Table of Contents

[CHAPTER I: INTRODUCTION TO THE STUDY 2](#_Toc199669609)

[Research Questions 2](#_Toc199669610)

[Limitations of Study 2](#_Toc199669611)

[CHAPTER II: REVIEW OF LITERATURE 3](#_Toc199669612)

[Introduction to the Problem 3](#_Toc199669613)

[The main issue: Skew 7](#_Toc199669614)

[Mixed Dn 8](#_Toc199669615)

[Heteroskedastic 8](#_Toc199669616)

[Need to address, lack of agreement how 9](#_Toc199669617)

[Lack of documentation 10](#_Toc199669618)

[Why disregard for importance of detecting outliers 14](#_Toc199669619)

[Defining/Handling/Detection 16](#_Toc199669620)

[Defining/Detecting 17](#_Toc199669621)

[Outlier vs Influential Observations 19](#_Toc199669622)

[Addressing / Treating 25](#_Toc199669623)

[Treating the data directly 25](#_Toc199669624)

[Location 26](#_Toc199669625)

[Why OLS is insufficient for the problem 26](#_Toc199669626)

[CHAPTER III: METHODOLOGY 29](#_Toc199669627)

[Introduction 29](#_Toc199669628)

[Empirical Study 29](#_Toc199669629)

[Simulation Plan 34](#_Toc199669630)

[Research Question 1 Analysis Plan 36](#_Toc199669631)

[Research Question 2 Analysis Plan 36](#_Toc199669632)

[Research Question 3 Analysis Plan 36](#_Toc199669633)

[Chapter III Summary 36](#_Toc199669634)

[CHAPTER IV: RESULTS 37](#_Toc199669635)

[Simulation Overview 37](#_Toc199669636)

[Percent Difference Formulas 37](#_Toc199669637)

[Research Question 1 Raw Bias Percentage Results 37](#_Toc199669638)

[Condition X1 Raw Bias Percentage Results 37](#_Toc199669639)

[Condition X2 Raw Bias Percentage Results 37](#_Toc199669640)

[Condition X3 Raw Bias Percentage Results 37](#_Toc199669641)

[Research Question 2 Empirical Standard Errors Results 37](#_Toc199669642)

[Condition X1 37](#_Toc199669643)

[Condition X2 37](#_Toc199669644)

[Condition X3 37](#_Toc199669645)

[Research Question 3 XX Results 37](#_Toc199669646)

[Chapter IV Summary 37](#_Toc199669647)

[LIST OF ABBREVIATIONS 38](#_Toc199669648)

[REFERENCES 38](#_Toc199669649)

# CHAPTER I: INTRODUCTION TO THE STUDY

Text

## Research Questions

The following list of research questions will be answered by comparing the performance of the three transformed models.

|  |  |
| --- | --- |
| Q1 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of raw bias percentage? |
| Q2 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of empirical standard errors? |
| Q3 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of confidence interval coverage? |

## Problem

Policy decisions regarding health care resource allocation rely on accurate, reliable determinations about the costs of medical care and expenditures. Several key characteristics of health care cost and utilization data, however, render accurate estimation of utilization outcomes (e.g. cost, visits, length of stay) challenging: they typically exhibit a large mass at zero, a heavy right-skew, and heteroskedasticity. These characteristics have been observed for decades, running rampant across datasets, contexts, and even fields. Rather than enjoying a unified approach to greet these familiar qualities, the ubiquity and scope of this issue has instead produced a surplus of methods and recommendations. It is not uncommon to find papers that use a combination of data processing methods and modeling techniques to accommodate these characteristics, often with little justification or acknowledgement. Though recent literature has largely endorsed the two-part model for estimation of cost outcomes to accommodate the mass at zero, many authors address the skew by also applying a variety of data processing techniques (such as top-coding), and frequently with little discussion or justification for their data transformation selection. This study will contribute to the literature by examining one of the most common data processing methods, top-coding, and its effects on estimates when used in conjunction with a two-part generalized linear model, following recent papers on this topic (CITE).

## Background of the empirical data

To further investigate the effects of applying a two-part model with generalized linear model in conjunction to top-coding the outcome variable at various values, settings, and sample sizes, the candidate was able to extract the data used in a relevant paper *Modeling Health Care Expenditures and Use* (Deb & Norton, 2018) and follow their modeling recommendations.

## Purpose of the study

The purpose of this study is to model health care expenditures with a common technique (top-coding, tpm with generalized linear model) and examine XXX.

## Brief summary of Chapter 2

Chapter two start with a deep dive into the nature of health care cost data and the source of these data hallmarks, particularly why these data (and those of other fields) give rise to substantial mass at zero and exhibit long, right tails, and why these values must not be excluded from the dataset. The majority of this paper will focus on the handling of the long, right tails (or extreme values); as such, the mass at zero will not be discussed as thoroughly, and addressed using the two-part model. The chapter will continue to explore methods of identifying and handling the right-skew typical of these datasets. First, it will provide an overview of the various data transformation techniques that have been suggested and applied in the literature (e.g. removal, winsorization, and top-coding methods), as well as their inconsistent application and documentation in the literature. Next, this paper will explore the variety of implementation of common transformation functions (e.g. log transformation, inverse hyperbolic sine, cox proportional hazard model, and generalized linear models), as well as discuss the merits and demerits of one-part (single index) and two-part models. Finally, examples will be provided of combinations of methods: implementation of the two-part model accompanied by some direct data transformation of the outcome variable.

## Brief summary of the methodological study

## Significance of the study

This study will contribute to the literature …

## Limitations of Study

* Does not address mass at zero (only follows tpm) or implications of not addressing mass at zero

## Definition of key terminologies

* Right Skew
* Heteroskedasticity
* Truncation/Trimming
* Top-coding
* Transformation, Transformation functions
* Log Transformation

# CHAPTER II: REVIEW OF LITERATURE

The goal of this section is to provide a theoretical overview of transformations for XX studies. The chapter will begin with a discussion on health care expenditure distributions and modeling. In the following section, there will be a review of XX. The chapter will conclude with review on three distinct types of models: 1, 2, and 3.

## Introduction to the Problem

“For decades, “how to contain health care costs?” has been a major recurring theme for health economists and policymakers around the world. Therefore, credible empirical analyses of the right tail of health expenditure distributions harbor great potential to answer some of the most pressing policy questions.” (Karlsson et al., 2024)

Policy decisions regarding allocation of health care resources depend on accurate, reliable determinations about the costs of medical care and expenditures. Health care cost and utilization data, however, often have several key characteristics that make accurate estimation of utilization outcomes (e.g. cost, visits, length of stay) more challenging, rendering them vulnerable to a surplus of methods and recommendations; in fact, many authors use a combination of data processing methods and modeling techniques to accommodate these characteristics. Though recent literature has largely endorsed the two-part model for estimation of cost outcomes, many authors continue to apply data processing techniques (such as top-coding) in combination with these modeling methods, and frequently with little discussion or justification for their data transformation selection. This study will contribute to the literature by examining one of the most common data processing methods, top-coding, and its effects on estimates when used in conjunction with a two-part generalized linear model, following recent papers on this topic (CITE).

Frequently, the main goal of empirical analysis in this area is to obtain reliable estimates of policy parameters like the marginal effects or average marginal effects of *x* on some function of the conditional distribution of *y* (typically *E[y|x]*) and its corresponding marginal effects. Most empirical analyses of healthcare cost data are regression based (analysis considers various features of the distribution of cost outcomes (*y*) conditional on covariates (*x)*, with most instances focusing on the conditional mean *E[y|x]* and the parameters describing it (e.g. *E[y|x]* = *g(b0 + b1x1 + b2x2 +…+ bnxn,* where *g(x)* refers to some functional form).(Mullahy, 2009).

Regression methods form an important part of many empirical statistical analyses of health care data, from adjusting treatment effects in randomised trials to adjusting for case-mix in observational studies to developing models for prediction (Barber & Thompson, 2004). Health care cost and utilization data, however, often have several key characteristics that make accurate estimation of utilization outcomes (e.g. cost, visits, length of stay) more challenging and yet must be accommodated, rendering them vulnerable to a surplus of methods and recommendations (Deb & Norton, 2018; Mihaylova et al., 2011; Mullahy, 1998): they are nonnegative, have a non-trivial mass at zero (often up to 20% of records), exhibit “heavy” upper tails (defined by Mullahy (2009) as exceeding 10x the sample mean), are right skewed, and tend to be heteroskedastic.

These non-negative data are often multimodal, with values clustering around two distinct sections of the distribution: a mass at zero and heavy, right-skewed tails. The substantial point mass at zero on the left side of the distribution represents the non-trivial proportion of the population with outcomes of true zero (i.e. members with no utilization for that observation) (Deb & Norton, 2018; Weichle et al., 2013). These distributions often also display substantial skewness, manifesting in empirical densities with long, thin right tails comprised of a small number of patients that use large amounts of health care resources (Holle et al., 2021; Malehi et al., 2015; Mullahy, 1998; Mullahy & Norton, 2024). In 2019, for example, healthcare spending for the U.S. civilian, noninstitutionalized population surpassed $2 trillion, yet the majority was concentrated in a small percentage of the population (Mullahy, 1998; Polgreen & Brooks, 2012; Weichle et al., 2013). Nearly 14% of the population had no expenditures at all (mass at zero), while 5% accounted for nearly half of all healthcare spending. In 2022, 1% of persons accounted for nearly 21% of total healthcare expenditures (HCE), with the bottom 50% accounting for only 3% (Mitchell, 2020). In a study reviewing national survey data covering a span of 17 years (2001-2018), Holle et al. (2021) found that 50% of annual health care expenditures consistently attributed to 4.2-4.8% of the U.S. population; the small annual fluctuations were consistent with nothing more than statistical noise. They further note that the proportion of nonspenders stayed consistently around 20%, identifying a slight decrease in 2017 and 2018, but overall demonstrating a sustained concentration of zero expenditures.

This severely right-skewed concentration with large mass of true zeros has persisted for decades, and is likely to continue to do so (Berk & Monheit, 2001; Cohen & Uberoi, 2001; French & Kelly, 2016; Holle et al., 2021). This issue has been documented globally: for example, a study in two Barcelona hospitals showed that high cost outliers accounted for 4.8% of admissions and almost 20% of hospital costs (Cots et al., 2000; Pirson et al., 2006). Costs for patients at the extreme ends of the distribution, such as those patients with rare health complications or a host of other factors such as comorbidities, presence of complications, higher severity scores within the diagnosis-related groups [DRG] (Pirson et al., 2006), and high expenditures, do not represent the typical experience and can disproportionately influence statistical point estimates (Polgreen & Brooks, 2012; Weichle et al., 2013). The large mass of values at zero and high extreme values on the right side of the distribution have the potential to exert a disproportionate influence on statisticaly analysis (i.e. high leverage); in a regression analysis, even a small number of cases can contribute to almost all of the predicted response (i.e. leverage of 1.00), reverse the statistical signficance of an analysis in either direction (Sullivan et al., 2021), or change the magnitude of regression coeffiencies and even the direction of coefficient signs (Choi, 2009). In their review of accounting research studies, Leone et al. (2019) write, “The validity of the inferences drawn in virtually every archival accounting research study has the potential to be affected by how influential observations are identified and treated.” Though their work is related to archival accounting research, the sentiment applies across fields. Choi (2009) writes from a political science perspective: “When researchers ignore abnormal observations, especially with respect to dependent variables, their empirical results can be misleading. Unfortunately, the fact remains underappreciated in studies of political science,” particularly in a field where the OLS regression model is often the standard method to fit prediction equations.

We will discuss more below, but it is important to note that these extreme values are neither classical outliers nor measurement errors: as discussed previously in this paper, these heavy tail features in the underlying distribution of the data represent the approximately 5% of heavy users who produce nearly half of spending costs, and simple (though common) methods to handle such as removal or trimming are often insufficient and will affect the reliability of the inferences drawn in the study (Karlsson et al., 2024; Leone et al., 2019). These handling methods will be discussed more thoroughly later in the paper in the Data Transformation section.

Karlsson (2024) warns that for these heavy or long tails, the population variance and higher-order moments (e.g. population skewness, kurtosis) could be very large and even infinite, violating the assumptions of ordinary least squares (OLS) estimation. In heavy-tailed or outlier-prone distributions, predicted regression lines tend to track outlying observations more closely, fitting them at the expense of the rest of the sample data in order to minimize the sum of squared errors, and producing estimated results that are likely to encourage researchers to make incorrect inferences about the data (Choi, 2009).

In a recent paper, Geurts (2022) provides further insight to the impacts that these high-cost values can have on estimates when evaluating health care expenditures. They describe a study similar to this study (and many other contemporary investigations regarding policy and implementation effects): a research agency is asked by health authorities to estimate potential savings realized through some intervention. (This paper will be estimating potential savings realized through a major change to the Affordable Care Act [ACA], described in more detail below). As is common in similar investigations, a difference-in-differences approach was used to analyze changes in mean spending in the treatment versus risk-adjusted mean changes in a control group, where the latter represent the expected outcome had the project not been implemented. They found that savings were estimates with large standard errors, very sensitive to high-cost patients:

Part of the explanation of the large standard errors is the strongly skewed distribution of health care costs with a small fraction of the population accounting for a large share of the total health spending and mean costs well above median cost. Variance is even more concentrated in the high-cost cases, and exacerbated by the presence of large outliers. As a result, even sophisticated models with large sets of control variables are faced with a portion of high costs not well captured by the model. (Geurts et al., 2022)

Choi et al. (2009) also reexamined a published political study and found that by accounting for influential outliers, they could help reconcile otherwise contradictory findings on the democracy- foreign direct investment link the original study identified, providing evidence that erroneous operationalization of the dependent variable (in this case, inflows of foreign direct investment) caused a serious outlier problem that demanded address. They found that only four observations in the dependent variable (*y*-outliers) foreign direct investment data lead to the spurious results: When the team implemented remediating measures to accounted for these extreme values (creating an indicator variable and using robust regression methods rather than OLS), the results showed that, contrary to the original findings, democratic countries did in fact attract more foreign direct investment than authoritarian ones. Note that they did not drop the outliers, as is often done to improve distributional symmetry, as they felt that most published works in their field (using samples of only 40 to 60 countries) would all show a marked change by omitting four or five of the most outlying observations. In their study, they found that results incorporating the remediating actions were similar to results when they removed the outliers, lending further evidence that removal was not the best option.

### The main issue: Skew

For these reasons, Malehi (2015) calls skewed data “the main issue in statistical modeling of healthcare costs”. These issues affect researchers across health care fields: Mullahy (2023) points out that outcomes that are non-negative, highly-skewed, and have a non-trivial fraction to zero (where zeros are true zeros as opposed to missing) are featured heavily in health economics regarding studies of healthcare use (Mullahy, 1998; Mullahy & Norton, 2024), spending on health services (Manning et al., 1987), and a host of heath outcome measures. As one can imagine, researchers in health care are not alone in their endeavors to find best practices to model these data: labor economics, international trade, agricultural economics, (Mullahy & Norton, 2024), political science (Choi, 2009, 2022), accounting (Leone et al., 2019) and many other fields face outcomes with similar measurement properties. Leone (2019) documents the pervasiveness of extreme values in archival accounting studies, due to the wide cross-sectional and intertemporal variation in events captured by financial report and capital market data, as well as accounting rules that produces numbers based on frequently-shifting estimates and assumptions. Choosing an appropriate method for regression analyses of cost data challenges researchers as it must focus on population means while simultaneously accounting for the typically-skewed data distribution (Barber & Thompson, 2004).

### Mixed Distributions

These continuous distributions with positive mass at one or more points, such as in healthcare cost and utilization, are often considered mixed distributions: they’re not quite continuous, nor are they discrete (Blough et al., 1999). Berlotti (2007) points out that many outcomes (*yi)* in empirical analyses are mixed discrete-continuous random variables with two basic statistical features: 1) *yi>=0*, and 2) *yi=0* observed often enough that there are compelling substantive and statistical reasons for special treatment, with the mass at point zero rendering a single index model undesirable. These features not only impose difficulties applying standard statistical analysis (e.g. linear regression), but doing so would cause unreliable results (Malehi et al., 2015).

### Heteroskedastic

A third and final characteristic of note inherent in these data are that they are frequent violators of the assumption of homoscedasticity (constant variance): expense data often exhibit variability that increase as the mean expense increases (Blough & Ramsey, 2000; Deb & Norton, 2018).

### Need to address, lack of agreement how

The accommodation of these distinct features in the design of any econometric healthcare cost analysis is of great interest as many of the datasets that would be the object of such analysis provide outcomes sharing these characteristics (Mullahy, 2009). These key feature characteristics of healthcare data must be addressed (or at least implications / handling considered) in order to develop a valid model, particularly for researchers interested in estimating incremental costs associated with any specific treatment, condition, or patient or provider characteristic, as is so often the focus of any policy paper (Bhattarai, 2013; Blough & Ramsey, 2000; Mihaylova et al., 2011; Polgreen & Brooks, 2012). Perhaps due to the persistence of the issue in so many areas or growth in options available to researchers as access to knowledge and computational advancements are made, there is a wide body of literature regarding the specifications for models with these features, with the field seeming to collect around common approaches to the large mass at zero, but very little agreement on how to handle, identify, or address the extreme values causing this ubiquitous right skew. Mullahy (2023) points out that the specification of regression models for nonnegative outcomes *y* has been a familiar and longstanding concern in empirical economics, and while they study two widely-used classes of models that transform the dependent variable *y*, they are quick to note that the alternatives they posit aren’t panaceas in all settings.

In fact, it appears the only things literature has agreed on regarding this issue of extreme right skew is the lack of a primary approach that provides both unbiased and efficient estimates (Manning et al., 2005; Sullivan et al., 2021) and the need to perform checks and validations to find a suitable model consistent with the particular scientific, policy, or decision question/s at hand (Bhattarai, 2013; Buntin & Zaslavsky, 2004; Mullahy, 2009; Sullivan et al., 2021). As Manning et al. (2009) note, “No current method is optimal or dominant for all cost applications.” Bhattarai (2013) concluded that no single approach of outlier handling works for every situation, though they did point out that the best model they evaluated was the one that did not remove outliers. Bhattarai strongly recommends that analysts try multiple models guided by economic theory and the underlying distribution of the expenditure data to see which model may serve best. In an attempt to illustrate how accounting researchers identify and treat influential observations, Leone et al. (2019) reviewed 1833 studies published from 2006 to 2014 across various elite journals[[1]](#footnote-1). Their findings are similar, writing: “We stress that there is no universally correct (i.e. best) solution when it comes to identifying and treating influential observations in every possible setting. All approaches invariable require judgment and care on researchers’ part.” Malehi et al.’s 2015 paper compared alternative estimators, such as OLS for ln(*y*) or log(*y*), when studying skewed healthcare cost data, and found there was not one bet model across all generated conditions. The authors did find that generalized linear models (especially the Gamma regression model) behaved well in the estimation of population means of healthcare costs, as did Deb (2018).

### Lack of documentation

The theme of making theoretical decisions (unique to the study, etc) to complement algorithmic processing / findings comes up in many papers and places, of course, but when it comes to handling extreme values, many applied papers either don’t address it at all or mention it but don’t provide theoretical reasoning or citations to justify their decisions. In 1999, Diehr identified that multiple real-world complexities in research can indeed generate inconsistent findings in research, and many may be reasonable: “Because the types of populations studied, the predictor variables, the number of observations, the statistical model, and the measures of ‘goodness’ differed among … studies, there are many possible explanations for discrepancies in findings.” And yet, despite the ubiquity of the existence of extreme tails and while recognizing that approaches will indeed vary due to several important reasons, there is little consensus on approaches for identification, handling, or even documentation of the issue.

In their 2021 analysis, *So many ways for assessing outliers: What really works and does it matter?*, Sullivan et al. set out to identify techniques and methods currently used by academic researchers in top business journals to identify or detect outliers by for best practices for handling extreme values, surveying nearly 3000 business journals during years 2007 to 2018. They found a “surprisingly diverse array of methods used to discover and treat outliers”, noting widely varying – or even more concerning, non-existent – approaches to outlier detection (Sullivan et al., 2021). They note that recent research has varied widely in how statistical outliers are identified and handled, and though many techniques are reported, most articles with empirical data don’t mention outliers, while others simply refer to their removal without details. They write that even the top business journals were, “strikingly deficient in consistently reporting whether outliers were considered and how they were detected, particulary regarding the non-disclosure of a level of significance. … This wide variety of methods and frequent non-disclosure of methods cannot represent best practice” (Sullivan et al., 2021).

Two years earlier, Leone et al. (2019) lamented something similar after reviewing studies in accounting: “There are a number of alternative approaches to identify and mitigate the effect of influential observations, but there is little (if any) systematic evidence in the accounting literature demonstrating the efficacy and trade-offs associated with these alternative approaches.” They found inconsistent disclosure and discussion of methods used to deal with influential observations, with several studies providing this important research design choice “scattered throughout the text and/or across multiple footnotes.” In studies using winsorization and/or truncation, they identified several studies that implemented some *ad hoc* rule for a subset of variables, while in other studies, procedures used were not clearly stated, rendering replication impossible. They also found a number of studies reporting using the Belsley et al. (1980) regression diagnostics without reporting sufficient detail as to the exact procedures and thresholds/cutoffs employed. “Such inconsistencies make it difficult, if not impossible, to precisely replicate those studies, to make comparisons with the results of related studies, or to reconcile conflicting results between studies” (Leone et al., 2019) They found that only 68% of papers reviewed even mentioned influential observations or describe their approach to identify or treat observations, leaving nearly one-third (32%) neither mentioning influential observations or not clearly describing their approach to identify such observations. While they, as many other authors, recognize the importance of researchers identifying the correct methods for their particular study and research questions, their broad review of accounting research found no clear evidence that the decision to adopt a particular approach for handling extreme values was driven systematically by special characteristics of the data or by the nature of the research questions themselves. While none of the textbooks they reviewed recommended variable-by-variable winsorization or truncation, these methods stood out in their research as two of the most common approaches in accounting (textbooks typically recommending examining the distribution of a regression model’s residuals as an indicator of potential problems with influential observations). Further, Mihaylova et al. (2011) point out that many of the diagnostics used in choosing alternative estimators (due to presence of extreme or influential values) have their own limitations that each require more careful study, but often aren’t addressed or mentioned in discussions.

Though many researchers today consider the two-part model best for fitting health care data with the above characteristics (INSERT CITATIONS), we will see later that analysts continue to handle the extreme values in a variety of ways, often still directly transforming the outcome variable in pre-processing steps in addition to using modeling methods that accommodate skew. Deb et al. (2018) is one such example: while the authors provide an in-depth look at modeling choices (and conclude that best practice is to implement a two-part model with gamma distribution to address the concerns we’ve featured above), and at least inform the reader that they chose to top-code their outcome variable, they proffer little explanation or discussion as to how they chose to do so. From their paper, *Modeling Health Care Expenditures and Use*:

The distribution of total expenditures is highly skewed with a large mass at zero (see Figure 2). More than one-third of observations have zero expenditures, and less than 5% have expenditures in excess of $9,000. In a very small fraction of observations, 35 to be precise, the expenditure values are greater than $50,000 (reaching a maximum of $2,226,997). Although our statistical models are designed to account for skewness, they are not designed to take extreme values such as these into account. Although it is tempting to drop these observations, we are reluctant to do so because we cannot be sure that they are outliers in any real sense. As a compromise, we topcode the value of the expenditure for each of these 35 observations at $50,000. (Deb & Norton, 2018)

### Why disregard for importance of detecting outliers

Sullivan et al. (2021) hypothesize several reasons for the “widespread disregard for the importance of detecting outliers”, all relevant to our discussion outlining the variety, strengths, and weaknesses of handling method. First, though there is a wide variety of methods available, each method has their own strengths and weaknesses, there is little acceptance of any method as a reasonable compromise between ease of implementation and effectiveness. For example, textbooks on statistical methods often do include sections on identification and treatment of extreme cases, but many that they reviewed advised that users implement several or all methods, which certainly can prove challenging for researchers up against time or financial constraints. Sullivan et al. (2019) call this the “toolkit technique”: they write that this dominant presentation of outlier detection methods describes many useful tools, and encourages the reader to use all of them or risk ignorance. Even where they are mentioned, textbooks or methodological suggestions have historically paid little attention to false detection rates or overall level of significance for any one method. As many researchers can attest to, using several methods collectively can lead to the familiar problem of multiple comparisons. This may discourage researchers by the complexity in applying any of the various methods, choosing which to use, knowing what to do if they obtain conflicting results, and handling excessive false positives. False positives become particularly problematic with large sample sizes when cutoff criteria are not appropriately adjusted. Sullivan et al. (2021) offter two examples from commonly used textbooks where the authors recommend that all reearchers answer in manuscripts how outliers were handled, but provide no recommendations for effective methods of identification, reporting critical values used, mention made of controlling Type I error rate, or differentiation made between types of outliers (in the *y* variable, *x* variables, univariate, multivariate, etc). Finally, they point out that many commonly used methods today have evolved from conventions or rules-of-thumb developed many years ago, often using small sample sizes, low dimensionality, and limited computational resources (Sullivan et al., 2021).

Another source of the lack of clarity that is particular to the healthcare field may derive from something Mihaylova et al. (2011) point out as they describe the history of statistical modeling of cost data. In their comprehensive 2010 review, *Review of statistical methods for analysing healthcare resources and costs,* Mihaylova describes two broad areas, pointing out that developed independently over time. In what they call the ‘randomized evaluation’ field, researchers examine the impact of interventions or policies on average costs in terms of mean cost difference by collecting health care resource use and cost data alongside randomised controlled trials (which are often of limited sample size). These studies evaluate the effectiveness, cost-effectiveness, and other effects of interventions to guide policy and treatment decisions. Potential pitfalls of this area include overfitting or misfitting data as they increase in complexity to take into consideration specific features of the data. The second area, ‘health econometrics’ uses large quantities of (mostly) observational data to model individual healthcare expenditures to understand how individual’s characteristics (e.g. health status, medical experiencies, demographics) influence overall costs, and can involve theoretical behavior models wtihin the framework to shed light on processes. Potential pitfalls in this area include biases in estimating effects due to non-random selection and confounding (Mihaylova et al., 2011). Though Mihaylova et al. write that these fields grew independently, it’s easy to see that there could be signficant overlaps in methodology and goals, and thus recommendations from one setting migrate into other settings that may not be appropriate or are appropriate but warrant further investigation.

Finally, the variation in terminology and operationalization of terms alone is enough to cause confusion in readers, preclude replication of results, and render a lack of transparency and clear documentation. For example, Leone et al. (2019) found substantial variation among the N=532 studies using winsorization: 63% (n=337) winsorized all variables, 6% (n=32) winsorized only the dependent variable, 10% (n=51) winsorized all the independent variables, 18% (n=95) wrote “Leave out some variables”, and 3% (n=17) reported winsorization but approach was not clearly stated. They further found that some studies reported using a winsorization rule for most vairables, but then used an *ad hoc* rule for others, such as setting values of discretionary accruals that were greater or less than 200% to +/- 2, or winsorizing all variables except the log of total asssets. These variations will be documented in the following section describing each of the various methods for handling extreme values.

Despite the variety in approaches to handling the issues above, researchers in the field of health economics have widely accepted that the statistic of interest to policy makers is the estimated population mean cost (Arrow & Lind, 1970; Mihaylova et al., 2011).

### Defining/Handling/Detection

As noted above, healthcare cost data often feature “heavy” upper tails that represent true values that might be upwards of 10x the mean, influencing the robustness with which some parameters are estimated. Positively skewed distributions can render the population variance and higher-order moments (such as skewness and kurtosis) very large and even infinite (situations never encountered in a normal or log-normal world (Mullahy, 2009), violating the assumptions of OLS estimation and almost ensuring poor finite sample performance (Karlsson et al., 2024). Violation of these assumptions, as in data on costs of episodes of care with values that are distinctly different from the rest of the sample, render these models inappropriate (Weichle et al., 2013). Random samples drawn from distributions with heavy tails tend to also include extreme, large values, and are often treated as outliers. While removing outliers is not uncommon in practice, in this instance it is ill-advised as the values are not the typical measurement errors, and frequently reflect very real values (Deb & Norton, 2018; Karlsson et al., 2024). Other common treatment methods found in literature such as trimming, topcoding, and winsorization will be explored below, following a more in-depth look at defining outliers.

Traditional outlier definitions, detection, and handling methods need to be examined to highlight the strengths of more contemporary methods, as well as point out areas for further research.

#### Outlier Definition

Traditionally, outliers within a variable are defined as a data value that is so unlike the other values in the sample that ignoring it is likely to lead to significantly incorrect estimates and may have the effect, particularly in smaller samples, of distorting the mean and variance (Chambers et al., 2004; Pirson et al., 2006; Sullivan et al., 2021). In regression analysis, the term “outlier” often refers to a case in which the residual is unusually large, which is often described as having a standardized residual larger than three in absolute value (Sullivan et al., 2021). In at least one study, the researchers used the term “statistical outlier” to denote values that are identified as highly unusual relative to other values part of the statistical analysis (Paddock et al., 2004). Outliers are commonly identified in two types: univariate and multivariate. Univariate outliers are extreme values in one feature of the dataset; multivariate outliers are uncommon values in multiple columns in the same row (Ibrahim et al., 2021). Many authors carefully note, however, that what we often see in healthcare cost data aren’t classical outliers, but are actually tails: natural extensions of a skewed distribution indicating that the data are not evenly distributed around the mean (Karlsson et al., 2024; Metcalf & Casey, 2016; Sullivan et al., 2021). A tail of a distribution, Metcalf & Casey (2016) describe, is the portion of the distribution that is far from the central part of the distribution, but should not be considered outliers, as the term ‘outlier’ is typically used to imply experimental error or some problem computing the value on a data point, and are often removed from datasets.

The importance of investigating outliers to confirm if they are entered erroneously and flagged for some form of correction (e.g. follow up with the respondent, removal) is particularly important for survey data, such as those carried out by national statistics agencies (such as MEPS, like in our study), as many errors can be introduced through the data capture process (Chambers, 1986), such as misunderstanding an item, typing in the item incorrectly, or putting the answer to a particular item in the wrong location. In a seminal paper, Chambers (1986) made the explicit distinction between non-representative outliers – outliers due to incorrect observations – and representative outliers – outliers defined to be correct observations from typical units which are extreme relative to the other data points (Welsh & Ronchetti, 1998). He recommended that non-representative outliers be dealt with via survey editing and imputation or by exclusion. He also points out that finite population parameters are typically based on all population elements, including representative outliers, and that these representative outliers can and should be expected to occur in both sampled and non-sampled portions of the population so that both the simplest expedients of either ignoring their presence (and including) or excluding these representative outliers can lead to very misleading estimates and conclusions (Chambers, 1986; Welsh & Ronchetti, 1998).

As one can imagine, detecting an outlier is much more difficult in small datasets (i.e. less than 20 observations), but in this review, even the smallest datasets far exceeded this case count and thus the identification of outliers in small datasets will not be discussed in this paper.

For the purposes of this paper, after investigating the data points, we will move forward by considering any outliers in *Y1…Yn* as representative outliers. To mimimize confusion moving forward and as we are specifically discussing patients with high costs that are truly part of the distribution and a frequent regular feature of healthcare cost datasets (rather than systematic errors, etc), know that moving forward in this paper, the terms “outlier”, “representative outlier”, and “extreme value” will be used interchangeably to refer only to true values far from the central portion of the distribution only (and not errors).

### Outlier vs Influential Observations

Another important distinction made in literature is the difference between an extreme value (or outlier, if using the broader definition that may capture systematic error) and influential value. Though we will not be identifying influential observations in our study, the discussion is relevant and helps shape our thinking about extreme values, particularly in the context in my of the papers referenced. Extreme values can refer to a variable or dataset with excess or inflated zeros, lots of positives, or extremely large or small values. In a regression context, not all outliers identified by evaluating the distribution of a variable (such as the observation of a long tail) are necessarily influential observations, although all influential observations are outliers. For example, if an extreme value of the dependent variable is paired with an extreme value of the independent variable and the pair falls in line with other data points, then that data point (*y, x)* will not necessary distort coefficient estimates in an OLS regression (Leone et al., 2019). For this reason, Leone (2019) points out that most statistics and econometrics textbooks often advocate examination of a regression model’s residuals, and generally refrain from recommonding winsorization or truncation (unless in the presence of data errors). Influential observations are values whose exclusion causes major change to fitted regression function or parameters and can affect statistical estimates (coefficients) and inferences (standard errors) in linear estimation methods such as OLS regressions (Leone et al., 2019; Weichle et al., 2013). Typically, influential observations exhibit high leverage (potential to influence regression results) and large residual (in absolute value) (Weichle et al., 2013). Leone (2019) further cautions that influential observations can lead to Type I or Type II errors. Again, identification methods for influential observations vary, but one example is from Weichle (2013), who considers an observation to be both influential and outlying only when the predicted DFBETA measurement exceeded 0.15 and the observation was a qualified box-plot outlier.

#### Detection/Identification

As discussed above, recommendations or mentions for detection or handling, when present, vary even among related studies. Leone (2019) stresses that there is no universally correct solution when it comes to detecting or treating influential observations across all possible settings, but notes that all approaches require researcher judgment and care. In Leone’s 2019 review, they found that the most detailed discussion of influential observation identification in statistical literature was found in Belsley et al. (1980), who provide criteria for identification alongside the mechanics of several diagnostic tests. Leone et al. highlight a few main points from the text that are relevant here as well. For example, Belsley et al. (1980, 3) write, ‘‘unusual or influential data points, of course, are not necessarily bad data points; they may contain some of the most interesting sample information.’’ Belsley et al. (1980, 3–4) go on to state that ‘‘only after such data points have been identified can their quality be assessed and appropriate action taken. Such an analysis must invariably produce regression results in which the investigator has increased confidence. Indeed, this will be the case even if it is determined that no corrective action is required, for then the investigator will at least know that the data showing the greatest influence are legitimate.’’

As we discussed above, several statistical techniques can be used to indentify and address outlying and/or influential cases in highly skewed datasets, and doing can potentially improve the precision and efficiency of analysis (Weichle et al., 2013). Again, the array of methods, documentation, and application varies from paper to paper, often with little evidence that the decision to adopt any approach was driven systematically by characteristics of the data (other than the existence of the extreme value) or by the nature of the research question (Leone et al., 2019). Even within each method, researchers implement them in various ways, often without justification or reasoning mentioned (for example, the Deb 2018 quote above). Weichle et al. (2013) observed that the number of observations their team identified as outlying and/or influential varied widely depending on the method employed, and that was using the same dataset; we can imagine that the issue is compounded when considering the diverse array of data and settings. First, we’ll look at commonly used methods found in literature and comprehensive reviews, then will move on to define each method and discuss merits and demerits.

In their systematic review of top business journals mentioned earlier, Sullivan et al.’s (2021) search through articles published from 2007 through 2018 (12 years of journal articles across 24 journals) using the search term *outl\** (and where the context was relevant) yielded 2,582 articles. They found a “surprisingly diverse array of methods used to discover and treat outliers,” highlighting the top five in their paper. Winsorization was the most frequently mentioned method at n=637 articles. Leone (2019) also found winsorization one of the most commonly used methods, noting that while it’s used in many applied empirical studies, the technique was not mentioned in any of the textbooks they reviewed; this points to just one example of inconsistencies between applied and methodological sources. Sullivan et al.’s second most commonly mentioned method was robust analysis methods (n=580) in which outliers were not identified outright, but the authors employed robust analysis methods such as transformations to minimize the effect of possible outliers in the data. Third, the team identified n=558 articles simply referring to the identification of extreme values (and removal, deletion, or ignoring), with no reference to the threshold used to conclude that the outlier was, in fact, extreme or influential. Trimming was the fourth most common (n=233 articles), followed by n=179 articles where the authors indicated that they checked for outliers, but provided no additional details. Trimming was also cited by other authors as a commonly used technique (Gregori et al., 2011; Weichle et al., 2013); Weichle notes that researchers use cutoff levels ranging anywhere from the upper 0.5% to 20% of the cost distribution. Further details will be described below.

#### Standard Deviation

One of the most canonical techniques for examining how spread out the variables in the distribution are is measuring the data point’s standard deviation from the mean (Bhattarai, 2013; Metcalf & Casey, 2016). The standard deviation (SD) measures how closely data points are clustered around the mean, typically denoted by *σ*. Commonly, values greater than *µ*+3*σ* or less than *µ*-3*σ* are considered outliers. Bhattarai points out that SD-based methods are often prone to a masking effect if few extremely large values happen to be together at the upper end. Related to the SD-method, many authors use percentile rank or a ranking method: Bhattarai suggests looking at values higher than the top 2.5 percentile, and another paper conducted a sensitivity analysis trimming values at the 99th percentile (Murphy et al., 2018).

#### Box-plot Analysis

Box-plot analysis, also known as the interquartile method, is another heavily referenced technique that involves the use of distributional characteristics to identify outliers (Cots et al., 2000; Pirson et al., 2006; Weichle et al., 2013). The most common use of this method is to compute the first quartile (Q1) at the 25th percentile and third quartile (Q3) at the 75th percentile, then compute the interquartile range (IQR) as Q3-Q1; points outside Q1-1.5\*IQR and Q3\*1.5\*IQR are tagged for inspection as outliers (Metcalf & Casey, 2016). For example, in their study to find factors that could explain high and low-resource use outliers of factors contributing to hospital length-of-stay, one research team identified high resource use and low resource use outliers using this method (Cots et al., 2000). Citing that they used the logarithmic link function in their generalized linear models, Weichle et al. (2013) incorporated the natural logarithm transformation to identify observations from the full cohort as outliers: outliers were identified where *ln(cost)>Q3+1.5\*IQR* or *ln(cost)<Q1\*IQR* (where *ln* refers to the natural logarithm). Bhattarai (2013) also compared outlier detection using this method against other methods, and also transformed costs to the natural log first to assume the distributions to be log-normal, but detected outliers separately for pre- and post-period expenditures. Bhattarai then removed observations with values larger than *Q3+1.5\*IQR(log of cost)*.

#### DFBETAs

Approaches to identify influential observations include DFBETAs: measures of standardized differences between regression coefficients when a given observation is included or excluded (Choi, 2009; Weichle et al., 2013). Researchers can use this measurement to determine an observation’s magnitude of influence on each regression parameter estimate. Weichle et al. (2013) identified an observation as influential if the absolute value of the predicted DFBETA measurements for two key cost-driving characteristics were greater than the size-adjusted cut-off value of (in their case, ), or approximately 0.03 (Belsley et al., 1980). The team also used 0.15 as a cut-off value for identifying an observation as influential, as a 10-15% change-in-estimate criteria are frequently used to assess confounding in epidemiological studies (Rothman et al., 2008).

Choi (2009) used DFBETAs along with a partial regression plot. Partial regression plots can be helpful visual ways of detecting potential outliers as they plot residuals of variables, providing a clear graphic representation of outliers over a simple scatter plot, displaying the relationship between the response variable and an additional independent variable, both adjusted for other independent variables. Further, Choi points out, while controlling for other variables, a partial regression plot can detect multivariate outliers, which are not normal errors and can potentially impact multiple regression estimates. Partial regression plots also enable visual detection of other potential OLS assumptions such as curvilinear relationships.

#### Cook’s distance

Cook’s distance is another measurement that can be used to identify influential observations. This metric deletes each obervation from the estimation before fitting model parameters, then measures the resulting aggregate change in estimated costs to summarize the influence of each observation on the fitted model (Weichle et al., 2013). Their team identified observations from the full cohort as influential if the predicted Cook’s distance was greater than *4/N*, the conventional size-adjusted cut-off value (Fox, 1991).

#### Other approaches

Other approaches identified in the literature review but not frequently seen in applied papers included selecting outliers based on geometric mean plus one or more standard deviations or using the interquartile method to remove outliers, then calculating the arithmetic mean baed on data that remain after trimming (Weichle et al., 2013), a range equalized at the top method (Bhattarai, 2013).

Sullivan et al. (2021) recommends detecting *y*-outliers using a hypothesis test in which the null model specifies a normal distribution for regression. See their paper for more details about *x*-outlier identification and handling methods.

The number of observations identified as outlying or influential can vary dramatically depending on the method employed (Weichle et al., 2013), and thus the method selection and documentation remains a vital piece of any paper.

### Why OLS is insufficient

Standard linear regression models are often used to predict average costs for patients as these models are easy to apply and results are easy to interpret (Weichle et al., 2013). These models, however, are based on the assumption that regression errors are normally distributed, homoscedastic (variance is the same for any fixed combination of the covariates), and have a linear relationship to the outcome variable (Barber & Thompson, 2004; Diehr et al., 1999; Paddock et al., 2004). Though these methods present results on the scale of interest and may provide unbiased estimates for randomised data, comparative studies have illustrated their sensitivity to extreme values and likelihood to be inefficient in small to medium samples in the presence of non-normality of the underlying distribution (Mihaylova et al., 2011; Weichle et al., 2013). Particularly when modeling cost data, the assumptions of OLS linear regression are unlikely to be met, as costs are usually non-normal, heteroscedastic, and relationships may not be sufficiently linear (Barber & Thompson, 2004). Utilization data can also fail to be independent: there can be multiple hospitalizations for one patient, and/or subjects may be clustered by clinic, hospital, or treating doctor (Diehr et al., 1999). Violating these assumptions, such as in data on costs of episodes of care with values that are markedly different from the rest of the sample either through values of zero or high positive values, render these models inappropriate (Weichle et al., 2013). These specific characterizations of healthcare costs cause a number of difficulties in using standard statistical analysis, such as the linear regressions providing unreliable results, and may mean that normality and efficiency of estimators are not achieved, so the model is not providing the best estimates of the average effects in the population (Barber & Thompson, 2004; Malehi et al., 2015; Mihaylova et al., 2011). Manning (2004) also warns that least squares could provide biased estimates of the mean response of outcome variables in the presence of heteroscedasticity in the log scale error. While it would be tempting to transform the costs and apply OLS analysis (e.g. log-transformed costs), we will see later that many of the approaches seen in literature are flawed as the analysis then is not concerned with population means, but instead with the mean on an alternative scale (Barber & Thompson, 2004). Recent years have seen extensive effort to propose a suitable regression method for the analysis of skewed healthcare data (Faddy et al., 2009; Gilleskie & Mroz, 2004; Gregori et al., 2011; Malehi et al., 2015; Manning et al., 2005), as well as several studies comparing various methods with a variety of conditions (Basu & Manning, 2009; Basu et al., 2004; Basu & Rathouz, 2005; Deb & Burgess, 2003; Malehi et al., 2015; Manning et al., 2005) and a few prominent reviews of statistical methods (Basu & Manning, 2009; Mihaylova et al., 2011). In none of the comparative papers reviewed for this dissertation did any of the authors find the gaussian or OLS models to be the preferred method, often finding poor comparative fit and unreliable results (Need to list all?) (Barber & Thompson, 2004).

The right skew causes a substantial barrier to OLS appropriateness. Mullahy et al. (2009) found that, “[...] heavy upper tails may influence the ‘robustness’ with which some parameters are estimated. Indeed, in worlds described by heavy-tailed Pareto or Burr–Singh–Maddala distributions some traditionally interesting parameters (means, variances) may not even be finite, a situation never encountered in, e.g., a normal or log-normal world.” Leone et al. (2019) writes that implement linear models based on OLS (or its variations), ignores the underlying causes of observations with extreme values, which is problematic as OLS is known to be very sensitive to extreme observations (both in the dependent and independent variables); as such, these observations are likely to affect OLS coefficients and standard errors. Though they point out that OLS will still yield unbiased estimates, even in the presence of skewness, such estimates reflect the ‘average’ effect, which is not typically the most appropriate statistic for hypothesis testing with skewed distributions.

Particularly when dealing with cost or utilization outcomes, we also must keep in mind not only the substantial skewness but the large mass at zero. These distributions with positive mass at one or more points are called mixed distributions: they are neither discrete nor exactly continuous (Blough et al., 1999). Many outcomes (*yi)* in empirical analyses have this underlying mixed discrete-continuous distribution with two basic statistical features: 1) *yi=0*, and 2) *yi>0* observed frequently enough that there are weighty substantive and statistical reasons for special treatment, particularly rendering a single index model undesirable (Belotti et al., 2015). As this paper will follow the lead of several recent, promient papers CITE and implement a two-part model to accommodate the mass at zeros, one limitation of this investigation is that it will not address or explore other methods of handling a mass at zero or the implications if not addressed properly.

Thanks to advances in computing, there are now better alternatives to using OLS to model skewed positive values; a large body of work points to GLMs, two-part models, Poisson regressions, negative binomial regressions, and hurdle models as superior to linear regression methods (Deb & Norton, 2018; Mullahy & Norton, 2024). Health econometrics literature has largely settled on the two-part model as the best way to model a DV with large mass at zero and many positive values (Belotti et al., 2015; Deb & Norton, 2018)

## Addressing / Treating

Several recent papers have addressed these modeling issues from a variety of points of view (e.g., Mullahy, 1998, Manning, 1998, Blough et al., 1999, Angrist, 2001, Manning and Mullahy, 2001). To our knowledge, there is no general empirical approach that simultaneously “makes irrelevant” the decision to transform or not and the choice of two-part versus one-part modeling while also allowing for possibly complex interactions of explanatory variables on the outcome of interest. (Gilleskie & Mroz, 2004)

In practice, recent research has used a combination of modeling and transformation. First we will thoroughly describe the methods most commonly found to treat the data directly (typically the dependent variable), followed by robust regression methods designed to accommodate the presence of outliers, followed by examples using a combination approach, where they treat the data directly in pre-processing in addition to using more complex statistical modeling techniques (beyond OLS linear regression). Our particular concern in this paper is with specifications that transform the outcome variable, often to deal with a mass at zero and right skew in positive outcomes.

### Data Transformation

Data transformation is a broad term describing the process of taking a mathematical funciton and applying it to the data (Metcalf & Casey, 2016). As our particular concern in this paper is with specifications that transform the outcome variable to deal with a mass at zero and right skew in positive outcomes, we will focus on methods found in literature and empirical research that seek the same goal. Mullahy et al. (2023) observe that linear models with transformed *y* are often used in applied work. Some of the most common methods of directly treating or transforming the dependent variable include truncation (removal) and/or winsorization (Leone et al., 2019), top-coding (Deb & Norton, 2018; Geurts et al., 2022; Kandilov et al., 2018; Randall P. Ellis, 2018), and power transformations (Box-Cox transformations, of which log-based transformations are a subset) (Cantoni & Ronchetti, 2006; Chen & Zhou, 2006; Diehr et al., 1999; Gilleskie & Mroz, 2004; Malehi et al., 2015; Manning et al., 2005; Metcalf & Casey, 2016; Mihaylova et al., 2011; Mullahy, 2009; Osborne, 2010; Polgreen & Brooks, 2012). Transformations such as the natural log, inverse hyperbolic sine, and power functions are frequently deployed by applied economists to handle dependent variables that are non-negative and skewed (Mullahy & Norton, 2024).

A canonical specification for scalar outcomes (denoted *y*) and an exogenous vector of covariates (denoted *x*) is , where *f(.)* is a monotonic transformation and *x’* indicates a variables related to, or derived from, the variable *x* (Mullahy & Norton, 2024). Mullahy & Norton (2024) write, however, that there is no reason why transformations must be used in these settings, and that they, along with various authors (Blough et al., 1999; Blough & Ramsey, 2000; Deb & Norton, 2018; Mullahy, 1998), argue in favor of specifying models that “rely on untransformed measures and accommodating their non-negative nature – and the strictly positive natures of their corresponding conditional means *E[y|x]* – using estimation strateglies like log-link generalized linear models (GLMs) with various distribution families. Ironically, though Deb & Norton (2018) argue for this, they also top-code their values (see quotation above – can include again).

Methods employing an initial data transformation are shown to provide potentially more efficient estimates in heavy-tailed data (Manning et al., 2005; Mihaylova et al., 2011; O'Hagan & Stevens, 2003). Ellis et al. (2018) also identified improved prescision in goodness-of-fit measures, and smaller confidence intervals (CIs) in heavy-tailed data. The team found that the CIs around the R2 when the data were not top-coded were close to 30% of the point estimate, even in large datasets (in their case, the large dataset had over 21 million records). They observe that the large CIs occur due to influence of outliers affecting unexplained variation in the dataset.

As in all statistics, there are trade-offs to any decision. The advantages reaped from decision to transform variables, either directly or through modeling, typically come with some cost, whether the data are transformed directly or through modeling. Sometimes, a back transformation is required (such as in log-transformed DVs), other times, the disadvantage is a a reduction in generalizability such that the dataset no longer truly represents the population. When analysts apply many of the approaches that directly transform the DV such as statistical trimming rules, cutoff levels, selecting outliers based on geometric mean plus one or more SD or the interquartile method, the analysis results are relevant only to the sample used, and findings can no longer be compared to other studies (Weichle et al., 2013). Winsorization and truncation particurlarly reduce inter-study comparability of results, and only allow sample-specific inferences (Leone et al., 2019). Other authors warn that many direct transformation methods are “neither statistically nor theoretically sound, as they simply mask the arbitrary nature of the inclusion or exclusion criteria and may severely bias the results” as they remove important observations or change the values of the dependent variable artificially (Randahl & Vegelius, 2024). Many transformations can perform poorly if an innapropriate transformation is used (Briggs & Gray, 1998; Briggs et al., 2005); for example, some transformations such as the log transformation are not appropriate for zero data (Mihaylova et al., 2011). Mullahy (2024) also reminds us of the importance of remembering that one of the main goals (frequently *the* main goal) of applied analysis is to obtain reliable estimates of policy parameters such as marginal effects or average marginal effects of *x* on some functional of the conditional distribution of *y* in its natural measurement units (typically *E[y|x]*). This paper’s emphasis is the same, as our focus will be on *E[y|x]* and its corresponding marginal effects. If we were to specify a linear transformation model, any estimates would have to be retransformed (or back-transformed) to the outcomes’ original (natural) scales. Mullahy (2024) writes that, though this was actually understood as early as 1968 (Goldberger, 1968), it came into widespread recognition in literature on smearing estimators developed during the RAND Health Insurance Experiment (Duan, 1983; Manning et al., 1987).

Most transformation methods found in my literature review related to transforming the dependent variable to accommodate (or remove) skew, researchers also grapple with the large mass at zero. In addition to methods we’ll discuss when implementing log-transformations (and their incompatibility with zero values), one author directly transformed the dependent variable by aggregating monthly costs into quarterly amounts and then divided by three to reduce the likelihood of getting a value of 0 for monthly expenses (Kandilov et al., 2018).

#### Truncation / trimming / removal

Though this paper has provided several reasons to not remove outliers from a dataset, the practice is so common that a more thorough review is warranted. Leone et al.’s 2019 survey of accounting literature indeed found that most studies attempted to mitigate the impact of influential observations through truncation. In reference to the frequency of this approach, it has been called a “traditional solution to increase precision” (Geurts et al., 2022), a “simple and popular approach” (Karlsson et al., 2024), and “the most widely used approaches in accounting” (Leone et al., 2019). Mihaylova’s review cites as a series of papers by Marazzi (Marazzi, 2002; Marazzi & Barbati, 2003; Marazzi et al., 1998; Marazzi et al., 2009) as examples illustrating the use of truncation to provide more robust parameter estimates.

The practice of truncation, also known as removal or trimming, entails discarding a specified fraction of the most extreme cases, usually based on sample percentiles and univariate considerations of the variables handled one-at-a-time (Sullivan et al., 2021). It is generally performed prior to conducting any statistical tests, depend on researcher-selected cut-offs, and can be done at one or both ends of the dataset, such as the 1st and 99th percentiles of each variable’s distribution (Geurts et al., 2022; Leone et al., 2019; Sullivan et al., 2021). When done at one end of the dataset (removal of high extreme values), truncation imparts a bias toward zero in OLS coefficients (Leone et al., 2019). The process of truncating both ends of the dataset is usually referred to as winsorization. Sullivan et al. (2021) describe two general processes for truncation / winsorization. The first, trimming by *r*%, involves removal of a total of r% of the cases, usually those below the r/2 percentile and those above the *1-r/2* percentile. The second method involves removing a total of *2\*r*% of the cases, *r*% on both sides. Though simple to implement, Sullivan et al. add to Leone et al.’s (2019) concern that there is often no accompanying theory or justification guiding the selection of the fraction to trim. They found that this process can simultaneously both fail to remove all outliers and also remove valid observations that belong in the dataset. They point out, as others have in previously in this paper (Belsley et al., 1980), that many extreme cases are indeed representative of the sampled population and their removal makes the sample less representative. As it is a univariate technique, it neglects incorporation of any correlations among variables, thus precluding any possibility of detecting *x*-outliers (Sullivan et al., 2021).

Researchers should anticipate these very large values and plan for them, as random samples drawn from a distribution with heavy tails (such as those in healthcare expenditures) will generally include high cost values. We have noted earlier that, in our case, these are neither classical outliers nor measurement errors; that being the case, then, truncating or removing them removes potentially valuable information, and “ignores precisely those individuals who are responsible for the lion’s share of per capita health spending” (Karlsson et al., 2024). Geurts (2022) makes the important point that reweighting or removing high-cost patients invites the risk of excluding a significant portion of the costs affected by the very initiatives researchers are attempting to measure, and could thus lead to systematic underestimations of realized savings. As integrated care initiatives commonly focus on populations with chronic illnesses, they will inevitably contain records from patients with high levels of expenditures. Short-term savings by integrated care are mainly expected from improvements in treatment methods or procedures, which are likely to be concentrated among these patients. In an attempt to obtain more robust estimates through removal of these data points, researchers may invariably be removing data points critical to answering their research question/s. Truncation or winsorization inserts the incorrect assumption that the data are contaminated in some way, a way that is not appropriate in the case of healthcare resource use and costs where high values or zero observations are true values and where exclusion of these cases can lead to substantial bias in estimation of averages (Mihaylova et al., 2011).

In theory, one could posit that extreme observations result from a different process inherent in the phenomena of interest, and perhaps make it appear problematic to include the most extreme observations in the analysis; in the healthcare setting, for example, patients with higher costs and/or utilization clearly have some distinctive reason for having values markedly higher than the rest of the population (Randahl & Vegelius, 2024). In Randahl et al. (2024), they include the example of episodes of one-sided violence, stating that in this context, researchers could argue that cases of countries with ongoing, active genocides be excluded, rationalizing that genocide arises from different processes than other forms of one-sided violence. Including these cases, as they are definitively larger compared to the central distribution, may mean they have a relatively larger impact on model estimates than the majority of other observations (though we saw in Leone (2019) that that might not, of course, always be the case). The team’s thought experiment comes to the conclusion, however, that excluding these most prominent cases of the phenomena the researcher is interested in seems “like a strange choice”.

Leone et al. (2019) offered some texture about how truncation is applied in the studies they surveys, as these were the first and second most common approaches they encountered across their broad review. Truncation, specifically in this case using cutoffs based on percentiles of a variable’s distribution, accounted for 30% of the studies reviewed (258 out of 851). Of these n=258 studies, 31% (n=79) truncated all variables; 9% (n=22) truncated only the dependent variable; 4% (n=10) truncated all independent variables; 51% (n=132) noted that they “left out some variables”; and 6% (n=15) had uninterpretable approaches due to lack of description / clarity. They also pointed out that, when choosing truncation rules, some studies truncated using non-percentile cutoffs such as truncation of earnings sclaed by total assets at +/- 3 or stock prices less than $1.

Many authors of the comparative studies included in this paper recommend that future studies shift away from truncation due not only to theoretical concerns but empirical ones as well. Murphy (2018) performed a sensitivity analysis with trimmed values at 99th percentile, and found results consistent with methods that instead accommodated the extreme values rather than deleting them. Geurts (2022) excluded the top 0.1% and top 1% of high-cost cases, and found that truncation and top-coding methods (which will be mentioned later) lead to an underestimation of realized savings at the top end of the distribution, effectively lowering incentives to implement cost reduction initiatives. Leone et al.’s simulation results (2019) indicate that truncation and winsorization are largely ineffective in dealing with truly influential observations, noting that influence diagnostics out-perform these methods, though they themselves are sensitive to effects of “masking” and “swamping” (problems that occur when extreme values are clustered together, leading to over or under-identification of outlying cases).

Not every instance of removal or truncation is unwarranted, however. In fact, Leone et al. (2019) point out that there are some special cases with economic or econometric reasons for a variable-by-variable cutoff; for example, studies may delete observations with a stock price less than $1 due to restriction on institutional ownership of low-priced shares, or in the case of errors. There are likely more instances where truncation or removal is both desirable and appropriate, but these are outside the scope of this paper as we are interested in modeling health care expenditures where the extreme values are tails of the naturally-occuring and real underlying distribution.

#### Winsorization

A related method to truncation, and listed by many authors as similar in popularlity, is the process of winsorization. Winsorization is a symmetric process that, rather then removing extreme values, instead replaces the smallest *k* and largest *k* data values with less extreme values at that cutoff point (Sullivan et al., 2021; Weichle et al., 2013). Typically based on counts, researcher can modify based on quantiles, such as the 5th and 95th percentiles, though using quantiles doesn’t necessarily lead to a symmetric process, and might result in potentially changing more values in one tail than another (Wicklin). This process can be used to transform the cost of outlier episodes to that of some pre-established percentile; using the example of 5th and 95th percentile above, winsorization would transform costs for patients above the 95th percentile to those of the costs of the patients in the 95th percentile, and those with costs below 5% to the costs of patients at the 5th percentile (Weichle et al., 2013). Randahl (2024) also calls this “censoring observations to a threshold value.” We can look again to Leone et al.’s review of studies using winsorization to see the smorgasbord of ways researchers implemented this method. Of the 532 studies they identified using winsorization, 63% (n=337) winsorized all variables; 6% (n=32) winsorized only the dependent variable; 10% (n=51) winsorized all of the independent variables; 18% (n=95) ‘‘Leave out some variables’’; and 3% (n=17) of the studies were considered “Unclear cases” (e.g., a study stated that they winsorized extreme cases but did not state which variables). They also found some studies report using a winsorization rule for most variables, but then use an *ad hoc* rule for other variables (such as setting values of discretionary accruals (as a percent of total assets) greater or less than 200 percent (-200 percent) to +/- 2, winsorizing all variables except the log of total assets, and setting values of effective tax rates that are greater than 1 (or negative) to 1 (0). (Leone, 2019)

Use Weichle’s paper to see their results: they winsorized costs at 2nd and 98th percentiles and at the 5th and 95th percentiles.

Again, while Leone et al. (2019) stress that winsorization is one of the most widely used approaches in accounting, their survey found no clear evidence that decisions to adopt this approach was driven systematically by any particular characteristics of the data or by the research questions. Further, the variation in application gives it the distinct advantage of reducing inter-study comparability of results and limits inferences to the sample studies, reducing generalizability. Their simulation results indicate that winsorization is “largely ineffective for dealing with truly influential observations, because winsorizing does little to mitigate the effect of influential observations.”

#### Topcoding

The sensitivity of results to right-skewed spending distributions is commonly addressed in healthcare expenditure data by top-coding high-cost cases (Geurts et al., 2022). Similar to winsorization in that it retains individuals with high levels of payments, topcoding methods essentially winsorizes only the right, positive side of the distribution; if a cost values exceeds some threshold, the portion above the threshold becomes the minimum of actual spending and threshold value (Geurts et al., 2022). Flexible, transparent, and less extreme than removal, top-coding the dependent variable is common in shared savings models to define spending benchmarks (Geurts et al., 2022), risk adjustment models (Randall P. Ellis, 2018), healthcare expenditures, researchers replace spending on individuals above a threshold (such as $250,000) with that threshold. Thresholds are often based on overall cost distributions and can be an element of negotiation with partners in a program, chosen on the basis of insights in specific care groups that might lead to high spending and as a function of the aims of the program, thereby rendering this one method for the treatment of high-cost cases that is not always implemented solely on the bases of statistical considerations but is also motivated by understandings of the actual scope, design, and goals of a given program (CITATION: I lost it! find where this came from!). Though most papers mentioned top-coding the dependent variable, at least one instance found in my literature search implemented top-coding in both the dependent and right-hand side variables at the same threshhold (Randall P. Ellis, 2018).

Top-coding effectively reduces the impact of high-cost cases while simultaineously retaining these records in the model, which is preferable to dropping high cost observations altogher, as extremely high costs are often predictably linked with associated conditions inherent in the study (Geurts et al., 2022; Randall P. Ellis, 2018). Top-coding these outliers not only increases the R2 of the model, but decreases confidence intervals to negligible amounts, increasing predictive power and rendering coefficient estimates stable (Geurts et al., 2022; Randall P. Ellis, 2018). Geurts (2022) points out that though this method is less drastic than removal, top-coding can still lower incentives for cost control, forcing researchers into facing another trade-off: top-coding to increase robustness of estimates, or risk reducing incentives for cost reduction.

Examples in the literature found alternate values for top-coding ranging from $50,000 to $1 million (Deb & Norton, 2018; Randall P. Ellis, 2018), or even set at percentiles such as the 99th percentile (Kandilov et al., 2018) as robustness checks for their model. Ellis (2018) found that top-coding at $250,000 reduced confidence bands to a close to zero range, and Kandilov (2018) found results in their robustness check consistent with findings without top-coding. In a study examining the prevalence and amount of care-related out-of-pocket expenditures of family caregivers, one team created an ordinal variable from their original outcome variable (expenditures) using the midpoint of ranges and a top-code of $3000, though I could not find mention of why they chose that value nor any citations regarding top-coding in the paper that might provide insights as to their rationale. They then created three ordinal responses from the top-coded values: less than $500 per month, $500-$2000 per month, and more than $2000 per month. In Deb et al. (2018), the team top-coded at $50,000, but again I could not find any rationale, either in citations, footnotes, or the text, to better understand their rationale for the selection. We will visit that paper in more detail later.

### Mean-preserved top coding

### Median-preserved top coding

### Transformation functions

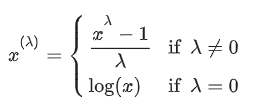
“Approaches to find an appropriate transformation have centered predominantly on the Box-Cox transformations (to achieve symmetry in error), with log transformation often prevailing when modelling costs.” (Mihaylova et al., 2011)

“OLS for ln(y) is one of the most prevalently used (and most prevalently misused) models for analyzing [healthcare expenditure] data.” (Manning et al., 2005)

Power transformations are a parametric family of data transformations offering a spectrum of opportunities for precisely tailoring a transformation to suit the data shape needs (Osborne, 2010; Ruppert, 2001). Power transformations include any transformation that raises numers to an exponent, or power; for example, a square root transformation can be characterized as *x1/2* andinverse transformations can be characterized as *x-1*. In empirical work, Mullahy et al. (2024) write that the most common power transformation in empirical work is the square root function (*λ*=0.5), as only the square root function is easily retransformed back to the original scale. Log transformations are themselves a form of power transformations, a class of transformations rather than a single transformation, and log-normal variables (i.e. variables that are normally distributed after log transformation) are popular across many fields of science (Osborne, 2010). Two methods can be used to analyze the data on the log scale: the first is to directly transform the data by applying a function to the outcome variable, then apply OLS, and the second is to choose a method that incorporates the logarithmic scale though a link funciton, such as in generalized linear models (Diehr et al., 1999). These two methods are very distinct as the first requires a constant to be added to the zero values and retransformation to the original scale. This section of the paper will discuss the first method (directly transforming the dependent variable in preprocessing) and the second method later.

Log transformations are a special case of the Box-Cox transformation when *λ*=0. Mullahy et al. (2009) write that the log-transformation of healthcare cost data is so ubiquitous, it almost seems “an automatic response to observed skewness in the data (considerations of 0 observations on cost notwithstanding).” Mullahy et al. (2023) highlight that the natural logarithm transformation’s long history in empirical economics, going back as far as 1983 (Duan, 1983; Duan et al., 1983; Manning et al., 1987), for outcomes that are positive, and often when researcher goals include computing marginal effects or elasticies. In 2006, one author noted the use of the log transformation (log-normal distribution) across such various contexts as the application of breath analysis to medical cost data and intensive care unit expenses (Chen & Zhou, 2006). For example, to determine Medicare reimbursement for teaching hospitals, the Centers for Medicare and Medicaid Services used a log transformation of the ratio of the number of interns and residents to beds (plus a constant) (Dalton & Norton, 2000; Mullahy & Norton, 2024; Rogowski & Newhouse, 1992). Mihaylova et al. agreed with this sentiment in their 2010 review of methods for handling extreme values, writing that, “Approaches to find appropriate transformation have centered predominantly on log transformation when modeling costs. Five years later, Malehi et al. (2015) continued to find it a “very common approach,” as did several authors writing well into 2022, 2023, and 2024 (Chen & Roth, 2024; Geurts et al., 2022; Mullahy & Norton, 2024). Polgreen (2012) writes that performing OLS on the natural logarithm of costs to decrease the importance of extreme cost observations in estimation is a “traditional approach to addressing problems of skewed costs.” As of 2005, Manning et al. called this approach, “By far the most prevalent estimation approach used in health economics and health services is to use ordinary least squares or a least-squares variant with ln(*y*) as the dependent variable,” citing that its popularity stems from its ability to obtain resulting error terms that are approximately normal. Power transformations de-emphasize extreme values so that they no longer dominate the estimation result, and can potentially obtain a bell-shaped distribution from skewed data to restore symmetry (Gilleskie & Mroz, 2004; Karlsson et al., 2024; Metcalf & Casey, 2016). Further, because of skewness, this transformation also helps at the bottom tail of the distribution, and the removal of the effect of outliers via use of the log-transformation will “decrease the confidence interval to negligible amounts” (Randall P. Ellis, 2018).

Osborne (2010) credits Tukey as the first to present the original idea that transformations can be conceptualized as a class or family of related mathematical functions in 1957, an idea updated in 1964 by Box and Cox to take the form of the Box-Cox transformation.

**(Bicego & Baldo, 2016)

Given that 𝜆 can take on an almost infinite number of values, we can theoretically calibrate a transformation to be maximally effective in moving a variable toward normality, regardless of whether it is negatively or positively skewed. Additionally, this family of transformations incorporates many traditional transformations: (Osborne, 2010)

|  |  |
| --- | --- |
| 𝜆 value | Transformation |
| 1.00 | No transformation needed; produces results identical to original data |
| 0.50 | square root transformation |
| 0.33 | cube root transformation |
| 0.25 | fourth root transformation |
| 0.00 | natural log transformation |
| -0.50 | reciprocal square root transformation |

These power functions proposed by Box and Cox were designed to render a continuous variable more symmetric and improve normality (Mullahy & Norton, 2024). Mullahy et al. (2024) focus on a specification in their paper *fp (y, c)=yλ* that they consider simpler than the original Box-Cox function *fBC(y, λ)=(yλ-1/ λ)*, which permitted the natural log transformation as a limiting case as lambda approached 0. As it is not possible to transform marginal effects back to the original scale for arbitrary values of *λ*, Mullahy et al. (2024) do not estimate power functions in their examples. They also note that power transformations also separate zeros from the positives. When logarithmic and square root transformations are used to transform utilization and cost data, both transformations will cause problems if it’s necessary to report results on the original scale as estimates will be biased by the process (Diehr et al., 1999).

#### Log transformation

In log transformations, target variables are replaced with *log(x)*, where the base of the log is determined by the analyst; common selections are base 10, base 2, and the natural log *ln* (Metcalf & Casey, 2016). Mathematically, the logarithm of a number less than 0 is undefined, and, just as in square root transformations, numbers between 0 and 1 require unique treatment than values greater than one (Metcalf & Casey, 2016; Mullahy & Norton, 2024). Since cost records often include values of 0, common practice is to add one to any values before taking logs (Chen & Roth, 2024; Osborne, 2010; Randall P. Ellis, 2018). The addition of any positive constant *c* (in most cases, *c=*1) is arbitrary and lacks any real justification or theoretical reason for choosing any positive constant (Mullahy & Norton, 2024). Mullahy et al. (2024) note that the practice is so widespread that “prominent statistical platforms offer built-in functions that perform the *log(y+1)* transformation directly, such as Stata’s *log1p(x)* function.”

Malehi et al. (2015) describe the regression model as follows:

(ChX.1)

where it was assumed that *E(xε)=0 and E(ε)=0*, since predicting costs on the original scale is the primary objective, so:

If the error term is distribution, it is a log-normal case, and then:

However, if the error term is not normally distributed, but is homoscedastic, then the smearing estimator is applied (Duan, 1983; Malehi et al., 2015).

Mullahy et al. (2024) write the case of the natural log transformation as follows:

Where the parameter *c>0* guarantees that the left-hand side of is defined when *y=0*.

Despite the ubiquity of the method in empirical studies, many of the authors of comparative or methodological papers reviewed came to general consensus that estimating the mean cost using a logarithmic regression model leads to biased estimation (Cantoni & Ronchetti, 2006; Deb & Burgess, 2003; Malehi et al., 2015; Mihaylova et al., 2011; Mullahy & Norton, 2024), the magnitude of which depends on the unknown distribution of *Yi* (Karlsson et al., 2024). Mihaylova et al. pointed out in 2010 that “Logarithmic transformations are not appropriate for zero data, and the device of replacing zero by a small number is not recommended.” Issues related to this practice have resurfaced in methodological literature in 2024, with papers by Chen & Roth (2024) and Mullahy & Norton (2024) both exploring the consequences of incorporating cases with values of zero or close to zero; the conclusion from both papers is that results tend to be very sensitive to methods of handling values of zero and that other modeling approaches are preferred (Karlsson et al., 2024).

Many authors wrote about concerns with the modeling results after applying the requisite retransformation to get the results back to original scale (Chen & Roth, 2024; Diehr et al., 1999; Gilleskie & Mroz, 2004; Randall P. Ellis, 2018). Chen et al. (2024) reference previous work illustrating simulations or select empirical applications that results for transformations such as *log(Y+1)* or *arcsinh(Y)* (discussed later) may be sensitive to the units of the outcome (Aihounton & Henningsen, 2021; De Brauw & Herskowitz, 2021), and that concurrent work by Mullahy & Norton (2024) show that, in theory, the marginal effects from linear regressions using these transformations are “sensitive to the scaling of the outcome, with the limits of the marginal effects approaching those of either a levels regression or a (normalized) linear probability model, depending on whether the units are made small or large.” Their work complements these studies by proving that scale dependence is a necessary feature of any identified average treatment effect (ATE) that is well-defined with outcomes including mass at zero, and that this dependence on units becomes increasingly problematic for transformations that approximate *log(Y)* when *Y* is large. They conclude that choosing “better” transformations or using different estimators can not fix the issues with these transformations. Geurts (2022) also points out that transformed data provide results on a log scale, and yet empirical applications largely require results expressed in terms of actual costs in the original scale; the issues inherent in retransformation are exacerbated in the presence of heteroskedasticity in data on the transformed scale. Ellis (2018) points out that in order to compare across specifications, predictions from log linear models must be transformed back into their original scale, and that these models “invariably do worse than linear models once this is done.”

#### Inverse Hyperbolic Sine

Mullahy et al. (2024) point out that recent years have seen an increase of use of the inverse hyperbolic sine transformation as a solution for dependent variables that are non-negative, skewed, and have a large mass at zero, citing examples from fields such as labour economics, environmental economics, public economics, charitable donations, and economic history. They attribute the popularity to the widespread perception that this method apprached the natural log model (as *k* gets large), but their research shows that this is not the case. The transformation is as follows (Aihounton & Henningsen, 2021):

Where *k* is implicitly assumed to equal 1 by most papers, but Aihounton & Henningsen (2021) explored the “behaviour of models that scale y multiplicatively, with *k>0*, before applying inverse hyperbolic sine transformations.”

I have more on this but can add later; not sure how relevant it is.

#### Single index / one-part models

Previously, this paper documented the issues with a single-index OLS regression, as well as direct data transformations commonly applied to improve symmetry. Now we will briefly outline several approaches that, rather than modify the data, assume some parametric density of *Yi* to accommodate the heavy tails in a one-part model or single-index model (Karlsson et al., 2024) such as the generalized gamma distribution (Diehr et al., 1999; Malehi et al., 2015; Manning et al., 2005) as it is a multiplicative model (Diehr et al., 1999), or the generalized beta of the second kind distribution (Jones et al., 2014), which covers the Pareto distribution and the Burr-Singh-Maddala distributions as special cases (Karlsson et al., 2024). Diehr et al. (1999) write that researchers looking to understand the effect of individual covariates on total costs might find a one-part model more insightful as it generates a single regression coefficient for each variable and ease of interpretation. Models regressed on the dollar scale specifically might be particularly easy to interpret, but assume an additive model, whereas the gamma distribution might be preferred as they are multiplicative models.

#### Cox Proportional Hazard Model

Numerous authors have proposed the use of the semiparametric Cox proportional hazards as a special, more flexible model when it comes to cost analysis, attractive in that the underlying distribution does not need to be specified for these models, and cost data are typically skewed and difficult to parameterize (Diehr et al., 1999; Malehi et al., 2015). The method, however, has been a controversial one in healthcare data modeling due to several important drawbacks. Based on survival and hazard functions rather than direct estimation of *E(y|x)*, the main issue researchers must grapple with in this model is the proportional hazards assumption, which states that the hazard ratio of two individuals is time invariant; the model is considered semiparametric as it does not require specification of the baseline hazard function:

where is the baseline hazard function, estimated via the Breslow method; as the proportional hazard assumptions means the hazard ratio of two individauals is independent of time, interpretation of coefficient estimates in these models are based on the hazard ratio rather than the covariate effect on the mean (Malehi et al., 2015).

Diehr et al. (1999) write that one issue is that the regression coefficients in this model pertain to the hazard ratio, which can be challenging to interpret in the medical cost context. “Rather than referring to a difference in means or ratio of costs”, they write, “the regression coefficients refer to the relationship of the cumulative distributions.” They provide the example that if FEMALE was a binary variable in the equation with a coefficient of 1.2, this output would mean that the “proportion of females with costs above *x* would be equal to the proportion of males with costs above *x*, raised to the *e1.2* power (for all values of x).” For these reasons, they opt to not use the Cox model in their comparative analysis.

In simulations studying sample size, Malehi et al. (2015) found that the Cox proportional hazard model revealed maximum MSE and less accurate 95% simulation intervals, even within proportional hazards data-generating scenario; investigation of the pattern of the residuals as a function of X (implemented by the mean of the residuals across deciles of X) showed more bias for this model across all generated data and sample sizes. In relatively smaller sample size conditions, results revealed that the Cox proportional hazard model exhibited poor estimation of population means of healthcare costs and the *β1*, even under proportional hazard data.

#### Generalized Linear Models

“Indeed, GLM captures particularities of health care spending distributions – including their long right tails and large mass points at zero spending, and it is more efficient than the transformed log model” (Karlsson et al., 2024).

This section will provide an overview of generalized linear models (GLMs), then discuss their application in single-index models. GLM integration in two-part models will be described later in the paper.

While researchers have frequently used logarithmic or other power transformations of the dependent variable in the face of significant skewness, more and more papers in the last 20 years have turned to GLMs to avoid the complicated and requisite process of transformations and corresponding retransformations (Buntin & Zaslavsky, 2004; Gilleskie & Mroz, 2004; Hardin & Hilbe, 2007; Polgreen & Brooks, 2012) as estimation is directly on the scale of the original raw data (Mihaylova et al., 2011). Largely following the lead of Mullahy (1998) and Blough (1999, 2000), the last 20 years have seen a rise in popularity of the use of generalized linear models, and now enjoys widespread use in health economics to model health expenditure distributions, particularly when deployed with the log-link function (Belotti et al., 2015; Buntin & Zaslavsky, 2004; Deb & Norton, 2018; Polgreen & Brooks, 2012).

For the regression of cost data, GLMs are particularly attractive as they provide parametric methods of analysis and allow researchers to specify a variety of non-normal distributions as well as the ability to alter the way covariates behave in the model (Barber & Thompson, 2004). Deb et al. (2018) write that these models “accommodate skewness in natural ways, give researchers considerable modeling flexibility, and fit healthcare expenditures extremely well.” GLMs offer a range of alternate functional forms that give enable researchers to match the relationship between the expected value of the dependent variable with the linear index of covariates, while explicitly modeling any heteroskedasticity by allowing “the variance of the outcome to be a function of its predicted value by the choice of an appropriate distribution family” (Deb & Norton, 2018). Mean function estimates from GLMs have been found to be largely robust to outliers and influential observations, and less sensitive to outliers or influential observations than linear regression models (Weichle et al., 2013).

GLMs explicitly accommodate skewness in large datasets through weighted variances; the term ‘Gaussian’ in GLMs is used when the mean and variance are unrelated (whereas in OLS, this term means errors are distributed normally) (Blough & Ramsey, 2000). When implementing these models, researchers specify an appropriate model for the mean and outcome variable as well as the correct mean-variance relationship (also called the variance function) (Mihaylova et al., 2011) through selection of a distributional family (Polgreen & Brooks, 2012). This approach “addresses linearity in response on the specified scale and accommodates skewnes through variance weighting” (Mihaylova et al., 2011). Further, the similarity of regression estimates indicates that GLMs are robust in handling skewness in large datasets; one team found evidence that estimates are robust to the specification of distribution among gamma, Poisson, and inverse Gaussian families (though warn that this might not be true in all situations and researchers must still run sensitivity tests and take care when specifying the variance function (Weichle et al., 2013). Another study found that mean function estimates were found typically robust even in the face of misspecification of the variance function, though such misspecification could lead to other inefficiences and loss of precision (Manning et al., 2005; Mihaylova et al., 2011; Weichle et al., 2013).

Can include findings from applied studies here if needed

Another strength of this modeling choice manifests in its robustness to outlier identification methods; though each of Weichle et al.’s (2013) methods of identifying outliers or influential cases identified different number of observations as such, the GLM method produced generally similar overall average costs as well as average costs by stage of diagnosis and colectomy receipt. In addition, they found the GLM-produced economic rates of return of key cost-drivers quantitatively and qualitatively similar and comparable in magnitude.

Of course, as is often true in research, there are trade-offs inherent in every decision. Even GLM models that produce unbiased estimates might display some imprecision in instances where log-scale errors are symmetric but heavy-tailed, or in the case of large (>1) log-scale error variances (Manning et al., 2005; Mihaylova et al., 2011; Mullahy, 1998; Weichle et al., 2013). Polgreen (2012) expounds further, writing that while these methods might mitigate problems associated with extreme residuals, other issues might be introduced:

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 IVs 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.(Polgreen & Brooks, 2012)

### GLM modeling

The central structure of a GLM is an exponential conditional mean (ECM) or log link relationship (Manning et al., 2005). The process described below is taken from Maheli et al. (2015). An invertible link function *(g(.))* converts the expectation of the dependent variable *E(Yi)* to the linear predictor:

This ECM model, a special type of GLM with log-link function, is equivalent to a nonlinear regression model:

They note that Weibull and Gamma regression models are two special types of this ECM model, where β values can be estimated via quasi-likelihood estimation; the exponential distribution was considered a special case of the Weibull and Gamma regression models where shape parameters are equal to 1.

GLMs require the researcher to select both a link function and distribution family (Deb & Norton, 2018). The link function is what actually relates the expected value of the outcome variable to the coefficients and linear index of covariates (such as ) (Diehr et al., 1999); in the case of health care expenditures, the natural logarithm is often the appropriate link selection (the log of the expected value of expenditures is the linear index), though other link functions are certainly possible (Deb & Norton, 2018). To see which power function would best suit their dependent variable (health care expenditures) to be closes to symmetric, Deb et al. (2018) used a Box-Cox test two ways: with and without controlling for covariates, and in both cases, limiting the test to positive observations. This approach tests which scalar power (*δ*) of the outcome variable (*yδ)* results in the most symmetric distribution; powers of *δ=1* indicate a linear model, *δ=0.5* indicate a square root transformation, and *δ=0* corresponds to the natural log transformation (CITATION FIND (Deb 2017 Health econometrics using stata)). The team found that results weren’t appreciably different: in both tests, estimated coefficients were close to zero, indicating the natural log transformation as the appropriate link. As this paper will be using the same dataset in their paper, my analysis will follow this method.

The next decision facing researchers in using the GLM is specification of the distribution, or relationship between the mean and variance, such as *σ2=µ2*. Again, this paper will follow Deb et al. (2018) results, obtained from the following process. As discussed previously, we saw that GLMs allow for allow for heteroskedasticity by allowing the variance to be specified as a function of the mean rather than held constant across all observations; one might expect, then, that a higher variance makes theoretical sense for datasets with higher expected values: for example, variances could be proportional to the square or cube of the mean. After running a GLM with log link and gamma distribution, the team used a modified Park test (Park, 1966) to empirically assess the relationship between the mean and variance by computing the expected value (mean) for each observation conditional on covariates as well as the squared error (variance) for each observation. Regressing the logarithm of the squared error on the expected mean values provides the test results. If the coefficient on the expected value is close to 0.0, the Gaussian distribution is appropriate as the value of 0.0 implies that the variance is independent of the mean. When results yield coefficients close to 1.0, 2.0, or 3.0, researchers should choose a Poisson-type distribution, the Gamma distribution, or the inverse-Gaussain distributions, respectively. In their sample, they observed an estimated coefficient of 1.83, supporting specification of the gamma distribution. This analysis will follow suit.

[Deb 2018 also has section on specification of explanatory variables: need??]

### Two-part models

“Assuming that the analyst is interested in estimates of *E(y|x)* and of ∂*E(y|x)/*∂x*,* the tpm is almost always an adequate (if not superior on precision grounds) way to model mixed discrete-continuous outcomes if there are no exclusion restrictions.” (Belotti et al., 2015)

As discussed, health care expenditure or cost data often feature a significant point mass of zero, accounting for the records of individuals who have observations with no utilization or resource use. Two-part models are based on these mixture models, and provide one approach to account for this mass of zeros (Belotti et al., 2015). It has become common among many health economists to implement two-part models despite the fact that the outcome of interest is often the unconditional outcome that includes the original zero values: health outcomes such as expenditures, number of doctor visits, and length of hospital stays often exhibit the skewness and point mass at zero that make two-part models attractive to researchers across various contexts. After enjoying a long history in empirical analysis, the two-part model has largely been setting on by health econometrics as “the best way” to model a dependent variable with these defining characteristics (Deb & Norton, 2018). Two-part models have been found to improve predictive performance over single-index models and, when appropriate, should be explored as one way of improving estimation and prediction (Kapitula, 2015). Particularly when models include covariates, GLMs offer some of the benefits of alternative distributions and/or transformations without the requisite retransformation (Mihaylova et al., 2011). For researchers whose goal is to understand a system, the two-part model allows investigators “to distinguish factors that affect the propensity to use any services from factors that affect volume of utilization once the person has entered the system” (Diehr et al., 1999), also known as extensive and intensive margins. Mullahy et al. (2024) write, “If primary interst is estimation of marginal effects at both extensive and intensive margins, then we recommend using a two-part model when there is a mass at zero for a continuous outcome.” Deb et al. (2018) points out that it is generally reasonable to expect heterogenous treatment effects across distributions of outcomes, particularly where an important source of heterogeneity springs from one that gives rise to zero versus nonzero resource use. Not only does making this modeling distinction improve effect estimates, but extensive margins (zero versus nonzero) and intensive margins (how much where nonzero) are often of substantive policy interest. Two-part and hurdle models allow researchers to not only estimate these margins separately (and in a way that single-equations can not), but also estimate overall program effects (Deb & Norton, 2018). For example, in Deb et al.’s 2018 analysis of health care expenditures, they found that changes in expenditures and office-based visits were on the intensive margin, while the change observed in emergency visits was attributable only to the probability of said visit. These types of findings are of great interest for implementation specialists and policy researchers.

The method, developed by Cragg in 1971 as an extension of the tobit model (Cragg, 1971), was used throughout the 1970’s by economists; during that same time period, meteorologists were using versions of the two-part model to model rainfall trends; it became more widely used in health economics and health services research after a team at RAND Corporation used it for the Health Insurance Experiment (Belotti et al., 2015; Duan et al., 1984). In Berlotti et al.’s 2015 paper, the team makes the distinction between the two-part model and the hurdle model, writing: “The two-part model has a commonly used counterpart for count data called the “hurdle” model. … We will use the term “two-part” model to distinguish models for continuous outcomes from models for count data.”

As mentioned previously, there are two ways to analyze data on a log scale: the first entails analyzing the log of cost (for only positive values or for all records after adding some arbitrary constant value) using OLS linear regression models. The second method is to select a model that incorporates a logarithmic scale through a link function, such as the GLM.

The basic structure of a two-part model is that, utilizing the entire data set, the model estimates the probability of a target variable being positive, then another model – a conditional outcome model - is run conditional on the target variable having a positive value (Kapitula, 2015). Two-part models are “based on a statistical decomposition of the density of the outcome into a process that generates zeros and a process that generates positive values” (Deb & Norton, 2018), with a logit or probit model typically estimating the probability of observing a positive-versus-zero outcome (Belotti et al., 2015). Conditional on a positive outcome, the researcher then fits an appropriate regression model. In the first step, the binary choice model derivative would show how the probability of observing a positive outcome changes along a linear index of covariates *x*. The derivative of the conditional outcome model would reveal how the expected value of *y* changes along *x* for positive outcomes. The GLM estimator used with a two-part model has been suggested across multiple papers for work in health economics (Blough et al., 1999; Manning, 1998; Manning & Mullahy, 2001), with the two-part model using a probit for expenditures, and in the conditional model, a quasi-likelihood function defined by where *g(.)* represents the natural log-link function and *F* is gamma distributed (Gilleskie & Mroz, 2004).

Belotti et al. (2015) outline the process in further detail, highlighting the process of how the two-part model permits zeros and nonzeros to be “generated by different densities as a special type of mixture model”. They write that the zeros are typically handles using a probability model of a positive outcome, modeled as follows:

where *x* is a vector of independent / explanatory variables, *δ* is the corresponding vector of parameters to be estimated, and *F* is the cumulative distribution function of i.i.d. error term, most frequently chosen to originate from normal (probit) or extreme value (logit) distributions. Deb et al. (2018) point out that the choice of logit or probit generally yield identical results, regardless of alternative specifications of the probability model.

For the conditional model (positive values), the model is usually represented as:

where *x* continues to represent a vector of explanatory variables, γ is the corresponding vector of parameters to be estimated, and *g* is the selected density function for positive values (*y|y>0*). The likelihood contribution for any given observation can be written as:

where *i(.)* represents the indicator function. Thus, the log-likelihood contribution is represented as:

The probability and conditional models can be estimated separately due to the fact that the γ and δ parameters are additively separable in the log-likelihood contribution for each record. The overall mean can thus be modeled as the product of these expectations:

Belotti et al. (2015) note that this function is derived from the first principles of statistics decomposition of a join distribution into extensive (marginal) and intensive (conditional) distributions, and is always true, with or without separability of *F* and *g(.)* specifications.

The positive values, or where *g(.)* denotes a density function are modeled using some regression framework for continuous outcome variables (e.g. OLS regression with GLM), with or without a transformation applied to positive values, with researchers more recently using the GLM framework to model *(y|y>0, x)* directly applying some nonlinear transformation of the linear index function (Deb & Norton, 2018). In that case:

where *g* is the link function in the generalized linear model (Deb & Norton, 2018).

Belotti et al. (2015) point out that a misconception existed, particularly in earlier research on two-part models, that the two-part model required the assumption of independence of binary outcomes that were conditional on positive, continuous outcomes, but prove that error terms in these two equations do not need to meet this assumption in order to obtain consistent parameter estimates for *δ* and *γ*.

Of the four main two-part modeling choices outlined in Deb et al.’s 2018 paper, we have already addressed the first three: 1) choice of logit or probit for the probability model, and for the second part of the model, 2) link function, and 3) distribution family. The team’s final decision comes down to specification of the linear index: whether to include interaction terms or allow flexible nonlinear specifications of any continuous variables, such as taking a curvilinear shape. Several authors advised that variables included in one part of the model be invariably included in the other part of the model, such that no variables are excluded in one part of the model that are included in the other model (Belotti et al., 2015; Deb & Norton, 2018).

Though several authors advocated the use of the GLM in the positive part of the model, Mullahy et al. (2024) point out that some applications (Jones & Yen, 2000; Yen & Jones, 1996, 1997) have also used the inverse hyperbolic sine and other power transformations.

### Three-part model

Recently, a paper by Karlsson et al. (2024), *Getting the right tails right*, advocated for a three part model, a model that would entail explicitly modeling value found in the the right tail.

[Add more about this if needed, or remove if not needed. Decide later.]

### Combinations

In practice, research has used a combination of modeling and transformations, with the familiar OLS for *ln(y)* or *log(y)* as a primary example (discussed previously) being used in single-index models but equally likely to be encountered in the second part of a two-part model (though a back transformation would be required to return units to the original scale) (Belotti et al., 2015). Belotti (2015) pointed out that the second part of two-part models (the conditional model for positive values) could be implemented with or without a transformation applied to *y|y>0*, as we saw in Deb et al.’s 2018 paper where they top-coded the dependent variable of health care expenditures at $50,000.

[Can include more examples here if needed, such as Kandilov]

### Difference-in-Differences

[Not sure where this will go or how to incorporate but have 2-3 pages saved for whereever I put this.]

# CHAPTER III: METHODOLOGY

The following list is a reminder of the research questions that will be answered by comparing the performance of the X models:

|  |  |
| --- | --- |
| Q1 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of raw bias percentage? |
| Q2 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of empirical standard errors? |
| Q3 | When applying a two-part model, which treatment of extreme values, topcoding, mean-adjusted topcoding, and median-adjusted topcoding, performs better under different sample size, proportion of extreme values, and magnitude of extreme values in terms of confidence interval coverage? |

## Introduction

Text

## Empirical Data

A real data situation modeling health care expenditures will be used with the candidate models and data scenarios. The data described in this section for the empirical study was acquired by the candidate based on descriptions by Deb et al (2018) contributing theoretical information on modeling health care expenditures with skewed distributions with a large mass at zero. One of the aims of the empirical study was to examine the effects of the young adult health insurance coverage expansion of the Patient Protection and Affordable Care Act (ACA) on health care expenditures. At the time, it was common practice for private insurers to drop nonstudent dependents at age 19 and student dependents at age 23; the young adult health insurance coverage expansion allowed dependents to remain on their parents’ private health insurance plan until they turn 26 years old. The provision took effect on September 23, 2010 (six months after the passage of the ACA), though, in practice, the expansion was implemented in January 2011 (corresponding to the first open enrollment period following that September). The authors wrote, “This expansion in coverage points to a powerful natural experiment using a difference-in-differences design”; treatment and control groups could be created from ages 23-25 and 27-29  respectively, and pre and post years could be created from 2008-2010 (pre-ACA) and 2011-2014 (post-ACA). The authors explored if people aged 23-25 were affected by the ACA policy using a treatment effect on the treated differences in differences model. Their article not only contributes an extension of prior analyses to test the effect of the ACA’s young adult expansion on different outcomes (including total health care expenditures) using nonlinear models but also “demonstrates current best econometric practice in modeling these kinds of outcomes, including those that have skewed distributions with a large mass at zero.” Deb et al. (2018) found that the ACA young adult expansion lowered health care expenditures, and that modeling the large mass of zeros through two-part models greatly improved the fit of the model and facilitated better understanding of the results.

Data Sources

To answer their research questions, Deb et al. required information on a large number of representative young American adults both younger and older than 26 years, with observations dated in the years before and after the rule change was implemented in 2010. To obtain accurate measures of health care expenditures in addition to detailed measures of health status and other observable member characteristics, the team utilized the Medical Expenditure Panel Survey (MEPS) (<https://meps.ahrq.gov/mepsweb/>). MEPS is a national survey that collects information regarding financing and use of medical care in the United States, and has been collected by the Agency for Healthcare Research and Quality (AHRQ), a federal government organization in the United States, every year since 1996. The data used in their examples was primarily sourced from the Household Component, which is drawn from a nationally representative subsample of households that participated in the prior year’s National Health Interview Survey, and contains data on a sample of families and individuals. From these data, AHRQ produces annual estimates for a variety of measures of health care expenditures and use, health status, health insurance coverage, and sources of payment for health services in the United States. The candidate was able to access this dataset; it will be used for the simulation.

### Data Processing

As the key independent variables for difference-in-differences analyses are indicators for treatment and control groups for pre and post periods, Deb et al (2018) assumed that adults aged 23-25 would be potentially affected by the ACA policy and were therefore in the treatment group. The control group comprised of adults 25-27 years old. Adults aged 26 years old were excluded due to partial coverage during the year. Per Ellis 2018, it is common (particularly for risk adjustment models) to be estimated on data after elimination of ‘troublesome’ records; this elimination often includes purging partial year (less than 12 month) eligibles, or, in prospective models, dropping records when the full year 12 months of prior-year claims are not available. ([Ellis, 2018](https://www.sciencedirect.com/science/article/pii/B9780128113257000038?via%3Dihub#fn5)) The pre period was defined as 2008 to 2010, and the post period defined as 2011 to 2014. The team was primarily interested in the effect of the ACA on young adults up to 26 (treatment effect on the treated). As such, they were able to compare those in the treatment group to those in the control group in the years before and after their policy change and evaluate various modeling techniques.

The final dataset used by the team contained 17,899 observations. MEPS data were restricted to observations of individuals who were in the scope of the survey design for the entire year and containing valid responses to family size and marital status items, using the last two as basic indicators of data reliability. The team excluded 2,529 observations with missing data for two health status variables (SF12 physical and mental health scales), explaining that they are important explanatory variables for their analysis.

### Variables

The dependent variable for modeling health care costs is the total annual health care expenditures. This includes out-of-pocket payments and third-party payments from all sources. It does not include insurance premiums, and is measured in nominal US dollars. In health care expenditures, total costs are often more variable than outpatient costs because of the relative rarity and high cost of hospital care. (Diehr, 1999) The distribution is highly skewed with a large mass at zero: more than one-third of the observations have zero expenditures, and less than 5% have expenditures exceeding $9,000. In 35 observations, values exceed $50,000 (with a maximum value of $2,226,997). The research team states that, “Although our statistical models are designed to account for skewness, they are not designed to take extreme values such as these into account”; rather than dropping the observations, they transform the raw data by topcoding the value of the expenditure for these 35 observations at $50,000.

The team used other covariates to adjust for demographics, education, poverty, and health status. The poverty category is scaled from 1 (poorest) to 5 (richest). Taking sampling weights into account, they found the datasets quite balanced across treated and control groups for several variables: 51% of the sample is female, and just under 20% of the population is Hispanic. In the control group, 13% are black, while 15% are black in the treated group. Physical and mental health statuses (derived from SF12 measures) range from 0 to 100, with higher scores indicating better health; weighted means of the physical and mental health scales were 51 and 51, respectively. Both distributions exhibited a left skew, with a median greater than the mean. They also included indicator variables for arthritis, asthma, cancer, high cholesterol, diabetes, high blood pressure, and presence of rare conditions, defined as any of five health conditions infrequently seen in this age group (angina, congestive heart disease, emphysema, myocardial infarction, and stroke).

### Descriptive Statistics

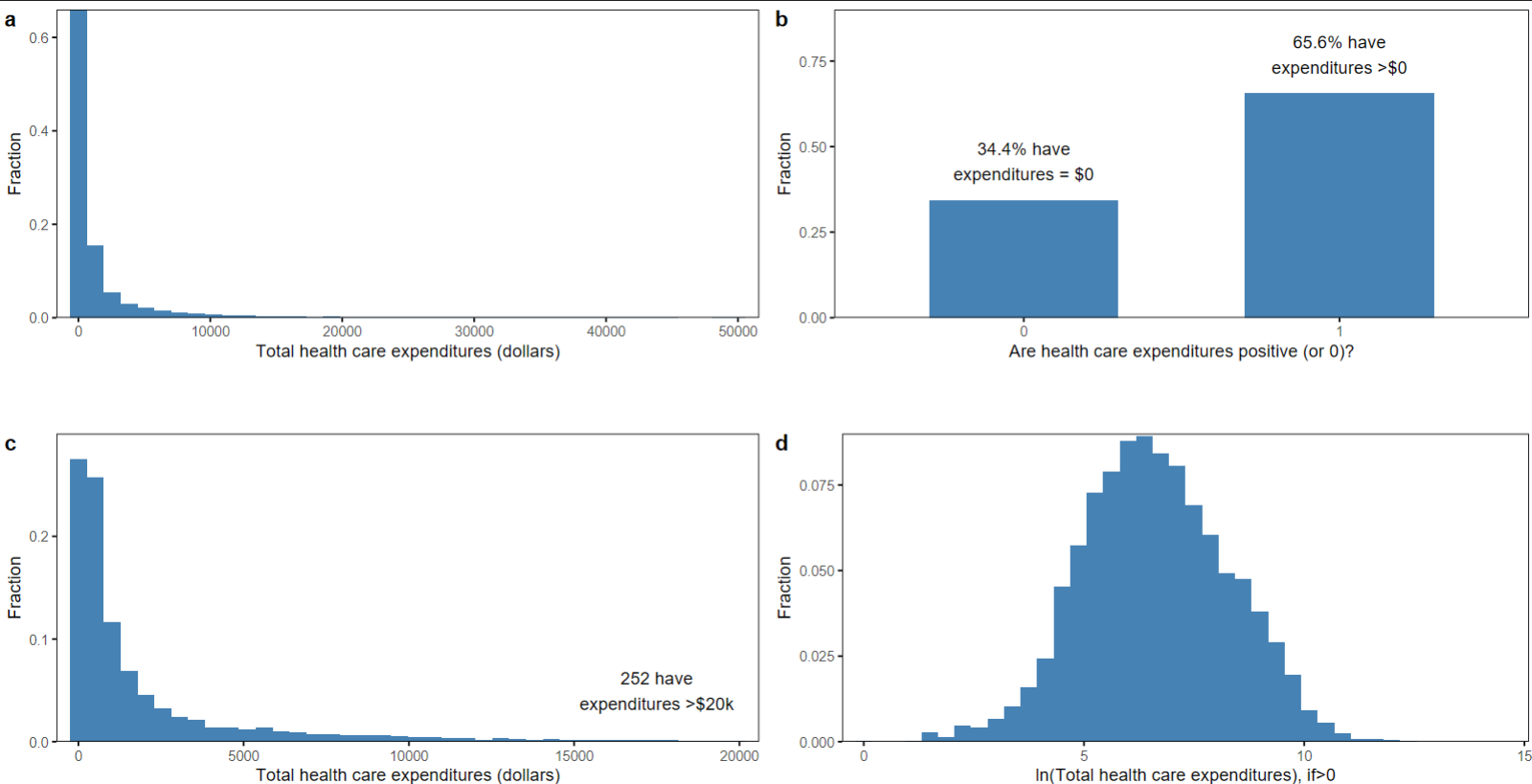
Summary statistics for the MEPS data used by Deb et al. (2018), with means of dependent and independent variables used in the analyses, stratified by control group (ages 27-29) and treatment group (ages 23-25), can be seen in TABLE 1 (XX).

Table 1: Caption

|  |  |  |
| --- | --- | --- |
| **Variables** | **Ages 27-29** | **Ages 23-25** |
| Total health care expenditures (US$) |  |  |
| Age (in years) |  |  |
| Female |  |  |
| Black |  |  |
| Hispanic |  |  |
| Married |  |  |
| Widowed, separated, divorced |  |  |
| High school education |  |  |
| Some college education |  |  |
| College graduate |  |  |
| Poverty category: Poor |  |  |
| Poverty category: Near poor |  |  |
| Poverty category: Low income |  |  |
| Poverty category: Middle income |  |  |
| Poverty category: High income |  |  |
| Physical health scale (0-100) |  |  |
| Mental health scale (0-100) |  |  |
| Arthritis |  |  |
| Asthma |  |  |
| Cancer |  |  |
| High cholesterol |  |  |
| Diabetes |  |  |
| High blood pressure |  |  |
| Rare diseases |  |  |
| Sample size N |  |  |

ADD MORE HERE

Figure 1: The distribution of total health care expenditures is displayed four ways: a) full histogram incluing zeros, b) the split between zeros and nonzeros, c) the histogram of just positive values, and d) the histogram of the natural log of positive values.



### Methodology

The team compared estimation and interpretation of the effect of the change in insurance policy on health care expenditures using OLS and a two-part model; we will be following their two-part model framework in analyzing the simulated data. To model these highly skewed expenditures, the team implemented generalized linear models, citing that health economists have increasingly been interested in applying GLMs to health care expenditures since their initial applications in the late 1990’s by Mullahy (1998) and Blough (1999). Using a GLM allowed them to address the heteroskedasticity issues inherent in these types of datasets by explicitly modeling heteroskedasticity, as “GLMs allow the variance of the outcome to be a function of its predicted value by the choice of an appropriate distribution family.”Though GLMs are often fit with log link (the natural logarithm of the expected value of the DV modeled as the linear index), the team checked the log link against several other functional form alternatives and found XX. The other main modeling challenge, the large proportion of observations with zero expenditures, was handled using a two-part model, first modeling the probability that a person has any health care expenditures with a logit model using the full sample, then estimating a GLM on the subset of observations with positive expenditures. The two-part model further allows for separate investigation of any potential effect of covariates on the extensive margin (logit model, if positive expenditure values) and on the intensive margin (GLM, amount of expenditures if any). No variables included in one part were excluded from the other.

* ATT will be reported: In nonlinear models, the interpretation of the interaction effect of two variables—such as between treatment and the implementation of the ACA as in this difference-in-differences study design—is complicated. Ai & Norton (1) showed that the full interaction effect, calculated by taking the double derivative with respect to both interacted variables, is not equal to the marginal effect of the change in just the interaction term. The full interaction effect can even be the opposite sign of the coefficient on the interaction term. However, Puhani (36) argued, using the potential outcomes framework, that the treatment effect on the treated in the difference-in-difference regression equals the expected value of the dependent variable for the treatment group in the post period with treatment compared with the hypothetical expected value of the dependent variable for the treatment group in the post period if they had not received treatment. In nonlinear models, the treatment effect on the treated equals the difference in two predicted values. It always has the same sign as the coefficient on the interaction term. Because we estimate many nonlinear models using a difference-in-differences study design, we report the treatment effect on the treated in all tables of results. (Deb, 2018)

The team employed the logit model (rather than the probit) for the first part of the two-part model, citing the choice as relatively inconsequential; specification tests for GLM supported the use of log link and gamma distribution. Specification tests for the preferred functional form of explanatory variables adjusting for demographics, education, and health did not reveal significant effects of these squared or interaction terms, and thus more variables were not added to the model.

## Simulation Plan

A simulation will be conducted to answer all the research questions by comparing the XX on a DiD outcome using the following methods to handle extreme values: topcoding, mean-adjusted topcoding, and median-adjusted topcoding. The purpose of the simulation is to compare a true model with XX balance to the selected missing methods under different sample sizes (XX, XX, and XX), varying proportion of extreme values (XX, XX, and XX), and introducing varying proportions of balance of the key predictor (xx, xx, and xx)  in terms of raw bias percentage, empirical standard errors (SE) and coverage Type I error rate. In the next section, a description of the simulation design will be outlined, followed by the process. Results will first start with a baseline model to serve as a benchmark: a two-part model with untransformed dependent variable.

Table : Conditions for The 3x3x3 Factorial Design of the Monte Carlo Simulation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Two-part model, treatment of DV at 99th percentile: | | | |
|  |  |  | 1 = No Treatment | 2 = Top-coding | 3 = Mean-preserved top-coding | 4 = Median-preserved top-coding |
| Sample Size | Magnitude of Extreme Value | Proportion of extreme value | Combination of Factorial Levels | | | |
| 1 = 1,000 | 1 | 1 | 1111 | 1112 | 1113 | 1114 |
| 1 = 1,000 | 1 | 2 | 1121 | 1122 | 1123 | 1124 |
| 1 = 1,000 | 1 | 3 | 1131 | 1132 | 1133 | 1134 |
| 1 = 1,000 | 2 | 1 | 1211 | 1212 | 1213 | 1214 |
| 1 = 1,000 | 2 | 2 | 1221 | 1222 | 1223 | 1224 |
| 1 = 1,000 | 2 | 3 | 1231 | 1232 | 1233 | 1234 |
| 1 = 1,000 | 3 | 1 | 1311 | 1312 | 1313 | 1314 |
| 1 = 1,000 | 3 | 2 | 1321 | 1322 | 1323 | 1324 |
| 1 = 1,000 | 3 | 3 | 1331 | 1332 | 1333 | 1334 |
| 2 = 2,000 | 1 | 1 | 2111 | 2112 | 2113 | 2114 |
| 2 = 2,000 | 1 | 2 | 2121 | 2122 | 2123 | 2124 |
| 2 = 2,000 | 1 | 3 | 2131 | 2132 | 2133 | 2134 |
| 2 = 2,000 | 2 | 1 | 2211 | 2212 | 2213 | 2214 |
| 2 = 2,000 | 2 | 2 | 2221 | 2222 | 2223 | 2224 |
| 2 = 2,000 | 2 | 3 | 2231 | 2232 | 2233 | 2234 |
| 2 = 2,000 | 3 | 1 | 2311 | 2312 | 2313 | 2314 |
| 2 = 2,000 | 3 | 2 | 2321 | 2322 | 2323 | 2324 |
| 2 = 2,000 | 3 | 3 | 2331 | 2332 | 2333 | 2334 |
| 3 = 10,000 | 1 | 1 | 3111 | 3112 | 3113 | 3114 |
| 3 = 10,000 | 1 | 2 | 3121 | 3122 | 3123 | 3124 |
| 3 = 10,000 | 1 | 3 | 3131 | 3132 | 3133 | 3134 |
| 3 = 10,000 | 2 | 1 | 3211 | 3212 | 3213 | 3214 |
| 3 = 10,000 | 2 | 2 | 3221 | 3222 | 3223 | 3224 |
| 3 = 10,000 | 2 | 3 | 3231 | 3232 | 3233 | 3234 |
| 3 = 10,000 | 3 | 1 | 3311 | 3312 | 3313 | 3314 |
| 3 = 10,000 | 3 | 2 | 3321 | 3322 | 3323 | 3324 |
| 3 = 10,000 | 3 | 3 | 3331 | 3332 | 3333 | 3334 |

## Research Question 1 Analysis Plan

## Research Question 2 Analysis Plan

## Research Question 3 Analysis Plan

## Chapter III Summary

# CHAPTER IV: RESULTS

## Simulation Overview

## Percent Difference Formulas

## Research Question 1 Raw Bias Percentage Results

### Condition X1 Raw Bias Percentage Results

### Condition X2 Raw Bias Percentage Results

### Condition X3 Raw Bias Percentage Results

## Research Question 2 Empirical Standard Errors Results

### Condition X1

### Condition X2

### Condition X3

## Research Question 3 XX Results

## Chapter IV Summary

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| Abbreviation | Term |
| ATE | Average Treatment Effect |
| CI | Confidence Interval |
| HCE | Health care expenditures |
| OLS | Ordinary Least Squares |
| SD | Standard Deviation |

# REFERENCES

Aihounton, G. B., & Henningsen, A. (2021). Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, *24*(2), 334-351.

Arrow, K. J., & Lind, R. C. (1970). Uncertainty and the Evaluation of Public Investment Decisions. *The American economic review*, *60*(3), 364-378. <https://unco.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwhV07T8MwED7RLrDwKEWUlsoDa4pfie2BgUerDkyIzlESOxIIUkSDEP-ecx5FLULMiRLrzr7vs333HYDgExpsxQQrTZIq3G6FQtpIZsIq7sWdEpHoXBi3Ka0HF21pjM-yrNIEq0t95Evpi7tEUFK4Ve5ARzNaa-L-BF_O1Jrxci-vV6cZboRa5IwNgMwOYN7-C3FkOXmyr-s66C1hxv_Gcwj7DYkk17XXj2DHFT3YbWuMVz3o3Cefx3C1QI9WN_7lF0kKS5Dskela3pssc1Kf2pFabcMfFJK7punOqg-L2fTxdh407RKCZ6RpZeB0omkmZZ5lQrlcylBwZ0PKMqUz66zgYeqXqAl17mhoUyly4_GaRSl1WooT6BbLwp0CcZS5MMq49m-pKE1FYjgSS2RzuddsHUC_skL8VktixI0JBjBqrRw3a2EVIwKaCGOqYgMY_n6sI4Mj0kqf_fHRIewxU7XHCSgbQbd8_3Dn0PXuGiP9vXkYV1PgG3sLs0Q>

Barber, J., & Thompson, S. (2004). Multiple regression of cost data: use of generalised linear models. *Journal of health services research & policy*, *9*(4), 197-204. <https://doi.org/10.1258/1355819042250249>

Basu, A., & Manning, W. G. (2009). Issues for the next generation of health care cost analyses. *Med Care*, *47*(7 Suppl 1), S109-114. <https://doi.org/10.1097/MLR.0b013e31819c94a1>

Basu, A., Manning, W. G., & Mullahy, J. (2004). Comparing alternative models: log vs Cox proportional hazard? *Health Econ*, *13*(8), 749-765. <https://doi.org/10.1002/hec.852>

Basu, A., & Rathouz, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, *6*(1), 93-109. <https://doi.org/10.1093/biostatistics/kxh020>

Belotti, F., Deb, P., Manning, W. G., & Norton, E. C. (2015). Twopm: Two-Part Models. *The Stata Journal*, *15*(1), 3-20. <https://doi.org/10.1177/1536867x1501500102>

Belsley, D. A., Welsch, R. E., & Kuh, E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley. <https://unco.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwdV1LC8IwDA46L3ryNXxN9gcmW1vX9SwOEbyI97GtnXiZB_8_mK7bUJnHUEgflCRfmnwFoGTnez82QTKRZhzh1p4yGbKcSk40uVNK06igQn1T67U9YjWXI1r4Z2V5H10_pmUB0WEK4vY-el8dEZFzm2ZBl8RE_dBVyXinEWZXvenoXDFYDlnDw9PIuh4aJ_nwM_EYLN17MIGeKqcwurSUqq8Z2Fd1NzWrpStNfZwemIMTH2-Hk6d1JXUyJmkWS4gNFgJ8tQCXF36mSWokRrqMKxGxXAYIHnmUySL15RLsbh2rfwNrGAYi8k2qYAODAq-ycsy2ttUBvQHt6Gwz>

Berk, M. L., & Monheit, A. C. (2001). The concentration of health care expenditures, revisited. *Health Aff (Millwood)*, *20*(2), 9-18. <https://doi.org/10.1377/hlthaff.20.2.9>

Bhattarai, G. (2013). *Understanding the outlier in healthcare expenditure data* 26th NorthEast SAS User Group Annual Meeting, <https://www.lexjansen.com/nesug/nesug13/116_Final_Paper.pdf>

Bicego, M., & Baldo, S. (2016). Properties of the Box–Cox transformation for pattern classification. *Neurocomputing (Amsterdam)*, *218*, 390-400. <https://doi.org/10.1016/j.neucom.2016.08.081>

Blough, D. K., Madden, C. W., & Hornbrook, M. C. (1999). Modeling risk using generalized linear models. *J Health Econ*, *18*(2), 153-171. <https://doi.org/10.1016/s0167-6296(98)00032-0>

Blough, D. K., & Ramsey, S. D. (2000). Using Generalized Linear Models to Assess Medical Care Costs. *Health services and outcomes research methodology*, *1*(2), 185. <https://doi.org/10.1023/A:1012597123667>

Briggs, A., & Gray, A. (1998). The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy*, *3*(4), 233-245. <https://doi.org/10.1177/135581969800300410>

Briggs, A., Nixon, R., Dixon, S., & Thompson, S. (2005). Parametric modelling of cost data: some simulation evidence. *Health Econ*, *14*(4), 421-428. <https://doi.org/10.1002/hec.941>

Buntin, M. B., & Zaslavsky, A. M. (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *J Health Econ*, *23*(3), 525-542. <https://doi.org/10.1016/j.jhealeco.2003.10.005>

Cantoni, E., & Ronchetti, E. (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *J Health Econ*, *25*(2), 198-213. <https://doi.org/10.1016/j.jhealeco.2005.04.010>

Chambers, R., Hentges, A., & Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society. Series A, Statistics in society*, *167*(2), 323-339. <https://doi.org/10.1111/j.1467-985X.2004.00748.x>

Chambers, R. L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, *81*(396), 1063-1069. <https://doi.org/10.1080/01621459.1986.10478374>

Chen, J., & Roth, J. (2024). Logs with Zeros? Some Problems and Solutions. *The Quarterly journal of economics*, *139*(2), 891-936. <https://doi.org/10.1093/qje/qjad054>

Chen, Y. H., & Zhou, X. H. (2006). Interval estimates for the ratio and difference of two lognormal means. *Stat Med*, *25*(23), 4099-4113. <https://doi.org/10.1002/sim.2504>

Choi, S.-W. (2009). The effect of outliers on regression analysis: regime type and foreign direct investment. *Quarterly journal of political science*, *4*(2), 153-165.

Choi, S.-W. (2022). Outlier analysis: Natural resources and immigration policy. *PloS one*, *17*(1), e0261533-e0261533. <https://doi.org/10.1371/journal.pone.0261533>

Cohen, S. B., & Uberoi, N. (2001). Differentials in the Concentration in the Level of Health Expenditures across Population Subgroups in the U.S., 2010. In *Statistical Brief (Medical Expenditure Panel Survey (US))*. <https://www.ncbi.nlm.nih.gov/pubmed/29708701>

Cots, F., Elvira, D., Castells, X., & Dalmau, E. (2000). Medicare’s DRG-weights in a European environment: the Spanish experience. *Health policy (Amsterdam)*, *51*(1), 31-47. <https://doi.org/10.1016/S0168-8510(99)00074-3>

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 829-844.

Dalton, K., & Norton, E. C. (2000). Revisiting Rogowski and Newhouse on the indirect costs of teaching: a note on functional form and retransformation in Medicare's payment formulas. *Journal of health economics*, *19*(6), 1027-1046.

De Brauw, A., & Herskowitz, S. (2021). Income variability, evolving diets, and elasticity estimation of demand for processed foods in Nigeria. *American journal of agricultural economics*, *103*(4), 1294-1313.

Deb, P., & Burgess, J. F. (2003). *A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures*. <https://unco.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV3PT8IwFH7R7eJJjBp_oHmJ5yG0W8dORAnEkwHCnbRrl5Dohsz9__ZtHQoHDp6btC9t8_X1_fg-AM56_eAAE3SYSBXb71bEQy3ClOuYEbmT5HKY8cTsU-s18bfSVVm2EFnjti5SCpk_D0RCXGuCs9HmKyARKUq2OkWNU_AZYbMH_uvkfbbYx9nIPR7Tc1DtOvYNKXpr_bnrgT4gZfy_LR3wZ3JjthdwYvJLWLzgvJLlOvhL4I_jneYgFhk2LckkqJUiKaJ9lGgdWGz6kpD6kpB4kCmlXdk_-RU8TSfL8VvQGrdy17Bc_VrGr8HLi9zcAJr-wEQiZcNIqzAWSnGZMOs1WlctI0LWW-gem-nu-PA9nNX1bnWUogve97YyD-DRBj-64_gBGKuk3Q>

Deb, P., & Norton, E. C. (2018). Modeling Health Care Expenditures and Use. *Annu Rev Public Health*, *39*, 489-505. <https://doi.org/10.1146/annurev-publhealth-040617-013517>

Diehr, P., Yanez, D., Ash, A., Hornbrook, M., & Lin, D. Y. (1999). Methods for analyzing health care utilization and costs. *Annu Rev Public Health*, *20*, 125-144. <https://doi.org/10.1146/annurev.publhealth.20.1.125>

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association*, *78*(383), 605-610. <https://doi.org/10.1080/01621459.1983.10478017>

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. *Journal of business & economic statistics*, *1*(2), 115-126. <https://doi.org/10.1080/07350015.1983.10509330>

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1984). Choosing Between the Sample-Selection Model and the Multi-Part Model. *Journal of business & economic statistics*, *2*(3), 283-289. <https://doi.org/10.1080/07350015.1984.10509396>

Faddy, M., Graves, N., & Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value Health*, *12*(2), 309-314. <https://doi.org/10.1111/j.1524-4733.2008.00421.x>

Fox, J. (1991). Regression Diagnostics. Newberry Park. In: CA: Sage Publications.

French, E., & Kelly, E. (2016). Medical Spending around the Developed World. *Fiscal studies*, *37*(3-4), 327-344. <https://doi.org/10.1111/j.1475-5890.2016.12127>

Geurts, K., Bruijnzeels, M., & Schokkaert, E. (2022). Do we care about high-cost patients? Estimating the savings on health spending by integrated care. *Eur J Health Econ*, *23*(8), 1297-1308. <https://doi.org/10.1007/s10198-022-01431-3>

Gilleskie, D. B., & Mroz, T. A. (2004). A flexible approach for estimating the effects of covariates on health expenditures. *J Health Econ*, *23*(2), 391-418. <https://doi.org/10.1016/j.jhealeco.2003.09.008>

Goldberger, A. S. (1968). The interpretation and estimation of Cobb-Douglas functions. *Econometrica: Journal of the Econometric Society*, 464-472.

Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. *Int J Qual Health Care*, *23*(3), 331-341. <https://doi.org/10.1093/intqhc/mzr010>

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. Stata press.

Holle, M., Wolff, T., & Herant, M. (2021). Trends in the Concentration and Distribution of Health Care Expenditures in the US, 2001-2018. *JAMA Netw Open*, *4*(9), e2125179. <https://doi.org/10.1001/jamanetworkopen.2021.25179>

Ibrahim, E., Shouman, M. A., Torkey, H., & El-Sayed, A. (2021). Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system. *Multimedia tools and applications*, *80*(13), 20125-20147. <https://doi.org/10.1007/s11042-021-10727-0>

Jones, A. M., Lomas, J., & Rice, N. (2014). Applying beta‐type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, *29*(4), 649-670.

Jones, A. M., & Yen, S. T. (2000). A Box–Cox double‐hurdle model. *The Manchester School*, *68*(2), 203-221.

Kandilov, A., Keyes, V., van Hasselt, M., Sanders, A., Siegfried, N., Stone, R., Edwards, P., Collins, A., & Brophy, J. (2018). The Impact of the Vermont Support and Services at Home Program on Healthcare Expenditures. *Cityscape (Washington, D.C.)*, *20*(2), 7-18. <https://unco.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1JT-UwDLbgcUFC7Ggem3xAcyvTNkmXEwIEggtCiEFwekrrVAJpWpZy4N9jdwPeYYQ4N1GjxLE_x_ZnABXu-96UTogiImXzgNgeGypsyLA7NpQ643ScNLzan6j12ve3ly7LsleRjd6mKpcn8z_st8RSUGvCg8cnT5pISbC166gxC3Oh6OYRzB2dXFxeDR6Y1Ey3YWZ2o_34tk9C_AIvpwzQTNBbm9MlyPqFsdGp9u_p31A0PcXi-PPFL8Nih0XxsBWeFZhx5SpsDiUs-Bvb4l1suUTe1uCOpQrPm8JKrApk8Ig3rNqrskZpD8pQHm1J2CsgtDVKH3a8bLPAsCrxbEg4Q6FZloj5K7v86_D39OT6-MzrmjN4D4EytUdRFluTZL5NdGakAZaLpZ-VyY0ly7jMpX7hCkYEOuXvfh7oICtslKc8RalIbcCClST-sm6K_egXIBVJmKuCQSNluoiSxCnlKNaaImIYS2PYaI5r8tgycUxCIb9nh28M2_1GT7o7-DL52OUx7H0-02F6EwdlgKy1mG0eFnxn2HFHni6kAfXm__-8BfOMsJL2zWYbRvXzq9uBkUjPbiecuzB7cZS-A6529ro>

Kapitula, L. R. (2015). When Two Are Better Than One: Fitting Two-Part Models Using SAS.

Karlsson, M., Wang, Y., & Ziebarth, N. R. (2024). Getting the right tail right: Modeling tails of health expenditure distributions. *J Health Econ*, *97*, 102912. <https://doi.org/10.1016/j.jhealeco.2024.102912>

Leone, A. J., Minutti-Meza, M., & Wasley, C. E. (2019). Influential Observations and Inference in Accounting Research. *The Accounting review*, *94*(6), 337-364. <https://doi.org/10.2308/accr-52396>

Malehi, A. S., Pourmotahari, F., & Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Econ Rev*, *5*, 11. <https://doi.org/10.1186/s13561-015-0045-7>

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of health economics*, *17*(3), 283-295.

Manning, W. G., Basu, A., & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ*, *24*(3), 465-488. <https://doi.org/10.1016/j.jhealeco.2004.09.011>

Manning, W. G., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics*, *20*(4), 461-494.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *Am Econ Rev*, *77*(3), 251-277. <https://www.ncbi.nlm.nih.gov/pubmed/10284091>

Marazzi, A. (2002). Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational statistics & data analysis*, *39*(4), 503-528.

Marazzi, A., & Barbati, G. (2003). Robust parametric means of asymmetric distributions: estimation and testing. *Estadistica*, *54*(162-163), 47-72.

Marazzi, A., Paccaud, F., Ruffieux, C., & Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical care*, *36*(6), 915-927.

Marazzi, A., Villar, A. J., & Yohai, V. J. (2009). Robust response transformations based on optimal prediction. *Journal of the American Statistical Association*, *104*(485), 360-370.

Metcalf, L., & Casey, W. (2016). *Cybersecurity and applied mathematics* (1 ed.). Syngress is an imprint of Elsevier. <https://doi.org/10.1016/C2015-0-01807-X>

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Econ*, *20*(8), 897-916. <https://doi.org/10.1002/hec.1653>

Mitchell, E. M. (2020). Concentration of Healthcare Expenditures and Selected Characteristics of High Spenders, U.S. Civilian Noninstitutionalized Population, 2019. In *Statistical Brief (Medical Expenditure Panel Survey (US))*. <https://www.ncbi.nlm.nih.gov/pubmed/35696512>

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ*, *17*(3), 247-281. <https://doi.org/10.1016/s0167-6296(98)00030-7>

Mullahy, J. (2009). Econometric Modeling of Health Care Costs and Expenditures: A Survey of Analytical Issues and Related Policy Considerations. *Medical care*, *47*(7), S104-S108. <https://doi.org/10.1097/MLR.0b013e31819c9593>

Mullahy, J., & Norton, E. C. (2024). Why Transform Y? The Pitfalls of Transformed Regressions with a Mass at Zero. *Oxford Bulletin of Economics and Statistics*, *86*(2), 417-447. <https://doi.org/https://doi.org/10.1111/obes.12583>

Murphy, S. M. E., Hough, D. E., Sylvia, M. L., Dunbar, L. J., & Frick, K. D. (2018). Key Design Considerations When Calculating Cost Savings for Population Health Management Programs in an Observational Setting. *Health Serv Res*, *53 Suppl 1*(Suppl Suppl 1), 3107-3124. <https://doi.org/10.1111/1475-6773.12832>

O'Hagan, A., & Stevens, J. W. (2003). Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Econ*, *12*(1), 33-49. <https://doi.org/10.1002/hec.699>

Osborne, J. W. (2010). Improving Your Data Transformations: Applying the Box-Cox Transformation. *Practical assessment, research & evaluation*, *15*(12), 12. <https://unco.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1LT8MwDLZguyAhnkMMxpQ_0LE2aZpyQWzqNHFACMG5SutU4sA6WJHGv8dpszF24MC5rhI5jv3Z8QOAB4Oht6UTpETkOveR7HGIhQ4IdkchxiY0IlJ1X-2N1npN_G3hsixXKrLW21jmNmR-HXApVd0v6nb-7tkhUvax1U3U2IV2YHVzC9qj5OHxae2BcW6H5P5ObK51r3QGZXII2Wptsivl4BXf1nXRW40a_7-_IzhwcJPdNfJxDDtmdmInNbusjlOYrqMK7IuImc0YZdUGmiWpvGEWq9p6KEZwkY3KpTcul1tUHXiZJM_jqeemK3jG9-PKE0jGmywkuYDcBFqHPBoqnSk_ypGcEB_DQppIK8FzlZkACyUyrskd8TMh6U9-BvvaZuHPqrpaD8-BEVGQ84JQH2aikEoZzmkFIVAi4VDsQsfyPp03nTTS5D7mXMZ-F3orLqbuDi3SHxZe_P35EvaaN30bGOlBq_r4NFfQsufXdxLQr_3ib2kbxrc>

Paddock, S. M., Wynn, B. O., Carter, G. M., & Buntin, M. B. (2004). Identifying and Accommodating Statistical Outliers When Setting Prospective Payment Rates for Inpatient Rehabilitation Facilities. *Health Services Research*, *39*(6p1), 1859-1880. <https://doi.org/10.1111/j.1475-6773.2004.00322.x>

Park, R. E. (1966). Estimation with Heteroscedastic Error Terms. *Econometrica*, *34*(4), 888-888. <https://doi.org/10.2307/1910108>

Pirson, M., Dramaix, M., Leclercq, P., & Jackson, T. (2006). Analysis of cost outliers within APR-DRGs in a Belgian general hospital: Two complementary approaches. *Health policy (Amsterdam)*, *76*(1), 13-25. <https://doi.org/10.1016/j.healthpol.2005.04.008>

Polgreen, L. A., & Brooks, J. M. (2012). Estimating incremental costs with skew: a cautionary note. *Appl Health Econ Health Policy*, *10*(5), 319-329. <https://doi.org/10.2165/11632430-000000000-00000>

Randahl, D., & Vegelius, J. (2024). Inference with Extremes: Accounting for Extreme Values in Count Regression Models. *International studies quarterly*, *68*(4). <https://doi.org/10.1093/isq/sqae137>

Randall P. Ellis, B. M., Sherri Rose,. (2018). Chapter 3 - Risk Adjustment for Health Plan Payment. In R. C. v. K. Thomas G. McGuire (Ed.), *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets*. Academic Press. <https://doi.org/B978-0-12-811325-7.00003-8>.

Rogowski, J. A., & Newhouse, J. P. (1992). Estimating the indirect costs of teaching. *Journal of health economics*, *11*(2), 153-171.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (Vol. 3). Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Ruppert, D. (2001). Statistical Analysis, Special Problems of: Transformations of Data. In P. B. B. Neil J. Smelser (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 15007-15014): Pergamon.

Sullivan, J. H., Warkentin, M., & Wallace, L. (2021). So many ways for assessing outliers: What really works and does it matter? *Journal of business research*, *132*, 530-543. <https://doi.org/10.1016/j.jbusres.2021.03.066>

Weichle, T., Hynes, D. M., Durazo-Arvizu, R., Tarlov, E., & Zhang, Q. (2013). Impact of alternative approaches to assess outlying and influential observations on health care costs. *Springerplus*, *2*, 614. <https://doi.org/10.1186/2193-1801-2-614>

Welsh, A. H., & Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *60*(2), 413-428. <https://doi.org/10.1111/1467-9868.00133>

Wicklin, R. (February 8, 2017). Winsorization: The good, the bad, and the ugly. *The DO Loop*. <https://blogs.sas.com/content/iml/2017/02/08/winsorization-good-bad-and-ugly.html>

Yen, S. T., & Jones, A. M. (1996). Individual cigarette consumption and addiction: a flexible limited dependent variable approach. *Health economics*, *5*(2), 105-117.

Yen, S. T., & Jones, A. M. (1997). Household consumption of cheese: an inverse hyperbolic sine double‐hurdle model with dependent errors. *American journal of agricultural economics*, *79*(1), 246-251.

1. Authors reviewed studies published in *The Accounting Review, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies,* and *Contemporary Accounting Resarch*. The studies they surveyed include (but not limited to): auditing, analysts’ forecasts, management compensation, earnings management, conservatism, taxes, disclosure, and the earnings-returns relation. [↑](#footnote-ref-1)