big k, small n

$(X'X)$ $(k+1)*(k+1)$

9.3 Multicollinearity

Multicollinearity in regression refers to the case when one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

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

covariates $(1, x_1, \dots, x_k)$

x_i

y

Detection

1. Significant correlations between pairs of independent variables in the model
2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy $H_0{:}\, \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is significant
3. Opposite signs (from what is expected) in the estimated parameters
4. A variance inflation factor (VIF) for a β parameter greater than 10, where

$$(VIF)_i = \frac{1}{1 - R_i^2} \qquad i = 1,2,\dots,k$$

$R_i^2 \geq 0.9$

$\Leftrightarrow (VIF)_i \geq 10$

and R_i^2 is the multiple coefficient of determination for the model

$$E(x_i) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{i-1}x_{i-1} + \alpha_{i+1}x_{i+1} + \cdots + \alpha_k x_k$$

R_i^2

5. Let the ordered eigenvalues of $X'X$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ where $p = k + 1$.

some eigenvalues $= 0$

if $\mathrm{Rank}(X) \leq k+1$

$\Rightarrow \kappa \to \infty$

Condition number: $\kappa = \sqrt{\dfrac{\lambda_1}{\lambda_p}}$

Condition indexes: $\sqrt{\dfrac{\lambda_1}{\lambda_i}}$ where $i = 2,\dots,p$

$\sqrt{\lambda_1/\lambda_i} \to \infty$ for some i

Multicollinear problem: Condition number or indexes ≥ 30

(note in some references, condition index is also called condition number)

Motivation example

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X1 denotes expenditures for point-of-sales displays in beauty salons and department stores (in thousand dollars), and X2 and X3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Y denotes sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X1 is increased by 1 thousand dollars and X2 and X3 are held constant, and was told to use an ordinary multiple regression model with linear terms for the predictor variables.

a) *State the regression model to be employed and fit it to the data.*

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

Computer output

Model Summary[b]

Model	R	R Square	Adjusted R Square
1	.861[a]	.742	.722

a. Predictors: (Constant), national_advertising, Exp_point_of_sales_display, local_advertising

b. Dependent Variable: Sales

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$; $H_1: \beta_j \neq 0$ for at least one of $j = 1, 2, 3$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	382.659	3	127.553	38.279	.000[a]
	Residual	133.286	40	3.332		
	Total	515.945	43			

H_0: $F_{3,40}$

p-value

Reject H_0

		coefficients	Std. Error	t	sig	VIF
1	(Constant)	1.023	1.203	.851	.400	
	Exp_point_of_sales_display	.966	.709	1.362	.181	20.072
	local_advertising	.629	.778	.808	.424	20.716
	national_advertising	.676	.356	1.900	.065	1.218

β_0 β_1 β_2 β_3

p-value for $H_0: \beta_j = 0$, $H_1: \beta_j \neq 0$

Refer to the above output, the regression equation is. $\hat{Y} = 1.023 + .966X_1 + .629X_2 + .676X_3$

b) *Test whether there is a regression relation between sales and the three predictor variables (use α = 0.05). State the alternatives, decision rule, and conclusion.*

$E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ H_a: not all $\beta_k = 0$ (k = 1, 2, 3).

Test Statistics: F = 38.279 and *p*-value < 0.001.

Therefore, conclude H_a. The model is useful for the prediction of sales.

$\beta_i = 0$?

c) *Test for each of ~~the regression coefficients~~ (equal to zero?) individually (use α = 0.05 each time). Do the conclusions of these tests correspond to that obtained in part (b)?*

Refer to part (a) output, all regression coefficients are not significant at the given α level. Therefore, these tests do not yield the same conclusion as in part (b). This is a consequence of the multi-collinearity problem.

d) *Obtain the correlation matrix of the X variables and comment on the suitability of the data for the research objective.*

From the correlation matrix below, we observe that the independent variables (X1 and X2) are highly correlated and the regression model is therefore not quite appropriate.

matrix

Correlations

		Sales	Exp point-of-Sales display	Local advertising	National advertising
y Sales	Pearson Correlation	1	.842**	.842**	.474**
	Sig. (2-tailed)		.000	.000	.001
x_1 Exp point-of-sales display	Pearson Correlation	.842**	1	.974**	.376*
	Sig. (2-tailed)	.000		.000	.012
x_2 Local advertising	Pearson Correlation	.842**	.974**	1	.410**
	Sig. (2-tailed)	.000	.000		.006
x_3 National advertising	Pearson Correlation	.474**	.376*	.410**	1
	Sig. (2-tailed)	.001	.012	.006	

**. Correlation is significant at the 0.01 level (2-tailed), *. Correlation is significant at the 0.05 level (2-tailed).

e) *Obtain the three variance inflation factors. What do these suggest about the effects of multicollinearity here?*

$$(VIF)_1 = 20.072$$
$$(VIF)_2 = 20.716$$
$$(VIF)_3 = 1.218$$

The problem is quite serious since two of the VIF are much larger than 10.

f) The assistant eventually decided to drop variables X2 from the model to clear up the picture. Fit the assistant's revised model. Is the assistant now in a better position to achieve the research objective?

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	380.481	2	190.241	57.579	.000[a]
	Residual	135.464	41	3.304		
	Total	515.945	43			

a. Predictors: (Constant), national_advertising, Exp_point_of_sales_display

b. Dependent Variable: Sales

			Coefficient	Std. Error	t	sig	VIF
	2	(Constant)	1.017	1.198	.849	.401	
x_1		Exp_point_of_sales_display	1.522	.170	8.948	.000	1.165
X_3		national_advertising	.736	.346	2.125	.040	1.165

Refer to the above output, the regression equation is

$$\hat{Y} = 1.017 + 1.522X_1 + .736X_3$$

The VIF indicates that the problem of multicollinearity disappears (much less than 10).

Consequences of multicollinearity

1. Difficult to test individual regression coefficients due to inflated standard errors.
2. Unstable coefficient estimates, sensitive to small change in the model.

Remedial measures

1. Drop one or several highly correlated independent variables (or by stepwise regression to select appropriate variables)
2. Combine variables (dimension reduction by the use of methods such as principle component procedures)
3. Shrinkage methods such as ridge regression

$$\hat{\beta}_{ridge} = (X'X + \delta I)^{-1} X'Y \qquad \delta \geq 0$$

Another Example (extracted from Linear Models with R, Julian J. Faraway)

Car drivers like to adjust the sear position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age.

Sample size: 38 (drivers)

Dependent variable: hipcenter (the horizontal distance of the midpoint of the hips from a fixed location in the car in mm)

Independent variables:

- Age
- Weight
- HtShoes (height with shoes)
- Ht (height without shoes)
- Seated (seated height)
- Arm (arm length)
- Thight (thigh length)
- Leg (lower leg length)

```
> library(faraway)
> data(seatpos)

> names(seatpos)
[1] "Age"       "Weight"    "HtShoes"   "Ht"        "Seated"    "Arm"       "Thigh"
[8] "Leg"       "hipcenter"
> reg1 <- lm(hipcenter~., seatpos)
> summary(reg1)

Call:
lm(formula = hipcenter ~ ., data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-73.827 -22.833  -3.678  25.017  62.337 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843  
Weight        0.02631    0.33097   0.080   0.9372  
HtShoes      -2.69241    9.75304  -0.276   0.7845  
Ht            0.60134   10.12987   0.059   0.9531  
Seated        0.53375    3.76189   0.142   0.8882  
Arm          -1.32807    3.90020  -0.341   0.7359  
Thigh        -1.14312    2.66002  -0.430   0.6706  
Leg          -6.43905    4.71386  -1.366   0.1824  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,	Adjusted R-squared:  0.6001 
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

*****F-test significant, but NOT individual t-test.*****

```
>round(cor(seatpos),4)
              Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg hipcenter
Age        1.0000  0.0807 -0.0793 -0.0901 -0.1702  0.3595  0.0913 -0.0423    0.2052
Weight     0.0807  1.0000  0.8282  0.8285  0.7756  0.6976  0.5726  0.7843   -0.6403
HtShoes   -0.0793  0.8282  1.0000  0.9981  0.9297  0.7520  0.7249  0.9084   -0.7966
Ht        -0.0901  0.8285  0.9981  1.0000  0.9282  0.7521  0.7350  0.9098   -0.7989
Seated    -0.1702  0.7756  0.9297  0.9282  1.0000  0.6252  0.6071  0.8119   -0.7313
Arm        0.3595  0.6976  0.7520  0.7521  0.6252  1.0000  0.6711  0.7538   -0.5851
Thigh      0.0913  0.5726  0.7249  0.7350  0.6071  0.6711  1.0000  0.6495   -0.5912
Leg       -0.0423  0.7843  0.9084  0.9098  0.8119  0.7538  0.6495  1.0000   -0.7872
hipcenter  0.2052 -0.6403 -0.7966 -0.7989 -0.7313 -0.5851 -0.5912 -0.7872    1.0000
```

Some very large pairwise correlations

remove the 1st column in X

```
> x <- model.matrix(reg1) [,-1]
> e <- eigen(t(x) %*% x)
> e$val
[1] 3.653671e+06 2.147948e+04 9.043225e+03 2.989526e+02 1.483948e+02 8.117397e+01
5.336194e+01
[8] 7.298209e+00
> sqrt(e$val[1]/e$val)  # compute condition indexes
[1]   1.00000  13.04226  20.10032 110.55123 156.91171 212.15650 261.66698 707.54911
```

Very large condition numbers

```
> round(vif(x), 3)
    Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg
  1.998   3.647 307.429 333.138   8.951   4.496   2.763   6.694
```

VIFs of HtShoes and Ht are extremely high

Q: How to use lm to calculate R_i^2 and $(VIF)_i$?

Rerun the regression with Age, Weight and Ht only

```
> reg2 <- lm(hipcenter~Age+Weight+Ht,seatpos)
> x <- model.matrix(reg2) [,-1]
> e <- eigen(t(x) %*% x)
> sqrt(e$val[1]/e$val)  # compute condition indexes
[1]  1.00000 11.50837 15.50904
> vif(x)
     Age   Weight       Ht
1.093018 3.457681 3.463303
> summary(reg2)

Call:
lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)

Residuals:
    Min      1Q  Median      3Q     Max
-91.526 -23.005   2.164  24.950  53.982

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 528.297729 135.312947   3.904 0.000426 ***
Age           0.519504   0.408039   1.273 0.211593
Weight        0.004271   0.311720   0.014 0.989149
Ht           -4.211905   0.999056  -4.216 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562,  Adjusted R-squared:  0.6258
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08


> round(cor(x),3)
          Age Weight     Ht
Age     1.000  0.081 -0.090
Weight  0.081  1.000  0.829
Ht     -0.090  0.829  1.000
```