Online Statistical Modeling (Regression Analysis) for Independent Responses

2017 J. Phys.: Conf. Ser. 855 012054

(http://iopscience.iop.org/1742-6596/855/1/012054)

Online Statistical Modeling (Regression Analysis) for Independent Responses

I Made Tirta[1], Dian Anggraeni[2], Martinus Pandutama[3]

[1,2] Jurusan Matematika FMIPA Universitas Jember, Indonesia
[3] Fakultas Pertanian Universitas Jember, Indonesia
itirta.fmipa@unej.ac.id

Abstract. Regression analysis (statistical analmodelling) are among statistical methods which are frequently needed in analyzing quantitative data, especially to model relationship between response and explanatory variables. Nowadays, statistical models have been developed into various directions to model various type and complex relationship of data. Rich varieties of advanced and recent statistical modelling are mostly available on open source software (one of them is R). However, these advanced statistical modelling, are not very friendly to novice R users, since they are based on programming script or command line interface. Our research aims to developed web interface (based on R and shiny), so that most recent and advanced statistical modelling are readily available, accessible and applicable on web. We have previously made interface in the form of e-tutorial for several modern and advanced statistical modelling on R especially for independent responses (including linear models/LM, generalized linier models/GLM, generalized additive model/GAM and generalized additive model for location scale and shape/GAMLSS). In this research we unified them in the form of data analysis, including model using Computer Intensive Statistics (Bootstrap and Markov Chain Monte Carlo/ MCMC). All are readily accessible on our online Virtual Statistics Laboratory. The web (interface) make the statistical modeling becomes easier to apply and easier to compare them in order to find the most appropriate model for the data.

Keywords: *additive models, linear models, MCMC Regression, online regression analysis, statistical models, web-interface*

1 Introduction

Regression analyss (statistical models) are among statistical methods which are frequently employed in analyzing quantitative data, especially to model dependences between response and several explanatory variables. Nowadays, statistical models have been developed into various directions to handle various type and complex relationship of data. Rich variety of advanced and recent statistical modelings are mostly available on open source software (one of them is R). However, these advanced statistical models, are mostly based on programming script or command line interface, which mean, that they are not easily accessed by applied or practical researchers. The gaps between developed and accessible statistical methods worried statisticians [1] that “practitioners continue to use inappropriate or suboptimal methods due to their being restricted to what is made available via GUIs”.

Therefore it is essential to build interface to make advanced and most recent statistical methods, especially statistical model on R, becoming more user friendly and easier to access and to use. Several GUIs have been developed for various purposes. Explicet, is a GUI designed for management, analysis and visualization of microbiome data [2] and it is claimed has made the analysis of complex microbiome datasets becoming “much more accessible to the growing number of investigators”. Microarray Я US, has been developed based on bioconductor R packages, mainly for researchers with no or little knowledge of R, to have a more reliable and accurate microarray data analysis [3]. Interactive web for statistics learning have also been developed. RwikiStat was developed by combining MediaWiki and Rweb [4] and combining theory with laboratory practice using Rweb, however user still need to have R scripting capabilities. Other types of statistics tutorial with combination of statistics theory and data analysis have been developed using R with shiny packages for spesific topic [5], [6]. This type of data analysis are accompanied with summary of theory and step by step choice analysis with example of interpretation to ensure users are doing analysis data with understanding but no need to master or understand R scripting.

Statistical models with general form $y_i = \mu_i + \varepsilon_i$ for i = 1, 2,3, ..., n have been extended into various directions. For model with independent errors, the model starti from (i) simple linear models (LM) having independent Gaussian errors, i.e., $\varepsilon_i \sim iid\, N(0, \sigma^2)$ and $\mu_i = \sum x_{ij}\beta_j$, for j = 0,1,2,..., p (with p number of regressor/ predictors) [7], (ii) when outliers exist, there are several methods available using robust linear models approaches (RLM) [8][9] (iii) Generalized linear model (GLM) extends LM to accommodate independent errors with wider class of distributions known as the exponential family distributions (i.e, having continuous, count, or binary responses) and possibly nonlinear relationship between response means and the linear predictors, i.e., continuous and differentiable link function g (such as log, logit, inverse/ reciprocal), such that $g(\mu_i) = \sum \beta_j\, x_{ij}$. [10][11]. Later, (iv) statistical model were again generalized to accommodate additive predictors (GAM) such that, $g(\mu_i) = f(x_i)$, for smooth function f (parametric or nonparametric). One of the most frequently applied nonparametric smooth functions are the family of spline smoothers [12][13][14], and (v) perhaps the most recent development of statistical model with independent errors are extension of GAM into GAMLSS [15][16]. GAMLSS accommodates wider type of distribution (with 1, 2, 3, up to 4 parameters, such as the mean, variance, skewness and kurtosis). In addition to modeling the mean, with wider type of distributions, GAMLSS, can also model all other parameters of distributions, each may have its own link function. Recently GAMLSS is extended with variables selection capabilities [17]. In addition to those main statistical models, for small sample, the model are also extended to employ Computer Intensive Statistics (CIS) techniques, such as Bootstrap regression [7] and Markov Chained Monte Carlo (MCMC) regression [18].

All the statistical models mentioned above are already implemented in various packages on R. However, for novice R users, they are not easy to apply since they are all based on command line interface (script). Moreover, in addition to those packages, users may need to upload and call other functions from other R packages for drawing graph or calculating goodness of fit. In this paper we report the development Web-based-GUI interface that unifies most statistical models for independent responses using R and enriched by various options for data exploration, graphical visualisation and goodness of fit measures utilizing several selected R packages.

2 Methods

We develop an interface for unified online (web-based) statistical models for independent responses, which covers LM, RLM, GLM, GAM, GAMLSS, Bootsrap and MCMC regression based on various previously mentioned R- packages. We mainly utilize shiny toolkits [18] to build the interface. There are severeal main steps to follow in building the interface: (i) selecting main and related packages, including the primary functions for the models and secondary functions for graphical visualization, such as scatter plot and correlation plot matrices [20,21], and scatter plot with various smoother [22], and other regression visualization [23]; (ii) identifying the input parameters of the functions, (iii) defining input functions and their options in ui.r file and output functions server.r file; (iv) checking the compatibility of loaded packages and related functions; (v) uploading the files to the web (shiny server) so that they are readily accessible by users.

3 Results and discussion

3.1 General Features of the Web GUI interface

At this stage, we have developed online statistical model fitting for independent responses, covering several models described previously with general features as follows (see Figure 1).

1) **Data Input**: internal database (for practical purposes), or import users' own data with csv or text format (for real data analysis). Users can load all chosen data, or only load small number of the

data (may be needed for illustration or practice with small size of data, such as CIS or Robust Linear Model)

2) **Data exploration** graphical representation on relationship among variables (correlation diagram and scatterplot diagram), graphical exploration of scatter plot with specific model (for examples distributions and link function for LM, RLM, GLM, and smoother for GAM, GAMLSS, see Figure 2, and Figure 3 as samples).
3) **Input Options** for statistics model, including choosing response, predictors (on mean for all models, and for shape and scale for GAMLSS), family and link function (for GLM, GAM and GAMLSS), smooth variable (for GAM and GAMLSS), number of simulations for CIS.
4) **Output Options**, including parameter estimates with their p-values, GOF (AIC, BIC, Adj-Rsquared), diagnostics and other visualization graphics and some selected detailed output (see Fig 4).

The web can be accessed at http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSI/. The summary of features for each model fitting is given in Table 1 and the appearance of the web can be seen in Figure 1.

Table 1. Summary of features of the model fitting

No	Parts	Input Option	Output Option
1	Input Data	• Internal database • Import data (.csv, .txt) • All or randomly select only small number of available data	• Summary of data • List of data
2	Exploratory Data	• General exploration • Smoother exploration (LM, RLM , GLM, GAM)	• Correlation matrix • Correlation Diagram • Scatter plot matrix • Scatter plot with various smoother
	Models Fitting		
3	LM	• Xs and Y • Factor (dummy)	• Estimates (with p-val), • Anova • GOF (AIC, BIC, R-sqr, Adj-RSq) • Diagnostik Graphic • Stepwise regression (variable selection)
4	RLM	• Xs and Y • Method M, MM, LTS • Factor (Dummy)	• Estimates, • Bonferroni test for outlier, • MSE (mean square error) • Graphic
5	GLM	• Xs and Y • Family (link) exponential family including Negative Binomial (log) • Factor (Dummy) • NS or BS smoother	• Estimates (with p-val), • Deviance Analysis • GOF (AIC, BIC) • Scatter plot • Diagnostic Graphic • Stepwise regression (variable selection)
6	GAM (based on mgcv package)	• Xs and Y • Family (link): exponential family including Negative Binomial (log) • Factor (Dummy) • Spline smoother (Cubic Splines, Penalized Spline, Thin	• Estimates (with p-val), • Deviance Analysis • GOF (GCV, AIC, BIC) • Scatter plot • Diagnostics Graphic

No	Parts	Input Option	Output Option
		Plate Spline)	
7	GAMLSS	• Xs and Y • Family (link) including Nornmal Family, Zero Inflated Poisson (log) • Spline smoothers (Cubic Splines, Penalized Spline, Thin Plate Spline) and limited local regression (loess) • Linear and single Predictor for LSS • Choice of Algorithms	• Responses distributions fit graphics • Estimates (with p-val), • Deviance Analysis • GOF (AIC, BIC) • Scatter plot • Diagnostic Graphic
8	CIS	• Bootstrap regression • MCMCRegress (for Gaussian responses) • MCMCpoisson (for Poisson/ count responses) • MCMClogit (for Binomial, especially binary responses)	Graphics of estimates and confidence interval based on 95% percentiles

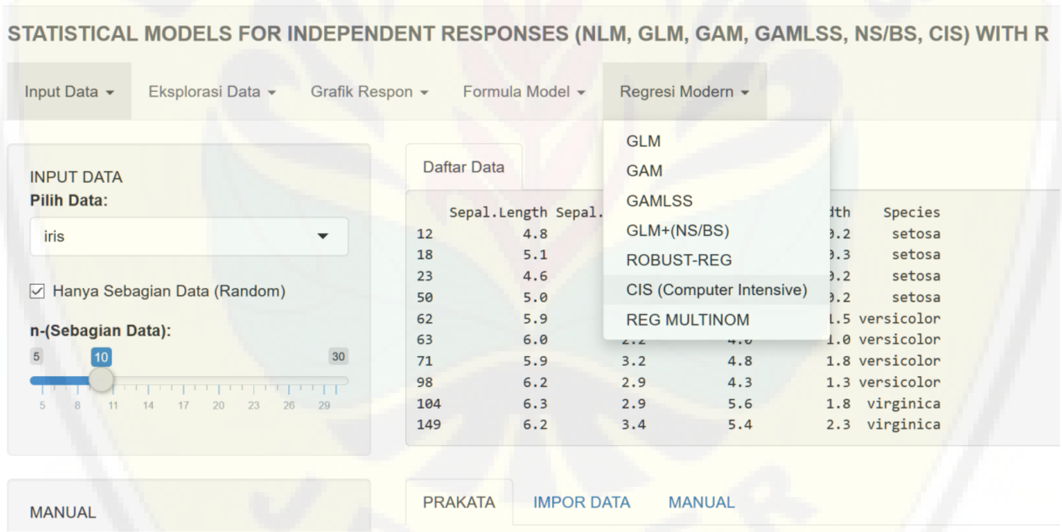


Figure 1. Web appearrance and the main menu (Navbar menu and sidebar menu)

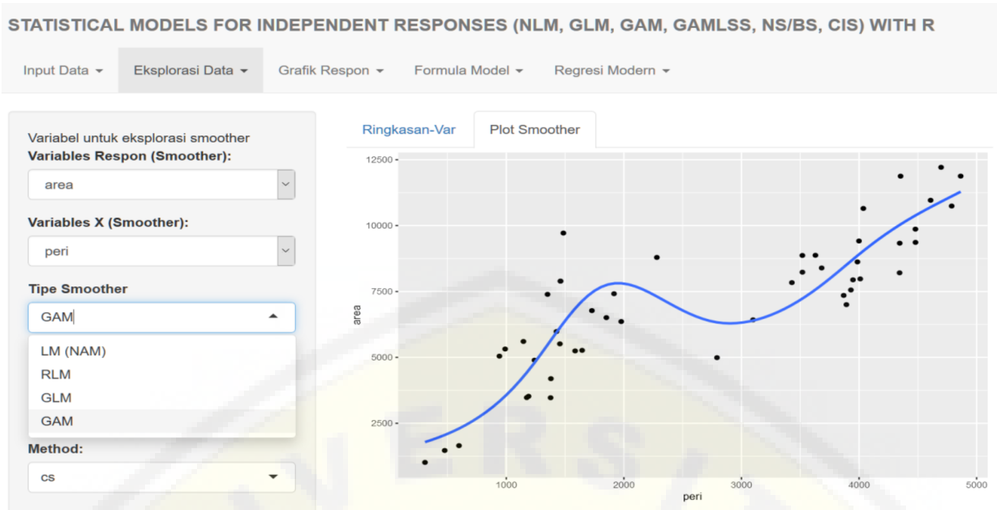


Figure 2. Data and model exploration using various type of smoothers (LM, RLM, GLM, GAM)

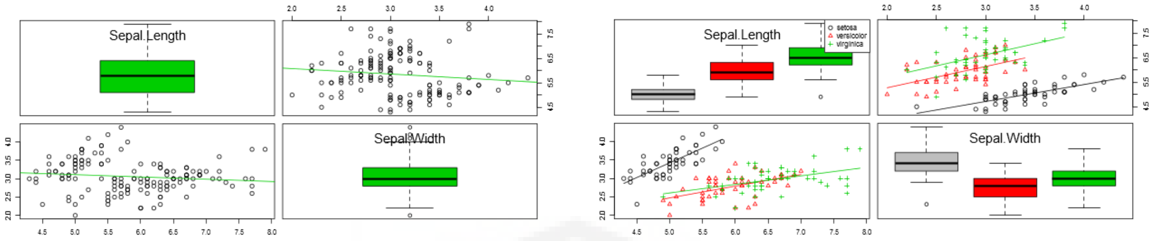
3.2 Numerical Illustrations

The following are numerical illustrations using iris data available on datasets package. The data were first published in 1935 [24]. The main purposes of the illustration are not to show the accuracy of the computation (results), since the results are the same if they are done via script, but to show that results (estimate and comparison among available model) can be done completely and more easily, using "point and click" on the web (see Figure 1). The fitting start from exploration to testing hypothesis about parameters of various alternatif models

3.2.1 Data Exploration

From summary of data we see that the data consist of 1 factor and 4 variables, means that Species as factor may worth considering in the model. Graphical exploration can be made by creating scatter plot, with various type of smoother (available on menu). The first graphics explorations utilize scatter plot matrix of the variables, to check whether factor (i.e., Species) worth considering in the model. For the seek of clarity we only focus on Sepal.Length and Sepal.Width (Figure 3). The plot show that inclusion of Species in the model changes the regression lines directions (regression coefficients or the lines' slope) significantly from negative and may near zero (Figure 3a), to positive for each Species (Figure 3b). The next exploration using various smoother (with ggplot2 package), give us idea that the data may be better fitted using more advanced regression (such as GLM or GAM). Figure 4 shows that applying other continuous distributions (Gamma families) with nonidentity link seem improve the fitness of model. These graphics appearance suggests that Species should be included in the model and more advance model (such as GLM, GAM, GAMLSS, should be considered). Graphics explorations are very beneficial for giving rough idea. However, for more accurate results, user should check and compare goodness of fit measures such as AIC or BIC which are calculated and informed for every choice of model. For illustration or practice with Robust and CIS type regressions, users may randomly load only small amout of the data, however in this paper the illustration for Robust and CIS are excluded. The following are the summary output of all the iris data.

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1   setosa    :50  
 1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3   versicolor:50  
 Median :5.80   Median :3.00   Median :4.35   Median :1.3   virginica :50  
 Mean   :5.84   Mean   :3.06   Mean   :3.76   Mean   :1.2                  
 3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8                  
 Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5                  
```

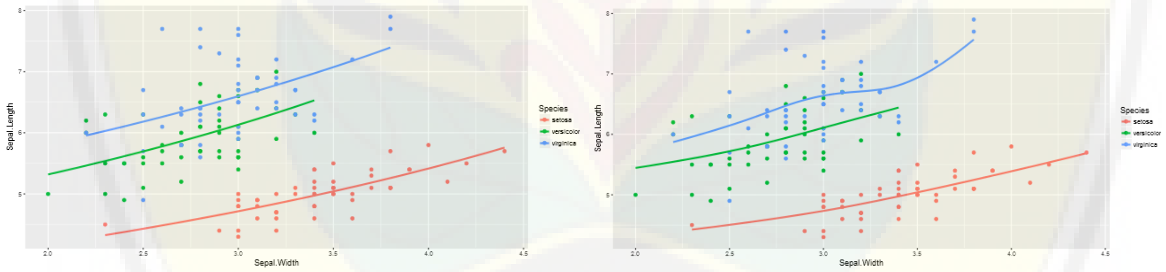


(a) (b)

Figure 3. Scatter plot for Sepal.Length vs Sepal.Width. (a). Without and (b). with factor/group separations)

3.2.2 Alternatives of model fittings

Using our online data analysis, users can easily employ various types of modelings and various combinations of model parameters. For illustration, we set Sepal.Length as response and other variables or factor as explanatory variabels. We fit several models (i) Gaussian distribution (LM) with and without factor, (ii) Gamma with Log link (GLM) and (iii) GAM (by giving smoother on some variables, (iv) GAMLSS (by modeling the scale parameter), and (v) GLM with Natural or B-Splines. All the models are easily set in our interface and the GOF are informed for each model. We describe some of the fitting and summarise the results of all fittings.



(a) (b)

Figure 4. Scatter plot for Sepal.Length vs Sepal.Width with Species as factor . (a) GLM with Gamma distribution and log-link; (b) GAM with Gamma distribution and log-link and additional cubic spline smoother

(i) Fiiting Linear Model without Species

```
Sepal.Length~Sepal.Width+Petal.length+Petal.Width

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.8560     0.2508    7.40  9.9e-12 ***
Sepal.Width    0.6508     0.0666    9.77  < 2e-16 ***
Petal.Length   0.7091     0.0567   12.50  < 2e-16 ***
Petal.Width   -0.5565     0.1275   -4.36  2.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.315 on 146 degrees of freedom
Multiple R-squared:  0.859,     Adjusted R-squared:  0.856
F-statistic:  296 on 3 and 146 DF,  p-value: <2e-16
```

IOP Conf. Series: Journal of Physics: Conf. Series **855** (2017) 012054 doi:10.1088/1742-6596/855/1/012054

```
AIC  BIC   RSq    AdjRsq
84.6 99.7  0.859  0.856
```

(ii) Fitting pararel Linear Model with Species as dummy

```
Sepal.Length~Sepal.Width+Petal.length+Petal.Width+Species-1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
Speciessetosa       2.1713     0.2798    7.76  1.4e-12 ***
Speciesversicolor   1.4477     0.2815    5.14  8.7e-07 ***
Speciesvirginica    1.1478     0.3536    3.25   0.0015 **
Sepal.Width         0.4959     0.0861    5.76  4.9e-08 ***
Petal.Length        0.8292     0.0685   12.10  < 2e-16 ***
Petal.Width        -0.3152     0.1512   -2.08   0.0389 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.307 on 144 degrees of freedom
Multiple R-squared:  0.997,    Adjusted R-squared:  0.997
F-statistic: 9.22e+03 on 6 and 144 DF,  p-value: <2e-16
```

```
AIC  BIC  RSq    AdjRsq
79.1 100  0.997  0.997
```

(iii) Fitting GLM with Gamma(log)

```
Sepal.Length~Sepal.Width+Petal.length+Petal.Width+Species,
family=Gamma(link=log)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.1122     0.0476   23.35   <2e-16 ***
Speciesversicolor  -0.0746     0.0409   -1.83    0.070 .
Speciesvirginica   -0.1220     0.0568   -2.15    0.033 *
Sepal.Width         0.0939     0.0147    6.41    2e-09 ***
Petal.Length        0.1283     0.0117   11.00   <2e-16 ***
Petal.Width        -0.0491     0.0257   -1.91    0.058 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.00273)

    Null deviance: 2.97256  on 149  degrees of freedom
Residual deviance: 0.39372  on 144  degrees of freedom
AIC: 74.95
               link  AIC BIC RSq   AdjRsq
family "Gamma" "log" 75  96  0.996 0.995
```

(iv). Fitting GAM with Cubic Splines

```
Family: Gamma
Link function: log

Formula:
Sepal.Length ~ s(Sepal.Width, bs = "cs", k = 5) + Petal.Length +
    Species
```

```
Parametric coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.3973     0.0166   84.14  < 2e-16 ***
Petal.Length        0.1243     0.0109   11.45  < 2e-16 ***
Speciesversicolor  -0.1257     0.0361   -3.48  0.00067 ***
Speciesvirginica   -0.1970     0.0480   -4.11  6.7e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(Sepal.Width) 1.87      4 9.92 7.9e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.862   Deviance explained = 86.8%
GCV = 0.0028247  Scale est. = 0.0027034  n = 150
```

(v) Fitting GAMLSS with Two Parameters Gamma and log link

We have variety of choices of parameters for GAMLSS (sunch as type of distributions; formula for mean, sigma, nu and Tau, and, type of smoothers). We only choose Gamma with two parameter (mu and sigma), so we only have choices to model mu (μ) and sigma (σ) (neither tau and nor nu). Apparently (for some rough choices) sigma does not dignificantly dependend upen some predictor. Therefore we only report model with constant sigma. Which mean interm of parameter model our GAMLSS does not differ significantly from GAM.

```
Family:  c("GA", "Gamma")
Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.1147     0.0460   24.26  < 2e-16 ***
cs(Sepal.Width, 3)   0.0912     0.0141    6.47  1.5e-09 ***
Petal.Length         0.1305     0.0112   11.65  < 2e-16 ***
Petal.Width         -0.0400     0.0248   -1.61   0.1093
Speciesversicolor   -0.0924     0.0393   -2.35   0.0200 *
Speciesvirginica    -0.1454     0.0546   -2.66   0.0087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.9907     0.0577   -51.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-------------------------------------------------------------------
Global Deviance:     55.2
            AIC:     75.2
            SBC:     105
```

3.2.3 Comparing the models

The estimate of each model and its GOF are summarized and compared in Table 2. We consider the best model (in term of number of parameters and value of likelihood) being the model with the smallest AIC or the biggest BIC.

Tabel 2. Comparison of models

No	Model	Parameters	Estimates	GOF
1	LM without Factor	Intercept	1.8560 (***)	AIC=84.6
		Sepal.Width	0.6508 (***)	BIC=99.7
		Petal.Length	0.7091 (***)	AdjRsq=0.856
		Petal.Width	-0.05565 (***)	
	LM with Factor (paralel model)	InterceptSetosa	2.1713 (***)	AIC=79.1
		InterceptVersicolor	1.4477 (***)	BIC=100
		InterceptVirginica	1.1478 (**)	AdjRsq=0.997
		Sepal.Width	0.4959 (***)	
		Petal.Length	0.8292 (***)	
		Petal.Width	-0.3152 (*)	
2	GLM (with Gamma, log-link)	Intercept	1.1122 (***)	AIC=74.95
		Versicolor	-0.0746 (NS)	BIC=96
		Virginica	-0.1220 (*)	RSq=0.996
		Sepal.Width	0.0939 (***)	AdjRSq=0.995
		Petal.Length	0.1283 (***)	
		Petal.Width	-0.0491 (NS)	
3	GAM with cubic spline smoother on Sepal.Width	Intercept	1.3950 (***)	AIC=73.1 AdjRsq=0.864
		SpeciesVersicolor	-0.0964 (*)	Deviance exlpained-= 87%
		SpeciesVirginica	-0.1501 (**)	(When Petal.Width excluded,
		Petal.Length	0.1308 (***)	AIC=73.7)
		Petal.Width	-0.0396 (NS)	
		s(Sepal.Width)	Edf=1.77 (***)	
4	GAMLSS with cubic spline and constant sigma	(Intercept)	1.1147 (***)	AIC=75.2
		Speciesversicolor	-0.0924 (*)	
		Speciesvirginica	-0.1454 (**)	
		cs(Sepal.Width, 3)	0.0912 (***)	
		Petal.Length	0.1305 (***)	
		Petal.Width	-0.0400(NS)	
		Log(sigma)	-2.9907 (***)	
5	GLM (with Gamma, log-link) and natural spline on Sepal.Width	Intercept	1.3273 (***)	AIC=74.9
		SpeciesVersicolor	-0.0939 (*)	BIC=102
		SpeciesVirginica	-0.1477 (*)	(When Petal.Width excluded
		Petal.Length	0.1308 (***)	AIC=109)
		Petal.Width	-0.0393 (NS)	
		ns(Sepal.Width, df = 3)1	0.0845 (**)	
		ns(Sepal.Width, df = 3)2	0.1948 (**)	
		ns(Sepal.Width, df = 3)3	0.2153 (***)	

Notes:
(***) : p-val $\leq$ 0.1%
(**) : 0.1 < p-val $\leq$ 1%
(*) : 1% < p-val $\leq$ 5%
(NS) : p-val >5%

There are some remarks can be drawn from the results.

(i) It is worth to consider including factors (grups) in the model, when data do not heve observed group, users can perform cluster analysis and take the clusters as group (buliding cluster using Kmeans is also available in our online analysis)

(ii) The significance of individual parameter depends upon the combination of other parameters in the model. The parameters of some variabels may not be significant, but removing them from model can worsen the model (increase the AIC). Therefore, the parameters or variabels may be retained in the model.

(iii) To include spline smoothers in the model (with explonential family distributions), users can choose GAM or GLM+Natural or B-Spline, where the later are easier to interprete in term of using the model for prediction. (In this illustration GLM with natural spline has the smallest AIC value).

3.3 Advantages and disadvantages using online data analysis

The method are placed in the web as part of Virtual Statistics Laboratory. The method can be accessed at http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSI/. There are some adventages for users in using this online model fitting including (i) no need to install R, (ii) no need to master R scripting, (iii) users are easier to surf from one model to another, checking the graphical appearance and the GOF of the model (iv) user can access (do data analysis) using various type of gadgets (hp, tablet notebook, etc) and do simple to advanced statistical modeling with R. The main discomfort in using online data analysis is related to the speed of the available internet network and the number of users accessing the web at the same time. At this stage, web performances (the speed on various gadgets and various web browsers) have not been critically examined. However for local lectures or laboratory practices, students experience no noticable disruptions.

3.4 Future developments

Some features have not been currently implemented namely (i) loess smoother for GAM (since they are conflicted with MGCV), (ii) nonlinear and multiple predictors model for the scale, shape and tau parameters in GAMLSS (iii) testing multicolinearity and models alternatives when it occurs in the predictors. These features, in near future, will be gradually included and tested.

4 Conclusion

Our online statistical model for independent responses, for LM, RLM, GLM, GAM , GAM LSS and CIS, has covered all main features (options) generally done using CLI (script programming), although for CIS types they are not illustrated. It enables users easier to do and compare various types of statistical modellings and choose the most appropriate model. In addition, user is also able to do various data explorations (scatter plot matrix, correlation plot matrix, and other visualization grafik). For GAM and GAMLSS more features are still to be added, and possibly extend the models to include multicollinearity.

Acknowledgement

This research is partly supported by PUPT Grant 2016, from DGHE Ministry of Research Technology and Higher Education Republic of Indonesia

References

[1]. Wallace B C, Dahabreh I J, Trikalinos TA, Lau J, Trow P, Schmid C H. Closing the Gap between Methodologists and End-Users: R as a Computational Back-End *Journal of Statistics Softwares.*Vol. 49, Issue 5, Jun 2012

[2]. Charles E. Robertson1, J. Kirk Harris, Brandie D. Wagner, David Granger, Kathy Browne, Beth Tatem 5, Leah M. Feazel 6, Kristin Park 1, Norman R. Pace 1 and Daniel N. Frank. Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics.* Vol. 29 no. 23 2013, pages 3100–3101

[3]. Dai, Y, Ling Guo, Meng Li and Yi-Bu Chen. Microarray Я US: a user-friendly graphical interface to Bioconductor tools that enables accurate microarray data analysis and expedites comprehensive functional analysis of microarray results. *BMC Research Notes* 2012, 5:282.

[4]. Subianto, M and H Sofyan. 2010. Interactive Statistics Learning with RwikiStat. *International Conference on Networking and Information Technology*

[5]. Tirta, IM. and D. Anggraini. 2015. Clustering: Analysis and Validation Using R-shiny web based interface. *ICOLIB: International Conference On Life Secience and Biotechnology*. 28-29 September 2015 URL: http://statslab-rshiny.fmipa.unej.ac.id/JORS/Cluster/

[6]. Tirta, IM. and D. Anggraini, L.C.Octaviani. 2016. Online and Interactive Web For Fitting GEE With Natural Splines For Longitudinal Data. *IBSC:International Basic Scince Conference*. FMIPA Universitas Jember. 26-27 September 2016. URL: http://statslab-rshiny.fmipa.unej.ac.id/JORS/GEE/

[7]. Fox, J. and S Weisberg. 2011. *An R Companion to Applied Regression*, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

[8]. Venables, W. N. & Ripley, B. D. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[9]. Rousseeuw P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, M. Maechler. 2016. *robustbase: Basic Robust Statistics*. R package version 0.92-6. URL http://CRAN.R-project.org/package=robustbase

[10]. McCullagh & Nelder. 1989. *Generalized Linear Models*. Chapman & Hall

[11]. Marschner,I. 2014. *glm2: Fitting Generalized Linear Models*. R package version 1.1.2. https://CRAN.R-project.org/package=glm2

[12]. Trevor Hastie 2016. gam: Generalized Additive Models. R package version 1.14. https://CRAN.R-project.org/package=gam

[13]. Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

[14]. Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1):3-36

[15]. Rigby R.A. and Stasinopoulos D.M. 2005. Generalized additive models for location, scale and shape,(with discussion), *Appl. Statist*., 54, part 3, pp 507-554.

[16]. Stasinopoulos D.M and Rigby R.A. 2008. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistitics Software*. Vol **23** (no 7). 1-46

[17]. Benjamin, H., Andreas Mayr, Matthias Schmid. 2016. gamboostLSS: An R Package for Model Building and Variable Selection in the GAMLSS Framework. *Journal of Statistitics Software*. Vol **74** (no 1). 1-31

[18]. Andrew D. Martin, Kevin M. Quinn, Jong Hee Park. 2011. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*. **42(9)**: 1-21. URLhttp://www.jstatsoft.org/v42/i09/.

[19]. Chang, W., J. Cheng, JJ Allaire, Y. Xie and J McPherson. 2015. *shiny: Web Application Framework for R*. R package version 0.11.1. http://CRAN.R-project.org/package=shiny

[20]. Fox,J. 2016. *RcmdrMisc: R Commander Miscellaneous Functions*. R package version 1.0-5. https://CRAN.R-project.org/package=RcmdrMisc

[21]. Revelle, W. 2016. *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.6.9.

[22]. Wickham. H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[23]. Patrick Breheny and Woodrow Burchett (2016). *visreg: Visualization of Regression Models. R package version 2.3-0*. https://CRAN.R-project.org/package=visreg.

[24]. Anderson, E. 1935. The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5.

Appendix:

Table 3. Menu Structure of Online Statistical Model for Independent responses

No	NavBar	Sub Menu	SideBar	Output
1	Input Data	-	Data Selection (internal & Import)	List of Data
2	Exploration	General	• Variable selection • Type of Diagonal plot (histogram, boxplot, qqplot, density) • Check and set for dummy	• Summary statistics • Correlation matrix • Correlation diagram

No	NavBar	Sub Menu	SideBar	Output
			• Form Kmean clustering	• Scatterplot matrix
		Smoother	• response and predictor selection • type of smoother and parameters (LM, RLM, GLM, GAM, Natural Spline)	• scatter plot with smoother
3	Response variable	• Respon Variable • Histogram • QQplot • Box-Plot	• Respon Variable Slection • For histogram ✓ Check for density estimate ✓ Check for density estimate ✓ Check for mean and median	Qqplot BoxPlot Histogram (with density estimate, mean and median)
4	Model Formula	Respon-predictors (for LM, GLM, GAM)	• Setting Response • Setting Predictor • Checking & Setting Dummy • Setting Predictor for scatter plot	OLS output • Summary • GOF • Anova • Scatter Plot • Visual Plot (from visreg packages) • Stepwise output • X matrix
		Family and link (For GLM, GLM+NS and GAM)	• Family selection • Link selection	• Scatter plot with smoother • GOF of GLM
		Formula for GAMLSS	• Response selection • Distributionsfamily • Predictor Selection • Predictor for Sigma • Predictor for Nu • Predictor for Tau	• Histogram of data and fitness of theoretical density
5	Detail of Modern Ression(output and further selection)	GLM	-	• Summary of fit • Deviance analysis • Fiited Plot (2D) • Stepwise • Diagnostic plot
		GAM	• Variable for nonparametric • Type of smoother • Df • Spesific object of GAM	• Summary • Fitted plot (2D) • GOF • Detail output of GAM
		GLM+NS	• Response • Predictor • Nonparametric • Df for NS	• Smoother plot (2D) • Summary of fit • Deviance analysis • GOF • Diagnosticplot • X matrix
		GAMLSS	(extension from Formula for GAMLSS) • Type of Smoother	• Summary • Plot

No	NavBar	Sub Menu	SideBar	Output
			• Df • Estimation method (RS, CG, Mixed) • Type of Plot (diagnostic, term, worm)	• GOF
		Robust-Reg	Method (M,MM,MF, LTS)	• Summary ofOLS • Summary of Robust • Bonferroni Test • GOF • Plot (OLS and Robust)
		CIS	• Response for CIS • Predictor for CIS • Type of CIS (Bootstrap, MCMC) • Type of MCMC (Gaussian, Poisson, Logit-Binomial) • Number of bootstraps • number of burned in (MCMC)	• Estimate • Plot of estimate • Bootstrap CI • Bootstrap Jakknife