Statistical Methods for Analysis of Health Care Costs

Xiao-Hua Andrew Zhou

azhou@u.washington.edu

HSR&D Center of Excellence, Seattle VA Medical Center

Department of Biostatistics, University of Washington

Outline

- Challenges in the analysis of health care costs.
- Misuse of statistical methods in analysis of costs in medical literature.
- Some new statistical methods for the analysis of health care costs.

Some challenges in analysis of costs

- A proportion of the population can usually be expected to incur no costs during a typical study period.
- Non-zero cost observations are typically skewed to the right; their distribution may be approximated by a log-normal distribution (Diehr et al., 2000; Annual Reviews).
- Health care costs typically exhibit heteroscedasticity.

Focus on population means

- Although it is a common practice to use the median instead the mean as the measure of central location in skewed data, many applications do call for the use of means.
- This is particularly common in the analyses of medical cost data, because the mean can be used to recover the total cost, which reflects the entire expenditure on health care in a given patient population.

Misuse of methods for analysis of costs in medical literature

- Zhou et al. (1997a, Annals of Internal Medicine) reviewed statistical methods in studies of medical costs published in medical journals between January, 1991 to January, 1996 and found that at least 26% of the studies might have wrong conclusions.
- Barber and Thompson (1998, British Medical Journal) critically reviewed the statistical methods used in analysis of health care costs in randomized trials, and they found that in at least two thirds of the published papers, the main conclusions regarding costs were not justified.

Some newly developed statistical methods

- One population
- Two populations
- Three or more populations
- Regression models.

Statistical problem in one population

- Let $W_1, \ldots, W_n$ be a random sample from a skewed distribution with mean θ and variance τ^2. We are interested in point and interval estimators for θ.

Point estimators of θ

- Let $W_1, \ldots, W_n$ be a random sample from a skewed distribution with mean θ and variance τ^2. We are interested in point and interval estimators for θ.

Point estimators of θ, cont

Assuming that $Y = \log W \sim N(\mu, \sigma^2)$ (Zhou, 1998, Stat Med), we have the following estimators for θ.

- The sample mean, $\bar{W}$.
- the ML estimator:

$$\exp(\bar{Y} + \frac{m}{2(m+1)} S^2),$$

  where $m = n - 1$, $\bar{Y}$ and S^2 are the sample mean and variance of $Y_1, \ldots, Y_n$.
- A uniformly minimum variance unbiased (UMVU) estimator:

$$exp(\bar{Y}) g_m(S^2/2),$$

  where

$$g_m(t) = \sum_{r=1}^{\infty} \frac{1}{r!} \frac{m+2r}{m} (\frac{m}{m+1} t)^r \prod_{i=1}^{r} \frac{m}{m+2i}.$$

- A conditionally minimal MSE estimator:

$$exp(\bar{Y}) g_m(\frac{m-3}{2m} S^2).$$

Results

- After deriving the mean square errors of the four estimators, we obtained the following results:
- The standard estimator, the sample mean, has the largest MSE.
- We recommend the use of the ML estimator when skewness is not high.
- Otherwise we recommend the use of the conditionally minimal MSE estimator.
- Future research problem: searching for better estimators, particularly when the log-normal distribution does not hold.

Interval estimation of θ

Assume that $Y_i = \log W_i \sim N(\mu, \sigma^2)$. A commonly used (naive) interval:

- Construct a confidence interval based on log-transformed outcome, Y_i's, $\bar{Y} \pm Z_{1-\alpha/2} \frac{S}{\sqrt{n}}$.
- Transform the interval back to the original outcome,

$$exp(\bar{Y} \pm Z_{1-\alpha/2} \frac{S}{\sqrt{n}}).$$

- Since this interval is for $exp(\mu)$, not for θ, the naive interval is biased.

Interval estimation of θ, based on the original data

- Large-sample central limit theory:

$$\bar{W} \pm Z_{1-\alpha/2}\sqrt{\widehat{\tau}^2 n},$$

  where $\bar{W}$ and $\widehat{\tau}^2$ are the sample mean and variance of the original observations, W_i's.

- Student's interval of θ, based on the original data:

$$\bar{W} \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{\tau}^2 n},$$

New Interval One

- Note that $\log(\theta) = \mu + \sigma^2/2$.
- The biased corrected ML estimator of $\log\theta$: $\bar{Y} + S^2/2$ with the variance estimate, $S^2/n + S^4/(2m)$.
- a $1-\alpha$ level confidence interval for θ:

$$exp(\bar{Y} + S^2/2 \pm Z_{1-\alpha/2}\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}}).$$

- A simulation study suggests this interval has good coverage probability (Zhou and Gao (1997, Stat in Med)).
- Or, the modified version (Olsson, 2005, Journal of Statistical Education):

$$exp(\bar{Y} + S^2/2 \pm t_{1-\alpha/2,n-1}\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}}).$$

- The modified version outperforms the original one when the sample size is small.

Generalized confidence intervals

- Krishnamoorthy and Mathew (2003) proposed an interval for θ using the idea of generalized confidence intervals (Journal of statistical planning and inference, 115, 103-121)

Simulation study

- Olsson (2005) compared the performance of the naive approach, the new interval approach, the modified method with t instead of z as multiplier, the generalized confidence intervals, and the large-sample central limit method.
- The large-sample method gives a consistently lower coverage than the nominal level.
- The easy to compute new and modified confidence intervals perform well with the modified method being better in small sizes.
- The generalized confidence interval approach also works well; a small disadvantage is that it requires a computer to simulate the sampling distribution.

A real example

- We illustrate the method using data from a study of the effect of obesity on hospital charges following knee replacement (KR) procedures.
- This dataset consists of hospital charges for 355 obese patients following KR operations.
- The distribution of costs was skewed significantly toward higher cost patients, and the log transformed data approximate a normal distribution.
- A formal Shapiro-Wilk test for the normality on the log-transformed data gives a p-value of 0.25.

Results

The 95% confidence intervals for the mean of hospital charges:

- The naive method,

$$[\$8,839.6, \$9,363.7]$$

- The new method,

$$[\$9,326.0, \$9,893.2]$$

A general skewed distribution

- If the distribution of Wi is unknown, some modified t-intervals that have achieved limited success have been proposed using an Edgeworth expansion of the standard t-statistic.
  - Here are some references: Hall (1992, Biometrika) Sutton (1993, JASA) Chen (1995, JASA), Zhou and Gao (2000, Amer. Statist.).
- Future research: finding better non-parametric intervals

Statistical problem in one population with additional zeros

- Let $W_1, \ldots, W_n$ be a random sample from a lognormal distribution containing additional zero values. That is, if $W_i > 0$, $logW_i$ has $N(\mu, \sigma^2)$.
- Construction of an interval for $\theta = E(W_i) = P(W_i > 0)E(W_i \mid W_i > 0)$.

Statistical problem in one population with additional zeros

- Let $W_1, \ldots, W_n$ be a random sample from a lognormal distribution containing additional zero values. That is, if $W_i > 0$, $logW_i$ has $N(\mu, \sigma^2)$.
- Construction of an interval for $\theta = E(W_i) = P(W_i > 0)E(W_i \mid W_i > 0)$.

Interval estimation

- Owen and DeRouen (1980, Biometrics) derived a minimum variance unbiased estimator (MVUE) confidence interval for θ.
- Zhou and Tu (2000, Biometrics) have proposed a percentile-t bootstrap interval based on the sufficient statistics, a biased-corrected maximum likelihood (ML) estimation using normal approximation, and an interval based on the signed log-likelihood ratio test statistic. The bootstrap and likelihood ratio confidence intervals were recommended for means of lognormal data with zeros. But the methods perform not so well in very small sample situations.
- Tian (2005, Stat in Med) proposed an alternative generalized inference approach for confidence interval estimation and hypothesis testing.

Two populations without any zeros

- W_{i1}: the outcome variable of the ith patient in the first sample, $i = 1, \ldots, n_1$. W_{i2}: the outcome variable of the jth patient in the second sample, $i = 1, \ldots, n_2$. $M_j = E(Wij)$, and $\sigma_j^2 = Var(Wij)$.
- Assume that $X_{ij} = logW_{ij} \sim N(\mu_j, \sigma_j^2)$.
- The null hypothesis of interest is $H_0 : M_1 = M_2$.

Some existing methods

- Student's test on W_{ij}'s; it is valid when both n_1 and n_2 are large.
- Wilcoxon non-parametric test, based on W_{ij}'s
- Student's test on $logW_{ij}$'s
- Two-sample bootstrap test.

A simple valid test

- Zhou et al (1997, Biometrics) proposed the following test statistics:

$$Z = \frac{\widehat{\mu}_2 - \widehat{\mu}_1 + (1/2)(S_2^2 - S_1^2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + (1/2)(\frac{S_1^4}{n_1 - 1} + \frac{S_2^4}{n_2 - 1})}},$$

  where $\widehat{\mu}_j$ and S_j^2 are the jth sample mean and variance.

- The two-sided p-value = $2\Phi(- \mid Z \mid)$.

Simulation results

- The Z-score method has the observed type I error rate that is the closest to the pre-set nominal level even with small sample sizes ($n_1 = n_2 = 25$).
- The type I error rates of the other existing four tests are all larger than the nominal level and become larger as the difference in two skewness increases.
- With unequal sample sizes, the Z-score method has a greater advantage over the other tests.

More complicated tests

- Modified signed Likelihood ratio tests for the ratio of means of two independent log normal distributions (Wu et al, 2002).

Two extensions

- Zhou and Tu (2000, Computational Statistics & Data Analysis) proposed new confidence intervals for the difference in and the ratio of the means of cost data with additional zeros.
- Zhou et al. (2000, Stat in Med) proposed new tests for comparing means of health care costs in a paired design study.

Three or more populations with additional zeros

- Let $W_{1j}, \ldots, W_{n_j j}$ be a random sample from the jth population containing additional zeros and $M_j = E(W_{ij})$, $j = 1, \ldots, K$.
- Assume that for $W_{ij} > 0$, $Y_{ij} = logW_{ij} \sim N(\mu_j, \sigma_j^2)$. The null hypothesis of interest is $H_0 : M_1 = \ldots = M_K$.

New Tests

- Zhou and Tu (1999, Biometrics) proposed a computationally more complicated likelihood ratio test.
- Tu and Zhou (1999, Stat in Med) proposed a computationally simple Wald-type test.
- Simulation results suggest (1) that the likelihood ratio test has the best type I error rate (closest to the nominal level), and is closely followed by the Wald test, (2) that for unequal sample sizes, the likelihood ratio test has better coverage accuracy than the Wald test, and (3) that when the sample sizes are large, the type I error rates of the two tests are quite close to the nominal level.

Regression models for costs

- We are interested in the effect of patient-level factors (such as patients' medication compliance and patients¡® satisfaction with their health care providers) on health care costs of patients.

Existing regression models for skewed data

In statistical literature, there are four common ways of modeling skewed cost data.

1. The standard linear regression model with Ordinary least squares (OLS) without any transformation.
2. The Cox proportional hazards model (Dudley et al, 1993).
3. A parametric skewed distribution family for ϵ, leading to generalized linear models with an exponential family (Blough et al (1999)).
4. Transformation model so that transformed costs have a particular type of distribution, e.g. normal, homoscedastic, symmetric distribution, or remove extreme skewness with more efficient estimation (Ruppert (2001)).

Comments on OLS linear models

Advantages

1. Easy
2. No retransformation problem
3. Easy to compute marginal and Incremental effects

Disadvantages

1. Clear violation on normality and homoscedasticity.
2. Not robust in small to medium sized data set or in large datasets with extreme observations
3. Can obtain predictions with negative costs

Comments on Cox regression models

Advantages

1. Semi=parametric model without assuming the normality assumption.

Disadvantages

1. The regression coefficients in the Cox proportional hazards model pertain to the hazard ratio, it is difficult to interpret them in the context of health care costs

Comments on GLM

Advantages

1. No retransformation problems
2. Gains in precision from estimator of the assumed model holds
3. Consistent even if the variance function is misspecified.

Disadvantages

1. Can suffer substantial precision losses if heavy-tailed (log) error term (i.e., log-scale residuals have high kurtosis (> 3) or if variance function is misspecified

Extension of GLMs

- A non-parametric GLM by Chiou and Muller (1998, JASA): $g(E(W)) = X'\beta$, and $Var(W) = \sigma^2(E(W))$, where both $g(.)$ and $\sigma(.)$ are unknown functions.
- A semi-parametric GLM by Basu and Rathouz (2005, Biostatistics): $g_\lambda(E(W)) = X'\beta$, and $Var(W) = h(E(W); \gamma_1, \gamma_2)$, a parametric function of the mean with two unknown parameters. Here $g_\lambda(y)$ is a Box-Cox transformation function, and $h(E(W); \gamma_1, \gamma_2) = \gamma_1(E(W))^{\gamma_2}$, a power family or $h(E(W); \gamma_1, \gamma_2) = \gamma_1 E(W) + \gamma_2(E(W))^2$, a quadratic variance function.

Comparison of these two approaches

- Usually, generalized linear models and transformation models lead to different non-linear regression relationships between $E(W)$ and X.
- Which model is correct will depend on a particular application and whether we have additive or multiplicative errors.
- Unlike the transformation model, with the GLM we do not have the problem of re-transformation bias.
- The GLM addresses skewness by the choice of a distribution family, a commonly used one being a Gamma, Poisson, or negative binomial distribution, and tackles the non-linearity by the choice of its link function, commonly used ones being a log or square root link.
- Estimation of GLM still only use the first two moments of cost data, ignoring the information from the third and higher moments.

Advantages of transformation models over GLM

- When the expected value of W is related to a vector of covariates, X, in a complex way, often a transformation of W will simplify this relationship by inducing linearities or removing interactions (Ruppert, 2000). Suppose that

$$W = \beta_0 X_1^{\beta_1} \dots X_J^{\beta_J} + \epsilon,$$

a complicated non-linear model, where ϵ is a small random error. Because $log(\mu + \epsilon) \approx log(\mu) + \epsilon/\mu$ for small values of ϵ, we obtain that $log(W)$ follows approximately the linear model

$$log(W) = \beta_0^* + \beta_1 X_1^* + \ldots + \beta_J X_J^* + (\beta_0 X_1^{\beta_1} \dots X_J^{\beta_J})^{-1} \epsilon,$$

where $\beta_0^* = log(\beta_0)$, and $X_j^* = log(X_j)$.

Advantages of transformation models over GLM, cont

- The transformation method can induce a particular type of distribution or remove extreme skewness so that more efficient estimators and more appropriate plotting can be obtained. For example, for highly skewed health care costs, where most of the data are crowded into the lower left-hand corner of the plot, it is hard to see what type of regression model is appropriate between the outcome and a predictor.
- A transformation that can induce the symmetric distribution of transformed data would make it much easier to see a relationship between the transformed outcome and a predictor.
- The transformation method for estimating $E(W)$ uses information from higher moments.

Re-transformation bias

- One complication in the use of a transformation model is possible re-transformation bias.
- As we are interested in dollar scale not in transformed-dollar scale, we have to transform regression results on the transformed scale back to results on original dollar-scale.
- If this re-transformation is not properly done, we may introduce bias in our results called "re-transformation bias".

Existing semi-parametric transformation model

- Welsh and Zhou (2006) and Zhou and Cheng (2008) proposes the following semi-parametric transformation model to analyze skewed and heteroscedastic variance data:

$$H(Y) = \boldsymbol{X}'\boldsymbol{\beta} + \sigma(\boldsymbol{X}'\gamma)\varepsilon, \tag{0.1}$$

  where Y is a scalar dependent variable, $H(\cdot)$ is a known increasing transformation function, $\sigma(\cdot)$ is the known variance function, $\boldsymbol{X}$ is a $q \times 1$ vector of observed explanatory variables with the first element being 1, $\boldsymbol{\beta}$ and γ are vectors of unknown parameters, and ε is an error term that has a unknown distribution F with mean 0 and variance 1. Here the error term ε is independent of $\boldsymbol{X}$.

- In the model (0.4), we allow the effect of $\boldsymbol{X}$ on the mean and variance to be different.

Software

- We developed a computer program in R.
- It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- Our program can run on R 2.0.0 and later releases.
- The program is available from http://faculty.washington.edu/∼azhou/Heter/.

Software, cont

- Our program reads in patient data files in plain text format
- Estimates regression parameters $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\theta}$
- Then computes both the externally weighted estimator $\widehat{u}^*$ and the internally weighted estimators $\widehat{u}$ of mean on the original scale.
- Welsh and Zhou (2005) showed that both the internally weighted estimator and the externally weighted estimator have very similar bias and MSE.
- While the internally weighted estimator has slightly smaller bias then the externally weighted estimator, the externally weighted estimator has slightly smaller MSE than the internally weighted estimator.
- Therefore, we give both estimators so that users can choose which one to use.

Software, cont

- In addition to these two mean estimators, our program also outputs their statistics, including standard deviation, asymptotic confidence interval, and an option for bootstrap confidence interval.
- Our program can run in both interactive mode and batch mode.

An example

- We illustrate the use of our computer program in a data set on hypertension patients from a prospective drug utilization review (DUR) study.
- A goal of our analysis is to estimate the average of in-patient charges of a patient given his/her age, gender, race, and general health status as measured by SF-36.
- Since the in-patient charges are zero for some patients, we apply a two-stage heteroscedastic regression model to our data set.

An example, cont

- Let Y_i be the in-patient charge of the ith patient, and corresponding covariates are defined as follows.
- X_{i1} is the age of the patient; X_{i2} represents the patient's race ($X_{i2} = 1$ for Caucasians and $X_{i2} = 0$ for African Americans); X_{i3} represents the gender of the patient ($X_{i3} = 1$ for males and $X_{i3} = 0$ for females); X_{i4} is the score based on 100 representing ith patient's general health status.

An example, cont

- For $i = 1, \ldots, n$, we model the probability of non-zero in-patient charge by the logistic regression model,

$$\log \frac{P(Y_i = 0 \mid X_{i1}, \ldots, X_{i4})}{P(Y_i > 0 \mid X_{i1}, \ldots, X_{i4})} = \alpha_0 + \alpha_1 X_{i1} + \ldots \alpha_4 X_{i4}, \tag{0.2}$$

- and we model the conditional magnitude of the positive charges Y_i given $Y_i > 0$ by the log-transformed, heteroscedastic linear regression model

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_4 X_{i4} + \exp\{(\theta_0 + \theta_1 X_{i1} + \ldots + \theta_4 X_{i4})/2\}\epsilon_i. \tag{0.3}$$

Input data

```
0         56 0 0 32
0         64 1 1 25
1952.05   68 1 1 42
 0        54 1 1 50
 ......
```

Input data

```
Please specify the name of input data file: example.dat Please
choose method for estimating parameters beta and theta
  1: MLE estimator, maximize log-likelihood function (default)
  2: MLE estimator, solve estimating equation
1
 Do you want to give initial guess of parameters (default: all
zeros)? (Y/N)n alpha.0 = 0 0 0 0 0 beta.0  = 0 0 0 0 0 gamma.0 = 0 0
0 0 0
 Please specify (1-r) confidence interval (default r=0.05):
  r =
  r =  0.05
Do you want to calculate bootstrap confidence interval? (Y/N)y
Please specify bootstrap sample size (default 100): 100
 Do you want to assign seed for random number generator? (Y/N)n
  Seed not assigned.
```

Commands

```
Please specify the name of input data file: example.dat
 Please choose method for estimating parameters beta and theta
  1: MLE estimator, maximize log-likelihood function (default)
  2: MLE estimator, solve estimating equation
1
 Do you want to give initial guess of parameters (default: all
zeros)? (Y/N)n alpha.0 = 0 0 0 0 0 beta.0  = 0 0 0 0 0 gamma.0 = 0 0
0 0 0
 Please specify (1-r) confidence interval (default r=0.05):
  r =
  r =  0.05
 Do you want to calculate bootstrap confidence interval? (Y/N)y
 Please specify bootstrap sample size (default 100): 100
 Do you want to assign seed for random number generator? (Y/N)n
 Seed not assigned.
```

Commands, cont

```
Please input covariates values: 1: 55 1 0 50 Covariate =  55 1 0 50
Do you want to see estimation results for another covariate? (Y/N) y
Please input covariates values: 1: 65 1 0 50 Covariate =  65 1 0 50
Do you want to see estimation results for another covariate? (Y/N) n
```

Results

```
estimator of parameter alpha = 1.513531 -0.008025662 0.4702371
 0.3740364 0.007434743
 std of alpha estimator = 0.7334805 0.01321834 0.3182745 0.2856736
 0.006093177
estimator of parameter beta = 9.538691 -0.004213263 -0.8231364
0.02954837 0.003566381
 std of beta estimator = 0.701519 0.0126878
0.3566817 0.2814012 0.005726485
 estimator of parameter theta =
-0.9736211 0.05375912 -1.048779 -0.5785688 -0.01572478
  std of theta
estimator = 0.9676426 0.01870368 0.4402539 0.3763115 0.00826068
```

Results, cont

```
Source data file:        example.dat Number of observations:  483
Number of covariates:    4
-------------------------------------------------------------
Covariate = 2 2 2 2
 Externally weighted estimator:
                              mean = 111.6968
                 standard deviation = 145.8094
           95% confidence interval = [0, 397.478]
95% bootstrap confidence interval = [0.7717479, 3523.01]
Internally weighted estimator:
                              mean = 111.3817
                 standard deviation = 144.0542
           95% confidence interval = [0, 393.7227]
95% bootstrap confidence interval = [0, 3539.867]
-------------------------------------------------------------
```

A new non-parametric transformation model

- We propose the following nonparametric transformation model to analyze skewed and heteroscedastic variance data:

$$H(Y) = \boldsymbol{X}'\boldsymbol{\beta} + \sigma(\boldsymbol{X}'\gamma)\varepsilon, \tag{0.4}$$

  where Y is a scalar dependent variable, $H(\cdot)$ is a unknown increasing transformation function, $\sigma(\cdot)$ is the known variance function, $\boldsymbol{X}$ is a $q \times 1$ vector of observed explanatory variables with the first element being 1, $\boldsymbol{\beta}$ and γ are vectors of unknown parameters, and ε is an error term that has a unknown distribution F with mean 0 and variance 1. Here the error term ε is independent of $\boldsymbol{X}$.

- In the model (0.4), we allow the effect of $\boldsymbol{X}$ on the mean and variance to be different.

Identifiability Assumptions

- To make the model (0.4) identifiable, we need to make the following assumptions.
- Like Horwitz (1996), we assume that there exists y_0 such that $H(y_0) = 0$.
- Re-arrange X so that its first component X_1 has the absolute continuous density conditional on $X_2, \ldots, X_p$. Let β_1 be the corresponding coefficient. $\mid \beta_1 \mid = 1$.

Some special cases

- Model (0.4) includes as special cases a large number of widely used and extensively investigated models that make stronger assumptions than the ones made in the paper about H and F.
- Linear regression models, log-linear regression models, the Cox proportional hazard model, and accelerated failure time models
- Transformation models in which H is specified up to a vector of finite-dimensional parameters (e.g.,Box and Cox, 1964; Bickel and Doksum, 1981)
- Transformation models in which F is specified up to a vector of finite-dimensional parameters, and H is nonparametric (Cheng et al.1995, Dabrowska and Doksum 1988).

Some notation

- Let $\{Y_i, \boldsymbol{X}_i, i = 1, \cdots, n\}$ be a random sample of (Y, X) that satisfies the model (0.4).
- Denote $Z_1 = \boldsymbol{X}'\boldsymbol{\beta}$, $Z_2 = \boldsymbol{X}'\gamma$, $Z_{1i} = \boldsymbol{X}_i'\boldsymbol{\beta}$, and $Z_{2i} = \boldsymbol{X}_i'\gamma$.
- Let $G(\cdot|z_1, z_2)$ be the cumulative distribution function (CDF) of Y conditional on $Z_1 = z_1$ and $Z_2 = z_2$, and $p(\cdot, \cdot)$ be the probability density function of (Z_1, Z_2).
- Assume that H, F, and G are differentiable with all their arguments.
- Define $h(y) = dH(y)/dy$, $f(y) = dF(y)/dy$, $p(y|z_1, z_2) = \partial G(y|z_1, z_2)/\partial y$, and $g_j(y|z_1, z_2) = \partial G(y|z_1, z_2)/\partial z_j, j = 1, 2$.

Estimation method

- Since under the model (0.4), Y depends on $\boldsymbol{X}$ only through the index Z_1 and Z_2, the model (0.4) implies that

$$G(y|z_1, z_2) = F(\frac{H(y) - z_1}{\sigma(z_2)}),$$

  and we can show that $p(y|z_1, z_2) = -g_1(y \mid z_1, z_2)h(y)$.

- Denote $g_1(y, z_1, z_2) = g_1(y|z_1, z_2)p(z_1, z_2)$ and $p(y, z_1, z_2) = p(y|z_1, z_2)p(z_1, z_2)$, we get

$$g_1(y, z_1, z_2)h(y) = -p(y, z_1, z_2). \tag{0.5}$$

- Hence we obtain that $h(y) = -\frac{\sum_{i=1}^n p(y, Z_{1i}, Z_{2i})}{\sum_{i=1}^n g_1(y, Z_{1i}, Z_{2i})}$, and

$$H(y) = -\int_{y_0}^{y} \frac{\sum_{i=1}^n p(u, Z_{1i}, Z_{2i})}{\sum_{i=1}^n g_1(u, Z_{1i}, Z_{2i})} du. \tag{0.6}$$

- The expression (0.6) forms the basis for the estimator of H proposed here.

Estimation of $H(.)$

- From (0.6), we see that to derive an estimator of $H(.)$, we need to estimate $p(z_1, z_2)$, $G(y|z_1, z_2)$ and derivatives of $G(y|z_1, z_2)$ when the values of $\boldsymbol{\beta}$ and γ are given.

Estimation of $H(.)$, cont

- We estimate $G(y|z_1, z_2)$ by the following kernel estimator:

$$G_n(y|z_1, z_2) = \frac{1}{nh_1h_2p_n(z_1, z_2)} \sum_{i=1}^{n} I(Y_i \leq y) K_1\left(\frac{Z_{1i} - z_1}{h_1}\right) K_2\left(\frac{Z_{2i} - z_2}{h_2}\right), \quad (0.7)$$

  where K_1 and K_2 be bounded and symmetric kernel functions with the support $[-1, 1]$ with h_1 and h_2 being bandwidths.

- Here $p_n(z_1, z_2)$ is the kernel density estimate of $p(z_1, z_2)$, and is given by

$$p_n(z_1, z_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K_1\left(\frac{Z_{1i} - z_1}{h_1}\right) K_2\left(\frac{Z_{2i} - z_2}{h_2}\right). \quad (0.8)$$

Estimation of $H(.)$, cont

- Since $g_1(y|z_1, z_2) = dG(y|z_1, z_2)/dz_1$, we obtain an estimator of $g_1(y|z_1, z_2)$ by differentiating $G_n(y|z_1, z_2)$ with respect to z_1,

$$g_{1n}(y|z_1, z_2) = \partial G_n(y|z_1, z_2)/\partial z_1. \tag{0.9}$$

- Although $p(y|z_1, z_2)$ is the probability density function of Y conditional on $Z_1 = z_1, Z_2 = z_2$, it can not be estimated by $\partial G_n(y|z_1, z_2)/\partial y$ because $G_n(y|z_1, z_2)$ is a step function of y.
- Instead, we use the following kernel density estimator for $p(y|z_1, z_2)$:

$$p_n(y|z_1, z_2) = \frac{1}{nh_1h_2h_0p_n(z_1, z_2)} \sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0}) K_1\left(\frac{Z_{1i} - z_1}{h_1}\right) K_2\left(\frac{Z_{2i} - z_2}{h_2}\right), \tag{0}$$

where K_0 be bounded and symmetric kernel functions with the support $[-1, 1]$ with h_0 being a bandwidth.

Estimation of $H(.)$, cont

By substituting (0.8), (0.9) and (0.10) into (0.6), we obtain the estimator H_n of H,

$$H_n(y) = -\int_{y_0}^{y} \frac{\sum_{i=1}^{n} p_n(u|Z_{1i}, Z_{2i}) p_n(Z_{1i}, Z_{2i})}{\sum_{i=1}^{n} g_{1n}(u|Z_{1i}, Z_{2i}) p_n(Z_{1i}, Z_{2i})} du. \tag{0.11}$$

Estimation of β and γ

- Since $E((H(Y) - Z_1)^2 \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\gamma)$, we can use the following estimating equations to simultaneously estimate $\boldsymbol{\beta}$ and γ:

$$\sum_{i=1}^{n} \frac{(H(Y_i) - \boldsymbol{X}_i'\boldsymbol{\beta})\boldsymbol{X}_i}{\sigma^2(\boldsymbol{X}_i'\gamma)} = 0, \tag{0.12}$$

and

$$\sum_{i=1}^{n} \left\{(H(Y_i) - \boldsymbol{X}_i'\boldsymbol{\beta})^2 - \sigma^2(\boldsymbol{X}_i'\gamma)\right\} \boldsymbol{X}_i = 0, \tag{0.13}$$

when given H.

- From the equation (0.12), we obtain a closed-form estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_n = \left(\sum_{i=1}^{n} \frac{\boldsymbol{X}_i\boldsymbol{X}_i'}{\sigma^2(\boldsymbol{X}_i'\gamma)}\right)^{-1} \sum_{i=1}^{n} \frac{\boldsymbol{X}_i H(Y_i)}{\sigma^2(\boldsymbol{X}_i'\gamma)}. \tag{0.14}$$

Estimation algorithm of $\boldsymbol{\beta}, \gamma$ and $H(\cdot)$.

1. Selection of initial values.
   (a) Initial values for $\boldsymbol{\beta}$ and H. We can still obtain consistent estimates for $\boldsymbol{\beta}$ and H even we misspecify the variance function, and hence, we can obtain reasonable starting values for $\boldsymbol{\beta}$ and H with estimates obtained under the homoscedasticity model,

$$H(Y) = \boldsymbol{X}'\boldsymbol{\beta} + \sigma\varepsilon. \tag{0.15}$$

   Under this homoscedastic model, we can estimate $\boldsymbol{\beta}$ by the maximum rank correlation (MRC) method proposed by Han(1987); that is, we estimate $\boldsymbol{\beta}$ with $\widetilde{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} W_n(\boldsymbol{\beta})$, where $W_n(\boldsymbol{\beta}) = \sum_{i \neq j} \{Y_i > Y_j\}\{\boldsymbol{X}_i'\boldsymbol{\beta} > X_j'\boldsymbol{\beta}\}$. And then we can estimate H using the proposed method with large enough h_2 so that $K_2(\frac{Z_{2i}-z_2}{h_2}) = 1$ for any z_2 and $i = 1, \cdots, n$.
   (b) Initial values for γ. Given $\boldsymbol{\beta}$ and H, we estimate γ by the equation (0.13).
2. Estimation of $H(.)$. Given $\boldsymbol{\beta}$ and γ, we estimate H by (0.11).
3. Estimation of $\boldsymbol{\beta}$ and γ. Given H, we estimate $\boldsymbol{\beta}$ and γ by (0.14) and (0.13).
4. Iteration. Repeat Steps 2 and 3 until two successive values of $\boldsymbol{\beta}$ and γ don't differ significantly.

Prediction of $\mu(\mathbf{x}) = \mathbf{E}(\mathbf{Y} \mid \boldsymbol{X} = \mathbf{x})$

- For given covariates $\boldsymbol{x}$ of a patient, we are interested in predicting $\mu(\boldsymbol{x})$. Under the model (0.4), we can write

$$\mu(\boldsymbol{x}) = \int H^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + \sigma(\boldsymbol{x}^T\gamma)u)dF(u). \tag{0.16}$$

-

- We propose to estimate F by the empirical distribution $\widehat{F}$ of the standardized residuals, $\widehat{e}_i = \frac{\widehat{H}(Y_i) - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i^T\widehat{\gamma})}$, where $\widehat{H}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ are the estimators of H, $\boldsymbol{\beta}$ and γ.

Non-parametric estimator

- Therefore, replacing H, $\boldsymbol{\beta}$, γ, and F by $\widehat{H}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}$, and $\widehat{F}$ in (0.16), we obtain the following estimator of $\mu(\boldsymbol{x})$:

$$\widehat{\mu}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}^{-1}\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}} + \sigma(\boldsymbol{x}'\widehat{\gamma})\frac{\widehat{H}(Y_i) - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i'\widehat{\gamma})}\right). \tag{0.17}$$

- This estimator can be considered as an extension of Duan's smearing estimator (Duan, 1982) to the heteroscedastic transformation model with the unknown transformation and error distribution functions.

Simulation studies

- We conduct a simulation study to assess the finite-sample performance of the proposed method.
- Unlike the existing parametric or the single-semiparametric models, where one of the transformation and error distribution functions is specified, the validity of our method does not rely on parametric specifications for both the transformation and error distribution functions.
- Hence we expect our estimators of the untransformed scale expectation and regression parameters are more robust than the ones derived under the existing parametric and single-semiparametric methods.
- We also want to know whether the added robustness is gained at the expense of reduced efficiency.

Simulation studies, cont

- To investigate this, we compare the performance of the proposed method with the following models:
  1. the CTCD model, where the transformation and error distribution functions are correctly specified by a parametric model, the case that serves as the gold standard,
  2. the CTMD model, where the transformation is correctly specified, but the error distribution is misspecified, and
  3. the MTCD model, where the error distribution is correctly specified, but the transformation function is misspecified. The CTCD model is used to investigate the efficiency of the proposed method, and the MTCD and CTMD models are used to investigate the robustness of the proposed method.
  4. Two existing methods, Basu's method and Chiou's method.

First simulation study

- In our first simulation study, we consider a true transformation regression model with one binary covariate, one continuous covariate, and a non-logarithm transformation function. For $n = 2000$ subjects, we generate covariates X_1 and X_2 from the binomial distribution with $p = 0.5$ and the uniform distribution on $[0, 2]$, respectively, the random error ε from the standard normal distribution. We let our outcome follow the following transformation model:

$$H(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \sqrt{0.4 + \gamma X_1}\varepsilon,$$

  where $H(y) = \Phi^{-1}\{\exp(y - 10)\}$, $\beta_0 = -1.8$, $\beta_1 = 1.4$ $\beta_2 = 1.4$, and $\gamma = -0.35$.
- We assess the performance of the various estimators of the regression parameters and the untransformed scale expectation in terms of standard deviation (SD), bias, and the root of mean squared error (RMSE).
- In the MTCV model, the transformation function is misspecified as a function $H(y) = exp(y - 10)$.

Simulation studies, cont

- We present the results of RMSE and related quantities for the untransformed scale expectation at the combination of $x_1 = 1, 1.5, 2$ and $x_2 = 0.5$ in Table below based on 200 simulated data sets.

First simulation results on predicted means

x_1	x_2	Average value	Method	Bias	SD	RMSE
0	0.0	6.5023	Proposed	0.1362	0.2166	0.2558
			CTCV	0.0015	0.0884	0.0884
			CTMV	0.0022	0.0874	0.0874
			MTCV	2.8103	0.0182	2.8104
			CHIOU	0.8027	0.2261	0.8340
			BASU	Failed to converge		
0	1.0	8.7948	Proposed	-0.0062	0.0420	0.0424
			CTCV	-0.0030	0.0277	0.0279
			CTMV	-0.0015	0.0253	0.0253
			MTCV	0.7702	0.0227	0.7705
			CHIOU	-0.3023	0.0627	0.3087
			BASU	Failed to converge		

First simulation results on predicted means, cont

x_1	x_2	Average value	Method	Bias	SD	RMSE
1	0.0	8.9171	Proposed	0.0384	0.0334	0.0509
			CTCV	0.0015	0.0358	0.0358
			CTMV	0.0019	0.0349	0.0350
			MTCV	0.6480	0.0050	0.6480
			CHIOU	0.1796	0.1890	0.2607
			BASU	Failed to converge		
1	1.0	9.8181	Proposed	-0.0057	0.0088	0.0105
			CTCV	-0.0001	0.0021	0.0021
			CTMV	-0.0001	0.0021	0.0021
			MTCV	-0.0724	0.0047	0.0726
			CHIOU	-0.0637	0.0227	0.0677
			BASU	Failed to converge		

Conclusion

- Among the two existing GLM-based estimators, we found that Baus & Rathouz's procedure failed to converge for all of 200 simulated samples, suggesting that the Baus & Rathouz's estimator is not stable.
- Chiou and Muller's estimator has much larger bias and SD than our newly proposed estimator, and is badly biased.
- By comparing results among the parametric CTCV, CTMV, and MTCV estimates, we conclude that misspecification of the transformation function can lead to large bias and large RMSE while misspecification of variance function has minimal effect on bias and RMSE.
- Comparing our new estimator with the gold standard estimator, the CTCV estimator, derived under correctly specified transformation and variance function, the empirical efficiency of our new estimator is around 60% on average.

Second simulation study

- In our second simulation study, we considers a true transformation regression model with a logarithm transformation function, a constant variance, and an asymmetrical distribution error.
- For $n = 2000$ subjects, we generate covariates X_1 and X_2 from the binomial distribution with $p = 0.5$ and the uniform distribution on $[0, 1]$, respectively, and the random error ε from the Gamma distribution with scale of 1 and shape of 4.
- We let our outcome follow the following transformation model:

$$H(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon/2,$$

where $H(y) = 2\log(y)$, $\beta_0 = -8$, $\beta_1 = 4$, and $\beta_2 = 4$.

Second simulation results on predicted means

x_1	x_2	Average cost	Method	Bias	SD	RMSE
1	0.0	0.4287	Proposed	0.0276	0.0191	0.0336
			CTCV	-0.0012	0.0125	0.0126
			MTCV	0.2068	0.0292	0.2088
			BASU*	-0.0035	0.0176	0.0180
			CHIOU**	0.1308	0.0639	0.1456
1	0.5	1.1653	Proposed	-0.0113	0.0496	0.0509
			CTCV	-0.0038	0.0238	0.0241
			MTCV	0.2009	0.0418	0.2052
			BASU*	-0.0078	0.0265	0.0276
			CHIOU**	0.1470	0.0466	0.1542

* The procedure failed to converge in 37 out of the 200 samples, the summaries are based on the 163 remaining samples. ** The procedure failed to converge for 23 out of the 200 samples, the summaries are based on the 177 remaining samples.

An example

- The sample used here was from a study on the effectiveness of the Improving Mood-Promoting Access to Collaborative Treatment (IMPACT) collaborative care management program for late-life depression (Unutzer, et al., 2002).
- In this talk, we focus on the total outpatient cost in the first year (Y), the mean and standard deviation of Y are $\$6258.442$ and $\$5065.507$, respectively, and the coefficients of skewness and kurtosis of Y are 3.36 and 26.94, respectively.
- We fit the data using our model with the outcome variable being outpatient costs in the first year, and the two independent variable, X_1 and X_2.
- Here X_1 was the binary treatment indicator, and X_2 was the mean score of the 20 depression items from the Symptom Checklist.

An example, cont

- We set the variance function to be a polynomial function $\sigma^2(x;p) = \sum_{k=0}^{p} \alpha_k x^k$, $p = 1, 2, \cdots$, where p was chosen to minimize the following $GF(p)$:

$$GF(p) = \min_{\gamma} \sum_{i=1}^{n} \left\{ \left(\tilde{H}(Y_i) - \boldsymbol{X}_i' \tilde{\boldsymbol{\beta}} \right)^2 - \sigma^2(\boldsymbol{X}_i' \gamma; p) \right\}^2,$$

  where $\widetilde{H}$ and $\tilde{\boldsymbol{\beta}}$ were the initial values of H and $\boldsymbol{\beta}$, respectively.

- The results showed that $GF(p)$ did not substantially change with p.
- Hence to assure $\sigma^2(x) \geq 0$, here we selected $\sigma^2(x;p) = x^2$.

The estimates of the expected costs of a patient

Rand	SCL	Proposed Expection(se)	CHIOU Expection(se)	BASU Expection(se)
1	0.04	5008.6(444.4)	5239.9(74.7)	5334.8(213.0)
0	0.04	4424.0(500.6)	4639.4(10.5)	4991.9(390.1)
1	1.50	6916.1(392.9)	6779.1(26.7)	6717.8 (57.6)
0	1.50	6172.1(216.4)	6177.7(40.7)	6167.4(2.2)
1	3.20	9177.3(1156.9)	8574.4(31.9)	9802.8(3921.1)
0	3.20	8349.5(816.3)	7971.8(97.5)	8622.3(1668.5)

The estimated transformation and its 95% confidence limits for IMPACT data.