application to random drug screen, *Biometrics, 45*, 1281-1297.

Srinivasan, S. (1997), *Advanced PERL programming*, Sebastopol, CA: O'Reilly.