Chapter 2 Multiple Regression (Part 4)

1 The effect of multi-collinearity

Now, we know to find the estimator

$$(X'X)^{-1} \text{ must exist!}$$

Therefore, n must be great or at least equal to $p+1$ (WHY?) However, even $n \geq p+1$ we the inverse may still not exist when there is multi-collinearity in the predictors.

Multi-collinearity means the correlation coefficients between predictor variables are close to +1 or -1 (positive or negative). In that case, the design matrix X will be ill-conditioned, i.e. the determination, $det(X'X)$ is close to 0, or the inverse of $X'X$ is not stable. It also cause other problems. below are some discussions

1.1 An example in which two predictor variables are perfectly uncorrelated

- Work crew size example revisited

Case	Crew Size	Bonus pay	Crew productivity
i	X_1	X_2	Y
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

- Effects on Regression Coefficients

Models	b_1	b_2
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	5.375	-
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$	-	9.250
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	5.375	9.250

- Extra sums of squares

SSR($X_1 \vert X_2$)	SSR(X_1)	SSR($X_2 \vert X_1$)	SSR(X_2)
231.125	231.125	171.125	171.125

- Unrelated predictor variables (not practical!)
  - correlation coefficient of X_1 and X_2 is zero. X_1 and X_2 are uncorrelated
  - Regression effect of one predictor variable is independent of whether other predictor variables are included in the model
  - Extra sums of squares are equal to regression sums of squares
  - in that case, we can consider each predictor separately!

1.2 An example in which two predictor variables are perfectly correlated

case	X_1	X_2	Y
1	2	6	23
2	8	9	83
3	6	8	63
4	10	10	103

Two fitted lines:

$$\hat{Y} = -87 + X_1 + 18X_2$$

$$\hat{Y} = -7 + 9X_1 + 2X_2$$

because $X_2 = 5 + .5X_1$

- sometimes regression model can still obtain a good fit for the data
- but best fitted line (least squares estimator) is not unique
- (indicate) larger variability/instabability of estimator
- the common interpretation of regression coefficient is not applicable, we can not vary one predictor variable while holding other constant.

1.3 Body fat example revisited

- 20 healthy females 25-34 years old

subject	X_1	X_2	X_3	Y
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
⋮	⋮	⋮	⋮	⋮
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

The correlation matrix is

r	X_1	X_2	X_3
X_1	1.0	0.924	0.458
X_2	0.924	1.0	0.085
X_3	0.458	0.085	1.0

- Effects on Regression Coefficients

Models	b_1	b_2
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	0.8572	-
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$	-	0.8565
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	0.2224	0.6594
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$	4.334	-2.857

- Inflated variability of estimator

Models	$s\{b_1\}$	$s\{b_2\}$
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	0.1288	-
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$	-	0.1100
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	0.3034	0.2912
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$	3.016	2.582

- Extra sums of squares

SSR($X_1 \mid X_2$)	SSR(X_1)	SSR($X_2 \mid X_1$)	SSR(X_2)	SSE(X_1, X_2)
3.47	352.27	33.17	381.97	109.95

  - no unique sum of squares ascribed to any one predictor variable
  - must take into account other correlated predictor variables already included in the model

1.4 Effect of Multicollinearity

- When the multicollinearity is not strong, i.e. $(\mathbf{X}'\mathbf{X})^{-1}$ exists, we can still use the model to make prediction.
- However, the multicollinearity will result in instability of the estimated coefficient, i.e. the S.E. of the estimated coefficient is large. Thus the model is unreliable.
- The interpretation of the coefficient is difficult. For example, β_1 for X_1 is interpreted the increasment of EY when X_1 increase by 1 unit IF the other predictor variable hold constant. The real situation is that the other predictor variable CANNOT hold constant when there is multicollinearity
- However, if the multicollinearity is too serious, e.g. $X_{i1} = X_{i2}$, for which $(\mathbf{X}'\mathbf{X})^{-1}$ does not exits. There are other methods (not discussed here) such as the ridge regression and regression with penalty

2 Polynomial regression models

- General regression model: $Y = f(X) + \epsilon$, or $Y = f(X_1, X_2, ..., X_{p-1}) + \epsilon$
- Linear regression model: $f(X) = \beta_0 + \beta_1 X$ or $f(X_1, X_2, ..., X_{p-1}) = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$
- Polynomial regression function

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k$$

- reasons for using polynomial regression model:

  a. true regression function is a polynomial function
  b. better approximation than linear function ($k = 1$)

- second order

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

- Third order: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$
- Higher order: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_k X_i^k + \epsilon_i$
  higher order, more parameters (less degrees of freedom)

- two predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2} + \epsilon_i$$

  β_{12}: interaction effect coefficient

- three predictors

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{33} X_{i3}^2 \\ &\quad + \beta_{12} X_{i1} X_{i2} + \beta_{13} X_{i1} X_{i3} + \beta_{23} X_{i2} X_{i3} + \epsilon_i \end{aligned}$$

- interpretation of interaction regression models

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

  regression effects of X_1 per unit when holding X_2 constant:

$$\beta_1 + \beta_3 X_2$$

  regression effects of X_2 per unit when holding X_1 constant:

$$\beta_2 + \beta_3 X_1$$

- Easy implementation as special case of multiple regression (see the example below)

- Use polynomial regression to test linearity of regression function

  First fit a third order model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \epsilon_i$$

  then use $SSR(X^3|X, X^2)$ or $SSR(X^3, X^2|X)$ to test whether we can drop X^3 or X^3, X^2

Example 1 Suppose we have data **Data** with two predictors X_1, X_2 and response Y. If we fit a linear regression model (see **Code**)

$$\text{(Reduced model)}: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

the estimated model is

$\hat{Y}$	=	-543.594	+	$61.211X_1$	-	$101.387X_2$
(S.E.)		(228.244)		(3.774)		(42.099)

$R^2 = 0.9535$, $R_a^2 = 0.948$, $\hat{\sigma} = 170.3$
F-value 174.2 with df 2, 17.

If we consider model

$$\text{(Full model)}: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \varepsilon$$

The estimated model is

$\hat{Y}$	=	-1.56	+	$1.05X_1$	-	$0.55X_2$	+	$1.00X_1^2$	-	$1.01X_1X_2$	-	$0.03X_2^2$
(S.E.)		(0.73)		(0.02)		(0.28)		(.001)		(.003)		(0.03)

$R^2 = 0.9999$, $R_a^2 = 0.9999$, $\hat{\sigma} = 0.0878$, F-value 2.751e+08 with df 5 and 14.

It seems that X_2 and X_2^2 can be removed from the model. Let consider a test

$$H_0 : \beta_2 = \beta_{22} = 0$$

we have

$$SSE(F) = 0.0878^2 * 14, SSE(R) = 0.1986^2 * 16$$

and

$$F^* = \frac{(SSE(R) - SSE(F))/2}{SSE(F)/14} = 33.93 > F(1 - 0.05, 2, 14)$$

Thus, we reject H_0

Thus, we need to remove one variable

$$H_0' : \beta_{22} = 0$$

Under which we consider model

$$\text{(Reduced model)}' \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \varepsilon.$$

we have

$$SSE(R) = 0.08832^2 * 15$$

and

$$F^* = \frac{(SSE(R') - SSE(F))/1}{SSE(F)/14} = 1.178203 < F(1 - 0.05, 1, 14)$$

concluding H_0'.

The estimated model is

$\hat{Y}$	=	-0.90	+	$1.04X_1$	-	$0.84X_2$	+	$1.00X_1^2$	-	$1.00X_1X_2$
(S.E.)		(0.42)		(0.02)		(0.10)		(.001)		(.003)

$R^2 = 0.9999$, $R_a^2 = 0.9999$, $\hat{\sigma} = 0.08832$, F-value 3.399e+08 with df 4 and 15.