STATE OF MAINE
DIRIGO HEALTH AGENCY

RE:	DETERMINATION OF	)
	AGGREGATE MEASURABLE	)
	COST SAVING FOR THE FOURTH	)
	ASSESSMENT YEAR (2009)	)

MAINE CHAMBER EXHIBIT # 11

Multiple regression of cost data: use of generalised linear models

Julie Barber, Simon Thompson[1]

Research and Development Directorate, University College London Hospitals R&D Directorate, London; [1]MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Objective: **Choosing an appropriate method for regression analyses of cost data is problematic because it must focus on population means while taking into account the typically skewed distribution of the data. In this paper we illustrate the use of generalised linear models for regression analysis of cost data.**

Methods: **We consider generalised linear models with either an identity link function (providing additive covariate effects) or log link function (providing multiplicative effects), and with gaussian (normal), overdispersed poisson, gamma, or inverse gaussian distributions. These are applied to estimate the treatment effects in two randomised trials adjusted for baseline covariates. Criteria for choosing an appropriate model are presented.**

Results: **In both examples considered, the gaussian model fits poorly and other distributions are to be preferred. When there are variables of prognostic importance in the model, using different distributions can materially affect the estimates obtained; it may also be possible to discriminate between additive and multiplicative covariate effects.**

Conclusions: **Generalised linear models are attractive for the regression of cost data because they provide parametric methods of analysis where a variety of non-normal distributions can be specified and the way covariates act can be altered. Unlike the use of data transformation in ordinary least-squares regression, generalised linear models make inferences about the mean cost directly.**

Journal of Health Services Research & Policy Vol 9 No 4, 2004: 197–204

Introduction

Regression methods form an important part of many statistical analyses of health care data, including adjusting treatment effects in randomised trials, adjusting for case-mix in observational studies and developing models for prediction. For cost data, choice of regression method is not straightforward. It is important to focus analysis on population mean costs, while using methods suitable for the often skewed distribution of cost data.[1–4]

When used for costs, assumptions of standard ordinary least-squares (OLS) linear regression are unlikely to be met. In particular, costs are usually non-normal and heteroscedastic (i.e. not of constant variance) and relationships may not be truly linear.[2] Violation of OLS assumptions may mean that normality and efficiency of estimators are not achieved, so not providing the best estimates of the average effects in the population. In order to meet model assumptions, it is tempting to transform the costs and apply OLS analysis (e.g. using log-transformed costs). Such an approach is flawed because analysis is then not concerned with population means but with the mean on an alternative scale. For example, log-transformed data provide an analysis of geometric mean costs[5] unless sophisticated back-transformation methods are adopted.[6]

In this paper, the usefulness of generalised linear models (GLMs) for regression analyses of cost data is described. These models provide appropriate and flexible methods for analysis of mean costs that explicitly allow for the non-normal distribution of the data. Some examples are discussed in detail, in the context of adjusting treatment effect estimates from randomised trials. Methods for adjusting effects in observational studies and development of models for predicting costs raise closely related issues.

Julie Barber PhD, Lecturer in Medical Statistics, Research and Development Directorate, University College London Hospitals R&D Directorate, 1st Floor, Maple House, 149 Tottenham Court Road, London W1P 9LL, UK. Simon Thompson DSc, Director, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Correspondence to: JB.

Methods

Generalised linear models (GLMs)

GLMs have a variety of forms characterised by two features: a distribution function (F) for the outcome data (in this case costs) and a link function (g) which describes the scale on which covariates in the model are related to the outcome.[7] Suppose y_i is the observed outcome (cost) for individual $i = 1 \ldots n$ and $\mu_i = E(y_i)$ is its expected (mean) value from the model. If there are K covariates, and writing x_{ik} as the kth covariate for subject i, the general form of a GLM is:

$$g(\mu_i) = x_i\beta,\ y_i \sim F$$

where $x_i\beta = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik}$ and β_0 (the intercept) and $\beta_1 \ldots \beta_K$ are the $K+1$ regression coefficients. This family of models is particularly attractive for costs because, in terms of both prediction and estimation of effects, focus is always on the means μ_i. Also, since the distribution function can be chosen as any from the 'exponential family', various distributions for the data can easily be specified. GLMs are weighted regression models[7] that can be fitted in most standard statistical software packages.

The standard OLS model

$$\mu_i = x_i\beta,\ y_i \sim \text{gaussian}$$

is a GLM having an identity link function and a gaussian (normal) distribution function. The identity link means that covariates act additively on outcome, the coefficients describing the change in mean cost per unit increase in a covariate (e.g. per year increase in age). For cost data, it would be more appropriate to employ a skewed distribution function, such as a gamma or inverse gaussian distribution.[8]

Changing the distribution function, but still using the identity link, leaves interpretation of the coefficients unchanged from the OLS model. Changing the link function of the GLM alters the way in which covariates are assumed to act on the outcome and thus alters the interpretation of the coefficients. For example, a log link will provide GLMs of the form

$$\log(\mu_i) = x_i\beta,\ y_i \sim F$$

or equivalently

$$\mu_i = \exp(x_i\beta),\ y_i \sim F$$

In this model, covariates act multiplicatively on the mean, which in some cases may be more realistic than assuming additive effects. An exponentiated (anti-logged) coefficient provides a ratio of means, which can be re-expressed as the percentage increase in mean cost per unit increase in the covariate (e.g. per year increase in age). Interpretation of such estimates is different from the ratio of geometric means provided by analysis of log-transformed data; these are not of interest in analysis of costs.[5] Other possible links useful for costs might include reciprocal and power functions.[7] Although for some datasets these may provide better fitting models, they are less attractive since interpretation of coefficients on such scales is not straightforward.

The GLM approaches above require specification of a distribution function that fully describes the distribution of the outcome, including its shape and the relationship between its variance and mean: the variance, $\text{var}(y_i)$, is proportional to some function of μ_i, called the variance function. For example, the variance function for the gaussian distribution is 1, because variance and mean are independent, whereas for the skewed gamma and inverse gaussian distributions the variance functions are quadratic ($\text{var}(y_i) \propto \mu_i^2$) and cubic ($\text{var}(y_i) \propto \mu_i^3$), respectively. The flexibility of GLMs can be extended further to fit models without specifying the exact distribution for the response variable. This involves a quasi likelihood approach[7] that requires only specification of link and variance functions. This allows fitting models with power variance functions of the form $\text{var}(y_i) \propto \mu_i^\lambda$ which are not covered by the exponential family. For example, $\lambda = 1$ (overdispersed poisson model) or $\lambda > 3$ may be useful for cost data.

Examples

The usefulness of GLM methods is illustrated using data from the large UK700 trial ($n = 667$) and from a smaller trial of child self-poisoning ($n = 149$).

The UK700 trial was a large multi-centre randomised trial designed to investigate the cost-effectiveness of intensive (caseloads of 10–15 patients) compared with standard case management (caseloads of 30–35 patients) in caring for mentally ill patients in the community.[9,10] The trial randomised 708 patients who were followed up for 2 years. Information collected on the use of hospital and community services, contacts with the general practitioner, staffed accommodation and drop-in centre use was used to estimate a total two-year cost per patient. Complete cost data were available for 667 (94%) patients (Table 1).

The child self-poisoning trial aimed to assess the cost-effectiveness of a home-based social work intervention for young people who have deliberately poisoned themselves.[11,12] The trial randomised 162 children to either routine care or routine care plus social work intervention. Over the six months after randomisation, information on the use of health education, social and voluntary services was collected and used to estimate a total cost for each patient (Table 1).

These examples are typical of cost data in practice, being severely positively skewed: the means are substantially greater than the medians (Table 1).

Table 1 Summary of two-year costs in the UK700 trial and six-month costs in the self-poisoning trial

Costs (£)	Mean	SD	Median	Interquartile range
UK700 trial[a]				
Intensive case management ($n = 336$)	24553	23408	15911	6079 to 38438
Standard case management ($n = 332$)	22704	22000	14735	4428 to 37366
Self-poisoning trial[b]				
Social work intervention ($n = 74$)	1455	1587	1137	644 to 1602
Routine care ($n = 75$)	1752	2531	946	535 to 1627

[a]41 patients had missing cost data.
[b]13 patients had missing cost data.
SD, standard deviation.

Implementation

Only identity and log links are considered because these provide easily interpretable coefficient estimates. Distributions investigated are the gaussian, overdispersed poisson, gamma and inverse gaussian (corresponding to variance function powers $\lambda = 0,1,2,3$, respectively) with extensions in the second trial to other variance function powers based on quasi likelihood. Emphasis in all models is on the treatment effect estimate, based on intention-to-treat in the randomised trial, with confidence intervals derived from the usual asymptotic standard errors.

GLMs with defined distributions were fitted in Stata,[13] and in SAS[14] for quasi likelihood models. Models are compared by examining plots of deviance residuals and using the Akaike information criterion (AIC)[15] defined as:

$$-2(\log\text{likelihood}) + 2(\text{number of fitted parameters})$$

AICs provide an objective indication of which models are most appropriate for the data, with smaller AIC values indicating the preferred models. Differences in AIC of, say, 10 or more indicate strong evidence for the better fitting model.[16]

Results

The UK700 trial

Models with just treatment group

Table 2 shows results for analyses of the UK700 trial where the model includes only a covariate for randomised group. This model is given by:

$$g(\mu_i) = \beta_0 + \beta_1(\text{GROUP})_i,\ y_i \sim F$$

where $(\text{GROUP})_i = 1$ for the intensive group and 0 for the standard group. The treatment effect estimate, β_1, compares the intensive with the standard group and, although the mean cost is greater in the former, all results indicate no significant evidence of a difference.

Table 2 UK700 trial ($n = 687$): comparison of generalised linear models with identity and log links and various distributions for models with just a treatment group covariate[a]

	Treatment effect coefficient (£)	95% CI	AIC
Identity link			
Gaussian	1849	−1599 to 5297	15278
Overdispersed poisson	1849	−1597 to 5295	14836[b]
Gamma	1849	−1600 to 5298	14782
Inverse gaussian	1849	−1606 to 5307	14980
Log link (exponential of coefficients)			
Gaussian	1.08	0.93 to 1.25	15278
Overdispersed poisson	1.08	0.93 to 1.25	14836[b]
Gamma	1.08	0.93 to 1.25	14782
Inverse gaussian	1.08	0.93 to 1.25	14980

[a]All results use dispersion parameter estimated as Pearson χ^2/degrees of freedom.
[b]Calculated using extended quasi log likelihood.[5]
AIC, Akaike information criterion; CI, confidence interval.

For models with an identity link, the treatment coefficient estimate is the difference in arithmetic means; with a log link its exponential is the ratio of arithmetic means. Thus the mean cost is estimated as £1849 or 8% greater in the intensive group than in the standard group.

For a particular link function, estimates of treatment effect are the same whichever distribution function is used. This is because, for these simple models, the weights estimated in the GLM procedure are identical for all observations in a particular treatment group, and cancel out in estimating the regression coefficients. For a continuous covariate this would not be the case. The confidence intervals given in Table 2 differ very slightly depending on the distribution used; differences for log link models are not apparent because of the number of significant figures reported.

Figure 1 shows normal plots of deviance residuals for the various models with identity link (plots for log link models are identical). For appropriate models the deviance residuals should follow an approximately normal distribution (indicated by the straight line on the normal plots). Figure 1 clearly indicates how badly the gaussian distribution represents these data and that the gamma distribution may be much more appropriate. The AICs in Table 2 also indicate a strong preference for the gamma model. For this simple case, the AICs are identical for models with the same distribution regardless of the link function, because predicted values used in calculating the likelihoods are the same.

Models with treatment group and a continuous covariate

Results from fitting various GLMs, including covariates for treatment group and age at randomisation, are shown in Table 3. These models are of the form:

$$g(\mu_i) = \beta_0 + \beta_1(\text{GROUP})_i + \beta_2(\text{AGE})_i,\ y_i \sim F$$

where $(\text{AGE})_i$ is a continuous covariate measured in years. All models show no significant evidence of a treatment effect after adjusting for age but strong evidence of a negative association between age and cost.

Unlike the previous simple model (Table 2), estimates differ depending on the distribution used. From the gamma model, for example, the mean cost is estimated as £1533 or 7% greater in the intensive group than in the standard group, and is estimated to decrease by £2622 or 12% per decade of age. For both link functions, estimates for the gaussian distribution are substantially larger than those from other models. In the gaussian case, each observation is given equal weight, whereas for the other models observations with high predicted costs are down-weighted in calculating regression coefficients.

Plots of deviance residuals against fitted values for each identity link model (Figure 2) clearly show the gamma distribution to be most appropriate (this having

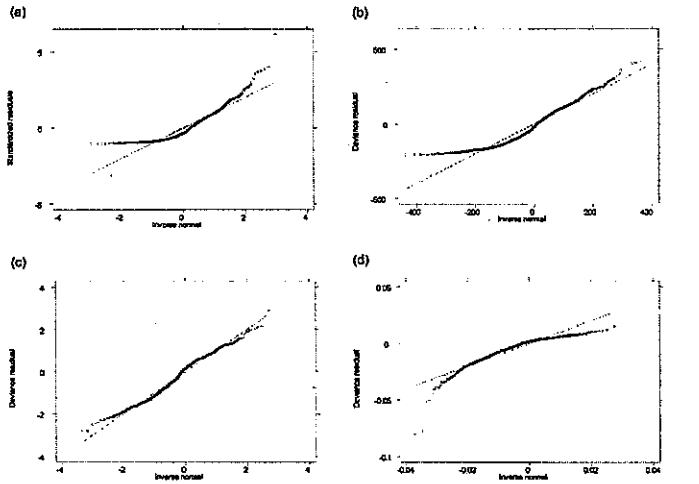


Figure 1 UK700 trial. Normal plots of deviance residuals for the identity link generalised linear models described in Table 2: (a) gaussian; (b) overdispersed poisson; (c) gamma; (d) inverse gaussian.

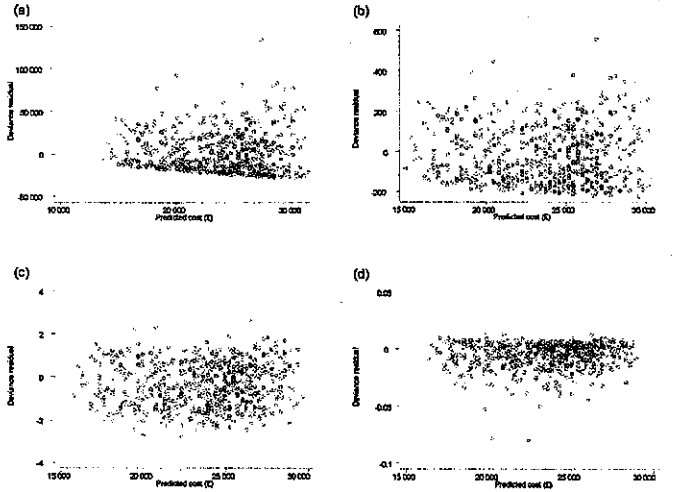


Figure 2 UK700 trial. Plots of deviance residuals against fitted values for models with identity link including treatment and age covariates (Table 3). Models with (a) gaussian, (b) overdispersed poisson, (c) gamma and (d) inverse gaussian distributions.

Table 3 UK700 trial ($n = 667$): comparison of generalised linear models with identity and log links and various distributions for models including covariates for randomised group and age[a]

	Treatment effect coefficient	95% CI	Age (per decade)	95% CI	AIC
Identity link					
Gaussian	2032	−1371 to 5435	−3324	−4812 to −1836	15259
Overdispersed poisson	1758	−1575 to 5091	−2918	−4236 to −1596	14816[b]
Gamma	1533	−1746 to 4813	−2622	−3975 to −1270	14765
Inverse gaussian	1361	−1877 to 4601	−2415	−3740 to −1091	15924
Log link (exponential of coefficients)					
Gaussian	1.10	0.95 to 1.27	0.84	0.79 to 0.90	15256
Overdispersed poisson	1.09	0.94 to 1.25	0.86	0.81 to 0.92	14813[b]
Gamma	1.07	0.93 to 1.24	0.88	0.82 to 0.93	14763
Inverse gaussian	1.07	0.93 to 1.23	0.89	0.84 to 0.95	15924

[a]All results use dispersion parameter estimated as Pearson χ^2/degrees of freedom.
[b]Calculated using extended quasi log likelihood.
AIC, Akaike information criterion; CI, confidence interval.

a random scatter of points indicating homoscedastic and normally distributed residuals). Plots for other distributions indicate severely non-normal residuals. Similar plots are obtained with a log link. AICs (Table 3) strongly support use of the gamma distribution, but do not allow clear conclusions regarding choice of link function.

Adjusting treatment effect for several covariates

The various GLMs were refitted adjusting the treatment effect for 11 baseline factors (listed in footnote of Table 4). With this number of covariates, GLMs with overdispersed poisson or inverse gaussian distributions and identity link did not converge.

The gaussian and gamma models for both links give very different estimates and confidence intervals for the treatment effect (Table 4). For example, with the identity link, the estimate for the gamma model is substantially larger and confidence interval much narrower than for the gaussian case.

Residual plots and AICs (Table 4) again indicate that the gamma distribution is most appropriate. In this analysis the AIC is lower for the gamma identity link than log link model, showing a convincing preference for the former.

Table 4 UK700 trial ($n = 512$): comparison of treatment effect estimates for generalised linear models with identity and log links and gaussian or gamma distributions[a]. Model with covariates for randomised group and 11 other baseline factors[b]

	Treatment effect coefficient	95% CI	AIC
Identity link			
Gaussian	1310	−2296 to 4917	11607
Gamma	2170	−545 to 4884	10563
Log link (exponential of coefficients)			
Gaussian	1.10	0.95 to 1.27	11610
Gamma	1.06	0.90 to 1.25	10762

[a]All results use dispersion parameter estimated as Pearson χ^2/degrees of freedom.
[b]Adjusted for centre, sex, age, ethnic group, source (from hospital or community), Comprehensive Psychiatric Rating Scale, Disability Assessment Schedule Score, social class, length of illness, number of days in hospital in previous two years, Research Diagnostic Criteria diagnosis.[9]
AIC, Akaike information criterion; CI, confidence interval.

The self-poisoning trial

Table 5 shows results for the self-poisoning trial where the treatment effect is adjusted for parental General Health Questionnaire (GHQ) score, a continuous variable. Coefficients compare the social work intervention group with the routine care group. Estimates of the treatment effect differ less than was seen for models including age in the UK700 example because of a weaker association between costs and GHQ score. All models indicate no evidence of a difference in mean costs between the randomised groups.

Although normal plots show that inverse gaussian is the best of the standard distributions (Figure 3), the plot for this model is not entirely satisfactory, and a variance function increasing even more rapidly with the mean may be preferable. A model fitted using quasi likelihood with a quartic variance function (power 4) shows a considerable improvement in residual plots (Figure 3) and AICs (Table 5). AICs do not, however, clearly distinguish between the log and identity links.

Discussion

In general, OLS models are not appropriate for cost data. Residuals are likely to be non-normal and not to have constant variance, relationships may not be truly linear and models could predict impossible negative values. Violation of OLS assumptions may mean that normality and efficiency of estimates are not achieved, so that inferences and predictions from such models are potentially misleading. Clearly, we would prefer a model that appropriately allows for the skewed nature of cost data. Basic statistical texts advise refitting the model following transformation of the data (e.g. a log transformation).[17] For cost data, however, such an approach is not generally appropriate, since analysis on transformed scales does not provide inferences about population mean costs which are of primary interest.

Table 5 Self-poisoning trial ($n = 149$): comparison of generalised linear models with identity and log links and various error distributions for models adjusting the treatment effect for General Health Questionnaire (GHQ) score[a]

	Treatment effect coefficient	95% CI	GHQ (per unit score)	95% CI	AIC
Identity link					
Gaussian	−311	−996 to 373	9.5	−34 to 53	2708
Gamma	−291	−961 to 378	7.5	−35 to 50	2550
Inverse gaussian	−283	−954 to 387	6.4	−35 to 48	2463
Quartic variance (var $\propto \mu^4$)	−282	−971 to 407	5.3	−37 to 48	2444[b]
Log link (exponential of coefficients)					
Gaussian	0.82	0.53 to 1.26	1.01	0.98 to 1.03	2708
Gamma	0.83	0.55 to 1.26	1.01	0.98 to 1.03	2550
Inverse gaussian	0.83	0.55 to 1.26	1.00	0.98 to 1.03	2463
Quartic variance (var $\propto \mu^4$)	0.84	0.56 to 1.27	1.00	0.98 to 1.03	2445[b]

[a]All results use dispersion parameter estimated as Pearson χ^2/degrees of freedom.
[b]Calculated using extended quasi log likelihood.
AIC, Akaike information criterion; CI, confidence interval.

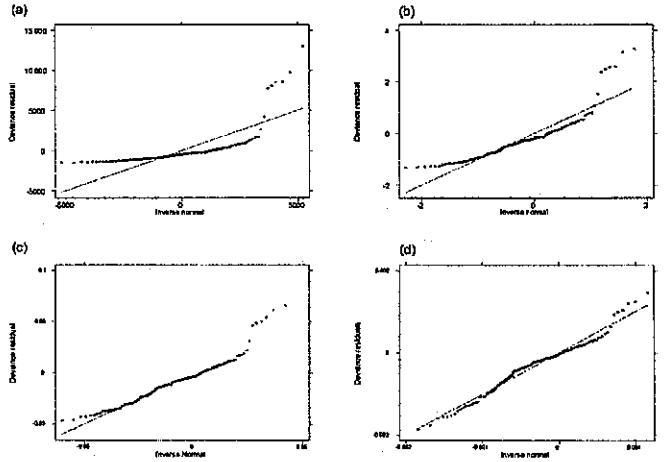


Figure 3 Self-poisoning trial. Normal plots of deviance residuals for the identity link generalised linear models described in Table 5: (a) gaussian, (b) gamma, (c) inverse gaussian and (d) quartic variance function.

Our examples have shown how GLMs provide a useful and flexible approach for regression of cost data. These models can allow appropriately for non-normal distributions, and covariate effects can be specified as, for example, additive or multiplicative. The gamma distribution clearly fitted the UK700 data better than any other distribution considered, and a quartic variance function provided an appropriate model for the child self-poisoning data. For simple comparisons of groups, use of these models compared with OLS had no impact on the treatment effect estimate, and had little impact on confidence intervals. In multiple regressions, or for a continuous variable, estimates of the treatment effect from the better fitting models sometimes differed substantially from those of OLS regression.

The flexibility of GLMs means that there are a number of possible link and variance function combinations to choose from. An appropriate variance function should be chosen, so that substantial losses in precision are avoided,[18] by comparing log likelihoods (or AICs) and residual plots. An alternative might be to use a modified Park test.[18] Choice of link function is more of a challenge. In this paper, only log and identity links have been considered because these provide easier interpretations for covariate effects, but comparisons of AICs provided little assistance in choosing link functions for

models including few covariates. Only when there are more plentiful (or more prognostic) variables in the model, such as the 12-covariate model for the UK700 example, does it become possible to distinguish the fit of different link functions to the data.

More detailed exploration of link and variance functions jointly could have been achieved using extended quasi likelihood, for example by fitting models with various power link and variance functions and comparing extended log likelihoods.[7,19,29] However, although the log normal distribution has been advocated for analysis of cost data,[3] it is not a member of the exponential family and so cannot be used within a GLM.

In our examples we have chosen to report confidence intervals based on the usual asymptotic estimates of standard error. There are, however, other approaches that could be used. For GLMs with non-normal distributions it has been suggested that likelihood-based confidence intervals are more appropriate,[15] particularly for smaller sample sizes. Where there is some doubt about model assumptions, confidence intervals could be derived using robust estimates of standard error.[21] These are calculated without assuming the model is true and relaxing distributional assumptions.

Non-parametric bootstrapping has been recommended previously as an appropriate check on inferences from standard normal distribution analyses of costs.[2,4,22] In using GLMs for costs, however, the aim is to identify well-fitting models so that bootstrapping is unnecessary. In cases where a satisfactory GLM cannot be found, there could be a role for bootstrapping the best-fitting (albeit not perfect) GLM. This would be preferable to a simple bootstrap of the OLS model, since the better-fitting GLM is likely to provide more appropriate coefficient estimates.

GLMs may be used for other more complex investigations. These include adjusting for case-mix in observational studies, assessment of covariate interactions, building predictive models for costs, or indeed for analysing cost data with a more complex structure, such as that from cluster randomised trials using the related generalised estimating equations approach.[23] A further important application of GLMs might be in adjusting net benefit analyses[24] for covariates. As for costs, interest is in analysis of population means, and typically net benefits have a skewed distribution.[25] Covariate-adjusted cost-effectiveness acceptability curves[26] could be derived from such analyses. Other cases for which GLMs could be useful include analyses of resource use, length of stay and sickness absence data.

Technical exposition of the use of GLMs for analysis of cost data has been presented previously.[18,20] This work focused mainly on prediction of costs in models with a log link, whereas the usefulness of GLMs with an identity link to provide direct estimates and confidence intervals for cost differences has received less attention. Other methods proposed for regression analysis of cost data include a retransformation technique called 'smearing' introduced by Duan.[6] This converts predicted geometric means back to estimated population means after an analysis on the log scale. The technique has been discussed and expanded on by others[18,27,28] and widely applied in practice. Use of two part models in cases where there are a large proportion of zero costs has also been discussed.[19,28,29] These methods have been concerned primarily with the prediction of expected costs, and do not directly provide simple estimates of how covariates affect population mean costs. For censored cost data, modified survival techniques have been described.[30] More recently, Manning and Mullahy[18] have compared regression approaches, including retransformation methods and GLMs with log links. They concluded that there are important trade-offs in terms of bias and precision depending on the method used; retransformation methods were shown to be biased for heteroscedastic variances and GLMs were imprecise if an appropriate variance function was not identified.

We conclude that GLMs are attractive for the regression of cost data because they provide parametric methods of analysis for which a variety of non-normal distributions can be specified and the way covariates act can be altered. Unlike the use of data transformation in OLS regression, GLMs make inferences about the mean cost directly as is appropriate for health economic decision-making.

References

1. Thompson SG, Barber JA. How should cost data in randomised controlled trials be analysed? BMJ 2000; 320: 1197–1200
2. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. Journal of Health Services Research & Policy 1998; 3: 233–245
3. Zhou X, Melfi CA, Hui SL. Methods for comparison of cost data. Annals of Internal Medicine 1997; 127: 752–756
4. Barber JA, Thompson SG. Analysis of cost data in randomised trials: an application of the non parametric bootstrap. Statistics in Medicine 2000; 19: 3219–3236
5. Barber JA, Thompson SG. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. BMJ 1998; 317: 1195–1200
6. Duan N. Smearing estimate: a non-parametric retransformation method. Journal of the American Statistical Association 1983; 78: 605–610
7. McCullagh P, Nelder JA. Generalized linear models. London: Chapman and Hall, 1989
8. Evans M, Hastings N, Peacock B. Statistical distributions. New York: Wiley, 1993
9. UK700 Group. The outcome of intensive versus standard case management for severe psychotic illness: the UK700 case management trial. Lancet 1999; 353: 2185–2189
10. UK700 Group. Cost-effectiveness of intensive versus standard case management for severe psychotic illness: the UK700 case management trial. British Journal of Psychiatry 2000; 176: 537–543
11. Harrington R, Kerfoot M, Dyer E, McNiven F, Gill J, Harrington V et al. Randomised trial of a home based family intervention for children who have deliberately poisoned themselves. Journal of the American Academy of Child and Adolescent Psychiatry 1998; 37: 512–518
12. Byford S, Harrington R, Torgerson DJ, Kerfoot M, Dyer E, Harrington V et al. Cost-effectiveness analysis of a home-based social work intervention for children and adolescents who have deliberately poisoned themselves: the results of a randomised controlled trial. British Journal of Psychiatry 1999; 174: 56–62
13. Stata reference manual: release 6. Texas: Stata Corporation, 1999
14. SAS reference manual: release 6.12. Cary, NC: SAS Institute Inc., 1996
15. Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. Statistics in Medicine 1998; 17: 59–68
16. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer-Verlag, 2002
17. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991
18. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? Journal of Health Economics 2001; 20: 461–494
19. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. Journal of Health Economics 1999; 18: 153–171
20. Blough DK, Ramsey SD. Using generalized linear models to assess medical care costs. Health Services and Outcomes Research Methodology 2000; 1: 185–202
21. Huber PJ. The behaviour of maximum likelihood estimates under non-standard conditions. Berkeley, CA: University of California Press. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967: 221–233
22. Desgagne A, Castilloux A, Angers J, LeLorier J. The use of the bootstrap statistical method for the pharmacoeconomic cost analysis of skewed data. Pharmacoeconomics 1998; 13: 487–497
23. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13–22
24. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Medical Decision Making 1998; 18 (suppl 2): S68–S80.
25. Hoch JS, Briggs A, Willan A. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Economics 2002; 11: 415–430
26. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Economics 2001; 10: 779–787
27. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 1998; 17: 283–295
28. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics 1998; 17: 247–281
29. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand of medical care. Journal of Business and Economic Statistics 1983; 1: 115–126
30. Lin DY. Linear regression analysis of censored medical costs. Biostatistics 2000; 1: 35–47

SUBJECT REVIEW

Regression Methods in the Empiric Analysis of Health Care Data

GRANT H. SKREPNEK, PhD

ABSTRACT

OBJECTIVE: The aim of this paper is to provide health care decision makers with a conceptual foundation for regression analysis by describing the principles of correlation, regression, and residual assessment.

SUMMARY: Researchers are often faced with the need to describe quantitatively the relationships between outcomes and predictors, with the objective of explaining trends, testing hypotheses, or developing models for forecasting. Regression models are able to incorporate complex mathematical functions and operands (the variables that are manipulated) to best describe the associations between sets of variables. Unlike many other statistical techniques, regression allows for the inclusion of variables that may control for confounding phenomena or risk factors. For robust analyses to be conducted, however, the assumptions of regression must be understood and researchers must be aware of diagnostic tests and the appropriate procedures that may be used to correct for violations in model assumptions.

CONCLUSION: Despite the complexities and intricacies that can exist in regression, this statistical technique may be applied to a wide range of studies in managed care settings. Given the increased availability of data in administrative databases, the application of these procedures to pharmacoeconomics and outcomes assessments may result in more varied and useful scientific investigations and provide a more solid foundation for health care decision making.

KEYWORDS: Claims database analysis, Pharmacoeconomics, Outcomes assessment, Regression analysis

J Manag Care Pharm. 2005;11(3):240-51

Researchers from a wide range of disciplines routinely use regression analyses to understand the mathematical relationships between variables, for purposes of description, hypothesis testing, and prediction.[1] Regression extends the capability provided by other statistical procedures by quantifying the level (amount) of change in the outcome or dependent variable that would be expected based upon a given level of change in a predictor or independent variable.

Applied to health care, particularly in managed care settings where administrative claims or ambulatory care databases are readily available, regression techniques may be useful in conducting pharmacoeconomic or outcomes analyses.[2,3] These methods provide the ability to assess cost functions and treatment effectiveness while simultaneously controlling for potential confounding factors such as comorbidities, prior health care utilization, and demographic characteristics. Researchers are afforded the ability to understand and describe the association of various predictor variables with phenomena such as the level of demand or consumption of medical services or changes in clinical outcomes for a given therapeutic intervention.

Regression is routinely applied to many aspects of research and reporting. As such, the objective of this paper is to establish a conceptual foundation for understanding regression techniques, to assist in their application and understanding within the context of health care decision making. Correlation is introduced initially because it serves as a major tenet of basic regression. Following this is a presentation of regression, including the diagnostic tools needed for sound interpretation. Although broad extensions and complexities exist in a discussion of the topic, this article centers on the transparent aspects of regression analysis and then presents general considerations that may be useful for researchers wishing to incorporate these methods into practice within managed care settings.

Correlation

Empirical investigations often seek to gain an understanding of the associations between variables. Specifically, the strength of a linear association between 2 variables is termed correlation. Although there are several measures of correlation, the most commonly reported is the *Pearson product-moment correlation coefficient* (typically abbreviated with a lowercase "*r*").[4] The Pearson correlation is an expression of the association between 2 continuous variables; other correlations may be reported according to differing characteristics of the variables (e.g., *Spearman's rho*, which is a correlation coefficient for ranked variables). Correlation coefficients measure 2 facets of a relationship: (1) the magnitude (with absolute coefficient values ranging from zero to one) and (2) the direction (either positive or negative). For the Pearson correlation, the larger the absolute value, the stronger the linear association: a correlation of –1.0 indicates a perfectly negative linear association, 0.0 indicates no linear association, and +1.0 indicates a perfectly positive linear association.[1] Illustrating the overall concept of correlation, Figure 1 is a graphical representation of hypothetical scatter plots of data for 2 variables.

Example. A positive correlation may exist between a patient's overall health care expenditures and the number of physician visits or hospital stays incurred (i.e., expenditures increase as a function of demand for medical services). Conversely, a negative correlation may exist between the number of errors occurring in practice and the level of quality assurance measures implemented in the workplace (i.e., errors would be expected to decrease as the number of quality assurance measures increased).

It should be emphasized that the Pearson correlation measures a *linear* association between 2 variables. Thus, in the presence of skewed distributions of data, curvilinear functions between variables, or instances with extreme outliers, the correlation coefficient may not be a reliable indicator of the presence of a relationship. For example, if the correlation between 2 variables is found to equal zero, this does not preclude the possibility that a relationship between the variables exists.[5] For this reason, although statistical software packages routinely report correlations and related tests of statistical significance, graphical analyses of scatter plots similar to those shown in Figure 1 should be reviewed when ascertaining the validity of any given value for a correlation.

Correlation is often viewed as a precursor to regression techniques because the 2 topics are conceptually related. However, clear differences between correlation and regression exist. Whereas correlation centers upon expressing the strength of a linear association between variables, regression involves the estimation of a mathematical equation that is based upon a theoretical specification (selecting an appropriate functional form, or model, that includes all relevant variables). Thus, correlation is not considered to be dependent upon the scales of measurement between variables, whereas altering the units of measure in regression may change the interpretation of the results dramatically.[6]

In regression analysis, statistical tools based on the principle of correlation are used to assess regression results. The *coefficient of determination* is particularly important because it represents the amount of variability of a *dependent* (or outcome) variable explained by an *independent* (synonymous with predictor or explanatory) variable.[6] Being the squared term of the correlation coefficient, the coefficient of determination is often referred to simply as "*r-squared*" ("r^2"). The mathematical properties of this statistic dictate values that range from 0 to 1 and are always less in magnitude than the correlation coefficient. In instances involving more than 2 variables, the *coefficient of multiple determination*, or "*R-squared*" ("R^2"), is often reported. The *adjusted R-squared* coefficient may also be reported, and corrects for the number of independent variables that are specified in a regression equation when more than one predictor variable is being analyzed. The R^2 coefficient provides useful information to researchers in assessing the overall "goodness of fit" of a given regression equation—a higher value indicates a better fit.[1] Researchers are also encouraged to analyze the overall F statistic from regression analyses in order to assess the statistical significance of a regression equation. Regardless of the magnitude or statistical significance of a correlation, researchers must continually be aware that only an association between variables is being measured. These statistical measures represent only one available means of assessing regression results, and their interpretation must

Author

GRANT H. SKREPNEK, PhD, is an assistant professor and investigator, University of Arizona College of Pharmacy, Center for Health Outcomes and PharmacoEconomic Research, Tucson.

AUTHOR CORRESPONDENCE AND ARTICLE REPRINTS: Grant H. Skrepnek, PhD, University of Arizona College of Pharmacy, 1703 East Mabel St., Tucson, AZ 85721. Tel: (520) 626-3930; Fax: (520) 626-7355; E-mail: skrepnek@pharmacy.arizona.edu

FIGURE 1 Scatterplots Illustrating Various Degrees of Correlation[1]

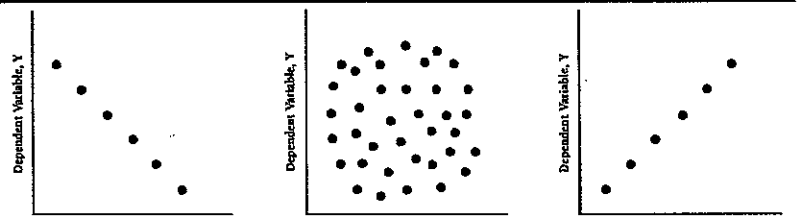
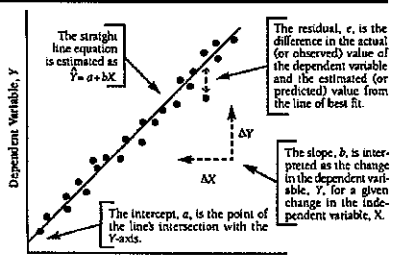


FIGURE 2 Scatterplot of Data Points and a Line of Best Fit[4,A9]



be made within the context of other statistics and diagnostics. Thus, findings in this regard cannot be used to establish causation—experimental methodologies are required to establish a causative relationship.[4,10-13]

The Linear Regression Model

The development of a regression model begins with a literature review to identify the appropriate variables. In describing this process, Motulsky (1995) defined a *model* as simply a "mathematical abstraction that is an analogy of events in the real world."[14] One recommended strategy in conducting regression analysis is to perform the following in a stepwise fashion: (1) literature review, (2) model specification (i.e., selection of appropriate dependent and independent variables and functional forms), (3) data collection, (4) model estimation and evaluation, and (5) result documentation.[1] Without a proper theoretical grounding, *specification errors* (e.g., using an improper functional form, excluding relevant variables, or including irrelevant variables or measurement biases) may result in biased statistical estimates.[1] Thus, Glantz and Slinker (1995) stated:

> If the [regression] model is not correctly specified, the entire process of estimating parameters and testing hypotheses about these parameters rests on a false premise and may lead to a misunderstanding of the system or process being studied. In other words, an incorrect model is worse than no model.[9]

Gujarati (1995) stated the importance of relying upon theoretical grounds for model development:

> Detecting the presence of an irrelevant variable(s) is not a difficult task. But it is very important to remember that in carrying out these tests of significance we have a specific model in mind. . . . [The] data mining technique or regression fishing, grubbing, or number crunching, is not to be recommended, for if [a given variable] legitimately belonged in the model it should have been introduced to begin with. Excluding [a given variable] in the initial regression would then lead to the omission-of-relevant-variable bias. . . . This point cannot be overemphasized: Theory must be the guide to any model building.[15]

Regression itself embodies a diverse range of analytical techniques, such as multivariate, proportional hazards, logistic, and nonlinear methods.[16] Despite the existence of more advanced methods, linear regression is the most often utilized. It describes a dependent variable as a straight-line function with respect to one or more independent variables. *Simple* linear regression refers to a specific case wherein a linear relationship is examined between only one dependent variable and one independent variable. An extension of simple regression is a *multivariate* (or multiple) regression that involves more than one explanatory variable. Adding predictor variables such as comorbidities, risk factors, and demographics thus establishes a multivariate model.

The premise behind regression is the estimation of a "line of best fit" through a set of data points and is modeled from a *population* equation.[9] For a simple linear regression, this population model would be as follows:

$$Y = \alpha + \beta X + \varepsilon$$

where Y is the dependent variable of interest; α is the intercept, or constant, of the equation; β is the slope coefficient, often referred to simply as "beta"; X is the independent variable of interest; and ε is the *residual*, disturbance, or error term of the equation.[6] In simple linear regression (with 1 independent variable), the line of best fit represents a 1-dimensional line drawn in a 2-dimensional space (an XY graph). In a multivariate regression analysis, the line of best fit actually refers to a 2-dimensional plane when there are 2 independent variables and an n-dimensional object when there are n independent variables, while the β estimates represent the effects of 1 independent variable on the dependent variable while all other independent variables are held constant. When a *sample* of data is used to estimate the population parameters α and β, the estimates are denoted a for α and b for β; the estimated value of the dependent variable is often expressed with a caret as $\hat{Y}$ and called "*Y-hat*."[9] Although numerous methods may be used, that of least squares (termed ordinary least squares [OLS]) is commonly used to estimate the coefficients of the regression model. Other than stating that the OLS criterion renders estimates for α and β that minimize the sum of the squared error term, the mathematical foundations of this approach are beyond the scope of this paper. As such, researchers are encouraged to consult other sources for a more comprehensive treatment of the fundamentals of coefficient estimation.[6,16,17]

A graphical depiction of a set of data points with a line of best fit, shown in Figure 2, builds upon the presentation of the simple linear model equation and illustrates several important components of regression. The intercept of the regression (i.e., the α coefficient) is defined as the point estimate wherein the line of best fit intersects the Y-axis, thus being the value of the dependent variable when the independent variable is equal to zero. The slope, or β coefficient, indicates the change in the dependent variable per change in the independent variable. The slope is fundamentally related to correlation: a positive β corresponds to a positive correlation and a negative β corresponds to a negative correlation. Beyond this, however, the β coefficient is interpreted differently. In a straightforward application of the simple linear model without transformations, β is interpreted as follows: given that the dependent variable (i.e., Y) is modeled by a linear function with the independent variable (i.e., X), Y increases by β-units for every 1 unit increase in X. The interpretation of the coefficient estimates changes when different types or transformations of variables are introduced into the model, a topic that is addressed in the section "Logarithmic Transformation of Data" later in this paper.

> **Example.** In a simple linear model in which the dependent and independent variables are continuous and linearly related (i.e., $Y = a + bX$), the value for the intercept, a, is merely the value of Y when X equals zero. The interpretation of the β coefficient, b, is that if X increases by 1 unit, Y increases by b units. To illustrate using a simple, hypothetical example, if a dependent variable is overall health care expenditure (measured in dollars) and an independent variable is pharmacy expenditures (measured in dollars), an estimated regression equation can be found to be $\hat{Y} = 225.00 + 1.07X$. Thus, if no pharmacy expenditures are incurred, a patient would be predicted to have a total health care expenditure of $225 (i.e., the value of the intercept, wherein the independent variable equals zero). For every dollar of pharmacy claims expended, the total health care expenditure would increase beyond $225 by $1.07. According to the estimated equation, if $300 in medications were utilized, the patient would be predicted to have a total consumption of $546 (i.e., $546 total health care expenditures = $225 + 1.07 [$300]).

The results of a regression analysis generally should include the coefficient of multiple determination (i.e., R^2 or $R^2_{adjusted}$, as appropriate) and should consider whether the overall regression equation was statistically significant via an F statistic. It is necessary to report the following statistics for specific coefficients: the parameter estimate, the standard error, the t-statistic value (i.e., the unstandardized parameter estimate divided by the standard error), and the P value. Given that the parameter coefficient obtained from the analysis is a point estimate, lower- and upper-bound confidence intervals may also be reported so that an estimated range of values likely to include a population parameter is presented. Additionally, the practical significance versus statistical significance of the findings must also be assessed.[18,19]

Assumptions of Regression Analysis

Researchers must remain aware of the assumptions of regression models since violations may bias and render false interpretations of the coefficient estimates. The key assumptions required for least squares analysis were reviewed by Gujarati (1995) and are consistent with those of Johnston (1984) and Greene (1997):[15-17]

1. The parameter coefficients of the model describe a linear relationship between the dependent and independent variables, described as follows:
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
2. The predictors of the model, X_i are understood to be *nonstochastic* (i.e., nonrandom);
3. The *conditional mean* of the error term (the expected value of the residual ε_i for any given value of the independent variable) is equal to zero, as follows:
$$E(\varepsilon_i | X_i) = 0$$
4. The residual term has a constant variance across observations, denoted as lacking *heteroscedasticity* (nonconstant variance). This is represented mathematically as the conditional variance, *var*, equal to the constant squared term of the standard deviation, σ, as follows:
$$\text{var}(\varepsilon_i | X_i) = \sigma^2$$
5. The residual terms of any 2 observations (i.e., X_i and X_j where $i \neq j$) are independent and unrelated, indicating a random distribution or lack of *autocorrelation* (i.e., correlation between observations). Mathematically, the lack of correlation in the residuals of 2 observations is represented as follows, wherein the conditional covariance of 2 observations with respect to the residual terms is equal to zero:
$$\text{cov}(\varepsilon_i, \varepsilon_j | X_i, X_j) = 0$$
6. The residual term is unrelated to all independent variables. That is, the conditional covariance with respect to an independent variable and the residual is equal to zero:
$$\text{cov}(\varepsilon_i, X_i) = 0$$
7. The residual term, ε_i, is normally distributed;
8. No perfect linear correlation exists between any of the independent variables (i.e., *romulticollinearity* is present).[15-17]

Gujarati (1995) also stated that, in order to derive parameter estimates, the number of observations must exceed the number of parameters estimated.[15] The number of observations per independent variable required for reasonable precision of the estimates is a function of the effect size that exists in the population (i.e., the slope of Y versus X_i). A general guideline offered by statisticians in the social sciences is that a minimum number of 15 observations per independent variable is required. The adequacy of this value is also contingent upon the magnitude of the multiple population correlation coefficient (defined as quantifying the degree of linear relationship between one variable and several others)[16,21] and other factors that must be considered in a robust power analysis (i.e., an analysis of the ability of the study to demonstrate an association if one exists).[24]

Given that parametric methods (methods that depend upon assumption[s] about the distribution of the data) are often used in conjunction with regression techniques, the following assumptions for the use of parametric test statistics should also be considered when conducting regression analyses: (1) random sampling of data, (2) independence of data in the sample, (3) normal distribution of sample means, and (4) *interval scale* data (an example of an interval scale is the Fahrenheit temperature scale).[16] The assumption of normally distributed sample means is often achieved through the *law of large numbers* (a theorem that states that a mean value from a randomly drawn sample converges in probability to a population mean when the sample size is large).[17] Within regression, numerous *categorical* variables can be added through *dummy codes*, which ultimately loosen the interval scale requirement and add increased applicability to the procedure.[22] Essentially, a dummy code creates a binary categorization by assigning either a 0 or a 1 to respective groups such as male and female or one therapeutic agent versus another; investigators should refer to other sources for a more complete discussion of this issue.[6,9,23] Dummy coding establishes the inclusion of an important qualitative component in multivariate regressions, allowing researchers to investigate processes beyond continuous variables. Furthermore, dummy coding allows the investigator to test for differences between 2 or more groups, thus employing elements of a t test or F test.

A linear relationship between the dependent and independent variable (item 1 in the above list of key assumptions) is often simply assumed—other functions (e.g., exponential terms) or transformations (e.g., logarithmic transformations) may be more appropriately employed to model the data and should often be explored. However, when modeling the variables with other mathematical functions, the other assumptions of OLS analysis continue to play an important role, especially those concerning the residuals (items 3–7 in the above list of key assumptions).

The assumptions and diagnostic tools of regression rest heavily upon the residual term of the model. In the linear regression model the residuals are assumed to be normally and independently distributed, with a mean of zero and a constant variance, which is typically denoted by econometricians as follows:

$$\varepsilon \sim NID\ (0, \sigma^2)$$

where ε is the residual, $\sim$ denotes "approximately," *NID* refers to "normally and independently distributed," 0 is the expected average value of the residual (i.e., zero), and σ^2 is the (constant) variance of the residual, defined as the squared value of the standard deviation; notably, the standard deviation of the residual is also another measure of a model's "goodness of fit."[16,17] According to Gauss-Markov theory, the best linear unbiased estimates (BLUE) or minimum variance linear unbiased estimates (MVLUE) for the coefficients of the linear regression model follow a similar form, with both the alpha and beta coefficients assumed to be normally distributed, N, with a mean estimate of α or β, and a constant variance:

$$a \sim N\left(\alpha, \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right]\right)$$

and

$$b \sim N\left(\beta, \frac{\sigma_\varepsilon^2}{\sum x^2}\right)$$

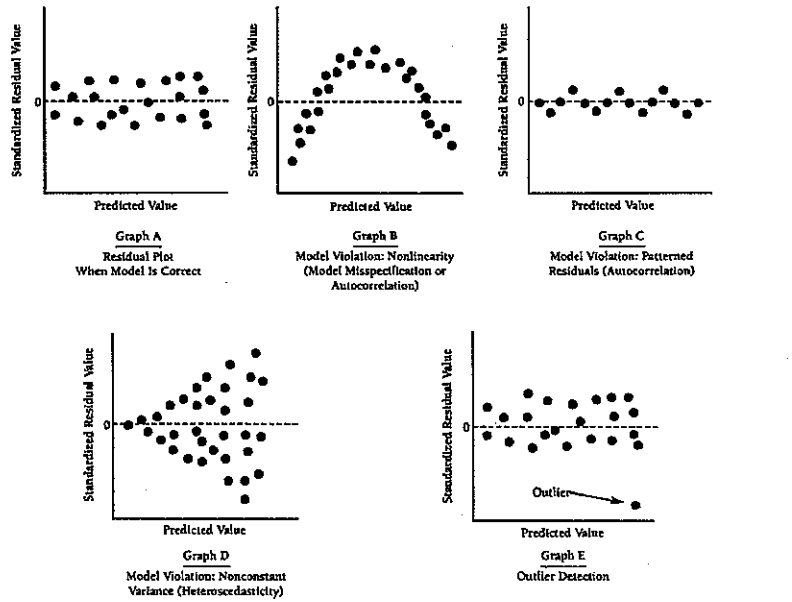
where a is the parameter estimate for the Y-intercept; α is the population parameter for the Y-intercept; b is the parameter estimate for the slope; β is the population parameter for the slope; σ_ε^2 is the variance of the residual term; x is the deviation of the independent variable from the sample mean, or $(X - \bar{X})$; $\bar{X}$ is the mean value for the independent variable X; and n is the number of observations.[16,17] A major goal, then, of assessing residuals is to ensure that BLUE/MVLUE is achieved concerning the parameter estimates a and b.

The assumption of *multicollinearity* (item number 8 in the above list of key assumptions) specifically involves instances in which the independent variables are highly or perfectly correlated with each other and the individual effects of each cannot be estimated with precision.[13,17,24] Thus, intercorrelated independent variables may decrease both the precision and accuracy of coefficients that are estimated in the regression model. Multicollinearity may arise from the following:

- the method of data collection (e.g., employing a sampling technique with an inappropriately limited number of values),
- regression model constraints or constraints in the sample population,
- model misspecification, and
- over-determined modeling (wherein a regression model has more independent variables than the number of observations).[25]

However, multicollinearity substantially affects the coefficient estimates only when it is present at high levels.[13] The only substantial effect of multicollinearity is that the standard errors of the regression coefficients may be unusually large and an ideal and efficient solution may not exist.[26] Several methods are available for detection of multicollinearity, which may include observing (1) a large R^2 value with relatively few statistically significant t ratios, (2) high pairwise correlations between independent variables, or (3) either a high value for the *condition index* or the *variance inflation factor* (VIF) (e.g., values of the condition index ≥ 15 or VIF ≥ 10).[1,15-17] Researchers should remain aware that ignoring multicolinearity may yield large variances for parameter estimates, resulting in statistically insignificant parameter coefficients and wide confidence intervals. Despite being able to recognize the presence of this correlation, controlling for it is challenging and may include transforming or combining variables, imposing a priori restrictions, increasing the number of observations in the study, or adding or dropping a variable.[13]

FIGURE 3 Scatterplot of Residuals Versus Predicted Values[1,4,7]



Regression Diagnostics via Residual Analyses

Residual analyses afford researchers the ability to test explicitly for the assumptions of constant variance (homoscedasticity, or same scattering), independence, linearity, and normality.[17] These diagnostics are warranted in any investigation that employs regression procedures. Residuals are defined as the error terms from a regression equation and represent the difference between (a) the actual value of the dependent variable and (b) the value of the dependent variable that is estimated from the line of best fit (i.e., $e = Y_{observed} - \hat{Y}_{predicted}$). If the assumptions of linear regression are violated or if the regression equation inappropriately models a set of data, anomalies will be observed in the error term. Outliers within the dataset can also be identified through a residual assessment. The relevance of diagnostic analyses extends beyond those that assess the goodness of fit (e.g., analysis of overall F tests and R^2 values in addition to graphical and

formal statistical analyses of scatter plots and residuals). Based upon the findings of the residual analysis, researchers may be directed toward examining alternate methods of properly specifying the regression equation, which may require the use of advanced mathematical and transformational techniques (e.g., quadratic terms, exponentials, or logarithmic transformations).[4,27]

The values observed in the error term should be stochastic or random in nature. Obtaining a graph of the residual plots is a preliminary method for verifying that the assumptions of the regression model have been met and that the parameters are the best linear unbiased estimates. Investigators should initially plot the standardized residuals versus the predicted values of the dependent variable; nonstandardized residual plots may also be utilized and follow a similar interpretation.[7]

Figure 3 presents various scatter plots of the standardized residual (error term) against the predicted value of the dependent variable. An ideal pattern would be one that is stochastic and randomly distributed around an average value of zero (Figure 3A). When a straight-line regression equation is used to model a set of data that is nonlinear, or when the residuals are systematically related to other cases within the set (i.e., autocorrelated), a curved pattern in the residuals may be observed, as illustrated in Figure 3B. Another example of autocorrelated residuals appears in Figure 3C, as a distinct relationship appears in the error term (in this instance, a sine wave).

The presence of a funnel-shaped pattern (Figure 3D) may indicate a violation of the constant variance assumption; the illustration depicts an increased dispersion or variance of the residuals as the predicted value increases. Importantly, logarithmic transformations of variables often simultaneously stabilize the variance and normalize the data. Tests of normality may be conducted by viewing a histogram of the error terms or normal quantile-comparison plots in addition to formal statistical methods (e.g., correlation test for normality).[4,6,16,17] The presence of an outlier in the set is depicted in Figure 3E. Researchers should analyze individual cases to determine if they may be attributed to error or if they are indeed valid observations.

A number of advanced techniques exist that allow the researcher to detect and correct for violations of the regression model although a formal description of these techniques is beyond the scope of this paper. Residual analyses provide a means of testing many assumptions of a regression model, but the challenge often resides in how to control for violations once they have been identified. The appropriate techniques are often highly specific for a given violation and researchers must first begin by understanding any underlying theory behind the nature of the variables of interest. Relatively simple transformations of one or more of the variables may rectify common violations.

Serial correlation or *autocorrelation* is a particular violation of the independence assumption (e.g., if the residuals are related between observations).[17,28] The most common formal statistical test for autocorrelated residual terms is the Durbin-Watson d test, whereas correcting for autocorrelation often includes employing generalized least squares (GLS) methods.[3,6,16,17] Other common methods used to quantitatively test for the presence of serial correlation include the runs test (Geary test) and the Breusch-Godfrey test.[29-31] The method of rectifying autocorrelation is dependent upon the specific interrelationship and structure of the serial correlation itself. Two steps are typically required: first, obtain or estimate a *coefficient of autocovariance*, and, second, use these values to transform the original regression equation. In calculating the coefficient of autocovariance, time-related processes may follow autoregressive functions that involve distributive lag functions (wherein the dependent variable is systematically related to past values of the dependent variable), moving average schemes (that involve the dependent term's being functionally related to past values of the error term), combinations (autoregressive moving average), and numerous others.[32] Transformations to regression equations that incorporate estimations of the coefficient of autocovariance typically follow applications of GLS, such as feasible generalized least squares.

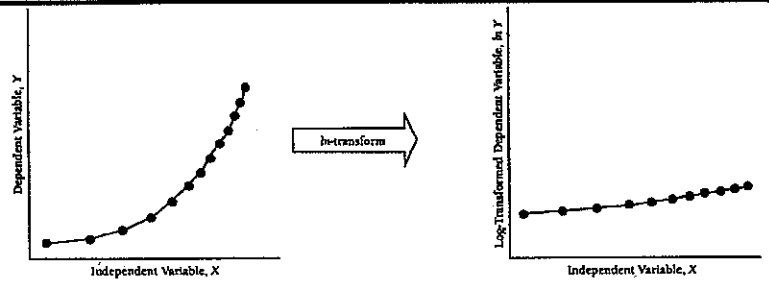
When the constant variance assumption is violated (e.g., if the variance of the residual is not equal between cases), *heteroscedasticity* (i.e., unequal variance, or different scattering) is present. Detection of heteroscedasticity may be undertaken through quantitative procedures such as the following: Park test, Glejser test, Breusch-Pagan/Godfrey test, Goldfeld-Quandt test, or White's general test.[33-37] The large number of statistical tests from which to choose is due to the diverse nature of heteroscedasticity. In a method that is somewhat similar to controlling for autocorrelation, GLS is often used to correct for unequal variance via weighted least squares estimators. Heteroscedasticity-robust statistical tests have also been developed and may be considered by analysts (e.g., Eicker-Huber-White standard errors).[38-40] Additionally, procedures have been reported that control for concomitant violations of both serial correlation and unequal variance (e.g., autoregressive conditional heteroscedasticity model).[41]

Logarithmic Transformation of Data

All phenomena do not necessarily conform to straight-line functions. Health care utilization or cost data are rarely normally distributed and are often characterized by heteroscedasticity (different scattering) and distributions that are positively skewed, thus warranting transformation to reach the assumption of normality.[42-45] Departures from linearity may be investigated via regression, providing that the proper mathematical operands are expressed within the model. Various transformations of variables may also be employed to create linear functions in instances that involve curvilinear or nonlinear processes, skewed distributions, or heteroscedastic residuals; transformations may additionally resolve violations of the OLS estimation of a regression model and can essentially change a nonlinear form to a linear one. To illustrate, if an observed relationship exists between the variables as depicted in Figure 4A, a logarithmic transformation (often with a Naperian or natural logarithm, *ln*) may establish a linear function between variables as indicated in Figure 4B. A representation of the residuals for each of the models appears in Figure 5. A curvilinear relationship in the residual plot of the untransformed data (Figure 5A) is due to model misspecification, (i.e., the incorrect assignment of a linear regression model to nonlinear data).

When the dependent variable of the model is transformed via a natural logarithm or Briggsian/common logarithm (log, base 10), the resulting model is commonly denoted as a

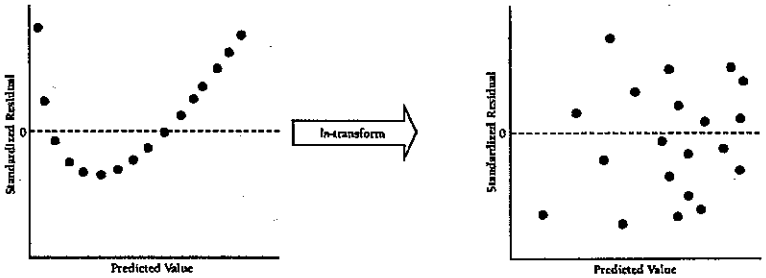
FIGURE 4 Logarithmic Transformation of Data to Yield a Linear Function



Graph A
Logarithmic Function

Graph B
Log Transformation
to Yield a Linear Function

FIGURE 5 Residual Plots Following a Logarithmic Transformation to Yield a Linear Function



Graph A
Residual Plot of the
Logarithmic Function

Graph B
Residual Plot of the
Logarithmic-Transformed Function

TABLE 1 Summary of Functional Forms Involving Logarithms for a Simple Regression Equation[1,45]

Model	Regression Equation	Interpretation of the β Coefficient Estimate, *b*	
lin-lin	$Y = \alpha + \beta X$	$\Delta Y = b\Delta X$	A 1-unit change in X is associated with a b-unit change in Y.
lin-log	$Y = \alpha + \beta \ln(X)$	$\Delta Y = \left(\frac{b}{100}\right)\%\Delta X$	Dividing the β coefficient estimate by 100 yields the following interpretation: a 1% change in X is associated with a $(b/100)$-unit change in Y.
log-lin	$\ln(Y) = \alpha + \beta X$	$\Delta Y = (100 \cdot b)\Delta X$	Multiplying the β coefficient estimate by 100 yields the following interpretation: a 1-unit change in X is associated with a $b \cdot 100$ percent change in Y (i.e., semielasticity).
log-log	$\ln(Y) = \alpha + \beta \ln(X)$	$\%\Delta Y = b\%\Delta X$	A 1% change in X is associated with a b percent change in Y (i.e., constant elasticity).

"log-linear" or "log-lin" model. In economics, the coefficients are loosely defined as a semielasticity.[35,36] When both the dependent and independent variables have been log-transformed, the resulting coefficients are interpreted as a constant elasticity or, simply, an elasticity.

If transformations are required to meet the assumptions of a linear model, the researcher must interpret the coefficient estimates properly. Whenever transformations are implemented, the data units change, so that coefficient estimates from logarithmic models cannot be interpreted in an identical fashion to coefficients derived from simple linear models. Table 1 presents a summary of the interpretations of various coefficients given log-transformed variables. Importantly, when logarithmic transformations are used, researchers cannot estimate actual values of a variable by simply applying an antilog. This method introduces a substantial retransformation bias, which may often be overcome by applying a smearing estimation, as proposed by Duan (1983).[37-39] Duan's smearing estimate has received increased attention in pharmacoeconomics and outcomes research, as health care decision makers often desire regression parameter coefficients to be presented in units rather than as percentage comparators. However, an important caveat in the use of Duan's smearing estimator is that it has been found to introduce a substantial bias when heteroscedasticity is present. Furthermore, robust estimators (e.g., Eicker-Huber-White standard errors) often do not correct the bias. In instances where unequal variance is noted, alternatives to Duan's approach may include natural log transformations with an OLS estimation, GLS and related extensions (e.g., gamma or Weibull regression with log link), or the Cox proportional hazards model.[30-32] Unfortunately, empirical analyses also indicate that no single best model is robust under all specifications.[30-32]

Example. In reference to Table 1, if a β coefficient estimate obtained from a set of data was +0.250 and statistically significant, a lin-lin interpretation would be that a 1-unit increase in the independent variable (X) is associated with a 0.250-unit increase in the dependent variable (Y). For a lin-log equation wherein only the independent variable is log-transformed, the interpretation is that a 1% increase in the independent variable (X) is associated with a 0.0025-unit change in the dependent variable (Y) (i.e., 0.250 ÷ 100). The log-lin interpretation of a log-transformed dependent variable is such that a 1-unit change in the independent variable (X) is associated with a 25% increase in the dependent variable (Y) (i.e., 0.250×100), and is considered a semielasticity. Finally, for a log-log model wherein both dependent and independent variables are log-transformed, the interpretation is that a 1% increase in the independent variable (X) is associated with a 0.25% increase in the dependent variable (Y), which is interpreted as a constant elasticity.

A concern when employing transformations to variables occurs if there are any values in the dataset that are not subject to the mathematical function.[37,38] To illustrate, the logarithm of zero or any negative value is mathematically undefined. Thus, researchers attempting to apply a logarithmic transformation to a variable that has negative values or is zero would essentially exclude these cases, which may appear as a clustered region in the residual plots. This issue is of particular concern for data involving health care utilization or expenditure data, as many patients or medical beneficiaries go through sustained periods without utilization or expenditures. Addressing this issue often involves the use of advanced statistical procedures.[4,5,8,19] Methods used by researchers have included employing a log plus constant model (e.g., $1 added to all costs to allow a logarithmic transformation), a 2-part model (wherein the first model uses a logistic regression to predict the probability of costs or utilization and a second model estimates the actual level of use for subjects that incur costs or utilization), survival analysis techniques, or generalized linear models with related variants.[4,40] Consideration of gamma, negative-binomial, or Poisson distributions of cost or utilizaiton data may be incorporated to yield best linear unbiased estimates. Although all of these methods have appeared in the literature, no single best approach may be recommended nor do the methods represent an exhaustive list. Applying any given method without assessing the fundamental assumptions of the statistical test may produce misleading results and conclusions.

Regarding broader elements of forecasting, censored data (wherein data truncation occurs due to death or lack of follow-up) is a cause of concern.[35,39] Empiric research has found that the censored nature of costs may result in biased estimates if not appropriately controlled.[36] Remedial measures that have been suggested involve applications of survival analysis although research has demonstrated that these techniques may not necessarily be appropriate and that nonparametric methods may be better suited.[37]

Applications in Managed Care

In a purely hypothetical scenario with relevance to managed care or formulary decision making, an analyst may want to use a managed care dataset to ascertain whether health care cost differences exist between therapeutic options for patients with heart failure. Thus, the dependent variable of interest would be total health care costs. Furthermore, the analyst may also be interested in determining the related predictors of hospitalization, defining whether hospitalization occurred as a second dependent variable that is dichotomous (i.e., hospitalization=1, no hospitalization=0). Given that retrospective analysis of administrative claims data lacks randomization and fully experimental methodologies, relevant confounding variables must be included within the regression model to statistically control for differences between study groups.[11] Thus, in this example, simply comparing unadjusted total health care costs (or whether a patient was or was not hospitalized) between the therapeutic options without controlling for confounders is inappropriate; regression analyses are required.

Contingent upon a thorough literature review of predictors of cost and/or hospitalization in heart failure patients, the specification of a regression model in this hypothetical case should provide a theoretical basis for the variables that are ultimately included in the analysis. In pharmacoeconomic or outcomes research, these independent variables may include risk adjustment measures to control for case mix severity (e.g., Chronic Disease Score [CDS] based upon prescription drug use, Charlson Index based upon *International Classification of Diseases, 9th Revision, Clinical Modification* codes), patient characteristics (e.g., age, sex), comorbid conditions, pretreatment costs, and treatment groups.[39,40] Additionally, medication adherence (e.g., Medication Possession Ratio) and insurance or provider characteristics (e.g., type of health care organization, Medicare, Medicaid, prescriber specialty) may be deemed important to consider within a research question.[41] Incorporating some of these considerations, a theoretical model may be proposed in comparing treatment options (*Treatment One*, *Treatment Two*, and *Treatment Three*) as

Expenditures = f(Risk, Patient, Comorbidities, Pretreatment Cost, Treatment Group),

which appears in regression form as

$$Expenditures = \alpha + \beta_1 CDS + \beta_2 Age + \beta_3 Sex + \beta_4 Comorbidity + \beta_5 Treatment_{One} + \beta_6 Treatment_{Two} + \varepsilon$$

where *Sex*, *Comorbidity*, *Treatment One*, and *Treatment Two* are defined as dichotomous dummy variables (variables with only 2 values: 0 or 1). *Treatment Three* is considered the baseline used for comparison. Given that the total number of dummy variables required is equal to 1 less than the total number of categories to be compared, a single dummy variable is used for *Sex*, while 2 dummy variables are used for the treatment groups. Comorbidities may also be coded as dummy variables, with the presence of the condition coded as 1 and absence coded as 0. Examples relevant to heart failure may include a past history of myocardial infarction or stroke and can extend to include the presence of diabetes, atrial fibrillation, renal disease, or hypertension.

Concerning the second dependent variable in the example (i.e., hospitalization), when binary categorical variables are used as a dependent variable, a logistic regression may be considered; this may also be extended to ascertain predictors of treatment success (i.e., treatment success=1, treatment failure=0).[42]

Example 1. Armstrong and Malone (2002) conducted an assessment of asthma-related costs associated with fluticasone versus leukotriene modifier use, in which the dependent variable was log-transformed post-asthma cost and independent variables included age, sex, log-transformed pre-asthma cost, CDS, presence of chronic obstructive pulmonary disease, treatment group (i.e., fluticasone or leukotriene modifier), and the use of certain medications prior to the study period (i.e., number of short-acting β-agonist canisters used and dummy codes representing the use of long-acting β-agonists, theophylline or mast cell stabilizers, or oral corticosteroids).[43]

Example 2. McLaughlin, Eaddy, and Grudzinski (2004) analyzed depression-related charges associated with treatment with sertraline and citalopram.[44] The dependent variable, treatment charges, was transformed via natural logarithm specifically due to the detection of heteroscedasticity. The independent variables included age, sex, geographic region, natural log of treatment charges prior to the study period, the presence of comorbidities (either mental or nonmental health), the type of managed care institution, the physician specialty, the use of emergency department or hospital services prior to the study period, and the year of initial diagnosis.

Conclusion

Regression is a powerful and commonly used statistical technique that gives researchers the ability to quantify mathematical relationships for purposes of description, hypothesis testing, or prediction. Before a proper analysis can begin, an understanding is necessary of correlation, of the assumptions of a classical linear regression model, and of the importance of residual assessments. Despite its complexities, the flexibility

of regression makes it especially applicable in certain settings, enabling decision makers to analyze complex phenomena and answer questions that other statistical methods inadequately address. Contingent upon the specific research question, a number of extensions to basic regression techniques may be explored so that this method can be used appropriately in pharmacoeconomic or outcomes research and assist in formulary decision making. The increased availability of administrative databases containing medical and pharmacy claims data may provide those in managed care settings with a greater ability to evaluate treatments and practice patterns. Given that administrative data are observational rather than experimental, it is critical that analysts and decision makers be versed in appropriate statistical methods to design investigations or evaluate empiric findings.

DISCLOSURES

No outside funding supported this study. The author discloses no potential bias or conflict of interest relating to this article.

REFERENCES

1. Studenmund AH. *Using Econometrics: A Practical Guide.* 4th ed. Boston: Addison Wesley; 2001.

2. Armstrong EP, Manuchehri F. Ambulatory care databases for managed care organizations. *Am J Health Syst Pharm.* 1997;54:1973-83.

3. Johnson N. The six-step process for conducting outcomes analyses using administrative databases. *Formulary.* 2002;37:362-64.

4. Gandi SK, Salmon JW, Kong SX, Zhao SZ. Administrative databases and outcomes assessment: an overview of issues and potential utilities. *J Manag Care Pharm.* 1999;5(3):215-22.

5. Else BA, Armstrong EP, Cox ER. Data sources for pharmacoeconomic and health services research. *Am J Health Syst Pharm.* 1997;54:2601-08.

6. Neter J, Kutner MH, Nachtscheim CJ, Wasserman W. *Applied Linear Statistical Models.* 4th ed. Chicago: Irwin, McGraw-Hill; 1996.

7. Shott S. *Statistics for Health Professionals.* Philadelphia: W.B. Saunders; 1990.

8. Dawson B, Trapp RG. *Basic and Clinical Biostatistics.* 3rd ed. New York: Lange Medical Books, McGraw-Hill; 2001.

9. Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance.* New York: McGraw-Hill; 1990.

10. Cook TD, Campbell DT. *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin; 1979.

11. Motheral BR, Fairman KA. The use of claims databases for outcomes research: rationale, challenges, and strategies. *Clin Ther.* 1997;19:346-66.

12. Motheral BR. Research methodology: hypotheses, measurement, reliability, and validity. *J Manag Care Pharm.* 1998;4(4):382-88.

13. Richards KM, Shepherd MD. Subject review: claims data and drawing appropriate conclusions. *J Manag Care Pharm.* 2002;8(2):152.

14. Motulsky H. *Intuitive Biostatistics.* New York: Oxford University Press; 1995.

15. Gujarati DN. *Basic Econometrics.* 3rd ed. New York: McGraw-Hill; 1995.

16. Johnston J. *Econometric Methods.* 3rd ed. New York: McGraw-Hill; 1984.

17. Greene WH. *Econometric Analysis.* 4th ed. Upper Saddle River, NJ: Prentice-Hall; 2000.

18. O'Brien BJ, Drummond MF. Statistical versus quantitative significance in the socioeconomic evaluations of medicines. *Pharmacoeconomics.* 1994;5:389-98.

19. Drummond M, O'Brien B. Clinical importance, statistical significance, and the assessment of economic and quality-of-life outcomes. *Health Econ.* 1993;2: 205-12.

20. Stevens J. *Applied Multivariate Statistics for the Social Sciences.* 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 1996.

21. Park C, Dudycha A. A cross validation approach to sample size determination for regression models. *J Am Stat Assoc.* 1974;69:214-18.

22. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: Lawrence-Erlbaum Associates; 1988.

23. Hardy MA. *Regression with Dummy Variables.* Newbury Park, CA: Sage Publications; 1993.

24. Frisch R. *Statistical Conference Analysis by Means of Complete Regressions Systems.* Publication Number 5. Oslo: University of Oslo, Institute of Economics; 1934.

25. Montgomery D, Peck E. *Introduction to Linear Regression Analysis.* New York: John Wiley & Sons; 1982.

26. Achen CH. *Interpreting and Using Regression.* Beverly Hills, CA: Sage Publications; 1982.

27. Carrol RJ, Rippert D. *Transformation and Weighting in Regression.* New York: Chapman and Hall; 1988.

28. Durbin J, Watson GS. Testing for serial correlation in least-squares regression. *Biometrika.* 1951;38:159-71.

29. Geary RC. Relative efficiency of a count of sign changes for assessing residual autoregression in least squares regression. *Biometrika.* 1970;57:123-27.

30. Breusch TS. Testing for autocorrelation in dynamic linear models. *Australian Econ Pap.* 1978;17:334-55.

31. Godfrey LG. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica.* 1978;46:1293-1302.

32. Hamilton JD. *Time Series Analysis.* Princeton: Princeton University Press; 1994.

33. Park RE. Estimation with heteroscedastic error terms. *Econometrica.* 1966; 34:888.

34. Glejser H. A new test for heteroscedasticity. *J Am Stat Assoc.* 1969;64:316-23.

35. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. *Econometrica.* 1979;47:1287-94.

36. Goldfeld SM, Quandt RE. *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland Publishing Company; 1972.

37. White H. A heteroscedasticity consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica.* 1980;48:817-18.

38. Eicker F. Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press; 1967:59-82.

39. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press; 1967:221-33.

40. White H. A heteroscedasticity-constant covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica.* 1980;48:827-38.

41. Bera A, Higgins M. ARCH models: properties, estimation, and testing. *J Econ Surveys.* 1993;7:305-66.

42. Fox J. *Regression Diagnostics.* Newbury Park, CA: Sage Publications; 1991.

43. Diehr P, Yanez D, Ash A, Hornbrook M, Yin DY. Methods for analyzing health care utilization and costs. *Annu Rev Public Health.* 1999;20:125-44.

44. Zhou X, Melfi CA, Hui SL. Methods for comparison of cost data. *Ann Intern Med.* 1997;127:752-56.

45. Wooldridge JM. *Introductory Econometrics: A Modern Approach.* 2nd ed. Stamford, CT: South-Western, Thomson; 2003.

46. Barnett RA, Ziegler MR. *Applied Calculus for Business, Economics, Life Sciences, and Social Sciences.* 5th ed. New York: MacMillan; 1994.

47. Duan N. Smearing estimate: a nonparametric retransformational method. *J Am Stat Assoc.* 1983;78:605-10.

48. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ.* 1998;17:283-95.

49. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ.* 1998;17:247-81.

50. Manning WG, Mullahy J. Estimating log models: to transform or not to transform. *J Health Econ.* 2001;20:461-94.

51. Basu A, Manning WG, Mullahy J. Comparing alternative models: log versus Cox proportional hazard? *Health Econ.* 2004;13:749-65.

52. Berry WD, Feldman S. *Multiple Regression in Practice.* Newbury Park, CA: Sage Publications; 1985.

53. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika.* 2000;87:329-43.

54. Lin DY. Linear regression analysis of censored medical costs. *Biostatistics.* 2000;56:775-78.

55. Zhao H, Tian L. On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics.* 2001;57:1002-08.

56. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. The analysis of censored treatment cost data in economic evaluation. *Med Care.* 1995;33:851-63.

57. Etzioni RD, Feuer EJ, Sullivan SD, Lin D, Hu C, Ramsey S. On the use of survival analysis techniques to estimate medical care costs. *J Health Econ.* 1999;18:365-80.

58. Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *J Clin Epidemiol.* 1992;45:197-203.

59. Clark DO, Von Korff M, Saunders K, Baluch WM, Simon GE. A chronic disease score with empirically derived weights. *Med Care.* 1995;33:783-95.

60. Charlson ME, Pompei P, Ales KL, McKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40:373-83.

61. Solar DA, Chin A, Skaer TL, et al. Effect of health education in promoting prescription refill compliance among patients with hypertension. *Clin Ther.* 1991;13:489-95.

62. Abarca J, Armstrong EP. Multiple logistic regression: a statistical method to evaluate predictors of hospitalizations. *Formulary.* 2000;35:832-41.

63. Armstrong EP, Malone DC. Fluticasone is associated with lower asthma-related costs than leukotriene modifiers in a real-world analysis. *Pharmacotherapy.* 2002;22:1117-23.

64. McLaughlin TP, Eaddy MT, Grudzinski AN. A claims analysis comparing citalopram with sertraline as initial pharmacotherapy for a new episode of depression: impact on depression-related treatment costs. *Clin Ther.* 2004;26:115-24.