mcvis: Multi-collinearity Visualisation

Kevin Wang

Initiated on 2020 Aug 22, compiled on 2020 Sep 23

Table of Contents

# Introduction

What is the most popular statistical model for a statistician? We think most statisticians would say the linear regression model. Indeed, linear regression is both rich in theory and powerful in practice with many flexible extensions such as generalised linear models and linear mixed models.

But what about the model limitations? Multicollinearity is perhaps the most obvious limitation of regression models, which can cause the least squares regression to produce unreasonable coefficient estimates and standard errors.

# Multicollinearity high correlation!

So what is multicollinearity? Formally, it is defined as the phenomenon when a group of predictor variables are exactly or approximately linearly dependent. In other words, if one predictor variable can be approximately expressed as a linear combination of other predictor variables, then we have a situation of multicollinearity.

**It is a common misconception that multicollinearity equals to high correlation between variables!** Consider the following example where we generate normally distributed predictor variables. The first predictor is constructed as plus some random noise. Looking at the correlation matrix, we don’t necessarily see anything alarming with all correlation coefficient being below 0.6. This might give us the false sense of security that no multicollinearity exist when in fact, the correlation of and the sum is approximately 0.998, enough to trigger multicollinearity for regression!

set.seed(123)  
p = 5; n = 50  
X = matrix(rnorm(n\*p), ncol = p)  
X[,1] = X[,2] + X[,3] + X[,4] + rnorm(n, 0, 0.1)  
round(cor(X), 2)

## [,1] [,2] [,3] [,4] [,5]  
## [1,] 1.00 0.42 0.57 0.55 0.02  
## [2,] 0.42 1.00 -0.16 -0.16 -0.13  
## [3,] 0.57 -0.16 1.00 -0.01 -0.04  
## [4,] 0.55 -0.16 -0.01 1.00 0.20  
## [5,] 0.02 -0.13 -0.04 0.20 1.00

cor(X[,1], X[,2] + X[,3] + X[,4])

## [1] 0.9978913

If we blindly fit a linear regression model not knowing of the existence of multicollinearity, then two things can happen. One is with the coefficient estimate itself, where we can obtain an estimate of that is opposite in sign to how we generated our response variable. Another consequence is on the standard error of the coefficient estimate. Looking at the summary output of the “full model” (with all predictors fitted), we see that the standard error of is ten times larger than that of the “true model” with only fitted.

y = 1 \* X[,1] + rnorm(n)  
full\_model = lm(y ~ X)  
broom::tidy(full\_model)

## # A tibble: 6 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.0148 0.162 0.0911 0.928  
## 2 X1 -0.318 1.67 -0.191 0.850  
## 3 X2 1.41 1.66 0.847 0.401  
## 4 X3 1.28 1.64 0.779 0.440  
## 5 X4 1.27 1.69 0.752 0.456  
## 6 X5 0.0618 0.167 0.370 0.713

true\_model = lm(y ~ X[,1])  
broom::tidy(true\_model)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.00192 0.147 0.0131 9.90e- 1  
## 2 X[, 1] 0.997 0.103 9.66 7.66e-13

# Classic ways of diagnosing multicollinearity

So why do these strange phenomena occur? In least squares regression, both the estimator and its variance depend on the precision matrix . However, when the model design matrix, , is numerically “close” to having linearly dependent columns, as we do have when multicollinearity occurs, the matrix is not numerically invertible. Hence, elements of the precision matrix, regression estimators and standard errors start to behave erratically from a numeric perspective.

One way to “diagnose” multicollinearity is to compute the eigenvalues of the matrix and examine the value , where and are the maximum eigenvalue and the minimum eigenvalue, respectively. This ratio is known as the “condition number” of the design matrix and a rule of thumb is that if a condition number is larger than 30, then this indicates a severe case of multicollinearity (see e.g. Belsley et al. (1980, Section 3.2)). In our data example, we do indeed have such a severe case of multicollinearity.

evalues = eigen(t(X) %\*% X)$values  
(condition\_number = sqrt(evalues[1]/evalues[p]))

## [1] 35.51142

However, if we didn’t simulate the data above and we obtained such a large condition number, how would we know which variables are the culprits, i.e. variables that contribute the most to causing multicollinearity? After all, the condition number only tells us of the existence of multicollinearity, but not which predictors are causing this. Knowing this information may help us to decide if any variables should be dropped or alternative modelling strategies might be needed (e.g. averaging these variables).

This is where mcvis can help!

# Introducting mcvis

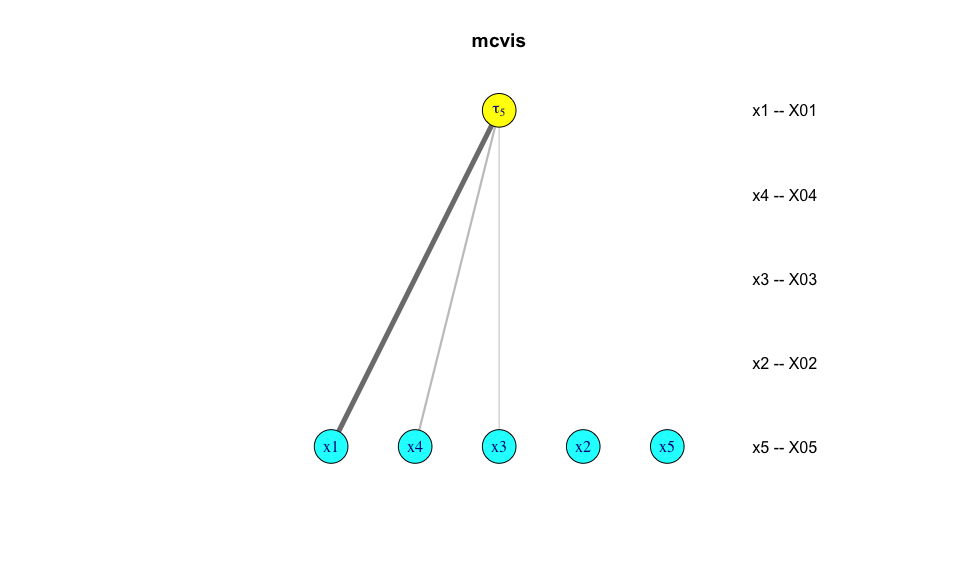
The recently developed mcvis package computes a new statistic called the MC-index which can diagnose multicollinearity. The theory behind this MC-index is published in Lin et. al. (2020). Briefly, MC-index is a vector of length matching that of the number of columns of . Each element of the MC-index vector is a number between zero and one, **connecting the number** [what does connecting the number mean? and tau\_p = lambda\_min?] , which measures ’s proximity to linear dependence, with a regression predictor variable. A larger MC-index value indicates that a variable is causing more multicollinearity than another variable. The computation of the MC-index uses a bootstrap resampling scheme to stabilise estimations. The process is simple to implement through the mcvis function in the mcvis R package which is available on CRAN.

# install.packages("mcvis")  
library(mcvis)  
mcvis\_result = mcvis(X)  
print(mcvis\_result)

## X01 X02 X03 X04 X05  
## tau1 0.59 0.23 0.10 0.06 0.02  
## tau2 0.10 0.14 0.03 0.01 0.72  
## tau3 0.55 0.15 0.25 0.02 0.02  
## tau4 0.06 0.02 0.01 0.04 0.87  
## tau5 0.51 0.12 0.15 0.22 0.00

In the output above, we see that , the first column of , is identified as the main variable causing multicollinearity. [Why? what is it about the table that indicates this? there are larger values in the X05 column for example. Or should we only look in the tau5 row?] We can further visualise this via a (bipartite) graph that shows this connection between and the predictor variables. The size and colour of the lines are categorised by the magnitude of the MC-index value. This plot gives us a quick indication of the sources of the multicollinearity, ordered by the magnitude. [The thickness of the line represents magnitude?]

plot(mcvis\_result)



# Breast cancer

We will illustrate the use of the mcvis package using a real example from the dslabs data package. The brca data contains various important biopsy features for breast cancer cell nuclei. The original data measures features such as the radius for a collection of cell nuclei in a tumour and summarises these into mean, standard error and the worst value. The response variable is a factor with two levels denoting whether a tumour is malignant (“M”) or benign (“B”). To simplify interpretations, we will only focus on the first ten predictors, which correspond to the mean measurement of cell nuclei.

We do expect multicollinearity in this data because measures such as radius, perimeter and area of nucle[i?] should be highly correlated. Performing mcvis on this data, we do indeed see that this is the case and we will demonstrate that more than one group of variables are linearly highly correlated. We add some extra plotting parameters for mcvis to improve visibility.

library(dslabs)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

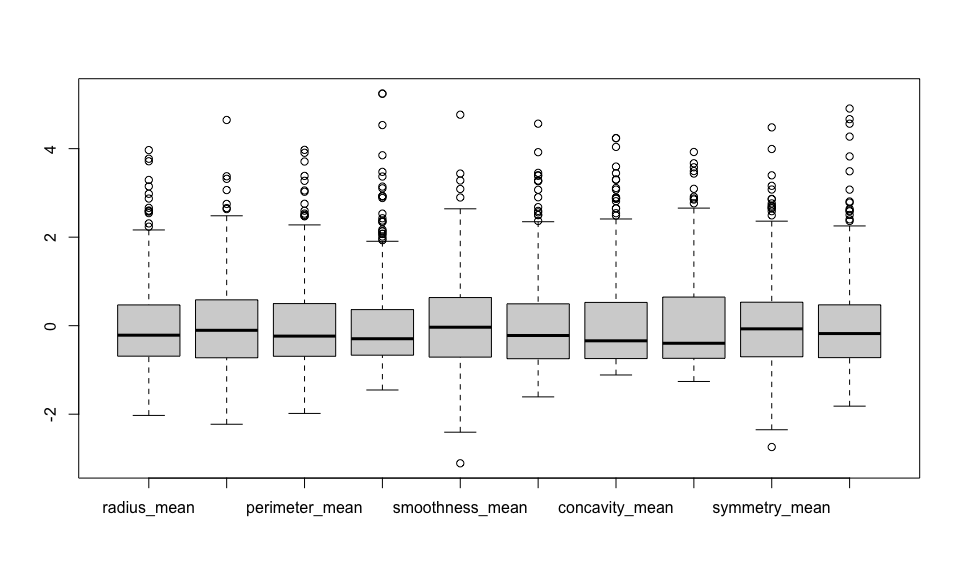
library(ggcorrplot)

## Loading required package: ggplot2

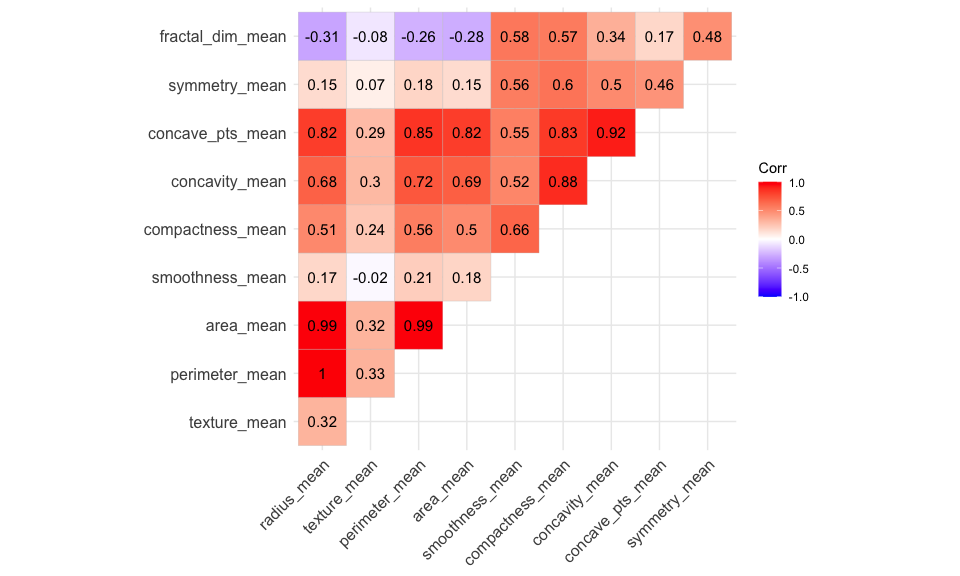
data(brca)  
x = as.data.frame(brca$x[,1:10])  
colnames(x)

## [1] "radius\_mean" "texture\_mean" "perimeter\_mean" "area\_mean"   
## [5] "smoothness\_mean" "compactness\_mean" "concavity\_mean" "concave\_pts\_mean"  
## [9] "symmetry\_mean" "fractal\_dim\_mean"

boxplot(scale(x))



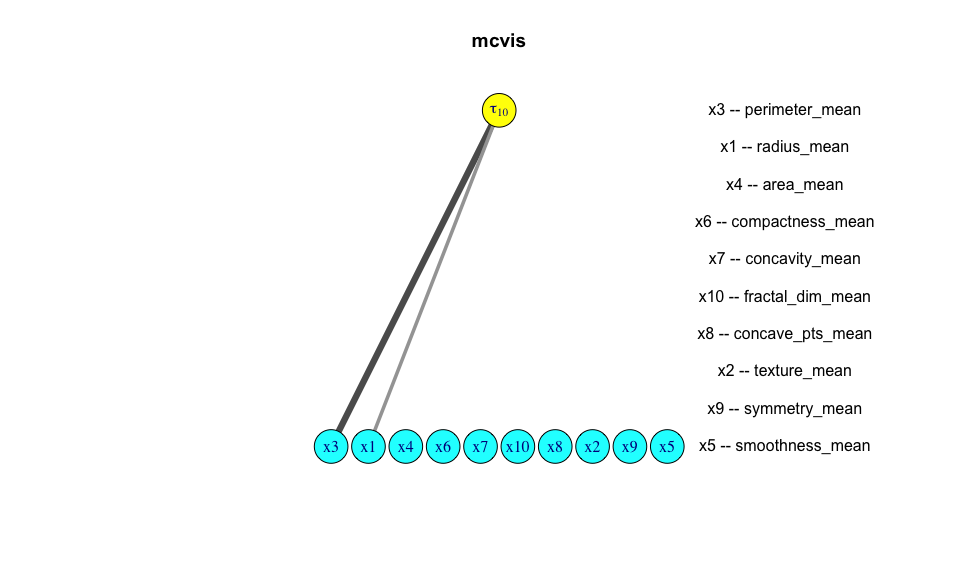
ggcorrplot(cor(x), type = "upper", lab = TRUE)



(mcvis\_result = mcvis::mcvis(x))

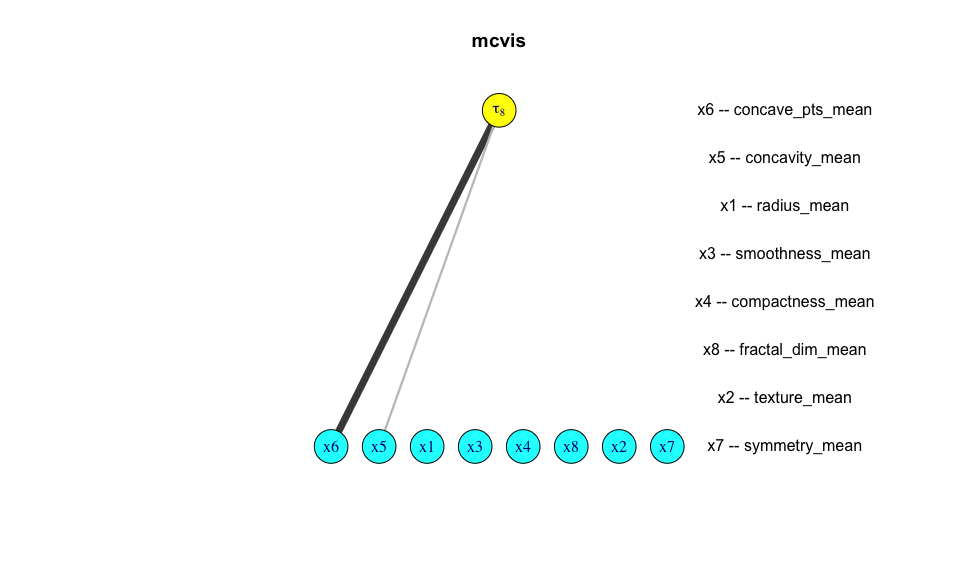
## radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean  
## tau1 0.08 0.05 0.07 0.04 0.01  
## tau2 0.05 0.02 0.04 0.03 0.07  
## tau3 0.09 0.68 0.08 0.08 0.01  
## tau4 0.00 0.00 0.00 0.01 0.00  
## tau5 0.02 0.03 0.02 0.02 0.74  
## tau6 0.18 0.03 0.18 0.01 0.07  
## tau7 0.05 0.01 0.07 0.01 0.01  
## tau8 0.01 0.00 0.01 0.00 0.00  
## tau9 0.05 0.00 0.06 0.87 0.00  
## tau10 0.36 0.00 0.64 0.00 0.00  
## compactness\_mean concavity\_mean concave\_pts\_mean symmetry\_mean  
## tau1 0.09 0.09 0.11 0.09  
## tau2 0.11 0.03 0.08 0.01  
## tau3 0.01 0.01 0.01 0.02  
## tau4 0.00 0.01 0.01 0.96  
## tau5 0.01 0.12 0.01 0.02  
## tau6 0.06 0.10 0.07 0.03  
## tau7 0.16 0.64 0.01 0.01  
## tau8 0.00 0.14 0.82 0.00  
## tau9 0.00 0.01 0.00 0.00  
## tau10 0.00 0.00 0.00 0.00  
## fractal\_dim\_mean  
## tau1 0.38  
## tau2 0.55  
## tau3 0.01  
## tau4 0.00  
## tau5 0.01  
## tau6 0.28  
## tau7 0.02  
## tau8 0.00  
## tau9 0.01  
## tau10 0.00

plot(mcvis\_result, var\_max = 5, label\_dodge = TRUE)



The MC-index identifies two variables as having a “Strong” MC-index value. Therefore, in this case we might be tempted to remove the perimeter\_mean and area\_mean from our design matrix as they provide very similar information as radius\_mean. Once the strongest sources of multicollinearity is addressed, additional groups of variables causing multicollinearity may be identified. We can repeat mcvis to identify concavity\_mean and concave\_pts\_mean as highly correlated variables.

x2 = dplyr::select(x, -perimeter\_mean, -area\_mean)  
mcvis\_result2 = mcvis::mcvis(x2)  
plot(mcvis\_result2, var\_max = 5, label\_dodge = TRUE)



mcvis\_result2$MC[,c(5,6)] %>% round(2)

## concavity\_mean concave\_pts\_mean  
## tau1 0.05 0.05  
## tau2 0.09 0.08  
## tau3 0.04 0.01  
## tau4 0.00 0.00  
## tau5 0.01 0.02  
## tau6 0.68 0.06  
## tau7 0.02 0.01  
## tau8 0.23 0.72

Repeating mcvis again, we can also find that concave\_pts\_mean and concavity\_mean as another potential source of multicollinearity and as their definition are very similar according to the documentation, we may decide to remove one of these.

Of course, depending on data context, we might not want to simply remove the collinear variables like we have done here. Some common alternatives include averaging collinear predictors or using models that can decorrelate the predictor variables (e.g. random forest). Nonetheless, the diagnosis of multicollinearity is an important part of choosing which modelling strategy one should take.

# Conclusion

The linear regression model is arguably the most powerful tool in statistics. However, using it in practice may require extra checks and practical considerations with one often ignored issue being multicollinearity. mcvis is a tool that can help statisticians to identify different sources of multicollinearity to better decision making and the application of linear regression model.

# Reference

* Belsley, D. A., Kuh, E. & Welsch. R. E. Regression Diagnostics. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1980.
* Lin, C., Wang, K. & Mueller, S. mcvis: A new framework for collinearity discovery, diagnostic and visualization. Journal of Computational and Graphical Statistics In Press, (2020). DOI: 10.1080/10618600.2020.1779729