

A simple goodness-of-fit test for continuous conditional distributions

Peter Veazie*
Zhiqiu Ye†

Abstract

This paper presents a pragmatic specification test for conditional continuous distributions with uncensored data. We employ Monte Carlo (MC) experiments and the 2011 Medical Expenditure Panel Survey data to examine coverage and the power to discern deviations from the correct model specification in distribution and parameterization. We carry out MC experiments using 2000 runs for sample sizes 500 and 1000. The experiments show that the test has accurate coverage under correct specification, and that the test can discern deviations from the correct specification in both the distributional family and parameterization. The power increases as sample size increases. The empirical example shows the test's ability to identify specific distributions from other candidates using real cost data. Although the test can be used as a goodness-of-fit test for marginal distributions, it is particularly useful as an easy-to-use test for conditional continuous distributions, even those with one observation per pattern of explanatory variables.

Keywords: Goodness-of-fit test; model specification test; conditional continuous distributions.

2010 AMS subject classification: 62F03‡

* University of Rochester, Rochester New York, USA; peter_veazie@urmc.rochester.edu.

† University of Rochester, Rochester New York, USA; sophieye999@gmail.com.

‡ Received on June 10th, 2020. Accepted on December 17th, 2020. Published on December 31st, 2020. doi: 10.23755/rm.v39i0.524. ISSN: 1592-7415. eISSN: 2282-8214. ©Peter Veazie et al. This paper is published under the CC-BY licence agreement.

1 Introduction

To determine whether a probability model is statistically adequate for representing a data generating process (DGP), it is common to test whether the model fits with a data set produced by that process. The investigation into the model specification of a conditional distribution is fundamental for methods such as Maximum Likelihood Estimation (MLE), which is consistent and asymptotically efficient only if the distribution is correctly specified (Amemiya, 1985). However, there are two key challenges for a general test of continuous conditional distribution models, if it is to be broadly adopted in applied sciences such as social and health sciences: First, is the sparse empirical information regarding the conditional distribution when patterns of the explanatory variables have few corresponding observations. Second, is the ease of use: many researchers do not have the background, time, or inclination to engage in complicated programming in order to implement a statistical test—to be useful to such researchers a test must be easily implemented.

Regarding sparse information, consider the data shown in Figure 1: although some data points appear close to each other, for most of the data there is no more than one observation at each value of x . Consequently, the empirical distribution of random variable Y conditioned on such a value for variable X is based on a trivial point mass. How then can we test a model of the conditional distribution of Y for such sparse data?

Regarding ease of use, existing tests for conditional distributions require more mathematical and computational skill than many applied researchers may have to make their implementation generally accepted. Some of these tests require the use of kernel or local polynomial functions with arbitrary smoothing parameters (Zheng, 2000, Fan et al., 2006). Others, such as the Conditional Kolmogorov Test, compare model and distribution functions additionally incorporating the empirical distribution functions of the conditioning set of variables (Andrews, 1997). Transformations to the unit interval have been applied to construct tests for goodness of fit such as the Rincon-Gallardo et al. test for multivariate normality (Rincon-Gallardo et al., 1979). However, their method is also technically difficult and computationally intense in general applications due to procedures involved in the transformation (O'Reilly and Quesenberry, 1973). Additionally, some are dependent on the order of the data being transformed (O'Reilly and Stephens, 1982); therefore, researchers may obtain variant test results if the same data were ordered differently. What is needed for the applied researcher who does not have the mathematical or programming skills to meaningfully implement complex algorithms is a simple pragmatic test. This paper presents a pragmatic

A Simple Goodness-of-fit Test

general goodness-of-fit statistic for continuous conditional models using uncensored data.

In the next section, we introduce the goodness-of-fit test and the rationale behind it. We then evaluate the performance of the goodness-of-fit statistic in Section 3 using two groups of Monte Carlo experiments. The first group of experiments focuses on discerning deviations from correct specification in the distributional family; the second group focuses on discerning deviations in parameterization. We choose these investigations because they represent the two misspecification issues in estimating conditional probability models. In Section 4, we apply the goodness-of-fit test to the 2011 Medical Expenditures Panel Survey (MEPS) dataset, modeling three health care expenditure outcomes as functions of patient characteristics. Finally, in Section 5, we conclude our paper with a summary of the findings and discussions about the applications of the goodness-of-fit test. The Appendix provides the expected value of the statistic and the procedure for the calculation of the degrees of freedom for the test statistics, the data generating process for each Monte Carlo experiment, and the analyses modelling cost data from MEPS.

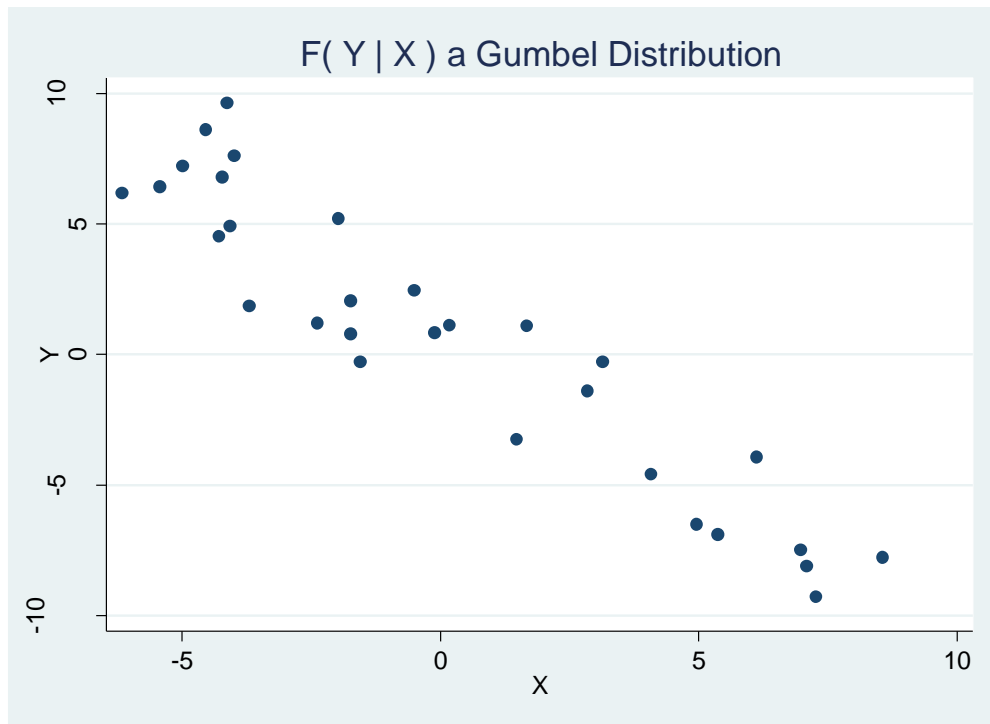


Figure 1. A typical conditional Gumbel distribution with sparse observations for each conditional value of observed X.

2 A proposed goodness-of-fit test

The Pearson Chi-square goodness-of-fit statistic is based on comparing the number of actual observations within each set of a partition of the random variable's range to the number of observations that would be expected to show up in those sets if the model correctly represents the DGP (Schervish, 1995). If the model is correct, then the expected number of observations is the expected number for the DGP; consequently, the observed and predicted number in each set should be different merely by random variation.

The Chi-square goodness-of-fit statistic for continuous distributions is created by partitioning the range of a continuous random variable Y into K regions. Denote each region $k \in \{1, 2, \dots, K\}$ as

$$R_k = \{y: y_{k-1} < y \leq y_k\},$$

the number of observations with values of y in region R_k as N_k , and the total sample size as N . The probability of an observation with y in region R_k is then

$$P_k = \int_{y_{k-1}}^{y_k} f_Y(y) dy,$$

in which $f_Y(y)$ is the probability density for y associated with a cumulative distribution function (CDF) $F_Y(y)$. The Chi-square statistic is defined as

$$C = \sum_k \frac{(N_k - N \cdot P_k)^2}{N \cdot P_k}.$$

The corresponding sample statistic is

$$C_N = \sum_k \frac{(N_k - N \cdot \hat{P}_{k,N})^2}{N \cdot \hat{P}_{k,N}}$$

in which $\hat{P}_{k,N}$ is a consistent estimator of P_k . If $F_Y(y; \theta)$ accurately represents the data generating process, C_N converges to C with increasing N and the corresponding asymptotic distribution of C_N is a Chi-square with degrees of freedom equal to the number of groups in the partition minus the number of estimated parameters plus one (Schervish, 1995).

If we are interested in a model of the conditional distribution $F(y/x; \phi)$, the preceding statistic is not generally applicable because N_k can contain insufficient observations to inform the conditional distribution. Indeed, with x

A Simple Goodness-of-fit Test

containing precisely measured continuous variables, there may be only one value y for some observed x values (see Figure 1 as an example). However, we can take advantage of the probability integral transform and consequent fact that the CDF of a continuous random variable is itself a random variable with a uniform distribution on the unit interval. Because the uniform distribution is the same regardless of underlying CDF, a set of random variables from independent observations with different conditional distributions can all be converted by their CDFs to the same uniform distribution. We can use this fact to construct a test of the conditional distribution; even if each observation has a different conditioning value (i.e. the data in Figure 1 will pose no problem for this test).

Because the CDF for each random variable has a uniform distribution, the CDF values of sample results from a correctly specified model for each random variable will produce a single realization from a uniform distribution. Therefore, the full sample results should together provide a histogram that deviates only by chance from a uniform distribution. We can use a Pearson Chi-square type statistic applied to the uniform distribution to test of the specification for the conditional distributions.

The process is quite simple. For each observation i we have a model specification for the distribution $F(y_i|x_i)$ and therefore can obtain from the estimated model the sample quantity $u_i = F(y_i|x_i)$ for which (x_i, y_i) are the observed values for observation i . The random variable underlying u_i has a uniform distribution on the unit interval if F is correctly specified. We can construct a goodness-of-fit test by partitioning the unit interval into K subintervals defined by equally spaced boundary points, which for $K = 10$ is

$$R_k = \left\{ u: \frac{k-1}{10} < u \leq \frac{k}{10} \right\}, \text{ s.t. } k \in \{1, 2, \dots, 10\}.$$

The statistic is then

$$U = \sum_k \frac{(N_k - N \cdot P_k)^2}{N \cdot P_k}$$

for which N is the total sample size, and N_k is the observed number of u values in the R_k interval. This statistic can alternatively be written as

$$U = N \cdot \sum_k \frac{(\hat{P}_k - P_k)^2}{P_k}.$$

in which \hat{P}_k is the observed proportion in interval R_k . Because the statistic is based on a partition of the uniform into K equal sized intervals, $P_k = 1/K$; therefore,

$$U = K \cdot N \cdot \sum (\hat{P}_k - \frac{1}{K})^2 .$$

As shown in the Appendix, the expected value of U , which is the degrees of freedom for its approximating Chi-square distribution, is equal to the degrees of freedom for the usual Pearson Chi-square test (i.e. $K - 1$) minus a factor due to the estimation of model parameters.

Since P_k is known, which in the case of $K = 10$ intervals is 0.1, we can simply state the statistic for $K = 10$ as

$$U = 10 \cdot N \cdot \sum_k (\hat{P}_k - 0.1)^2 .$$

The selection of $K = 10$ is arbitrary, as it is with the Hosmer-Lemeshow test for logistic regression (Hosmer and Lemeshow, 1980). For other values of K , the degrees of freedom can be directly estimated as shown in the Appendix or determined by Monte Carlo simulation (see Box 2).

The U statistic has a distribution proportional to the sum of gamma random variables with different parameters. Specifically, denoting $\hat{P}_k - \frac{1}{K}$ as z_k , as shown in the appendix z_k is asymptotically normally distributed with mean 0 and variance σ_k^2 . Consequently, the ratio of z_k squared to σ_k^2 has an asymptotic Chi-square distribution with degrees of freedom 1, which is a Gamma distribution with parameters 0.5 and 2 (i.e. $\Gamma(0.5, 2)$). Therefore, z_k^2 has a distribution $\sigma_k^2 \cdot \Gamma(0.5, 2)$, which is $\Gamma(0.5, 2 \cdot \sigma_k^2)$. U is therefore proportional to the sum of K differently scaled gamma random variables. Moschopoulos shows that sum of such variates can be express as a gamma series in which the series coefficients can be recursively determined (Moschopoulos, 1985). The use of this recursive coefficient determination and gamma series is overly complex for the practical application of this statistic among many applied researchers. However, ease of use is the purpose of this goodness-of-fit statistic. Fortunately, the Monte Carlo experiments presented below indicate that for a correct specification the statistic is approximately Chi-square in distribution with degrees of freedom 7.5 when $K = 10$ and calculated as shown in the Appendix or as shown in Box 2 if K is not 10.

3 Simulation experiments

3.1 Methods

We investigated finite sample performance of the proposed statistic using Monte Carlo experiments of conditional Normal, Gumbel, Gamma, and Weibull models, each applied to data generating processes based on the same set of distributions. The first set of experiments comprised a total of sixteen model/DGP comparisons. We evaluated each model/DGP pair for sample sizes 500 and 1000, each using 2000 Monte Carlo samples from the DGP (see Appendix Table A1 for parameter specifications). We inspected rejection rates for significance levels spanning between 0 and 0.2 for each comparison. For each correct model/DGP pair (i.e. Normal/Normal, Gumbel/Gumbel, Gamma/Gamma, and Weibull/Weibull), the plot of the empirical cumulative distribution function (eCDF) of the calculated p values, across the 2000 MC samples, should approximately match the significance level (i.e. this plot should be approximately a straight line). For example, the use of a significance level of 0.01 should reject the model for approximately 1 per cent of the 2000 samples; using a significant level of 0.05 should reject approximately 5 per cent of the samples; and a 0.1 significance level should result in approximately 10 per cent rejections. For mismatched pairs (e.g. Weibull/Gumbel), if the fit test is useful it should produce rejection rates that are higher than the significance levels and increase with sample size; consequently, the eCDF of the test's p-value should be above the significance level.

The second set of experiments compared models in which parameters are specified as linear in conditioning variables to the DGP having the same distributional family but with parameters quadratic in the conditioning variables (see Appendix Table A2 for parameter specifications). For the normal distribution, we estimated models with homoscedasticity and heteroscedasticity. In the case of heteroscedasticity both the mean and variance were generated as quadratic in X in the DGP, but they were modeled as linear in the misspecified model. Similarly, we carried out experiments for sample sizes 500 and 1000. These experiments provide evidence regarding whether the test can identify deviations in parameterization as well as distributional family. In the Monte Carlo experiments reported below, we applied the steps presented in Box 1 to obtain p-values for each of 2000 data sets generated for each model/DGP being considered. We calculated both a p-value using degrees of freedom equal to 7.5 and also using the mean of the 2000 calculated U values for each DGP considered when using the correct model (remember that the degrees of freedom are associated with the distribution of U given the model is correct).

3.2 Results

Because we tested continuous conditional distributions, it is difficult to see the differences between the model and DGP for all patterns of explanatory variables. However, Table 1 shows the probability density functions for the true DGP (in the solid line) and the estimation model (the dotted line, using the average parameter values across the 2000 estimated models) evaluated at the mean of X . This gives some sense of the differences between the distributions being tested in the first set of experiments; however, the deviation of the model from the underlying distribution that drives larger values of U may be from other regions of the conditioning set than at the mean of X .

BOX 1. How to calculate U and its p-value using $K = 10$

Step 1. For a candidate model $F(y_i | x_i ; \theta)$, estimate the parameters, obtaining $\hat{\theta}$.

Step 2. Calculate the CDF value of $u_i = F(y_i | x_i ; \hat{\theta})$ for each observation (y_i, x_i) in the data.

Step 3. Calculate the proportion (\hat{P}_k) of u_i in each of the ten intervals R_k for $k \in \{1, 2, \dots, 10\}$.

Step 4. Calculate the statistic U using the equation

$$U = 10 * N \cdot \sum_k (\hat{P}_k - 0.1)^2 .$$

Step 5. Calculate the p-value as the upper tail area of a Chi-square distribution with degrees of freedom set to 7.5 or set to the estimated value determined by the equations presented in the Appendix or the model-specific Monte Carlo determined empirical degrees of freedom (see Box 2 for the algorithm).

Tables 2 and 3 present the eCDFs of the statistic's p-values for each indicated model applied to the indicated DGP plotted for significance levels up to 0.2. Table 2 presents results for sample sizes of 500; Table 3 presents results for sample sizes of 1000. The thin straight lines show the points where the eCDFs would be if it corresponded to the significance level. The thick dark lines (or curves) are the eCDFs associated with p values based on degrees of freedom set to 7.5. The thick light lines (or curves) are the eCDFs associated with the Monte Carlo based empirical degrees of freedom. We determined the 7.5 degrees of freedom approximation by the average of the four empirical degrees of freedom across the DGPs using sample sizes of 1000. We also ran Monte Carlo experiments for correct model specifications using 10 correlated explanatory variables (results not presented); these

A Simple Goodness-of-fit Test

experiments showed that the empirical degrees of freedom remained around 7.5 in multivariable models. Specifically, the means of the U statistics, and therefore the degrees of freedom, in these experiments for the Normal, Gamma, Weibull, and Gumbel were 7.52, 7.32, 7.54, and 7.27 respectively.

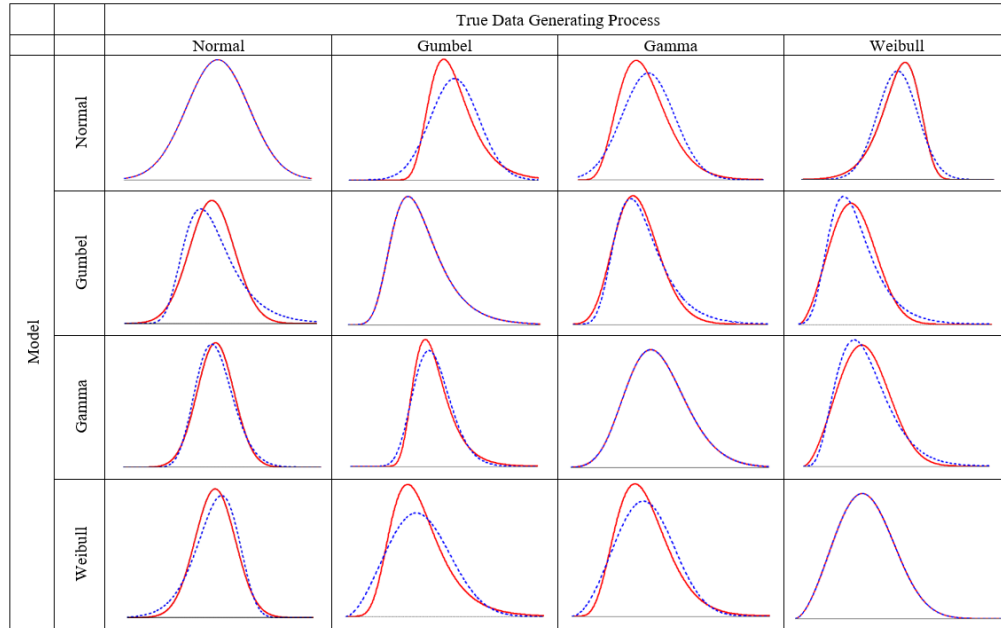


Table 1. Probability density functions of the true data generating process (solid curve) and the estimated model (dashed curve) evaluated at $X=0$ for the Monte Carlo simulations.

The figures on the diagonals of Tables 2 and 3 show the coverage of the test when the model is correctly specified. The results fell along the line representing accurate coverage: the eCDF corresponds to the significance level. Not surprisingly, the empirical degrees of freedom (the thick lighter line) were more accurate than using the approximate degrees of freedom of 7.5; however, the differences were slight, particularly up to the 0.1 significance level.

The off-diagonal figures in Tables 2 and 3 show the rejection rate for the test of misspecified models across significance levels. The test was sufficiently powerful for some of the model/DGP combinations to reject the model for all 2000 samples at all significance levels greater than 0.001. Results for these combinations are simply indicated by the phrase ‘ALL DATA SETS REJECTED AT SIGNIFICANCE LEVEL 0.001’. Not surprisingly, comparing Table 2 to Table 3, the curve has a greater departure from the straight line in Table 3; it is evident that the power of the test increases with sample size. It is also clear that using the approximate 7.5

degrees of freedom provide similar results to that of using the Monte Carlo determined empirical degrees of freedom.

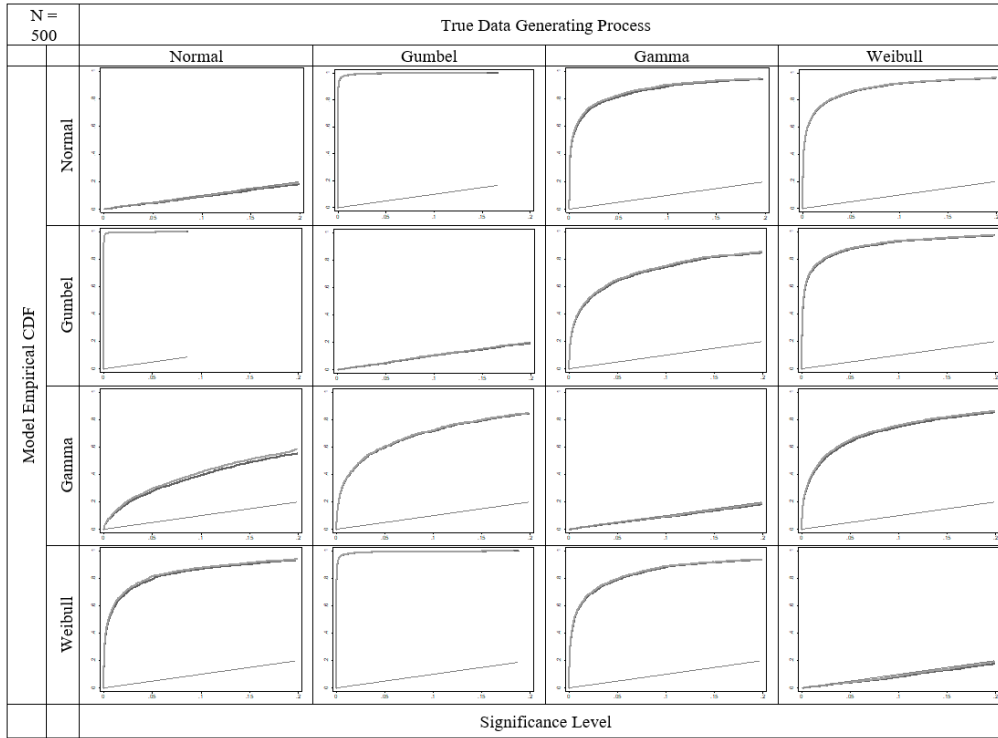


Table 2. Monte Carlo Simulation: Empirical CDFs of Experiments on Distribution Specifications (N=500).

Table 4 presents results for the second set of experiments, which tested deviations from correct specification in the parameterization. The upper two rows show results for sample sizes of 500; the lower two rows show results for sample sizes of 1000. Similar to the first set of experiments, results showed accurate coverage for the test when the model was correctly specified and the ability to discern deviations from correct specification in parameterization. As the sample size went up, the power of the test to discern such deviations increased. The approximate 7.5 degrees of freedom yields results that were similar to the Monte Carlo calculated empirical degrees of freedom.

4 Example

To present an example with real data, we used a random sample of 2000 individuals from the Household Component of the 2011 Medical Expenditure Panel Survey data file (MEPS). As one of the largest national health survey,

A Simple Goodness-of-fit Test

MEPS has been widely used to study the patterns of health care access, utilization and expenditures in the United States (Cohen et al., 2009). We modeled each of the three outcomes – annual total health care expenditure, total office-based visits expenditure, and total dental care expenditure – as a function of individual demographics, socioeconomic status, self-rated health status, common chronic conditions, presence of usual source of care provider, and health insurance coverage. These covariates were selected in accordance with prior studies focusing on modeling health care costs using MEPS survey data (Fenton et al., 2012, Fleishman and Cohen, 2010).

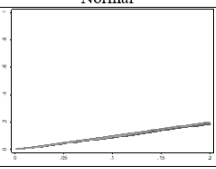
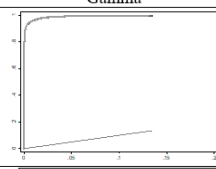
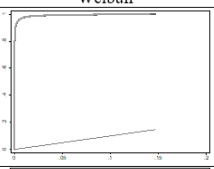
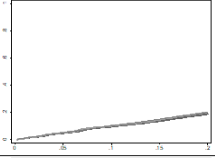
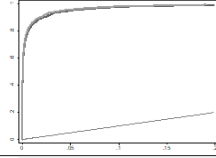
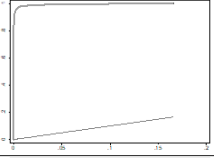
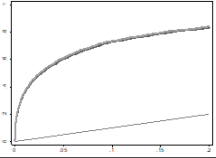
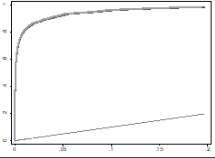
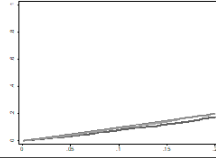
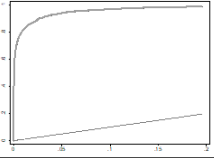
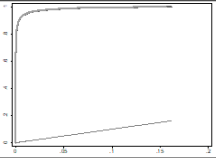
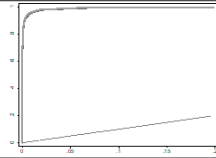
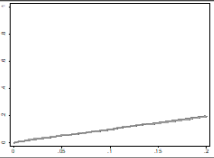
N = 1000		True Data Generating Process			
		Normal	Gumbel	Gamma	Weibull
Model Empirical CDF	Normal		ALL DATA SETS REJECTED AT SIGNIFICANCE LEVEL 0.001		
	Gumbel	ALL DATA SETS REJECTED AT SIGNIFICANCE LEVEL 0.001			
	Gamma				
	Weibull		ALL DATA SETS REJECTED AT SIGNIFICANCE LEVEL 0.001		
		Significance Level			

Table 3. Monte Carlo Simulation: Empirical CDFs of Experiments on Distribution Specifications (N=1000).

For each model, we included all individuals who reported an expense on the outcome of interest and took the log of the expenditure as the dependent variable. There were 1527 and 1215 individuals reporting expenses on health care and office-based services, which represented 76.4% and 60.8% of the total sample, respectively. Much fewer individuals reported any expenses on dental care (N= 724, 36.2%). Appendix Table A3 presents the descriptive statistics and distribution of the outcome variables and the covariates that we employed in the model.

We used Pregibon’s link test (Pregibon, 1980) to identify a statistically

adequate specification of the explanatory variables for each model. We then computed U to test the hypothesis that the specified distribution was correct. This allows us to use the test to focus on testing deviations in the distributional family. We calculated the p-value based on the approximate degrees of freedom of 7.5 and the empirical degrees of freedom calculated from the parameter estimates of the specified model, based on 500 Monte Carlo samples. The algorithm for computing the empirical degrees of freedom is shown in Box 2. Table 5 presents the results from the empirical example for the three health care expenditure outcomes. The test clearly discerns the goodness-of-fit performance of different distributions. Results for the model of the logarithm of total health care expenditure strongly rejected the hypothesis

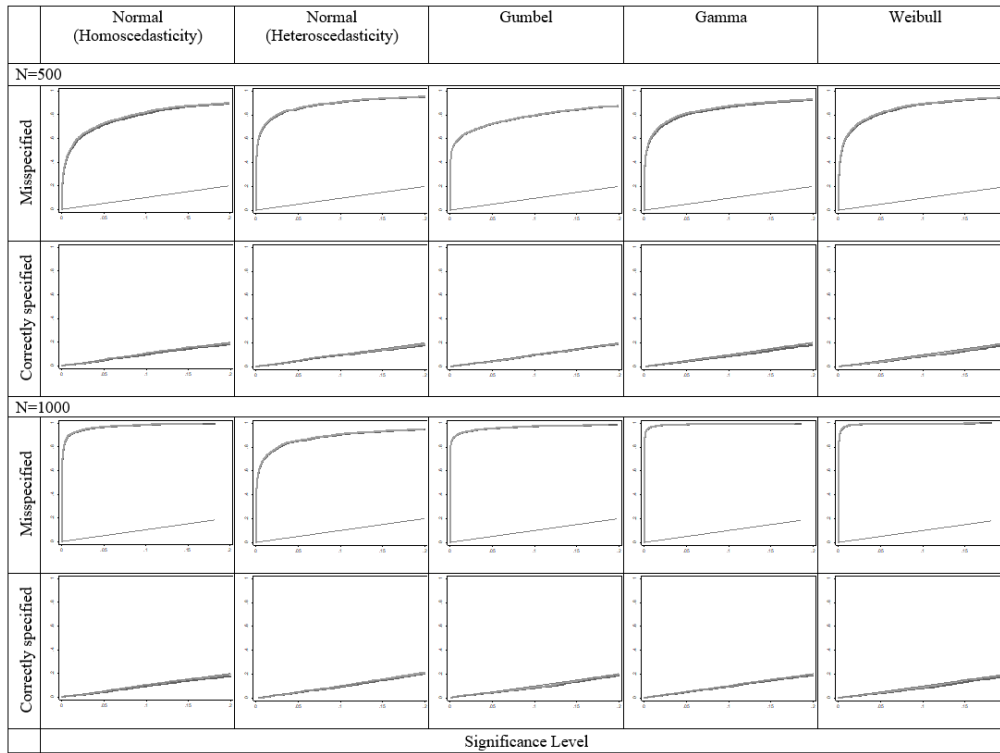


Table 4. Monte Carlo Simulation: Empirical CDFs of Experiments on Parameter Specifications.

that the conditional distribution follows a Gamma, Weibull or Gumbel distribution (U ranges from 21.985 to 116.578, p-value < 0.001 for all), and unequivocally failed to reject the hypothesis for normal ($U = 5.304$, p-value = 0.676 with approximate degrees of freedom of 7.5). For the model of office-based visits expenditure, we strongly rejected the hypotheses for the Gumbel and Weibull distribution (U equals to 59.626 and 33.776, respectively, p-value

A Simple Goodness-of-fit Test

< 0.001 for both) and fail to reject the Normal ($U=12.467$, p -value = 0.107) or Gamma ($U = 8.897$, p -value = 0.305). For the model of dental care expenditure, we rejected all distributions except for the Gumbel ($U = 14.333$, p -value = 0.058 with degrees of freedom of 7.5). Figures A1-A3, in the Appendix, show the histograms of the residuals obtained from these models, standardized by the estimated standard deviations. Figure A1 shows the symmetry expected of a Normal distribution, which was not rejected by the test that unambiguously rejected the other distributions. Figure A2 shows a right-skewedness characteristic of a Gamma distribution (Model 1), but it is insufficiently skewed to reject the Normal at a significance level of 0.05. However, U is smaller in the Gamma indicating a better fit to the data. Under certain circumstances (i.e. shape parameter sufficiently large, >15), the Gamma distribution is approximately a Normal distribution (Rothschild and Logothetis, 1986). In this real-data example, the estimated shape parameter equaled to 35 in the model assuming Gamma distribution. It is therefore not surprising the test did not reject either the Gamma or the Normal distributions. Figure A3 demonstrates the clear right-skewedness of the residual from the model of dental care expenditure, which is expected of a Gumbel distribution. The calculated Monte Carlo empirical degrees of freedoms were approximately 7.5 for all three outcomes and therefore yielded similar results. As there were 21 variables in the empirical model, these results again show that the degrees of freedom for the statistic distribution based on 10 categories is approximately 7.5 in multivariable models.

	<i>Total Health Care Expenditure (N=1527)</i>				<i>Total Office-Based Visits Expenditure (N=1215)</i>				<i>Total Dental Care Expenditure (N=724)</i>			
	U	p-value (DF=7.5)	p-value (DF=edf)	edf	U	p-value (DF=7.5)	p-value (DF=edf)	edf	U	p-value (DF=7.5)	p-value (DF=edf)	edf
<i>Normal</i>	5.30	0.676	0.686	7.60	12.47	0.107	0.099	7.30	89.27	<0.001	<0.001	7.19
<i>Gumbel</i>	116.58	<0.001	<0.001	7.79	59.63	<0.001	<0.001	7.71	14.33	0.058	0.054	7.36
<i>Gamma</i>	23.76	0.002	0.002	7.40	8.90	0.305	0.307	7.53	43.36	<0.001	<0.001	7.38
<i>Weibull</i>	21.99	0.004	0.004	7.59	33.78	<0.001	<0.001	7.41	90.94	<0.001	<0.001	7.57

Note. DF, degrees of freedom; edf, empirical degrees of freedom.

Table 5. Empirical Example: Goodness-of-Fit Tests on Conditional Probability Models for Log-Transformed Health Expenditures from MEPS.

5 Conclusion

In this paper, we presented a simple specification test for conditional continuous distributions using uncensored data (see Box 1). We showed, using simulation experiments, that the test has accurate coverage under correct specification, and that the test can discern deviations from correct specification in both the distributional family as well as parameterization. The empirical example shows its ability to distinguish specific distributions from other candidates using real data.

The results of our analysis indicate that U is approximately distributed Chi-square with degrees of freedom 7.5. We also provide a Monte Carlo method for an empirical determination of degrees of freedom in Box 2 and a direct estimator in the Appendix should the researcher not wish to use the approximating 7.5, for example when the p-value using the approximating 7.5 degrees of freedom is close to the test's designated significance level. However, comparing the empirical degrees of freedom to 7.5 across all Monte Carlo experiments and real-data analyses of our study, the differences were slight and not likely to impact inferences. If a researcher does not wish to approximate the distribution using a Chi-square, a p-value based on the Monte Carlo distribution of statistic values generated in the process of Box 2 can be used as a parametric bootstrap test (Davison et al., 2003).

Because the test discerns deviations in parameterization as well as the distributional family, an extra step is required to investigate the distributional

BOX 2. How to calculate the empirical degrees of freedom

- Step 1. Obtain the parameter estimates predicted from the estimated model ($\hat{\theta}$).
- Step 2. Generate outcome values as random draws from the distribution defined by the estimated parameters $\hat{y}_i \sim F(Y | X = x_i; \hat{\theta})$ for all x_i in the data.
- Step 3. Re-estimate the model using the generated outcomes.
- Step 4. Obtain the predicted parameter estimates ($\tilde{\theta}$) from using the 'correctly' specified model in Step 3.
- Step 5. Calculate the value of $\hat{u}_i = F(\hat{Y}_i | X_i; \tilde{\theta})$ for each observation.
- Step 6. Calculate U .
- Step 7. Repeat the steps 2 through 6 multiple times (e.g. we repeated 500 times in the empirical example), saving the statistic values.
- Step 8. Set the degrees of freedom to the mean of the calculated U values.

family alone. Specifically, the researcher should engage in standard tests to identify the best parameter specification within each proposed model (e.g. we used Pregibon's link test in the preceding example). Using the best within-family model specification, the test will then primarily be identifying deviations in the distributional family.

It is important to note that our results using multiple explanatory variables in the models indicate the degrees of freedom for the statistic's distribution is not a function of the number of estimated parameters. This is different from

A Simple Goodness-of-fit Test

the direct application of the Pearson Chi-square test to distributions with multiple parameters in which the degrees of freedom depend on the number of parameters m . This is an advantage since the degrees of freedom in the latter case is typically $K-m-1$, which implies m must be less than $K-1$ for those applications (Schervish, 1995): our test does not have this constraint.

Although our test can be used as a goodness-of-fit test for marginal distributions, it is particularly useful as an easy-to-use model fit test of continuous conditional distributions for uncensored data, particularly in the case of few observations, indeed even one observation per pattern of explanatory variables, such as a time-series.

References

- AMEMIYA, T. 1985. *Advanced Econometrics*, Cambridge, MA, Harvard University Press.
- ANDREWS, D. W. K. 1997. A Conditional Kolmogorov Test. *Econometrica*, 65, 1097-1128.
- COHEN, J. W., COHEN, S. B. & BANTHIN, J. S. 2009. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. *Med. Care.*, 47, S44-50.
- DAVISON, A. C., HINKLEY, D. V. & YOUNG, G. A. 2003. Recent developments in bootstrap methodology. *Statistical Science*, 18, 141-157.
- FAN, Y. Q., LI, Q. & MIN, I. 2006. A Nonparametric bootstrap test of conditional distributions. *Economet. Theor.*, 22, 587-613.
- FENTON, J. J., JERANT, A. F., BERTAKIS, K. D. & FRANKS, P. 2012. The cost of satisfaction: a national study of patient satisfaction, health care utilization, expenditures, and mortality. *Arch. Intern. Med.*, 172, 405-11.
- FLEISHMAN, J. A. & COHEN, J. W. 2010. Using Information on Clinical Conditions to Predict High-Cost Patients. *Health. Serv. Res.*, 45, 532-552.
- HOSMER, D. W. & LEMESHOW, S. 1980. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics Part A-Theory and Methods*, 10, 1043-1069.
- MOSCHOPOULOS, P. G. 1985. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37, 541-544.
- O'REILLY, F. J. & QUESENBERY, C. P. 1973. The Conditional Probability Integral Transformation and Applications to Obtain Composite Chi-Square Goodness-of-Fit Tests. *Ann. Statist.*, 1, 74-83.
- O'REILLY, F. J. & STEPHENS, M. A. 1982. Characterizations and Goodness of Fit Tests. *J. Roy. Statist. Soc. Ser. B*, 44, 353-360.
- PREGIBON, D. 1980. Goodness of Link Tests for Generalized Linear Models. *J. Roy. Statist. Soc. Ser. C*, 29, 15-24.
- RINCON-GALLARDO, S., QUESENBERY, C. P. & O'REILLY, F. J. 1979. Conditional Probability Integral Transformations and Goodness-of-Fit Tests for Multivariate Normal Distributions. *Ann. Statist.*, 7, 1052-1057.
- ROTHSCHILD, V. & LOGOTHETIS, N. 1986. *Probability distributions*, Wiley.
- SCHERVISH, M. J. 1995. *Theory of Statistics*, New York, Springer-Verlag.
- ZHENG, J. X. 2000. A consistent test of conditional parametric distributions *Economet. Theor.*, 16, 667-691.

Appendix

A1 The Expected Value of the U-Statistic

The expected value of U is the expected value associated with the distribution of the standard Pearson Chi-square goodness-of-fit statistic minus a factor due to estimating the parameters of the model. In this appendix we provide the determination of the expected value, and we provide an estimator for the adjustment factor and thereby an estimator of the expected value for the proposed statistic.

The expected value of U is proportional to the sum of expected values across the K equal-length regions of the partition of the unit interval being considered:

$$\begin{aligned} E(U) &= E[K \cdot N \cdot \sum_k (\hat{P}_k - P_k)^2] \\ &= K \cdot N \cdot \sum_k E[(\hat{P}_k - P_k)^2] \end{aligned}$$

The expected values under the summation sign on the right-hand side of this equation are variances. This is seen by denoting an indicator of whether observation i falls in region k as

$$I_{k,i} = \begin{cases} 1 & u_i \in ((k-1) \cdot 0.1, k \cdot 0.1) \\ 0 & \text{Otherwise} \end{cases}$$

and noting that the expected value of the estimated proportion in category k is

$$\begin{aligned} E[\hat{P}_k] &= \int E[\hat{P}_k | \hat{\theta}] dF(\hat{\theta}) \\ &= \int E\left[\frac{1}{N} \sum_{i=1}^N I_{k,i} | \hat{\theta}\right] dF(\hat{\theta}) \\ &= \frac{1}{N} \cdot \int \sum_{i=1}^N E[I_{k,i} | \hat{\theta}] dF(\hat{\theta}) \\ &= \frac{1}{N} \cdot \int \sum_{i=1}^N P_{k,i}(\hat{\theta}) dF(\hat{\theta}) \\ &= \int P_k(\hat{\theta}) dF(\hat{\theta}), \text{ for } P_{k,i} = P_k \text{ for all } i \\ &= E(P_k(\hat{\theta})) \end{aligned}$$

To determine $E(P_k(\hat{\theta}))$, consider a first order Taylor series approximation around the true value θ

$$P_k(\hat{\theta}) = P_k(\theta) + \frac{\partial P_k}{\partial \theta'} \cdot (\hat{\theta} - \theta),$$

which yields

$$\sqrt{N} \cdot (P_k(\hat{\theta}) - P_k(\theta)) = \frac{\partial P_k}{\partial \theta'} \cdot \sqrt{N} \cdot (\hat{\theta} - \theta).$$

For an estimator, such as the maximum likelihood estimator, for which $\sqrt{N} \cdot (\hat{\theta} - \theta)$ converges to a normal distribution $N(0, \Sigma)$ by a central limit theorem, the left-hand side converges in distribution to a normal as well:

$$\sqrt{N} \cdot (P_k(\hat{\theta}) - P_k(\theta)) \xrightarrow{d} N\left(0, \frac{\partial P_k}{\partial \theta'} \Sigma \frac{\partial P_k}{\partial \theta}\right).$$

Therefore, $P_k(\hat{\theta})$ has an asymptotic distribution with expected value of $E[P_k(\hat{\theta})] = P_k(\theta)$ and variance of $V[P_k(\hat{\theta})] = \frac{1}{N} \cdot \left(\frac{\partial P_k}{\partial \theta'} \Sigma \frac{\partial P_k}{\partial \theta} \right)$. Consequently, since $E[P_k(\hat{\theta})] = P_k(\theta)$,

$$E[(\hat{P}_k - P_k)^2] = V[\hat{P}_k].$$

The expected value of the U is then proportional to the sum of variances:

$$E[K \cdot N \cdot \sum_k (\hat{P}_k - P_k)^2] = K \cdot N \cdot \sum_k V[\hat{P}_k].$$

The variance terms under the summation sign on the right-hand side are

$$\begin{aligned} V[\hat{P}_k] &= \int V[\hat{P}_k | \hat{\theta}] dF(\hat{\theta}) \\ &= \int V\left[\frac{1}{N} \cdot \sum_{i=1}^N I_{k,i} | \hat{\theta}\right] dF(\hat{\theta}) \\ &= \frac{1}{N^2} \cdot \int \sum_{i=1}^N V[I_{k,i} | \hat{\theta}] dF(\hat{\theta}), \text{ for independent observations} \\ &= \frac{1}{N} \cdot \int P_k(\hat{\theta})(1 - P_k(\hat{\theta})) dF(\hat{\theta}) \\ &= \frac{1}{N} \cdot \int (P_k(\hat{\theta}) - P_k(\hat{\theta})^2) dF(\hat{\theta}) \\ &= \frac{1}{N} \cdot [E(P_k(\hat{\theta})) - E(P_k(\hat{\theta}))^2 - V(P_k(\hat{\theta}))] \\ &= \frac{1}{N} \cdot [P_k(\theta) - P_k(\theta)^2 - V(P_k(\hat{\theta}))] \end{aligned}$$

Therefore,

A Simple Goodness-of-fit Test

$$\begin{aligned}
 E[K \cdot N \cdot \sum_k (\hat{P}_k - P_k)^2] &= K \cdot N \cdot \sum_k \frac{1}{N} \cdot [(P_k - P_k^2) - V(P_k(\hat{\theta}))] \\
 &= K \cdot N \cdot \sum_k \frac{1}{N} \cdot \left[\left(\frac{1}{K} - \frac{1}{K^2} \right) - V(P_k(\hat{\theta})) \right] \\
 &= (K - 1) - K \cdot \sum_k V(P_k(\hat{\theta}))
 \end{aligned}$$

The expected value of U is the degrees of freedom for a common Pearson Chi-Square test statistic (i.e. $K - 1$) minus a factor due to estimation of the distribution parameters. For $K = 10$, the expected value of U is then $9 - 10 \cdot \sum_k V(P_k(\hat{\theta}))$.

A2 Estimation of the Shrinkage Factor

The variance terms in the shrinkage factor can be estimated by using consistent estimators for the derivatives $\frac{\partial P_k}{\partial \theta}$ and the covariance matrix Σ . The derivative of P_k is determined by noting that

$$P_k = \int [F(y_k^*(x) | x; \theta) - F(y_{k-1}^*(x) | x; \theta)] dF_x(x),$$

for which y_k^* are the critical values

$$y_k^*(x) = F^{-1}\left(\frac{k}{K} | x\right).$$

Therefore, assuming we can interchange the order of integration and differentiation,

$$\frac{\partial P_k}{\partial \theta} = \int \frac{\partial}{\partial \theta} F(y_k^*(x) | x; \theta) dF_x(x) - \int \frac{\partial}{\partial \theta} F(y_{k-1}^*(x) | x; \theta) dF_x(x).$$

Estimating the integrals on the right-hand side of the equation by sample means yields the estimator

$$\frac{\partial P_k}{\partial \theta} = \frac{1}{N} \cdot \sum_{i=1}^N \frac{\partial}{\partial \theta} F(y_k^*(x_i) | x_i; \hat{\theta}) - \frac{1}{N} \cdot \sum_{i=1}^N \frac{\partial}{\partial \theta} F(y_{k-1}^*(x_i) | x_i; \hat{\theta}).$$

The estimator for the variances in the shrinkage factor is therefore

$$\hat{V}[P_k(\hat{\theta})] = \frac{1}{N} \cdot \left(\frac{\partial P_k}{\partial \theta'} \cdot \hat{\Sigma} \cdot \frac{\partial P_k}{\partial \theta} \right).$$

For the maximum likelihood estimator, note that the scaled deviation of the estimator converges in distribution to a normal:

$$\sqrt{N} \cdot (\hat{\theta} - \theta) \xrightarrow{d} N(0, [-E(\frac{1}{N} H(\theta))]^{-1}),$$

for H denoting the matrix of second derivatives of the log-likelihood with respect to the parameters. Therefore,

$$\begin{aligned} \Sigma &= [-E(\frac{1}{N} H(\theta))]^{-1} \\ &= N \cdot [-E(H(\theta))]^{-1} \end{aligned}$$

Using the sample mean for the expectation of the Hessian, evaluated at the estimated parameter values, yields the estimator

$$\hat{\Sigma} = N^2 \cdot [-H(\hat{\theta})]^{-1}.$$

The estimated variance of $P_k(\hat{\theta})$ is then

$$\hat{V}[P_k(\hat{\theta})] = N \cdot \left(\frac{\partial P_k}{\partial \theta'} \cdot [-H(\hat{\theta})]^{-1} \cdot \frac{\partial P_k}{\partial \theta} \right).$$

For example, consider the Weibull distribution specified in Table A1. The Weibull CDF is

$$F(y | x) = 1 - e^{-(e^{a_0+a_1 \cdot x}) \cdot y^{b_0+b_1 \cdot x}}.$$

The derivatives with respect to the parameters are

$$\begin{aligned} \frac{\partial F}{\partial a_0} &= D \\ \frac{\partial F}{\partial a_1} &= D \cdot x \\ \frac{\partial F}{\partial b_0} &= D \cdot (e^{b_0+b_1 \cdot x}) \cdot \ln(y) \\ \frac{\partial F}{\partial b_1} &= D \cdot (e^{b_0+b_1 \cdot x}) \cdot \ln(y) \cdot x \end{aligned}$$

where,

$$D = y^{e^{b_0+b_1 \cdot x}} \cdot e^{-y^{e^{b_0+b_1 \cdot x}} \cdot e^{a_0+a_1 \cdot x}} \cdot e^{a_0+a_1 \cdot x}.$$

A Simple Goodness-of-fit Test

Evaluating each of these derivatives and each observation in the sample $i \in \{1, \dots, N\}$ at the estimated parameter values, data values x_i , and the corresponding critical values $y_0^*(x_i)$, and $y_k^*(x_i)$ for each $k \in \{1, \dots, 10\}$ creates variables for which the sample means can be used to determine $\frac{\partial P_k}{\partial \theta}$. These estimated derivatives combined with the estimated parameter covariance matrix $\hat{\Sigma}$ provide the information to calculate the shrinkage factor as shown above.

Table A0 presents the means of the estimated expected value of U using the above equations and means of the calculated U values across 100,000 data sets of sample sizes 100, 1000, and 10,000. The mean estimated $E(u)$ was very similar to the mean of U values, rounding to 7.37 for each. An alternative for estimating the expected value of U (i.e. degrees of freedom for an approximating Chi square distribution) is the Monte Carlo method shown in Box 2 of the main text.

Sample Size	Mean Estimated E(u)	Mean U-statistic
100	7.367	7.369
1000	7.373	7.367
10000	7.374	7.374

Table A0. Mean estimated $E(u)$ and mean U across 100,000 samples.

A3 Additional Tables and Figures

		True Data Generating Process			
		Normal	Gumbel	Gamma	Weibull
Model Empirical CDF	Normal	$\mu = e^{(2+0.1x)}$ $\sigma^2 = \mu^2$	Location $\mu = 10 + x$ Scale $\beta = e^{(0.1x)}$	Shape $\alpha = e^{(2+0.2x)}$ Scale $\beta = e^{(-0.2+0.1x)}$	Shape $\alpha = e^{(2.5+0.1x)}$ Scale $\beta = e^{(0.1+0.2x)}$
	Gumbel	$\mu = 10 + x$ $\sigma^2 = e^{(0.5+0.1x)}$	Location $\mu = 5 + x$ Scale $\beta = e^{(0.6+0.1x)}$	Shape $\alpha = e^{(3+0.2x)}$ Scale $\beta = e^{(-0.6+0.1x)}$	Shape $\alpha = e^{(1+0.1x)}$ Scale $\beta = e^{(0.1+0.2x)}$
	Gamma	$\mu = e^{(1.03+0.001x)}$ $\sigma^2 = 0.04\mu^2$	Location $\mu = 8 + x$ Scale $\beta = e^{(0.5+0.1x)}$	Shape $\alpha = e^{(3+0.2x)}$ Scale $\beta = e^{(-0.6+0.1x)}$	Shape $\alpha = e^{(1+0.1x)}$ Scale $\beta = e^{(0.2+0.2x)}$
	Weibull	$\mu = 8 + 0.1x$ $\sigma^2 = e^{(0.05+0.1x)}$	Location $\mu = 5 + x$ Scale $\beta = e^{(0.6+0.1x)}$	Shape $\alpha = e^{(2+0.2x)}$ Scale $\beta = e^{(0.1x)}$	Shape $\alpha = e^{(1+0.1x)}$ Scale $\beta = e^{(0.1+0.2x)}$

Table A1. Simulation process: conditional distribution of the data for the test of incorrect distributional family.

True Data Generating Process
<p>Normal/Normal (homoscedasticity)</p> $\mu = 6 + 0.1x + 0.65x^2$ $\sigma^2 = 1$
<p>Normal/Normal (heteroscedasticity)</p> $\mu = 6 + 0.1x + 0.35x^2$ $\sigma^2 = e^{(1+0.2x+0.35x^2)}$
<p>Gumbel/Gumbel</p> <p>Location $\mu = 5 + x + 0.24x^2$</p> <p>Scale $\beta = e^{(0.6+0.1x+0.24x^2)}$</p>
<p>Gamma/Gamma</p> <p>Shape $\alpha = e^{(2+0.2x+0.1x^2)}$</p> <p>Scale $\beta = e^{(0.1x+0.1x^2)}$</p>
<p>Weibull/Weibull</p> <p>Shape $\alpha = e^{(1+0.1x+0.47x^2)}$</p> <p>Scale $\beta = e^{(0.2+0.2x+0.47x^2)}$</p>

Table A2. Simulation process: conditional distribution of the data for the test of incorrect parameterization

A Simple Goodness-of-fit Test

Variables	Sample for positive total health care expenditure (N=1527)	Sample for positive total office-based visits expenditure (N=1215)	Sample for positive total dental care expenditure (N=724)
<i>Cost-related outcome variables*</i>			
Total health care expenditure, median (IQR), \$	947 (291-3359)	---	---
Total office-based visits expenditure, median (IQR), \$	---	371 (145-1053)	---
Total dental care expenditure, median(IQR), \$	---	---	225 (113-501)
<i>Explanatory variables</i>			
Age, median (IQR), y	35 (16-55)	39 (17-58)	33 (14-55)
Female sex	849 (55.60)	686 (56.46)	388 (53.6)
Race/Ethnicity			
White	709 (46.43)	584 (48.07)	384 (53.04)
African American	307 (20.10)	232 (19.09)	114 (15.75)
Hispanic	390 (25.54)	301 (24.77)	170 (23.48)
Other	121 (7.92)	98 (8.07)	56 (7.73)
Education			
<High school	359 (26.03)	265 (24.40)	197 (29.27)
Some high school	107 (7.76)	89 (8.20)	32 (4.75)
High school graduate	349 (25.31)	279 (25.69)	125 (18.57)
Some college	254 (18.42)	207 (19.06)	117 (17.38)
College graduate and above	310 (22.48)	246 (22.65)	202 (30.01)
Self-rated health status			
Excellent	480 (31.54)	365 (30.17)	258 (35.73)
Very good	467 (30.68)	356 (29.42)	230 (31.86)
Good	383 (25.16)	318 (26.28)	168 (23.27)
Fair	148 (9.72)	132 (10.91)	60 (8.31)
Poor	44 (2.89)	39 (3.22)	6 (0.83)
Chronic diseases (≥ 3)	140 (9.17)	129 (10.62)	676 (6.63)
Usual source of care	1235 (82.44)	1037 (86.34)	595 (83.45)
Household income relative to percentage of FPL			
<100	307 (20.10)	237 (19.51)	128 (17.68)
100-124	92 (6.02)	71 (5.84)	33 (4.56)
125-199	251 (16.44)	194 (15.97)	95 (13.12)
200-399	450 (29.47)	359 (29.55)	212 (29.28)
>400	427 (27.96)	354 (29.14)	256 (35.36)
Health insurance coverage			
Private	926 (60.64)	751 (61.81)	473 (65.33)
Public	466 (30.52)	377 (31.03)	215 (29.70)
None	135 (8.84)	87 (7.16)	36 (4.97)

Note. Numbers are number (%) unless otherwise indicated. * The cost variables are log-transformed before entered into the MLE model. IQR, interquartile range. FPL, federal poverty level.

Table A3. Distribution of the cost-related outcome variables and patient characteristics.

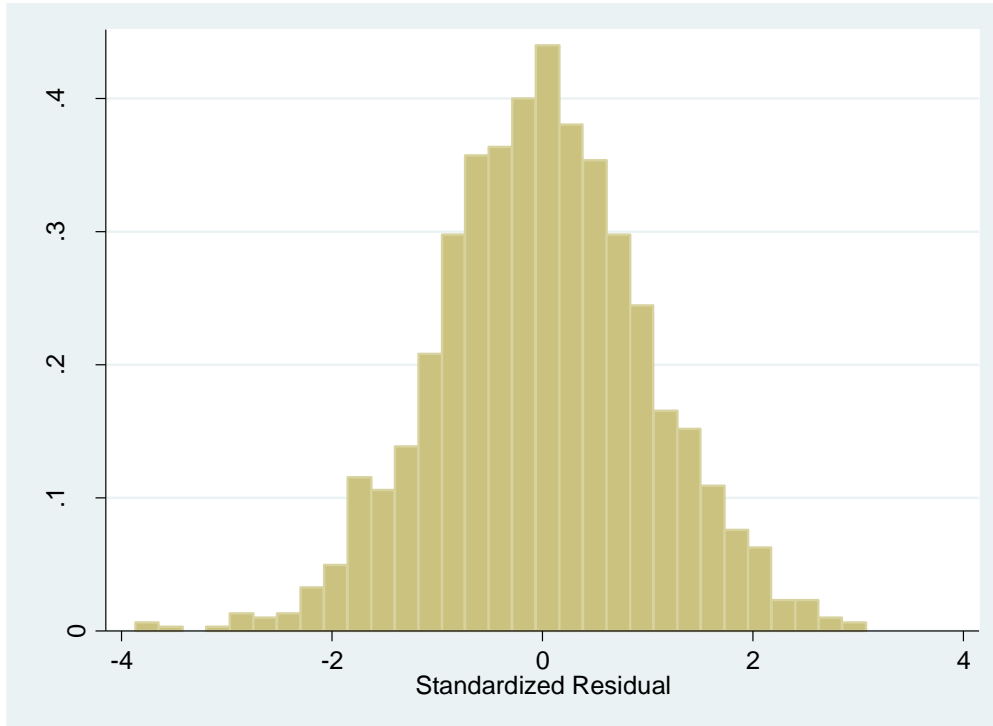
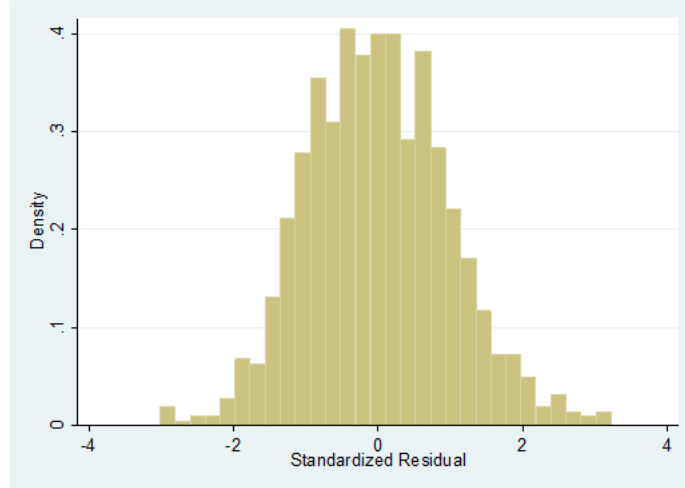


Figure A1. Histogram of the Standardized Residual from the Model for Annual Total Health Care Expenditure. Model: MLE assuming Normal distribution with heteroskedasticity.

A Simple Goodness-of-fit Test

Model 1: MLE assuming Gamma distribution:



Model 2: MLE assuming Normal distribution with heteroskedasticity:

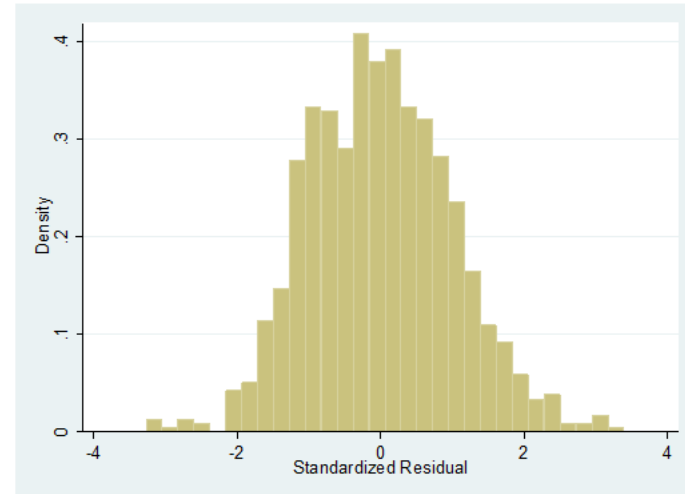


Figure A2. Histogram of the Standardized Residual from the Model for Annual Total Expenditures on Office-Based Visits.

Model: MLE assuming Gumbel distribution:

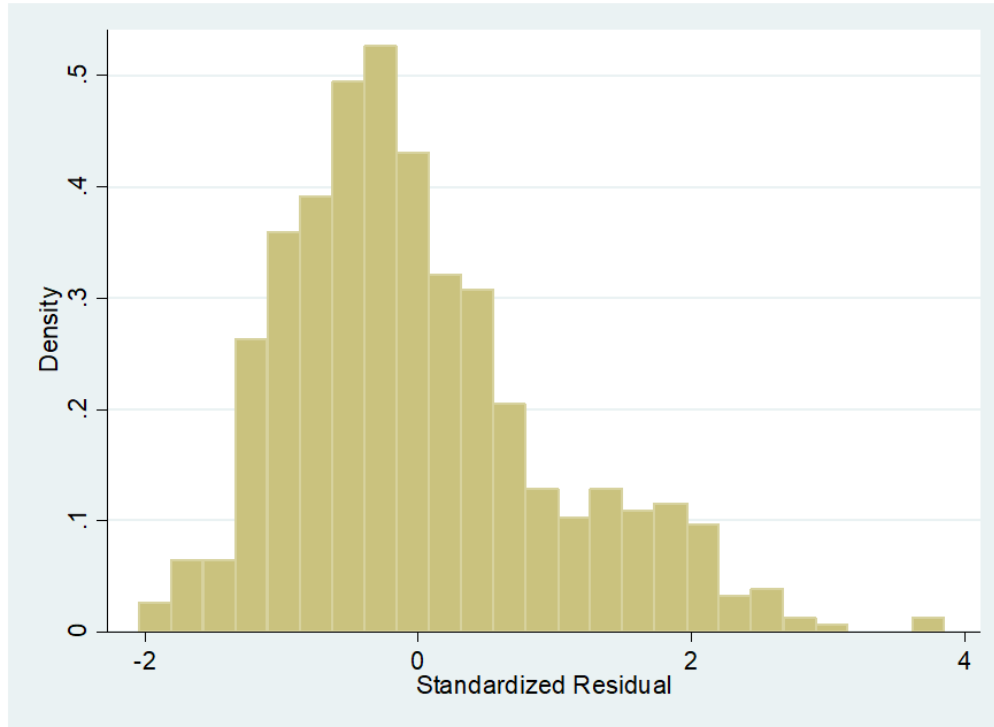


Figure A3. Histogram of the Standardized Residual from the Model for Annual Total Expenditures on Dental Care.