mcvis: multicollinearity visualisation

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

Kevin Y. X. Wang

5 December 2019, Adelaide

THE UNIVERSITY OF SYDNEY

Acknowledgement

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

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**.

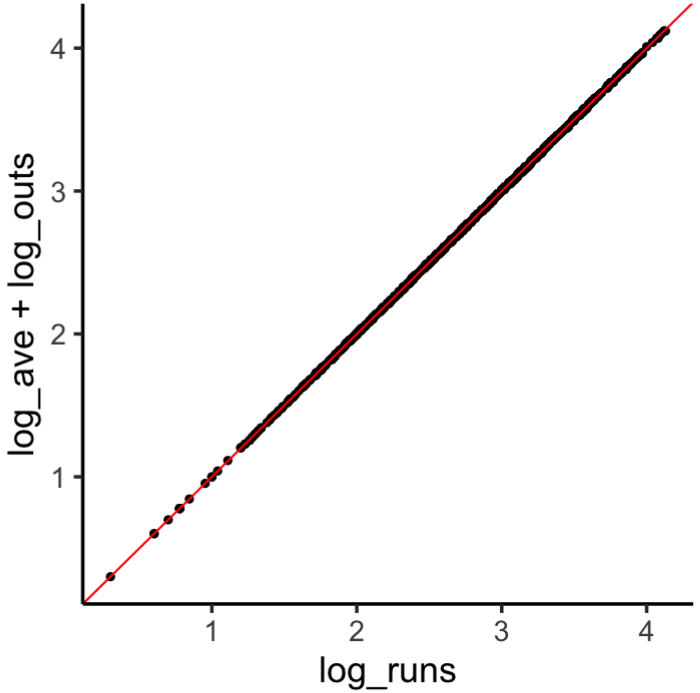
```
glimpse(X)
```

```
## Observations: 810
## Variables: 8
## $ log_runs  <dbl> 2.20, 1.56, 2.84, 2.68, 2.01, 3.21, 2.03, 2.65, 3.13, 2.68,…
## $ log_ave   <dbl> 1.160, 0.778, 1.430, 1.360, 0.778, 1.600, 1.030, 1.160, 1.4…
## $ log_outs  <dbl> 1.040, 0.778, 1.410, 1.320, 1.230, 1.610, 1.000, 1.490, 1.7…
## $ log_fours <dbl> 1.280, 0.301, 1.830, 1.830, 1.040, 2.160, 1.110, 1.650, 2.0…
## $ log_sixes <dbl> 0.000, 0.000, 0.477, 0.845, 0.301, 0.602, 0.000, 0.477, 0.6…
## $ log_ducks <dbl> 0.699, 0.477, 0.602, 0.602, 1.040, 0.778, 0.602, 0.845, 0.6…
## $ log_hs    <dbl> 2.07, 1.26, 2.02, 2.00, 1.41, 2.10, 1.48, 1.52, 2.10, 1.95,…
## $ log_100   <dbl> 0.301, 0.000, 0.301, 0.301, 0.000, 0.699, 0.000, 0.000, 0.4…
```

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)
```

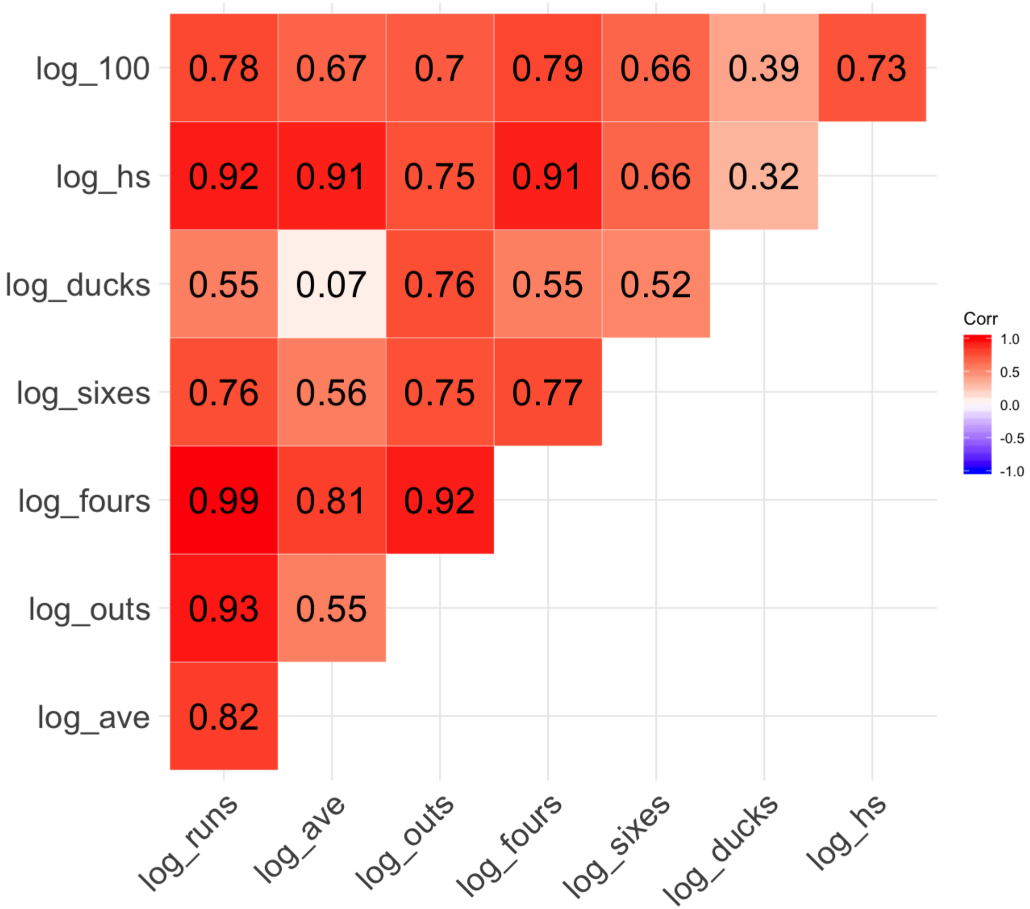
```
## # A tibble: 8 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -0.365      0.0902   -4.05   5.67e- 5
## 2 log_runs    -1.92       1.95     -0.984  3.25e- 1
## 3 log_ave      1.84       1.96      0.943  3.46e- 1
## 4 log_outs     1.61       1.96      0.826  4.09e- 1
## 5 log_fours    0.647      0.0969    6.68   4.58e-11
## 6 log_sixes    0.131      0.0264    4.96   8.57e- 7
## 7 log_ducks    0.00357    0.0497    0.0718 9.43e- 1
## 8 log_hs      -0.0187     0.0753   -0.248  8.04e- 1
```

Consequence of multi-collinearity

- We will proceed with rounding all variables to 3 significant figures.

	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**	-0.36	0.09	**<0.001**
log_runs	-1.92	1.95	0.325				-0.08	0.12	0.491
log_ave	1.84	1.96	0.346	-0.08	0.12	0.530			
log_outs	1.61	1.96	0.409	-0.31	0.11	**0.004**	-0.23	0.10	**0.019**
log_fours	0.65	0.10	**<0.001**	0.64	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	0.00	0.05	0.943	0.00	0.05	0.922	0.00	0.05	0.934
log_hs	-0.02	0.08	0.804	-0.02	0.08	0.811	-0.02	0.08	0.837

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 methods

1. Variance inflation factors (VIFs)

Introduced in Marquaridt (1970) and elsewhere:

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

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 as suggested 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() %>% scale()
eigen = svd(t(Xmat) %*% Xmat)
round(eigen$d, 3)
```

```
## [1] 4839.921  928.325  303.818  252.626   91.953   45.354    9.982    0.020
```

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:

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

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

Larger τ_p value indicates larger 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.

```
solve(eigen$u * eigen$u)[1:2,1:5]
```

```
##                [,1]         [,2]         [,3]          [,4]         [,5]
## [1,] -3.761500e+14 1.070173e+15 3.546325e+14 -7.096193e+13 1.581491e+14
## [2,] -1.113496e+13 3.167974e+13 1.049799e+13 -2.100647e+12 4.681601e+12
```

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!

$VIF_1, \ldots, VIF_p$

$\tau_1, \ldots, \tau_p$

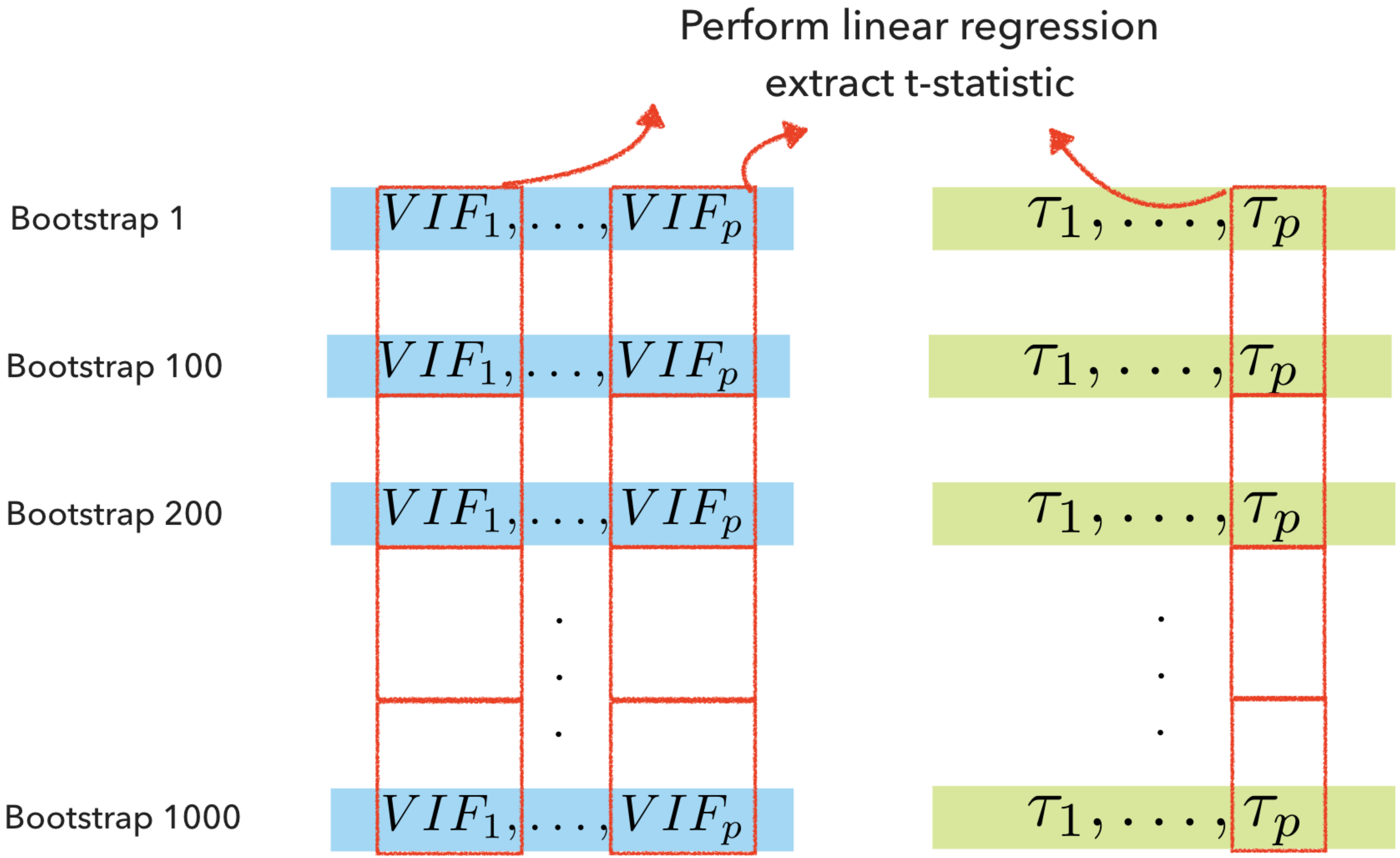
Bootstrap 1	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
Bootstrap 100	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
Bootstrap 200	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
	$\vdots$	$\vdots$
Bootstrap 1000	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$

Perform linear regression
extract t-statistic

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

Perform linear regression
extract t-statistic

Bootstrap 1	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
Bootstrap 100	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
Bootstrap 200	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$
	$\vdots$	$\vdots$
Bootstrap 1000	$VIF_1, \ldots, VIF_p$	$\mathcal{T}_1, \ldots, \mathcal{T}_p$



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
$VIF_1, \ldots, VIF_p$
$\tau_1, \ldots, \tau_p$
Bootstrap 1000
$VIF_1, \ldots, VIF_p$
$\tau_1, \ldots, \tau_p$

Bootstrap 1

Bootstrap 100

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

Bootstrap 200

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

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

Bootstrap 1000

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

Bootstrap 1

Bootstrap 100

$K = 1$ $\quad t_{1,1}, \dots, t_{p,1}$

Bootstrap 200

$K = 2$ $\quad t_{1,2}, \dots, t_{p,2}$

$K = 10$ $\quad t_{1,10}, \dots, t_{p,10}$

Bootstrap 1000

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

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

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

The `mcvis` package

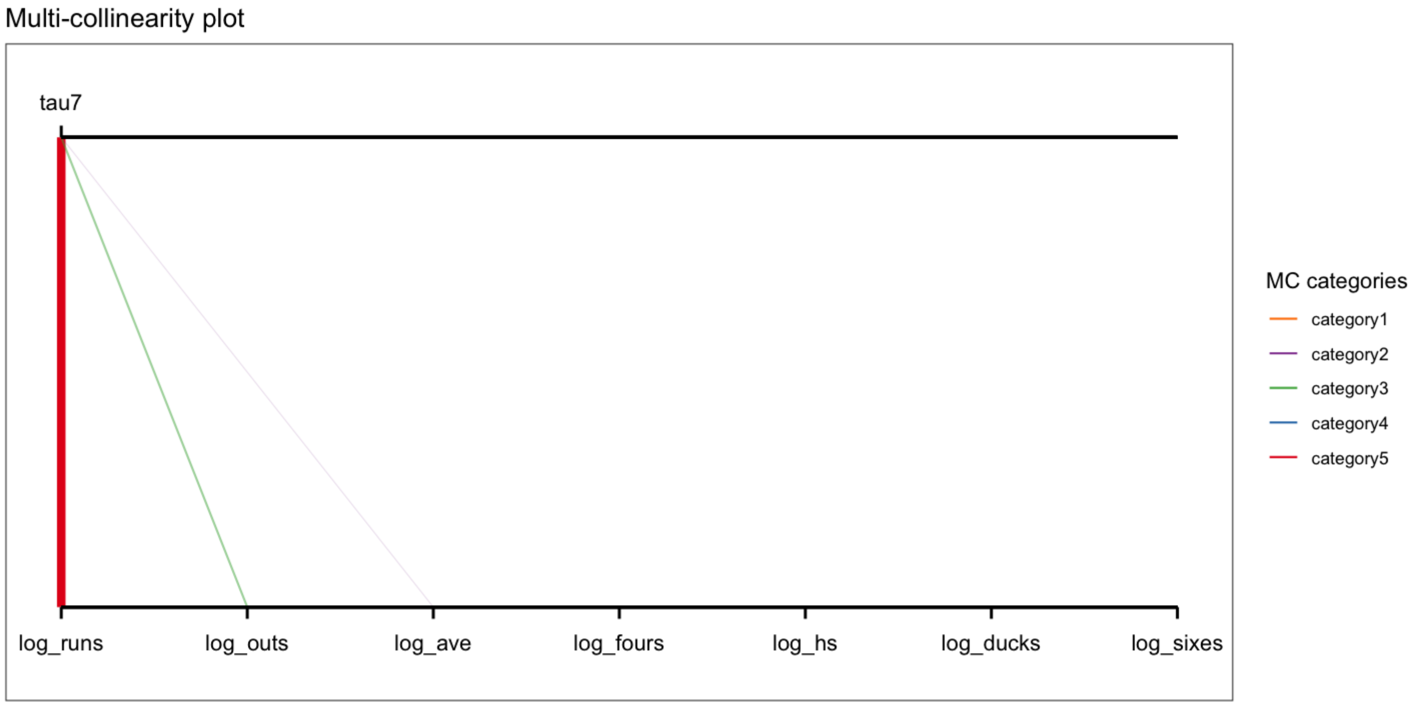
1. MC-index

```
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

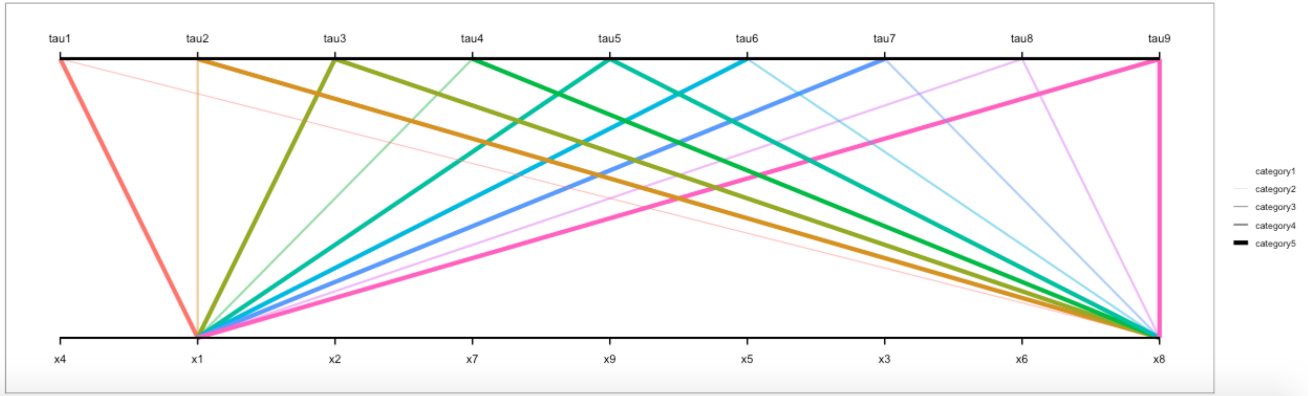
Show 10 entries

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

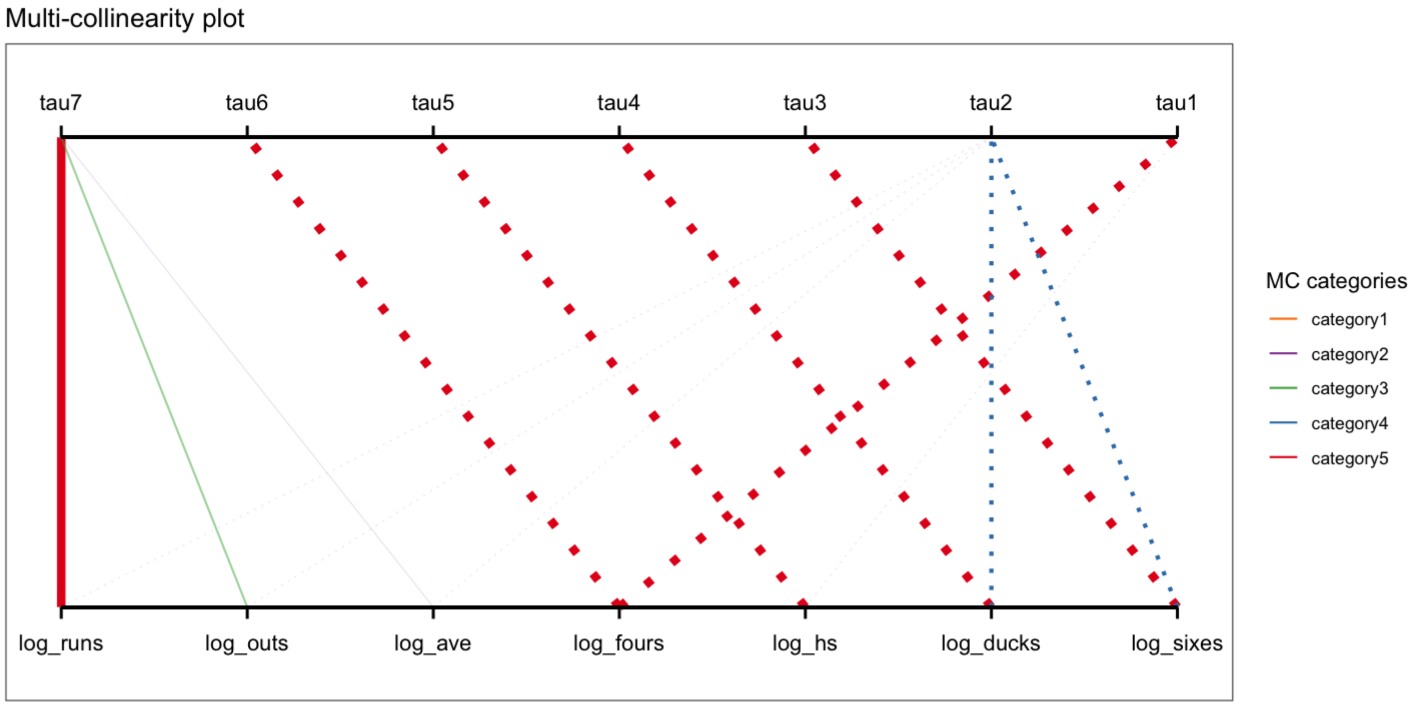
Showing 1 to 9 of 9 entries

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

Marquaridt, D. W. (1970). “Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation”. In: Technometrics 12.3, pp. 591–612.

Sheather, S. (2009). A Modern Approach to Regression with R. Springer Texts in Statistics. New York, NY: Springer New York.

Belsley, D. A. (1984). “Demeaning Conditioning Diagnostics through Centering”. In: The American Statistician 38.2, pp. 73–77.

Friendly, M. and E. Kwan (2009). “Statistical computing and graphics where's Waldo? Visualizing collinearity diagnostics”. In: The American Statistician 63.1, pp. 56–65.

O'Brien, R. M. (2007). “A Caution Regarding Rules of Thumb for Variance Inflation Factors”. In: Quality & Quantity 41.5, pp. 673–690.

Stewart, G. W. (1987). “Collinearity and Least Squares Regression”. In: Statistical Science 2.1, pp. 68–84.