Evaluation of driving risk at different speeds

Guangyuan Gao[1]

School of Statistics, Renmin University of China

CAS annual meeting in Honolulu, November, 2019

[1]joint work with Hanfang Yang (Renmin University) and Mario V. Wüthrich (ETH Zurich).

Main conclusions based on our data set

1. Driving style is much more related to claims frequency than driving habit.
2. The driving style in $(0, 20]$km/h is most related to claims frequencies among the four speed buckets, and it also reflects the driving style at other speeds.

Table of Contents

Telematics car driving data

- Every second we receive the current speed and the acceleration in all directions from the internal sensor installed in the cars.
- We select the recorded speed and the recorded longitudinal acceleration to form the v-a heatmaps.
- We consider the telematics data of $n = 973$ cars during three months of driving experience from 01/05/2016 to 31/07/2016.
- An assumption is that a driver's driving characteristics remain the same during his/her policy period, since we apply the same telematics covariates for all policies of a given driver.

Partition of v-a rectangle

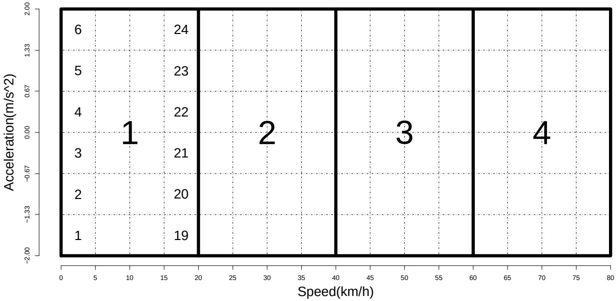

Figure 1: The partition of $R = (0, 80] \times [-2, 2]$.

For each speed bucket $m = 1, \ldots, 4$, we divide the v-axis (speed) into 4 intervals and the a-axis (acceleration) into 6 intervals, which results in 24 sub-rectangles $(R_{m,j})_{j=1:24}$ in each speed bucket m (see the numbers in speed bucket 1 in Figure 1).

Normalization in each speed bucket

- For each driver i, we denote the amount of time spent in $R_{m,j}$ by $t_{i,m,j}$.
- Given a speed bucket m, for each driver i we calculate the relative amount (normalized amount) of time spent in $R_{m,j}$ as

$$z_{i,m,j} = \frac{t_{i,m,j}}{t_{i,m}} \geq 0, \tag{1}$$

  where $t_{i,m} = \sum_{j=1}^{24} t_{i,m,j}$ is the total amount of time spent in speed bucket m by driver i.
- Equation (1) induces an empirical discrete distribution $\boldsymbol{z}_{i,m} = (z_{i,m,1}, \ldots, z_{i,m,24})'$ on speed bucket m.
- $\boldsymbol{z}_{i,m}, m = 1, \ldots, 4$ can be illustrated by v-a heatmaps.

v-a heatmaps of three drivers

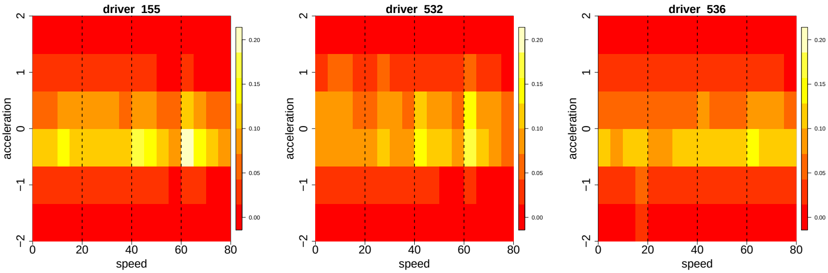

Figure 2: v-a heatmaps of drivers 155, 532and 536.

Driving style

- The *driving style* of every car driver i is described by a J-vector $\boldsymbol{x}_i = (\boldsymbol{z}_{i,1}', \ldots, \boldsymbol{z}_{i,4}')' \in \mathbb{R}^J$ containing the four discrete distributions $\boldsymbol{z}_{i,m}$ on the rectangle $m = 1, \ldots, 4$.
- Note that the dimension of $\boldsymbol{x}_i$ is $J = 24 \times 4 = 96$.

Driving habit

- *Driving habit* of driver i is defined to be the relative amount of time spent in each speed bucket m:

$$h_{i,m} = \frac{t_{i,m}}{t_i}, \qquad \text{for } m = 1, \ldots, 4, \tag{2}$$

  where $t_i = \sum_{m=1}^{4} t_{i,m}$ is the total amount of time spent in the entire speed interval $(0, 80]$km/h by driver i.

- Another driving habit covariate is the average driving hours in $(0, 80]$km/h per week, defined as

$$ave\_hours_i = \frac{t_i \times 7}{3600 \times 92},$$

  which indicates the intensity of driving.

Driving habit v.s driving style

- Suppose that a commuting driver i and an off-peak driver i' had the same driving style, we would have $h_{i,1} > h_{i',1}, h_{i,4} < h_{i',4}$, but $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$.

Design matrix of driving style

- For each speed bucket m, we stack the vectors $\boldsymbol{z}_{i,m}, i = 1, \ldots, n$, to form the $n \times 24$ design matrix $\boldsymbol{X}_m \in \mathbb{R}^{n \times 24}$.
- For the four speed buckets altogether, we stack the vectors $\boldsymbol{x}_i, i = 1, \ldots, n$, to form the $n \times J$ design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times J}$.
- Denote the normalized design matrices by $(\boldsymbol{X}_m^0)_{m=1:4}$ and $\boldsymbol{X}^0$ (all column means are set to zero and variances are normalized to one).
- Denote the corresponding i-th row by $(\boldsymbol{z}_{i,m}^0)_{m=1:4}$ and $\boldsymbol{x}_i^0$.

Singular value decomposition

Singular value decomposition of $\boldsymbol{X}^0$ is as follows:

$$\boldsymbol{X}^0 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}',$$

where $\boldsymbol{U}$ is an $n \times J$ orthogonal matrix, $\boldsymbol{V}$ is a $J \times J$ orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}(g_1, \ldots, g_J)$ is a $J \times J$ diagonal matrix with singular values.

- The w-th column of the rotation matrix $\boldsymbol{V}$ is the w-th principal component loading vector (or right-singular vector) $\boldsymbol{v}_w = (v_{1,w}, \ldots, v_{J,w})', w = 1, \ldots, J$.
- The w-th principal component of driver i is the projected value of $\boldsymbol{x}_i^0$ onto the direction $\boldsymbol{v}_w$

$$p_{i,w} = \sum_{j=1}^{J} v_{j,w} x_{i,j}^0.$$

The first two loading vectors

- We illustrate the proportion of explained variance in $\boldsymbol{X}^0$ by the principal components in Figure 3 (left).
- The first 20 principal components explain around 95% of the total variance in $\boldsymbol{X}^0$. Therefore, we only consider the first 20 principal components in claims frequency modeling.
- In Figure 3 we show the first and second loading vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ in its corresponding sub-rectangle.

The first two loading vectors

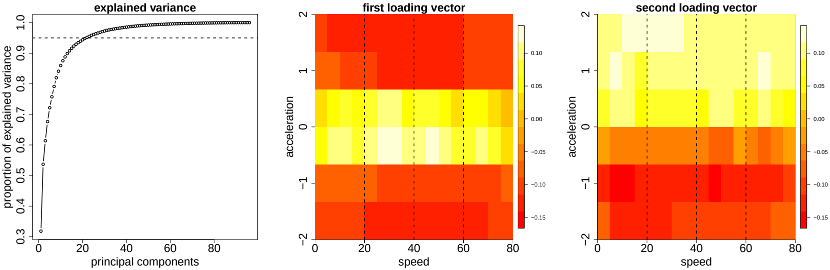

Figure 3: The proportion of explained variance by the principal components (left). The first and second loading vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (middle and right).

- The first principal component reflects the degree of concentration on the zero acceleration rate.
- The second principal component illustrates the frequency difference between acceleration and braking.

The principal components in each speed bucket

- We apply the principal component analysis to the matrices $(\boldsymbol{X}_m^0)_{m=1:4}$, respectively.
- We denote by $p_{i,w}^m, w = 1, \ldots, 24, m = 1, \ldots, 4$, the w-th principal component of driver i in speed bucket m.
- In Table 2, we calculate the coefficient of correlation among the first two principal components $p_{i,1}^m, p_{i,2}^m$.

The principal components in each speed bucket

Table 2: The coefficients of correlation among the first two principal components $p_{i,1}^m, p_{i,2}^m$.

	$p_{i,1}^1$	$p_{i,1}^2$	$p_{i,1}^3$	$p_{i,1}^4$	$p_{i,2}^1$	$p_{i,2}^2$	$p_{i,2}^3$	$p_{i,2}^4$
$p_{i,1}^1$	1.00	0.86	0.69	0.55	0	-1.2$\times 10^{-2}$	1.1$\times 10^{-2}$	3.0$\times 10^{-2}$
$p_{i,1}^2$	0.86	1.00	0.87	0.70	-2.2$\times 10^{-2}$	0	1.5$\times 10^{-2}$	3.9$\times 10^{-2}$
$p_{i,1}^3$	0.69	0.87	1.00	0.92	-8.3$\times 10^{-2}$	-4.6$\times 10^{-2}$	0	2.3$\times 10^{-2}$
$p_{i,1}^4$	0.55	0.70	0.92	1.00	-1.3$\times 10^{-1}$	-8.9$\times 10^{-2}$	-2.4$\times 10^{-2}$	0
$p_{i,2}^1$	...	...	...	...	1.00	0.95	0.91	0.86
$p_{i,2}^2$	...	...	...	...	0.95	1.00	0.96	0.89
$p_{i,2}^3$	...	...	...	...	0.91	0.96	1.00	0.93
$p_{i,4}^4$	...	...	...	...	0.86	0.89	0.93	1.00

It shows that the driving characteristics in different speed buckets are quite similar in terms of the first two principal components.

Three aspects to be investigated

1. The predictive performance of driving habit covariates $(h_{i,m})_{m=1:4}$ and ave_hours_i;
2. The predictive performance of driving style covariates $(p_{i,w})_{w=1:20}$;
3. The predictive performance of the covariates $(p^m_{i,w})_{w=1:7}$ in each speed bucket m.

Claims data

- We consider the compolsory third party policies purchased by these $n = 973$ cars (these policies have all the same coverage limit of CNY $122,000$).
- We record the number of reported claims from 01/01/2014 to 29/06/2017. The total exposure is $2,179.5$ years-at-risk with the empirical frequency of 0.24.

Four classical risk factors

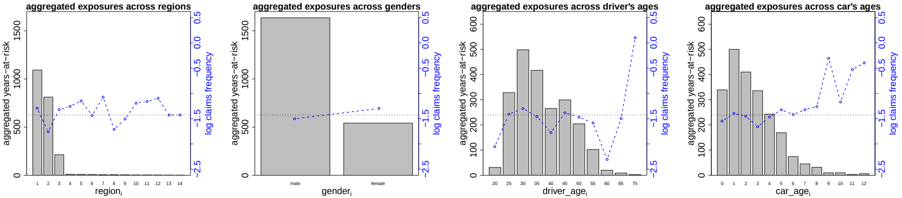

Figure 4: Distribution of aggregated years-at-risk (left axis) and the corresponding logarithm of the empirical claims frequencies (right axis) across the four classical risk factors: regions, gender, driver's age, and car's age.

Driving habit covariates and driving style covariates

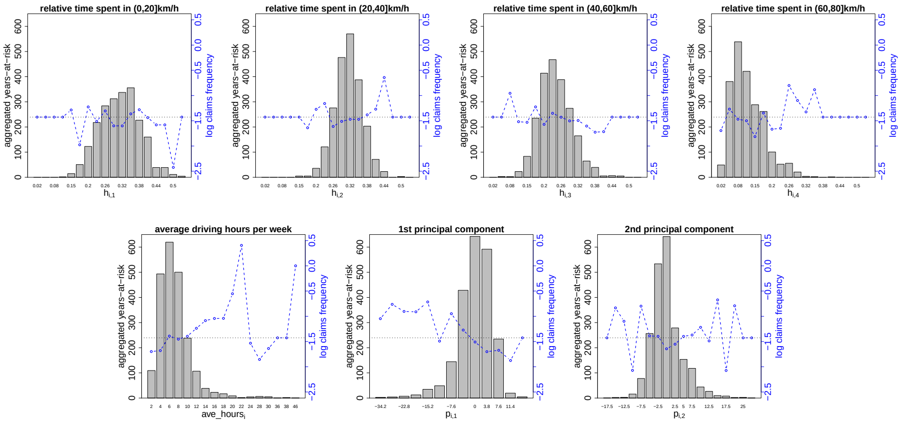

Figure 5: Distribution of aggregated years-at-risk and the corresponding logarithm of the empirical claims frequencies across the driving habit covariates and the selected driving style covariates.

General setting

We assume that the number of claims Y_i of driver i follows a Poisson distribution with an underlying expected claims frequency of λ_i per year:

$$\begin{aligned} Y_i &\overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with} \\ \log \lambda_i &= \beta_0 + \alpha_{u_i} + \beta_1 v_i + s(w_i; \boldsymbol{\beta}_2, \delta), \end{aligned} \tag{4}$$

- $e_i \in [1, 3.5]$ years-at-risk is the total exposure of driver i.
- The non-linear effect of w_i is described by a penalized thin plate regression spline s with regression parameters $\boldsymbol{\beta}_2$ and smoothing parameter δ. By using the penalized thin plate regression splines, we do not need to specify the knots (Section 4.1.5 of Wood [17]).

Backward elimination, cross validation

- We always start with a full model containing all the considered covariates.
- Then we sequentially drop the single covariate with the highest non-significant p-value from the model and refit the model until all the covariates are significant.
- We randomly partition the data of all cars $\mathcal{N}$ into 10 roughly equally-sized disjoint parts, denoted by $\mathcal{T}_1, \ldots, \mathcal{T}_{10}$.
- We estimate the average Poisson deviance loss by 10-fold cross validation as

$$\widehat{D} = \frac{1}{10} \sum_{l=1}^{10} D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l}), \qquad (5)$$

where $D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l})$ is the average Poisson deviance loss on the data $\mathcal{T}_l$ using the estimated claims frequencies $\lambda_i(\hat{\theta}_{-\mathcal{T}_l})$

$$D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l}) = \frac{2}{|\mathcal{T}_l|} \sum_{i \in \mathcal{T}_l} Y_i \left[\frac{\lambda_i(\hat{\theta}_{-\mathcal{T}_l}) e_i}{Y_i} - 1 - \log \left(\frac{\lambda_i(\hat{\theta}_{-\mathcal{T}_l}) e_i}{Y_i} \right) \right].$$

GAM with the classical risk factors

- We start with the model

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1) + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) \tag{7}$$

- We apply the backward elimination to model (7) to remove driver's age and gender sequentially. The resulting model is

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2). \tag{8}$$

- We also fit an intercept model for comparison:

$$\log \lambda_i = \beta_0. \tag{9}$$

GAM with driving habit covariates

- A starting point of backward elimination is to include linear terms of $(h_{i,m})_{m=1:4}$ and the smooth term of ave_hours_i:

$$\begin{aligned}\log \lambda_i =& \beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1) + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) \\ &+ \beta_1^h h_{i,1} + \beta_2^h h_{i,2} + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h).\end{aligned} \tag{10}$$

- Note that we have removed $h_{i,4}$ in the model because there is a constraint of $\sum_{m=1}^4 h_{i,m} = 1$ and most cars spend the least time in $(60, 80]$km/h.
- The backward elimination leads to the following regression function:

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h). \tag{11}$$

GAM with driving habit and driving style covariates

- If we start with smooth terms of driving habit and style covariates, the backward elimination leads to the following model:

$$\begin{aligned}\log\lambda_i =&\beta_0 + \alpha_{region_i} + \beta_2^h h_{i,2} + \beta_3^h h_{i,3} + \beta_5^h ave\_hours_i \\ &+ \beta_1^p p_{i,1} + \beta_7^p p_{i,7} + \beta_{15}^p p_{i,15} + \beta_{16}^p p_{i,16} \\ &+ r_8(p_{i,8}; \boldsymbol{\beta}_8^p, \delta_8^p) + r_{10}(p_{i,10}; \boldsymbol{\beta}_{10}^p, \delta_{10}^p) + r_{12}(p_{i,12}; \boldsymbol{\beta}_{12}^p, \delta_{12}^p).\end{aligned} \tag{12}$$

- If we start with linear terms of driving habit and style covariates, the backward elimination leads to the following model:

$$\begin{aligned}\log\lambda_i =&\beta_0 + \alpha_{region_i} + \beta_3^h h_{i,3} + \beta_5^h ave\_hours_i \\ &+ \beta_1^p p_{i,1} + \beta_3^p p_{i,3} + \beta_7^p p_{i,7} + \beta_{10}^p p_{i,10}.\end{aligned} \tag{13}$$

- We calculate the weight for sub-rectangle j as $\hat\beta_1^p v_{j,1} + \hat\beta_3^p v_{j,3} + \hat\beta_7^p v_{j,7} + \hat\beta_{10}^p v_{j,10}$ for $j = 1, \ldots, J$. We plot these weights in the v-a rectangle according to their signs in Figure 7.

GAM with driving habit and driving style covariates

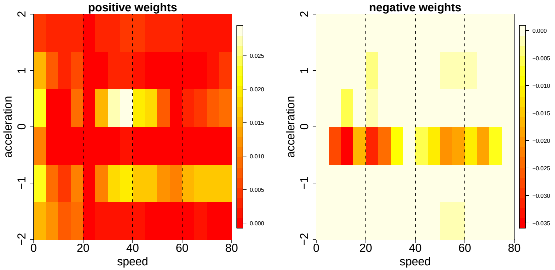

Figure 7: The weights on the v-a rectangle in model (13).

- Most sub-rectangles in $(0, 20]$km/h are highlighted, indicating that $(0, 20]$km/h is important in predicting claims frequency.
- Hard brake and acceleration have the positive effect on claims frequency, while smooth brake and acceleration have the negative effect on claims frequency.

GAM with driving style covariates in each speed bucket

For each speed bucket m, we either start with the model

$$\begin{aligned}\log \lambda_i =&\beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1) + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2)\\ &+ f_1(h_{i,1}; \boldsymbol{\beta}_1^h, \delta_1^h) + f_2(h_{i,2}; \boldsymbol{\beta}_2^h, \delta_2^h) + f_3(h_{i,3}; \boldsymbol{\beta}_3^h, , \delta_3^h) + f_4(h_{i,4}; \boldsymbol{\beta}_4^h, \delta_4^h)\\ &+ f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h)\\ &+ r_1^m(p_{i,1}^m; \boldsymbol{\beta}_1^m, \delta_1^m) + \ldots + r_7^m(p_{i,7}^m; \boldsymbol{\beta}_7^m, \delta_7^m),\end{aligned} \tag{14}$$

or start with the model with only driving style covariates

$$\log \lambda_i = \beta_0 + r_1^m(p_{i,1}^m; \boldsymbol{\beta}_1^m, \delta_1^m) + \ldots + r_7^m(p_{i,7}^m; \boldsymbol{\beta}_7^m, \delta_7^m), \tag{15}$$

GAM with driving style covariates in each speed bucket

The backward elimination leads to the following models:

1 The first speed bucket $(0, 20]$km/h.

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h) + \beta_1^1 p_{i,1}^1. \quad (16)$$

$$\log \lambda_i = \beta_0 + \beta_1^1 p_{i,1}^1. \quad (17)$$

2 The second speed bucket $(20, 40]$km/h.

$$\begin{aligned}\log \lambda_i =& \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h) \\ &+ \beta_1^2 p_{i,1}^2 + r_7^2(p_{i,7}^2; \boldsymbol{\beta}_7^2, \delta_7^2).\end{aligned} \quad (18)$$

$$\log \lambda_i = \beta_0 + \beta_1^2 p_{i,1}^2. \quad (19)$$

3 The third speed bucket $(40, 60]$km/h.

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h) + \beta_1^3 p_{i,1}^3. \quad (20)$$

$$\log \lambda_i = \beta_0 + \beta_1^3 p_{i,1}^3 + \beta_4^3 p_{i,4}^3. \quad (21)$$

4 The forth speed bucket $(60, 80]$km/h.

$$\log \lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}_5^h, \delta_5^h) + \beta_1^4 p_{i,1}^4. \quad (22)$$

$$\log \lambda_i = \beta_0 + \beta_1^4 p_{i,1}^4. \quad (23)$$

The selected representative models

Table 4: The selected representative models.

model index	covariates in the model	equation
1	no covariates	(9)
2	classical	(8)
3	classical, driving habit	(11)
4	classical, driving habit, driving style (in smooth terms)	(12)
5	classical, driving habit, driving style (in linear terms)	(13)
6	classical, driving habit, driving style of $(0, 20]$km/h	(16)
7	classical, driving habit, driving style of $(20, 40]$km/h	(18)
8	classical, driving habit, driving style of $(40, 60]$km/h	(20)
9	classical, driving habit, driving style of $(60, 80]$km/h	(22)
10	driving style of $(0, 20]$km/h	(17)
11	driving style of $(20, 40]$km/h	(19)
12	driving style of $(40, 60]$km/h	(21)
13	driving style of $(60, 80]$km/h	(23)

UBRE, AIC and average Poisson deviance loss

We plot the UBRE, the AIC and the average Poisson deviance loss with 90% interval for these selected models in Figure 8.

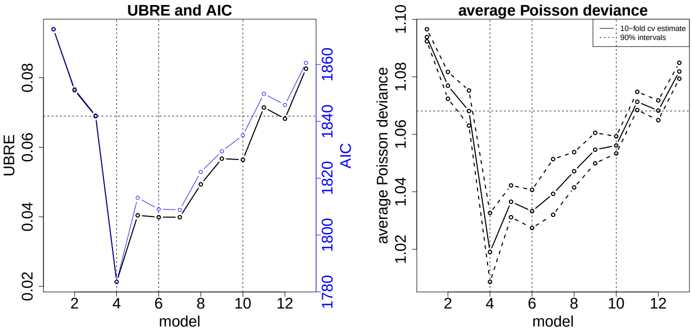

Figure 8: The UBRE, the AIC and the average Poisson deviance loss with 90% interval for the models in Table 4.

Main:

- Driving style is much more related to claims frequency than driving habit.
- The driving style in $(0, 20]$km/h is most related to claims frequencies among the four speed buckets, and it also reflects the driving style at other speeds.

- Ayuso, M., Guillen, M., Pérez-Marín, A.M. (2016). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks,* **4/2**, article 10.
- Ayuso, M., Guillen, M., Nielsen, J.P. (2018). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, DOI: 10.1007/s11116-018-9890-7.
- Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101/473**, 119-137.
- Boucher, J.-P., Côté, S., Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, **5**, article 54.
- Gao, G., Meng, S., Wüthrich, M.V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, **2019/2**, 143-162.

Guillen, M., Nielsen, J.P., Ayuso, M., Pérez-Marín, A.M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, **39/3**, 662-672.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction,* second edition. New York: Springer-Verlag.

Hung,W.T., Tong, H.Y., Lee, C.P., Ha, K., Pao, L.Y. (2007). Development of practical driving cycle construction methodology: a case study in Hong Kong. *Transportation Research Part D: Transport and Environment*, **12/2**, 115-128.

Kaufman, L., Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

Denuit, M., Guillen, M., Trufin, J. (2019). Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, DOI: 10.1017/S1748499518000349.

Paefgen, J., Staake, T., Fleisch, E. (2014). Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data. *Transportation Research A: Policy and Practice*, **61**, 27-40.

Reynolds, A., Richards, G., de la Iglesia, B., Rayward-Smith, V. (1992). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, **5/4**, 475-504.

Verbelen, R., Antonio, K., Claeskens, G. (2018). Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67/5**, 1275-1304.

Wang, Q., Huo, H., He, K., Yao, Z., Zhang, Q. (2008). Characterization of vehicle driving patterns and development of driving cycles in Chinese cities. *Transportation Research Part D: Transport and Environment*, **13/5**, 289-297.

- Weidner, W., Transchel, F.W.G., Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, **6/1**, 3-24.
- Weidner, W., Transchel, F.W.G., Weidner, R. (2016). Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science*, **11/2**, 213-236.
- Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, second edition. New York: Chapman & Hall.

Thank you!

Q & A