NBER WORKING PAPER SERIES

GETTING THE RIGHT TAIL RIGHT: MODELING TAILS OF HEALTH EXPENDITURE DISTRIBUTIONS

Martin Karlsson
Yulong Wang
Nicolas R. Ziebarth

Working Paper 31444
http://www.nber.org/papers/w31444

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2023, revised June 2024

We thank the editor Maarten Lindeboom, two anonymous referees, Anirban Basu, Rita Ginja, John Mullahy, Edward Norton, and Lei Liu for their extremely helpful comments and suggestions. Further, we thank all participants of the Annual Health Econometrics Workshop at Emory University and at the 2022 Essen Health Conference for very helpful comments. We also thank representatives of the German Association of Private Health Insurers for invaluable help with the private insurer claims dataset. We are grateful to Sarah McNamara for copyediting this manuscript. We do not have financial interests that would constitute any conflict of interest with this research. Generous funding by the German Federal Ministry of Education and Research (FKZ: 01EH1602A) is gratefully acknowledged. All the remaining errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Getting the Right Tail Right: Modeling Tails of Health Expenditure Distributions
Martin Karlsson, Yulong Wang, and Nicolas R. Ziebarth
NBER Working Paper No. 31444
July 2023, revised June 2024
JEL No. C10,C13,I10,I13

ABSTRACT

Health expenditure data almost always include extreme values, implying that the underlying distribution has heavy tails. This may result in infinite variances as well as higher-order moments and bias the commonly used least squares methods. To accommodate extreme values, we propose an estimation method that recovers the right tail of health expenditure distributions. It extends the popular two-part model to develop a novel three-part model. We apply the proposed method to claims data from one of the biggest German private health insurers. Our findings show that the estimated age gradient in health care spending differs substantially from the standard least squares method.

Martin Karlsson
CINCH, University of Duisburg-Essen
Weststadttürme Berliner Platz 6-8
45127 Essen
Germany
martin.karlsson@uni-due.de

Nicolas R. Ziebarth
ZEW Mannheim
D-68161, Mannheim
Germany
and ZEW Mannheim
nicolas.ziebarth@zew.de

Yulong Wang
Syracuse University
Maxwell School of Citizenship and Public Affairs
Economics Department
127 Eggers Hall
Syracuse, NY 13244
ywang402@syr.edu

1 Introduction

Around the world, it is a stylized fact that roughly 50% of total population health care spending falls on 5% of the sickest individuals in society (French and Kelly, 2016; Karlsson, Klein, and Ziebarth, 2016; Finkelstein, 2020). Large proportions of extreme values imply that the underlying expenditure distributions feature heavy right tails (Handel, Kolstad, and Spinnewijn, 2019). With such heavy tails, the population variance and higher-order moments such as population skewness and kurtosis, could be very large and even infinite, thus violating the assumptions of ordinary least squares (OLS) estimation. Such distributional features may lead to poor finite sample performance.

If a distribution has a heavy tail, generally, random samples drawn from this distribution likely include very large values. A simple and popular approach is to treat extreme values as outliers and trim them or top code them. However, these extreme values are neither classic outliers nor measurement errors. Hence, simply deleting or ignoring them implies potentially ignoring valuable information. What's more, the researcher then ignores precisely those individuals who are responsible for the lion's share of per capita health care spending. 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.

As an alternative to trimming, the existing literature has developed various methods to accommodate extreme values, such as taking the logarithm or modeling higher-order moments. In the next section, our literature review provides more details and discusses our proposed method in the context of existing estimation approaches. All existing methods combine extreme values with the remaining data and focus on the *mid-sample* features of the underlying distribution, such as the mean, median, and mid-sample quantiles. These mid-sample features are indeed of high relevance in applied research, and extreme values are usually a nuisance for estimating them. As stated in Mullahy (2009),

> "[...] 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."

Our paper follows this lead. We study situations where the (population) variance and other higher-order moments of health expenditure distributions are potentially infinite. On the other hand, researchers may be interested in the tail features themselves – extreme quantiles and marginal effects for individuals whose health care spending is in the right tail. Such features are of importance to policymakers and health insurers; for example, to assess how population aging affects the level and distribution of health care spending (cf. Karlsson et al., 2016; Karlsson, Iversen, and Øien, 2018). Hence, we propose separating the right tail of the data and explicitly studying extreme values. We do so by using both simulations and high-quality claims data which contain 620 thousand policyholders from one of the biggest German private health insurers. Note that large insurance pools are essential when the focus is on the top percentiles of spenders.

Specifically, we study heavy tails as follows. In the first step, using log-rank-log-size plots, we show that the tails of our claims data exhibit clear features of a Pareto distribution. This implies a highly nonlinear relationship between individual i's predictors X_i and their health care spending Y_i. Moreover, we estimate the Pareto exponent, which characterizes the heaviness of the tail of the underlying distribution. We find that the Pareto exponent is around 2 in our data, implying that the finite second moment condition of OLS is very likely violated, leading to poor performance of OLS and t-tests. Using simulations, we then study the behavior of OLS under the Pareto heavy tail. Moreover, we show that the Pareto and heavy tail features bias OLS estimates of the coefficients and marginal effects due to strong non-linearities implied by the Pareto distribution. It also results in rejection errors when conducting inference.

In the next step, we propose an alternative method that leads to unbiased estimation and asymptotically correct statistical inference. To do so, we exploit the Pareto tail feature and introduce a maximum likelihood estimator (MLE) for the pseudo-true parameter and, more importantly, the marginal effects. This method was initially proposed and studied by Wang and Tsai (2009) and Wang and Li (2013) in the statistics literature. We tailor their approach to the health expenditure context and benchmark it against a simple linear specification. Further, we incorporate our method into the widely used two-part model (cf., Manning, 1998; Mullahy, 1998), which employs a binary outcome model along with a conditional model for positive spending. We propose a novel *three-part model* by incorporating our tail MLE as the third part. We also provide applied users with a cookbook recipe of the various steps to implement our method.

After that, using German claims data, we estimate marginal effects and calculate standard errors for exogenous spending predictors such as age and gender, both for the standard OLS estimator and for our proposed approach. We provide explicit evidence on the relevance of extreme expenditures for the robustness of OLS estimation. In line with the literature, we confirm that OLS is sensitive to extreme values. Further, consistent with our simulation results, we find that the OLS point estimates of the age-spending nexus lie below the marginal effects of our proposed method. In other words, the estimates differ along the entire age distribution from age 35 to 75.

The paper is organized as follows. Section 2 reviews the existing literature and summarizes our contribution. Section 3 first previews the heavy tail features in our data; then introduces our proposed method that explicitly accommodates the heavy tail; and finally extends the two-part model to develop a novel three-part model. Section 4 introduces our claims dataset and presents descriptive statistics. Section 5 contains empirical results. In particular, using Monte Carlo simulations, we first show that the commonly used least squares method performs poorly when data exhibit heavy tails. Second, we apply the proposed method and present the empirical findings. The mathematical details, additional simulation results, and robustness tests are in the Appendix.

2 Literature and Contribution

This paper speaks to a large literature in health economics. It complements the existing toolbox for studying heavy tail features of health expenditure data. For a clear comparison, we briefly categorize the existing methods into distinct groups and discuss them individually. More comprehensive overviews can be found in Jones (2011), Manning (2012), and Mihaylova, Briggs, O'Hagan, and Thompson (2011).

The first group of approaches modifies the data by taking the *logarithm* of Y_i so that its extreme values no longer dominate the estimation result. The generalized linear model (GLM) with a log-link function is a widely used approach (Mullahy, 1998; Manning and Mullahy, 2001; Manning, Basu, and Mullahy, 2005; Deb, Norton, and Manning, 2017). 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 (Manning and Mullahy, 2001; Buntin and Zaslavsky, 2004). However, modeling the logarithm of Y_i as a (linear) function of X_i could introduce additional bias, the magnitude of which depends on the unknown distribution of Y_i. In comparison, we show below that once the data exhibit a linear feature in a log-log plot,[1] the Pareto tail feature is reasonably satisfied. Hence, modeling the Pareto exponent as a function of X_i becomes a natural choice. We present extensive simulation exercises which suggest that taking the logarithm of Y_i instead could lead to a significant bias in this scenario.

On a side note, issues related to the practice of using log-transformed versions of the main dependent variable have resurfaced in recent methodological literature. Two papers by Chen and Roth (2024) and Mullahy and Norton (2024) discuss the consequences of incorporating observations close to or equal to zero in such approaches. The main conclusion from both papers is that results tend to be very sensitive to how observations with zeros are handled; therefore, other modeling approaches are preferred. As noted, our paper provides further reasons to be cautious about log transformations.

Instead of modifying the data, a second group of approaches assumes some *parametric density* of Y_i that accommodates heavy tails. For example, Manning et al. (2005) propose using the generalized gamma distribution, and Jones, Lomas, and Rice (2014) propose using the generalized beta of the second kind (GB2) distribution, which covers the Pareto distribution and the Burr-Singh-Maddala distribution as special cases. Using Monte Carlo simulations, Jones, Lomas, and Rice (2015) and Jones, Lomas, Moore, and Rice (2016) evaluate the empirical performance of a range of different empirical techniques for modeling the distribution of health expenditures. One of their performance indicators is how accurately they represent the right tail. While all these methods model the *whole* distribution, we focus on modeling the *tail* part. As our method uses the Pareto tail approximation, which does not necessarily hold for the entire distribution, it entails more robustness to misspecification originating in the non-tail part. In addition, exploiting the Pareto approximation safeguards against

[1] A log-log plot displays the natural logarithm of an observation's rank as a function of the natural logarithm of its value; cf. Figure 1 below.

misspecification of the distribution within the right tail.

A third group of approaches *nonparametrically* estimates the density and moments of expenditures conditional on covariates. In addition to the standard kernel and sieves methods, Gilleskie and Mroz (2004) propose a flexible estimator of the conditional density of expenditures within a number of set intervals. These nonparametric estimators typically require a large sample size; hence their performance in the tail might not be satisfactory. In comparison, our proposed method takes advantage of the Pareto tail approximation. Hence it can be considered semiparametric. Monte Carlo simulations show that our estimator performs well in finite samples.

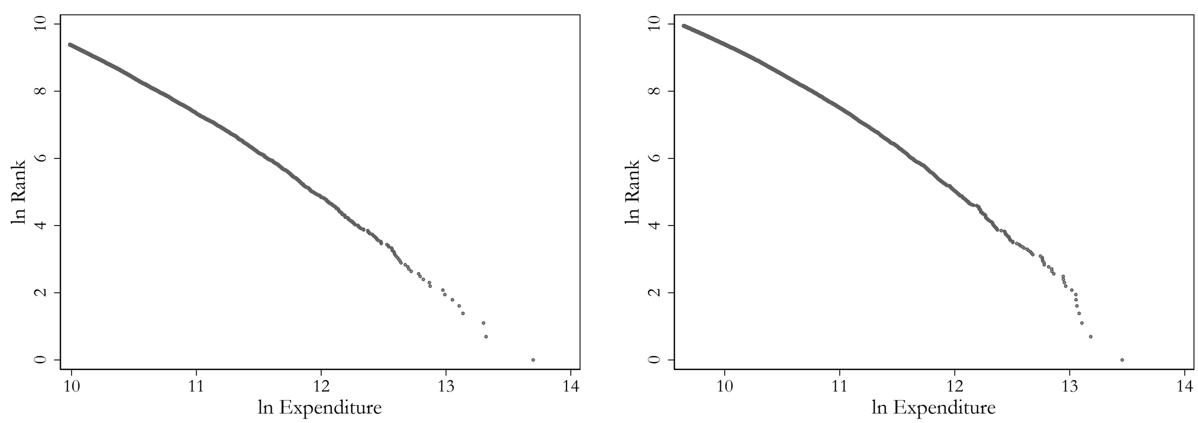
Another feature of health expenditure data is the large proportion of individuals with zero expenditures. The *two-part model* employs a binary outcome model for the extensive margin along with a conditional model for positive spending (Newhouse and Phelps, 1976; Manning, Newhouse, Duan, Keeler, and Leibowitz, 1987; Mullahy, 1998). As mentioned, we extend the canonical two-part model and propose a *three-part model*. Accordingly, our three-part model essentially divides the positive spending part into extreme values ($Y_i > y_{\min}$ for some tail cutoff $y_{\min}$) and non-extreme (but still positive) values ($Y_i \in (0, y_{\min})$). For the extreme values, we propose using the Pareto tail approximation and tail index regressions (Wang and Tsai, 2009). For the non-extreme values, we propose using linear regressions since Y_i now has compact support. Then, we combine all three parts, i.e. $Y_i = 0$, $Y_i \in (0, y_{\min})$, and $Y_i > y_{\min}$ to estimate the overall mean conditional on X_i. Note that our three-part model is different from the four-part model proposed by Duan, Manning, Morris, and Newhouse (1982). We focus on the distribution of Y_i itself and divide it based on cutoffs of Y_i. In contrast, Duan et al. (1982) define the parts based on types of health care used, distinguishing nonusers, ambulatory-only users, and inpatient users. Therefore their four parts are based on additional covariates.

We note that there also exists a rich literature on how to model heavy tailed distributions in the field of risk management and insurance. Recently published papers, among many others, include Yuen, Stoev, and Cooley (2020) who propose a novel inference method to estimate Value-at-Risk statistics using a Pareto-type tail assumption. Goegebeur, Guillou, Pedersen, and Qin (2022) estimate conditional tail moments also using Pareto-type tail assumptions and apply the method to reinsurance ratings. By contrast, Mao, Stupfler, and Yang (2023) use Pareto-type tail assumptions to estimate a novel risk measure that is based on behavioral economic theory. Most papers in this field, including the ones just referenced, assume that the Pareto exponent is a constant that does not vary across covariates. In comparison, Goegebeur, Guillou, and Qin (2021) use a nonparametric kernel estimator proposed by Goegebeur, Guillou, Ho, and Qin (2023) to estimate reinsurance premiums and allow the Pareto exponent to depend on covariates.

3 Modeling Pareto Tails

3.1 Preview of the Pareto Tail

We start by presenting the Pareto tail feature in our health care claims data. Let Y_i denote health care expenditures of individual i for $i = 1, \ldots, n$, where n denotes the total sample size. Also, let $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ be the descending and ordered expenditure values whose ranks are accordingly $1, 2, ..., n$. Figure 1 plots the natural logarithms of the rank i against $\ln Y_{(i)}$ for the largest 5% of all values. We separately show the plots for females (left) and males (right).



(a) Females

(b) Males

Figure 1: Rank-Size Plots of Natural Logarithms of Rank against Expenditures

Notes: The left graph shows plots for females and the right graph shows plots for males. See Section 4 for more details about the German health care claims data.

Both Figure 1a (Females) and b (Males) suggest a linear fit in the rank-size plots. As has been extensively shown (e.g., Gabaix, 2009), this pattern implies that the underlying distribution exhibits a Pareto tail, or equivalently, the power law. More specifically, if Y_i has a Pareto distribution beyond some cutoff value $y_{\min}$, we have that

$$\mathbb{P}\left(Y_i > y | Y_i \geq y_{\min}\right) = \left(\frac{y}{y_{\min}}\right)^{-\alpha}, \tag{3.1}$$

where α is called the *Pareto exponent*, a positive parameter that uniquely characterizes the heaviness of the tail. Given the Pareto tail assumption, the slope of the linear fit in the rank-size plot is equal to $-\alpha$. Furthermore, the parameter $y_{\min}$ determines the cutoff location of the tail, above which the Pareto distribution serves as a good approximation of the underlying distribution. We discuss such Pareto tails from two perspectives.

First, from a theoretical perspective, it has been established in the statistics literature (e.g., Smith, 1987) that many commonly used distributions can be well approximated by Pareto distributions as long as one focuses on a sufficiently far tail region, that is, by considering a sufficiently large $y_{\min}$. Examples include the Student-t, the F, and the Cauchy distributions, among many others.[2]

Accordingly, we can treat $y_{\min}$ as a *tuning parameter* that determines the precision of the Pareto tail approximation. Note that this concept is close in spirit to the choice of the bandwidth parameter in nonparametric kernel estimations. In practice, we set $y_{\min}$ as the top 5% (equivalently 95%) percentile and present robustness checks with alternative cutoffs in Appendix A1.

Second, from an empirical perspective, the Pareto tail has been widely documented in many other datasets in economics and finance, such as stock returns, city size, firm size, and income, see Gabaix (2009, 2016) for reviews of other datasets that exhibit Pareto tails.

The presence of a Pareto tail causes two problems for the standard OLS method: One is a bias in estimating the marginal effect due to the strong non-linearity implied by the Pareto distribution. The other is a potentially infinite variance due to the heavy tail.

Regarding non-linearity, let X_i denote a vector of exogenous individual expenditure predictors such as age and gender. As the Pareto distribution (3.1) is uniquely characterized by the exponent α, $\alpha = \alpha(X_i)$ captures the effect of X_i on Y_i in the *tail*. The power function $y^{\alpha(X_i)}$ naturally generates a

[2] In particular, the Pareto exponent α is equal to the degree of freedom when the underlying distribution is Student-t.

nonlinear effect of X_i on the expected value of Y_i. Conversely, a model based on a linear specification, such as a simple OLS, could produce substantially biased results. We demonstrate such bias in a simple simulation study in Section 5.1 below.

Regarding infinite variances, the slope in Figure 1 is around -2, implying that the underlying distribution of expenditures has a very heavy tail. In particular, the tail of the Pareto distribution (3.1) is heavier with a smaller α. Moreover, the Pareto distribution implies that for any $r > 0$ (Mikosch, 1999):

$$\mathbb{E}[Y_i^r] < \infty \text{ if } r < \alpha \text{ and } \mathbb{E}[Y_i^r] = \infty \text{ if } r > \alpha.$$

Accordingly, when α is less than two, the tail is so heavy that $\mathbb{E}\left[Y_i^2\right]$ becomes infinite! Recall that the asymptotic normality of the OLS estimator and the t-statistic requires that the second moment of Y_i is finite, that is, $\mathbb{E}\left[Y_i^2\right] < \infty$. The heavy tail feature in the data then compromises this population moment condition, even though the sample variance is always defined. Hence the OLS estimator, the standard t-test, and estimates of any other higher-order moments such as skewness and kurtosis may perform poorly. To evaluate such effect in finite samples, we run simulation studies in Section 5.1.

We remark that the classic definition of bias, i.e., the expected value of the estimator minus the true value, might not be well-defined when the underlying distribution has a heavy tail. Specifically, the first order moment becomes infinite if $\alpha \leq 1$, given the regular variation condition above. In this situation, estimators such as OLS are inconsistent and the estimation result is meaningless. An advantage of our proposed method is that it does not suffer from this issue, although we find empirically that $\alpha \in (1, 2)$ is more relevant. For readability, we maintain the terminology of a "bias" throughout the paper (unless it would cause obvious confusion). In the simulation study below, we use the mean absolute deviation as a measure of "bias", which is well-defined regardless of the value of α.

3.2 The Proposed Maximum Likelihood Estimator

Given the failure of OLS under a Pareto tail, we move forward to construct a valid alternative that explicitly accommodates extreme values. For illustrative purposes, we preview the new approach in this subsection by assuming an exact Pareto tail. In fact, the econometric derivation only requires an approximate Pareto tail, which holds for many commonly used distributions such as Student-t, F, Beta, *et cetera*. For reasons of readability, we relegate the technical details and the primitive assumptions to Appendix A3.

Our method is based on the tail index regression proposed by Wang and Tsai (2009). First, assuming Y_i has an exact Pareto tail above $y_{\min}$, we obtain

$$\mathbb{P}\left(Y_i > y | Y_i > y_{\min}, X_i = x\right) = \left(\frac{y}{y_{\min}}\right)^{-\alpha(x)}, \tag{3.2}$$

where $\alpha\left(x\right)$ is the Pareto exponent that depends on the characteristics $X_i = x$. We adopt the model $\alpha = \exp\left(X_i'\beta_0\right)$, where β_0 again denotes the pseudo-true coefficient. The exponential function guarantees that the Pareto exponent is always positive, and the linear index form is adopted mainly for computational simplicity. Using only the observations above $y_{\min}$, we then obtain the following negative log-likelihood function:

$$\mathcal{L}\left(\beta\right) = n^{-1}\sum_{i=1}^{n}\left\{\exp(X_i'\beta)\log\left(Y_i/y_{\min}\right) - X_i'\beta\right\}\mathbf{1}\left[Y_i > y_{\min}\right], \tag{3.3}$$

where $\mathbf{1}\left[\cdot\right]$ denotes the indicator function. Then, the maximum likelihood estimator (MLE) of β_0 is

$$\hat{\beta} = \arg\max_{\beta}\mathcal{L}\left(\beta\right).$$

We can estimate its asymptotic variance as

$$\hat{\Sigma}_\beta = \left(n_0^{-1}\sum_{i=1}^{n}X_iX_i'\mathbf{1}\left[Y_i > y_{\min}\right]\right)^{-1}, \tag{3.4}$$

where $n_0 = \sum_{i=1}^{n}\mathbf{1}\left[Y_i > y_{\min}\right]$ denotes the total number of tail observations.

We provide two remarks about the proposed MLE. First, the Pareto tail assumption (3.2) implies that the conditional expectation of health expenditures beyond $y_{\min}$ is

$$\mathbb{E}\left[Y_i | Y_i > y_{\min}, X_i = x\right] = y_{\min} \frac{\alpha\left(x'\beta_0\right)}{\alpha\left(x'\beta_0\right) - 1}. \tag{3.5}$$

Again, the tail cutoff $y_{\min}$ is a tuning parameter chosen by the econometrician. In our subsequent analysis, we use the 95% quantile as $y_{\min}$. Appendix A3 provides more details about the choice of this parameter.

Given the Pareto tail (3.2), the marginal effect of X_i on tail expenditures is

$$\begin{aligned} M\left(x;\beta_0\right) &\equiv \frac{\partial \mathbb{E}\left[Y_i | Y_i > y_{\min}, X_i = x\right]}{\partial x} \\ &= -y_{\min} \frac{\exp\left(x'\beta_0\right)}{\left(\exp\left(x'\beta_0\right) - 1\right)^2} \beta_0 \\ &= -\frac{\mathbb{E}\left[Y_i | Y_i > y_{\min}, X_i = x\right]}{\left(\exp\left(x'\beta_0\right) - 1\right)} \beta_0, \end{aligned} \tag{3.6}$$

which we estimate by replacing β_0 with our MLE $\hat{\beta}$ in (3.3).

As seen, the marginal effect is a function of X_i. Hence the proposed estimator allows for a nonlinear impact of individual characteristics, such as age, on expected health care expenditures in the tail. This may be a desirable feature as average health care spending increases with age at a faster rate for seniors. In contrast, an OLS specification assumes that the marginal effect is constant, unless higher-order polynomials are included in the regression equation. We emphasize that the non-linearity is not generic but specifically due to the Pareto tail. One could include higher-order and interaction terms in $\alpha(x)$ for more flexibility. It should also be noted that whenever the model includes several independent variables, the assumption is that their interaction effects are non-zero (apart from the special case where the marginal effect is zero).

Using the Delta method, we can construct the standard errors for the marginal effects. In particular, for the marginal effect of the jth component of X_i, we have that

$$
\begin{aligned}
\nabla_j M(x, \beta_0) &\equiv \left. \frac{\partial M(x;\beta)}{\partial \beta_j} \right|_{\beta=\beta_0} \\
&= -y_{\min} \left[e_j \frac{\exp(x'\beta_0)}{(\exp(x'\beta_0)-1)^2} \right. \\
&\quad \left. -2x \frac{\exp(2x'\beta_0)}{(\exp(x'\beta_0)-1)^3} \beta_{0j} + x \frac{\exp(x'\beta_0)}{(\exp(x'\beta_0)-1)^2} \beta_{0j} \right],
\end{aligned}
$$

where e_j denotes the jth standard unit vector. The estimate of the standard error is then

$$
\hat{\Sigma}_{M_j} = \nabla_j M\left(x, \hat{\beta}\right)' \hat{\Sigma}_\beta \nabla_j M\left(x, \hat{\beta}\right). \tag{3.7}
$$

In summary, we propose the following steps:

1. Given $y_{\min}$, say the 95% percentile of Y_i, select all Y_i's that are larger than $y_{\min}$.
2. Construct the MLE by numerically solving (3.3) and estimating the standard error using (3.4).
3. Estimate the marginal effect (3.6) and the standard error (3.7).
4. Perform robustness check by using different $y_{\min}$.
5. Generate the counterfactual of the conditional tail expectation using (3.5).

Choice of $y_{\min}$. We close this subsection by discussing the choice of $y_{\min}$. As a tuning parameter, the econometrician chooses the tail threshold $y_{\min}$ which likely matters for the results, e.g. due to differences in statistical power, especially when the sample size is only moderate. However, the optimal selection of $y_{\min}$ is challenging. It has stimulated a large literature in statistics and econometrics (e.g., Danielsson, de Haan, Peng, and de Vries, 2001; Guillou and Hall, 2001): On the one hand, a large $y_{\min}$ ensures that the tail Pareto approximation performs well, and hence the bias is small. On the other hand, a small $y_{\min}$ ensures enough tail observations for asymptotic normality, and hence the variance is small. This bias-variance trade-off is inherent for the choice of $y_{\min}$ (and equivalently the number of tail observations n_0). See de Haan and Ferreira (2006, Chapter 3) for more discussions about the effect of the threshold on the bias and variance.

It turns out that a theoretically optimal choice of $y_{\min}$ does not exist—if no other condition is imposed on the true underlying distribution (Müller and Wang, 2017). Therefore, we recommend to vary $y_{\min}$ for sensitivity analysis. In our analysis of real claims data, we follow this strategy and contrast our baseline $y_{\min}$ to alternatives that are larger and smaller, as reported below.

In addition, Wang and Tsai (2009) provide a data-driven method of choosing $y_{\min}$; however, its theoretical properties need further investigation. If $Y|Y > y_{\min}, X = x$ is exactly Pareto distributed, $\exp(X'\beta_0)\ln(Y/y_{\min})$ is standard exponentially distributed conditional on $Y > y_{\min}$. Therefore, $\exp\{-\exp(X'\beta_0)\ln(Y/y_{\min})\}$ is standard uniform on $[0, 1]$. If $y_{\min}$ is too small, the distribution of this term deviates substantially from the standard uniform distribution. Hence the Pareto approximation becomes poor. If $y_{\min}$ is too large, the variance of our estimator becomes large due to few tail observations.

To implement this, define $\hat{U}_i = \exp\{-\exp(X'\hat{\beta})\ln(Y/y_{\min})\}$ where $\hat{\theta}$ is the estimator (3.3) using the threshold $y_{\min}$. We can measure the deviation from uniform distribution by

$$\hat{D}(y_{\min}) = n_0^{-1}\sum_{i=1}^{n}\left(\hat{U}_i - \hat{F}_n(\hat{U}_i)\right)^2 \times \mathbf{1}[Y_i > y_{\min}],$$

where $\hat{F}_n(\cdot)$ is the empirical distribution of $\{\hat{U}_i : Y_i > y_{\min}\}$. The data-driven choice of $y_{\min}$ is then selected as $y^*_{\min} = \arg\min_{y_{\min}} \hat{D}(y_{\min})$. Note that this data-driven method is designed for estimating the coefficients in the tail but not the marginal effect. In our empirical analysis below, we add a specification with a data-driven threshold as an additional robustness check.

3.3 Extension to a Three-Part Model

So far, we have focused on the tail solely using observations $Y_i > y_{\min}$. In this subsection, we generalize the previous analysis to model the whole distribution and extend the existing two-part model (cf., Mullahy, 1998) to a three-part model.[3]

In particular, the widely used two-part model is designed to capture that many observations of Y_i are zero. To model this, consider

[3] We thank Anirban Basu and Edward Norton for proposing this extension.

$$\mathbb{E}[Y_i|X_i = x] = \mathbb{P}(Y_i > 0|X_i = x) \times \mathbb{E}[Y_i|Y_i > 0, X_i = x], \tag{3.8}$$

provided that $Y_i \geq 0$ almost surely.

In the first part, the binary outcome $1[Y_i = 0]$ is fitted with a standard logit or probit model. Then, the partial effect on $\mathbb{P}(Y_i > 0|X_i = x)$ of X_i is estimated. In the second part, regressions of Y_i (or $\ln Y_i$) on X_i are conducted. The overall marginal effect $\partial \mathbb{E}[Y_i|X_i = x]/\partial x$ is obtained by combining the estimates from both parts. Given the Pareto tail, we can extend (3.8) and propose the following three-part model:

$$\begin{aligned} \mathbb{E}[Y_i|X_i = x] \quad &= \quad \mathbb{E}[Y_i|0 < Y_i \leq y_{\min}, X_i = x] \times \mathbb{P}[0 < Y_i \leq y_{\min}|X_i = x] \\ &\quad + \mathbb{E}[Y_i|Y_i > y_{\min}, X_i = x] \times \mathbb{P}[Y_i > y_{\min}|X_i = x]. \end{aligned} \tag{3.9}$$

Three-Part Model. In the first part, we estimate the conditional probabilities $\mathbb{P}[0 < Y_i \leq y_{\min}|X_i = x]$ and $\mathbb{P}[Y_i > y_{\min}|X_i = x]$ by running some pseudo maximum likelihood estimation. More specifically, denote $Y^* = 0, 1, 2$ if $Y_i = 0, Y_i \in (0, y_{\min}), Y_i > y_{\min}$, respectively. Imposing some parametric function F_j for $j = 0, 1, 2$ on the conditional choice probability, we can estimate some pseudo true parameter θ by fitting

$$\mathbb{P}(Y_i^* = j|X_i = x) = F_j(x; \theta).$$

For example, the multinomial logistic (MNL) regression assumes

$$\mathbb{P}(Y_i^* = j|X_i = x) = \frac{\exp(x'\theta_j)}{1 + \sum_{j=0}^{2} \exp(x'\theta_j)}, \tag{3.10}$$

for $j = 1, 2$ and θ_0 is understood as zero for normalization. Denote the estimated coefficient $\hat{\theta}_j$. In the second part, we run a linear regression of Y_i on X_i with observations $Y_i \in (0, y_{\min})$. Denote the regression coefficient as $\hat{\gamma}$. Given the upper bound $y_{\min}$, there is no need to employ a transformation such as $\ln Y_i$. In the third part, we implement the MLE method as described in the previous subsection.

Combining all three parts, we then estimate the conditional expectation by

$$
\begin{aligned}
\hat{\mathbb{E}}[Y_i|X_i = x] &= x'\hat{\gamma} \times \frac{\exp(x'\hat{\theta}_1)}{1 + \sum_{j=0}^{2} \exp(x'\hat{\theta}_1)} \\
&\quad + y_{\min} \frac{\exp(x'\hat{\beta})}{\exp(x'\hat{\beta}) - 1} \times \frac{\exp(x'\hat{\theta}_2)}{1 + \sum_{j=0}^{2} \exp(x'\hat{\theta}_2)}.
\end{aligned}
$$

Finally, we obtain the partial effect $\partial\mathbb{E}[Y|X = x]/\partial x$ by taking the derivative and obtain the standard error by bootstrapping.

Compared with the two-part model, our three-part model implies more marginal effects, on both the extensive and intensive margins. Specifically, here the extensive margin corresponds to the marginal effects on the conditional probabilities, that is, $\partial\mathbb{P}(Y_i^* = j|X_i = x)/\partial x$. The intensive margin corresponds to the marginal effects on the conditional moments of Y_i given that $Y_i \in (0, y_{\min})$ or given that $Y_i > y_{\min}$. The extensive margin is estimated via the standard MNL regression or other existing methods for ordered discrete choice models. The intensive margin at the non-tail part ($Y_i \in (0, y_{\min})$) is estimated via standard OLS.

We speculate that the principal component that differentiates our three-part model from the widely used two-part model lies in the intensive margin at the tail part ($Y_i > y_{\min}$), especially in our example. As health expenditure data typically exhibit heavy tails, its intensive margin in the tail part becomes non-negligible or may even dominate the overall marginal effect. See Section 5.2 for such comparison in our data analysis.

Limitations and Extensions. In the existing two-part model (3.8), it is assumed that $\mathbb{P}(Y_i > 0|X_i = x)$ is represented by a parametric binary probability model such as logit or probit. Additionally, it is assumed that $\mathbb{E}[Y_i|Y_i > 0, X_i = x]$ follows a linear or log-linear function of x. For further details, refer to Mullahy (1998) and Manning (1998).

In a similar fashion, our three-part model assumes that (i) $\mathbb{P}(Y_i \in (0, y_{\min})|X_i = x)$ and $\mathbb{P}(Y_i > y_{\min})|X_i = x)$ are governed by a parametric multinomial logit model, (ii) that $\mathbb{E}[Y_i|Y_i > y_{\min}, X_i = x]$ is governed by the Pareto tail as in (3.5), and (iii) that $\mathbb{E}[Y_i|Y_i \in (0, y_{\min}), X_i = x]$ is linear in x. Note that this model is different from the standard ordered logit model in which the thresholds need to be estimated. In contrast, our thresholds are determined at zero and $y_{\min}$. Obviously, these assumptions are stronger than those for the two-part model.

Specifically, the MNL model (3.10) is valid under the Independence of Irrelevant Alternatives (IIA) assumption. In our framework, IIA implies that the probability of one category such as $\mathbb{P}(Y_i^* = 2|X_i) = \mathbb{P}(Y_i > y_{\min}|X_i)$ does not impact the relative probabilities of other alternatives, that is,

$$\frac{\mathbb{P}(Y_i^* = 1|X_i)}{\mathbb{P}(Y_i^* = 0|X_i)} = \frac{\mathbb{P}(Y_i \in (0, y_{\min})|X_i)}{\mathbb{P}(Y_i = 0|X_i)}.$$

This IIA property implicitly restricts the underlying distribution of Y_i given X_i. For example, if the probability of incurring extreme health expenditures increases, it is very likely that the probability of incurring zero expenditure decreases by a larger margin than the probability of incurring moderate expenditures. More precise characterizations of such restrictions depend on both the tail and the non-tail parts of the underlying distribution of Y_i given X_i, which is unknown.

There are a number of alternatives to the MNL model which do not require the IIA assumption. One such estimator is the ordered generalized extreme value (OGEV) model proposed by Small (1987). This OGEV model relaxes the IIA condition at the cost of adding more parameters. Our preferred alternative is, however, a sequential logit specification (cf. Tutz, 1991) due to its simplicity. It estimates $\mathbb{P}(Y_i > 0|X_i = x)$ and $\mathbb{P}(Y_i > y|Y_i > 0, X_i = x)$ in separate steps and hence requires a slight reformulation of equation (3.9) and all the derivations that follow. But all extensive margins can be estimated in an analogous manner. A comparison thus provides a natural test of the robustness to the IIA assumption. See also Greene and Hensher (2010, Chapter 3) for reviews of other methods.

Generalizations and extensions of this three-part model raise a deeper question: how to reconcile the tail and non-tail parts in a data generating process? An extensive study of this question is beyond the scope of this paper, but we would like to end this section with a heuristic discussion. In principle, we aim for the conditional distribution $F_{Y|X=x}(y)$ and its functionals such as conditional mean and marginal derivative. Flexible estimation can be done by modeling $Y_i = m(X_i, U_i)$ for some unknown function $m(\cdot, \cdot)$ and resorting to nonparametric methods. The function $m(\cdot, \cdot)$ could be even nonadditive to accommodate the nonlinearity (e.g., Matzkin, 2003). These fully nonparametric methods are robust to mis-specifications at the cost of requiring large samples. However, extreme values are sparse by construction, making fully nonparametric methods less applicable. This observation motivates our MLE based on the Pareto tail approximation.

Specifically, our underlying model assumption is that the tail of $\mathbb{P}(Y_i > y|X_i = x)$ can be well approximated by the Pareto distribution (see (A3.1) in Appendix A3). This assumption is satisfied by a broad class of distributions such as Student-t, F, and Cauchy distributions, with their shape parameters being functions of X. Therefore, the robustness of our Pareto tail approximation lies in the fact we leave the non-tail part unspecified. Now combining the non-tail part, the three-part model requires additional assumptions beyond the tail. One could again use nonparametric methods to model $F_{Y|Y<y_{\min},X=x}(y)$ and maintain a high level of robustness. The MNL or other models serve as parsimonious simplifications of the non-tail part, which inevitably involve additional assumptions. In summary, we consider our proposed three-part model as one of the potential extensions to the popular two-part model, one that is specifically designed to accommodate the heavy tail feature of health expenditure data.

4 Data

This section describes the claims data used in this paper. The main working sample focuses on individuals privately insured in the German health care system. It is important to note that the policyholders do not have supplemental private insurance, but rather *comprehensive* long-term health insurance throughout their life cycles until death. For more details on the German two-tier health care system and German private health insurance, please refer to Atal, Fang, Karlsson, and Ziebarth (2023).

The claims data are administrative records on the universe of insurance plans and claims between 2005 and 2011 from one of the largest private health insurers in Germany. In total, our dataset includes more than 2.6 million enrollee-year observations from 620 thousand unique policyholders along with detailed information on plan parameters such as premiums, claims, and diagnoses. Atal, Fang, Karlsson, and Ziebarth (2019) provide more details about the dataset. The data also contain the age and gender of all policyholders as well as their occupational group. We convert all monetary values to 2016 U.S. dollars (USD).

Sample Selection. We focus on primary policyholders. In other words, we disregard insured children and those who are younger than 25 years (555,690 enrollee-year observations).[4] Moreover, due to a 2009 portability reform (Atal et al., 2019), we disregard inflows after 2008 (253,325 enrollee-year observations). The final sample consists of 1,867,465 enrollee-year observations from 362,783 individuals.

Descriptive Statistics. Table 1 presents the descriptive statistics. The mean age of the sample is 45.5 years. The oldest policyholder is 99 years old. 34% of the sample are high-income employees, 49% are self-employed and 13% are civil servants. The majority of policyholders (72 percent) are male because women are underpresented among the self-employed and high-income earners in Germany. On average, policyholders have been clients of the insurer for 13 years and have been enrolled in their current health plan for 7 years.[5] The majority of individuals join private insurance around the age of 30, when most Germans have fully entered the labor market but are still healthy and thus charged moderate premiums in this risk-rated market (risk rating is only imposed at contract inception and all subsequent premium increases are community rated).

Table 1 shows that the average *annual premium* is \$4,749 and slightly lower than the average premium for a single plan in the U.S. group market at the time (Kaiser Family Foundation, 2019). Note that the *annual premium* is the total premium—including employer contributions for privately insured high-income earners.[6] The average *deductible* is \$675 per year.

In terms of benefits covered, we simplify the rich data and focus on a plan generosity indicator provided by the insurer. It classifies plans into three coverage tiers: *TOP*, *PLUS*, and *ECO* plans.

[4]Children obtain their own individual risk-rated policies. However, if parents purchase the policy within two months of birth, no risk rating applies. Under the age of 21, insurers do not have to budget and charge for old-age provisions.

[5]Our insurer doubled the number of clients between the 1980s and 1990s and thus has a relatively young enrollee population, compared to all privately insured in Germany. Gotthold and Gräber (2015) report that a quarter of all privately insured are either retirees or pensioners.

[6]Employers cover roughly one-half of the total premium and the self-employed pay the full premium.

ECO plans are the lowest coverage tier; they lack coverage for services such as single rooms in hospitals and treatments by a leading senior M.D. For *ECO* and *PLUS* plans, a 20% coinsurance rate applies if enrollees see a specialist without a referral from their primary care physician. About 38% of all policyholders have a *TOP* plan, 34% a *PLUS* plan, and 29% an *ECO* plan. Because these plan characteristics have mechanical effects on claim sizes and correlate with policyholders' age, we control for them in our estimation of health care costs.

Table 1: Summary Statistics: German Claims Panel Data

	Mean	SD	Min	Max	N
Health Plan Parameters					
Total Claims (USD)	3,289	8,577	0	2,345,126	1,867,465
Annual premium (USD)	4,749	2,157	0	33,037	1,867,318
Deductible (USD)	675	659	0	3,224	1,867,465
Annual risk penalty (USD)	157	453	0	21,752	1,867,465
TOP Plan	0.377	0.485	0.0	1.0	1,867,465
PLUS Plan	0.338	0.473	0.0	1.0	1,867,465
ECO Plan	0.285	0.451	0.0	1.0	1,867,465
Socio-Demographics					
Age (in years)	45.5	11.4	25.0	99.0	1,867,465
Female	0.276	0.447	0.0	1.0	1,867,465
Policyholder since (years)	6.5	5.0	1.0	40.0	1,867,465
Client since (years)	12.8	11.0	1.0	86.0	1,867,465
Employee	0.336	0.473	0.0	1.0	1,867,465
Self-Employed	0.486	0.500	0.0	1.0	1,867,465
Civil Servant	0.132	0.338	0.0	1.0	1,867,465
Health Risk Penalty	0.358	0.480	0.0	1.0	1,867,465
Pre-Existing Condition Exempt	0.016	0.126	0.0	1.0	1,867,465

Source: German Claims Panel Data. *Policyholder since* is the number of years since the policyholder has enrolled in her current plan; *Client since* is the number of years since the client joined the insurer. *Employee* and *Self-Employed* are dummies for the policyholders' current occupation. *Health Risk Penalty* is a dummy that is one if the initial underwriting led to a health-related risk penalty on top of the factors age, gender, and type of plan; *Pre-Existing Conditions Exempt* is a dummy that is one if the initial underwriting led to exclusions of pre-existing conditions. The mutually exclusive dummies *TOP Plan*, *PLUS Plan* and *ECO Plan* capture the generosity of the plan. *Annual premium* is the annual premium, and *Annual Risk Penalty* is the amount of the health risk penalty charged. *Deductible* is the deductible and *Total Claims* the sum of all claims in a calendar year. See Section 4 for further details.

5 Results

This section presents empirical results. Section 5.1 conducts Monte Carlo studies to evaluate the performance of OLS and GLS when data have heavy tails. Section 5.2 examines the claims data applying our proposed method and the standard OLS method.

5.1 Monte Carlo Simulation Studies

5.1.1 Ordinary Least Squares

The Effect of Pareto Tails on Coefficient Bias. We now perform a simple simulation study to illustrate the effect of the Pareto tail. First, we focus on the potential bias of the OLS estimator due to the nonlinearity. To this end, we generate Y_i from the standard Pareto distribution (3.1) such that

$$\mathbb{P}(Y_i > y | Y_i > y_{\min}, X_i = x) = \left(\frac{y}{y_{\min}}\right)^{\alpha(x)},$$

where X_i is an independent draw from the absolute value of the standard normal distribution. We set $\alpha(x) = \exp(1 + x\beta_0)$ with $\beta_0 = 1$ as the pseudo-true parameter. This setup guarantees that $\alpha(X_i)$ is always positive. To capture the Pareto distribution in the tail, we set $y_{\min} = 1.645$, which is the 95% quantile of the standard normal distribution. This setup implies that Y_i has a normal distribution in the non-tail part and a Pareto tail. [7] Moreover, the minimum value of $\alpha(x)$ is $\exp(1) = 2.718 > 2$, implying that the variance of Y_i, given X_i, is always finite. Therefore, the potential bias of the OLS method could only originate from misspecification due to nonlinearities in the tail, as we will see in Figures 2 and 3 below.

The Pareto distribution implies that $\mathbb{E}[Y_i|X_i = x] = y_{\min}\alpha(x)/(\alpha(x) - 1)$. Then the marginal effect of X_i on the average of Y_i is

$$\frac{\partial \mathbb{E}[Y_i|X_i = x]}{\partial x} = -y_{\min}\frac{\alpha(x)}{(\alpha(x) - 1)^2}\beta_0,$$

[7] The tail is *not* invariant to an additive transformation like, e.g., the amount of spending above a uniform deductible. However, such transformations are of limited relevance when studying the right tail of health expenditures.

This is the main object of interest. When β_0 is positive, a larger x leads to a larger $\alpha(x)$ and hence a thinner tail. Then, accordingly, the expectation of Y_i conditional on being in the tail is smaller.

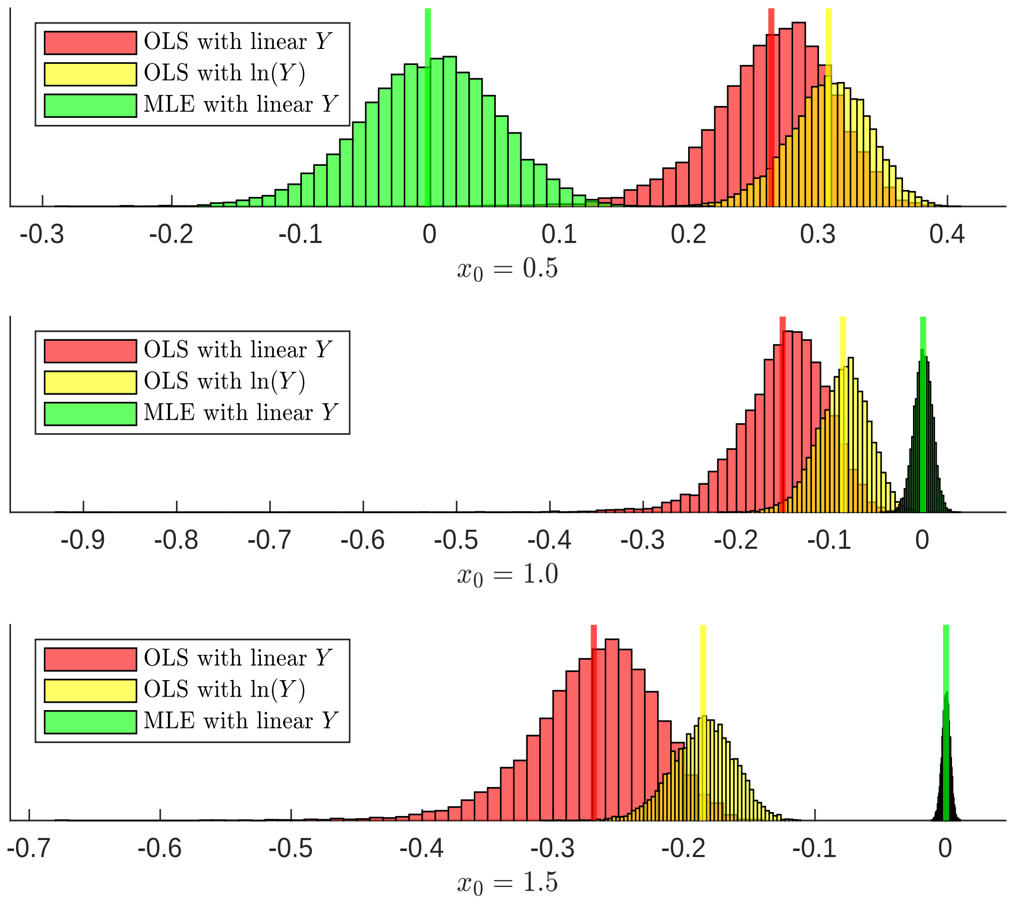
To estimate the marginal effect, we implement three methods. First, we use the standard OLS estimator, regressing Y_i on X_i (and a constant). The OLS coefficient estimates the marginal effect. By construction, such an estimated marginal effect is constant regardless of which value of X_i we condition on (red in Figures 2 and 3). Second, we regress $\ln Y_i$ on X_i (and a constant) to obtain the coefficients $(\hat{\beta}_0, \hat{\beta}_1)$. Given the logarithm, the marginal effect of X_i on Y_i evaluated at $X_i = x_0$ is then estimated as $\exp(\hat{\beta}_0 + x_0\hat{\beta}_1)\hat{\beta}_1\overline{\exp(\hat{u})}$, where $\overline{\exp(\hat{u})}$ is the average of the exponential of the residuals (cf. Duan, 1983); hence, the estimated marginal effect (yellow in Figures 2 and 3) varies with X_i even though β does not. Third, we implement our proposed MLE method (green in Figures 2 and 3).

Figure 2 depicts the histograms of the OLS estimators—the true marginal effect subtracted from $\hat{\beta}_1$ where 0 indicates no bias. The figure shows the results of the first method using Y_i (red color) as well as the second method using $\ln Y_i$ (yellow color) and our proposed MLE (green color). The histograms are based on $n_0 = 500$ observations in each simulation and 10,000 simulation draws. The top/middle/bottom panel corresponds to the marginal effect evaluated at $x_0 = 0.5/1/1.5$ and $\beta_0 = 2$.

We find the following: It is evident that the OLS estimator is substantially biased—regardless of whether we use Y_i or $\ln Y_i$; moreover, none of these misspecified estimators dominates the other. By contrast, our proposed MLE estimator is unbiased. Here, the bias is due to the fact that the marginal effect is highly nonlinear in X_i, while the OLS method specifies a linear model. We emphasize that such a bias exists only in the tail but not necessarily below $y_{\min}$ where a linear model is more reasonable and OLS could still perform well. Therefore, we consider our proposed method as a useful complement for studying tail features of heavily skewed distributions such as medical spending.

Next, we repeat the previous analysis with data generated from the same process as in Figure 2. We maintain that $x_0 = 2$, but now vary $\beta_0 = 0, 0.5, 1$. The histograms of the OLS estimators with Y_i and $\ln Y_i$ and our proposed MLE are in Figure 3. In the top panel, where $\beta_0 = 0$, we know that—by construction—X_i does not have any effect on Y_i. Therefore $\mathbb{E}[Y_i|X_i]$ is linear in X_i. In this scenario, the OLS method does not suffer from any misspecification due to nonlinearity. Hence, the histograms are basically identical, as expected. As β_0 increases from zero to one when moving to the bottom panel

Figure 2: Histograms of OLS with Linear or Natural Logarithms of Y and the proposed MLE

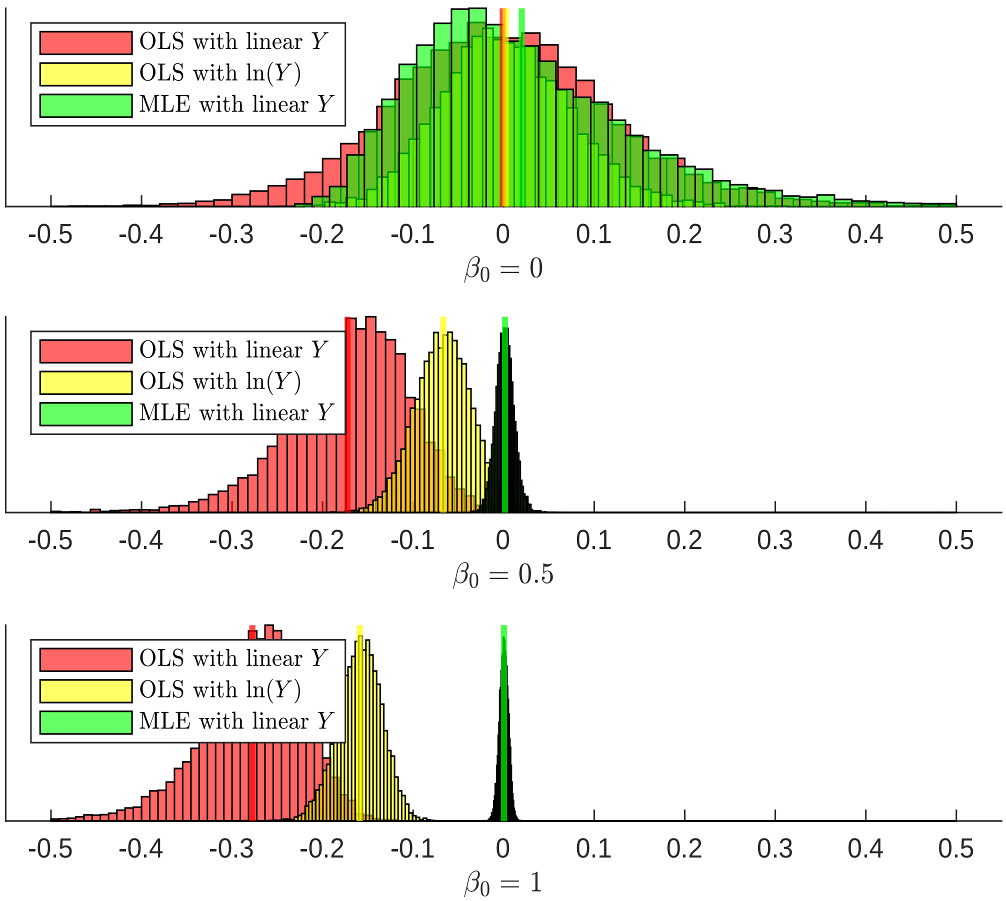


Notes: This figure depicts the histograms of the OLS estimator (subtracted by the true value) with Y_i (red color) or $\ln Y_i$ (yellow color) and the MLE (green color) for a marginal effect evaluated at $x_0 = 0.5, 1, 1.5$ and $\beta_0 = 2$. The vertical line depicts the averages of the estimators. Results are based on 10,000 simulation draws. See the main text for more details about the data generating process.

in Figure 3, the nonlinearity becomes more significant. Hence the bias of the OLS method becomes more severe.

In summary, we know from Figure 1 that Y_i exhibits a Pareto tail. This implies that $\mathbb{E}[Y_i|Y_i > y_{\min}] = y_{\min}\alpha/(\alpha - 1)$. Considering that $\alpha = \alpha(X_i)$ is a function of X_i, the Pareto tail imposes a nonlinear effect of X_i on the tail expectation of Y_i. Such a nonlinear effect cannot be well approximated by the linear regression model, except in some special cases. This observation is the first motivation

Figure 3: Histograms of OLS with Linear or Natural Logarithms of Y and the proposed MLE



Notes: This figure depicts the histograms of the OLS estimator (subtracted by the true value) with Y_i (red color) or $\ln Y_i$ (yellow color) and the MLE (green color) for the marginal effect with $\beta_0 = 0, 0.5, 1$. The vertical line depicts the averages of the estimators. Results are based on 10,000 simulation draws. See the main text for more details about the data generating process.

for our proposed MLE that explicitly takes advantage of the Pareto tail regardless of its heaviness.

The Effect of Heavy Tails on Variance. After having examined the bias, we now evaluate the effect of a heavy tail on the variance of the OLS estimation. To rule out that the effect stems from nonlinearities, we generate data from the standard linear regression model that

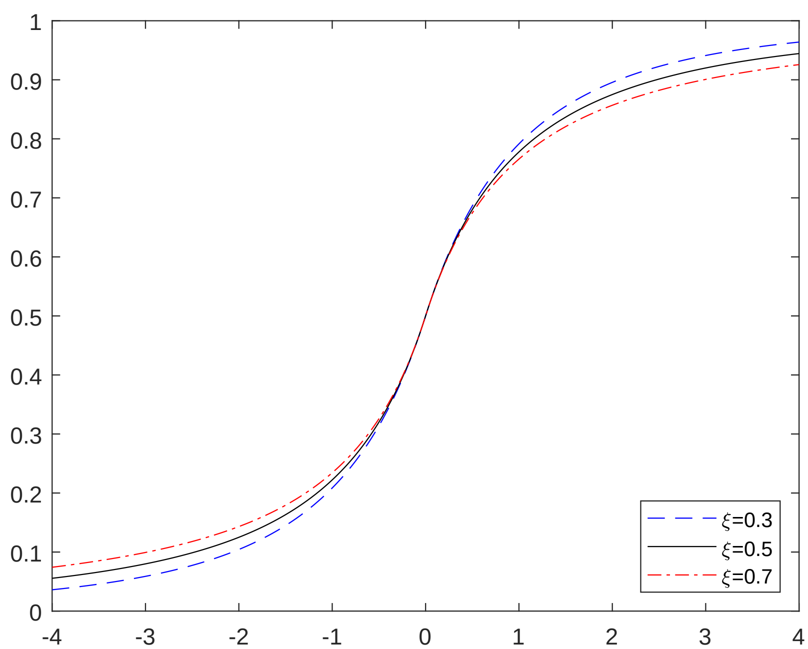
$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

with $(\beta_0, \beta_1) = (1, 0)$ and X_i being i.i.d. standard normal. To characterize the potential heavy tail, we independently generate u_i from the two-sided generalized Pareto distribution. In particular, the cumulative distribution function of u_i is such that

$$\mathbb{P}(u_i \leq u) = \left(\frac{1}{2}(1-\xi u)^{-1/\xi}\right)\mathbf{1}[u \leq 0] + \left(1 - \frac{1}{2}(1+\xi u)^{-1/\xi}\right)\mathbf{1}[u > 0].$$

The functions with $\xi = 0.3, 0.5, 0.7$ are plotted in Figure 4.

Figure 4: Plot of the CDF of the Two-sided Generalized Pareto Distribution.



Notes: This figure depicts the cumulative density function (CDF) of the two-sided generalized Pareto distribution for $\xi = 0.3, 0.5, 0.7$. See the main text for more details.

The parameter ξ is called the tail index and equals the reciprocal of the Pareto exponent α. Using the standard OLS, we estimate the coefficients β_0 and β_1 and construct the standard t-statistic and the 95% confidence interval based on heteroskedasticity robust standard errors. We implement our simulations with a wide range of sample sizes $n \in \{500, 1000, 5000, 10^4, 10^5, 10^6\}$. Let $\hat{\beta}_1(s)$ denote the OLS estimator for β_1 in the sth simulated draw. Further, let $\hat{\sigma}(s)$ denote the estimated robust standard error of the OLS estimator $\hat{\beta}_1(s)$. Accordingly, let $t(s) = \hat{\beta}_1(s)/\hat{\sigma}(s)$ denote the t-statistic.

Panel A of Table 2 depicts the mean absolute deviation (MAD): $S^{-1}\sum_{s=1}^{S}|\hat{\beta}_1(s) - \beta_1|$ across $S = 10,000$ simulation draws. Panel B depicts the root mean squared error (RMSE): $(S^{-1}\sum_{s=1}^{S}|\hat{\beta}_1(s) - \beta_1|^2)^{1/2}$ of the OLS estimator. Panel C depicts the average rejection rate for the null that $\beta_1 = 0$, that is $S^{-1}\sum_{s=1}^{S}1[|t(s)| > 1.96]$. Finally, Panel D of Table 2 depicts the average length of the 95% confidence intervals, that is, $S^{-1}\sum_{s=1}^{S}2 \times 1.96\hat{\sigma}(s)$.

Table 2: OLS Simulation Results with Generalized Pareto Distribution

n	500	1000	5000	10^4	10^5	10^6	500	1000	5000	10^4	10^5	10^6
$\xi(1/\alpha)$			Panel A: MAD						Panel B: RMSE			
0.09	0.06	0.04	0.02	0.01	0.00	0.00	0.07	0.05	0.02	0.02	0.01	0.00
0.19	0.07	0.05	0.02	0.02	0.01	0.00	0.09	0.06	0.03	0.02	0.01	0.00
0.29	0.09	0.06	0.03	0.02	0.01	0.00	0.12	0.08	0.04	0.03	0.01	0.00
0.39	0.13	0.09	0.04	0.03	0.01	0.00	0.18	0.12	0.05	0.04	0.01	0.00
0.49	0.18	0.14	0.07	0.05	0.02	0.01	0.27	0.25	0.12	0.09	0.02	0.01
0.59	0.32	0.24	0.13	0.10	0.05	0.02	1.31	0.68	0.25	0.16	0.08	0.03
0.69	0.63	0.50	0.28	0.23	0.12	0.06	5.09	3.36	0.89	0.79	0.47	0.49
0.79	1.14	0.94	0.74	0.64	0.38	0.24	4.47	3.98	5.43	4.63	1.84	1.40
0.89	2.87	5.19	5.02	3.43	1.88	1.19	46.9	260	291	147	30.0	13.5
0.99	5.61	6.69	5.49	5.68	5.00	7.75	44.5	87.7	44.1	38.7	37.1	156
$\xi(1/\alpha)$			Panel C: Rejection Prob.						Panel D: Length of 95% CI			
0.09	0.05	0.05	0.05	0.05	0.05	0.05	0.28	0.20	0.09	0.06	0.02	0.01
0.19	0.05	0.05	0.05	0.05	0.05	0.05	0.34	0.25	0.11	0.08	0.02	0.01
0.29	0.05	0.05	0.05	0.05	0.05	0.05	0.44	0.31	0.14	0.10	0.03	0.01
0.39	0.05	0.05	0.05	0.05	0.05	0.05	0.59	0.43	0.20	0.14	0.05	0.01
0.49	0.04	0.04	0.05	0.05	0.05	0.05	0.85	0.65	0.32	0.24	0.08	0.03
0.59	0.04	0.03	0.04	0.04	0.04	0.04	1.43	1.08	0.58	0.44	0.18	0.07
0.69	0.03	0.03	0.04	0.04	0.03	0.04	2.75	2.17	1.23	1.01	0.51	0.27
0.79	0.03	0.03	0.03	0.03	0.03	0.03	4.81	3.97	3.14	2.73	1.62	1.03
0.89	0.03	0.03	0.03	0.02	0.03	0.03	11.8	20.8	20.1	13.8	7.70	4.92
0.99	0.02	0.02	0.02	0.02	0.02	0.02	22.9	27.2	22.3	23.1	20.4	31.1

Notes: The table depicts the average mean absolute deviation (MAD), average root mean squared error (RMSE), average rejection probability of the standard t-test, and the average length of the standard 95% confidence intervals. The results are based on 10,000 simulation draws. See the main text for details about the data generating process.

Table 2 shows the following. First, note that the error term u_i has finite variance when $\xi < 0.5$ ($\alpha > 2$). Thus, in the first five rows of Panels A and B, the MAD and the RMSE are reasonably small. In comparison, they become substantially larger in the bottom five rows where $\xi > 0.5$.

Second, when the variance of u_i is finite, we expect the t-statistic to be approximately normally distributed, as implied by the central limit theorem. Therefore, we would expect that the rejection probability is around 5%, as seen in the first five rows of Panel C. However, when $\xi > 0.5$, the rejection probability becomes substantially smaller than 5%.

Third, following the previous point, the underrejection results from large standard errors, as reflected in the long confidence intervals in Panel D. Remember that the confidence interval is expected to shrink at the root-n rate when $\xi < 0.5$. However, when $\xi > 0.5$, the standard error is not well-defined and hence the confidence interval becomes too wide to be informative. More specifically, since the variance of u_i is infinite, we may alternatively consider its inter-quantile range as a benchmark.

In our data generating process, the inter-quantile range of u_i is one when $\xi = 1$ and the standard deviation of X_i is always one. So the average length of the confidence interval should be of the order of magnitude $n^{-1/2}$ if the central limit theorem provides a good approximation. However, the average length of the standard CI is substantially larger than $n^{-1/2}$. In the last row of Table 2, where ξ (and α) is approximately one, the average length of the confidence interval is even above 20. Therefore, the standard OLS-based inference is not performing satisfactorily under the heavy tail distribution.

Finally, our simulations in Table 2 assume a correctly specified model. Given the poor performance of the linear model in the presence of heavy tails, an applied researcher might be tempted to follow the common practice of taking the logarithm of Y_i as the dependent variable as in Figure 2. The performance of such a transformed specification crucially depends on the true data generating process, which is typically unknown.

In Appendix A2, we conduct simulations based on a logarithmic specification applied to a linear data generating process with heavy tails. We find that approximating the linear model with a heavy-tailed error by the log-linear model could lead to substantial misspecification errors, which also do not diminish as sample size increases.

5.1.2 Generalized Linear Model

Next, we repeat the previous exercise using the generalized linear model (GLM). The GLM is also widely used in health economics to model health expenditure distributions; see, for instance, Manning and Mullahy (2001) and Buntin and Zaslavsky (2004). More specifically, we generate the data from

$$Y_i = \exp(\beta_0 + \beta_1 X_i) + u_i, \tag{5.1}$$

with $(\beta_0, \beta_1) = (1, 0)$ and (X_i, u_i) following the same distribution as before. This model implies that the conditional mean $\mathbb{E}[Y_i|X_i = x] = \exp(\beta_0 + \beta_1 x)$, which is nonlinear in X_i. To discipline the estimators, we impose the infeasible bound that the estimators are within $[-50, 50]$.

An alternative data generating process could be $Y_i = \exp(\beta_0 + \beta_1 X_i)\exp(u_i)$. This model implies that $Y_i > 0$ almost surely and hence taking natural logarithms on both sides yields that $\ln Y_i = \beta_0 + \beta_1 X_i + \mathbb{E}[\ln u_i]$. Note that Appendix A2 presents additional simulation results when taking natural logarithms of Y_i. Since the expenditure data usually involve $Y_i = 0$, we stick to (5.1) in this simulation.

Table 3 depicts the same performance measures as in Table 2 for the GLM method. The findings are also similar. In particular, the GLM estimator has a small bias and RMSE when the error term does not have a heavy tail and $\xi < 0.5$ in Panels A and B. The confidence interval is short and shrinking with the sample size. In contrast, in Panel C, the t-test substantially overrejects when the error term has a heavy tail. Moreover, the confidence interval is 'exploding' when ξ exceeds 0.6. Such wide confidence intervals originate from the poor standard error estimates in the GLM. To see this, note that we obtain the GLM estimator $\hat{\beta}$ by solving the nonlinear least squares problem

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \exp(\beta_0 + \beta_1 X_i))^2.$$

Some derivation shows that the asymptotic variance of $\hat{\beta}$ becomes

$$\mathbb{E}[u_i^2] \begin{pmatrix} \mathbb{E}[\exp(2(\beta_0 + \beta_1 X_i))] & \mathbb{E}[\exp(2(\beta_0 + \beta_1 X_i))X_i] \\ \mathbb{E}[\exp(2(\beta_0 + \beta_1 X_i))] & \mathbb{E}[\exp(2(\beta_0 + \beta_1 X_i))X_i^2] \end{pmatrix}^{-1}.$$

When the error has a heavy tail, $\mathbb{E}[u_i^2]$ becomes extremely large. Furthermore, in this case, the estimator can be numerically unstable such that the above matrix is not invertible. Both features lead

Table 3: GLM Simulation Results with Generalized Pareto Distribution

n	500	1000	5000	10^4	10^5	10^6	500	1000	5000	10^4	10^5	10^6
$\xi(1/\alpha)$	Panel A: MAD						Panel B: RMSE					
0.09	0.02	0.02	0.01	0.00	0.00	0.00	0.03	0.02	0.01	0.01	0.00	0.00
0.19	0.03	0.02	0.01	0.01	0.00	0.00	0.03	0.02	0.01	0.01	0.00	0.00
0.29	0.04	0.02	0.01	0.01	0.00	0.00	0.07	0.03	0.01	0.01	0.00	0.00
0.39	0.05	0.03	0.02	0.01	0.00	0.00	0.27	0.05	0.02	0.01	0.00	0.00
0.49	0.07	0.05	0.03	0.02	0.01	0.00	0.30	0.19	0.04	0.03	0.01	0.00
0.59	0.12	0.10	0.05	0.04	0.01	0.01	0.53	0.46	0.11	0.07	0.03	0.01
0.69	0.23	0.16	0.10	0.08	0.04	0.02	0.92	0.46	0.30	0.29	0.09	0.06
0.79	0.39	0.33	0.21	0.18	0.11	0.08	1.27	1.14	0.60	0.40	0.24	0.22
0.89	0.62	0.56	0.42	0.39	0.28	0.23	1.89	1.54	1.03	0.99	0.55	0.45
0.99	0.96	0.89	0.72	0.70	0.61	0.52	2.44	2.12	1.51	1.45	1.12	0.89
$\xi(1/\alpha)$	Panel C: Rejection Prob.						Panel D: Length of 95% CI					
0.09	0.05	0.05	0.05	0.05	0.05	0.05	0.10	0.07	0.03	0.02	0.01	0.00
0.19	0.05	0.05	0.05	0.05	0.05	0.05	0.13	0.09	0.04	0.03	0.01	0.00
0.29	0.05	0.05	0.05	0.05	0.05	0.05	0.16	0.12	0.05	0.04	0.01	0.00
0.39	0.05	0.04	0.05	0.05	0.05	0.05	0.22	0.16	0.07	0.05	0.02	0.01
0.49	0.04	0.04	0.04	0.04	0.05	0.05	0.33	1.03	0.12	0.09	0.03	0.01
0.59	0.05	0.05	0.04	0.04	0.04	0.04	66.6	0.42	0.23	0.16	0.07	0.03
0.69	0.06	0.06	0.05	0.05	0.04	0.04	$> 10^3$	$> 10^3$	409	$> 10^3$	1.94	$> 10^3$
0.79	0.08	0.08	0.07	0.07	0.05	0.04	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
0.89	0.10	0.11	0.11	0.11	0.09	0.08	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
0.99	0.13	0.13	0.14	0.14	0.13	0.14	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$

Notes: The table depicts the average mean absolute deviation (MAD), average root mean squared error (RMSE), average rejection probability of the standard t-test, and the average length of the standard 95% confidence intervals. The results are based on 10,000 simulation draws. See the main text for details about the data generating process.

to large standard errors and hence wide and uninformative confidence intervals. The results become even worse when we relax the restriction that $\hat{\beta}_j \in [-50, 50]$ for $j = 0, 1$.

Note that GLM might not even be well-defined when data exhibit heavy tails. In particular, GLM requires the second moment of Y_i to be finite, and accordingly $\alpha \geq 2$. Otherwise, the asymptotic distribution of the GLM estimator is non-Gaussian and its finite sample performance could be very unstable.

In summary, ignoring heavy tails in heavily skewed data, such as health expenditure data, can lead to substantial estimation biases as well as rejection errors in the statistical inference of unknown parameters. As shown, such errors could become even more severe in nonlinear GLM than linear

OLS models, motivating our proposed method that explicitly focuses on the heavy tail. In Appendix A2, we repeat the above simulation studies with Student-t distributed errors. The findings are very similar.

5.2 Application to Real Data: The Marginal Effect of Age

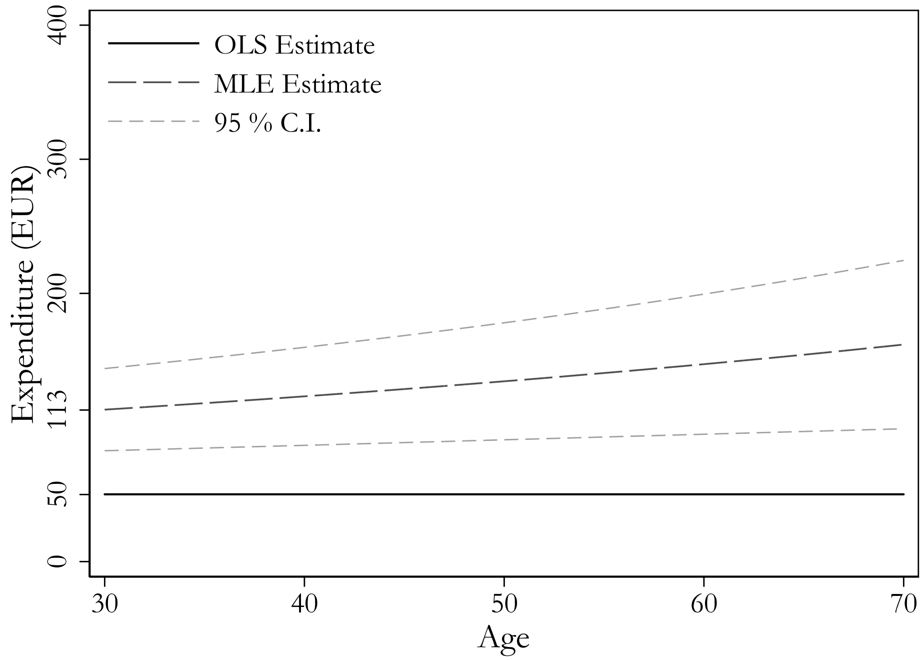
We now examine the health care expenditure data introduced in Section 4. In particular, we implement our MLE as described in equation (3.6) and set the tail cutoff $y_{\min}$ at the 95% quantile of Y_i as the benchmark value. In addition to age, the specification controls for gender, plan generosity, and also includes six year dummies. We use the final year 2011 as the reference category. We note that different disciplines have different approaches to interpreting age as an exogenous characteristic. While it is obvious that age cannot be influenced by health care spending, behaviors, or policies, and is a "good control" in the sense of Angrist and Pischke (2008), it is also clear that various unobservables that could drive spending are correlated with age. Figure 5 depicts estimates of marginal age effects on health care spending.

We summarize the results in Figure 5 as follows: First, the OLS estimate of the marginal age effect is biased: for all ages, it is outside the 95% confidence interval of the MLE estimate. Second, the bias is large and meaningful. The smallest marginal effect of the MLE, which is 112 Euro for 29-year-olds, is more than twice the OLS estimate of 51 Euro.

In Figure 6, we split the sample by gender and compare results for males and females. The OLS estimates for females in Figure 6a paint a slightly different picture compared to Figure 5 above: over the entire age range, the OLS estimate is within the confidence interval of the MLE estimate. On the other hand, the bias is larger in this subsample: for the youngest females, the OLS estimate is 77% downward biased, compared to 54% in the pooled sample. In Figure 6b for males, the relative bias is slightly lower at 51%; however, also within this subsample, the OLS estimate lies outside the MLE confidence interval for all ages.

Following the suggestion in Section 3.2, we also consider how results are affected when we vary the threshold $y_{\min}$. In Appendix Figures A2 and A3 we consider 3% and 7% as alternative thresholds. We further implement the data-driven determination of $y_{\min}$ proposed in Section 3.2. Varying $y_{\min}$

Figure 5: Comparison of OLS and MLE Estimates of Marginal Age Effects.

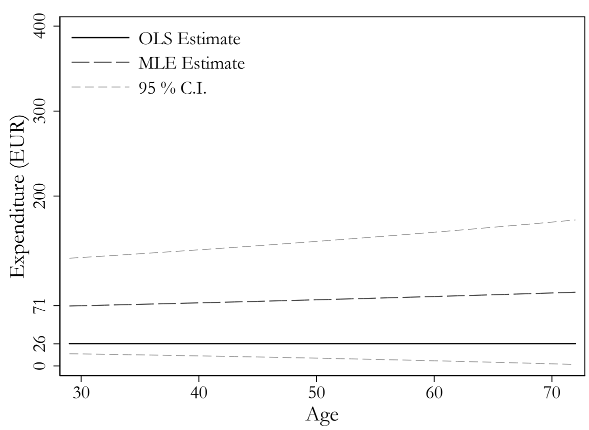


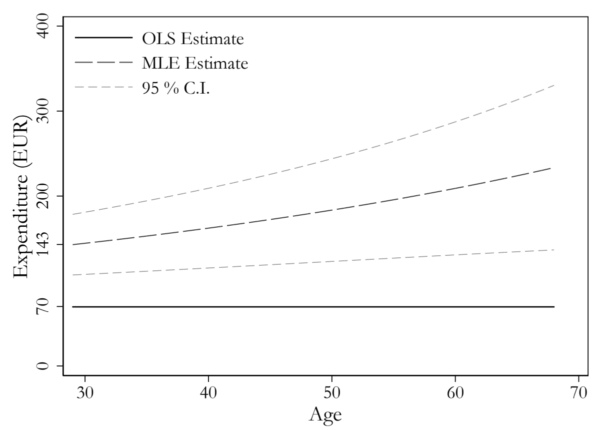
Notes: See Section 4 for more details about the German claims panel data. The graph compares age as a predictor of health care spending, showing marginal effects along with 95% confidence intervals. The marginal MLE effect is larger across the whole age domain and the difference to OLS increases with age.

Figure 6: Marginal Age Effects by Sex, OLS versus MLE.

(a) Females

(b) Males





Notes: Own calculations based on German claims panel data. The graph compares age as a predictor of health care spending, showing marginal effects along with 95% confidence intervals.

over a fine grid between top 1% and 7% percentiles, we find that the preferred threshold $y^*_{\min}$ is at the top 2.4% percentile for males, top 3.6% percentile for females, and top 2.9% percentile for the pooled sample. Results are presented in Figures A4 and A5. The main results are in general robust to these sensitivity checks, even though more restrictive cutoffs lead to a notable loss in statistical power, in particular when we split the sample by gender. Finally, we investigate whether a spline function in age can mitigate the difference between MLE and OLS estimates. Appendix Figure A6 presents the results for 10-year splines. Apparently, it aligns much better in the female subsample, whereas considerable discrepancies persist for males.

Next, we investigate some of the empirical implications of the three-part model. Whereas the traditional two-part model distinguishes between an extensive and an intensive margin, the proposed three-part model leads to two extensive and two intensive margins. Table 4 shows estimates of the marginal effects of age. The top rows provide estimates of the two extensive margins—the probabilities of having expenditures in the range $(0, y_{\min}]$ and $(y_{\min}, \infty)$ respectively. Assuming a linear index in the multinomial logit, increasing age by one year is associated with an increase in the first probability by 0.4 percentage points. The effect is more pronounced for males than for females. Likewise, the probability of having expenditure in the right tail increases by 0.17 percentage points, with similar estimates for males and females.

As mentioned in Section 3.3, the multinomial logit rests on the IIA assumption, which may be restrictive in our case given the clear ordering between the alternatives. In Appendix Table A1 we test the robustness of our results to relaxing the IIA assumption, contrasting our baseline results with estimates from a sequential logit model. All estimates from the alternative specifications are very close to our baseline specification with the possible exception of the tail extensive margin for males.

In the lower part, Table 4 compares estimates of the intensive margins. We estimate the two intensive margins and compare them to a "combined" effect which represents a weighted average of the marginal effect below and above $y_{\min}$, while also taking the probability of entering the tail ("Extensive 2") into account. The estimated marginal effects differ from each other: in the pooled sample the MLE estimate is twice as large as the OLS estimate. Accordingly, the combined intensive margin estimate is always outside the confidence interval of the intensive margin estimates (based on non-tail observations with positive expenditures). In the bottom panel, we present the intensive margin estimate that a standard two-part model with linear intensive margin would suggest. More precisely, we run OLS using all observations with $Y_i > 0$. Interestingly, these estimates are comparable in magnitude to the MLE estimates for the tail. This finding suggests that the intensive margin in the tail part largely dominates. Therefore, the OLS estimates could be severely biased if we combine tail with non-tail observations. Note that this finding does not reflect general properties of these estimators: in other settings, an OLS estimate of the combined intensive margins may well provide a more balanced approximation of the marginal effects.

Table 4: Marginal Effects of Age: Extensive and Intensive Margins

	Pooled	Females	Males
Extensive 1	0.0041	0.0011	0.0060
	(0.0001)	(0.0001)	(0.0001)
Extensive 2	0.0017	0.0010	0.0010
	(0.0000)	(0.0000)	(0.0000)
Intensive 1 (OLS)	70.852	54.685	77.854
	(0.63)	(1.13)	(0.74)
Intensive 2 (MLE)	141.632	79.207	202.582
	(24.95)	(36.34)	(38.28)
Combined	75.380	56.079	86.468
Intensive 1+2 (OLS)	152.220	100.067	192.789
	(1.68)	(2.72)	(2.13)

Source: Own calculations based on German claims panel data. Standard errors in parentheses. "Extensive 1" reports marginal age effects for the probability $\mathbb{P}(0 < Y_i \leq y_{\min} \mid X_i)$ and "Extensive 2" for the probability $\mathbb{P}(Y_i > y_{\min} \mid X_i)$. "Intensive 1" reports marginal effects of age on expenditure in the interval $(0, y_{\min})$ and "Intensive 2" reports the corresponding MLE estimates for the tail. The 'Combined' row provides a weighted average of the intensive margin, combining OLS and MLE estimates with effects on the probability of being in the tail, derived from "Extensive 2". "Intensive 1+2 (OLS)" presents OLS estimates of a linear model applied along the entire range of positive expenditure, as in the canonical two-part model.

5.3 Discussion

The previous simulations and empirical results suggest a substantial difference between our proposed MLE and the classic OLS methods. Both specifications are based on some functional form assumptions regarding the relationship between health care spending and age. More specifically, the OLS assumptions are that expenditures are linear in age and that the variance is finite, whereas the MLE estimate allows for a non-linear relationship but requires the Pareto tail, see equation (3.5).

In general, both these and any other functional form assumptions may be incorrect. When the distributional assumptions for OLS are satisfied, we know that OLS represents the best linear approximation to $\mathbb{E}[Y_i|X_i]$ (Angrist and Pischke, 2008). However, when the true distribution has an exact Pareto tail above $y_{\min}$, the MLE estimator will be the most efficient one among all consistent estimators. In order to get guidance on this issue, we compare the RMSEs based on the MLE and OLS estimates. Recall that the Pareto tail approximation (3.5) implies that

$$\mu(x) := \mathbb{E}[Y_i|Y_i > y_{\min}, X_i = x] = y_{\min}\frac{\alpha(x)}{\alpha(x) - 1}.$$

Given our MLE estimate $\hat{\beta}_{MLE}$, by plugging in $\alpha(X_i) = \exp(X_i'\hat{\beta}_{MLE})$, we compute the estimated conditional mean $\hat{\mu}(X_i)$ and then obtain the RMSE as

$$R_{MLE} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0}(Y_i - \hat{\mu}(X_i))^2\right)^{1/2},$$

where we use only the tail observations $\{Y_i > y_{\min}\}$. The RMSE based on OLS is standard as

$$R_{OLS} = \left(\frac{1}{n_0}\sum_{i=1}^{n_0}(Y_i - X_i'\hat{\beta}_{OLS})^2\right)^{1/2}.$$

In our dataset, we find that $R_{OLS} - R_{MLE} > 10^8$ for all three samples (pooled, females, males), suggesting that our MLE has a substantially better fit than the OLS estimate.

We close this section with some heuristic discussions about the underlying distributional assumptions and provide some empirical guidance. To compare the proposed MLE and the OLS methods, one needs to address two issues: (i) whether expected health expenditures can be well approximated by a linear function of age and other controls, and (ii) whether the claims data exhibit a sufficiently heavy tail such that the variance is possibly infinite.

Note that the *sample* variance is always finite given any dataset, while the unknown *population* variance could be infinite. In this scenario, the best linear approximation property fails. Then the sample variance and any other higher moment, such as sample skewness and kurtosis, are not informative about their population analogs.

The first issue (i), regarding the functional form, depends on the true data generating process, which is usually unknown. In order to shed some light on this issue, we return to the log-size-log-rank plot in Figure 1. If the distribution of expenditures has a Pareto-type tail, we expect to see a linear fit in the plot. In this scenario, age and other control variables could only affect expenditures through the Pareto exponent $\alpha = \alpha(X)$ and hence

$$\mathbb{E}[Y_i|Y_i > y_{\min}, X_i = x] = y_{\min}\frac{\alpha(x)}{\alpha(x) - 1}.$$

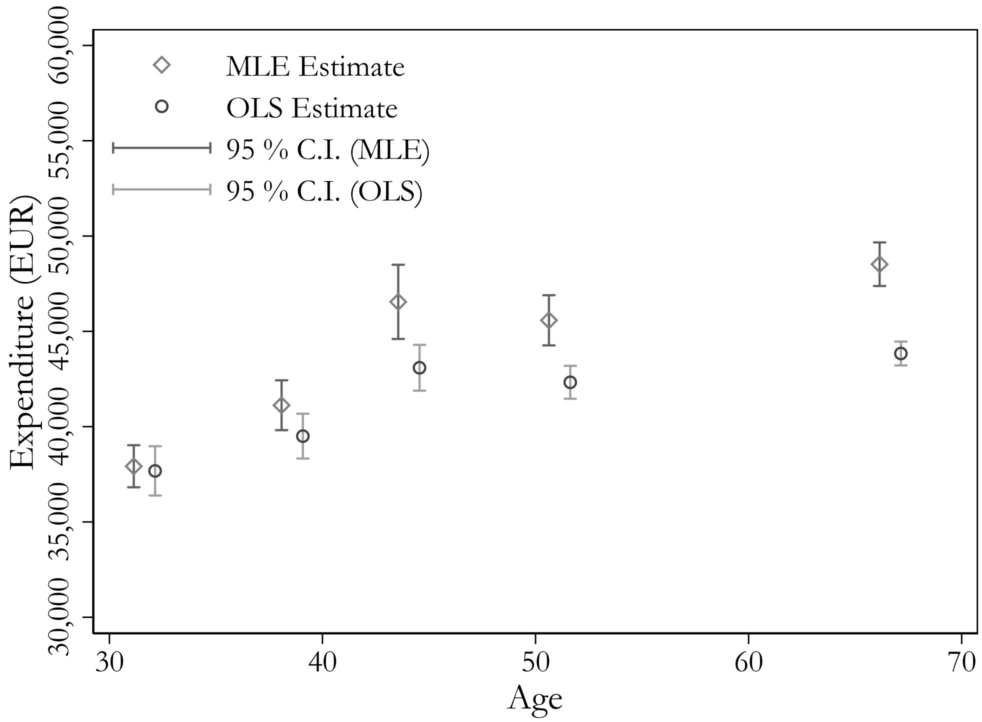
Although the functional form of $\alpha(x)$ is unknown, the conditional mean is generally nonlinear in x, leading to bias in OLS estimation. Consequently, when the data exhibit a clear Pareto pattern in the tail, researchers should apply the proposed MLE. Moreover, we follow Wang and Tsai (2009) to consider the single-index form $\alpha(X) = \exp(X'\beta)$. This additional assumption facilitates the estimation but could be restrictive.

As a robustness analysis, we relax the functional form assumptions imposed so far. To do so, we summarize the age information by five age quintile dummies. We then use the entire sample to construct point estimates and confidence intervals for the predicted expenditure values at each age quintile.[8] We plot predicted values as the partial effects of age dummies are poorly defined in the MLE specification. Figure 7 presents the results. For the youngest group, aged 31.6 years on average, the OLS estimate is close to the MLE estimate of EUR 37,900. For the second youngest group, the OLS point estimate is just outside the 95% CI of the MLE estimate, and there is still considerable

[8]Thereby it is asserted that the other independent variables representing gender, year, and plan type are constant across age quintiles.

overlap between the CI's of the two estimators. For the three oldest groups, however, the two CI's are completely disconnected, and the OLS estimates are substantially below their MLE counterparts. This finding is consistent with our simulation results.

Figure 7: Predicted Health Spending by Age Quintiles.



Note: Own calculations based on German claims panel data.

As a remark, quantile regression is another commonly adopted method to study nonlinear effects. However, we argue that it might not be appropriate in our context. To see why, let $Q_{Y|X=x}(\tau)$ denote the quantile function of Y_i conditional on $X_i = x$ for $\tau \in (0, 1)$. The classical quantile regression model imposes that

$$Q_{Y|X=x}(\tau) = x'\beta(\tau)$$

for some pseudo-true coefficient $\beta(\tau)$. Wang and Li (2013, Proposition 2.1) show that the Pareto exponent $\alpha(x)$ has to be constant over all values of x for the above quantile regression model to be appropriate. However, Figure 1 clearly shows different slopes for males and females, suggesting that the Pareto exponent changes at least across genders. This feature is not coherent with quantile regression and hence our proposed MLE is more suitable, at least in our data.

Regarding the second issue (ii), about the tail heaviness, we can estimate the unconditional Pareto exponent of Y_i using existing estimators. Common choices include Hill (1975) and Gabaix and Ibragimov (2011). The negative slope in Figure 1 is also a consistent estimator of α. As mentioned, the second moment $\mathbb{E}[Y_i^2]$ is infinite if the true α is less than two. Therefore, Figure 1 raises the concern of an infinite variance as the slope is merely above two. Sasaki and Wang (2023) provide a formal test about the finite moment condition.

Finally, our proposed MLE is specially designed for learning about the *tail* properties of health care expenditures, but not the *whole* sample. For the non-tail observations satisfying $Y_i < y_{\min}$, the support and the variance are naturally bounded, and hence OLS may perform well.

We summarize these discussions in the following guidance for an empirical implementation.

5.4 Guidance for Empirical Analysis

Step 1 Given any data with potentially heavy tails, run the log-size-log-rank plot as in Figure 1. A linear fit in the tail suggests that the underlying distribution has a Pareto-type tail.

Step 2 If the slope of the linear fit and other estimators of the Pareto exponent such as Hill (1975) and Gabaix and Ibragimov (2011) are below (or approximately) two, the underlying distribution might have a heavy tail, and the population variance might be infinite.

Step 3 Select the tail part of the data, $Y_i > y_{\min}$, and the associated X_i. Run our proposed MLE as in Section 3.2 to estimate the conditional expectation $\mathbb{E}[Y_i|Y_i > y_{\min}, X_i]$.

Step 4 For the non-tail part, estimate the three-part model as described in Section 3.3.

6 Conclusion

Health expenditure data typically involve extreme outliers. They represent heavy tail features in the underlying distribution of the data. Simple truncation of such extreme values could lead to substantial bias when estimating any type of effect on health expenditures. In general, extreme values can be a threat to the commonly adopted least squares methods.

In this paper, using simulation studies, we first show that when the underlying health expenditure distribution has heavy tails, the commonly used OLS and GLS methods may suffer from large biases. Further, the corresponding confidence intervals may be too wide to be informative. Second, to accommodate extreme values, we propose a new econometric method that allows us to recover information about the right tail of health expenditure distributions, which is entirely ignored in many standard approaches such as top coding. Third, we apply the proposed method to high-quality claims data from one of the biggest German private health insurers with 620 thousand policyholders.

We estimate the marginal effects of exogenous age predictors that substantially differ from those of the biased least squares methods. In general, OLS tends to underestimate the age gradient in health spending. However, both estimators require careful consideration of functional form assumptions. Following that, we extend the standard two-part model and propose a novel three-part approach to represent health expenditure distributions. Finally, we provide guidance and a cookbook recipe for applied economists on how to test for heavy tail features, and how to implement our proposed method.

Our findings underscore the significance of considering extreme values in empirical analysis. They may have a substantial impact on parameter estimates, particularly when the focus is on tail features. Our method provides a more nuanced understanding of the relationship between individual predictors and health care spending by effectively capturing the complex tail behavior. High and further rising health care expenditures remain a pivotal concern for both economists and policymakers. This is especially true with regard to the 5% of heavy users who produce 50% of all spending. Thus, refining methodological approaches to account for heavy tails and extreme values offers a pathway toward more accurate and insightful policy recommendations.

References

Angrist, J. D. and J.-S. Pischke (2008): *Mostly Harmless Econometrics*, Princeton University Press.

Atal, J., H. Fang, M. Karlsson, and N. R. Ziebarth (2019): "Exit, voice or loyalty? An investigation into mandated portability of front-loaded private health plans," *Journal of Risk and Insurance*, 86, 697–727.

——— (2023): "Long-term health insurance: Theory meets evidence," Tech. Rep. 26870, https://www.nber.org/papers/w26870, retrieved June 22, 2023.

Buntin, M. B. and A. M. Zaslavsky (2004): "Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures," *Journal of Health Economics*, 23, 525–542.

Chen, J. and J. Roth (2024): "Logs with zeros? Some problems and solutions," *Quarterly Journal of Economics*, 139, 891–936.

Danielsson, J., L. de Haan, L. Peng, and C. G. de Vries (2001): "Using a bootstrap method to choose the sample fraction in tail index estimation," *Journal of Multivariate analysis*, 76, 226–248.

de Haan, L. and A. Ferreira (2006): *Extreme Value Theory: An Introduction*, Springer Series in Operations Research and Financial Engineering, NY: Springer.

Deb, P., E. C. Norton, and W. G. Manning (2017): *Health Econometrics Using Stata*, Stata Press College Station, TX.

Duan, N. (1983): "Smearing estimate: a nonparametric retransformation method," *Journal of the American Statistical Association*, 78, 605–610.

Duan, N., W. Manning, C. Morris, and J. Newhouse (1982): "A comparison of alternative models for the demand for health care," *Rand Corp. Publication R-2754-HHS*.

Finkelstein, A. (2020): "A strategy for improving US health care delivery—conducting more randomized, controlled trials," *New England Journal of Medicine*, 382, 1485–1488.

French, E. and E. Kelly (2016): "Medical spending around the developed world," *Fiscal Studies*, 37, 327–344.

Gabaix, X. (2009): "Power laws in economics and finance," *Annual Review of Economics*, 1, 255–293.

——— (2016): "Power laws in economics: An introduction," *Journal of Economic Perspectives*, 30, 185–206.

Gabaix, X. and R. Ibragimov (2011): "Rank-1/2: a simple way to improve the OLS estimation of tail exponents," *Journal of Business Economics and Statistics*, 29, 24–39.

Gilleskie, D. B. and T. A. Mroz (2004): "A flexible approach for estimating the effects of covariates on health expenditures," *Journal of health economics*, 23, 391–418.

Goegebeur, Y., A. Guillou, N. K. L. Ho, and J. Qin (2023): "A Weissman-type estimator of the conditional marginal expected shortfall," *Econometrics and Statistics*, 27, 173–196.

Goegebeur, Y., A. Guillou, T. Pedersen, and J. Qin (2022): "Extreme-value based estimation of the conditional tail moment with application to reinsurance rating," *Insurance: Mathematics and Economics*, 107, 102–122.

Goegebeur, Y., A. Guillou, and J. Qin (2021): "Extreme value estimation of the conditional risk premium in reinsurance," *Insurance: Mathematics and Economics*, 96, 68–80.

Gotthold, K. and B. Gräber (2015): "So kommen Sie zurück in die gesetzliche Kasse," https://www.welt.de/finanzen/verbraucher/article138148142/So-kommen-Sie-zurueck-in-die-gesetzliche-Kasse.html, last retrieved on October 18, 2018.

Greene, W. H. and D. A. Hensher (2010): *Modeling Ordered Choices: A Primer*, Cambridge University Press.

Guillou, A. and P. Hall (2001): "A diagnostic for selecting the threshold in extreme value analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 293–305.

Handel, B. R., J. T. Kolstad, and J. Spinnewijn (2019): "Information frictions and adverse selection: Policy interventions in health insurance markets," *Review of Economics and Statistics*, 101, 326–340.

Hill, B. M. (1975): "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, 3, 1163–1174.

Jones, A. M. (2011): "Models for health care," in *The Oxford Handbook of Economic Forecasting*, ed. by M. P. Clements and D. F. Hendry, Oxford University Press, chap. 44, 473–480.

Jones, A. M., J. Lomas, P. Moore, and N. Rice (2016): "A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: An application to healthcare costs," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 179, 951–974.

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

——— (2015): "Healthcare cost regressions: going beyond the mean to estimate the full distribution," *Health economics*, 24, 1192–1212.

Kaiser Family Foundation (2019): *2019 Employer Health Benefits Survey*, https://www.kff.org/health-costs/report/2019-employer-health-benefits-survey/, retrieved on December 14, 2019.

Karlsson, M., T. Iversen, and H. Øien (2018): "Aging and Healthcare Costs," in *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press.

Karlsson, M., T. J. Klein, and N. R. Ziebarth (2016): "Skewed, persistent and high before death: medical spending in Germany," *Fiscal Studies*, 37, 527–559.

Manning, W. G. (1998): "The logged dependent variable, heteroscedasticity, and the retransformation problem," *Journal of Health Economics*, 17, 283–295.

——— (2012): "Dealing with skewed data on costs and expenditures," in *The Elgar Companion to Health Economics*, ed. by A. M. Jones, Edward Elgar Publishing, chap. 44, 473–480, 2 ed.

Manning, W. G., A. Basu, and J. Mullahy (2005): "Generalized modeling approaches to risk adjustment of skewed outcomes data," *Journal of Health Economics*, 24, 465–488.

Manning, W. G. and J. Mullahy (2001): "Estimating log models: To transform or not to transform?" *Journal of Health Economics*, 20, 461–494.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health insurance and the demand for medical care: Evidence from a randomized experiment," *American Economic Review*, 77, 251–27.

Mao, T., G. Stupfler, and F. Yang (2023): "Asymptotic properties of generalized shortfall risk measures for heavy-tailed risks," *Insurance: Mathematics and Economics*, 111, 173–192.

Matzkin, R. L. (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1339–1375.

Mihaylova, B., A. Briggs, A. O'Hagan, and S. G. Thompson (2011): "Review of statistical methods for analysing healthcare resources and costs," *Health Economics*, 20, 897–916.

Mikosch, T. (1999): *Regular Variation, Subexponentiality and their Applications in Probability Theory*, vol. 99, Eindhoven University of Technology Eindhoven, The Netherlands.

Mullahy, J. (1998): "Much ado about two: Reconsidering retransformation and the two-part model in health econometrics," *Journal of Health Economics*, 17, 247–281.

——— (2009): "Econometric modeling of health care costs and expenditures: A survey of analytical issues and related policy considerations," *Medical Care*, 47, S104–S108.

Mullahy, J. and E. C. Norton (2024): "Why transform Y? The pitfalls of transformed regressions with a mass at zero," *Oxford Bulletin of Economics and Statistics*, 86, 417–447.

Müller, U. K. and Y. Wang (2017): "Fixed-k asymptotic inference about tail properties," *Journal of the American Statistical Association*, 112, 1334–1343.

Newhouse, J. P. and C. E. Phelps (1976): "New estimates of price and income elasticities of medical care services," in *The Role of Health Insurance in the Health Services Sector*, National Bureau of Economic Research, Inc, NBER Chapters, chap. 7, 261–320.

Sasaki, Y. and Y. Wang (2023): "Diagnostic testing of finite moment conditions for the consistency and root-n asymptotic normality of the gmm and m estimators," *Journal of Business & Economic Statistics*, 41, 339–348.

Small, K. A. (1987): "A discrete choice model for ordered alternatives," *Econometrica*, 55, 409–424.

Smith, R. L. (1987): "Estimating tails of probability distributions," *Annals of Statistics*, 15, 1174–1207.

Tutz, G. (1991): "Sequential models in categorical regression," *Computational Statistics & Data Analysis*, 11, 275–295.

Wang, H. and C.-H. Tsai (2009): "Tail index regression," *Journal of the American Statistical Association*, 104, 1233–1240.

WANG, H. J. AND D. LI (2013): "Estimation of extreme conditional quantiles through power transformation," *Journal of the American Statistical Association*, 108, 1062–1074.

YUEN, R., S. STOEV, AND D. COOLEY (2020): "Distributionally robust inference for extreme Value-at-Risk," *Insurance: Mathematics and Economics*, 92, 70–89.

Appendix

The Appendix consists of three sections. Appendix A1 contains additional empirical robustness checks. Appendix A2 contains additional simulation results. Appendix A3 contains econometric details about the proposed estimator.

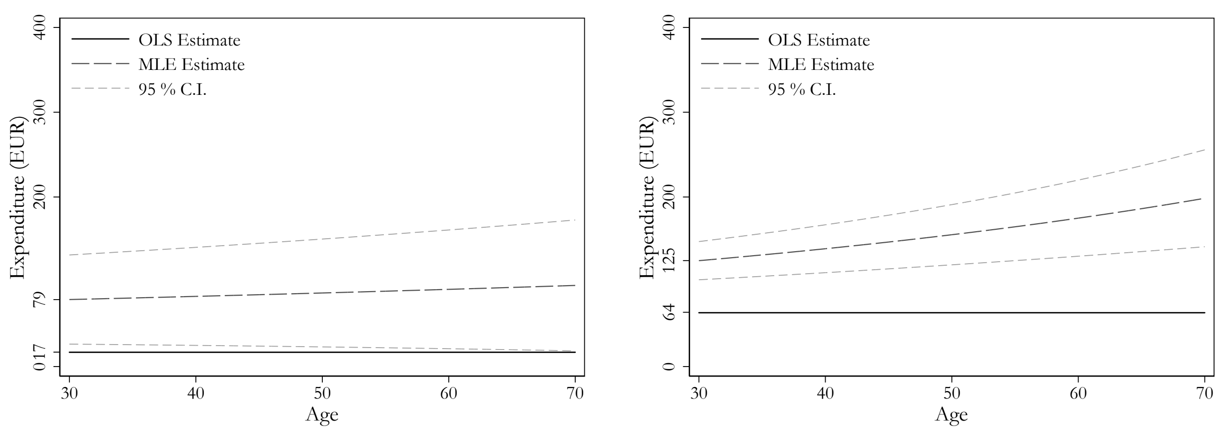
A1 Robustness Checks

We first check the robustness of the choice of $y_{\min}$. In Section 5, we use the top 5% percentile. In this section, we vary the threshold from 3 to 7% and also implement the data-driven threshold proposed in Section 3.2.

Figure A1 provides sensitivity analysis for the pooled sample. Figures A2 and A3 repeat the analysis with female and male subsamples. The findings are similar to those reported in Section 5. In particular, the OLS estimates are substantially below the MLE regardless of the subsample and the choice of $y_{\min}$.

Figure A1: Marginal Age Effects for Different Cutoffs – Pooled. OLS versus MLE.

(a) Top 3% (b) Top 7%

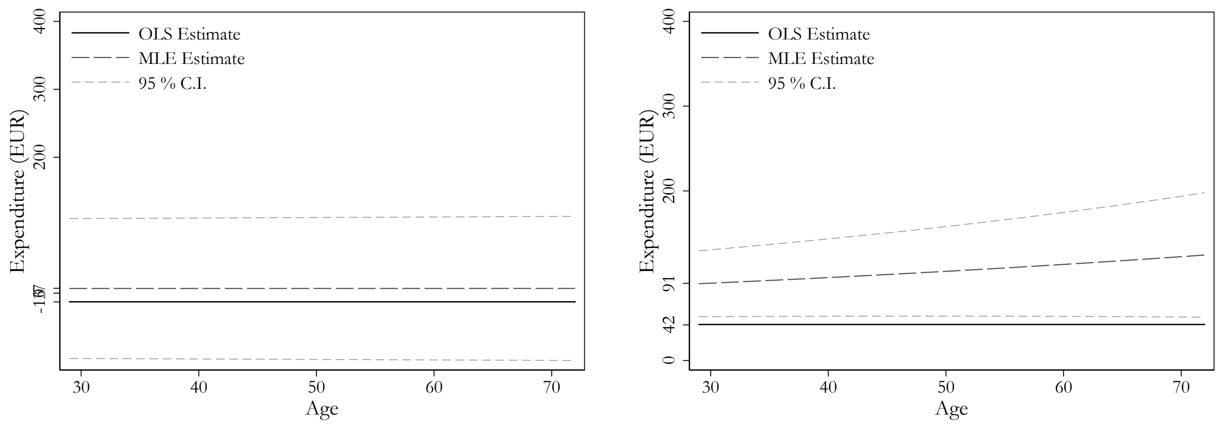


Notes: Own calculations based on German claims panel data. The graph compares the marginal age effects on health expenditures, along with 95% confidence intervals.

Figure A4 implements the data-driven y_{min} in the pooled sample and Figure A5 does implement

Figure A2: Marginal Age Effects for Different Cutoffs – Females. OLS versus MLE.

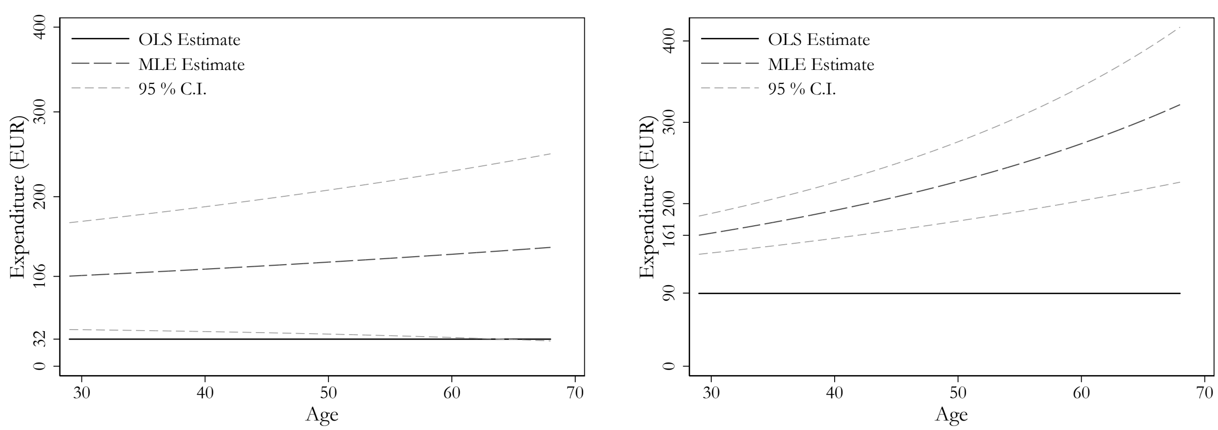
(a) Top 3% (b) Top 7%



Notes: Own calculations based on German claims panel data. The graph compares the marginal age effects on health expenditures, along with 95% confidence intervals.

Figure A3: Marginal Age Effects for Different Cutoffs – Males. OLS versus MLE.

(a) Top 3% (b) Top 7%

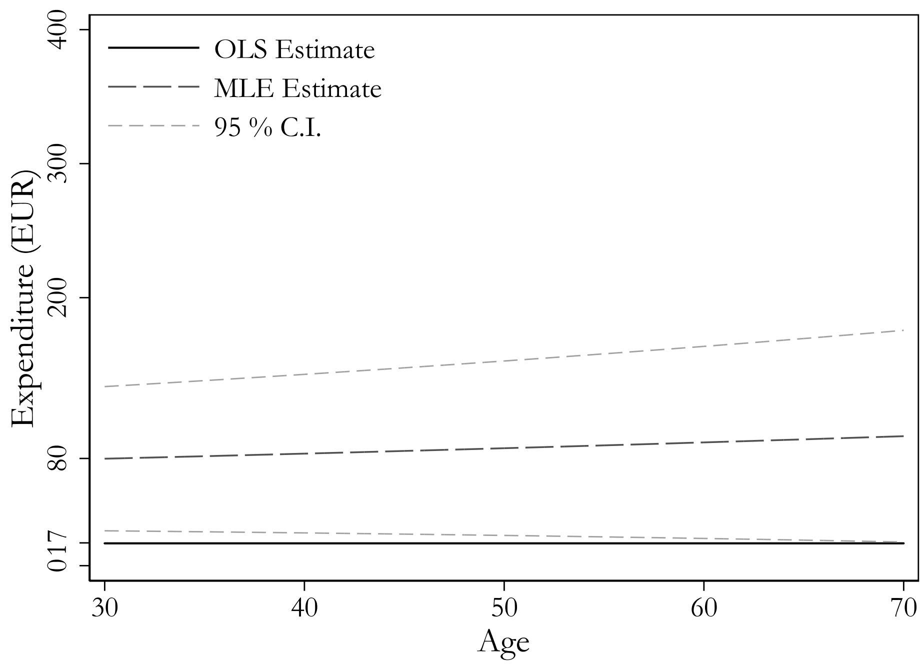


Notes: Own calculations based on German claims panel data. The graph compares the marginal age effects on health expenditures, along with 95% confidence intervals.

the data-driven thresholds that apply in the male and female subsamples, respectively.

Second, we relax the assumption of a global marginal affect along a wide range of ages; Figure A6 provides estimates based on 10-year splines in age, contrasting MLE and OLS estimates. For the female subsample, the results suggest that most of the discrepancy betwenn OLS and MLE is removed,

Figure A4: Comparison of OLS and MLE Estimates of Marginal Age Effects: Data-Driven Cutoff



Notes: See Section 4 for more details about the German claims panel data. The graph compares the marginal age effects on health care spending, along with 95% confidence intervals, for a data-driven cutoff corresponding to quantile 0.971. The marginal MLE effect is larger across the whole age domain and the difference to OLS increases with age.

whereas the opposite holds for the male subsample.

Finally, Table A1 contrasts extensive margin effects from the MNL model to the corresponding estimates based on a sequential logit model. These estimates thus give an indication of sensitivity to the IIA assumption discussed in section 3.3.

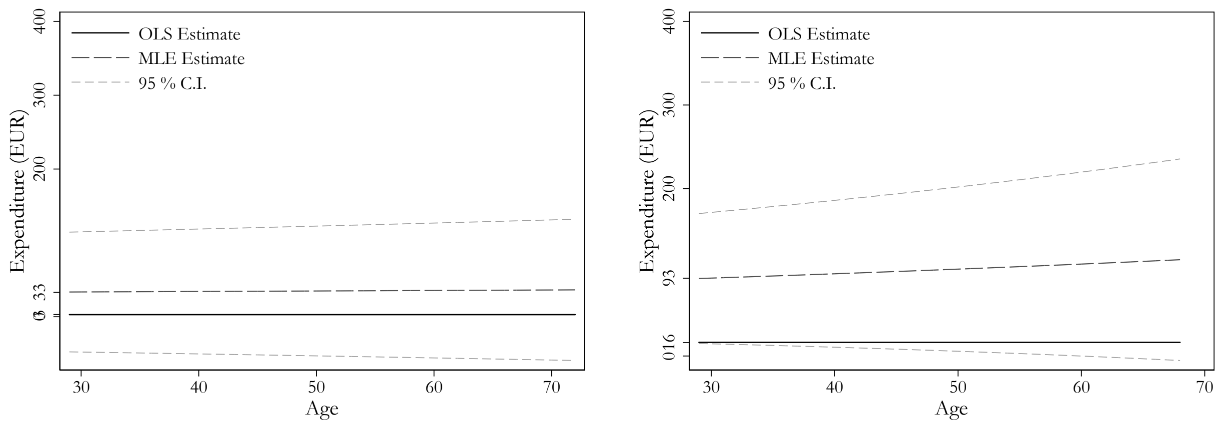
A2 Additional Simulations

A2.1 OLS with Log-Transformed Data

This section presents additional simulation results. First, we conduct simulation studies of the OLS estimator with log-transformed data. The data are still generated from

Figure A5: Marginal Age Effects for Data-Driven Cutoffs, OLS versus MLE.

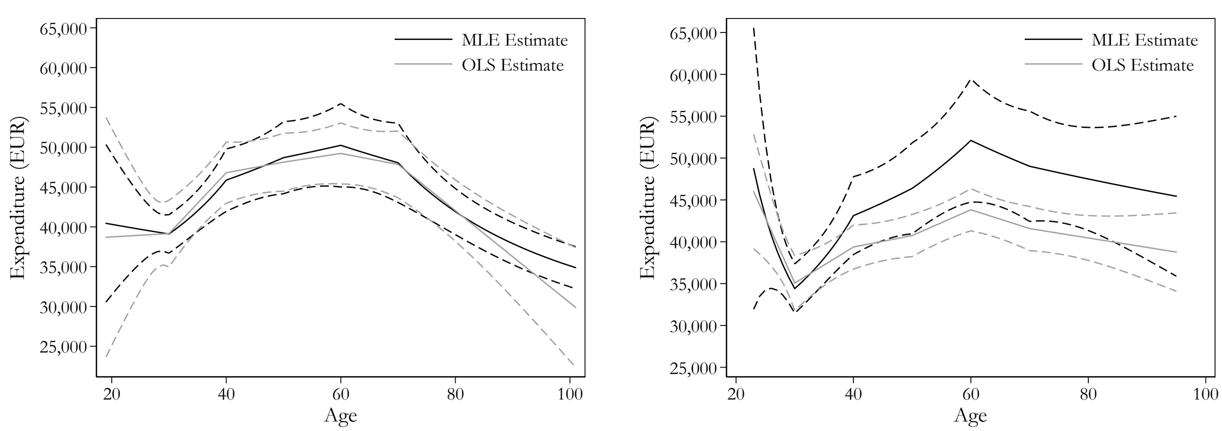
(a) Females (b) Males



Notes: Own calculations based on German claims panel data. The graph compares the marginal age effects on health expenditures, along with 95% confidence intervals, for data-driven cutoffs corresponding to quantiles 0.964 (females) and 0.976 (males).

Figure A6: Predicted Health Expenditures by Age: 10-Year Splines

(a) Females (b) Males



Notes: Own calculations based on German claims panel data.

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{A2.1}$$

with $(\beta_0, \beta_1) = (1, 1)$. To make sure that Y_i is positive, we generate X_i from the absolute value of the standard normal distribution, and u_i from the standard Pareto distribution with exponent $1/\xi$, that

Table A1: Sensitivity Analysis – Multinomial versus Sequential Logit

	Pooled	Females	Males
Multinomial Logit			
Extensive 1	0.0041	0.0011	0.0060
	(0.0001)	(0.0001)	(0.0001)
Extensive 2	0.0017	0.0010	0.0010
	(0.0000)	(0.0000)	(0.0000)
Sequential Logit			
Extensive 1	0.0042	0.0012	0.0058
	(0.0003)	(0.0001)	(0.0002)
Extensive 2	0.0018	0.0011	0.0026
	(0.0000)	(0.0000)	(0.0000)

Source: Own calculations based on German claims panel data. Standard errors in parentheses. "Extensive 1" reports marginal age effects for the probability $\mathbb{P}(0 < Y_i \leq y_{\min} \mid X_i)$ and "Extensive 2" for the probability $\mathbb{P}(Y_i > y_{\min} \mid X_i)$.

is, $\mathbb{P}(u_i \geq u) = (1 + \xi u)^{-1/\xi}$ for $u > 0$. The other model specifications are the same as in Tables 2 and 3.

Note that the OLS estimator of β_1—when regressing $\ln(Y_i)$ on X_i (and a constant)—is *not* measuring the marginal effect of X_i on Y_i, given the nonlinear setup. To make a reasonable comparison, we treat $\bar{Y}\hat{\beta}_1$ as the estimator of the marginal effect, where $\bar{Y}$ denotes the sample average of Y_i, and compare it with the true parameter β_1. The standard error of the estimator of the marginal effect is adjusted accordingly by multiplying $\bar{Y}$ to that of $\hat{\beta}_1$. As the marginal effect depends on the value of X_i, this estimator essentially estimates the average marginal effect over X_i. We also implement the estimator for $X_i = 1$ with $\beta_1 = 1$ as in Figures 2 and 3. The results are very similar and hence omitted (but available upon request).

Table A2 depicts the performance of such a marginal effect estimator in terms of MAD, RMSE, average rejection probability of the standard t-test, and the average length of the standard 95% confidence intervals.

We summarize the results as follows. First, the large bias and RMSE indicate that the misspecification error of approximating the linear model with a heavy-tailed error term by the log-linear model

Table A2: Simulation Results with $\ln Y$ and Generalized Pareto Distribution

n	500	1000	5000	10^4	10^5	10^6	500	1000	5000	10^4	10^5	10^6
$\xi(1/\alpha)$	Panel A: MAD						Panel B: RMSE					
0.09	0.07	0.06	0.05	0.05	0.05	0.05	0.09	0.07	0.05	0.05	0.05	0.05
0.19	0.09	0.08	0.08	0.08	0.08	0.08	0.11	0.10	0.08	0.08	0.08	0.08
0.29	0.13	0.12	0.12	0.12	0.12	0.12	0.16	0.14	0.13	0.12	0.12	0.12
0.39	0.19	0.18	0.18	0.18	0.18	0.18	0.22	0.20	0.19	0.19	0.18	0.18
0.49	0.28	0.27	0.27	0.27	0.27	0.27	0.33	0.30	0.28	0.28	0.27	0.27
0.59	0.41	0.41	0.41	0.41	0.41	0.41	0.50	0.46	0.42	0.42	0.41	0.41
0.69	0.66	0.64	0.64	0.66	0.64	0.64	1.57	1.07	0.81	1.88	0.68	0.65
0.79	1.61	1.03	1.23	1.11	1.09	1.10	47.9	2.65	11.2	4.48	1.27	1.29
0.89	1.88	2.05	2.13	2.12	2.07	2.26	8.90	19.7	9.34	11.8	3.48	6.63
0.99	4.43	4.37	6.47	9.83	6.04	6.02	65.5	61.5	129	342	104	37.6
$\xi(1/\alpha)$	Panel C: Rejection Prob.						Panel D: Length of 95% CI					
0.09	0.10	0.15	0.57	0.86	1.00	1.00	0.27	0.19	0.09	0.06	0.02	0.01
0.19	0.15	0.27	0.87	0.99	1.00	1.00	0.32	0.23	0.10	0.07	0.02	0.01
0.29	0.24	0.44	0.98	1.00	1.00	1.00	0.37	0.27	0.12	0.08	0.03	0.01
0.39	0.35	0.62	1.00	1.00	1.00	1.00	0.45	0.32	0.14	0.10	0.03	0.01
0.49	0.49	0.77	1.00	1.00	1.00	1.00	0.54	0.38	0.17	0.12	0.04	0.01
0.59	0.61	0.89	1.00	1.00	1.00	1.00	0.68	0.48	0.22	0.15	0.05	0.02
0.69	0.73	0.95	1.00	1.00	1.00	1.00	0.91	0.63	0.28	0.20	0.06	0.02
0.79	0.81	0.98	1.00	1.00	1.00	1.00	1.94	0.88	0.43	0.29	0.09	0.03
0.89	0.86	0.99	1.00	1.00	1.00	1.00	2.02	1.51	0.68	0.48	0.15	0.05
0.99	0.90	1.00	1.00	1.00	1.00	1.00	4.63	3.03	1.84	1.94	0.38	0.12

Notes: The table depicts the average mean absolute deviation (MAD), the average root mean squared error (RMSE), the average rejection probability of the standard t-test, and the average length of the standard 95% confidence intervals. The results are based on 10,000 simulation draws. See the main text for details about the data generating process.

can be substantial (Panel A and B). Such misspecification error is small when the tail is thin, i.e, ξ is close to zero. This is what has been documented in the existing literature. However, when the tail of the error term becomes heavy, the misspecification error is amplified substantially. Second, accordingly, the rejection probability of the t-test becomes much larger than the nominal 5% level (Panel C). When ξ is above 0.5, the variance of the error term is not well-defined and hence the t-test is hardly informative. Third, although the confidence interval shrinks with the sample size, the bias and the overrejection do not (Panel D). This suggests that the estimator substantially deviates from the true value with the bias strictly dominating the randomness.

A2.2 Student-t Distribution

The simulation results in Tables 2, 3, and A2 are based on u_i with exact Pareto tails. Now we repeat the analysis using Student-t distribution, whose tail is only approximately Pareto. In particular, Table A3 adopts the same data generating process as in Table 2, that is,

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

with $(\beta_0, \beta_1) = (1, 0)$ and X_i being i.i.d. standard normal. Now we generate u_i from the Student-t distribution with α degrees of freedom. Note that Student-t distribution can be well approximated by Pareto in the tail with the Pareto exponent equal to α. Using the same parameter space as in Table 2, we implement our simulations with a wide range of sample sizes $n \in \{500, 1000, 5000, 10^4, 10^5, 10^6\}$ and $\xi = 1/\alpha \in (0.09, 0.19, ..., 0.99)$. The variance of u_i is finite if $\xi < 0.5$ (and equivalently $\alpha > 2$) and infinite otherwise. The findings are very similar to those in Table 2.

Next, we repeat the exercise in Table 3 using the generalized linear model (GLM) with data generated from the following model:

$$Y_i = \exp(\beta_0 + \beta_1 X_i) + u_i,$$

with $(\beta_0, \beta_1) = (1, 0)$, X_i is standard normal, and u_i is Student-t with α degrees of freedom. Table A4 presents the same measure as in Table 3. Again, the findings are very similar.

A3 More Details about the MLE

Our maximum likelihood estimator is based on the tail index regression proposed by Wang and Tsai (2009). The key condition is that the conditional distribution of Y_i on $X_i = x$ has an approximate Pareto tail. In particular, we assume that uniformly over x,

$$\mathbb{P}\left(Y_i > y | X_i = x\right) = c\left(x\right) y^{-\alpha(x)} \left(1 + o(1)\right) \text{ as } y \to \infty, \tag{A3.1}$$

Table A3: OLS Simulation Results with Student-t Distribution

n	500	1000	5000	10^4	10^5	10^6	500	1000	5000	10^4	10^5	10^6
$\xi(1/\alpha)$	Panel A: MAD						Panel B: RMSE					
0.09	0.04	0.03	0.01	0.01	0.00	0.00	0.05	0.03	0.02	0.01	0.00	0.00
0.19	0.05	0.03	0.01	0.01	0.00	0.00	0.06	0.04	0.02	0.01	0.00	0.00
0.29	0.06	0.04	0.02	0.01	0.00	0.00	0.07	0.05	0.02	0.02	0.00	0.00
0.39	0.07	0.05	0.02	0.02	0.01	0.00	0.09	0.07	0.03	0.02	0.01	0.00
0.49	0.10	0.08	0.04	0.03	0.01	0.00	0.15	0.14	0.05	0.04	0.01	0.00
0.59	0.17	0.14	0.07	0.05	0.02	0.01	0.33	0.60	0.15	0.11	0.05	0.03
0.69	0.32	0.26	0.17	0.14	0.07	0.03	1.15	1.32	0.85	0.82	0.31	0.11
0.79	0.62	0.56	0.43	0.34	0.24	0.15	1.98	2.77	2.42	1.36	1.78	1.48
0.89	1.54	1.48	1.17	1.44	0.91	0.81	10.8	14.1	6.45	20.1	7.41	14.3
0.99	3.96	6.11	3.61	5.93	3.90	5.83	67.8	233	33.0	157	44.3	155
$\xi(1/\alpha)$	Panel C: Rejection Prob.						Panel D: Length of 95% CI					
0.09	0.05	0.05	0.05	0.05	0.05	0.05	0.19	0.14	0.06	0.04	0.01	0.00
0.19	0.05	0.05	0.05	0.05	0.05	0.05	0.22	0.16	0.07	0.05	0.02	0.00
0.29	0.05	0.05	0.05	0.05	0.05	0.05	0.27	0.19	0.09	0.16	0.02	0.01
0.39	0.05	0.05	0.05	0.05	0.05	0.05	0.34	0.25	0.11	0.08	0.03	0.01
0.49	0.05	0.05	0.05	0.05	0.04	0.05	0.49	0.36	0.18	0.13	0.04	0.01
0.59	0.04	0.04	0.04	0.04	0.04	0.04	0.77	0.63	0.33	0.25	0.10	0.04
0.69	0.03	0.03	0.04	0.04	0.04	0.04	1.42	1.14	0.75	0.59	0.29	0.14
0.79	0.03	0.03	0.03	0.03	0.03	0.03	2.64	2.40	1.84	1.46	1.02	0.64
0.89	0.03	0.02	0.03	0.02	0.02	0.02	6.36	6.11	4.89	5.45	3.76	3.30
0.99	0.02	0.02	0.02	0.02	0.02	0.02	16.1	24.5	14.6	23.8	15.7	23.4

Notes: The table depicts the average mean absolute deviation (MAD), average root mean squared error (RMSE), average rejection probability of the standard t-test, and the average length of the standard 95% confidence intervals. The results are based on 10,000 simulation draws. See the main text for details about the data generating process.

for some constant $c(x) > 0$ and $\alpha(x) > 0$. This condition is mild and satisfied by many commonly used distributions such as Student-t, Gamma, and F distributions. In particular, if Y is Student-t distributed conditional on $X = x$ with $v(x)$ degrees of freedom, then $\alpha(x)$ is simply $v(x)$. Chapter 1 in de Haan and Ferreira (2006) provides a complete review of the literature.

Note that condition (A3.1) requires that the right tail of the conditional distribution is well approximated by a Pareto distribution with component $\alpha(x)$. Such an approximation becomes more accurate as we move further towards the tail (i.e., $y \to \infty$). In this sense, we consider our method a semiparametric method that does not hinge on any specific distribution. This is important for empirical applications as, *a priori*, researchers do not know the true health expenditure distribution.

Table A4: GLM Simulation Results with Student-t Distribution

n	500	1000	5000	10^4	10^5	10^6	500	1000	5000	10^4	10^5	10^6
$\xi(1/\alpha)$	Panel A: MAD						Panel B: RMSE					
0.09	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.00
0.19	0.02	0.01	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.00
0.29	0.02	0.01	0.01	0.00	0.00	0.00	0.03	0.02	0.01	0.01	0.00	0.00
0.39	0.03	0.02	0.01	0.01	0.00	0.00	0.03	0.03	0.01	0.01	0.00	0.00
0.49	0.04	0.03	0.01	0.01	0.00	0.00	0.21	0.10	0.03	0.01	0.00	0.00
0.59	0.07	0.05	0.03	0.02	0.01	0.00	0.25	0.22	0.06	0.05	0.02	0.01
0.69	0.14	0.10	0.06	0.05	0.02	0.01	0.75	0.41	0.13	0.24	0.08	0.04
0.79	0.23	0.19	0.13	0.11	0.07	0.02	0.87	0.66	0.36	0.36	0.19	0.16
0.89	0.41	0.34	0.29	0.26	0.20	0.15	1.28	0.98	0.75	0.72	0.48	0.32
0.99	0.68	0.61	0.55	0.51	0.47	0.40	1.92	1.63	1.29	1.12	0.93	0.70
$\xi(1/\alpha)$	Panel C: Rejection Prob.						Panel D: Length of 95% CI					
0.09	0.06	0.05	0.05	0.05	0.05	0.05	0.07	0.05	0.02	0.02	0.01	0.00
0.19	0.06	0.05	0.05	0.05	0.05	0.05	0.08	0.06	0.03	0.03	0.01	0.00
0.29	0.05	0.05	0.05	0.05	0.05	0.05	0.10	0.07	0.03	0.02	0.01	0.00
0.39	0.05	0.05	0.05	0.05	0.05	0.05	0.13	0.09	0.04	0.05	0.02	0.01
0.49	0.05	0.05	0.04	0.05	0.05	0.05	0.18	0.13	0.06	0.05	0.02	0.01
0.59	0.04	0.04	0.04	0.05	0.04	0.04	$> 10^3$	0.24	0.12	0.09	0.04	0.01
0.69	0.05	0.04	0.04	0.04	0.04	0.03	395	31.0	13.9	$> 10^3$	0.13	0.05
0.79	0.06	0.06	0.05	0.05	0.04	0.04	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	7.55	$> 10^3$
0.89	0.08	0.09	0.08	0.08	0.07	0.06	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
0.99	0.11	0.11	0.12	0.13	0.13	0.12	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$

Notes: The table depicts the average mean absolute deviation (MAD), average root mean squared error (RMSE), average rejection probability of the standard *t*-test, and the average length of the standard 95% confidence intervals. The results are based on 10,000 simulation draws. See the main text for details about the data generating process.

Under Condition (A3.1), we can further approximate the conditional probability density of Y given X and $Y > y_{\min}$ for some large tail threshold $y_{\min}$ by $\alpha(x)(y/y_{\min})^{-\alpha(x)}y^{-1}$. This leads to the negative likelihood function in Equation (3.3). Under (A3.1) and some additional technical conditions, Wang and Tsai (2009) establish that by solving (3.3), the MLE $\hat{\beta}$ is consistent and asymptotically normal. The standard error can be estimated by (3.4).