THE TAMING OF THE SKEW: MODELS FOR HEALTH CARE COSTS (ATTRIBUTABLE TO SUPPLEMENTAL INSURANCE IN MEPS)

ANDREW M. JONES

HEALTH, ECONOMETRICS AND DATA GROUP (HEDG) & DEPARTMENT OF ECONOMICS AND RELATED STUDIES
UNIVERSITY OF YORK

SOME NOTATION:

- $s$: size of total population
- $s_n$: number of (ever-) “treated”
- $\pi$: proportion of treated
- $s_m$: average medical costs for treated
- $s_l$: average attributable medical costs for treated
- $s_n$: average non-attributable medical costs for treated
- $s$: population average health care costs
- $M$: population total health care costs ($=n \cdot s$)
- $d$: binary indicator for ever treated
- $S$: vector for treatment history

ATTRIBUTABLE COSTS – SOME DEFINITIONS

Total attributable costs (object of interest):

\[ C = n_s s = n \pi s \]  

[C.0]

Attributable fraction:

\[ \theta = \frac{E(m_s | S, d = 1) - E(m_s | 0, d = 1)}{E(m_s | S, d = 1)} = \frac{l + s - l}{l + s} = \frac{s}{m_s} \]

Which implies:

\[ s = \theta (l + s) = \theta m_s \]

and hence:

\[ C = n_s s = n \pi s = n \pi \theta m_s \]  

[C.1]

ESTIMATION

Using [C.1]:

\[ C = n_s s = n \pi s = n \pi \theta m_s \]

Can use population (census) data for $n$, sample data for $\pi$ (prevalence of treatment), and sample data for $m_s$ (average costs for treated).

Use model to estimate $\theta$ (attributable fraction).
Estimation of attributable fraction:

\[
\theta = \frac{E(m \mid S, d = 1) - E(m \mid 0, d = 1)}{E(m \mid S, d = 1)}
\]

This involves estimation of the counterfactual:

\[
E(m \mid 0, x, d = 1)
\]

Using treated parameter values and x values (actual data), with S set to zero. Can let form of \(E(m \mid .)\) differ between treated and never-treated (e.g. interactions between d and x). Identified by variation in x (risk factors) within treated. S identified within treated and relative to never-treated.

Attributable Fractions with exponential (ECM) specifications

\[
\frac{E(M \mid S, X) - E(M \mid 0, X)}{E(M \mid S, X)} = 1 - \frac{E(M \mid 0, X)}{E(M \mid S, X)}
\]

1-part ECM:

For

\[
E(M \mid S, X) = \exp(\delta S + \beta'X + \gamma'S X)
\]

\[
\frac{E(M \mid S, X) - E(M \mid 0, X)}{E(M \mid S, X)} = \frac{\exp(\delta + \beta'X + \gamma'S X) - \exp(\beta'X)}{\exp(\delta + \beta'X + \gamma'S X)}
\]

\[
= 1 - \frac{\exp(\delta + \gamma'S X)}{\exp(\delta + \gamma'S X)}
\]

\[
= 1 - \exp(-\delta - \gamma'S X)
\]

STUDY DESIGN: for observational study of attributable costs

Take account of complex survey design (stratification, clustering, over-sampling etc)

Missing data (patterns of missing data, imputation etc)

Nonparametric preprocessing (Using matching to reduce imbalance in empirical distribution of covariates prior to parametric modelling)

Cost models (estimate counterfactual costs using (semi-) and parametric regression methods)
The MEPS Data

The practical application is based on the datasets from *Microeconometrics using Stata* by Cameron and Trivedi (Stata Press, 2009).

The original source of the data is the US Medical Expenditure Survey (MEPS).

The MEPS is a set of surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.) and employers across the US. The surveys collect data on the use of health services (e.g. frequency and cost) and whether individuals hold health insurance.

For more information on the MEPS refer to the website: [http://www.meps.ahrq.gov/mepsweb/](http://www.meps.ahrq.gov/mepsweb/).

Working Dataset

The working dataset is based on the MEPS sample used in Chapter 3 (p71) of Cameron and Trivedi. The data is “mus03data.dta”. Cameron and Trivedi describe the data as follows:

“We analyze medical expenditure of individuals aged 65 years and older who qualify for health care under the U.S. Medicare program. … Medicare does not cover all medical expenses. For example, copayments for medical services and expenses of prescribed pharmaceutical drugs were not covered for the time period studies here. About half of eligible individuals therefore purchase supplementary insurance in the private market that provides insurance coverage against various out-of-pocket expenses.” (p71)

Health care expenditures are measured in US dollars. Sociodemographic and health-status measures are also available together with insurance status.

Description of total health expenditure per annum

```
. tabstat y, stat (n mean med sd cv iqr sk k mi ma) by(d)
Summary for variables: y
by categories of: d
          | N    | mean     | p50     | sd   | cv     | iqr   | skewness
---------+------+----------+----------+------+--------+-------+----------
 0       | 1207 | 6824.303 | 2779    | 11425.94 | 1.674301 | 5462 | 3.868345
 1       | 1748 | 7611.963 | 3648.5  | 12358.83 | 1.623607 | 6387.5 | 4.232459
---------+------+----------+----------+------|--------+-------+----------
  Total   | 2955 | 7290.235 | 3334    | 11990.84 | 1.644781 | 6064 | 4.113807
```

```
. graph vars(d costs), title(Description of total health expenditure per annum)
```

Histograms for total expenditure

```
. graph vars(d costs), title(Histograms for total expenditure)
```

Andrew M Jones, June 2009
Issues with modelling health care costs

- Typically non-normal:
  - Skewness and kurtosis (long & heavy right-hand tails)
  - Severe patients may attract substantial services
  - Relatively rare events/procedures might be very expensive
  - Minority of patients responsible for a high proportion of health care costs
  - Mean >> median
- Heteroskedastic
- Might have a mass point at zero (truncated at zero)
- The relationship between costs and covariates may not be linear
**Linear regression (OLS on y)**

```stata
. regress y $xs //non-robust
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 2955</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4.941e+10</td>
<td>7</td>
<td>7.058e+09</td>
<td>F( 7, 2947) = 55.42</td>
</tr>
<tr>
<td>Residual</td>
<td>3.7532e+11</td>
<td>2947</td>
<td>127357036</td>
<td>R-squared = 0.1163</td>
</tr>
<tr>
<td>Total</td>
<td>4.2473e+11</td>
<td>2954</td>
<td>143780269</td>
<td>Adj R-squared = 0.1142</td>
</tr>
</tbody>
</table>

```

```
| y    | Coef.  | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|------|--------|-----------|-------|-------|----------------------|
| d    | 724.8632 | 433.8887  | 1.67  | 0.095 | -125.8925            |
| phylim | 2389.019 | 534.7384  | 4.47  | 0.000 | 1340.52              |
| actlim | 3900.491 | 582.9913  | 6.69  | 0.000 | 2757.379             |
| totchr | 1844.377 | 172.9187  | 10.67 | 0.000 | 1505.323             |
| actlim | 3900.491 | 582.9913  | 6.69  | 0.000 | 2757.379             |
| female | -1383.29 | 427.4854  | -3.24 | 0.001 | -2221.49             |
| income |   6.46894 | 9.56821 | 0.68  | 0.499 | -12.29211            |
| _cons | 8358.954 | 2597.715  | 3.22  | 0.001 | 3265.434             |
```

**Copas Test (for over-fitting)**

```stata
. program copas, eclass
1.         version 10.1
2.         tempname b
3.         tempvar run
4.         tempvar yfa
5.         tempvar yfb
6.         tempvar ea
7.         tempvar eb
8.         gen `run'=runiform()
9.         regress y $xs if `run'<=0.5, vce(robust)
10.        predict `yfa'
11.        regress y $xs if `run'>0.5, vce(robust)
12.        predict `yfb'
13.        replace yfc=`yfa' if `run'>0.5
14.        replace yfc=`yfb' if `run'<=0.5
15.        regress y yfc
16.        matrix `b'=e(b)
17.        ereturn post `b'
18.        end
```

**Bootstrap results (OLS on y)**

```stata
. bootstrap _b, reps(100) seed(12345) nowarn nodots: copas
```

```
Bootstrap results                               Number of obs      =      2955
Replications       =       100

------------------------------------------------------------------------------
|     Observed  Bootstrap                         Normal-based              |
|      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| yfc    | .9826713 | .0252514  | 38.92 | 0.000 | .9331795    1.032163 |
| _cons  | 127.4473 | 183.2907  | 0.70  | 0.487 | -231.7958    486.4904 |
```

```
. test yfc==1
( 1)  yfc = 1

chi2( 1) = 0.47
Prob > chi2 = 0.4926
```
**Counterfactual predictions - attributable fraction and costs**

```
. * Counterfactual predictions - attributable fraction and costs
. preserve
. replace d=0
. predict y0 if e(sample)
. gen attf=(yf-y0)/yf
. gen attc=yf-y0
. summ y yf y0 attf attc
. summ y yf y0 attf attc if d_raw==1
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1748</td>
<td>7611.963</td>
<td>12358.83</td>
<td>3</td>
<td>125610</td>
</tr>
<tr>
<td>yf</td>
<td>1748</td>
<td>7611.963</td>
<td>4068.397</td>
<td>502.9237</td>
<td>22559</td>
</tr>
<tr>
<td>y0</td>
<td>1748</td>
<td>6887.1</td>
<td>4068.397</td>
<td>-221.9395</td>
<td>21834.14</td>
</tr>
<tr>
<td>attf</td>
<td>1748</td>
<td>.1412443</td>
<td>.1307571</td>
<td>.0321319</td>
<td>1.441298</td>
</tr>
<tr>
<td>attc</td>
<td>1748</td>
<td>724.8632</td>
<td>.0002073</td>
<td>724.8623</td>
<td>724.8643</td>
</tr>
</tbody>
</table>

. restore

So the estimated attributable fraction, \( \theta \) = 0.095

---

**Linear Regression on Transformations**

- Consider transforming data to produce a more symmetric distribution
  - Log transformation
  - Square-root transformation
  - Other power functions
- No longer working on raw cost scale

---

**Log transformation & retransformation**

- Model:
  \[
  \ln y_i = \beta x_i + \epsilon
  \]
- If error term is normally distributed, \( N(0, \sigma^2) \)
  - then can recover estimate of \( E[y|x] \) as:
  \[
  E[y|x] = \exp(\beta x_i + 0.5\sigma^2)
  \]
- Lognormal prediction
  - gen exp_yf=exp(yf+0.5*e(rmse)^2)

- If not normally distributed, but still homoskedastic, then above will be biased. Instead use Duan’s smearing estimator.

---

**Duan’s smearing estimator**

- Estimate using:
  \[
  E[y|x] = \hat{\phi} \times \exp(\hat{x}_i \hat{\beta})
  \]
- where \( \hat{\phi} \) is the estimated smearing factor:
  \[
  \hat{\phi} = N^{-1} \sum_i \exp(\hat{\epsilon}_i)
  \]
- And \( \hat{\epsilon}_i = \ln y_i - \hat{x}_i \hat{\beta} \)
- Typically smearing factor lies between 1.5 and 4.0
- Applicable to homoskedastic errors
If error is heteroskedastic

- Duan’s smearing estimator will lead to bias, with the bias being a function of $x$
- In log normal case: $E[y \mid x] = \exp(x'\hat{\beta} + 0.5\sigma^2(x))$
- In general case: $E[y \mid x] = \rho(x)\exp(x'\hat{\beta})$
- Accordingly, to eliminate bias in predictions requires knowledge of the form of heteroskedasticity
- Straightforward if have a limited number of binary regressors
- Difficult if number of regressors is large and contains continuous measures

Square-root transformations

- Similar approach to log models:
- Model: $\sqrt{y_i} = x_i'\beta + \varepsilon$
- Duan’s smearing estimator: $E[y \mid x] = \hat{\phi} + (x'\hat{\beta})^2$
- with smearing factor: $\hat{\phi} = N^{-1}\sum_i \hat{\varepsilon}_i^2$
- Heteroskedastic case: $E[y \mid x] = \rho(x) + (x'\hat{\beta})^2$

Box-Cox regression

- Rather than imposing a particular transformation a Box-Cox model can be used.
- Model: $y_{i(\theta)}^{(\theta)} = \frac{y_i^\theta - 1}{\theta} = x_i'\beta + \varepsilon$
- This includes levels ($\theta=1$) and logs ($\theta=0$) as special cases.
- Assuming $\varepsilon$ has a normal distribution, the model can be estimated by mle in Stata (more general models are also available that apply Box-Cox to covariates as well).

Exponential Conditional Mean (ECM) models

- ECM directly assumes a nonlinear relationship:
  \[ E[y \mid x] = \mu = \exp(x'\beta) \]
- Or, more generally
  \[ E[y \mid x] \propto \exp(x'\beta) \quad \text{i.e.,} \quad E[y \mid x] = \exp(x'\beta)\phi \]
- Note that this implies that the effect of covariates is “proportional” rather than additive (as in a proportional hazard model).
- The ECM, and extensions, can be estimated in a variety of ways:
  - Nonlinear least squares (NLS)
  - Poisson ML estimator (QML/PML)
  - Using hazard models (exponential, Weibull, generalised gamma)
  - Generalised linear models (GLM).
Generalised Gamma Model

- Generalised gamma is used as a flexible parametric distribution for survival models e.g. using \texttt{streg} in Stata. Proposed by Manning, Basu & Mullahy (2005, JHE) for use with cost data.
- Generalised gamma has density function (using notation from Stata manual):

$$ f(y; \kappa, \mu, \sigma) = \frac{y^{\gamma}}{\sigma \Gamma(y)} \exp\left(\frac{z - u}{\gamma}\right) $$

where $\gamma = |\kappa|^2$, $z = \text{sign}(\kappa)\{\ln(y) - \mu\}$, $u = \gamma \exp(\kappa|x|)$, $\mu = x'\beta$

$$ E(y | x) = \exp(x'\beta) \cdot \left[ \kappa \gamma / \kappa + 1 \right] \cdot \frac{\Gamma(1 + \kappa)}{\gamma \Gamma(1/\kappa)} = \exp(x'\beta) \phi $$

- Special cases are Weibull ($\kappa = 1$), exponential ($\kappa = 1, \sigma = 1$), and lognormal ($\kappa = 0$).

*using Anirban Basu's gengam2 code (equivalent to \texttt{streg})*

. gengam2 $xs, \text{robust nolog}$

Tests for identifying distributions

<table>
<thead>
<tr>
<th>Distributions</th>
<th>chi2</th>
<th>df</th>
<th>Prob&gt;chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Gamma (kappa = sigma)</td>
<td>359.85</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Log Normal (kappa = 0)</td>
<td>16.24</td>
<td>1</td>
<td>0.0001</td>
</tr>
<tr>
<td>Weibull (kappa = 1)</td>
<td>258.27</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>Exponential (kappa = sigma = 1)</td>
<td>412.52</td>
<td>2</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Andrew M Jones, June 2009
Generalized linear models (GLM)

• GLMs specify the conditional mean directly:
  \[ E[y \mid x] = \mu = f(x'\beta) \]
  – For example, with a “log link”:
  \[ E[y \mid x] = \exp(x'\beta) \]

• Advantages of GLM:
  – Predictions are made on the cost scale (no retransformation)
  – They allow for heteroskedasticity through the choice of distributional family (albeit limited to functions of the mean)

The GLM framework

• Requirements:
  – A link function \( g(.) \) that relates the conditional mean to the covariates:
  \[ g(\mu) = x'\beta \]
  \[ \Rightarrow \mu = g^{-1}(x'\beta) = f(x'\beta) \]
  – A distribution (D), that belongs to the linear exponential family, is used to specify the relationship between the variance and the mean:
  \[ y \sim D \]
  \[ \Rightarrow \text{Var}(y \mid x) = \nu(\mu) \]

Distributions

• Distributions:
  – Used to describe the relationship between the variance and conditional mean:
  \[ \text{var}[y \mid x] \approx (E[y \mid x])^\lambda \]
  – Distributional families:
    – Gaussian: constant variance; \( \lambda = 0 \)
    – Poisson: variance proportional to the mean; \( \lambda = 1 \)
    – Gamma: variance proportional to the square of the mean; \( \lambda = 2 \)
    – Inverse Gaussian: variance proportional to cube of the mean; \( \lambda = 3 \)
  – Can apply distributions and link functions independently (although there are canonical links)

Estimation of GLMs

Estimation is based on the classical “estimating equations” (score functions):

\[ \sum_i \left( \frac{y_i - \mu_i(\beta)}{\nu_i(\mu)} \right) \frac{\partial \mu_i}{\partial \beta} = \sum_i \left( \frac{r_i}{\sqrt{\nu_i(\mu)}} \right) \frac{\partial \mu_i}{\partial \beta} = 0 \]

Where \( r \) is the Pearson residual.

As GLMs are based on the linear exponential family they have the quasi-ML property (consistent so long as mean is correctly specified).
Park test

- Idea: GLM distribution should reflect the relationship between the variance and the mean:
  \[ \text{var}[y|\mu] = \alpha \left[ E(y|\mu) \right]^\lambda \]

- Park test exploits this by regressing \( \ln((y_i - \hat{y}_i)^2) \) on \( \ln(\hat{y}_i) \) and a constant.
- The estimated coefficient on \( \ln(\hat{y}_i) \) provides guidance on the appropriate distributional family:
  - Gaussian: \( \lambda = 0 \)
  - Poisson: \( \lambda = 1 \)
  - Gamma: \( \lambda = 2 \)
  - Inverse Gaussian: \( \lambda = 3 \)

Extended Estimating Equations (EEE)

- Proposed by Basu & Rathouz (2005, *Biostatistics*)
- Combines a Box-Cox transformation for the link function:
  \[ \frac{\mu^\lambda - 1}{\lambda} \]
- With a general power function for the variance:
  \[ \theta_1 \mu^{\theta_2} \]

* * Extended Estimating Equations (EEE) *

```
* EXTENDED ESTIMATING EQUATIONS (EEE) *
sum y, meanonly
gen scy = y/r(mean)
global sc = r(mean)
* EXTENDED ESTIMATING EQUATIONS (EEE) *
pglm scy $xs

Iter: 1 Max % Diff: 6.7972928 Rel Diff: 93.202707
Half step applied. Reset Max % Diff
Iter: 9 Max % Diff: .00028327 Rel Diff: .00062624
Iter: 10 Max % Diff: .00008658 Rel Diff: .00019668

Extended GEE with Power Variance Function
No of obs = 2955
Optimization: Fisher’s Scoring
Residual df = 2944

Variance: (theta1*mu^theta2)
Link: (mu^lambda - 1)/lambda
Std Errors: Robust
```
predict fitted values, linear index & residuals
.pglmpredict yfc if e(sample), mu scale($sc)
predict xb if e(sample), xb
gen eee_yf=yfc if e(sample)
gen ey = y - yfc

* Counterfactual predictions - attributable fraction and costs

summ y eee_yf y0 attf attc if d_raw==1

Variable |       Obs        Mean    Std. Dev.       Min        Max -------------+--------------------------------------------------------
y |      1748    7611.963    12358.83          3     125610
eee_yf |      1748    7597.822    4237.654   2095.638   27122.07           y0 |      1748    6847.585    4052.656   1663.242   25746.22
attf |      1748    .1163505    .0350095   .0507279   .2063313
attc |      1748    750.2375    187.6568   432.3958   1375.846
So \( \hat{\theta} = 0.098 \) (recall OLS gave 0.095!)

FINITE MIXTURE MODELS

Deb and Trivedi (1997) proposed the use of finite mixture models (FMM) as an alternative to the hurdle model in the empirical modelling of health care utilisation.

In a FMM the population is assumed to be divided in \( C \) distinct components in proportions \( \pi_1, \ldots, \pi_C \), where \( \sum_{j=1}^{C} \pi_j = 1 \) and \( 0 \leq \pi_j \leq 1 \), \( j=1, \ldots, C \). The C-point finite mixture model is given by:

\[
f(y_i | \cdot) = \sum_{j=1}^{C} \pi_j f_j(y_i | \cdot),
\]

where the mixing probabilities \( \pi_j \) are estimated along with all the other parameters of the model. Also, \( \pi_C = 1 - \sum_{j=1}^{C-1} \pi_j \).

**FINITE MIXTURE MODELS (FMM)**

* Finite mixture model using fmm command (Partha Deb)

.fmm y $xs [pweight=wt], vce(robust) components(2) mixtureof(gamma)

Fitting Gamma regression model:
Iteration 9:  log pseudolikelihood = -28503.5

2 component Gamma regression  Number of obs = 2955
Log pseudolikelihood = -28503.5  Prob > chi2 = 0.0000

|                | Coef.   Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------------|------------------|-------|-------|-----------------------|
| Robust component1 |                    |       |       |                       |
| d               | 0.2982341        | 0.0538204 | 5.54 | 0.000 | 0.1927486 | 0.4037197 |
| phylim          | 0.2875844        | 0.0664254 | 4.33 | 0.000 | 0.1573934 | 0.4177755 |
| actlim          | 0.3067102        | 0.0773966 | 3.96 | 0.000 | 0.1550157 | 0.4584047 |
| totchr          | 0.3590686        | 0.0227799 | 15.76 | 0.000 | 0.3144207 | 0.4037165 |

|                | Coef.   Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------------|------------------|-------|-------|-----------------------|
| Robust component2 |                    |       |       |                       |
| d               | 0.0355457        | 0.011813 | 0.39 | 0.697 | -0.1431664 | 0.2142578 |
| phylim          | 0.3974127        | 0.1021512 | 3.89 | 0.000 | 0.1972 | 0.5976253 |
| actlim          | 0.4100909        | 0.1035322 | 4.05 | 0.000 | 0.2161764 | 0.6201533 |
| totchr          | 0.2087046        | 0.0876243 | 2.38 | 0.000 | 0.0310302 | 0.3814666 |
| age             | -0.0163599       | 0.0076243 | -2.15 | 0.032 | -0.0313032 | -0.0014166 |
| female          | -0.2716229       | 0.0892164 | -3.04 | 0.002 | -0.446437 | -0.096762 |
| income          | -0.0005708       | 0.0024328 | -0.23 | 0.815 | -0.002389 | 0.0053987 |
| _cons           | 5.951575         | 0.6161151 | 16.60 | 0.000 | 5.307571 | 6.659579 |

alpha1 | 1.415317 | 0.0598543 | 25.36 | 0.000 | 1.302733 | 1.537629 |
alpha2 | 1.063922 | 0.0892222 | 12.02 | 0.000 | 0.984306 | 1.142125 |
/lnalpha1 | 0.437538 | 0.0422904 | 8.21 | 0.000 | 0.354641 | 0.523144 |
/lnalpha2 | 0.0639604 | 0.0554762 | 1.19 | 0.234 | -0.046799 | 0.176618 |

/lnogitpi1 | 1.001253 | 0.004055 | 2.50 | 0.012 | 0.001964 | 0.018054 |
/lnogitpi2 | 0.0035457 | 0.011813 | 0.39 | 0.697 | -0.1431664 | 0.2142578 |
alpha1 | 1.415317 | 0.0598543 | 25.36 | 0.000 | 1.302733 | 1.537629 |
alpha2 | 1.063922 | 0.0892222 | 12.02 | 0.000 | 0.984306 | 1.142125 |
/lnalpha1 | 0.437538 | 0.0422904 | 8.21 | 0.000 | 0.354641 | 0.523144 |
/lnalpha2 | 0.0639604 | 0.0554762 | 1.19 | 0.234 | -0.046799 | 0.176618 |

So \( \hat{\theta} = 0.098 \) (recall OLS gave 0.095!)
Predictions for the two components (note cost scale!)

---

**Discrete Conditional Density Estimator**

- Divides support of y into fixed number (K) of discrete intervals eg. deciles.
- Interested in:
  \[
  E[h(y) \mid x] = \int h(y)f(y \mid x)dy = \sum_k h^*(k)p[y_{k-1} \leq Y < y_k \mid x]
  \]
  Where \( h^*(k) \) is an approximation within \( k \)th interval – they use simple mean
- They use a discrete hazard specification for probabilities. Here use a multinomial logit.

---

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1748</td>
<td>7611.963</td>
<td>12358.83</td>
<td>3</td>
<td>125610</td>
</tr>
<tr>
<td>yf</td>
<td>1748</td>
<td>6508.435</td>
<td>4160.099</td>
<td>2226.253</td>
<td>40631.48</td>
</tr>
<tr>
<td>y0</td>
<td>1748</td>
<td>5724.134</td>
<td>3608.32</td>
<td>1950.773</td>
<td>35190.02</td>
</tr>
<tr>
<td>attf</td>
<td>1748</td>
<td>.117273</td>
<td>.0138686</td>
<td>.0822406</td>
<td>.1720743</td>
</tr>
<tr>
<td>attc</td>
<td>1748</td>
<td>784.3005</td>
<td>563.7233</td>
<td>250.2881</td>
<td>5441.469</td>
</tr>
</tbody>
</table>

So \( \theta = 0.120 \)
<table>
<thead>
<tr>
<th>Link</th>
<th>t p value</th>
<th>OTest F p value</th>
<th>Pearson t p value</th>
<th>COPAS (βχ p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS on y</td>
<td>0.133</td>
<td>0.009</td>
<td></td>
<td>0.98 (0.493)</td>
</tr>
<tr>
<td>OLS on ln(y) + Duan</td>
<td>0.000</td>
<td>0.000</td>
<td>0.535 (0.000)</td>
<td></td>
</tr>
<tr>
<td>OLS on √x + Duan</td>
<td>0.712</td>
<td>0.325</td>
<td>1.207 (0.000)</td>
<td></td>
</tr>
<tr>
<td>ECM - nls</td>
<td></td>
<td>0.350</td>
<td>0.983 (0.744)</td>
<td></td>
</tr>
<tr>
<td>ECM – Poisson-nl</td>
<td></td>
<td>0.158</td>
<td>0.888 (0.481)</td>
<td></td>
</tr>
<tr>
<td>Generalised gamma</td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>GLM log-gamma</td>
<td>0.123</td>
<td></td>
<td></td>
<td>0.741 (0.000)</td>
</tr>
<tr>
<td>GLM log-normal</td>
<td>0.052</td>
<td></td>
<td></td>
<td>1.004 (0.945)</td>
</tr>
<tr>
<td>GLM log-poisson</td>
<td>0.071</td>
<td></td>
<td></td>
<td>0.883 (0.007)</td>
</tr>
<tr>
<td>EEE</td>
<td></td>
<td>0.690</td>
<td></td>
<td>0.954</td>
</tr>
<tr>
<td>FMM gamma</td>
<td></td>
<td>0.935</td>
<td></td>
<td>0.926 (0.482)</td>
</tr>
<tr>
<td>CDE</td>
<td></td>
<td></td>
<td></td>
<td>0.639</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link</th>
<th>R² (y-scale)</th>
<th>RMSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS on y</td>
<td>0.116</td>
<td>11270</td>
<td>6225</td>
</tr>
<tr>
<td>OLS on ln(y) + Duan</td>
<td>0.090</td>
<td>11797</td>
<td>6569</td>
</tr>
<tr>
<td>OLS on √x + Duan</td>
<td>0.114</td>
<td>11314</td>
<td>6237</td>
</tr>
<tr>
<td>ECM - nls</td>
<td>0.113</td>
<td>11296</td>
<td>6267</td>
</tr>
<tr>
<td>ECM – Poisson-nl</td>
<td>0.110</td>
<td>11312</td>
<td>6196</td>
</tr>
<tr>
<td>Generalised gamma</td>
<td>0.093</td>
<td>11769</td>
<td>5408</td>
</tr>
<tr>
<td>GLM log-gamma</td>
<td>0.106</td>
<td>11390</td>
<td>6254</td>
</tr>
<tr>
<td>GLM log-normal</td>
<td>0.113</td>
<td>11295</td>
<td>6267</td>
</tr>
<tr>
<td>GLM log-poisson</td>
<td>0.110</td>
<td>11312</td>
<td>6196</td>
</tr>
<tr>
<td>EEE</td>
<td>0.116</td>
<td>11274</td>
<td>6179</td>
</tr>
<tr>
<td>FMM gamma</td>
<td>0.106</td>
<td>11395</td>
<td>5575</td>
</tr>
<tr>
<td>CDE</td>
<td>0.117</td>
<td>11264</td>
<td>6180</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link</th>
<th>Attributable fraction</th>
<th>Pruned sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS on y</td>
<td>0.095</td>
<td>0.078</td>
</tr>
<tr>
<td>OLS on ln(y) + Duan</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td>OLS on √x + Duan</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>ECM - nls</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>ECM – Poisson-nl</td>
<td>0.091</td>
<td></td>
</tr>
<tr>
<td>Generalised gamma</td>
<td>0.206</td>
<td></td>
</tr>
<tr>
<td>GLM log-gamma</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td>GLM log-normal</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>GLM log-poisson</td>
<td>0.091</td>
<td></td>
</tr>
<tr>
<td>EEE</td>
<td>0.099</td>
<td>0.082</td>
</tr>
<tr>
<td>FMM gamma</td>
<td>0.121</td>
<td>0.087</td>
</tr>
<tr>
<td>CDE</td>
<td>0.052</td>
<td>0.053</td>
</tr>
</tbody>
</table>