

How is Healthcare Cost Data Distributed? Using Proc Univariate to Draw Conclusions about Millions of Different Customers

Aran Canes, Cigna¹, Raleigh, NC

ABSTRACT

The modelling of health care cost and utilization data has received substantial attention from the academic community.² Different methodological approaches have been used, tested and contrasted on various health care datasets. Some of the more simple approaches include Ordinary Least Squares, Generalized Linear Methods and log transforms of the dependent variables. In addition, more sophisticated non-parametric methods have been proposed and tested. The conclusion of most researchers is that different approaches will work best for different datasets.

However, despite a plethora of methodological comparisons few papers report how health care cost data is actually distributed. This may be due to difficulties accessing data because of privacy concerns, or because many analysts simply assume that different slices of cost will have different distributions.

While this paper does not try to falsify the hypothesis that, in some instances, different slices of healthcare cost data are distributed differently, it reaches a surprising conclusion: a wide range of healthcare cost data slices among Cigna's customers: all are distributed approximately log normally once the substantial part of the population that is zero-cost is excluded. This is true whether one looks at pharmacy or medical cost, different ways of purchasing insurance among customers or comparing customers who stayed eligible for a full customer year versus customers who only temporarily have coverage. These results are reached using the histogram and quantile measurements available to all SAS® users in PROC UNIVARIATE.

Since there is a lack of empirical information regarding how the cost data of large customer populations is distributed, this paper should provide significant help in assessing the validity of different methodological approaches. If the results are confirmed elsewhere, methodological discussions could be conducted with the knowledge that, if healthcare cost data has a substantial number of customers, the Probability Density Function that best matches the data will be the lognormal.

Key words: Health Care Cost Data, Probability Density Function, OLS Regression, Generalized Linear Models, Lognormal Distribution, Proc Univariate

INTRODUCTION

Performing regressions to predict future health care cost or to assess the effect of different variables on health care cost is a bedrock technique in healthcare data analytics. At first glance, the method of choice would seem to be ordinary least squares since this is the simplest to implement in SAS, and the assumptions of OLS seem to be met in many circumstances. These assumptions are:

- **IID observations:** Observations are independent and identically distributed
- **No Perfect Multicollinearity:** There are not combinations of predictor variables that contain the same information as another predictor variable
- **Exogeneity:** The predictor variables are not themselves dependent on the dependent variable
- **Homoscedasticity:** The variance of the error term does not change across the different observations³

Why then are more complicated techniques used like Generalized Linear Methods which employ a different distribution to model the dependent variable? Perhaps one might think that the significance tests depend upon a normally distributed error term and so, if one thinks that the significance tests may be biased, they could continue to use Ordinary Least Squares to determine the coefficients and make predictions. If the statistical significance of coefficients is needed, however, one could use a Generalized Linear Method.

¹

I would argue however, despite much contrary practice to the contrary, that this approach is wrong. While the normality of the error terms strictly speaking is not an assumption of OLS and it is true that the only direct bias occurs in the significance testing when this assumption does not hold, this does not mean that the distribution of the dependent variable is a matter of secondary significance as implied by the frequent use of Ordinary Least Squares.

This is because of a fundamental assumption of all linear regression; that is, that the dependent variable is related in a linear manner to the independent variables. For multivariate linear regression we try to find the coefficients which best fit the model:

$$y = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n$$

This makes sense if, using the textbook regression picture, the distribution of the dependent variable looks like this:

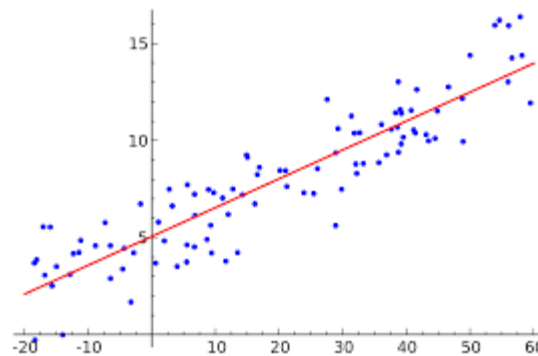


Figure 1- Textbook Picture of Regression

However, assume that your dependent variable is distributed according to a Poisson distribution (as shown below).⁴

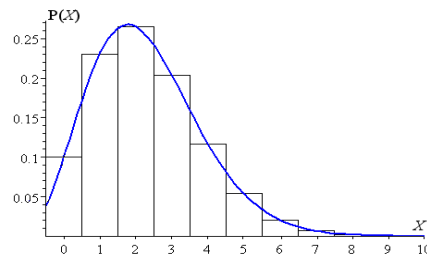


Figure 2- Poisson Distribution

How likely is it that, if you use OLS regression, your predictors will be linearly related to the dependent variable if the dependent variable is distributed as above⁵? If your data violates the linearity assumption, not only will the significance tests on the predictor variables be biased, but the actual coefficients will be biased because the model is not correctly specified.

This motivates the use of such techniques as Generalized Linear Models not only because they are useful but also because they are necessary to obtain correct results when modeling health care cost data. This also then motivates the topic of this paper: How is health care cost data actually distributed? If we agree that, prior to performing a regression, we need to know what distribution best describes the behavior of the dependent variable, are there any patterns in health care cost data that can assist us a starting point in our search for a distribution?

I will present evidence to suggest that, while individual cases may vary, when one looks at healthcare cost data with a sufficient number of customers, for a sufficient length of time and excluding those customers with zero cost, the distribution that best fits the dependent variable is generally the lognormal distribution. I believe this is an important finding because privacy and access concerns make it difficult for many scholars to obtain large amounts of healthcare data so arguments about techniques often have to be conducted on very small samples of real data. Working at Cigna, I fortunately can access data covering millions of different customers and, sufficiently aggregated and de-identified, can publish the results of different experiments. I hope that this paper will cause other researchers with similar access to large amounts of data to similarly contribute to answering the empirical question: how is healthcare cost data distributed?

METHODOLOGY

Fortunately, the code to test different probability density functions is relatively simple to implement in SAS. One of the most basic statistical functions, Proc Univariate, is often sufficient to determine a distribution. The code for determining distributions is as follows:

```
PROC UNIVARIATE DATA=dis.rxdistributionel3;

HISTOGRAM pay_amt / BETA (COLOR=YELLOW)
WEIBULL (COLOR=RED) EXPONENTIAL (COLOR=BLUE) GAMMA (COLOR=GREEN) NORMAL (COLOR=BLACK)
LOGNORMAL (COLOR=ORCHID);
WHERE 0<pay_amt;
QUIT;
```

In this code snippet, the histogram statement plays the major part. The statement creates a histogram of the specified variable with the different distributions overlaid on top of the histogram. Because the statistical tests for distributional match often are too sensitive to determine if a distribution is a good match, simply looking at the histogram and comparing the actual versus predicted values for different quantiles of the variable is often more useful than significance testing.⁶

The output of Proc Univariate will often look like this:

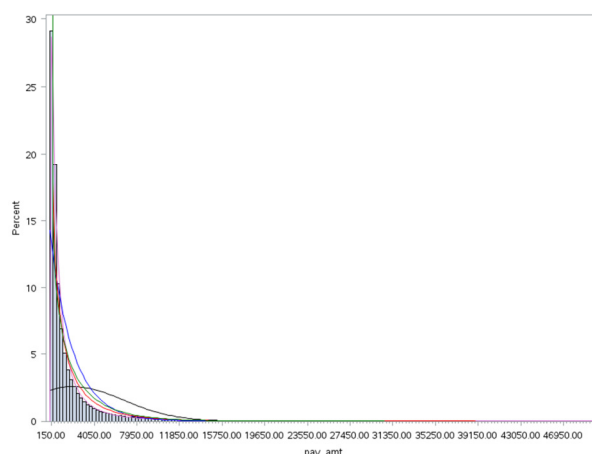


Figure 3- Histogram of Health Care Cost

As one can see, even with eliminating extreme values, the histogram can be hard to read. Before showing a means of making the graph more useful, let's look at the predicted versus actual values for the dependent variable if we try to fit it to a normal distribution.

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
1	\$ 104.25	\$ (8,764.33)
5	\$ 122.59	\$ (5,580.16)
10	\$ 149.60	\$ (3,882.69)
25	\$ 258.71	\$ (1,046.29)
50	\$ 642.76	\$ 2,105.15
75	\$ 1,861.51	\$ 5,256.59
90	\$ 4,776.56	\$ 8,092.98
95	\$ 8,461.55	\$ 9,790.45
99	\$27,169.26	\$ 12,974.62

Table 1-Observed vs. Estimated Quantiles for Normal Distribution

As indicated in the table, the normal distribution does a horrendous job of matching health care cost data, mostly, though not exclusively, because health care cost data does not go below zero. The normal distribution, based on the mean, predicts that the first, fifth, tenth, and twenty-fifth percentiles will all be negative values—clearly nonsense for health care cost data.

One method which can make the histogram charts more useful is to take the log of the variable under investigation. Of course, you are then trying to match the log of the data to the gamma, normal, lognormal, etc. so you have to be careful to remember that if the lognormal is not a good match since it is a comparison of the log of the variable to the

lognormal, not the actual variable to the lognormal, which one is are considering. If the normal distribution proves to be a good match then you have proved that the lognormal distribution is a good match—which you could have seen in the original histogram if there wasn't such a spike close to zero. Below is an example of a plot of log cost data.

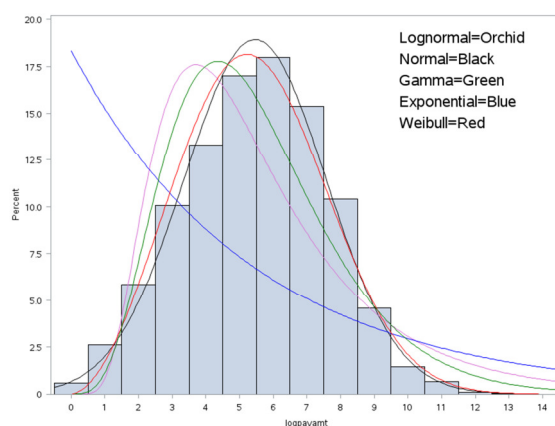


Figure 4—Histogram of Log Cost Data

As one can see, no distribution matches the data precisely, but the normal does the best job of capturing both the lower and upper tails of the distribution. This can be seen more precisely by comparing the actual quantiles of the normal with the next best fit, which is the Weibull distribution:

Quantiles for Normal Distribution			Quantiles for Weibull Distribution		
Percent	Quantile		Percent	Quantile	
	Observed	Estimated		Observed	Estimated
1	0.73237	0.55484	1	0.73237	1.18249
5	1.89612	1.99176	5	1.89612	2.11637
10	2.62249	2.75777	10	2.62249	2.73673
25	3.96916	4.03775	25	3.96916	3.91757
50	5.53038	5.4599	50	5.53038	5.36294
75	6.96482	6.88205	75	6.96482	6.86917
90	8.11669	8.16202	90	8.11669	8.23376
95	8.78719	8.92804	95	8.78719	9.04506
99	10.31702	10.36496	99	10.31702	10.54629

Table 2—Quantiles of Observed vs. Estimated for Normal and Weibull distributions for the log of cost data

At this point we have shown that there is a need to know the distribution of the dependent variable, demonstrated how to test this in Proc Univariate both by chart and quantiles, and showed that taking the log of the dependent variable often leads to easier to interpret results. What remains is to prove the thesis of this paper, that, for a large enough population over a long enough length of time, the lognormal distribution will best model healthcare cost data once those with zero cost are excluded.

RESULTS

Space does not permit me to show all the different slices of the Cigna data warehouse made to test the fit of the lognormal distribution. However, a few examples may suffice to illustrate the pattern, and other researchers then can confirm or falsify the hypothesis that large slices of healthcare cost data tend towards lognormal distributions.

The first and most obvious example is to look at annual medical cost for all Cigna customers who had a cost greater than zero. Below is the log of annual medical cost for all customers who made a claim with cost>0 in calendar year 2015.

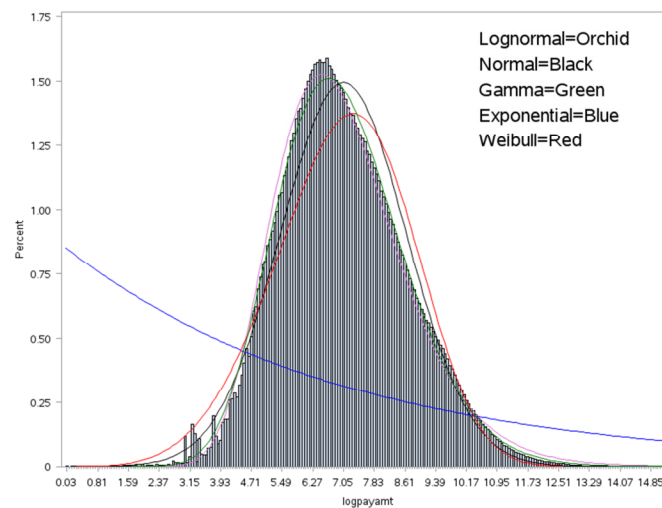


Figure 5-Log of Medical Cost for All 2015 Cigna Customers

While one can see that the normal (recall that we are looking at the log of cost instead of cost directly) does not capture the peak of the distribution exactly it does a good job of modeling the lower and upper tails. The other distributions, while approximating the lognormal, do not show as good a fit. Next is a similar chart looking at the log of medical cost for those customers who purchased Cigna insurance not through their employer, but as individuals. Again we see the normal distribution closely matching the log of the dependent variable.

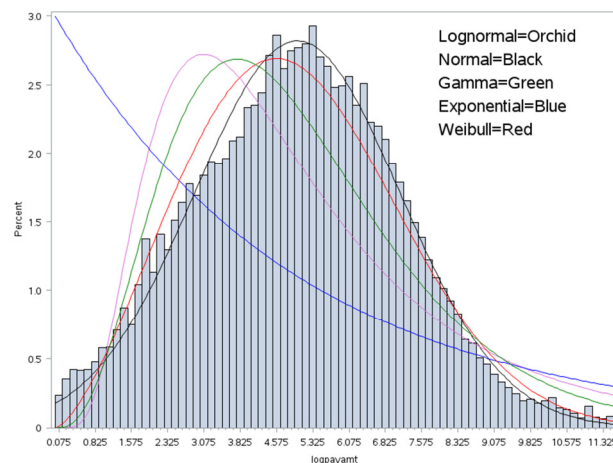


Figure 6 Log of Medical Cost for All Cigna Customers who Purchased Insurance as Individuals

Now we can look at pharmacy cost from two perspectives: all customers who had a claim > \$0 in 2015 and, in addition, those who were fully eligible for all of 2015. First let's look at a similar graph to the first one shown of medical cost except that here we are looking at the log of *pharmaceutical* spending:

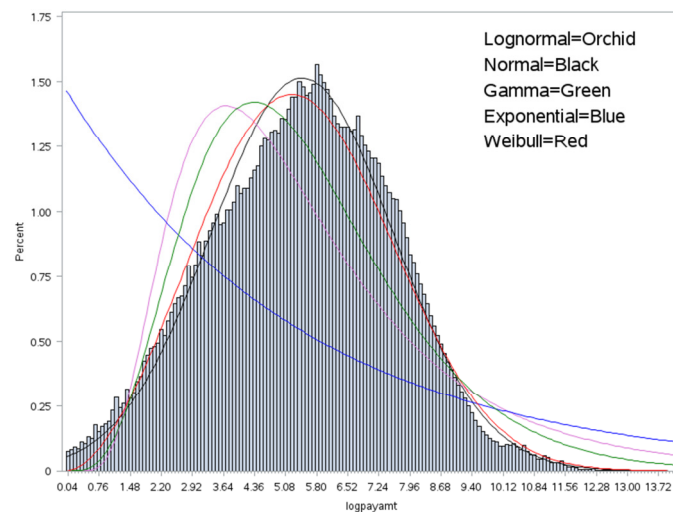


Figure 7 Log of Pharmaceutical Cost for All 2015 Cigna Customers

Again, we see the black line representing the normal distribution as the best fit to the log of pharmacy cost. Lastly, we look at fully eligible pharmacy customers in 2015. Fully eligible refers to those customers who had Cigna insurance throughout the calendar year.

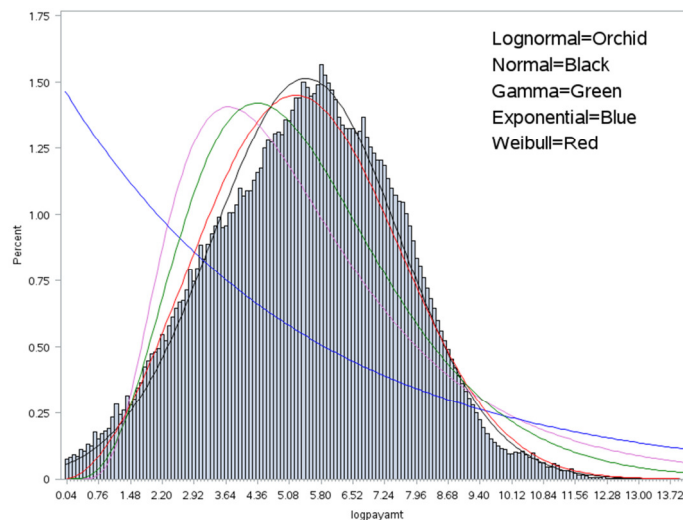


Figure 8-Log of Pharmaceutical Cost for All 2015 Fully Eligible Cigna Customers

The distribution which best captures the peak and right tail is again the normal distribution (recall we're looking at the log of cost) and so the lognormal distribution again provides the best fit of the data. Because graphical inspection is subject to human error it is always best to look at the quantile data produced by Proc Univariate to confirm the result.

Quantiles for Normal Distribution			Quantiles for Weibull Distribution		
Percent	Quantile		Percent	Quantile	
	Observed	Estimated		Observed	Estimated
1	0.73237	0.55484	1	0.73237	1.18249
5	1.89612	1.99176	5	1.89612	2.11637
10	2.62249	2.75777	10	2.62249	2.73673
25	3.96916	4.03775	25	3.96916	3.91757
50	5.53038	5.4599	50	5.53038	5.36294
75	6.96482	6.88205	75	6.96482	6.86917

90	8.11669	8.16202		90	8.11669	8.23376
95	8.78719	8.92804		95	8.78719	9.04506
99	10.31702	10.36496		99	10.31702	10.54629

Table 3- Quantiles of Observed vs. Estimated 2015 Log Pharmaceutical Cost for Fully Eligible Customers for Normal and Weibull Distribution

The normal distribution clearly is the closest match to the log of the pharmacy cost for all fully eligible 2015 customers making the lognormal the distribution which best fits the data.

While I have only included a sample of my investigations into health care cost: 2015 Medical Cost, 2015 Individual Segment Medical Cost, 2015 Pharmacy Cost and 2015 Fully Eligible Pharmacy Cost I have looked at many more slices of aggregate data: different lengths of time (e.g. calendar year vs. 6 months), different customer segments (e.g. individual vs. group), different years (2014 vs. 2015), different disease specific categories (e.g. customers with cardiovascular conditions) and in all my analyses I have found the lognormal distribution to be the best fit to aggregate customer data with one exception: HIV customers, who tend to be high cost because of the cost of antiviral therapy, were distributed closer to a gamma distribution.

I suggest as an additional hypothesis that there is a reason for this. The lognormal distribution is usually employed as a distribution to model growth factors in organisms.⁷ In other words, the heights of human beings of different ages exhibit a lognormal distribution. I would argue that annual health care cost, since it is additive and auto-correlated, could be expected to follow a distribution similar to that of growth in organisms. However, I welcome other researchers with access to aggregate cost data to analyze their data and see if it follows a similar distribution and so conforms to this hypothesis.

CONCLUSION

Because of the linearity assumption in all linear regressions, knowing the distribution of the dependent variable can be important, particularly if there are severe departures from normality. This paper has examined several different slices of Cigna cost data to see what distributions best model aggregate cost. What it finds is that the lognormal distribution appears to be the best fit if sufficiently long time spans and a large enough sample are used. Because the lognormal distribution is often used to model growth mechanisms in organisms, I have suggested that health care cost might be analogous since it is additive and auto-correlated. I have also suggested that other researchers with access to large amounts of customer data, perform their own exploratory analyses of their data to determine appropriate probability distributions to either confirm or refute these findings, and as an aid to those elsewhere who can develop statistically sophisticated models that are useful in real-world settings but which lack access to the data accessible to analytic professionals.

REFERENCES

1 "Cigna" as used herein refers to operating subsidiaries of Cigna Corporation, including Cigna Health and Life Insurance Company and Cigna Health Management, Inc. All Cigna products and services are provided exclusively by such operating subsidiaries.

2 See for example, P.Diehr, D. Yanez, A. Ash, M. Hornbrook and D.Y. Lin. *Methods for Analyzing Health Care Utilization and Costs*, Annual Review of Public Health, Vol.20 1999 125-144.

http://isites.harvard.edu/fs/docs/icb.topic79832.files/L08_O_E_Research_1_Cost_outcomes/Diehr.et.al.pdf

Dr. John Mullahy. *Econometric Modeling of Health Care Costs and Expenditures*, Medical Care, Vol. 47, No.7 pp. s104-s108.

https://www.jstor.org/stable/40221977?seq=1#fndtn-page_scan_tab_contents

3 These are drawn from Wikipedia's article on Ordinary Least Squares where a more technical description may be found: https://en.wikipedia.org/wiki/Ordinary_least_squares

4 This picture is taken from:

<http://www.intmath.com/blog/mathematics/maximum-value-of-a-poisson-distribution-4327>

5 This point is made in a book I highly recommend:

Piet de Jong and Gillian Z. Heller, 2008. *Generalized Linear Models for Insurance Data*. Cambridge, England. Cambridge University Press.

6 See SAS documentation on distributional hypothesis testing with Proc Univariate:

http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect008.htm

7 See Wikipedia's article on the Log-Normal Distribution:

https://en.wikipedia.org/wiki/Log-normal_distribution

ACKNOWLEDGMENTS

The author would like to thank Dr. Michael Manocchia, Dr. Saad Aslam and Mr. Jigar Shah for providing me the opportunity to conduct this analysis and for helpful suggestions throughout.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Aran Canes
Enterprise: Cigna Health and Life Insurance Company
Address: 701 Corporate Center Drive
City, State ZIP: Raleigh, NC 27607
Work Phone: 919-854-7261
E-mail: aran.canes@cigna.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.