1

STAT2008/STAT6038

Multicollinearity

Multicollinearity

2

Multicollinearity relates to the degree of interrelation among the predictor values.

A first diagnostic test for whether there is multicollinearity present in the data, we calculate the correlation matrix of the predictors, $R = (r_{ij})$, where

$$r_{ij} = Corr(X_i, X_j) = \frac{\sum_{k=1}^{n}(X_{ki} - \bar{x}_i)(X_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(X_{kj} - \bar{x}_j)^2}},$$

where X_i is the i^{th} column of the design matrix and $\bar{x}_i$ is just the average of all the values of the i^{th} predictor.

Multicollinearity

3

If any of the off-diagonal r_{ij}'s (i.e., those for which $i \neq j$) are large, then we know that the associated two predictor variables are highly correlated and this results in multicollinearity.

The problem may also arise because several of the predictors are strongly linearly related to each other. Generally speaking, a set of predictor values are said to be multicollinear if there exist any constants $c_1, \ldots, c_p$ (not all equal to zero) such that

$$\sum_{j=1}^{p} c_j X_j \approx 0.$$

Multicollinearity

4

The interpretation of the parameters in a multiple regression is a rate of change of the response with respect to a particular predictor while the rest of the predictors are held constant

If we have predictor values which are highly interrelated, then it is not possible to "hold the other predictors constant" within our dataset

Centred Regression Model

5

A multiple regression model:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

can be re-written as

$$Y = \beta_0^* + \beta_1(x_1 - \bar{x}_1) + \ldots + \beta_k(x_k - \bar{x}_k) + \epsilon = \beta_0^* + \beta_1 x_1^* + \ldots + \beta_k x_k^* = \beta_0^* + X^*\beta + \epsilon$$

where the intercept parameter is suitably transformed to $\beta_0^* = \beta_0 + \beta_1\bar{x}_1 + \ldots + \beta_k\bar{x}_k$ and X^* is the centered design matrix, without an initial column of ones for the intercept term.

Centred Regression Model

6

With this restructuring:

$$R = (X^*)^T X^*.$$

So, to estimate the β vector for this regression, we need to invert the matrix R, since $b = R^{-1}(X^*)^T Y$.

Suppose the correlation between two of the predictors is exactly one.

In such a case, the R matrix will not be invertible!

In other words, we cannot estimate the parameters at all.

Partial Regression Coefficients and Multicollinearity

7

Once the first of the two perfectly correlated predictors is put into the model, the least-squares fit will calculate an estimate of the associated parameter value, and then when the second predictor is incorporated, it adds nothing new to the model

Therefore should be assigned an estimated parameter value of zero.

If the entry order of the two predictors were reversed, then the first predictor would now be the one with the parameter estimate of zero.

Least-squares gets confused by completely collinear predictors!!!

"Nearly" Collinear

8

Even if the predictors are not perfectly correlated if the predictors are *nearly* collinear (e.g., if there are any off-diagonal elements of the R matrix which are very near unity), then the parameter estimates will be highly unstable and imprecise

Variance Inflation Factors

9

Suppose that we are going to conduct a study of the relationship of a response variable, Y, and two predictors, x_1 and x_2

Suppose also that we know that the assumptions of the linear model hold, so that $Y = X\beta + \epsilon$ and $Var(\epsilon) = \sigma^2 I$.

We are asked to choose between two possible sets of predictor values:

Predictor set 1:

x_1	10	10	10	10	15	15	15	15
x_2	10	10	15	15	10	10	15	15

Predictor set 2:

x_1	10.0	11.0	11.9	12.7	13.3	14.2	14.7	15.0
x_2	10.0	11.4	12.2	12.5	13.2	13.9	14.4	15.0

Example

10

If we were to estimate the parameters from a regression using each of these datasets, we would find that for the first set of predictor values, $r_{12} = 0$ so that:

$$(X^*)^T X^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \Longrightarrow \quad \{(X^*)^T X^*\}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

while for the second predictor set, $r_{12} = 0.99215$ so that:

$$(X^*)^T X^* = \begin{pmatrix} 1 & 0.99215 \\ 0.99215 & 1 \end{pmatrix} \quad \Longrightarrow \quad \{(X^*)^T X^*\}^{-1} = \begin{pmatrix} 63.94 & -63.44 \\ -63.44 & 63.94 \end{pmatrix}$$

Variance of the estimates

11

The least-squares parameter estimates we get from the first set of predictors will have variances:

$$Var(b_1) = Var(b_2) = \sigma^2,$$

While the parameter estimates from the second set of predictor values will have variances:

$$Var(b_1) = Var(b_2) = 63.94\sigma^2.$$

Note that this inflation of the variance has nothing to do with the response variable, or with the intrinsic variability of the errors, ϵ, in the population. It is purely an artifact of the correlation structure of the predictors.

Variance Inflation Factor

12

The value 63.94 in the previous example is the *variance inflation factor*, $VIF_1 = VIF_2$, since it shows the proportion by which the variance has been increased over a set of *orthogonal* predictor values (i.e., predictor values for which $R = I$).

In general, the variance inflation factor associated with the i^{th} predictor is:

$$VIF_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination for a regression of the predictor x_i as the response variable on the rest of the predictors, $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$

VIF_i

13

The stronger the linear relationship between x_i and the rest of the predictors, the larger the value of R_i^2 and thus the larger the value of VIF_i.

So the parameter estimates associated with predictor variables with large VIF_i's will have very low precision

Any quantities which rely heavily on these parameter estimates (e.g., various predicted values and general linear hypotheses) may also tend to be highly imprecise.

Why?

14

Suppose that the true underlying model relating a response to two predictors is $Y = 2x_1 + 3x_2 + \epsilon$

Suppose $x_1 = x_2$ for all data points. For such a set of predictors, we note that

$$\begin{aligned} Y &= 2x_1 + 3x_2 + \epsilon \\ &= 2x_1 + 3x_1 + \epsilon \\ &= 5x_1 + 0x_2 + \epsilon \\ &= 2x_2 + 3x_2 + \epsilon \\ &= 0x_1 + 5x_2 + \epsilon \\ &= 7x_1 - 2x_2 + \epsilon \\ &\vdots \end{aligned}$$

We must be extremely careful of our interpretation of parameter estimate values in the presence of multicollinearity.

Context and theory

15

Often the case that the problem can be solved by simply removing one or a few of the predictors.

The removal should ideally be based on both statistical and non-statistical grounds.

A good place to start is to remove the variable with the highest VIF_i and see if this solves the problem or else the predictor with the weakest simple regression with the response.

External considerations should be taken into account, as these variables may turn out to be extremely relevant in the context in which the data was gathered. (Correlation doesn't imply causation)

Example: Body Fat Data

16

```
> cor(cbind(Triceps,Thigh,Midarm))
          Triceps     Thigh    Midarm
Triceps 1.0000000 0.9238425 0.4577772
Thigh   0.9238425 1.0000000 0.0846675
Midarm  0.4577772 0.0846675 1.0000000


> diag(solve(cor(cbind(Triceps,Thigh,Midarm))))
 Triceps    Thigh   Midarm
708.8429 564.3434 104.6060

> R2i <- summary(lm(Triceps ~ Thigh + Midarm))$r.squared
> 1/(1-R2i)
[1] 708.8429
```

Standard errors large (as proportion of estimates)

17

```
> lsfit(cbind(Triceps,Thigh,Midarm),BodyFat)$coef
 Intercept    Triceps      Thigh     Midarm
117.084695   4.334092  -2.856848  -2.186060
> ls.diag(lsfit(cbind(Triceps,Thigh,Midarm),BodyFat))$std.err
                [,1]
Intercept 99.782403
Triceps    3.015511
Thigh      2.582015
Midarm     1.595499


> anova(lm(BodyFat ~ Triceps + Thigh + Midarm))
Analysis of Variance Table
Response: BodyFat
          Df Sum Sq Mean Sq F value    Pr(>F)
Triceps    1 352.27  352.27 57.2768 1.131e-06 ***
Thigh      1  33.17   33.17  5.3931   0.03373 *
Midarm     1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Removing the predictor with the largest *VIFi* and then refitting the model yields

18

```
> lsfit(cbind(Thigh,Midarm),BodyFat)$coef
   Intercept        Thigh       Midarm
-25.99695164   0.85088172   0.09602947
> ls.diag(lsfit(cbind(Thigh,Midarm),BodyFat))$std.err
              [,1]
Intercept 6.9973208
Thigh     0.1124482
Midarm    0.1613927
> anova(lm(BodyFat ~ Thigh + Midarm))
Analysis of Variance Table
Response: BodyFat
          Df Sum Sq Mean Sq F value    Pr(>F)
Thigh      1 381.97  381.97  58.441 6.737e-07 ***
Midarm     1   2.31    2.31   0.354    0.5597
Residuals 17 111.11    6.54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> diag(solve(cor(cbind(Thigh,Midarm))))
  Thigh  Midarm
1.00722 1.00722
```

Trial and error

19

We have solved the multicollinearity problem, however, the last variable in the model (Midarm) is still not significant

This predictor is also the one with the weakest simple regression relationship

But dropping it from the model and retaining the other two predictors will not solve multicollinearity problem.

We could try dropping the thigh circumference variable from the original three variable model, (since the largest correlation in the R matrix was between thigh circumference and triceps skinfold)

What does the science suggest?

20

```
> lsfit(cbind(Triceps,Midarm),BodyFat)$coef
Intercept   Triceps    Midarm
 6.791627  1.000585 -0.431442
> ls.diag(lsfit(cbind(Triceps,Midarm),BodyFat))$std.err
             [,1]
Intercept 4.4882871
Triceps   0.1282321
Midarm    0.1766156
> diag(solve(cor(cbind(Triceps,Midarm))))
 Triceps   Midarm
1.265118 1.265118
```

Multicollinearity has been fixed, and now both the predictors are significant.

But we now have a model which uses only measurements on the arm of the women