Deep Composite Regression

Mario V. Wüthrich

RiskLab, ETH Zurich

joint work with Tobias Fissler and Michael Merz

June 21-24, 2022

Actuarial Colloquia 2022

- **Motivation and introduction**

Regression modeling: an example

```
'data.frame'    :   339500 obs. of  9 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ WorkLeisure  : Factor w/ 2 levels "1","2": 1 1 2 2 2 1 2 2 2 1 ...
 $ LaborSector  : Factor w/ 24 levels "5","12","13",..: 5 10 13 7 12 13 4 21 1 4 ...
 $ AccQuarter   : int  3 2 1 3 4 4 1 2 1 3 ...
 $ RepDelay     : num  0 0 0 0 1 0 0 0 0 0 ...
 $ Age          : num  45 20 20 20 60 55 30 25 20 20 ...
 $ InjuryType   : Factor w/ 19 levels "1","2","3","4",..: 7 6 4 13 16 2 6 4 4 18 ...
 $ InjBodyPart  : Factor w/ 35 levels "1","2","3","4",..: 20 28 28 20 14 23 2 14 6 3 ...
 $ Claim        : num  562 6675 700 57 2382 ...
```

Goal.

- Find a regression function that describes the systematic effects in the claims Y as a function of the available covariates $\boldsymbol{x}$

$$\boldsymbol{x} \;\mapsto\; \mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{x}].$$

Difficulty in practice

- There is no (simple) model that fits the entire range of possible claims Y.
- To overcome this issue one uses (drop covariate $\boldsymbol{x}$ in the notation)
  - ⋆ either mixture models

$$Y \sim \sum_{k=1}^{K} p_k \, f_k(y),$$

  - ⋆ or composite models with (fixed) splicing points $m_{k-1} < m_k$

$$Y \sim \sum_{k=1}^{K} p_k \, \frac{f_k(y) \, \mathbb{1}_{\{m_{k-1} < y \leq m_k\}}}{F_k(m_k) - F_k(m_{k-1})}.$$

- Fitting uses Expectation-Maximization (EM) algorithm.

Our proposal

- We choose a composite model, but replace the fixed splicing point $m_k \in \mathbb{R}$ by a quantile splicing point $F^{-1}_{Y|\boldsymbol{x}}(\tau)$.

- Choose a quantile level $\tau \in (0,1)$ and consider regression function

$$\begin{aligned} \boldsymbol{x} \mapsto \mu(\boldsymbol{x}) &= \tau\, \mathbb{E}_{\boldsymbol{x}}\left[Y \,\middle|\, Y \leq F^{-1}_{Y|\boldsymbol{x}}(\tau)\right] + (1-\tau)\, \mathbb{E}_{\boldsymbol{x}}\left[Y \,\middle|\, Y > F^{-1}_{Y|\boldsymbol{x}}(\tau)\right] \\ &= \tau\, \mathrm{CTE}^{-}_{\tau}(Y|\boldsymbol{x}) + (1-\tau)\, \mathrm{CTE}^{+}_{\tau}(Y|\boldsymbol{x}). \end{aligned}$$

  CTE = conditional tail expectation

- This allows us to fit different models to tail and body of the data:
  - ⋆ Different distributional assumptions below and above $F^{-1}_{Y|\boldsymbol{x}}(\tau)$.
  - ⋆ Different regression functions below and above $F^{-1}_{Y|\boldsymbol{x}}(\tau)$.

- **Elicitability**

Functionals and action space

- Denote by $\mathcal{F}$ a fixed family of distribution functions F.
- Consider a functional $T : F \in \mathcal{F} \mapsto T(F) \in \mathbb{A}$ (action space).
- Examples:
  - ⋆ Mean functional

$$F \mapsto T(F) = \mathbb{E}_F[Y].$$

  - ⋆ Quantile for quantile level $\tau \in (0, 1)$

$$F \mapsto T(F) = F^{-1}(\tau).$$

  - ⋆ Upper conditional tail expectation

$$F \mapsto T(F) = \mathrm{CTE}_\tau^+(Y) = \mathbb{E}_F\left[Y \,\middle|\, Y > F^{-1}(\tau)\right].$$

Elicitability and consistency

The functional $T : \mathcal{F} \to \mathbb{A}$ is elicitable on $\mathcal{F}$ if there exists a loss function

$$L : \mathbb{R} \times \mathbb{A} \to \mathbb{R},$$

such that for all $F \in \mathcal{F}$

$$T(F) \;=\; \operatorname*{arg\,min}_{a \in \mathbb{A}} \mathbb{E}_F \left[L(Y, a) \right]. \tag{1}$$

- A loss function L satisfying (1) for all $F \in \mathcal{F}$ is called consistent for T.
- Elicitability means that we can learn the functional T by M-estimation (1), i.e., choosing a consistent loss function L for T.
- Having data $Y_i \overset{\text{i.i.d.}}{\sim} F$, we can estimate

$$\widehat{T}(F) \;=\; \operatorname*{arg\,min}_{a \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, a).$$

Examples elicitable

- Mean functional: is elicitable (Bregman divergences).

$$F \;\mapsto\; T(F) = \mathbb{E}_F[Y].$$

- Quantile for quantile level: $\tau \in (0,1)$ is elicitable (pinball losses).

$$F \;\mapsto\; T(F) = F^{-1}(\tau).$$

- Upper conditional tail expectation: is not elicitable.

$$F \;\mapsto\; T(F) = \mathrm{CTE}_\tau^+(Y) = \mathbb{E}_F\left[Y \,\middle|\, Y > F^{-1}(\tau)\right].$$

Composite triplet

Theorem (Fissler–Merz–W 2021). The composite triplet

$$\begin{aligned}&\big(\mathrm{CTE}_\tau^-(Y),\, F^{-1}(\tau),\, \mathrm{CTE}_\tau^+(Y)\big)\\ &\qquad = \ \big(\mathbb{E}_F\left[Y\,\middle|\,Y \le F^{-1}(\tau)\right],\, F^{-1}(\tau),\, \mathbb{E}_F\left[Y\,\middle|\,Y > F^{-1}(\tau)\right]\big).\end{aligned}$$

is jointly elicitable.

- We give a full characterization of all consistent loss functions L such that

$$\big(\mathrm{CTE}_\tau^-(Y),\, F^{-1}(\tau),\, \mathrm{CTE}_\tau^+(Y)\big) \ = \ \underset{(c^-,v,c^+)\in\mathbb{A}}{\arg\min}\ \mathbb{E}_F\left[L(Y;c^-,v,c^+)\right].$$

- This can be used for separate regression modeling of body and tail of the data.

- This estimation problem is solved with a (simple) gradient descent algorithm, i.e., no EM algorithm is necessary.

Example of a consistent loss function

- Example of a consistent loss function for the composite triplet

$$L(y; c^-, v, c^+) = \left[1 + \frac{\psi_1'(c^-)}{\tau} + \frac{-\psi_2'(c^+)}{1-\tau}\right] L_\tau(y, v) + L_{\psi_1}(y, c^-) + L_{\psi_2}(y, c^+),$$

  with

  - ⋆ pinball loss $L_\tau(y, v)$, and
  - ⋆ Bregman divergences $L_{\psi_1}(y, c^-)$ and $L_{\psi_2}(y, c^+)$ + assumptions on ψ_1 and ψ_2.

- This looks like a non-parametric regression, but different choices of the convex functions ψ_1 and ψ_2 reflect different properties in tail and body of the data.

- **Deep composite regression models**

Deep multi-task learning

- Assume independent observations Y_i being established with features $\boldsymbol{x}_i \in \mathcal{X}$.
- Choose a deep neural network

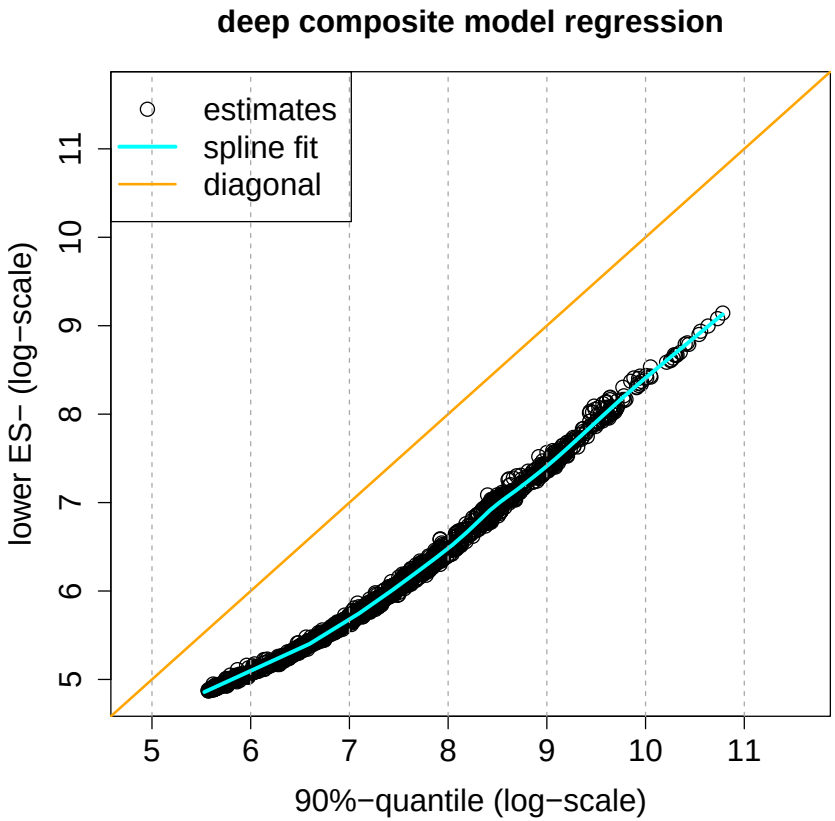
$$\boldsymbol{x} \in \mathcal{X} \;\mapsto\; \boldsymbol{z}^{(d:1)}(\boldsymbol{x}) = \left(\boldsymbol{z}^{(d)} \circ \cdots \circ \boldsymbol{z}^{(1)}\right)(\boldsymbol{x}) \;\in\; \mathbb{R}^{r_d}.$$

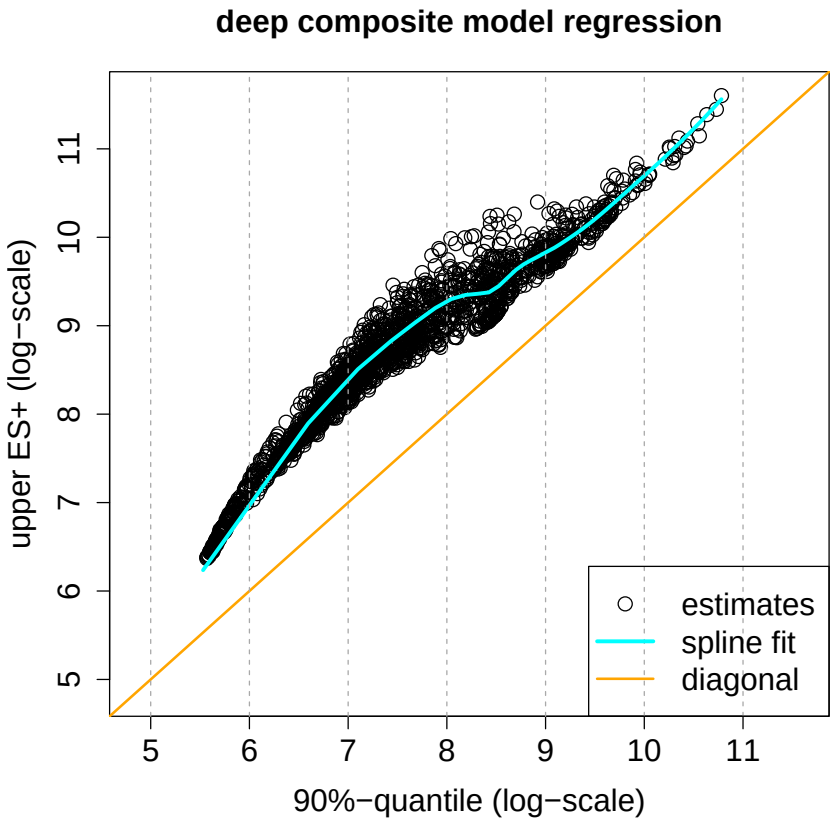
- Use this learned representation $\boldsymbol{z}^{(d:1)}(\boldsymbol{x})$ in a multi-output deep regression model

$$\begin{aligned}\boldsymbol{x} \;\mapsto\; & \left(\mathrm{CTE}_q^-(Y|\boldsymbol{x}),\; F_{Y|\boldsymbol{x}}^{-1}(q),\; \mathrm{CTE}_q^+(Y|\boldsymbol{x})\right) \\ & = \left(h\langle\boldsymbol{\beta}_1, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle,\; h\langle\boldsymbol{\beta}_2, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle, h\langle\boldsymbol{\beta}_3, \boldsymbol{z}^{(d:1)}(\boldsymbol{x})\rangle\right).\end{aligned}$$

- This multi-output network can be fitted with gradient descent and using a strictly consistent loss function L for the composite triplet.

Example: accident insurance data





Discussion

- Any functional $F \mapsto T(F)$ that is elicitable can be estimated with gradient descent (M-estimation).
- The composite triplet $\big(\mathrm{CTE}_\tau^-(Y),\, F^{-1}(\tau),\, \mathrm{CTE}_\tau^+(Y)\big)$ is elicitable.
- The composite triplet allows for estimating different regression functions in the body and the tail of the data.
- Different strictly consistent loss functions for the composite triplet implicitly imply different underlying distributional models. Asymptotically they all converge to the same limit, but they have different finite sample properties, see Gourieroux–Montfort–Trognon (1984).

References

- Fissler, Merz, Wüthrich (2021). Deep quantile and deep composite model regression. *arXiv*:2112.03075
- Fissler, Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics* **4474**, 1680-1707.
- Gneiting (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106/494**, 746-762.
- Gourieroux, Montfort, Trognon (1984). Pseudo maximum likelihood methods: theory. *Econometrica* **52/3**, 681-700.
- Nelder, Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135/3**, 370-384.
- Saerens (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks* **11**, 1263-1271.
- Savage (1971). Elicitable of personal probabilities and expectations. *Journal of the American Statistical Association* **66/336**, 783-810.
- Thomson (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory* **20**, 360-380.
- Wüthrich, Merz (2021). *Statistical Foundations of Actuarial Learning and its Applications.* SSRN Manuscript ID 3822407.