Topics in model fitting statistics: A focus on robust regression and model diagnostic statistics with application to maternal anaemia data in Malawi

Potiphar M. Damiano   Tsirizani M. Kaombe

Department of Mathematical Sciences, University of Malawi

February 12, 2024

Contents

1. Introduction
2. Statistical methods
3. Results
4. Conclusion
5. Selected references

Background

Maternal Anaemia

- Anaemia is a blood disorder characterized by low concentration of haemoglobin (Meena et al., 2019)
- Significant risk factors include education level, body mass index (BMI), wealth index, place of residence, contraception method during pregnancy, water source (Talukder et al., 2022)
- Commonly associated with physiological changes in pregnancy, gravidity, Age, nutritional deficiencies, infection (Malaria, HIV and Hookworm) (Munasinghe & van den Broek, 2006)
- The burden of maternal anaemia is high in sub-Saharan Africa, which derails safe motherhood campaign efforts in the region (Kassebaum et al., 2016)

Goal of the study

Maternal anaemia data modelling gaps and study objective

- Absence/limitations of studies that applied robust regression methods and diagnostic statistics on maternal anaemia data to observe their performance.
- Regression Diagnostic statistics and robust regression methods help to provide better estimates in presence of unusual observations (Ayinde et al., 2015)
- Hence, their application provide comparable quality in the estimates of the risk factors of maternal anaemia
- Thus, there is need to evaluate performance of these methods when applied to the same data

Goal of the study

study objective

- The study compares performance of robust regression methods and diagnostic statistics when applied to both simulation study and maternal anaemia data in Malawi
- 21,935 mothers aged 15-49 years who participated in 2015-16 Malawi Demographic Health Survey (DHS) and had hemoglobin level known were studied

Regression parameter estimation

Multiple linear regression model

- Let Y be the hemoglobin level outcome and X_i, for $i = 1, 2..., p$ be the explanatory variables, then multiple linear regression model is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1}$$

- where ϵ is the error term, assumed to be normally distributed, $N(0, \sigma^2 I_n)$ and independent
- The ML estimator for β and σ expressed by $\hat{\beta_N} = (X^T X)^{-1} X^T Y$ and $(\hat{\sigma_N})^2 = \frac{1}{2\sigma^2} \sum (y_i - x_i \hat{\beta_N})^2$ respectively
- Such that the vector of fitted values is represented by:

$$\hat{Y_N} = X(X^T X)^{-1} X^T Y = HY \tag{2}$$

Regression parameter estimation

Quantile regression

- Performs better than OLS regression when the data is skewed, it minimizes the median than mean
- For Hemoglobin level (Y) and it's distribution function $F(y) = Pr(Y \leq y)$, the θ-th, for $0 < \theta < 1$, quantile is defined as $Q(\theta) = inf(x : F(Y) \geq \theta)$
- Quantile regression model is given by:

$$Q(y_i) = \beta_0(\theta) + \beta_1(\theta)X_{i1} + \beta_2(\theta)X_{i2} + \ldots + \beta_p(\theta)X_{ip} \tag{3}$$

- Where $i = 1, ..., n$ and $\beta_j(\theta)$ is estimated by minimizing the problem,helped by R QUANTREG package; $\sum \rho_\theta(y_i - \beta_0(\theta) - \sum x_{ij}\beta_j\theta)$
- Where $\rho_\theta(r)$ is the check loss given by $\rho_\theta(r) = \theta max(r, 0) + (1 - r)max(-1, 0)$

Diagnostic Statistics tools

Outlier and Leverage measures

- At the i-th data point the unstandardized residual is $e_i = y_i - \hat{y}$ and $Var(e_i) = \sigma^2(1 - h_i)$
- Where $h_i = X_i(X^T X)^{-1})X_i^T$, i-th diagnonal element is interpreted as amount of Leverage or influence exerted by Y_i on $\hat{Y}_i$, h_i is large if $h_i \geq 2\frac{p}{n}$ where $p = \sum_1^n h_i$
- The standardized (Studentized) residuals, t_i, given by: $t_i = \frac{e_i}{s(1-h_i)^{\frac{1}{2}}}$
- Where $s = [\frac{\sum_1^n e_i^2}{n-p}]^{\frac{1}{2}}$ is the estimate for σ

Diagnostic Statistics tools

Influence measures

- To examine the effect of the observations on the parameter, Cook's distance, D_i shows the effect of i-th deleted case on all fitted values, $D_i > 1$ is considered influential, $D_i = \frac{(\hat{Y}-\hat{Y}_i)^T(X^TX)(\hat{Y}-\hat{Y}_i)}{ps^2} = \frac{n-p}{p}\frac{h_i}{1-h_i}t_i$
- DFFITS diagnostic combines the information in the leverage h_i, and the Studentized residual e_i, $DFFITS_i$ is considered large if $DFFITS_i \geq 2[\frac{p}{n}]^{\frac{1}{2}}$: $DFFITS_i = \frac{(\hat{Y}-\hat{Y}_i)}{s_i h_i^{\frac{1}{2}}} = [\frac{h_i}{1-h_i}]^{\frac{1}{2}} t_i$
- DFBETAS measures influence of i-th case on each regression coefficients, b_k, DFBETAS is considered large if is greater than 1 (small data) or $\frac{2}{\sqrt{n}}$ (large data): $DFBETAS_i = \frac{b_k - b_{k(i)}}{\sqrt{MSE_i C_{kk}}}$ where where C_{kk} is the k-th diagonal element of $(X^TX)^{-1}$

Robust Regression statistics

M-Estimators

- The M-estimator's goal is to minimise a function of the errors, ρ rather than the sum of squared errors. The objective function of the M-estimate is:

$$Min \sum_{i=1}^{n} \rho(\frac{e_i}{s}) = Min \sum_{i=1}^{n} \rho(\frac{Y_i - X_i\beta}{s}) \tag{4}$$

- Where s is estimate of scale often formed from linear combination of the residuals
- A reasonable ρ should have the following properties: $\rho(e) \geq 0$, $\rho(0) = 0, \rho(e) = \rho(-e)$, and $\rho(e_i) \geq \rho(e_i^T)$ for $|e_i| = |e_i^T|$
- Minima solution associated with equation (3) is obtained by taking Gauss-Newton iterations, helped by R ROSEPACK package: $\sum_{i=1}^{n}(\phi)(\frac{Y_i - X_i\beta_i}{s})X_i$ where ϕ is a derivative of ρ.

Robust Regression statistics

S-Estimators and MM-Estimators

- S-estimator is defined by minimization of dispersion of residuals: minimize $S(e_1(\theta), ..., e_n(\theta)$, defined as solution of $\frac{1}{n}\sum_1^n \rho(\frac{e_i}{s}) = K$
- Where $s(\theta)$ is a type of robust M-Estimate of scale of residuals, K is a constant and $\rho(\frac{e_i}{s})$ is the residual function.
- MM-estimators combine the high asymptotic relative efficiency of M-estimators with the high breakdown of class of estimators called S-estimators
- MM-estimator $\hat{\beta}$ defined as a solution to:

$$\sum_{i=1}^{n} x_{ij}(\phi_1)(\frac{y_i - x_i\beta_i}{s_n})x_i forj = 1, 2, ..., p \tag{5}$$

- Where j=1,2,...,p, $\phi_1(\mu) = \frac{\partial \rho_1(\mu)}{\partial \mu}$

Robust Regression statistics

LST Estimators

- LTS Estimator minimizes the sum of trimmed squared residuals and is given by:

$$\hat{\beta}_{LTS} = Min \sum_{i=1}^{n} e_i^2 \tag{6}$$

- Such that $e_{(1)}^2 \leq e_{(2)}^2 \ldots \leq e_{(n)}^2$ are the ordered squares residuals and h is defined in the range $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, with n and p being sample size and number of parameters respectively.
- The largest squared residuals are excluded from the summation in this method

Descriptive statistics of Hemoglobin level and selected covariates

Mean	Median	Std deviation	Min	Max	Skewness	Kurtosis
12.53	12.7	1.74	2	23	-0.52	4.85

Variable	Category	Frequency	Percentage
BMI	Underweight	1185	5.42
	Normal Weight	10958	50.09
	Overweight	6733	30.78
	Obese	3,001	13.72
Age	15-24	2655	12.10
	25-49	19280	87.90
Highest education level	No education	4451	20.29
	Primary	13958	63.63
	Secondary	3269	14.90
	Tertiary	257	1.17
Wealth Index	Poor	8699	39.66
	Middle	4403	20.07
	Rich	8833	40.27
Water source	Unsafe Water	3010	13.79
	Safe Water	18821	86.21
Distance to health centre	Big problem	12133	55.31
	No problem	9802	44.69
Residence	Urban	3440	15.64
	Rural	18495	84.32
Contraceptive use	Modern method	12567	57.29
	Traditional method	272	1.24
	Non-users	9096	41.47

Model Estimates

Multiple Linear and Quantile Regression ML Estimates

Variable	Category	OLS p-value	Q_{25} p-value	Q_{50} p-value	Q_{75} p-value	Q_{90} p-value
Intercept		11.86 (< 0.0001)	10.97 (< 0.0001)	12.59 (< 0.0001)	13.46 (< 0.0001)	14.1 (< 0.0001)
Residence	Urban*					
	Rural	-0.54 (0.001)	-0.32 (0.119)	-0.53 (0.004)	-0.92 (< 0.0001)	-1 (0.001)
Education	No education*					
	Primary	0.84 (< 0.0001)	0.59 (< 0.0001)	0.82 (< 0.0001)	1.1 (< 0.0001)	1.40 (< 0.0001)
	Secondary	0.65 (< 0.0001)	0.32 (0.160)	1.01 (< 0.0001)	0.64 (0.015)	0.80 (0.013)
	Tertiary	0.33 (0.546)	-0.36 (0.599)	-0.16 (0.799)	1.01 (0.205)	1.10 (0.256)
Gravidity		0.08 (0.003)	0.09 (0.005)	0.08 (0.005)	-0.002 (0.950)	$-2.5e^{7}$ (1.00)
Current pregnant duration		-0.02 (< 0.0001)	-0.24 (< 0.0001)	-0.02 (< 0.0001)	-0.024 (< 0.0001)	-0.20 (< 0.0001)
Distance to Health Centre	Big problem*					
	No problem	0.2 (0.043)	0.11 (0.367)	0.06 (0.560)	0.10 (0.492)	$8.86e^{7}$ (1.00)
BMI	Underweight*					
	Normal	0.15 (0.693)	0.15 (0.752)	-0.27 (0.519)	0.15 (0.785)	0.60 (0.366)
	Overweight	0.19 (0.616)	0.28 (0.542)	-0.11 (0.790)	0.31 (0.565)	1.00 (0.134)
	Obese	0.52 (0.199)	0.71 (0.153)	0.03 (0.955)	0.69 (0.235)	1.00 (0.158)
Wealth Index	Poor*					
	Middle	0.20 (0.128)	0.32 (0.048)	0.17 (0.228)	0.43 (0.022)	-0.30 (0.188)
	Rich	-0.20 (0.100)	-0.25 (0.098)	-0.48 (< 0.0001)	-0.15 (0.387)	-0.7 (0.001)
Age	15-24*					
	25-49	-0.10 (0.495)	-0.09 (0.615)	-0.45 (0.004)	-0.11 (0.585)	-0.10 (0.683)
Water source	Unsafe*					
	Safe	-0.24 (0.064)	-0.13 (0.393)	-0.26 (0.069)	-0.25 (0.164)	-0.80 (< 0.0001)
AIC		4352.5	4459.0	4297.5	4487.4	4897.6

Outliers in the models

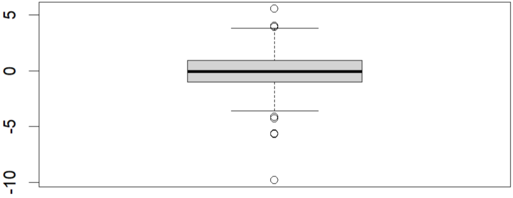

Figure: Box plot for OLS regression model

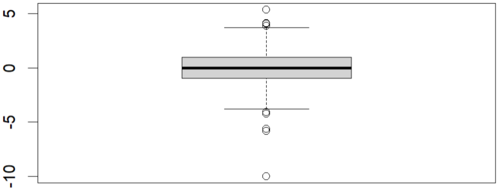

Figure: Box plot for Quantile regression model

Work in progress

- Apply simulation techniques to compare efficiency of estimates from maximum likelihood estimation and robust regression for a linear model
- To apply simulation methods to compare effectiveness of robust regression and diagnostic statistics in detecting outliers and influential data points to a linear model
- To apply diagnostic statistics and robust regression on maternal anaemia data in Malawi to compare the extent of detection of outliers and influential observations by each method

Conclusion

- The study observes performance of robust regression methods and diagnostic statistics on quality of estimates and detection of unusual observations when applied to both simulated data and maternal anaemia data in Malawi
- Akaike Information Criterion (AIC) was used to determine the best model fit for the dataset and p-values were used to determine statistical significance
- 50th Quantile regression model was the best fitted model to the dataset as it has small AIC compared to 25th, 75th, 90th and OLS
- Quantile regression model identified 8 outliers while OLS regression model identified 6 outliers in the dataset

Selected references

Ayinde, K., Lukman, A. F., Arowolo, O., et al. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics*, *5*(04), 273.

Kassebaum, N. J., Collaborators, G. . A., et al. (2016). The global burden of anemia. *Hematology/oncology clinics of North America*, *30*(2), 247–308.

Meena, K., Tayal, D. K., Gupta, V., & Fatima, A. (2019). Using classification techniques for statistical analysis of anemia. *Artificial intelligence in medicine*, *94*, 138–152.

Munasinghe, S., & van den Broek, N. (2006). Anaemia in pregnancy in malawi-a review. *Malawi Medical Journal*, *18*(4), 160–175.

Talukder, A., Paul, N., Khan, Z. I., Ahammed, B., Haq, I., & Ali, M. (2022). Risk factors associated with anemia among women of reproductive age (15–49) in albania: A quantile regression analysis. *Clinical Epidemiology and Global Health*, *13*, 100949.