

## Modeling Transformed Health Care Cost with Unknown Heteroskedasticity

O. Başer\*

*Assistant Professor of Surgery, University of Michigan, Ann Arbor, MI, 48104*  
*President, STATinMED Research, Ann Arbor, MI, 48104*  
*(formerly an employee of Thomson Healthcare)*

### SUMMARY

Log models are widely used to deal with skewed outcome such as health expenditure. They improve the precision of the estimates and diminish the influence of outliers. Retransformation is generally required after estimation and the evidence of heteroskedasticity complicates the process. Smearing estimation suggested in the literature only works for homoskedastic errors or heteroskedastic errors due to categorical variables. Generalized linear models have been proposed as an alternative approach for log models when there exists unknown forms of heteroskedasticity. Recent literature shows that log models are superior to generalized linear models under certain conditions. We present a method for applying transformation that accounts for any form of heteroskedasticity. Our proposed model assumes that errors achieve normality. Heteroskedasticity is modeled separately. Simulation studies are conducted. We also used the Medstat MarketScan Database to estimate healthcare costs for asthma patients. Finally, a comparison of the method with smearing estimators and generalized linear model (GLM) estimators is established. Log-transformed health care costs of asthma patients were normal. There was an evidence of heteroskedasticity. Confirming the simulation study, heteroskedasticity adjusted retransformed costs had the lowest mean squared error relative to estimators from smearing retransformation or generalized linear model. This study shows that if log-transformed costs are normally distributed, heteroskedasticity adjusted retransformation produces more efficient results.

Copyright © 2007

### 1. INTRODUCTION

The use of a log transformation is common practice for dealing with skewness in the health care cost. Although log transformation improves precision and diminishes the influence of outliers, log scale results, per se, are of little interest. Policy decisions are almost always based on actual dollar values requiring retransformation of log scaled predictions.

Retransformation presents no problem when errors achieve linearity, normality and homoscedasticity assumptions. When one of these does not hold, retransformation bias arises when we try to revert back to the original scale. Since the log-transformed model results in

---

\*Correspondence to: 110 Depot Street, Ann Arbor, MI, 48104

†Tel: 734-646-7991; Fax: 734-332-4246; Email: obaser@statinmed.com; web: <http://www.statinmed.com>

geometric means rather than arithmetic means, log scale predictions may provide a biased estimate of the impact of an explanatory variable on the arithmetic mean.

The complications of using log transformed models are well documented in the literature. [1, 2, 3, 4, 5, 6] Duan[1] proposed a smearing transformation which achieves linearity and homoskedasticity, but not necessarily normality on a known transformation. Carroll and Rubert[2] relaxed known transformation assumptions and let the data guide the transformation type. Taylor[3] proposed parametric estimation when transformation additionally achieves normality. Duan et al. [4] proposed extension of the smearing transformation which covers heteroskedasticity due to categorical variables. Manning[5] clearly presented evidence that the failure to account for heteroskedasticity can have misleading policy implications. Mullahy[6] showed that a homoscedasticity assumption after nonzero responses yields inconsistent inferences about important policy parameters. Our purpose in this paper is to present a methodology which enables us to handle retransformation bias when transformation achieves linearity, additivity and normality but not necessarily homoscedasticity. In the presented model, heteroskedasticity can be due to categorical or continuous variables, and the form of heteroskedasticity does not need to be known.

Heteroskedasticity is not an exception in health care costs. Cost responses on the original scale (actual dollars) are usually an increasing function of the mean. Transforming the dependent variable reduces the heteroskedasticity in the response on the scale of estimation, but may not be sufficient to eliminate all of the heteroskedasticity in the error term. In practice, the form of heteroskedasticity is rarely known. Therefore common practice is to switch to generalized linear models (GLM) from log-transformed models after detection of heteroskedasticity. [7] However, GLM estimators can be quite imprecise if the log-scale error was symmetric but heavy tailed or the log-scale error variance is greater than one.[7, 8] Generalized Gamma regression is an alternative for estimating skewed dependent variables under less restrictive assumption. [9] Basu and Rathouz [11] extended the model where the link function is determined by the data. Overall, the primary reason for switching to models other than log-transformed was due to unknown heteroskedasticity. Here we will show that this does not have to be the case under the assumption of normality.

We will review a heteroskedasticity test and present the appropriate transformation in Section 2. In Section 3, simulation studies show that a retransformation of the log transformed model; when adjusted for unknown form of heteroskedasticity, has good finite properties. We illustrate the application of the retransformation to Medstat Market Scan Data. We compare our estimates with smearing estimates (ignoring heteroskedasticity) and GLM estimators with Gaussian family and logarithmic link according to average squared prediction error method and mean squared error method.

## 2. General Framework and Methods

Assume that there is a population represented by the random vector  $(\mathbf{x}, y)$  where  $\mathbf{x}$  is a  $1 \times K$  vector of explanatory variables (ex. patient and clinical characteristics),  $y$  is the scalar response variable (ex. health care costs), and  $\boldsymbol{\beta}$  a  $K \times 1$  vector of unknown parameters.

Suppose that the population model of interest is

$$\log(y) = \beta_1 + \beta_2 x_2 + \dots + \beta_K x_K + u = \boldsymbol{\beta} \mathbf{x} + u, \quad (1)$$

where we define  $x_1 = 1$  and  $u$  as the error term. Heteroskedasticity exist if

$$Var(u|\mathbf{x}) = \sigma^2 h(\mathbf{x}), \tag{2}$$

where  $h(\mathbf{x})$  is some function of the explanatory variables that determines the heteroskedasticity. In practice, the Breusch-Pagan test helps us to detect heteroskedasticity.[12] Let  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, N$  be a random sample from the population. Then,

- Estimate the model (1) and obtain the squared residuals,  $\hat{u}_i^2$ .
- Estimate the equation:

$$\hat{u}_i^2 = \delta_1 + \delta_2 x_{i2} + \dots + \delta_K x_{iK} + \epsilon_i. \tag{3}$$

- Form F-statistics and compute the p-value.
- If the p-value is sufficiently small, there is evidence of heteroskedasticity.

If heteroskedasticity is due to one of the categorical variables, extension of smearing can be applied [4]. But if there are several variables significant in equation (3) or significance is due to some continuous variables, or they are jointly significant, we have to account for general heteroskedasticity in order to eliminate retransformation bias. Next, we will show how to do this with normality of log transformed dependent variable.

*2.1. Adjustment for General Heteroskedasticity with variant of Park Test*

It has been shown that expectations of  $y$ , under the assumption of  $u \sim N(0, \sigma^2 h(\mathbf{x}))$  is [16]

$$E(y|\mathbf{x}) = \int e^{(\beta\mathbf{x}+u)} dF(u) = e^{\beta\mathbf{x}} \int e^u dF(u) = e^{\beta\mathbf{x}+0.5\sigma^2 h(\mathbf{x})}. \tag{4}$$

Therefore, for the unbiased transformation, we need the estimates of  $\beta, \sigma$  and  $h(\mathbf{x})$ . The estimation  $\beta$  is simply the estimated coefficients from equation (1). We will model  $\sigma^2$  and  $h(\mathbf{x})$  and use data to estimate unknown parameters in the model.

Assume equation (3) holds and

$$h(\mathbf{x}) = exp(\delta_1 + \delta_2 x_2 + \dots + \delta_K x_K). \tag{5}$$

We choose exponential function to ensure the positivity of variance function, but one can assume other parametric or nonparametric approaches to model  $h(\mathbf{x})$ . Then we can write

$$u^2 = \sigma^2 h(\mathbf{x})v = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k)v, \tag{6}$$

where  $E(v|\mathbf{x}) = 1$ . With the assumption of independence between  $\mathbf{x}$  and  $v$ , we can write

$$log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + e, \tag{7}$$

where  $\alpha_0 = log(\sigma^2) + \delta_0$ ,  $e$  has a mean of zero and is independent of  $\mathbf{x}$ . We need to replace unobserved  $u$  with the residuals  $\hat{u}$  for estimation of equation (7). Let  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, N$  be a random sample from the population. We summarize the steps:

- Run regression  $log(\hat{u}_i^2)$  on  $x_{i1}, x_{i2}, \dots, x_{iK}$ , and estimate fitted values  $\hat{f}_i$ .
- Estimate the  $\sigma_i^2 h_i(\mathbf{x})$  as  $exp(\hat{f}_i)$ .
- Then the unbiased transformation is :  $E(y_i|\mathbf{x}_i) = exp(\hat{\beta}\mathbf{x}_i + \frac{1}{2}exp(\hat{f}_i))$ .

## 3. Simulation

To compare the proposed method with the smearing transformation ignoring heteroskedasticity, and GLM, a Monte Carlo simulation was conducted. We investigated how well the estimate minimized the residual error by using the mean square error (MSE). For each replicate  $r$ , the

$$MSE = \frac{1}{N} \sum (y_{ri} - \hat{y}_{ri}). \quad (8)$$

We also compare these three models with a cross-validation type technique.[1, 4] We randomly split the sample parts, the training subsample and test subsample; each of the seven models are fitted on the training subsample and then used to form predictions for all individuals in the test subsample. The prediction of each individual's expenditure is compared with that individual's actual expenditure. The average squared prediction error (ASPE) was then computed for each model:

$$ASPE = \frac{1}{N} \sum (E(Y_k) - Y_k)^2,$$

where  $E(\cdot)$  is the expected value of the term in the parenthesis, the index  $k$  runs through the  $m$  individuals in the test subsample. The ASPEs of different models can be compared directly, the model producing smaller ASPE being the better one.

Each model is evaluated on 1000 random samples, with each having a sample size of 10000. We hold the specific draws of underlying random numbers constant when comparing the estimators. This process is shown to be efficient since it decreases the Monte Carlo simulation variance. [7] All models are evaluated in each replicate of a data generating mechanism.

We focus on the case in which there is only one regressor in the equations. Therefore, equation (7) is described by  $\log(u^2) = \alpha_0 + \delta_1 x_1 + \epsilon$  and equation (1) is described by  $\log(y) = \beta_0 + \beta_1 x_1 + u$ , where we set  $\alpha_0 = 0, \delta_1 = \beta_0 = \beta_1 = 1$  and simulate  $x_1 \sim U(0, 5)$  and  $\epsilon \sim N(0, 1)$ .

Simulation results are listed in Table 1. Confirming Manning & Mullahy [7], we find evidence that GLM can be very imprecise. Some of the GLM with a Gamma family and log link did not converge, yielding unresanable mean squared error (MSE). Again confirming Manning [5], smearing retransformation ignoring heteroskedasticity yielded bias estimates. Heteroskedasticity adjusted retransformation had the lowest MSE and ASPE.

Table I. Simulation Results

Dependent Variable	Bias	Std. Err.	MSE	ASPE
Heteroskedasticity Adjusted Retransformation	8.61	18.63	60026.91	20352.82
Smearing Retransformation	18.96	28.14	158265.80	98547.22
GLM Model with Gamma Family and Log Link	103931	302056	2.15e+13	1.47e+13
GLM Model with Gaussian Family and Log Link	43.79	60.37	115832.14	45874.85

#### 4. An empirical example

This example draws upon an analysis, which sought to estimate the health care expenditures of patients with asthma. We choose this example because it involved health care expenditure data which is known to suffer heteroskedasticity and require log-transformation.

##### *4.1. Data source and sample*

Data for this study came from the Medstat MarketScan Commercial Claims and Encounters Database. In 2005, this database contained information on approximately thirteen million persons who were covered by private insurance. The following five variables were available in this database: age, gender, International Classification of Diseases, 9th revision (ICD-9), plan type and geographic region. Charlson Comorbidity Index and major diagnosis counts are calculated from the data source. Our analytic sample consisted of persons who satisfied the following characteristics:

- Had at least two outpatient claims with a primary or secondary diagnosis of asthma; or
- Had at least one emergency room claim with a primary diagnosis of asthma, and a transaction for an asthma drug 90 days prior to, or 7 days following, the emergency room claim, or
- Had at least one inpatient claim with a primary diagnosis of asthma; or
- Had a secondary diagnosis of asthma and a primary diagnosis of respiratory infection in an outpatient or inpatient claim, or
- Had at least one drug transaction for anti-inflammatory agents, oral anti-leukotrienes, long-acting bronchodilators or inhaled or oral short-acting beta-agonistics.

To ensure that individual records were complete, and that the analytic sample would be representative of the population of patients of interest, a number of exclusions were imposed. Individuals were excluded if they:

- Had a diagnosis of chronic obstructive pulmonary disease, emphysema or chronic bronchitis,
- Were pregnant at some stage during the study period,
- Were not continuously enrolled in the health plan for 24 months,
- Were in health Maintenance Organizations (HMOs) and capitated point of service (POS) plans,
- Were elderly, defined as ages 65 and over.

The dependent variable was the log transformed total health expenditure calculated as the sum of inpatient; outpatient, and pharmaceutical expenditures for all medical services. This included all services paid for by insurance, as well as co-payments and deductibles paid out-of-pocket. These values are derived from MarketScan Data.

A further, more in-depth discussion of these variables and their creation can be found elsewhere in the literature [13].

##### *4.2. Analysis*

We estimated total expenditures as a function of patients' demographic and clinical factors. In particular, the following variables were posited as explanatory: age, gender, geographic region,

overall medical mental health comorbidities (number of unique ICD-9 codes), and Charlson Comorbidity Index.

The Shapiro-Wilk W test is used to determine normality of the residuals. [14] Evidence for heteroskedasticity is detected using the variant of Park test described in equation 7.

The retransformation of estimated log transformed cost is adjusted for heteroskedasticity following the steps described in the Methods section. In order to compare our results, we did the retransformation using the smearing estimation (ignoring heteroskedasticity) and we used GLM models which do not require transformation. The Park test described in Manning and Mullahy [7] suggested that if GLM is chosen, the family should be Gaussian, therefore we used Gaussian family with log link in our GLM estimation.

#### 4.3. Results

The final analytic sample, which was used for subsequent analysis, consisted of 1184 patients for the year 1999. This sample decreased to 1132 after removing outliers determined by Cook's

Table II. Summary of Patient Characteristics

Variable	Mean	Std. Err
Age	27.86	16.17
Female	0.51	0.50
Capitated	0.72	0.44
North Central	0.24	0.43
South	0.38	0.48
West	0.10	0.30
Other Region	0.02	0.13
MDC count	5.38	2.78
CCI index	0.98	0.85

Table III. Estimation of Log-Transformed Cost

Variable	Coefficients	Std. Err	P-values
Age	0.01	0.001	0.00
Female	0.03	0.05	0.43
Capitated	-0.33	0.05	0.00
North Central	0.101	0.06	0.12
South	0.09	0.06	0.10
West	-0.18	0.08	0.04
Other Region	-0.50	0.18	0.01
MDC count	0.18	0.02	0.00
CCI index	0.18	0.03	0.00
N=	1132		
<i>Prob &gt; F</i>	0.000		
$R^2$	0.485		

distance[15]. Table 2 presents the characteristics of the analytical sample. Sample members were on average 28 years of age and balanced between male and female. 72% of the sample belonged the capitated plan. Patients reside mostly in South. In terms of clinical factors, the mean for overall medical mental health comorbidities was 5.4, and the mean Charlson Comorbidity Index was 0.97.

The Shapiro-Wilk  $W$  test for normality failed to reject that the residuals are normally distributed ( $Prob > Z = 0.21034$ ).

Estimation of log transformed-cost are given in Table 3. As expected age, higher number of overall medical mental health comorbidities, and Charlson Comorbidity Index significantly increased the health care expenditure. The effect of belonging to a capitated plan and residing other regions as opposed to North Central (used as reference) was negative and significant. Coefficients were jointly significant ( $Prob > \chi^2 = 0.000$ ) and  $R^2$  suggested that 48.5% of variation is explained with the set of covariates used in the analysis. Table 4 shows the effect of each covariate on the variance. Age, MDC count, and residence in "other regions" are the main contributors to heteroskedasticity. F-statistics show that the set of covariates jointly contributes to heteroskedasticity. We analyzed the residuals of the regression presented in table 4, there was no evidence of heteroskedasticity.

The retransformation results are presented at Table 5. The first row shows the unadjusted total health care expenditure for asthma patients, the second row shows the results with adjusted unknown heteroskedasticity according to the method described in this paper. The smearing transformation ignoring heteroskedasticity is shown in the third row, and GLM

Table IV. Heteroskedasticity Test [Equation 7]

Variable	Coefficients	Std. Err	P-values
Age	-0.003	0.001	0.04
Female	0.05	0.05	0.25
Capitated	-0.08	0.05	0.09
North Central	-0.02	0.06	0.71
South	-0.01	0.06	0.85
West	-0.13	0.08	0.11
Other Region	-0.48	0.18	0.01
MDC count	-0.026	0.01	0.00
CCI index	0.89	0.08	0.00
$Prob > F$	0.000		

Table V. Retransformed Health Care Expenditures

Model	Mean	Std. Err	ASPE	MSE
Unadjusted	\$ 7,686	\$ 11,3556		
Case 2	\$ 8,352	\$9,690	5.37+07	4.31+07
Smearing	\$ 8,066	\$9,979	5.69+07	4.89+07
GLM	\$ 7,310	\$9,786	5.59+07	4.46+07

results are presented in the last row. The method described in this paper has the lower standard error, ASPE and MSE relative to the others.

## 5. Discussion

One of the biggest disadvantages of log-transformed cost estimation is the retransformation problem since the interest lies in level cost rather than log-transformed cost. Duan's smearing estimation solves the problem when errors are homoskedastic. If heteroskedasticity is due to categorical variables, the extended version of smearing is available. In practice however, the form of heteroskedasticity is rarely known. Therefore a log-transformed model is not appropriate unless suitable transformation is applied. GLM models are the only alternative but the precision of these models diminish more as kurtosis increase. This study shows that if log-transformed costs are normally distributed, there is a way to account for any case of heteroskedasticity, allowing consistent retransformation and producing more efficient results than those produced by GLM or smearing methods that ignore heteroskedasticity.

When the health care cost data contain zero values, retransformation bias can be treated with the model proposed by Welsh and Zhou[18]. They presented two distribution free estimators to estimate the mean of a dependent variable after fitting a semiparametric two-part heteroskedastic regression model to a transformation of the dependent variable. Zhou et al. [19] applied the model to estimate VA total health care cost.

Although our study uses residuals for heteroskedasticity test it has been showed that using the residuals themselves for heteroskedasticity might be problematic. For small to moderate sizes or cases where some of the characteristics in explanatory variables are rare, the residuals will give the appearance of heteroskedasticity when there is none in the underlying error. Koenker suggested that one should use either the standardized or studentized residuals, multiplied by the appropriate estimate of the variance. Since we have a relatively large sample, we believe the differences would be minor.

In our study we removed the outliers from the sample. Determination of outliers is a difficult task. The deletion of outliers might be inappropriate unless they are known to be erroneous data and leads to systematic bias in the estimates. This is particularly an issue for health expenditures, where a small number of cases account for a substantial amount of raw mean, and where there may be some skewness in the residuals after log transformation. These catastrophic cases often real and important. Therefore we did sensitivity analysis where we run our model with and without outliers. The results were not sensitive to inclusion of outliers in the sample.

## REFERENCES

1. Duan, N. Smearing estimate: a nonparametric retransformation method. *J. Amer. Statist. Assoc.* 1983;78:605-610.
2. Carroll RJ, Ruppert D. Comment on "The analysis of transformed data" by D.V. Hinkley, G. Runger. *J. Amer. Statist. Assoc.* 1984;79:312-313.
3. Taylor JMG. The transformed mean after a fitted power transformation. *J. Amer. Statist. Assoc.* 1986;81:114-118.
4. Duan N, Manning WG, Willard G, et al. A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Studies.* 1983; 1:115-26.
5. Manning WG. The logged dependent variable, heteroskedasticity, and the retransformation problem. *Journal of Health Economics.* 1998;17: 283-295.
6. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econometrics.* 1998;17:247-281.
7. Manning WG, Mullahy H. Estimating log models: to transform or not to transform? *Journal of Health Economics.* 2001; 20:461-94.
8. Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics.* 2004; 23: 525-42.
9. Manning WG, Basu A, Mullahy J. Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data. *Journal of Health Economics.* 2005; 24: 465-88.
10. Basu A, Manning WG, Mullahy J. Comparing alternative model: log vs Cox proportional hazard? *Health Economics.* 2004;13(8):749-65.
11. Basu A, Rathouz P. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics.* 2005;6(1):93-109.
12. Breusch T, Pagan A. A Simple test for heteroscedasticity and random coefficient variation. *Econometrica.* 1979;47: 1287-1294.
13. Crown WH, Berndt ER, Baser O, et al. Benefit plan design and prescription drug utilization among asthmatics: Do patient copayments matter? In: Cutler D, Garbe AM, eds. *Frontiers in Health Policy Research.* Cambridge: The MIT Press. 2004:95-127.
14. Shapiro SS, Wilk MB. An analysis of variance test for normality. *Biometrika.* 1965;52(3):591-599.
15. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics,* New York, NY: John Wiley & Sons, Inc; 1980.
16. Wooldridge JM. *Introductory Econometrics: A Modern Approach.* South-Western College Publishing. 2000.
17. Koenker R, Gilbert B. Robust Tests for Heteroskedasticity based on regression quantiles. *Econometrica.* 1982;50(1):43-61.
18. Welsch AH, Zhou XH. Estimating the retransformed mean in a heteroskedastic two-part model. *Journal of Statistical Planning and Inference.* 2006; 136:560-881.
19. Zhou XH, Qin G, Maciejewski ML. Estimating the VA total health care cost using a semi-parametric heteroscedastic two-part model.