University of Zurich[UZH]

IFSPM Institut für Sozial- und Präventivmedizin

Conditional Transformation Models

Or: More Than Means Can Say

Torsten Hothorn, Universität Zürich
Thomas Kneib, Universität Göttingen
Peter Bühlmann, Eidgenössische Technische Hochschule Zürich

2013-09-18

“We are mean lovers.”

The famous top three reasons to become a statistician:

1. Deviation is considered normal.
2. We feel complete and sufficient.
3. We are ’mean’ lovers.

with the last point referring of course to our obsession with means.

Conceptually, statisticians are obsessed with distributions, but when there are many distributions to look at simultaneously, we tend to cut some corners, i.e., higher moments.

Conditional Transformation Models

- We observe response $Y \in \mathbb{R}$ and explanatory variables $\boldsymbol{X} = \boldsymbol{x} \in \chi$.
- We are interested in the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$.
- Instead, many regression models focus on the conditional mean $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.
- Conditional transformation models estimate the conditional distribution function $\mathbb{P}(Y \leq \upsilon|\boldsymbol{X} = \boldsymbol{x})$ directly.

A Scene of Contemporary Regression

- Let $Y_{\boldsymbol{x}} = (Y|\boldsymbol{X} = \boldsymbol{x}) \sim \mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ denote the conditional distribution of response Y given explanatory variables $\boldsymbol{X} = \boldsymbol{x}$; $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ (being dominated by some measure μ) with conditional distribution function $\mathbb{P}(Y \leq \upsilon|\boldsymbol{X} = \boldsymbol{x})$.
- A regression model describes the distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$, or certain characteristics of it, as a function of the explanatory variables $\boldsymbol{x}$.
- We estimate such models based on random variables $(Y, \boldsymbol{X}) \sim \mathbb{P}_{Y,\boldsymbol{X}}$.
- A regression model consists of signal and noise, *i.e.* , some error term $Q(U)$ with $U \sim \mathcal{U}[0, 1]$ and $Q : \mathbb{R} \to \mathbb{R}$ being a quantile function.

A Scene of Contemporary Regression

– There are two common ways to look at the problem

$$\begin{aligned} Y_{\boldsymbol{x}} &= r(Q(U)|\boldsymbol{x}) \quad \text{"mean or quantile regression models" and} \\ h(Y_{\boldsymbol{x}}|\boldsymbol{x}) &= Q(U) \quad \text{"transformation models".} \end{aligned}$$

– For each $\boldsymbol{x} \in \chi$, the regression function $r(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$ transforms the error term $Q(U)$ in a monotone increasing way.

– The inverse regression function $h(\cdot|\boldsymbol{x}) = r^{-1}(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$ is also monotone increasing. Because h transforms the response, it is known as a transformation function, and models in the second form are called transformation models.

A Scene of Contemporary Regression

- A major assumption underlying almost all mean or quantile regression models is the additivity of signal and noise:

$$r(Q(U)|\boldsymbol{x}) = r_{\boldsymbol{x}}(\boldsymbol{x}) + Q(U).$$

- When $\mathbb{E}(Q(U)) = 0$, we get $r_{\boldsymbol{x}}(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, *e.g.* linear or additive models depending on the functional form of $r_{\boldsymbol{x}}$.
- Model inference is commonly based on the normal error assumption, *i.e.* $Q(U) = \sigma\Phi^{-1}(U)$, where $\sigma > 0$ is a scale parameter and $\Phi^{-1}(U) \sim \mathcal{N}(0, 1)$.
- We often call σ a "nuisance parameter", but in fact this is an euphemism for "we simply ignore higher moments".

A Scene of Contemporary Regression

For transformation models, additivity is assumed on the scale of the inverse regression function h.

- $h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = h_Y(Y_{\boldsymbol{x}}) + h_{\boldsymbol{x}}(\boldsymbol{x}) = Q(U)$
- When $\mathbb{E}(Q(U)) = 0$, we get

$$-h_{\boldsymbol{x}}(\boldsymbol{x}) = \mathbb{E}(h_Y(Y_{\boldsymbol{x}})) = \mathbb{E}(h_Y(Y)|\boldsymbol{X} = \boldsymbol{x}).$$

- The monotone transformation function $h_Y : \mathbb{R} \to \mathbb{R}$ does not depend on $\boldsymbol{x}$.
- h_Y might be known in advance (Box-Cox transformation models with fixed parameters, accelerated failure time models).
- h_Y is commonly treated as a nuisance parameter (Cox model, proportional odds model).
- One is usually interested in estimating the function $h_{\boldsymbol{x}} : \chi \to \mathbb{R}$, *i.e.* , the negative conditional mean of the transformed response $h_Y(Y)$.

Towards Conditional Transformation Models

Some thoughts about transformation models:

- The transformation function h_Y is typically treated as an infinite dimensional nuisance parameter.
- But h_Y contains information about higher moments of $Y_{\boldsymbol{x}}$!
- An attractive feature of transformation models is their close connection to the conditional distribution function:

$$\mathbb{P}(Y \le \upsilon | \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(h(Y|\boldsymbol{x}) \le h(\upsilon|\boldsymbol{x})) = F(h(\upsilon|\boldsymbol{x})); \quad F = Q^{-1}.$$

Towards Conditional Transformation Models

- For additive transformation functions $h = h_Y + h_{\boldsymbol{x}}$, we have $F(h(\upsilon|\boldsymbol{x})) = F(h_Y(\upsilon) + h_{\boldsymbol{x}}(\boldsymbol{x}))$.
- Therefore, higher moments only depend on the transformation h_Y and thus cannot be influenced by the explanatory variables.
- Consequently, one has to avoid the additivity in the model $h = h_Y + h_{\boldsymbol{x}}$ to allow the explanatory variables to impact also higher moments.

Conditional Transformation Models

- To avoid the additivity $h = h_Y + h_{\boldsymbol{x}}$ in transformation model, we suggest a novel transformation model based on an alternative additive decomposition of the transformation function h into J partial transformation functions for all $\boldsymbol{x} \in \chi$:

$$h(\upsilon|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(\upsilon|\boldsymbol{x}).$$

- The transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x})$ and the partial transformation functions $h_j(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$ are conditional on $\boldsymbol{x}$ in the sense that not only the mean of $Y_{\boldsymbol{x}}$ depends on the explanatory variables.
- Therefore, we coin these models *Conditional Transformation Models* (CTMs).

Estimation

- Well-known “trick”: Use the mean regression hammer to nail the problem:

$$\mathbb{P}(Y \leq \upsilon | \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(I(Y \leq \upsilon) | \boldsymbol{X} = \boldsymbol{x}).$$

- Fit model $\mathbb{E}(I(Y \leq \upsilon) | \boldsymbol{X} = \boldsymbol{x})$ for a grid of υ values separately.
- This is similar to fitting multiple quantile regression models.
- Better: find an appropriate risk function that allows the whole conditional distribution function to be obtained in one step.

Estimation: Risk Function

- Let ρ denote a function of measuring the loss of the probability $F(h(\upsilon|\boldsymbol{X}))$ for the binary event $Y \leq \upsilon$, for example

$$\begin{aligned}
\rho_{\text{bin}}((Y \leq \upsilon, \boldsymbol{X}), h(\upsilon|\boldsymbol{X})) &:= -[I(Y \leq \upsilon) \log\{F(h(\upsilon|\boldsymbol{X}))\} + \\
&\quad \{1 - I(Y \leq \upsilon)\} \log\{1 - F(h(\upsilon|\boldsymbol{X}))\}] \\
\rho_{\text{sqe}}((Y \leq \upsilon, \boldsymbol{X}), h(\upsilon|\boldsymbol{X})) &:= \frac{1}{2}|I(Y \leq \upsilon) - F(h(\upsilon|\boldsymbol{X}))|^2 \\
\rho_{\text{abe}}((Y \leq \upsilon, \boldsymbol{X}), h(\upsilon|\boldsymbol{X})) &:= |I(Y \leq \upsilon) - F(h(\upsilon|\boldsymbol{X}))|.
\end{aligned}$$

- ρ_{sqe} is also known as the Brier score.
- Now define the loss function ℓ for CTM estimation as integrated loss ρ with respect to the measure μ dominating the conditional distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$:

$$\ell((Y, \boldsymbol{X}), h) := \int \rho((Y \leq \upsilon, \boldsymbol{X}), h(\upsilon|\boldsymbol{X}))\, d\mu(\upsilon).$$

Estimation: Risk Function = Scoring Rule

- In the context of scoring rules, the loss ℓ based on ρ_{sqe} is known as the continuous ranked probability score (CPRS) or integrated Brier score and is a proper scoring rule for assessing the quality of probabilistic or distributional forecasts.
- Define the corresponding risk function as

$$\mathbb{E}_{Y,\boldsymbol{X}}\ell((Y,\boldsymbol{X}),h) = \int\int \rho((y \leq \upsilon, \boldsymbol{x}), h(\upsilon|\boldsymbol{x}))\, d\mu(\upsilon)\, d\mathbb{P}_{Y,\boldsymbol{X}}(y,\boldsymbol{x}).$$

- $\mathbb{E}_{Y,\boldsymbol{X}}\ell((Y,\boldsymbol{X}),h)$ is convex in h and attains its minimum for the true conditional transformation function h with $\rho = \rho_{\text{bin}}$ and $\rho = \rho_{\text{sqe}}$ (but not with $\rho = \rho_{\text{abe}}$).

Estimation: Empirical Risk Function

– The corresponding empirical risk function defined by the data is

$$\hat{\mathbb{E}}_{Y,\boldsymbol{X}}\ell((Y,\boldsymbol{X}),f) = \int\int \rho((y \le \upsilon, \boldsymbol{x}), h(\upsilon|\boldsymbol{x}))\, d\mu(\upsilon)\, d\hat{\mathbb{P}}_{Y,\boldsymbol{X}}(y,\boldsymbol{x}).$$

– Use i.i.d. random sample $(Y_i, \boldsymbol{X}_i) \sim \mathbb{P}_{Y,\boldsymbol{X}}, i = 1,\dots,N$ to define $\hat{\mathbb{P}}_{Y,\boldsymbol{X}}$.

– For computational convenience, approximate the measure μ by the discrete uniform measure $\hat{\mu}$, which puts mass n^{-1} on each element of the equi-distant grid $\upsilon_1 < \dots < \upsilon_n \in \mathbb{R}$ over the response space.

– The weighted empirical risk is then

$$\begin{aligned}\hat{\mathbb{E}}_{Y,\boldsymbol{X}}\ell((Y,\boldsymbol{X}),h) &= \sum_{i=1}^{N} w_i n^{-1} \sum_{\imath=1}^{n} \rho((Y_i \le \upsilon_\imath, \boldsymbol{X}_i), h(\upsilon_\imath|\boldsymbol{X}_i)) \\ &= n^{-1}\sum_{i=1}^{N}\sum_{\imath=1}^{n} w_i \rho((Y_i \le \upsilon_\imath, \boldsymbol{X}_i), h(\upsilon_\imath|\boldsymbol{X}_i)).\end{aligned}$$

This risk is the weighted empirical risk for loss function ρ evaluated at the observations $(Y_i \le \upsilon_\imath, \boldsymbol{X}_i)$ for $i = 1,\dots,N$ and $\imath = 1,\dots,n$.

Estimation: Empirical Risk Minimisation

- We can now use empirical risk minimisation for fitting $\hat{h}$.
- Of course we need to smooth a bit here:
  - $\hat{h}_j(\upsilon|\boldsymbol{x})$ should be smooth in υ-direction (no steps in the conditional distribution function).
  - $\hat{h}_j(\upsilon|\boldsymbol{x})$ should also be smooth in $\boldsymbol{x}$-direction (conditional distribution varies smoothly in the explanatory variables).
- In principle, any algorithm for minimising risk functions defined by ρ can be used.
- Componentwise boosting comes in very handy here: Smoothing and variable / component selection are (almost) free (as in “free beer”).

Boosting: Base Learners

- Parameterise the partial transformation functions for all $j = 1, \dots, J$ as

$$h_j(\upsilon|\boldsymbol{x}) = \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_0(\upsilon)^\top\right)\boldsymbol{\gamma}_j \in \mathbb{R}, \qquad \boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_0},$$

  where $\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_0(\upsilon)^\top$ denotes the tensor product of two sets of basis functions $\boldsymbol{b}_j : \chi \to \mathbb{R}^{K_j}$ and $\boldsymbol{b}_0 : \mathbb{R} \to \mathbb{R}^{K_0}$.
- $\boldsymbol{b}_0$ is a basis along the υ values.
- The basis $\boldsymbol{b}_j$ defines how this transformation may vary with certain aspects of the explanatory variables.
- h_j needs to be smooth in both arguments; therefore the bases are supplemented with appropriate, pre-specified penalty matrices $\boldsymbol{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\boldsymbol{P}_0 \in \mathbb{R}^{K_0 \times K_0}$, inducing the penalty matrix $\boldsymbol{P}_{0j} = (\lambda_0 \boldsymbol{P}_j \otimes \boldsymbol{1}_{K_0} + \lambda_j \boldsymbol{1}_{K_j} \otimes \boldsymbol{P}_0)$ with smoothing parameters $\lambda_0 \geq 0$ and $\lambda_j \geq 0$ for the tensor product basis.
- The base-learners are now Ridge-type linear models with penalty matrix $\boldsymbol{P}_{0j}$.

Boosting: Algorithm

1. Initialise $\gamma_j^{[0]} \equiv 0$ for $j = 1, \dots, J$, the step-size $\nu \in (0, 1)$ and the smoothing parameters $\lambda_j, j = 0, \dots, J$. Define the grid $\upsilon_1 < Y_{(1)} < \dots < Y_{(N)} \leq \upsilon_n$. Set $m := 0$.
2. Compute the negative gradient:

$$U_{i\imath} := - \frac{\partial}{\partial h} \rho((Y_i \leq \upsilon_\imath, \boldsymbol{X}_i), h)\Big|_{h=\hat{h}_{i\imath}^{[m]}}$$

with $\hat{h}_{i\imath}^{[m]} = \sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{X}_i)^\top \otimes \boldsymbol{b}_0(\upsilon_\imath)^\top\right) \gamma_j^{[m]}$. Fit the base-learners for $j = 1, \dots, J$:

$$\hat{\boldsymbol{\beta}}_j = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^{K_j K_0}} \sum_{i=1}^{N} \sum_{\imath=1}^{n} w_i \left\{ U_{i\imath} - \left(\boldsymbol{b}_j(\boldsymbol{X}_i)^\top \otimes \boldsymbol{b}_0(\upsilon_\imath)^\top\right) \boldsymbol{\beta} \right\}^2 + \boldsymbol{\beta}^\top \boldsymbol{P}_{0j} \boldsymbol{\beta}$$

with penalty matrix $\boldsymbol{P}_{0j}$. Select the best base-learner $j^\star$.

3. Update the parameters $\gamma_{j^\star}^{[m+1]} = \gamma_{j^\star}^{[m]} + \nu \hat{\boldsymbol{\beta}}_{j^\star}$ and keep all other parameters fixed.
4. Iterate 2. and 3.
5. Stop if $m = M$.

Boosting

Skip the algorithmic details today! It just works.

Childhood Nutrition in India

- Childhood undernutrition is one of the most urgent problems in developing and transition countries.
- Childhood nutrition is usually measured in terms of a Z score that compares the nutritional status of children in the population of interest with the nutritional status in a reference population.
- We will focus on stunting, *i.e.* insufficient height for age, as a measure of chronic undernutrition and estimate the whole distribution of this Z score measure for childhood nutrition in India.
- The analysis is based on India's 1998–1999 Demographic and Health Survey on 24166 children.

Childhood Nutrition in India

- The simplest conditional transformation model allowing for district-specific means and variances reads

$$\mathbb{P}(Z \leq \upsilon | \text{district} = k) = \Phi(\alpha_{0,k} + \alpha_k \upsilon), \quad k = 1, \dots, 412.$$

- The base-learner is defined by a linear basis $\boldsymbol{b}_0(\upsilon) = (1, \upsilon)^\top$ for the grid variable and a dummy-encoding basis $\boldsymbol{b}_1(\text{district}) = (I(\text{district} = 1), \dots, I(\text{district} = k))^\top$ for the 412 districts.
- The resulting 824-dimensional parameter vector $\boldsymbol{\gamma}_1$ of the tensor product base-learner then consists of separate intercept and slope parameters for each of the districts of India.
- Note that since we assume normality for the linear function $\alpha_{0,k} + \alpha_k Z \sim \mathcal{N}(0, 1)$, also the Z score is assumed to be normal with both mean and variance depending on the district.

Childhood Nutrition in India

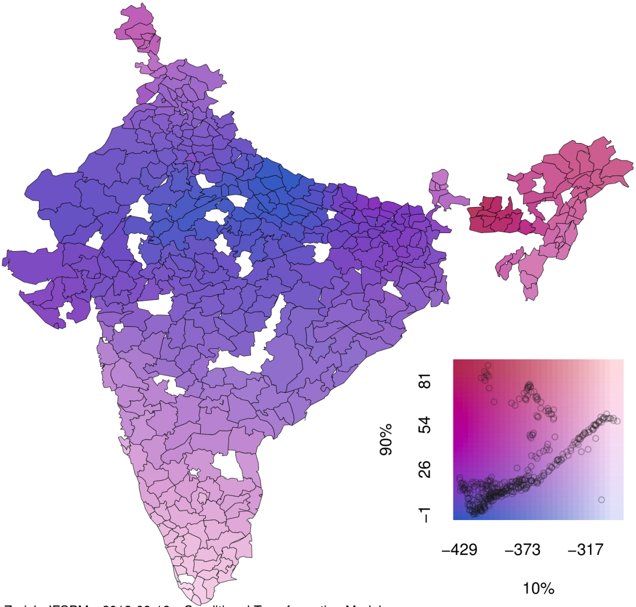
- We relax the normal assumption on Z by allowing for more flexible transformations

$$\mathbb{P}(Z \leq \upsilon | \text{district} = k) = \Phi(h(\upsilon | \text{district} = k)), \quad k = 1, \dots, 412.$$

- Now $\boldsymbol{b}_0(\upsilon)$ is a vector of B-spline basis functions evaluated at υ and $\boldsymbol{b}_1$ remains as above.
- To achieve smoothness of these non-parametric effects along the υ-grid, we specify the penalty matrix $\boldsymbol{P}_0$ as $\boldsymbol{P}_0 = \boldsymbol{D}^\top \boldsymbol{D}$ with second-order difference matrix $\boldsymbol{D}$.
- It makes sense to induce spatial smoothness on the conditional distribution functions of neighbouring districts. To implement spatial smoothness the penalty matrix $\boldsymbol{P}_1$ is chosen as an adjacency matrix of the districts.
- From the estimated conditional distribution functions, we compute quantiles of the Z score for each district via

$$\hat{Q}(\tau | \text{district} = k) = \inf\{\upsilon : \Phi(\hat{h}(\upsilon | \text{district} = k) \geq \tau\}.$$

Childhood Nutrition in India



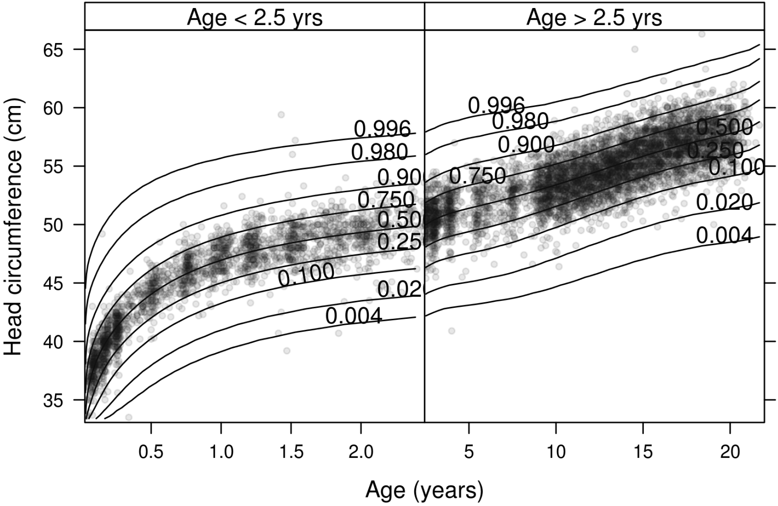
Head Circumference Growth

- The Fourth Dutch Growth Study is a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 22 years.
- We look at head circumference (HC) and age of 7040 males and estimate the whole conditional distribution function via

$$\mathbb{P}(\text{HC} \le \upsilon | \text{age} = x) = \Phi(h(\upsilon | \text{age} = x)).$$

- The base-learner is the tensor product of B-spline basis functions $\boldsymbol{b}_0(\upsilon)$ for head circumference and B-spline basis functions for age$^{1/3}$.
- The penalty matrices $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ penalise second-order differences, and thus $\hat{h}$ will be a smooth bivariate tensor product spline of head circumference and age.
- It is important to note that smoothing takes place in both dimensions.

Head Circumference Growth



Odds and Ends

- The corresponding paper will appear in JRSS B 75(5) http://dx.doi.org/10.1111/rssb.12017.
- The paper establishes the convergence of $\hat{h}$ to the true h.
- It furthermore contains simulation experiments comparing the performance of CTMs with GAMLSS, kernel conditional distribution estimation (package **np**), and additive quantile regression.
- More examples are contained in the extended paper version available from http://arxiv.org/abs/1201.5786 and in the IWSM proceedings.
- Conditional Transformation Models are implemented in packages `ctm` and `ctmDevel`, both available from http://R-forge.R-project.org.
- The source code for producing the results shown here is contained in these packages.

Thank you...

...for your attention!