Journal of Reliability and Statistical Studies (ISSN: 0974-8024)
Vol. 2, Issue 1(2009): 52 - 59

REGRESSION MODEL WITH POWER TRANSFORMATION WEIGHTING: APPLICATION TO PEAK EXPIRATORY FLOW RATE

Girdhar G Agarwal[1] and Rashmi Pant[2]
1. Department of Statistics, Lucknow University, Lucknow, U.P, India-226007.
2. Jaipuria Institute of Management Studies, Lucknow, U.P, India-26010.
E-mail: stats.quo@gmail.com

Abstract

This paper presents the development of weighted least squares (WLS) method in regression modeling when data are characterized by a high degree of heteroscedasticity in the response variable. An algorithm is developed to obtain the weighting parameter in the WLS model. Tests for coincidence and parallelism in the WLS model are studied. Finally results are demonstrated empirically by modeling the effect of age, body mass index and sex, on Peak Expiratory Flow Rate (PEFR).

Keywords: Weighted Least Squares Regression, Coincidence, Parallelism, Peak Expiratory Flow Rate, Regression Diagnostics

1. Introduction

Model fitting by unweighted least squares is efficient if errors besides being normal are independent and have constant variance. It sometimes happens that the variances of the observations are not all equal i.e data exhibit a phenomenon called *heteroscedasticity*. Regression modeling in the presence of heteroscedasticity has been extensively dealt with in literature (Theil, [18]; Draper and Smith, [8]). Heteroscedasticity is generally viewed as no more than a hindrance to the correct estimation of regression coefficients. However its analysis also provides investigators with significant information, about the structure of data that would ordinarily go undetected, as well as evidence of model misspecification.

Power Transformations (Box and Cox, [4]) and weighted least square (WLS) techniques are used when the assumption of constant error variance in the linear model is violated. An advantage of the weighted least squares approach is that it can be used irrespective of the fact that heteroscedasticity was inherent in the data or was induced by transformation.

Usually, in an empirical analysis, the weights are not known and have to be generated by a “combination of prior knowledge, intuition, and evidence” (p.101, Chatterji, [6])**.** Here we have presented a method for finding the weights in weighted least squares analysis when the variance of the fitted response variable is a function of its expected value. The method is applied to data for the study of peak expiratory flow (PEF) for healthy men and women. The residuals obtained from the WLS model are used to assess the validity of our model. Peak expiratory flow is a useful and simple parameter for assessing the lung function status of a person in the general population and for making diagnosis and treating patients with bronchial asthma and chronic obstructive lung disease. Many studies on peak expiratory flow in the general population had been carried out (Selby and Read ,[17]; Johannson and Erasmos, [12];

Woolcock et al., [20]; Gregg, [10]; Raju et al., [15], [16]). These studies have used multiple regression analysis to explore the relationship between PEFR and age, height, and weight. However, these studies have failed to account for the presence of heteroscedasticity in PEFR. So there is a need to study lung function data from healthy population to establish physiological norms to predict peak expiratory flow values in people of different age, height, weight and sex group by using weighted least squares regression model. We also intend to assess whether separate regression models are required to predict PEFR for males and females and instead of height and weight, we use body mass index (BMI), a quantity that is calculated using height and weight measurements.

2. Methods

We have the linear model

$$y = \eta + \varepsilon \tag{1}$$

where $y = (y_1, y_2, \ldots\ldots, y_n)'$ is n x 1 vector of independent observations and $V(y_i) = c_i\sigma^2$

The original variable y_i is transformed such that the variance of transformed variable $g(y_i)$ is constant. In particular,

$$V\{g(y_i)\} = \sigma^2 \tag{2}$$

Now the variance of original variable y_i can be expressed as (p.88-92, Kendall & Stuart, [13])

$$V(y_i) \approx \sigma^2 \, \eta_i^{2-2\lambda} \tag{3}$$

where $\eta_i = E(y_i)$ is the mean response and λ is termed as the weighting parameter. So the weights w_i for the least square analysis are chosen as inversely proportional to $V(y_i)$:

$$w_i = \eta_i^{2\lambda-2}. \tag{4}$$

In practice η_i is unknown and is replaced by the fitted value $\hat{y}_i$ of y_i.

3. Evaluating the weighting parameter λ

We assume that $\eta_i = X_i\,\beta$, where X_i is $1 \times p$ vector $(x_{i1}, x_{i2}, \ldots\ldots, x_{ip})$ of predictor variables and β is $p \times 1$ vector of regression parameters. The normality assumption for y_i gives likelihood function.

$$\left(\prod_{i=1}^{n} w_i\right)^{1/2} [(2\pi)^{\frac{n}{2}}\sigma^n]^{-1} \exp\left\{-(y - X\beta)' \, W\,(y - X\beta)(2\sigma^2)^{-1}\right\} \tag{5}$$

where X is n × p matrix consisting of vectors X_i (i = 1,2,….,n) and W is n × n diagonal matrix consisting of elements w_i (i = 1,2,…..,n). We can find the maximum-likelihood estimates in two steps. First, for given λ, (5) is, except for a constant factor, the likelihood for a weighted least-squares problem. Hence the maximum-likelihood estimates of β's are the weighted least square estimates $\hat{\beta} = (X'WX)^{-1} X'Wy$ and the estimate of σ^2 is:

$$\hat{\sigma}^2 = y'\left[W - WX(X'WX)^{-1}X'W\right] y/n = S/n \tag{6}$$

where S is the residual sum of squares.

Thus for fixed λ, the maximized log likelihood is, except, for a constant,

$$L(\lambda) = (1/2)\sum_{i=1}^{n} \log \hat{y}_i^{2-2\lambda} - (n/2)\log\hat{\sigma}^2 \tag{7}$$

It will now be informative to plot L(λ) against λ for a trial series of values. From this plot the maximizing value $\hat{\lambda}$ may be read off. Alternatively $\hat{\lambda}$ can be obtained iteratively as follows (Box and Hill, [5]):

(i) initially, substitute y_i for η_i in equation (4) for w_i;

(ii) select a λ;

(iii) calculate $\hat{\beta}$, the weighted least squares estimate of β, and hence calculate L(λ) using (7);

(iv) repeat (ii) and (iii) until finding a value $\hat{\lambda}$ that maximizes L(λ);

(v) corresponding to maximizing value $\hat{\lambda}$ find $\hat{y}_i = x_i \hat{\beta}$ and substitute for η_i in equation (4) for w_i, return to (ii) and start new iteration. Stop operation, when two consecutive iterations give a common value for $\hat{\lambda}$.

In most of the cases, the iterative process converges in four or five steps.

5. A Study of PEFR

The present study was carried out on 772 non-smoking healthy persons from North India in the age-group 19-60 years.. Out of these, 618 were males and 154 females. Table- 1 gives the baseline characteristics of the variables classified by sex. All the variables, except for PEFR do not differ significantly among males and females (Table-1). The objective is to get the best representation of the relationship between PEFR and the demographic variables (Age, BMI and Sex,).

	Male (618) Mean, S.D	Female (154) Mean, S.D	t-statistic, p -value
Age	33.2, 12.3	31.9, 11.4	1.12, 0.26
BMI	21.3, 3.37	21.4, 5.9	-0.17, 0.865
PEFR	464.2, 104.2	389.9, 84.5	8.42, <0.001

Table 1: Comparison of baseline characteristics of persons by sex

The violation of the constant variance assumption can be easily depicted when Var(y) is plotted against age (one of the predictors). The graph is shown in Figure 1. The ratio of maximum of Var(y) to minimum of Var(y) for age is approximately 50 to 1. Similar kinds of differences in variances of response variable PEFR were found for other predictor variables. Hence it is evident that weighting is needed for efficient least squares estimation.

Comparison of Coefficients

We started with the full model

$$y = \beta_0 + \beta_1 Age + \beta_2 BMI + \beta_3 Sex + \beta_4 (Age * Sex) + \beta_5 (BMI * Sex) + \varepsilon \tag{8}$$

where dependent variable y is PEFR. An unweighted least squares analysis (OLS) was performed on the data, leading to the estimates of parameters shown in column 2 of Table 2.

To proceed with the analysis using the proposed algorithm, the weights were obtained using equation (4). A converged estimate of power parameter, $\hat{\lambda}$ =1.85 was achieved in three iterations. The results of weighted linear least squares analysis (WLS) are shown in column 3 of Table 2. One of the regression parameter estimates ($\hat{\beta}_3$) has significantly different values. But the differences in the estimates of other parameters are not dramatic in the two methods (OLS and WLS) of solution.

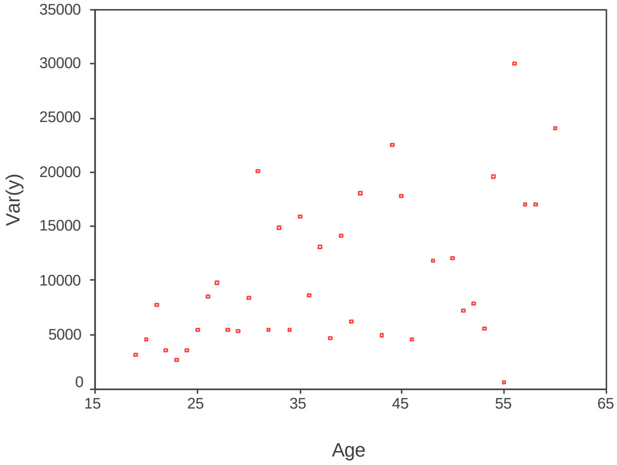

Figure 1: Plot of Var(y) against predictor age

Parameter Estimates	$\hat{\lambda}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Unweighted	1	503.607	-3.143	-0.72	-8.5	-1.003	5.7
Weighted	1.85	504.82	-3.13	-0.79	-2.64	-0.71	5.01

Table 2. Least squares results for 618 males (OLS and WLS)

Model Choice

The tests of coincidence and parallelism (Kleinbaum et al., [14]) are used to test if separate models are required for males and females. Test of coincidence implies testing for the reduced model which does not contain the interaction of the variable "Sex" with other independents. If the test of coincidence hypothesis is rejected, indicating that inclusion of variable 'Sex" and its interactions are important, then a test of parallelism would be used to ascertain whether separate models for males and females are necessary.

(a) Test for coincidence

The null hypothesis is

H_0: $\beta_3=\beta_4=\beta_5=0$

This is to test for the reduced WLS model

$$y=\beta_0+\beta_1 Age+\beta_2 BMI+\varepsilon$$

The weighting parameter being used is $\hat{\lambda}$ =1.85.

The $F_{(2,769)}$ value=31.24968, p-value<0.0001, hence the hypothesis for coincidence is rejected.

b) Test for parallelism

In the OLS model the test for parallelism using dummy variable for sex is not valid as the variances for the males and females are not equal. Proceeding to test for parallelism in our WLS model, the hypothesis is

H_0: $\beta_4=\beta_5=0$

We are thus testing for the reduced model

$$y = \beta_0 + \beta_1 Age + \beta_2 BMI + \beta_3 Sex + \varepsilon$$

The weighting parameter being used is $\hat{\lambda}$ =1.85.

The $F_{(2,768)}$ value=1.936, p-value=0.145, hence the hypothesis for parallelism is accepted and we conclude that the regression planes are parallel. A practical interpretation of the result leads us to the conclusion that the WLS model for males and females will differ *only in the intercept term.*

A re-estimation of the weighting parameter is now required for a single model for both sexes. The estimate obtained from the Fortran program converged at the third iteration yielding $\hat{\lambda}$ =1.9.

Using the above estimate the following WLS model will be valid for both sexes (both factors of the variable "sex"):

$$y = 435.95 - 3.77 Age + 3.34 BMI + 82.48 Sex + \varepsilon \qquad (9)$$

Validation of assumptions

Model fitting is incomplete without regression diagnostics (Anscombe, [1]; Atkinson, [2]). Such techniques are employed to validate the underlying assumptions and to assess the accuracy of computations for a multiple regression analysis. Regression diagnostics are performed with the help of residuals of fitted model (Atkinson, [3]; Cook, [7]; Tsai and Xizhi, [19]). Here, it is especially required to illustrate the necessity of weighting in the linear model (9). A plot of the unweighted residuals versus fitted values shown in Figure 2 shows the serious heteroscedasticity of variances. A plot of weighted residuals versus predicted values is shown in Figure 3. From Figure 3, it appears that the spread of residuals has evened out compared to Figure 2. Also there is no pattern in the plot of weighted residuals against the fitted value of y (Figure 3).

Based on the empirical analysis above, there is a clear suggestion that the corrective remedy is offered by the weighting procedure. The results for our final model are enumerated in Table 3. A plot of the unweighted residuals versus age (OLS) is shown in Figure 4, whereas plot of weighted residuals versus age (WLS) is shown in Figure 5.

6. Algorithm development

The algorithm for the computation of weights in the power transformation model contains the following steps.

Step 1. Specify sample size and number of predictors

Step 2. Define columns of the X-matrix and the response vector Y.

Step 3. Define the Weight matrix W and compute the weighting parameter λ (defined as LAMBDA) using algorithm proposed in sec.2.

Step 4. The new response vector and predictor matrix is computed using the weights used in step 3.

Step 5. The log-likelihood for the model is computed and if it is maximum then the algorithm is convergent else the algorithm is repeated until a value of LAMBDA is obtained for which the log-likelihood is maximized.

The algorithm was found to converge when variables were defined in double precision only.

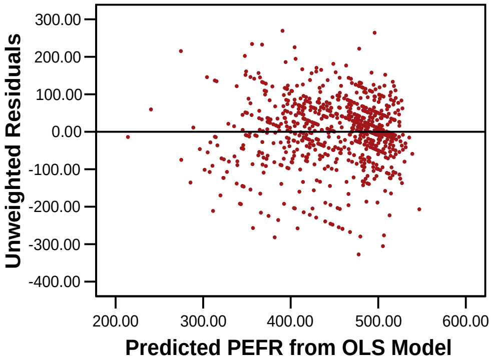

Figure 2: Plot of predicted values versus unweighted residuals

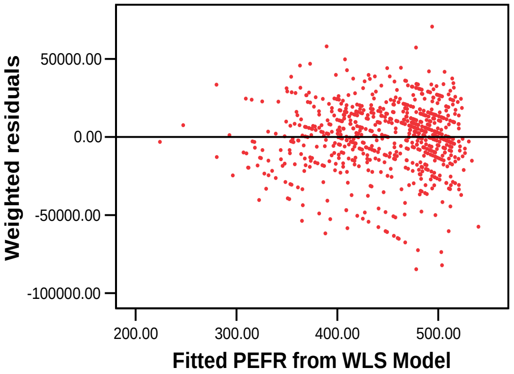

Figure 3: Plot of predicted values vs. weighted residuals from Model (9)

Parameter Estimates	$\hat{\lambda}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Unweighted	1.0	434.59	-4.02	3.82	81.427
Weighted	1.9	435.95	-3.77	3.34	82.48

Table 3: Least squares results for Model with Age, BMI & Sex as independents (OLS and WLS)

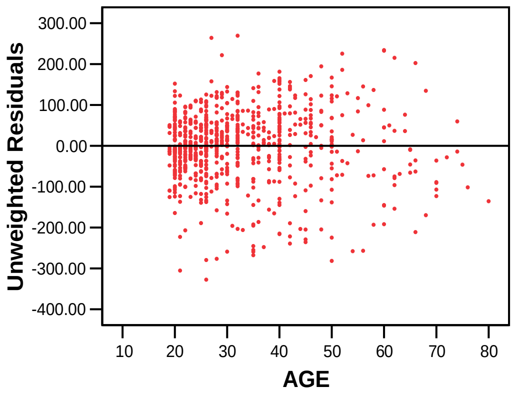

Figure 4: Plot of Age vs unweighted residuals

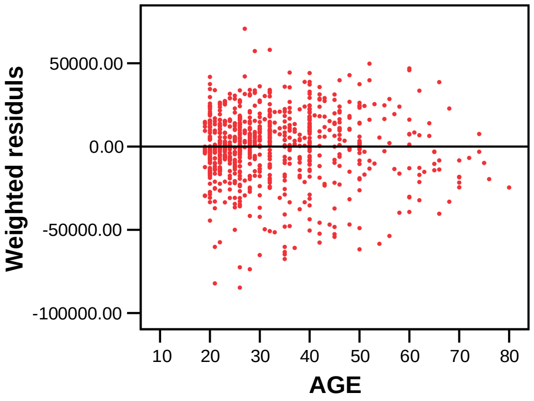

Figure 5: Plot of Age vs weighted residuals

7. Conclusion

Our purpose in studying this example is to illustrate that efficient estimation, such as weighted least squares in the presence of heteroscdeasticity can improve the quality of the investigative process. Also when one of the independents is categorical (dichotomous in our example) the requirement of separate prediction models for the different factors can be assessed by the tests for coincidence and parallelism. As such it is found that in the same population, different prediction models for peak expiratory flow rate are not required for the two sexes.

References

1. Anscombe, F.J., Tukey, J.W. (1963). The examination and analysis of residuals. Technometrics, **5**, 141-60.
2. Atkinson, A.C. (1982). Regression Diagnostics, Transformations and Constructed Variables (with discussion). Journal of the Royal Statistical Soc., Ser. B**, 44**, 1-36.
3. Atkinson, A.C. (1986). Diagnostic Tests for Transformations, Technometrics, **25**, 29-38.

4. Box, G.E.P.and Cox, D.R. (1964). The Analysis of Transformations. J. Roy. Statist. Soc., Series B, **26**, 211-252.
5. Box, G.E.P.and Hill, W.J. (1974). Correcting Inhomogeneity of variance with Power transformation weighting. Technometrics, **16**, 385-389.
6. Chatterjee, S. and Price, B. (1977). Regression analysis by example, New York: John Wiley.
7. Cook, R.D. and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression. Biometrika, **70**, 1-10.
8. Draper, N.R. and Smith, H. (1966). Applied Regression Analysis. 3rd edn. John Wiley and Sons, Inc., New York.
9. Durbin, J. and Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression I, Biometrika, **37**, 409-428.
10. Gregg, N. J. (1973). Peak Expiratory Flow Rates in Normal Subjects, British Medical Journal, **3**, 282-284.
11. Harvey, A.C. (1990). The Econometric Analysis of Time Series, Second Edition, MIT Press.
12. Johansson, Z.M, and Erasmos, L. D. (1968),.Clinical Spirometry in Normal Bantu. Amer. Rev. Respiratory Diseases, **97,** 585-589.
13. Kendall, M.G. and Stuart, A. (1966). The Advanced Theory of Statistics, Vol 3. Griffin & Co., London.14. Kleinbaum, D.G., Kupper, L.L., Muller, K. E. (1988). Applied Regression Analysis and Other Multivariable Methods. PWS- KENT Publishing Co., Boston,.
15. Raju, P. S., Prasad, K.V., Ramana, Y.V., Ahmed, S. K., Murthy, K.J. (2003). Study on lung function tests and prediction equations in Indian male children. Indian Journal of Pediatrics., **40,** 705-11.
16. Raju, P. S., Prasad, K.V., Ramana, Y.V., Murthy, K.J. (2004). Pulmonary function tests in Indian girls-Prediction equation. Indian J Pediatr., **71**, 893-97.
17. Selby, T., Read, J. (1961). Maximum expiratory flow rate in Australian adults. Austr Ann Med, **10**, 49-53.
18. Theil, H. (1971). Principles of econometrics. New York, Wiley.
19. Chih-Ling, Tsai; Xizhi , Wu. (1990). Diagnostics in Transformations and Weighted Regression, Technometrics, **32,** 315-322.
20. Woolcock, A.J., Colmen, M.H., Blackburn, C. R. B. (1972). Factors affecting normal values for ventilatory function. Amer. Rev Respir Dis, **106**, 692-709.