Multicollinearity occurs when several of the predictors under consideration \( x_1, x_2, \ldots, x_k \) are highly correlated with other predictors. Problems arising when this happens include:

1. Adding/removing a predictor changes estimated regression coefficients substantially, and hence some conclusions based on the model change substantially as well.

2. Sampling distribution of individual \( b_j \)'s may have hugely inflated variance, reflected in huge C.I.'s for some \( b_j \).

3. The standard interpretation of a \( b_j \) as the mean change in the response when \( x_j \) is increased by one is no longer valid. If \( x_2 \) is highly correlated with \( x_3 \), we can’t think of holding \( x_3 \) fixed while increasing \( x_2 \).

Multicollinearity does not pose a problem when the main use of the regression model is for prediction; predicted values and prediction intervals will not tend to change drastically when predictors correlated with other predictors are added to the model, when prediction is within the scope of the observed predictors.

Multicollinearity can be seen as a duplication of information and is often avoided simply by “weeding out” predictors in the usual fashion: use of the best-subsets \( C(p) \) statistic, “extra sums of squares” tests that a subset of regression coefficients are zero, etc.

A formal method for determining the presence of multicollinearity is the variance inflation factor (VIF). VIF’s measure how much variances of estimated regression coefficients are inflated when compared to having uncorrelated predictors. We will use the standardized regression model of Section 7.5.

Let \( Y_i^* = \frac{Y_i - \bar{Y}}{s_Y} \) and \( x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \),

where \( s_Y^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \), \( s_j^2 = n^{-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \), and \( \bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij} \). These variables are centered about their means and “standardized” to have Euclidean norm 1. For example, \( ||x_j^*||^2 = (x_{1j}^*)^2 + \cdots + (x_{nj}^*)^2 = 1 \).

Note that \( (Y^*)'(Y^*) = (x_j^*)'(x_j^*) = 1 \) and \( (x_j^*)'(x_i^*) = \text{corr}(x_j, x_i) \overset{def}{=} r_{js} \). Consider the standardized regression model

\[
Y_i^* = \beta_1^* x_{i1}^* + \cdots + \beta_k^* x_{ik}^* + \epsilon_i^*.
\]

Define the \( k \times k \) sample correlation matrix \( R \) for the standardized predictors, and the \( n \times k \) design matrix \( X^* \) to be:

\[
R = \begin{bmatrix}
1 & r_{21} & \cdots & r_{k1} \\
r_{12} & 1 & \cdots & r_{k2} \\
\vdots & \vdots & \ddots & \vdots \\
r_{1k} & r_{2k} & \cdots & 1
\end{bmatrix}, \quad \ X^* = \begin{bmatrix}
x_{11}^* & x_{12}^* & \cdots & x_{1k}^* \\
x_{21}^* & x_{22}^* & \cdots & x_{2k}^* \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1}^* & x_{n2}^* & \cdots & x_{nk}^*
\end{bmatrix}.
\]

Since \( (X^*)'(X^*) = R \), the least-squares estimate of \( \beta^* = (\beta_1^*, \ldots, \beta_k^*)' \) is given by \( b^* = R^{-1}(X^*)'Y^* \). Hence \( \text{Cov}(b^*) = R^{-1}(\sigma^*)^2 \).

Now note that if all predictors are uncorrelated then \( R = I_k = R^{-1} \). Hence the \( i \)th diagonal element of \( R^{-1} \) is how much the variance of \( b_i^* \) is inflated due to correlation between predictors. We call this the \( i \)th variance inflation factor: \( VIF_i = (R^{-1})_{ii} \). Usually the
The largest $VIF_i$ is taken to be a measure of the seriousness of the multicollinearity among the predictors, with $\max_i\{VIF_i\} > 10$ indicating that multicollinearity is unduly affecting the least squares estimates of the regression coefficients.

**Body Fat Example from your text**

---

* Body fat data from Chapter 7
---

data body;
input triceps thigh midarm bodyfat @@;
cards;
19.5 43.1 29.1 11.9 24.7 49.8 28.2 22.8
30.7 51.9 37.0 18.7 29.8 54.3 31.1 20.1
19.1 42.2 30.9 12.9 25.6 53.9 23.7 21.7
31.4 58.5 27.6 27.1 25.5 53.5 21.3 30.6
22.1 49.9 23.2 25.5 25.4 56.7 28.3 27.2
18.7 46.5 23.0 11.7 29.8 44.2 28.6 17.8
14.6 42.7 21.3 12.8 25.6 53.9 23.7 21.7
27.7 55.3 25.7 30.2 58.6 24.6 25.4
22.7 48.2 27.1 14.8 25.4 53.5 24.8 21.1
; run;

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>triceps</th>
<th>thigh</th>
<th>midarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>triceps</td>
<td>1.00000</td>
<td>0.92384</td>
<td>0.45778</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>0.0424</td>
<td></td>
</tr>
<tr>
<td>thigh</td>
<td>0.92384</td>
<td>1.00000</td>
<td>0.08467</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>0.7227</td>
<td></td>
</tr>
<tr>
<td>midarm</td>
<td>0.45778</td>
<td>0.08467</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0424</td>
<td>0.7227</td>
<td></td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>396.98461</td>
<td>132.32820</td>
<td>21.52</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>98.40489</td>
<td>6.15031</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>495.38950</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.47998 R-Square 0.8014

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| Type I SS | Inflation |
|----------|----|--------------------|----------------|---------|------|---------|-----------|
| Intercept| 1  | 117.08469          | 99.78240       | 1.17    | 0.2578 | 8156.76050 | 0         |
| triceps | 1  | 4.33409            | 3.01551        | 1.44    | 0.1699 | 352.26980  | 708.84291 |
| thigh   | 1  | -2.85685           | 2.58202        | -1.11   | 0.2849 | 33.16891   | 564.34339 |
| midarm  | 1  | -2.18606           | 1.59550        | -1.37   | 0.1896 | 11.54590   | 104.60601 |

This is an extremely interesting example in which we would not reject dropping any of the three predictors from the model, yet the overall $p$-value for $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ is zero to three decimal-places. The largest $VIF_j$ is 708.8, indicating a high degree of correlation among the predictors. The sequential extra sums of squares is given in the table: $SSR(x_1) = 352.3$; $SSR(x_2|x_1) = 33.2$, and $SSR(x_3|x_1,x_2) = 11.5$. Almost all of the $SSR(x_1,x_2,x_3) = 397.0$ is explained by $x_1$ (triceps) alone.

Also, note as required,

$$SSR(x_1,x_2,x_3) = 397.0 = 352.3 + 33.2 + 11.5 = SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1,x_2).$$