Published in final edited form as:

Appl Health Econ Health Policy. 2012 September 1; 10(5): 319–329. doi:
10.2165/11632430-000000000-00000.

Estimating Incremental Costs with Skew: A Cautionary Note

Linnea A. Polgreen and John M. Brooks

College of Pharmacy, University of Iowa, Iowa City, IA, USA 52242

Abstract

Background—Cost data in healthcare are often skewed across patients. Thus researchers have used either a log transformation of the dependent variable or generalized linear models (GLMs) with log links. However, frequently these non-linear approaches produce non-linear incremental effects: the incremental effects differ at different levels of the covariates, and this can cause dramatic effects on predicted cost.

Objectives—To demonstrate that when modeling skewed data, log link functions or log transformations are not necessary and have unintended effects.

Methods—We simulated cost data using a linear model with a “treatment”, a covariate, and a specified number of observations with excessive cost (skewed data). We also used actual data from a pain-relief intervention among hip-replacement patients. We then estimated cost models using various functional approaches suggested to handle skew and calculated the incremental cost of treatment at various levels of the covariate(s).

Results—All of these methods provide unbiased estimates of the incremental effect of treatment on costs at the mean level of the covariate. However, in some log-based models the implied incremental treatment cost doubled between extreme low and high values of the covariate in a manner inconsistent with the underlying linear model.

Conclusions—Although specification checks are always needed, the potential of misleading incremental estimates resulting from log-based specifications is often ignored. In this era of cost containment and comparisons of treatment effectiveness, it is vital that researchers and policy-makers understand the limitation of the inferences that can be made using log-based models for patients whose characteristics differ from the sample mean.

1. Introduction

Healthcare costs are often skewed: some patients have much-larger-than-average cost (e.g., patients with rare complications). As a result, researchers interested in estimating the incremental cost associated with a specific treatment, condition, or patient or provider characteristic must consider the implications of a skewed dependent variable in estimation.

Corresponding Author: Linnea Polgreen, Ph.D., University of Iowa, College of Pharmacy, 115 S. Grand Ave., Iowa City, IA 52246, Phone: 319-354-4820, Fax: 319- 353-5646, linnea-polgreen@uiowa.edu.

There are no potential conflicts of interest for either author, and both authors meet the criteria for authorship.

John Brooks simulated the data and also provided the hip-replacement data for these analyses. Linnea Polgreen performed the analyses. Both authors wrote and edited the manuscript. Linnea Polgreen is a guarantor for the overall content of the manuscript.

A traditional approach to addressing problems of skewed costs is to perform ordinary least squares estimation (OLS) on the natural logarithm of cost to decrease the importance of observations with extreme costs in estimation.^[1,2,3] This approach was made popular in papers describing results of the Rand Health Insurance Experiments, whose authors found that the log transformation solved their skewed data problem.^[4,5,6] A more recent approach is to use generalized linear models (GLM) to provide more general dependent variable transformations and error distributions.^[7,8,9,10] Specifically, many researchers have followed the findings in Manning and Mullahy^[11] that examined the relative merits of different approaches to dealing with models based on log transformation or GLM alternatives with a log link, and used a log link or log transformation.^[12,13,14,15,16] While these transformation methods mitigate the problems associated with extreme residuals, they introduce other problems, as log-transformed OLS models and GLM models with a log link *force* estimates of the incremental effect of each independent variable on cost to vary with the levels of the other independent variables in the model. As a result, the estimated interrelationships among the independent variables may not coincide with the underlying data generating process. Thus, estimation approaches to deal with skewed residuals may produce misleading inferences of the incremental costs associated with independent variables. Policymakers may conclude that the incremental cost associated with a specific treatment, condition or patient or provider characteristic varies with the other covariates in the model when it may not. This needs to be determined empirically.

The goal of this paper is to assess the properties of various estimators of incremental cost. We contrast the ability of four commonly-used estimators: GLM with a Gaussian family and an identity link (the Ordinary Least Squares model), GLM with gamma and Gaussian families with log links, and the Extended Estimating Equations (EEE) estimator^[17,18] to find the incremental cost changes associated with a randomized behavioral intervention in the pain treatment for hospitalized patients with hip fractures.^[19,20]

We also simulated data. We specifically estimate the cost of a discrete independent variable defined as a “treatment” when the treatment is linearly related to cost, and a portion of the population has excessive cost. Previous literature comparing the properties of log-transformed OLS models to various GLM specifications used underlying non-linear simulation models as the source of their comparisons.^[11,21,22,23,24] However, to the best of our knowledge, no one has examined the properties of various estimators when the underlying cost model is linear in the dependent variables, and the skew is caused by a percentage of patients having excessive costs not attributable to measured variables. In this study we performed a series of simulations in which we varied the size of the excessive cost and the percentage of the population with excessive cost. In addition, a portion of the simulated patients in these models receive a “treatment” that is linearly related to cost. Our goal is to estimate the incremental cost of this treatment from the simulated data. Simulations were performed with and without an additional measured covariate to assess the effect of covariates on the incremental treatment cost estimates from various estimators. Using the simulated data we estimated incremental treatment cost using three frequently used models: GLM with gamma and Gaussian families and log links and GLM with a Gaussian family and an identity link (OLS).

2. Strategies for Choosing Specifications for GLM-based Cost Models

Generalized linear models (GLM) provide an alternative to cost modeling that avoids the complications involved with retransformation.^[25,26,27] GLM models take the form:

$$E(y) = g^{-1}(x\beta) \quad \text{where } y \sim F \quad (1)$$

The function $g()$ is the link function, which relates the mean of the distribution to the linear predictor, and F is the distributional family; the family relates the mean to the variance function. The ordinary least squares (OLS) model can be represented by a GLM model with a Gaussian family and an identity link. An identity link means that $g()$ is an identity function such that $E(y) = x\beta$. Unlike OLS, a Gaussian model does not mean that the errors are distributed normally; it means that the variance function is unrelated to the mean. Different relationships between the mean and variance can be represented by other families. A model similar to the commonly-used log-transformed OLS model can be estimated with a GLM model with a Gaussian family and a log link, but estimating $E[\ln(y)]$ is not equivalent to estimating $\ln[E(y)]$. The GLM model is estimated on the raw (untransformed) scale: rather than logging the dependent variable, the are exponentiated, producing predicted values on the raw scale, and eliminating the need for retransformation. However, GLM with a log link is not used to normalize the error terms with skewed data, rather it is used to better align the variance function to the mean function in the data. Because of this, application of the wrong distributional family can create efficiency losses in estimation.

In addition to the Gaussian family, the gamma family is another common set of GLM models used by researchers.^[7,8,12,13,14,16] Gamma models require that the variance function goes up with the square of the mean function. Application of the gamma family is most often applied in healthcare for patient samples in which positive healthcare costs are ensured (e.g., patients all with a specific diagnosis that requires a minimum level of treatment).^[27]

For both the Gaussian and the gamma family of GLM models, we consider link functions that are most common in healthcare cost models: log and identity. However, for GLM models with a log link, the expected effect of x_i on y is a function of the level of the remaining independent variables in x , and this is true for all non-identity links. For example, a GLM with an identity link is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad (2)$$

and the incremental effect of x_1 , for example, is $\delta y / \delta x_1$, which equals β_1 . However, with a non-identity link, the GLM is of the form

$$y = g^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon). \quad (3)$$

In this case the incremental effect involves the chain rule such that

$$\delta y / \delta x_1 = \beta_1 \left[(\delta y / \delta g) g^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon) \right]. \quad (4)$$

The incremental effect depends on all the other covariates and coefficients in the model.

Choosing the correct combination of link function and distributional family is often difficult when using GLM. Various methods for determining the correct model have been proposed, and we employ two of the most common: the modified Park test^[11] and a grid-search method used in Blough et al.^[23] For our hip replacement data, we also employ the extended estimating equations (EEE) estimator, which estimates the most efficient link and variance functions simultaneously.

The Park test estimates the relationship between the mean and the variance of the data. Given the predicted values and residuals from a provisional model, the following model is estimated:

$$\ln(u_i^2) = \alpha_0 + \alpha_1 \ln(y_i) + e_i \quad (5)$$

where u represents the residuals from a GLM model with an identity link, and y is the dependent variable (cost in this case). The coefficient α_1 indicates which GLM variance function is appropriate. An α_1 of 0 implies constant variance (Gaussian), an α_1 of 1 implies a variance that is proportional to the mean (Poisson), and an α_1 of 2 implies that the variance is proportional to the mean squared (gamma). These tests say nothing about the correct link function.

Blough et al. (1999)^[23] presented a method to find the optimal GLM specification. Since many common link functions can be represented by a link from the power family (e.g., power = 1 is the identity link, and power = 0 is the log link), searching over multiple power links (from, for example, -2 to 2 by units of 0.5) and different families, one can determine the best model. The model with the lowest deviation is most desirable when searching within one family, and the model with the lowest AIC value is the most desirable when searching among families.

A more recent method for determining both the best GLM specification is the EEE estimator proposed by Basu and Rathouz (2005).^[27,28] The EEE estimator uses the power distributions to characterize the variance function. Specifically, $V(y_i) = \theta_1 \mu_i^{\theta_2}$. This specification nests the most popular variance functions: for example, it represents a Gaussian family when $\theta_2 = 0$ and a gamma family when $\theta_2 = 2$. The EEE estimator

characterize the link function. Specifically, $g(\mu_i; \lambda) = \frac{\mu_i^\lambda - 1}{\lambda}$, and $g(\mu_i; \lambda) = \log(\mu_i)$. In addition to the log function, many common link functions can be represented by this function including the identity ($\lambda = 1$), the square root ($\lambda = 0.5$), and the reciprocal ($\lambda = -1$). The EEE estimator uses 4 additional equations to simultaneously estimate the regression parameters, λ , θ_1 and θ_2 .

3 Data and Methods

We estimated incremental cost from a randomized behavioral intervention meant to promote evidence-based pain treatments for hospitalized patients aged over 65 years with hip

fractures. Further information on the intervention, data collection, and estimation are available elsewhere.^[17,18] One goal of that study was to estimate the effect of the intervention on inpatient costs. In the study, 12 hospitals were randomized into either an intervention or comparison group. The 6 intervention hospitals received a multi-faceted, interdisciplinary intervention promoting the adoption of evidence-based pain practices. The study collected hospital-stay data for individual patients with hip fractures from the intervention and control hospitals in two time periods – the year prior to the intervention (prior phase) and for a one-year period starting 90 days after the onset of the intervention (intervention phase). The final sample included 1,378 patients.

The data from this study are useful for our purposes here because the cost data are skewed, and the incremental cost associated with the intervention does not appear to be affected by the other covariates in the model. The distribution of costs in these data is pictured in Figure 1 in the appendix. The positive skew is apparent as the mean cost was \$8,050.38, almost \$1000 more than the median cost of \$7,145.99. The variance was 1.68×10^7 , and the skewness coefficient was 2.78. We also simulated data using linear models with skew, and with and without a measured (to the researcher) covariate. Specific information about the data can be found in the appendix. Summary statistics for the simulated data are found in Table I. Each set of observations was estimated using the models proposed earlier using the GLM command in STATA 11. The incremental effects associated with the treatment were calculated using the margins command in STATA, and the EEE estimates were obtained using the pglm package for STATA.

4 Results

4.1 Application to Hip Fracture Intervention Data

To examine the full effect of the intervention on cost, we examined the relationship between the covariates and the intervention. The Appendix contains OLS estimates of the effect of the behavioral intervention on cost using a specification in which the intervention was interacted with covariates -- the number of procedures used by the patient, the number of diagnoses reported by the patient, and patient age (Table AI). In the model without the interaction terms the intervention reduced inpatient costs by \$1,500 and inpatient cost increased with the number of procedures and diagnoses. Patient age had no effect on cost. In the second model, no statistically significant interactions between the intervention and procedures, diagnoses, and patient age were found, suggesting that the incremental effect of the intervention on inpatient costs does not vary with these covariates. In addition, the F-statistic for the interaction terms as a group was 1.45 ($p = 0.226$), which implies that the interaction terms as a group were not informative either.

The incremental effect of the intervention for the Gaussian-identity, Gaussian-log, gamma-log GLMs, and the EEE estimates are presented in Table II. For these data, the estimated incremental effect of the intervention differs: from a cost savings of \$1325.77 for the model selected by the EEE estimator, to cost savings of \$1612.13 for the GLM, Gaussian family, log link model. The standard error on the EEE model is the smallest.

A grid search concluded that the best model was a GLM model with a Gaussian family and a log link (power = 0). As shown in Table II, this model has the lowest AIC value. However, these models do not differ substantially in this regard: the AIC values are all between 18.721 and 19.921.

The Park test was performed on these data as well, and the value of a_1 is estimated to be 1.66 (std. error = 0.19) using residuals from an OLS model. The results of the Park Test suggest that the gamma family is the best choice for these data. This result highlights the difficulty in choosing the correct GLM model: although the Park Test indicates that the gamma family is the best choice for these data, using a grid search indicated that the Gaussian-log model is the best choice.

For the EEE model, $\lambda=0.3694$ ($p=0.111$; 95% CI: $-0.0952, 0.8341$); $\theta_1 = 0.1071$ ($p<0.001$, 95% CI: $0.0948, 0.1195$) and $\theta_2 = 1.5695$ ($p<0.001$, 95% CI: $1.088, 2.051$). In terms of standard GLM models, these EEE results would correspond to a cube-root link and a gamma family. Compound tests of these parameters rejected all of the other models considered in this paper except the gamma-log model ($p = 0.0980$). The identity-link models were rejected, but we cannot reject the log-link models. Although the results of this model agree with those generated by the Park test, they do not agree with the results from the grid search.

We examined the incremental effect of the intervention for various levels of covariates. For this exercise we chose covariates whose coefficients were significantly different from zero in the OLS model. Results are presented in Table III.

For all the non-identity-link models, the effect of the intervention on costs varied with the values of these covariates. These results run counter to the findings in the Appendix (Table AI) that show no statistically significant interactions between the intervention and these covariates on costs. Also note that, while not statistically significant, the signs of the interaction terms in the appendix suggest that the cost savings associated with the intervention fall as the number of diagnoses and procedures increase. The results of the Park Test suggest that the gamma-log model is appropriate here. However, in contrast with the identity link models, the incremental cost savings associated with the intervention increased with the number of diagnoses and procedures. In fact, the GLM gamma-log model suggests that the intervention cost saving nearly triples as the number of procedures goes from 0 to 4. Using this model, researchers could mistakenly conclude that the intervention in this study greatly reduces costs for patients with multiple procedures. For patients with 4 procedures, the GLM Gaussian-identity model suggests that the intervention reduced costs by \$1500.36, whereas using the GLM gamma-log model the cost savings estimate is more than double at \$3,059.08. Even using the more common GLM Gaussian-log model, the cost saving is estimated to be \$3036.81. These effects are less pronounced with the EEE model, where the cost savings estimates differ from those generated by the identity-link model, but not significantly.

4.2 Simulation Results for Model With One Dummy Variable

We also estimated the affect of treatment on cost using data from the simulation model when the only measured variable is a dummy variable representing treatment (see Appendix). We

analyzed twenty different simulation scenarios in which the cost of the unmeasured medical condition and percent of the population with the condition varied. Across scenarios we found that the log-linked models provided statistically significant parameter estimates of the relationship between treatment and cost, while only in scenarios with smaller unmeasured costs or a smaller percentage of patients with unmeasured costs were OLS estimates statistically significant. Notice in Table I that although these data were generated with a linear model, the distribution of costs exhibits large amounts of skew, and all are leptokurtic. Specifically, highly non-normal distributions can be derived from a linear data generating process.

Table IV reports the estimate of the incremental cost of the treatment and the standard error of the estimate. The estimate of the average treatment cost is between 999 and 1002 for all simulations considered. In addition, the standard errors for all of the models are similar. The Park Test and grid search methods were applied to these data, and in all cases, the Gaussian family models were chosen. For the grid search, the identity link was chosen as well. The EEE estimator was applied to these data, but it failed to converge.

4.3 Simulation Results for Model With an Additional Covariate

We now consider simulations in which 20% of the population had an extra cost of \$10,000. We chose the most conservative simulations: those with the least amount of skew. Even with the smallest amount of skew, the effects of the non-linear models on estimated incremental costs are clearly visible here. For comparison, we used more highly skewed data: a simulation in which 0.04% of the population had an extra cost of \$100,000 and the effects are more pronounced in the more-highly-skewed simulations. The first simulation contains a covariate with integer values between 1 and 10 inclusive, and the second and third simulations each have a covariate with values 10, 20, ..., 100. Results for the incremental effects of treatment on cost at the mean value of the covariate are presented in Table V. When the continuous covariate is added in the first simulation all models yield unbiased estimates of the treatment effect on cost at the mean value of the covariate. For the second simulation, whose covariate is simply the covariate in the first simulation multiplied by 10, the log-linked GLM models yield estimates of the treatment effect on cost that are obviously biased, and this bias increases with the increasing skew seen in the third simulation. For this second set of data, only the OLS model appears to be an unbiased, minimum variance estimator.

Again we examined the incremental effect of the intervention for various levels of covariates. We then assessed the relationship of the incremental estimated treatment cost across values of the covariate. In the GLM model with an identity link, the effect does not change. In the non-identity-link models, however, the incremental cost of the treatment varies with the value of the covariate. Table VI gives the incremental effects of the treatment for the first simulation in which the covariate value ranges from 1 to 10. Table VII gives the same results for the simulation in which the covariate value ranges from 10 to 100, and the simulation results for the same covariate but a more highly skewed sample are in the Appendix (Table VIII). For all log-based models, in Table VI the estimated incremental

effect of the treatment differs substantially between the lowest and highest covariate values: a difference of around \$100 for all of the log-link GLM models.

The only difference between the two simulations presented in Tables VI and VII is the scale of the covariate, but the scale of the covariate appears important when estimating incremental costs. In Table VII, for all non-identity-link models, the range in the estimated incremental effect of the treatment on costs between the lowest and highest covariate values is more than six times the range found in first simulation. In Table VIII, for the more highly skewed sample, this difference is five times.

A grid search was conducted to find the best model. Since more than one family of models was used, following Blough et al.,^[23] AIC values were used to find the best-fitting GLM model. Results indicated that for all simulations with covariates, the best model was a Gaussian model with a power = 1 link. This is the OLS model, and since our data were generated using a linear model, this was expected.

Results from Park Test also gave the expected result. Using the OLS model to calculate the residuals, the value of a_1 was 0.0000224 (0.4873856) for the simulation with covariate values from one to ten and -0.0000613 (0.1511228) for the simulation with covariate values from 10 to 100 (Standard errors are in parentheses.). Since these values of a_1 are closer to zero than any other number, the Park Test indicates that the Gaussian model fits these data best. Again, the EEE estimator failed to converge for these data sets.

In summary, the Park test and the grid search method selected the correct model for all simulated data – the Gaussian family – identity link model. This implies two important results: first, if the data are skewed, a log link is not always the correct choice. This model also implies no treatment heterogeneity, and using a model with an identity link does not impose heterogeneity where it does not exist. Using the hip-replacement-intervention data, the results are less clear, but both the grid search method and the EEE model suggest non-identity link functions, which, by design, imply treatment heterogeneity. This heterogeneity is imposed by the model and may not accurately represent the treatment effect. We estimated an OLS model with interaction terms and found that the implications of using a non-linear link in terms were inconsistent with that of the model with interaction terms. Both models should imply the effect of treatment on costs increases (or decreases) with changes in the covariates. If, as in this case, the interaction terms and the derivatives are inconsistent, a non-identity link model should not be used to investigate treatment heterogeneity.

5. Discussion

Calculating incremental effects of a treatment or intervention is important, especially in the current era of cost containment. However, in this paper we found that estimates of the incremental effect of treatment on costs vary substantially across models at given levels of model covariates. We found that in models using non-identity-link functions that estimates of incremental treatment effects vary with the level of other covariates in the model and the direct relationship between the covariates and cost, even if in the true underlying model the covariates do not modify the relationship between treatment and cost. This results from the

chain rule for partial derivatives: non-identity-link models depend not only on the other covariates in the model but also on their coefficients. These results may not be evident on a first pass: indeed, we found that all models produced unbiased estimates of the incremental effect of an intervention when evaluating the estimated functions at mean values of the covariates.

However, substantial differences appeared among models when incremental effects were estimated at values of the remaining model covariates other than their means. Using data from a hip fracture study, the incremental effects of the treatment varied with the levels of covariates, even though interactions between the covariates and the treatment were not significantly different from zero. These varying treatment levels were most pronounced using common GLM specifications such as the gamma family-log link or Gaussian family-log link.

Because we used a data-generating process with a linear relationship between the dependent and independent variable, the incremental effects should be constant. But when using non-identity-link models, widely variable incorrect incremental effects were estimated when values of the covariates differed from their means. These incremental effects sometimes more than doubled over the range of the covariate. Also, the scale of the covariate mattered: larger covariate values produced larger differentials in incremental effects, as did more highly skewed simulations.

Many researchers use log links reflexively. As we have shown here, even with skewed data, this may not be the best strategy. The data used in the simulations presented here were highly skewed, but using a log link was inappropriate: not all skewed data require log links. In addition, researchers need to be mindful of the effects of the log link on the incremental effects of the covariates as presented here.

This paper has some limitations in that the linear cost models may not be representative of many healthcare-related costs. For example, treating pneumonia may be more expensive for a patient with several co-morbidities than a healthier patient. However, the use of non-linear models may lead to estimates of cost difference across pneumonia patients that are artifacts of the functional form chosen and do not reflect actual cost differences.

6. Conclusion

Modeling cost has always been important, but with comparative effectiveness of treatments at the forefront of government policy, comparing the cost of treatments will become even more important. Policymakers will need accurate estimates of cost in order to make judgments about treatments, and researchers need to be careful about how they interpret and present their results to a non-technical audience. Investigators may be well aware of the limitations of their approaches, but their readers may not. Because many researchers engage in model selection by citation, the popularity and application of non-linear (especially log-link) models will most likely increase. Our results show that researchers need to consider if the variability in the incremental effects generated by non-linear models is appropriate for the cost data under consideration. Specifically, does it make sense that, like in our example,

cost savings should increase with the number of unrelated diagnoses? Failure to do so may lead to incorrect conclusions about the true cost of a treatment or intervention at levels of covariates other than the mean. As shown here, the effects can be dramatic.

Acknowledgments

This paper was funded in part by a grant from the National Institutes of Health, a University of Iowa KL2 (LAP).

Appendix

The model without a measured covariate had the following functional form:

$$C_i = 5000 + 1000 \cdot T_i + S \cdot K_i + \varepsilon_i \quad (A1)$$

where C_i represents the medical costs of an episode of care for patient “i” with a given medical condition over a fixed time period; T_i equals 1 if a patient receives a given treatment during the episode of care, 0 otherwise; \$5,000 equals the average cost of care for the given medical condition without T_i ; S equals the cost associated with the treatment of another expensive medical condition that may occur during the episode of care; K_i equals 1 if an expensive medical condition occurs, 0 otherwise; and ε_i is a random error. In this model the average cost associated with treatment T_i is constant -- \$1,000. K_i is assumed to be unmeasured by the researcher, and as a result the empirical error term will be $(S \cdot K_i + \varepsilon_i)$ which will be skewed. To focus on the effects of skew, our simulated variables T_i and K_i were constructed orthogonally. Note that no data transformations can fix the omitted variable bias that would occur if T_i and K_i were correlated and K_i was omitted from estimation. In our simulations we varied the percent of patients with an expensive medical condition ($K_i = 1$) using .04%, 1%, 5%, 10% and 20%, and the cost associated with the expensive medical condition using \$10,000, \$25,000, \$50,000, and \$100,000. The error term, ε_i , was distributed $N(0,10)$.

The model with a measured covariate had the following functional form:

$$C_i = 5000 + 1000 \cdot T_i + 100 \cdot X_i + 100,000 \cdot K_i + \varepsilon_i \quad (A2)$$

where X_i represents the measured covariate. We simulated this model using two different ranges for X_i (integers 1-10) and (integers 10, 20, ..., 100). To focus on the effects of X_i , in these simulations we fixed the percent of patients with an expensive medical condition ($K_i = 1$) at 20% and the amount of cost associated with the expensive medical condition at \$100,000.

Each set of data contained 500 sets of 1000 observations. For brevity, for each assumed cost level “S” in equation 5 we only report the simulation scenarios for the percentage of population $K_i = 1$ in which the OLS estimates crossover from being statistically insignificant to statistically significant. For example, with S equal to 100K, the estimated OLS parameter was statistically insignificant when 1% of the population had $K_i = 1$, but was statistically significant when 0.4% of the population had $K_i = 1$. When S was equal to 10K, the OLS estimates were statistically significant at all of the population percentages we analyzed.

References

1. M, Gyldmark; Morrison, GC. Demand for health care in Denmark: results of a national sample survey using contingent valuation. *Soc Sci Med*. 2001; 53(8):1023–36. [PubMed: 11556772]
2. Meng H, Wamsley BR, Eggert GM, VanNostrand JR. Impact of a health promotion nurse intervention on disability and health care costs among elderly adults with heart conditions. *J Rural Health*. 2007; 23(4):322–31. [PubMed: 17868239]
3. Shinall MC Jr, Koehler E, Shyr Y, Lovvorn HN III. Comparing cost and complication of primary and staged surgical repair of neonatally diagnosed Hirschsprung's disease. *J Pediatr Surg*. 2008; 43(12):2220–5. [PubMed: 19040939]
4. Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc*. 1983; 78:605–610.
5. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for health care. *Journal of Business and Economic Statistics*. 1983; 1(2):115–126.
6. Manning WG, Newhouse JP, Duan N, Keeler EB, Liebowitz A, Marquis MS. Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review*. 1987; 7(3):251–277. [PubMed: 10284091]
7. Chen YY, Wang FD, Liu CY, Chou P. Incidence rate and variable cost of nosocomial infections in different types of intensive care units. *Infect Cont Hosp Ep*. 2009; 30(1):39–46.
8. Delea TE, Hariwara M, Dalal AA, Stanford RH. Healthcare use and costs in patients with chronic bronchitis initiating maintenance therapy with fluticasone/salmeterol vs. other inhaled maintenance therapies. *Curr Med Res Opin*. 2009; 25(1):1–13. [PubMed: 19210134]
9. Jayadevappa R, Chhatre S, Wein AJ, Malkowicz SB. Predictors of patient reported outcomes and cost of care in younger men with newly diagnosed prostate cancer. *Prostate*. 2009; 69(10):1067–76. [PubMed: 19343738]
10. Shea AM, Curtis LH, Hammill BG, Kowalski JW, et al. Resource use and costs associated with diabetic macular edema in elderly persons. *Arch Ophthalmol*. 2008; 126(12):1748–54.
11. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ*. 2001; 20(4):461–494. [PubMed: 11469231]
12. Barnett PG, Chow A, Joyce VR, Bayoumi AM, Griffin SC, et al. for the OPTIMA team. Determinants of the cost of health services used by veterans with HIV. *Medical Care*. 2011; 49(9): 848–856. [PubMed: 21610542]
13. Meltzer D, Manning WG, Morrison J, Shah MN, Lei J, Guth T, Levinson W. Effects of physician experience on costs and outcomes on an academic general medicine service: results of a trial of hospitalists. *Annals of Internal Medicine*. 2002; 137(11):866–875. [PubMed: 12458986]
14. Harris RD, Hanson C, Christy C, Adams T, Banks A, Willis TS, Maciejewski ML. Strict hand hygiene and other practices shortened stays and cut costs and mortality in a pediatric intensive care unit. *Health Affairs*. 2011; 20(9):1751–1761. [PubMed: 21900667]
15. Escarce JJ, Kapur K. Racial and ethnic differences in public and private medical care expenditures among aged Medicare beneficiaries. *The Milbank Quarterly*. 2003; 81(2):249–275. [PubMed: 12841050]
16. Williams MD, Shah ND, Wagie AE, Wood DL, Frye MA. Direct costs of bipolar disorder versus other chronic conditions: an employer-based health plan analysis. *Psychiatric Services*. 2011; 62(9):1073–1078. [PubMed: 21885587]
17. Basu A, Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variable function models. *Biostatistics*. 2005; 6(1):93–109. [PubMed: 15618530]
18. Basu A. Extended generalized linear models: simultaneous estimation of flexible link and variance functions. *The Stata Journal*. 2005; 5(4):501–516.
19. Basu A, Arondekar BV, Rathouz PJ. Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Econ*. 2006; 15(10):1091–107. [PubMed: 16518793]
20. Hallinen T, Martikainen JA, Soini EJ, Suominen L, Aronkyto T. Direct costs of warfarin treatment among patients with atrial fibrillation in a Finnish health care setting. *Curr Med Res Opin*. 2006; 22(4):683–92. [PubMed: 16684429]

21. Montez-Rath M, Christiansen CL, Ettner SL, Lovel S, Rosen AK. Performance of statistical models to predict mental health and substance abuse cost. *BMC Med Res Methodol.* 2006; 6:53. [PubMed: 17067394]
22. Powers CA, Meyer CM, Roebuck MC, Vaziri B. Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Med Care.* 2005; 43(11):1065–72. [PubMed: 16224298]
23. Brooks JM, Titler MG, Ardery G, Herr K. Effect of evidence-based acute pain management practices on inpatient costs. *Health Serv Res.* 2009; 44(1):245–63. [PubMed: 19146567]
24. Titler MG, Herr K, Brooks JM, et al. Translating Research into Practice Intervention Improves Management of Acute Pain in Older Hip Fracture Patients. *Health Serv Res.* 2009; 44(1):264–287. [PubMed: 19146568]
23. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ.* 1999; 18:153–171. [PubMed: 10346351]
25. Buntin B, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *J Health Econ.* 2004; 23:525–542. [PubMed: 15120469]
26. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ.* 2005; 24:465–488. [PubMed: 15811539]
27. Hardin, JW.; Hilbe, JM. *Generalized Linear Models and Extensions.* 2nd. Stata Press; College Station(TX): 2007.

Key Points for Decision Makers

- Healthcare costs are often skewed, and in response, researchers have used log-based transformations.
- When using log-based models, the incremental effects differ at different levels of the covariates, and this can cause dramatic effects on predicted cost.

Figure 1.
Distribution of Costs: Hip-Fracture Data. Mean: 8050.382; Variance: 1.68×10^7 , Skewness:
2.777, Kurtosis: 18.616.

Table I

Cost Summary Statistics for Simulated Data Sets. (These numbers are averages for the 500 simulations in each group; each simulation had 1000 observations; and all values are given in US\$)

Data Group		Mean	Variance	Skewness	Kurtosis
Cost Associated with the Unmeasured Condition (S)	Percent of Population with the Unmeasured Condition.				
100K	.04	5899.998	4.01×10^7	15.570	244.956
100K	1	6500.008	9.93×10^7	9.812	97.532.
50k	1	6000.011	2.50×10^7	9.702	96.119
50K	5	8000.013	1.19×10^8	4.116	17.989
25K	10	8000.028	5.66×10^7	2.649	8.0659
25k	20	10499.99	1.00×10^8	1.494	3.249
10K	20	7500.013	1.63×10^7	1.466	3.242

Table II

The Incremental Effect of the hip fracture intervention and its Standard Error at the mean of the covariates for 4 models considered, measured in US\$. N=1378.

Model	GLM Gauss-ID (OLS)	GLM Gauss-Log	GLM Gamma- Log	EEE
ME (SE)	-1500.362 (341.55)	-1612.13 (423.26)	-1464.10 (324.40)	-1325.77 (323.44)
Predicted mean costs	8050.38	7643.49	7659.72	8049.57
AIC	18.787	18.721	19.921	19.338

Table III

The Incremental Effect of the Treatment and its Standard Error at different levels of two significant covariates for 4 models considered, measured in US\$. N = 1378.

Number of Diagnoses	GLM Gauss-ID (OLS)	Gauss-Log	Gamma-Log	EEE
0	-1500.36 (341.55)	-1270.50 (334.46)	-1225.79 (265.93)	-1912.12 (537.29)
1	-1500.36 (341.55)	-1319.38 (346.80)	-1260.89 (274.15)	-1940.51 (552.67)
2	-1500.36 (341.55)	-1370.14 (359.76)	-1296.98 (282.75)	-1968.07 (568.53)
3	-1500.36 (341.55)	-1422.85 (373.36)	-1334.12 (291.75)	-1995.79 (584.86)
4	-1500.36 (341.55)	-1477.59 (387.64)	-1372.31 (301.15)	-2023.67 (601.67)
5	-1500.36 (341.55)	-1534.43 (402.65)	-1411.60 (311.00)	-2051.72 (618.96)
6	-1500.36 (341.55)	-1593.46 (418.42)	-1452.01 (321.28)	-2079.92 (636.73)
7	-1500.36 (341.55)	-1654.76 (435.00)	-1493.58 (332.04)	-2108.28 (654.97)
	-1500.36 (341.55)	-1718.42 (452.43)	-1536.34 (343.28)	-2136.81 (673.70)
9	-1500.36 (341.55)	-1784.53 (470.76)	-1580.33 (355.04)	-2165.49 (692.90)
Number of Procedures				
0	-1500.36 (355.04)	-1319.36 (343.91)	-1158.74 (247.29)	-1129.72 (262.45)
1	-1500.36 (341.55)	-1625.09 (427.01)	-1477.01 (327.47)	-1345.75 (309.86)
2	-1500.36 (341.55)	-2001.66 (536.46)	-1882.73 (438.46)	-1577.39 (382.72)
3	-1500.36 (341.55)	-2465.50 (680.78)	-2399.88 (591.67)	-1824.21 (491.73)
4	-1500.36 (341.55)	-3036.81 (871.28)	-3059.08 (802.37)	-2085.82 (642.19)

Table IV

The Incremental Effect of “Treatment” and its Standard Error for 7 groups of simulated data and 3 models considered. The true value is 1000. (Results are the average of the 500 simulations in each group; each simulation had 1000 observations; and all values are given in US\$.)

Data Group (skew amount, % of population skewed)	GLM Gauss-ID (OLS)	GLM Gauss-Log	GLM Gamma-Log
10K, 20%	999.99 (0.65)	1000.48 (0.65)	1000.48 (0.65)
25K, 20%	999.98 (0.60)	1000.73 (0.60)	1000.73 (0.60)
25k, 10%	1000.02 (0.61)	1001.33 (0.61)	1001.33 (0.61)
50k, 5%	1000.05 (0.63)	1001.35 (0.64)	1001.35 (0.64)
50k, 1%	999.98 (0.64)	1002.31 (0.64)	1002.31 (0.64)
100k, 1%	1000.01 (0.64)	1001.99 (0.64)	1001.99 (0.64)
100K, 0.04%	1000.04 (0.64)	1002.44 (0.65)	1002.44 (0.65)

Table V

The Incremental Effect of the “Treatment” and its Standard Error at the mean of the covariate. (Results are the average of the 500 simulations in each group; each simulation had 1000 observations; and all values are given in US\$.)

Model	GLM Gauss-ID (OLS)	GLM Gauss- Log	GLM Gamma- Log	
10K, 20%, Covariate values 1, 2,...,10				
ME (SE)	999.95 (0.616)	999.97 (0.618)	1002.53 (0.619)	
10K, 20%, Covariate values 10, 20,...,100				
ME (SE)	999.96 (0.600)	956.08 (0.591)	1053.05 (0.649)	

Table VI

The Incremental Effect of the “Treatment” and its Standard Error at different levels (1,2,...,10) of the covariate for simulations where 20% of the sample had an extra \$10,000 expense. (Results are the average of the 500 simulations each with 1000 observations, and all values are given in US\$.)

Covariate Value	GLM Gauss-ID (OLS)	GLM Gauss-Log	GLM Gamma-Log
1	999.95 (0.616)	944.01 (0.588)	945.95 (0.588)
2	999.95 (0.616)	955.76 (0.594)	957.83 (0.594)
3	999.95 (0.616)	967.66 (0.600)	969.86 (0.601)
4	999.95 (0.616)	979.71 (0.607)	982.04 (0.608)
5	999.95 (0.616)	991.90 (0.614)	994.37 (0.615)
6	999.95 (0.616)	1004.25 (0.621)	1006.86 (0.623)
7	999.95 (0.616)	1016.76 (0.629)	1019.51 (0.631)
8	999.95 (0.616)	1029.42 (0.637)	1032.31 (0.640)
9	999.95 (0.616)	1042.23 (0.645)	1045.27 (0.648)
10	999.95 (0.616)	1055.21 (0.654)	1058.40 (0.658)

Table VII

The Incremental Effect of the “Treatment” on Cost and its Standard Error at different levels (10,20,...,100) of the covariate for simulations where 20% of the sample had an extra \$10,000 expense. (Results are the average of the 500 simulations, each with 1000 observations, and all values are given in US\$.)

Covariate Value	GLM Gauss-ID (OLS)	GLM Gauss-Log	GLM Gamma-Log
10	999.96 (0.600)	660.96 (0.413)	715.19 (0.442)
20	999.96 (0.600)		774.35 (0.478)
30	999.96 (0.600)	769.56 (0.459)	838.41 (0.518)
40	999.96 (0.600)	830.37 (0.516)	907.76 (0.560)
50	999.96 (0.600)	895.99 (0.556)	982.85 (0.606)
60	999.96 (0.600)	966.80 (0.599)	1064.15 (0.656)
70	999.96 (0.600)	1043.20 (0.645)	1152.18 (0.711)
80	999.96 (0.600)	1125.64 (0.696)	1247.49 (0.770)
90	999.96 (0.600)	1214.60 (0.750)	1350.68 (0.834)
100	999.96 (0.600)	1310.58 (0.809)	1462.41 (0.904)

Table VIII

The Incremental Effect of the “Treatment” on Cost and its Standard Error at different levels (the mean, 10,20, ...,100) of the covariate with highly skewed data, where 0.04% of the sample had an extra \$100,000 expense. (Results are the average of the 500 simulations, each with 1000 observations, and all values are given in US\$.)

	GLM Gauss- ID (OLS)	GLM Gauss- Log	GLM Gamma-Log
Mean	999.99 (0.653)	966.38 (0.641)	1038.98 (0.693)
10	999.99 (0.653)	703.26 (0.469)	747.69 (0.499)
20	999.99 (0.653)	751.32 (0.500)	800.55 (0.534)
30	999.99 (0.653)	802.76 (0.534)	857.14 (0.572)
40	999.99 (0.653)	857.52 (0.570)	917.74 (0.612)
50	999.99 (0.653)	916.13 (0.608)	982.62 (0.655)
60	999.99 (0.653)	978.74 (0.650)	1052.09 (0.701)
70	999.99 (0.653)	1045.6 (0.694)	1126.46 (0.751)
80	999.99 (0.653)	1117.09 (0.741)	1206.10 (0.804)
90	999.99 (0.653)	1193.44 (0.791)	1291.36 (0.861)
100	999.99 (0.653)	1275.01 (0.845)	1382.65 (0.923)

Table AI

OLS Estimates of the Effect of the Behavioral Intervention on Inpatient Costs (in US\$) With and Without Interaction Terms. N=1378. (Models also included variables for study phase, patient gender, discharge status, and hospital indicator variables).

	model without interaction terms ^a		model with interaction terms ^b	
	Estimate (robust std error)	p value	Estimate (robust std error)	p value
Intervention	-1,500.36 (341.55)	< .0001	-2,081.40 (2003.41)	0.299
number of procedures	2,500.90 (403.01)	< .0001	2,275.96 (428.01)	<.0001
number of diagnoses	269.00 (38.21)	<.0001	236.20 (41.20)	<.0001
patient age	1.50 (9.78)	0.879	4.81 (11.13)	0.666
intervention*number of procedures			913.48 (980.96)	0.352
intervention*number of diagnoses			131.80 (85.06)	0.122
intervention*patient age			-14.22 (21.68)	0.512
F-statistic	30.70	<.0001	27.05	<.0001
F-statistic for interaction terms			1.45	0.226
R-squared	0.5131		0.5170	