MIE237

Neil Montgomery
2016-03-29

Multicollinearity

We have seen (in the polynomial regression example) seemingly strange behaviour relating to p-values when new terms are added to a model.

The cause is "multicollinearity" - the existence of strong linear relationships among input variables.

Most regression datasets exhibit linear relationships among inputs to some extent. It is a MYTH that input variables must be "independent", either probabilistically or linear-algebraically.

In a nutshell: a strong enough linear relationship can make $\boldsymbol{X}'\boldsymbol{X}$ close to "singular" (determinant close to 0), which in turn inflates the variances of the $\hat{\beta}_i$, leading to model selection and interpretation challenges.

But this is a *numerical* problem and not a scientific problem.

The source of the problem

Recall:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y}$$

And:

$$Var(\hat{\beta}_i) = c_{ii}\sigma^2$$

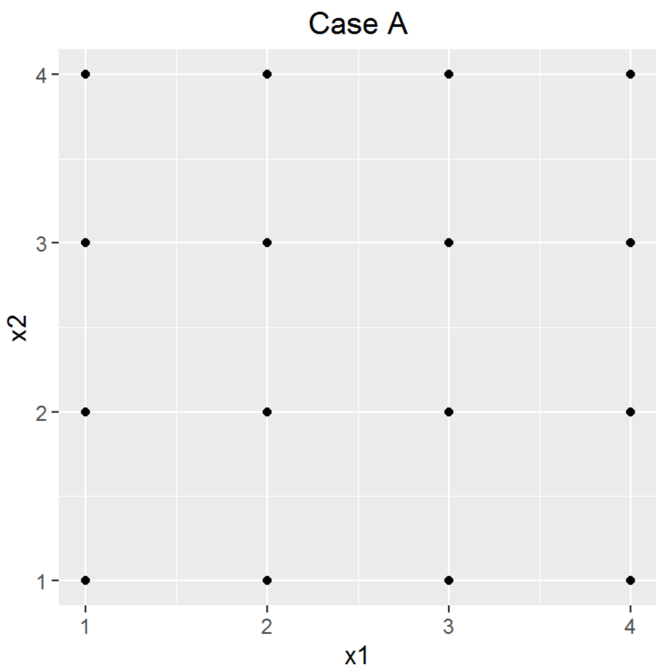
where c_{ii} is the ith diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$

Fact: the stronger the linear dependency among the columns of $\boldsymbol{X}$ are,the higher the c_{ii} for the $\hat{\beta}_i$ correspondingto those x_i involved in the dependency.

Usually the dependency is simply a matter of "correlation" among pairs of inputs, but complex multi-way dependencies are possible.

Illustration of the problem

Two cases:



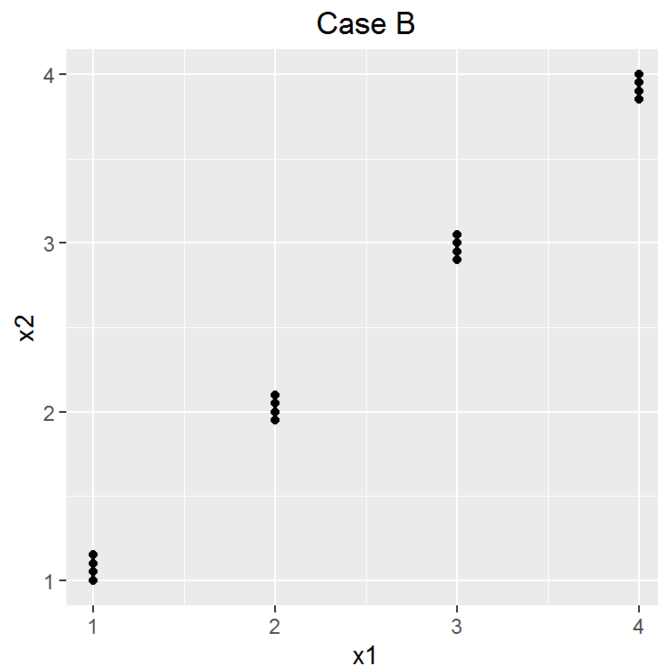


Illustration of the problem - the matrices

$$(\boldsymbol{X_A'X_A}) = \begin{bmatrix} 16 & 40 & 40 \\ 40 & 120 & 100 \\ 40 & 100 & 120 \end{bmatrix} \qquad (\boldsymbol{X_A'X_A})^{-1} = \begin{bmatrix} 0.69 & -0.13 & -0.13 \\ -0.13 & 0.05 & 0.00 \\ -0.13 & 0.00 & 0.05 \end{bmatrix}$$

$$(\boldsymbol{X_B'X_B}) = \begin{bmatrix} 16 & 40 & 40 \\ 40 & 120 & 119 \\ 40 & 119 & 118.1 \end{bmatrix} \qquad (\boldsymbol{X_B'X_B})^{-1} = \begin{bmatrix} 0.69 & 2.25 & -2.5 \\ 2.25 & 18.1 & -19 \\ -2.5 & -19 & 20 \end{bmatrix}$$

I'll generate some data from the same model in each case:

$$Y = 1 + 2x_1 + 3x_2 + \varepsilon, \qquad \varepsilon \sim N(0, 1)$$

Then fit the two datasets to regression models…

Case A

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = Case_A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1462 -0.7048 -0.1268  0.7506  1.8325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            1.9401     0.2744   7.069 8.43e-06 ***
## x2            2.8854     0.2744  10.513 1.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9251, Adjusted R-squared:  0.9135
## F-statistic: 80.25 on 2 and 13 DF,  p-value: 4.843e-08
```

Case B

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = Case_B)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1462 -0.7048 -0.1268  0.7506  1.8325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            4.1181     5.2218   0.789    0.444
## x2            0.7074     5.4890   0.129    0.899
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9591, Adjusted R-squared:  0.9528
## F-statistic: 152.3 on 2 and 13 DF,  p-value: 9.506e-10
```

Note the small p-value for the overall F test.

Note that multicollinearity is merely a problem

Case C: same model fit to the Case B situation but with $n = 288$

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = Case_C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31324 -0.65271 -0.04773  0.64939  2.77405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0510     0.1888   5.565 6.03e-08 ***
## x1            2.1419     0.9690   2.210  0.02787 *
## x2            2.8299     1.0186   2.778  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9663 on 285 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9691
## F-statistic:  4502 on 2 and 285 DF,  p-value: < 2.2e-16
```