mcvis: multicollinearity visualisation

https://kevinwang09.github.io/pres/mcvis_talk

Kevin Y. X. Wang

5th December 2019, Adelaide

THE UNIVERSITY OF SYDNEY

Acknowledgement

This is joint work with Chen Lin (Fudan Univeristy) and Prof Samuel Mueller (Sydney University).

Cricketers' career batting statistics

- Cricket is a bat-and-ball game.
- The aim of a batsman is to score as many **runs** as possible before getting **out**.

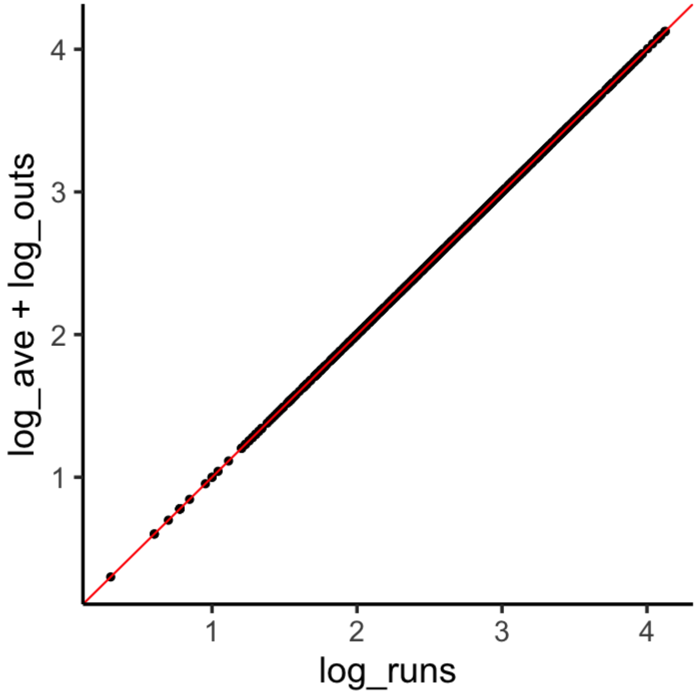
```
glimpse(X)
```

```
## Observations: 810
## Variables: 8
## $ log_runs  <dbl> 2.204120, 1.556303, 2.840106, 2.683947, 2.008600, 3.210051,…
## $ log_ave   <dbl> 1.1625644, 0.7781513, 1.4250449, 1.3617278, 0.7781513, 1.59…
## $ log_outs  <dbl> 1.0413927, 0.7781513, 1.4149733, 1.3222193, 1.2304489, 1.61…
## $ log_fours <dbl> 1.278754, 0.301030, 1.832509, 1.832509, 1.041393, 2.158362,…
## $ log_sixes <dbl> 0.0000000, 0.0000000, 0.4771213, 0.8450980, 0.3010300, 0.60…
## $ log_ducks <dbl> 0.6989700, 0.4771213, 0.6020600, 0.6020600, 1.0413927, 0.77…
## $ log_hs    <dbl> 2.071882, 1.255273, 2.021189, 2.004321, 1.414973, 2.103804,…
## $ log_100   <dbl> 0.3010300, 0.0000000, 0.3010300, 0.3010300, 0.0000000, 0.69…
```

Interesting feature in this data

There is a causal relationship:

$$\text{batting ave} = \frac{\text{runs}}{\text{no. of outs}}, \qquad \text{or equivalently,} \qquad \texttt{log\_runs} = \texttt{log\_ave} + \texttt{log\_outs}.$$



What is multi-collinearity (MC)?

MC occurs when columns of X are linear dependent (exactly or approximately).

```
M1 = lm(log_100 ~ ., data = X)
broom::tidy(M1)
```

```
## # A tibble: 8 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -0.365      0.0912   -4.01   6.74e- 5
## 2 log_runs    -2.63      71.4      -0.0368 9.71e- 1
## 3 log_ave      2.55      71.4       0.0357 9.72e- 1
## 4 log_outs     2.32      71.4       0.0325 9.74e- 1
## 5 log_fours    0.647      0.0978    6.61   6.90e-11
## 6 log_sixes    0.132      0.0264    4.98   7.87e- 7
## 7 log_ducks    0.00536    0.0498    0.108  9.14e- 1
## 8 log_hs      -0.0178     0.0752   -0.237  8.13e- 1
```

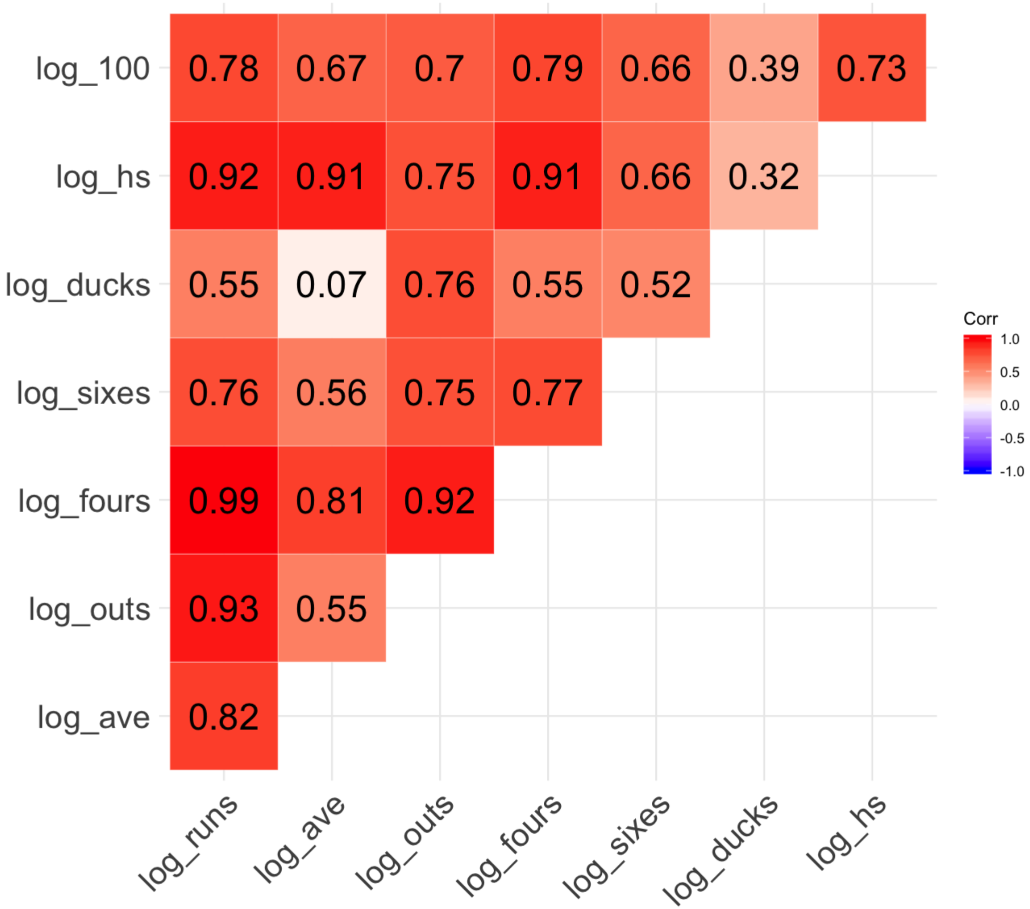
Consequence of multi-collinearity

- Numerical instability is a typical sympton of MC.

	Include all			**Remove log_runs**			**Remove log_ave**		
Predictors	*Estimates*	*std. Error*	*p*	*Estimates*	*std. Error*	*p*	*Estimates*	*std. Error*	*p*
(Intercept)	-0.37	0.09	**<0.001**	-0.37	0.09	**<0.001**	-0.37	0.09	**<0.001**
log_runs	-2.63	71.43	0.971				-0.08	0.12	0.517
log_ave	2.55	71.42	0.972	-0.08	0.12	0.517			
log_outs	2.32	71.45	0.974	-0.31	0.11	**0.003**	-0.23	0.10	**0.017**
log_fours	0.65	0.10	**<0.001**	0.65	0.10	**<0.001**	0.65	0.10	**<0.001**
log_sixes	0.13	0.03	**<0.001**	0.13	0.03	**<0.001**	0.13	0.03	**<0.001**
log_ducks	0.01	0.05	0.914	0.01	0.05	0.914	0.01	0.05	0.914
log_hs	-0.02	0.08	0.813	-0.02	0.08	0.813	-0.02	0.08	0.813

- We will proceed with rounding all variables to 2 significant figures.

High correlation $\neq$ multicollinearity



- By definition, it is the linear combination of variables that causes MC.
- The causal variables are not the most highly correlated.
- Thus, identifying high correlation does not always identify sources of MC.

> Diagnosis of multicollinearity requires specialised statistics.

Existing methodologies

1. Variance inflation factors (VIFs)

Introduced in Marquaridt (1970):

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, \ldots, p,$$

where R_j^2 is the coefficient of determination when the $\boldsymbol{x}_j$ independent variable is treated as a response variable against the remaining $p-1$ independent variables.

A **larger** value of VIF_j implies $\boldsymbol{x}_j$ can be highly predicted by other variables, and thus implies higher cause of MC by that variable.

```
M1 = lm(log_100 ~ ., data = X)
M1 %>% car::vif() %>% round(2)
```

```
##   log_runs    log_ave   log_outs log_fours log_sixes log_ducks     log_hs
##   23995.96    4666.15   11410.15     55.60      2.53      3.99      12.17
```

- Using a threshold of 5 endorsed by Sheather (2009), 5 MC-causing variables are identified.

2.Eigenvalues of $X^\top X$

Eigenvalues of the "uncentered covariance matrix" $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ offers a more linear algebra interpretation of MC.

A **smaller** value of λ_p produces a matrix determinant closer to 0, which implies linear dependence in X and thus MC (Stewart 1987).

```
Xmat = X %>% as.data.frame() %>% as.matrix()
eigen = svd(t(Xmat) %*% Xmat)
round(eigen$d, 3)
```

```
## [1] 16063.810   222.296   114.572    60.994    23.306     8.941     6.428
## [8]     0.006
```

Note: this only implicates the existence of MC, not which variable causes MC.

Relationships between the two measures

Suppose that X is standardised to have mean 0 and variance 1, and we decompose $(X^\top X)^{-1}$ into $G\,\mathrm{diag}(1/\lambda_1,\ldots,1/\lambda_p)G^\top$, then:

$$\begin{pmatrix} VIF_1 \\ \vdots \\ VIF_p \end{pmatrix} = \begin{pmatrix} g_{11}^2 & \cdots & g_{1p}^2 \\ \vdots & \ddots & \vdots \\ g_{p1}^2 & \cdots & g_{pp}^2 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_p \end{pmatrix} = (G \otimes G)\boldsymbol{\tau},$$

where $\tau_j = 1/\lambda_j, \quad j = 1,\ldots,p.$

> Larger τ_p value indicates great MC.

- It will be great if we have a formula of the form $\tau_p = f(VIF_1,\ldots,VIF_p)$ to reveal the relationship between every variable $\boldsymbol{x}_j$ and the cause of MC, τ_p.
- But $G \otimes G$ is generally not invertible.

The mcvis method

mcvis

We perform linear regression between τ_p and every VIF.

- By quantifying the linearity between τ_p and VIFs, we can diagnose MC-causing variables.
- How can we generate multiple "observations" of both τ_p and VIFs?
- Sampling!

$VIF_1, \dots, VIF_p$

$\tau_1, \dots, \tau_p$

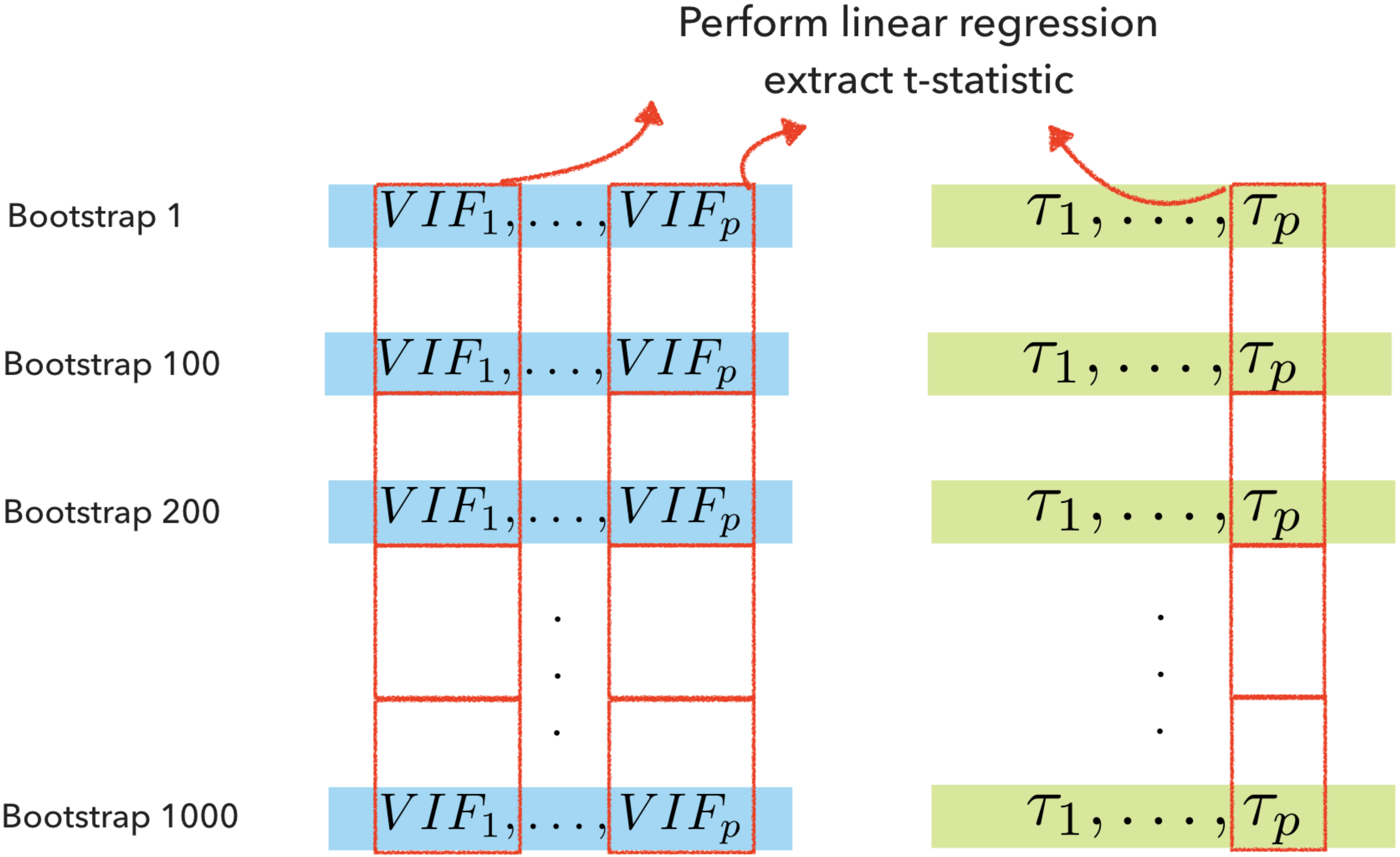
Bootstrap 1	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 100	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 200	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
	$\vdots$	$\vdots$
Bootstrap 1000	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$

Perform linear regression
extract t-statistic

Bootstrap 1	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 100	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 200	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
	$\vdots$	$\vdots$
Bootstrap 1000	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$

Perform linear regression
extract t-statistic

Bootstrap 1	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 100	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
Bootstrap 200	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$
	$\vdots$	$\vdots$
Bootstrap 1000	$VIF_1, \ldots, VIF_p$	$\tau_1, \ldots, \tau_p$


Perform linear regression
extract t-statistic
Bootstrap 1
Bootstrap 100
Bootstrap 200
Bootstrap 1000
$VIF_1, \ldots, VIF_p$
$\tau_1, \ldots, \tau_p$

Bootstrap 1

$$t_{1,1}, \ldots, t_{p,1}$$

Bootstrap 100

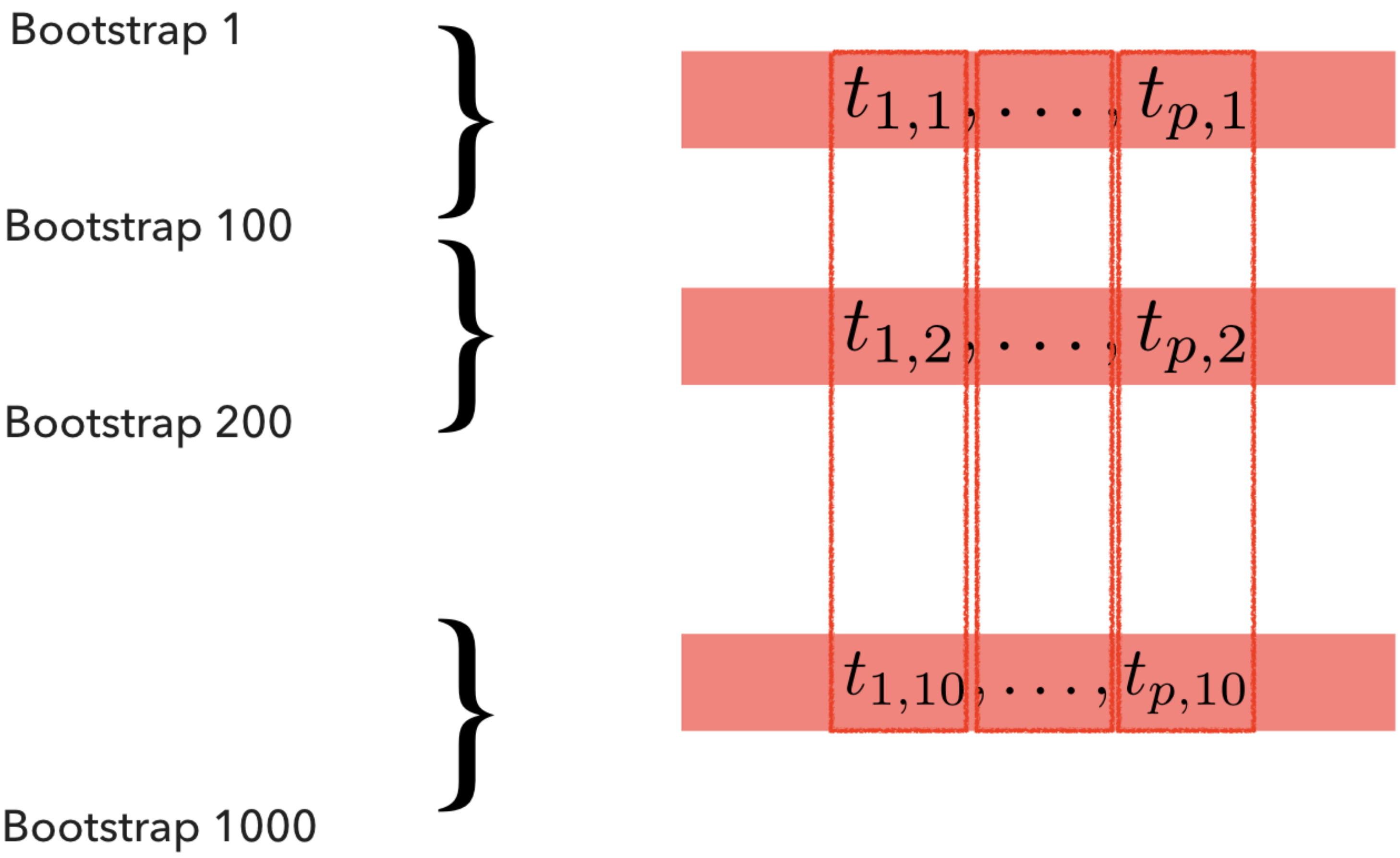
$$t_{1,2}, \ldots, t_{p,2}$$

Bootstrap 200

$$t_{1,10}, \ldots, t_{p,10}$$

Bootstrap 1000

$$\overline{t_j^2} = \left(\sum_{k=1}^{K} t_{j,k}^2\right) / K$$



$$\overline{t_j^2} = \left(\sum_{k=1}^{K} t_{j,k}^2 \right) / K$$

$$\overline{t_1^2}, \ \overline{t_2^2}, \ \ldots, \ \overline{t_p^2}$$

$$\boxed{MC_j = \frac{\overline{t_j^2}}{\sum_{j=1}^{p} \overline{t_j^2}}}$$

The `mcvis` package

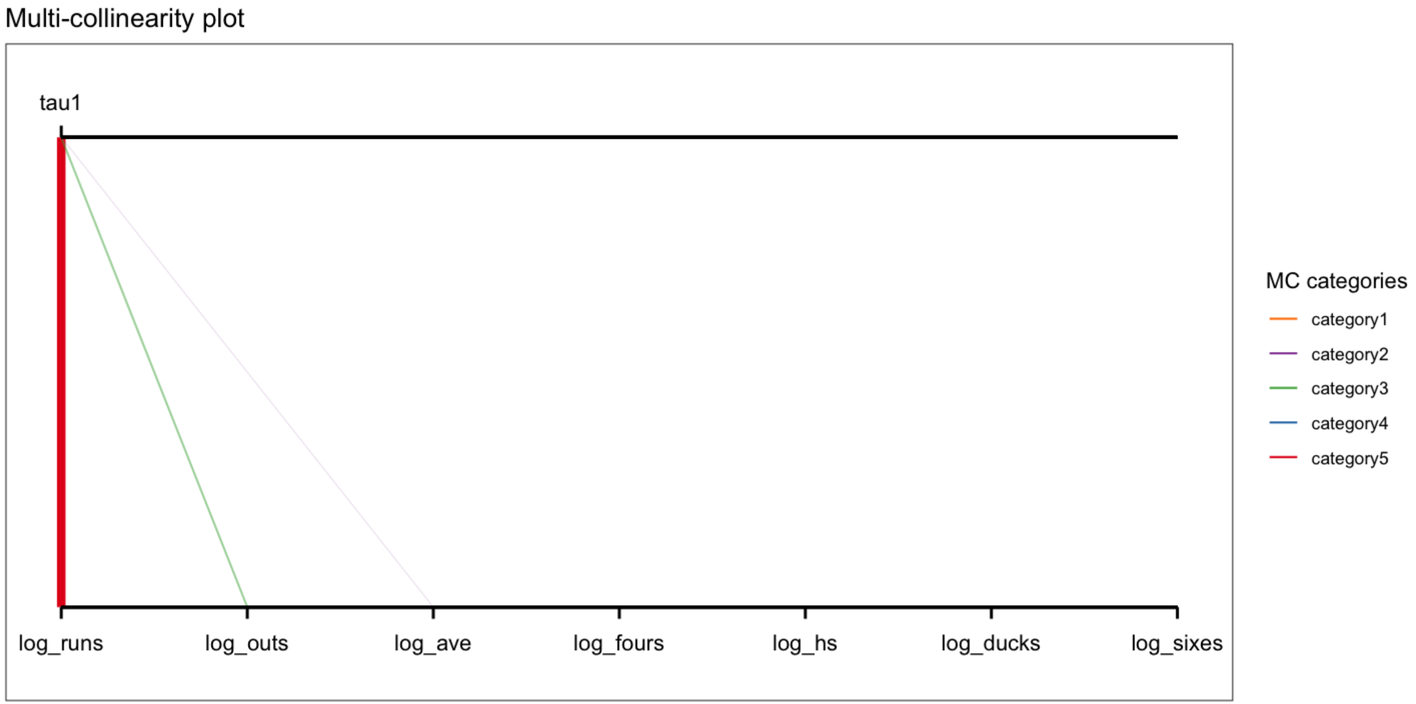
1. MC-index

```r
library(mcvis)
set.seed(13)
p = ncol(X)
mcvis_result = mcvis(X[,-p])
round(mcvis_result$MC[p-1,], 2)
```

```
##   log_runs    log_ave   log_outs log_fours log_sixes log_ducks     log_hs
##       0.69       0.14       0.16      0.00      0.00      0.00       0.00
```

2. MC visualisation

```
ggplot_mcvis(mcvis_result)
```



3. Shiny app for interactive exploration of data

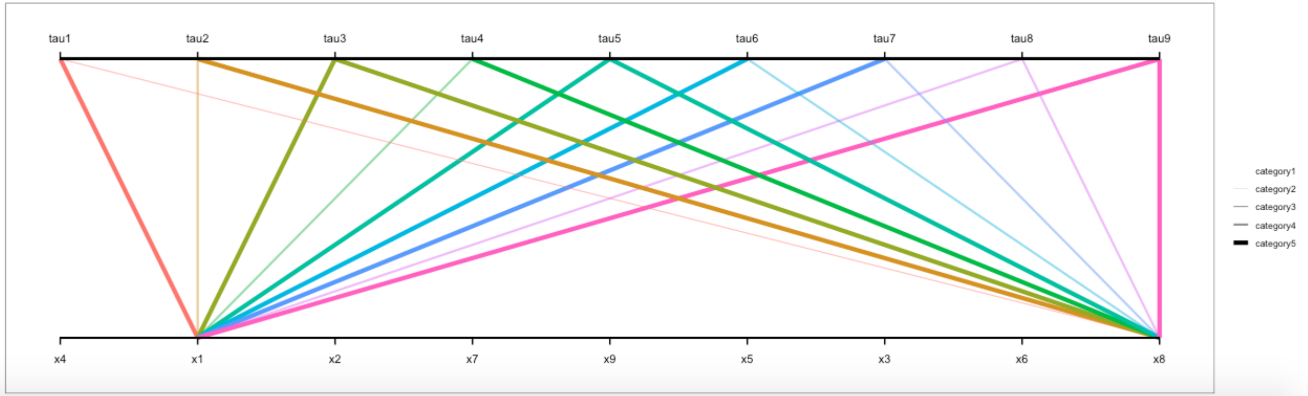
Show 10 entries

Search:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	1	50	0.096	2	0.1	0.15	1.9	-5.8	4.1	9.9	-0.31	0.16	0.29
2	2	50	-0.28	2.7	-0.35	-0.37	3.3	-5.4	6.9	12	0.34	-0.29	0.39
3	3	50	-0.09	2.4	-0.15	-0.082	2.1	-5.7	8.5	14	0.43	1.7	0.35
4	4	50	-0.51	3.2	-0.85	-0.44	2.5	-10	5.5	16	-0.31	0.47	0.45
5	5	50	-0.43	3.1	-0.45	-0.39	3	-9.3	6.6	16	-0.2	0.54	0.44
6	6	50	-0.4	3.9	-0.6	-0.36	4.1	-9.3	8.1	17	-0.094	-0.68	0.55
7	7	50	-0.2	2.6	0.15	-0.082	2.7	-6.7	5.5	12	-0.39	-0.15	0.36
8	8	50	1	5.6	1.1	0.8	6.3	-9.4	18	27	0.49	0.2	0.79
9	9	50	0.53	2.9	0.5	0.49	3.6	-4.8	5.8	11	0.12	-0.97	0.4

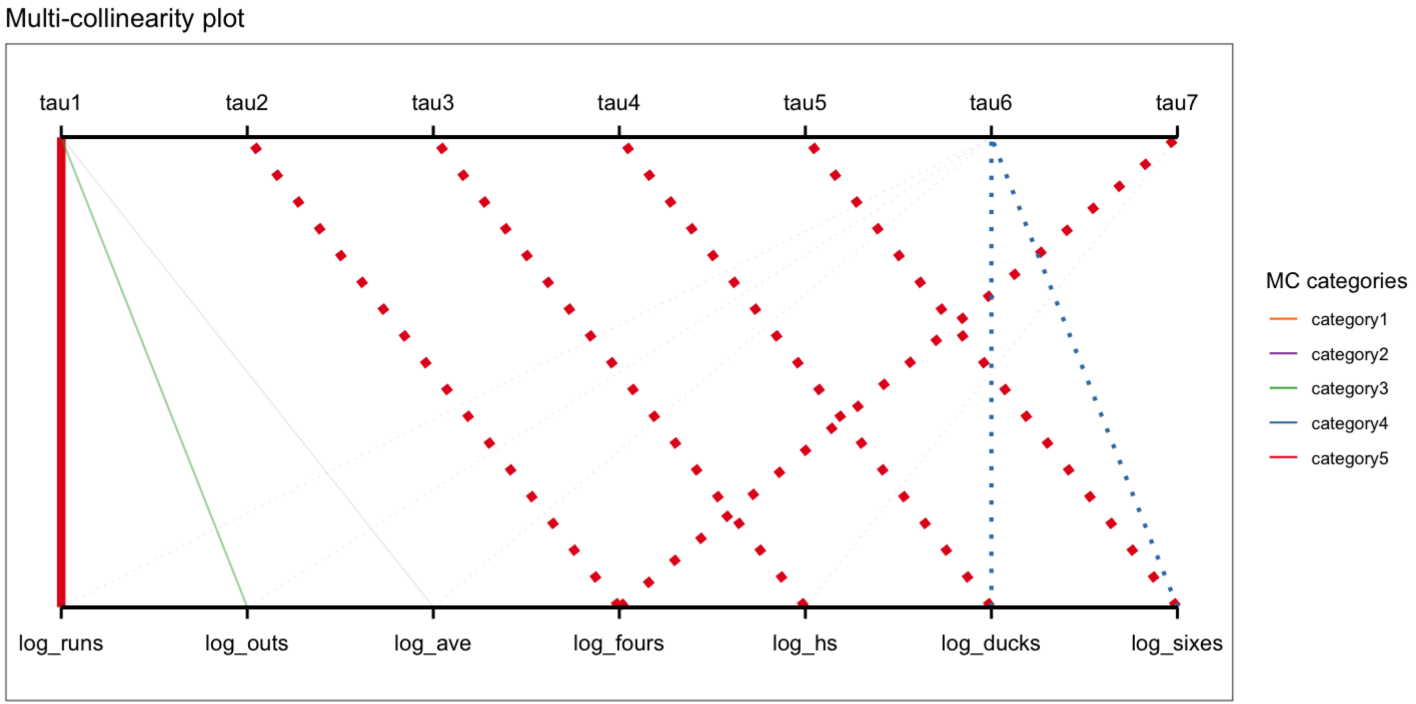
Showing 1 to 9 of 9 entries

Previous 1 Next



Extension work: Multiple τ's

```r
ggplot_mcvis(mcvis_result, eig_max = 7)
```



Final remarks

- mcvis provides a new MC-index and a visualisation of multicollinearity in linear regression.
- mcvis builds on top of classical statistics under a resampling framework and uncovers new sources of collinearity with an understanding of variability.
- Learn more from:
  - leaffur/mcvis
  - kevinwang09/mcvispy
  - samuel.mueller@sydney.edu.au
  - @KevinWang009 and @SamuelMuller74

Bibliography