Chapter 7 Multiple Linear Regression (II)

Topics:

- Extra Sums of Squares
- General Linear Tests
- Partial Correlations
- Multicollinearity

Extra sum of squares

- Extra sum of squares are used to Measure the effect of adding variables to a regression model
- Suppose x_1 is in the regression model and we add the predictor variable x_2, what is the effect of adding x_2 to the model that already contain x_1?
- $SSE(x_1)$ = SSError for the model containing x_1
  $SSR(x_1)$ = SSRegression for the model containing x_1,
  $SSR(x_1)$ gives a measure of the effect of x_1 alone
  $SSE(x_1) + SSR(x_1) = SSTO$

- Adding x_2 to the model, we get

$$SSE(x_1, x_2)$$

and

$$SSR(x_1, x_2)$$

- $SSR(x_1, x_2)$ gives a measure of the effect of both x_1, x_2
- The effect of adding x_2 to the model that contains x_1 is measured by

$$SSE(x_1) - SSE(x_1, x_2)$$

or

$$SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1)$$

- Define $SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1)$
- $SSR(x_2|x_1)$ measures the "marginal" effect of adding x_2 to the model that already contains x_1
- The definition of extra sum of squares can be extended to any number of variables

$$SSR(x_3|x_1, x_2) = SSR(x_3, x_1, x_2) - SSR(x_1, x_2)$$

$$SSR(s_2(\mathbf{x})|s_1(\mathbf{x})) = SSR(s_2(\mathbf{x}), s_1(\mathbf{x})) - SSR(s_1(\mathbf{x})),$$

where $s_1(\mathbf{x})$ is a set of predictor variables, $s_2(\mathbf{x})$ is another set of predictor variables and there are no variables in common to $s_1(\mathbf{x})$ and $s_2(\mathbf{x})$.

Decomposition of SSR

- $$\begin{aligned} SSR(x_1, x_2, x_3) &= SSR(x_1) + SSR(x_1, x_2) - SSR(x_1) \\ &+ SSR(x_1, x_2, x_3) - SSR(x_1, x_2) \\ &= SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) \end{aligned}$$

- Degrees of Freedom

$$df(SSR(s_2(\mathbf{x})|s_1(\mathbf{x})))$$

$$= df(SSR(s_1(\mathbf{x}), s_2(\mathbf{x}))) - df(SSR(s_1(\mathbf{x})))$$

$$= \text{number of predictors in} \quad s_2(\mathbf{x})$$

- ANOVA table for decompostion with x_1 first, then x_2 added, then x_3

Source	SS	df	MS
x_1	$SSR(x_1)$	1	$MSR(x_1) = SSR(x_1)/1$
$x_2 \vert x_1$	$SSR(x_2 \vert x_1)$	1	$MSR(x_2 \vert x_1) = SSR(x_2 \vert x_1)/1$
$x_3 \vert x_1, x_2$	$SSR(x_3 \vert x_1, x_2)$	1	$MSR(x_3 \vert x_1, x_2) = SSR(x_3 \vert x_1, x_2)/1$
Regression	SSR	3	MSR = SSR /3
Error	SSE	$n-4$	$MSE = SSE/(n-4)$
Total	SSTO	$n-1$	

- Sequential sum of squares

$$SSR(x_1)$$

$$SSR(x_2|x_1)$$

$$SSR(x_3|x_1, x_2)$$

Test whether all $\beta_k = 0$

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$$

v.s

$$H_\alpha : \beta_k \neq 0, \text{for at least one} \quad k = 1, 2, \cdots p-1$$

- If H_0 is true, $F = MSR/MSE$ has an F distribution with $(p-1, n-p)$ degrees of freedom.
- Reject H_0, if $F > F(1-\alpha, p-1, n-p)$

Test whether a single $\beta_k = 0$

- $$H_0 : \beta_k = 0 \quad \text{v.s} \quad H_\alpha : \beta_k \neq 0$$

  $$t^* = \frac{b_k}{s\{b_k\}}$$

  reject H_0 if $|t^*| > t(1 - \alpha/2; n - p)$

- Equivalently, we can use general linear test approach

  $$F^* = \frac{MSR(x_k | x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_{p-1})}{MSE(x_1, x_2, \cdots, x_{p-1})}$$

  Reject $H_0 : \beta_k = 0$ if

  $$F^* > F(1 - \alpha; 1, n - p)$$

- This F^* or t^* test is a "marginal" test of the predictor x_ks effect in the model, containing all the other $p-2$ predictors.

- Example senic data

  want to test

$$H_0 : \beta_3 = 0 \quad \text{v.s} \quad H_\alpha : \beta_3 \neq 0$$

	Estimator	se	t-value	p-value
intercept	1.001162	1.314724	.761	.448003
stay	0.308181	0.059396	5.189	9.88e-07 ***
age	-0.023005	0.023516	-0.978	0.330098
xray	0.019661	0.005759	3.414	0.000899 ***

$$t^* = \frac{b_k}{s\{b_k\}} = \frac{.019661}{.005759} = 3.414$$

P-value is less than .05, reject $H_0 : \beta_3 = 0$.

From ANOVA table,

	df	SS	MS	F-value	p-value
stay	1	57.305	57.305	48.6920	2.444e-10 ***
age	1	2.075	2.075	1.7632	0.1870031
xray	1	13.719	13.719	11.6568	0.0008992 ***
Residuals	109	128.281	1.177		

$SSR(x_3|x_1, x_2) = 13.719$, $MSE = 1.177$,

$$F^* = MSR(x_3|x_1, x_2)/MSE = 13.719/1.177 = 11.656$$

P-value is less than .05, reject $H_0 : \beta_3 = 0$.

Test whether several $\beta_k = 0$

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

v.s

$$H_\alpha : H_0 \quad \text{is false}$$

$$F^* = \frac{MSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})}{MSE(x_1, \cdots, x_{q-1}, x_q, \cdots, x_{p-1})}$$

where

$$MSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})$$

$$= SSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})/(p-q)$$

Reject H_0 if

$$F^* > F(1-\alpha; p-q, n-p)$$

- Example senic data

  want to test

$$H_0 : \beta_2 = 0, \beta_3 = 0 \quad \text{v.s} \quad H_\alpha : H_0 \quad \text{is false}$$

  i.e. We wish to test in the senic data that if variable age and xray can be dropped from the regression model.

$$F^* = \frac{MSR(x_2, x_3|x_1)}{MSE(x_1, x_2, x_3)}$$

$$\begin{aligned} SSR(x_2, x_3|x_1) &= SSR(x_2|x_1) + SSR(x_3|x_1, x_2) \\ &= 2.075 + 13.719 \\ &= 15.794 \end{aligned}$$

$$F^* = \frac{15.794/2}{1.177} = 6.71$$

compare to $F(.95, 2, 109) = 3.08$ Reject H_0, conclude that at least one of the variables age and xray should not be dropped from the regression model.

Summary of tests concerning regression coefficients

Test whether all $\beta_k = 0$

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$$

$$H_\alpha : \text{ not all } \beta_k (k = 1, 2, \cdots p-1) \text{ equal zero}$$

- Test statistic is

$$\begin{aligned} F^* &= \frac{SSR(X_1, \cdots, X_{p-1})}{p-1} \div \frac{SSE(X_1, \cdots, X_{p-1})}{n-p} \\ &= \frac{MSR}{MSE} \end{aligned}$$

- If H_0 is true, $F^* \sim F(p-1, n-p)$, reject H_0, if $F^* > F(1-\alpha, p-1, n-p)$

Test whether a single $\beta_k = 0$

- $$H_0 : \beta_k = 0 \quad \text{v.s} \quad H_\alpha : \beta_k \neq 0$$

  $$t^* = \frac{b_k}{s\{b_k\}}$$

  reject H_0 if $|t^*| > t(1 - \alpha/2; n - p)$

- Equivalently, we can use general linear test approach

  $$F^* = \frac{MSR(x_k | x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_{p-1})}{MSE(x_1, x_2, \cdots, x_{p-1})}$$

  Reject $H_0 : \beta_k = 0$ if

  $$F^* > F(1 - \alpha; 1, n - p)$$

Test whether several $\beta_k = 0$

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

$$H_\alpha : H_0 \quad \text{is false}$$

$$F^* = \frac{MSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})}{MSE(x_1, \cdots, x_{q-1}, x_q, \cdots, x_{p-1})}$$

where

$$MSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})$$

$$= SSR(x_q, \cdots, x_{p-1} | x_1, \cdots, x_{q-1})/(p-q)$$

Reject H_0 if

$$F^* > F(1-\alpha; p-q, n-p)$$

Note: If $q = 1$, the test is whether all regression coefficients equal zero.

If $q = p - 1$, the test is whether a single regression coefficient equals zero.

Coefficient of Partial Determination

- Measures the marginal (or partial) contribution of one predictor variable in reducing variability in y when other predictor variables are already in the regression model
- Example

$$R^2_{Y2|1} == \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

$SSE(X_1)$ measures the variation in Y when X_1 is included in the model

$SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model

$R^2_{Y2|1}$ represents the proportion reduction in variation in Y given X_1 is in the model that is gained by adding X_2 to the model.

General Form

$$R^2_{Yi|1,2,\cdots,i-1,i+1,\cdots,p-1}$$

$$= \frac{SSR(X_i|X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{p-1})}{SSE(X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{p-1})}$$

Comments:

$$0 \leq R^2_{Yi|1,2,\cdots,i-1,i+1,\cdots,p-1} \leq 1$$

Interpretation of $R^2_{Y2|1}$

–(a) Regress Y on X_1 and find the residuals

$$e_i(Y|X_1) = Y_i - \hat{Y}_i(X_1)$$

where $\hat{Y}_i(X_1)$ is the predicted value of Y based on the regression on X_1

–(b) Regress X_2 on X_1 and find the residuals

$$e_i(X_2|X_1) = X_{i2} - \hat{X}_{i2}(X_1)$$

where $\hat{X}_{i2}(X_1)$ is the predicted value of X_2 from the regression on X_1

–(c) The coefficient of simple determination R^2 between these two sets of residuals is $R^2_{Y2|1}$

–(d) $R^2_{Y2|1}$ is the proportion of variation in Y accounted for by X_2 after both Y and X_2 have been adjusted for their linear relationship with X_1

- Example Senic data

  want to calculate $R^2_{Y2|1}$

	df	SS	MS	F-value	p-value
stay	1	57.305	57.305	48.6920	2.444e-10 ***
age	1	2.075	2.075	1.7632	0.1870031
xray	1	13.719	13.719	11.6568	0.0008992 ***
Residuals	109	128.281	1.177		

$$SSE(X_1) = SSTO - SSR(X_1) = (57.305 + 2.075 + 13.719 + 128.281) - 57.305 = 144.075$$

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = 2.075/144.075 = .014$$

Multicolinearity and It's Effect

When the predictor variables are correlated among themselves, we say that there is multicollinearity.

1. The estimate of any parameter, say β_2, depends on all the variables that are included in the model.

2. The sum of squares for any variable, say x_2, depends on all the other variables that are included in the model. For example, none of $SSR(x_2)$, $SSR(x_2|x_1)$, and $SSR(x_2|x_3, x_4)$ would typically be equal.

3. A moderate amount of collinearity has little effect on predictions and therefore little effect on SSE, R^2, and the explanatory power of the model. Collinearity increases the vairance of the $\hat{\beta}_k$s, making the estimates of the parameters less reliable. Sometimes a large amount of collinearity can have an effect on predictions.

4. Suppose the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

is fitted and we obtain t statistics for each parameter.

If the t statistic for testing $H_0 : \beta_1 = 0$ is small, we are led to the model

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

If the t statistic for testing $H_0 : \beta_2 = 0$ is small, we are led to the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i.$$

However, if the t statistics for both tests are small, we are not led to the model

$$y_i = \beta_0 + \beta_3 x_{i3} + \epsilon_i.$$

Multicollinearity can greatly affect

- the regression coefficient
- the variance of the regression coefficients
- our understanding of the predictor variables and their effect on the response

Two types of SS from SAS

- Type I SS: Add one variable at a time in order (also called sequential SS)
- Type II SS: Each variable given all others are in the model

Two special case

- Uncorrelated Predictor variables: Assume columns in the design matrix were uncorrelated ($r = 0$). In that case Type I and Type II SS will be the same. The contribution of each explanatory variable to the model is the same whether or not the other explanatory variables are in the model.

  —-Example from book (page 279)

- Predictor variables are perfectly correlated: The Type II SS for the predictor variables involved will be zero because when one is included the other is redundant. It explains NO additional variation over the other variables.

  —Example from book (page 281)

Underlying reason for multicollinearity

- the matrix $\mathbf{X}^T\mathbf{X}$ is close to singular and is difficult to invert accurately.
- Or there is too much correlation among the explanatory variables and it is therefore difficult to determine the regression coefficients.
- If we can refine a model to move redundancy in the explanatory variables, we can solve the multicollinearity problem.

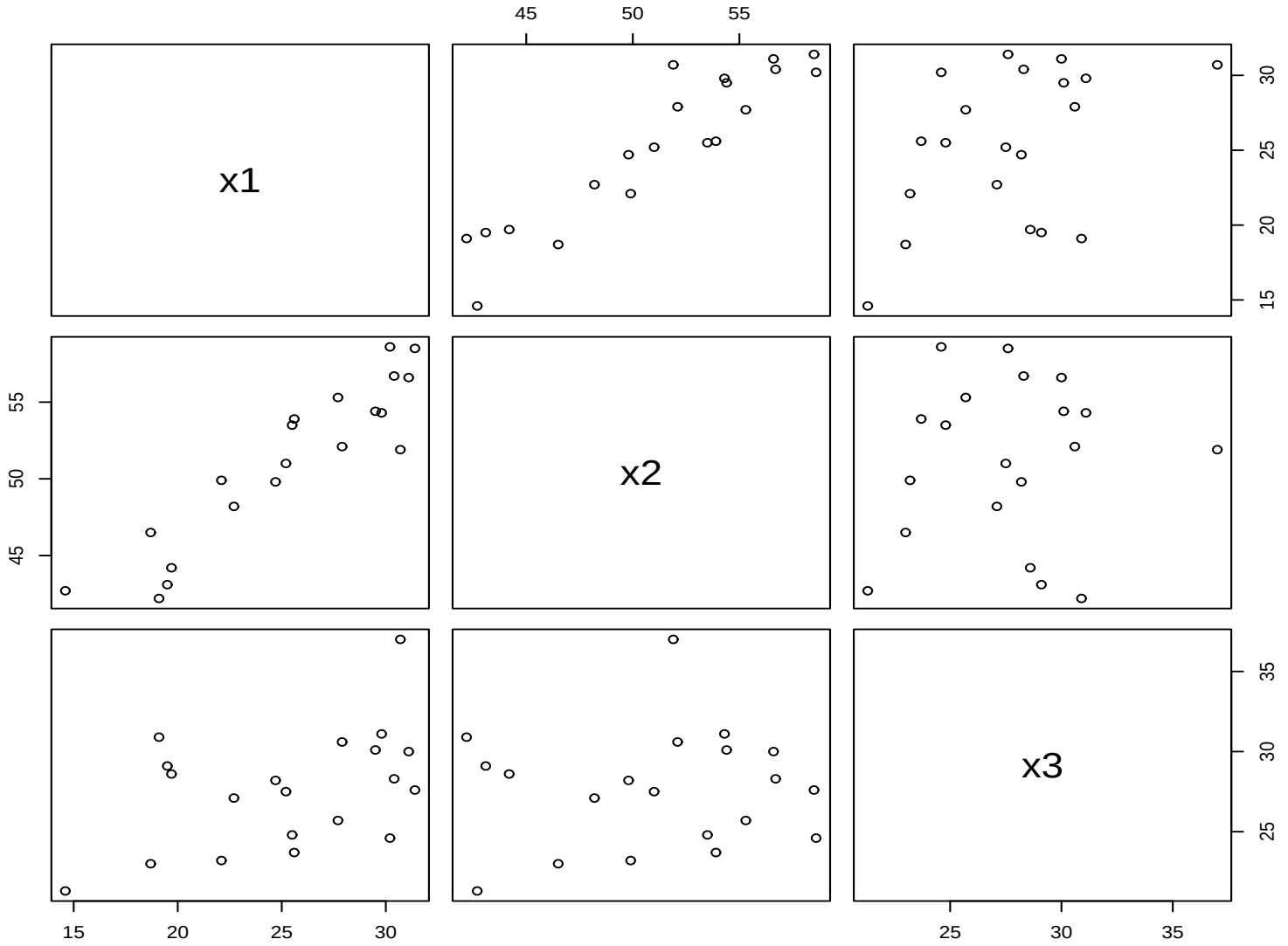
Example: Body fat data

bodyfat data

- The data is a portion of data for a study of the relation of amount of body fat (y) to several possible predictor variables, based on a sample of 20 healthy females $25 - 34$ years old.
- Predictor variables are triceps skinfold thickness ($x1$), thigh circumference ($x2$), and midarm circumference ($x3$).
- Response variable is y. The amount of body fat for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water.

- Because it is hard to get the response variable “body fat”, it would be very helpful if a regression model with some or all of these predictor variables could provide reliable estimates of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.
- Scatter plot matrix and correlation matrix of the predictor variables

Figure 1: scatterplot of three predictor variables—body fat example



- Correlation matrix of $\mathbf{X}$ variables

$$r_{\mathbf{xx}} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

	Estimator	se	t-value	p-value
(Intercept)	117.085	99.782	1.173	0.258
x1	4.334	3.016	1.437	0.170
x2	-2.857	2.582	-1.106	0.285
x3	-2.186	1.595	-1.370	0.190

Notice that the overall F statistic is 21.52 on 3 and 16 DF, p-value: 7.343e-06. But none of the individual t's are significant. This indicates multicollinearity problem.

Effects on Regression Coefficients

Variables in Model	b_1	b_2
x_1	.8572	none
x_2	none	.8565
x_1, x_2	.2224	.6594
x_1, x_2, x_3	4.334	-2.857

- Regression coefficient for x_1, triceps skinfold thickness, varies markedly depending on which other variables are included in the model

- Regression coefficient for x_2, also varies markedly depending on which other variables are included in the model. b_2 even changes sign when x_3 is added to model that includes x_1 and x_2.

- When variables are correlated, the regression coefficient of any one variable depends on which other predictor variables are included in the model and which ones are left out.
- A regression coefficient does not reflect any inherent effect of the particular predictor variables on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model.

Effects on Extra sum of Squares

$$SSR(x_1) = 352.27$$

$$SSR(x_1|x_2) = 3.47$$

$$SSR(x_2) = 381.97$$

$$SSR(x_2|x_1) = 33.17$$

- $SSR(x_1|x_2)$ is small compared with $SSR(x_1)$, since x_1 and x_2 are highly correlated with each other and with the response variable.

- When x_2 is included in the regression model, the marginal contribution of x_1 in reducing the error sum of squares is comparatively small because x_2 contains much of the same information as x_1.

- multicollinearity also affects the coefficients of partial determination through its effects on the extra sums of squares.

  —Example: x_1 is highly related to y:

$$R^2_{Y1} = \frac{SSR(x_1)}{SSTO} = \frac{352.27}{495.39} = .71$$

  however, the coefficient of partial determination between y and x_1, when x_2 is already in the regression model, is much smaller:

$$R^2_{Y1|2} = \frac{SSR(x_1|x_2)}{SSE(x_2)} = \frac{3.47}{113.43} = .03$$

- x_1 and x_2 are highly correlated with each other and with the response variable. Hence, x_1 provides only relatively limited additional information beyond that furnished by x_2.

- The extra sum of squares for a predictor variable after other correlated predictor variables are in the model need not necessarily be smaller than before these other variables are in the model.

Effects on $s\{b_k\}$

Variables in Model	$s(b_1)$	$s(b_2)$
x_1	$.8572_{(.1288)}$	none
x_2	none	$.8565_{(.1100)}$
x_1, x_2	$.2224_{(.3034)}$	$.6594_{(.2912)}$
x_1, x_2, x_3	$4.334_{(3.016)}$	$-2.857_{(2.582)}$

- As more predictor variables are added to the regression model, the estimated regression coefficients b_1 and b_2 becomes more imprecise

- The high degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients

Effects on Fitted Values and Predictions

Table 1: MSE comparison of different sets of predictor variables

Variables in Model	MSE
x_1	7.95
x_2	6.3
x_1, x_2	6.47
x_1, x_2, x_3	6.15

Table 2: Fitted value at $x_{h1} = 25, x_{h2} = 50, x_{h3} = 29$ comparison of different models

Variables in Model	fitted value	$s\{\hat{y}_h\}$
x_1	19.93	.632
x_2	19.19	.5758
x_1, x_2	19.35	.624
x_1, x_2, x_3	19.19	.621

Comments:

- As long as points 1 through 4 are kept in mind, a moderate amount of collinearity is not a big problem.
- The overall F statistic is significant, but none of the individual ts are significant. This indicates multicollinearity problem.
- High multicollinearity among the predictor variables does not prevent the mean square error, measuring the variability of the error terms, from being steadily reduced as additional variables are added to the regression model.
- The precision of fitted values within the range of the observations on the predictor variables is not eroded with the addition of correlated predictor variables into the regression model.

Remedies for Multicollinearity

- Variable selection
- Use biased regression methods such as ridge regression or principle components regression, especially there is an interest in the regression coefficients themselves.