Chapter 13: Multicollinearity

What is the problem when predictor variables are strongly correlated? (or when one predictor variable is a linear combination of the others).

If $X_1$ and $X_2$ are perfectly correlated, then they are said to be collinear. For example, suppose $X_2 = 5 + 0.5X_1$
Multicollinearity is the situation where one or more predictor variables are “nearly” linearly related to the others.

If one of the predictors is almost perfectly predicted from the set of other variables, then you have multicollinearity.

How does this affect an analysis?
• Effects of multicollinearity:

1) the fitted values \( \hat{Y}_i \)'s are probably fine (not greatly affected by the fact that there is multicollinearity)

2) the estimated slopes have high variability; the \( \hat{\beta}_j \)'s have large standard errors

3) the \( \hat{\beta}_j \)'s have great sensitivity to minor changes in the model/data (e.g. when a variable or case is removed)

3) the fitted model may not reflect the general population that the sample was drawn from (i.e. a new sample may suggest a very different model)

5) estimation of non-multicollinear coefficients \( \hat{\beta}_j \)'s are not affected
• What does this mean about the analysis?

1) If you’re ONLY interested in predicting for this data set, i.e. the $\hat{Y}_i$s, (i.e. no testing), then it’s not much of a concern

2) High std. err. of $\hat{\beta}_j$s may mean few $\beta_j$s are significant (testing $H_0 : \beta_j = 0$), even when a relationship truly exists.

3) The estimated regression coefficients can vary widely from one sample to the next.
How does high correlation in your predictor variables lead to unstable estimates of the $\hat{\beta}_j$s?

**Example:** Consider a hypothetical situation...

- Suppose we’re measuring the white blood-cell count (WBC) in a population of sick patients (the response, as count per microliter of blood).
- Higher white blood cell counts coincide with sicker patients.
- We also have measures on two other quantities in the body:
  1) quantity of Organism A in the body, $X_1$
  2) quantity of Organism B in the body, $X_2$

The researchers don’t know it, but Organism A attacks the body (and increases WBC), and Organism B is inert (a harmless by-product of Organism A).
What does the (simulated) data look like?

- The relationship in the top left plot is what the researchers would like to uncover.

- But when we consider the data, there also seems to be a relationship of WBC with $X_2$. 
• Generate 10 simulated data sets, and fit the model using both predictors:

<table>
<thead>
<tr>
<th>Run</th>
<th>$\hat{\beta}_1$</th>
<th>Std. error</th>
<th>$p$ value</th>
<th>$\hat{\beta}_2$</th>
<th>Std. error</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.89</td>
<td>20.19</td>
<td>0.28</td>
<td>0.36</td>
<td>2.00</td>
<td>0.86</td>
</tr>
<tr>
<td>2</td>
<td>-16.75</td>
<td>22.51</td>
<td>0.46</td>
<td>4.14</td>
<td>2.24</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>20.33</td>
<td>17.98</td>
<td>0.26</td>
<td>0.53</td>
<td>1.80</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>25.82</td>
<td>16.09</td>
<td>0.11</td>
<td>0.05</td>
<td>1.60</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>4.87</td>
<td>21.36</td>
<td>0.82</td>
<td>1.95</td>
<td>2.14</td>
<td>0.36</td>
</tr>
<tr>
<td>6</td>
<td>38.61</td>
<td>19.30</td>
<td><strong>0.05</strong></td>
<td>-1.40</td>
<td>1.93</td>
<td>0.47</td>
</tr>
<tr>
<td>7</td>
<td>-3.92</td>
<td>19.90</td>
<td>0.84</td>
<td>3.10</td>
<td>1.98</td>
<td>0.12</td>
</tr>
<tr>
<td>8</td>
<td>46.74</td>
<td>23.98</td>
<td><strong>0.05</strong></td>
<td>-2.19</td>
<td>2.41</td>
<td>0.37</td>
</tr>
<tr>
<td>9</td>
<td>-10.74</td>
<td>17.45</td>
<td>0.54</td>
<td>3.65</td>
<td>1.75</td>
<td><strong>0.04</strong></td>
</tr>
<tr>
<td>10</td>
<td>40.90</td>
<td>18.12</td>
<td><strong>0.03</strong></td>
<td>-1.48</td>
<td>1.83</td>
<td>0.42</td>
</tr>
</tbody>
</table>

• The $\hat{\beta}$s are very variable (they change a lot from one simulation to the next).
• The signs on $\hat{\beta}$s are even sometimes “−”
• Tests of $H_0 : \beta_j = 0$ are not often significant (eventhough a relationship exists).
• Our interpretation of MLR coefficients is lost when multicollinearity exists because we can’t hold one variable constant while changing the other (they MUST move simultaneously).
Visual of instability of coefficients

If we have instability of coefficient estimates, then the fitted plane is very dependent on the particular sample chosen. See the 4 runs below:
Investigators studied physical characteristics and ability in 13 football punters.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>Distance traveled in feet</td>
</tr>
<tr>
<td>Hang</td>
<td>Time in air in seconds</td>
</tr>
<tr>
<td>R_Strength</td>
<td>Right leg strength in pounds</td>
</tr>
<tr>
<td>L_Strength</td>
<td>Left leg strength in pounds</td>
</tr>
<tr>
<td>R_Flexibility</td>
<td>Right leg flexibility in degrees</td>
</tr>
<tr>
<td>L_Flexibility</td>
<td>Left leg flexibility in degrees</td>
</tr>
<tr>
<td>O_Strength</td>
<td>Overall leg strength in pounds</td>
</tr>
</tbody>
</table>

We’ll consider *Distance* as the dependent variable, which is the variable to be predicted.

Variable investigation:

```r
> football.data=read.csv("football_punts.csv")
> attach(football.data)
> round(cor(football.data),4)

<table>
<thead>
<tr>
<th></th>
<th>Distance</th>
<th>Hang</th>
<th>R_Strength</th>
<th>L_Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>1</td>
<td>0.8189</td>
<td>0.7915</td>
<td>0.7440</td>
</tr>
<tr>
<td>Hang</td>
<td>0.8189</td>
<td>1</td>
<td>0.8321</td>
<td>0.8622</td>
</tr>
<tr>
<td>R_Strength</td>
<td>0.7915</td>
<td>0.8321</td>
<td>1</td>
<td>0.8957</td>
</tr>
<tr>
<td>L_Strength</td>
<td>0.7440</td>
<td>0.8622</td>
<td>0.8957</td>
<td>1</td>
</tr>
<tr>
<td>R_Flexibility</td>
<td>0.8063</td>
<td>0.8451</td>
<td>0.7747</td>
<td>0.8141</td>
</tr>
<tr>
<td>L_Flexibility</td>
<td>0.4077</td>
<td>0.5327</td>
<td>0.3569</td>
<td>0.4232</td>
</tr>
<tr>
<td>O_Strength</td>
<td>0.7962</td>
<td>0.7558</td>
<td>0.6065</td>
<td>0.5231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>R_Flexibility</th>
<th>L_Flexibility</th>
<th>O_Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>0.8063</td>
<td>0.4077</td>
<td>0.7962</td>
</tr>
<tr>
<td>Hang</td>
<td>0.8451</td>
<td>0.5327</td>
<td>0.7558</td>
</tr>
<tr>
<td>R_Strength</td>
<td>0.7747</td>
<td>0.3569</td>
<td>0.6065</td>
</tr>
<tr>
<td>L_Strength</td>
<td>0.8141</td>
<td>0.4232</td>
<td>0.5231</td>
</tr>
<tr>
<td>R_Flexibility</td>
<td>1</td>
<td>0.6895</td>
<td>0.6903</td>
</tr>
<tr>
<td>L_Flexibility</td>
<td>0.6895</td>
<td>1</td>
<td>0.4081</td>
</tr>
<tr>
<td>O_Strength</td>
<td>0.6903</td>
<td>0.4081</td>
<td>1</td>
</tr>
</tbody>
</table>

Fairly high correlation between many variables.

> plot(football.data,pch=16)
Linearity looks reasonable, except for the *left flexibility* variable. This variable seems to have some outliers, and has the lowest correlation with the other variables (most kickers are right-footed?)
Marginally (i.e. in the XY scatterplots), every potential predictor is positively correlated with Distance.

Let’s fit a model with only Right Leg Strength and Overall Leg Strength:

```r
> lm.out=lm(Distance ~ R_Strength + O_Strength)
> summary(lm.out)
```

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------------|----------|------------|---------|-----------|
| (Intercept)    | 12.7676  | 24.9926    | 0.511   | 0.6205    |
| R_Strength     | 0.5563   | 0.2104     | 2.644   | 0.0246 *  |
| O_Strength     | 0.2717   | 0.1003     | 2.709   | 0.0220 *  |

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 13.21 on 10 degrees of freedom  
Multiple R-Squared: 0.7845, Adjusted R-squared: 0.7414  
F-statistic: 18.2 on 2 and 10 DF,  p-value: 0.0004645

Both predictors are significant and their regression coefficients are positive.
Let’s fit a model with multiple predictors, many of which are strongly correlated:

```r
> lm.out.2=lm(Distance ~ R_Strength + L_Strength + R_Flexibility + Hang + O_Strength)
> summary(lm.out.2)
```

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | -29.6979 | 68.9639    | -0.431  | 0.680    |
| R_Strength          | 0.3242   | 0.4573     | 0.709   | 0.501    |
| L_Strength          | 0.1024   | 0.5712     | 0.179   | 0.863    |
| R_Flexibility       | 0.7482   | 1.1471     | 0.652   | 0.535    |
| Hang                | -0.4941  | 25.0200    | -0.020  | 0.985    |
| O_Strength          | 0.2327   | 0.1625     | 1.432   | 0.195    |

Residual standard error: 14.94 on 7 degrees of freedom
Multiple R-Squared: 0.8069, Adjusted R-squared: 0.669
F-statistic: 5.851 on 5 and 7 DF, p-value: 0.01924

NONE of the predictors are significant in the individual \( t \)-tests, yet the \( R^2 \) is high.

The regression coefficient for *hang* is negative (eventhough we know it has a positive relationship with *Distance*).
Diagnosis of multicollinearity
(and collinearity)

• examine pairwise correlations

• look for $\hat{\beta}_j$s with unusual slopes (+/-)

• notice if there is great sensitivity of $\hat{\beta}_j$s to removal of a case

• use the Variance Inflation Factor (VIF) to make the call:

$$VIF_j = \frac{1}{1 - R^2_j}$$

- Recall that $R^2_j$ is the % of variability in $X_j$ explained by all the other predictors (Compute $R^2_j$ by regressing $X_j$ on the other variables).
\[ VIF_j = \frac{1}{1 - R_j^2} \]

- When \( R_j^2 \) is close to 1 (i.e. you can predict \( X_j \) very well from the other predictors), \( VIF_j \) will be large.

- \( VIF > 10 \) indicates a problem

- \( VIF > 100 \) indicates a BIG problem

COMMENT: The slopes have specific units, so finding the largest standard error does not necessarily mean you’ve found the variable with the largest VIF.
Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -29.6979 | 68.9639    | -0.431  | 0.680    |
| R_Strength           | 0.3242   | 0.4573     | 0.709   | 0.501    |
| L_Strength           | 0.1024   | 0.5712     | 0.179   | 0.863    |
| R_Flexibility        | 0.7482   | 1.1471     | 0.652   | 0.535    |
| Hang                 | -0.4941  | 25.0200    | -0.020  | 0.985    |
| O_Strength           | 0.2327   | 0.1625     | 1.432   | 0.195    |

Residual standard error: 14.94 on 7 degrees of freedom
Multiple R-Squared: 0.8069, Adjusted R-squared: 0.669
F-statistic: 5.851 on 5 and 7 DF, p-value: 0.01924

What are the VIFs for this model?

> vif(lm.out.2)

R_Strength  L_Strength  R_Flexibility  Hang  O_Strength
5.837192    9.217625    4.295805    8.041294    3.243021

Or getting the VIF manually for **Hang**:

> R.hang.squared=summary(lm(Hang ~ R_Strength + L_Strength +
| R_Flexibility + O_Strength))$r.squared

> 1/(1-R.hang.squared)

[1] 8.041294

**NOTE:** ‘vif’ is a function in the *car* library.
Remedies of multicollinearity (and collinearity)

- Could use the model only for prediction $\hat{Y}_i$s (but probably not so useful)

Notice how the fitted planes may be quite different, but the predicted values in the realm of observed $(x_1, x_2)$ combinations don’t change much.

- Drop variables that are highly correlated (there are choices to be made here)
• Create composite (combined) variables:
  i) If $X_1$, $X_2$, and $X_3$ are strongly correlated, 
  use $X_{new} = 0.25X_1 + 0.25X_2 + 0.50X_3$

ii) Principal components analysis

These are dimension reduction techniques.

• Use Ridge Regression which will introduce some bias to the regression coefficient estimates, but can greatly reduce the standard errors.
General idea of Ridge Regression:

Allow for a small amount of bias in the estimated $\hat{\beta}$ and get a greatly reduced variability in your estimate (below, $b = \hat{\beta}$)

This is a trade-off between ‘unbiased-ness’ and variability, but you think it’s a net gain.

Other courses will deal with this.
Multicollinearity (and Collinearity)

- It is not an assumption violation, but it is important to check for it.

- The problem can be a pair of highly correlated variables, or a large group of moderately correlated variables.

- It can cause a lot of problems in an analysis.

- Inclusion of interaction terms introduces some collinearity to a model, but their inclusion is necessary if the interaction effect is important.
Multicollinearity & Centering of variables

Consider data on hourly wage:

\[
\text{logHW} - \text{The natural log of hourly wage.}
\]

\[
\text{EDUCL} - \text{education level in years.}
\]

\[
\text{AGE} - \text{age in years.}
\]
The additive model, no interaction:

\[
> \text{lm.out.1}=\text{lm}(\text{logHW} \sim \text{EDUCL} + \text{AGE}) \\
> \text{summary(lm.out.1)}
\]

Coefficients:

\[
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & t \text{ value} & \text{Pr(>|t|)} \\
(\text{Intercept}) & 0.097646 & 0.356786 & 0.274 & 0.784684 \\
\text{EDUCL} & 0.122022 & 0.021740 & 5.613 & 8.66e-08 *** \\
\text{AGE} & 0.014351 & 0.004213 & 3.406 & 0.000834 *** \\
\end{array}
\]

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5986 on 159 degrees of freedom
Multiple R-squared: 0.2038, Adjusted R-squared: 0.1938
F-statistic: 20.35 on 2 and 159 DF,  p-value: 1.347e-08

\[
> \text{vif(lm.out.1)}
\]

\[
\begin{array}{cccccc}
\text{EDUCL} & \text{AGE} \\
1.004588 & 1.004588 \\
\end{array}
\]
The model with interaction:

```r
> lm.out.2 = lm(logHW ~ EDUCL + AGE + EDUCL:AGE)
> summary(lm.out.2)
```

Coefficients:

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 3.224920 | 1.203000   | 2.681   | 0.00813 ** |
| EDUCL     | -0.101347| 0.084931   | -1.193  | 0.23455  |
| AGE       | -0.063481| 0.028943   | -2.193  | 0.02975 * |
| EDUCL:AGE | 0.005579 | 0.002054   | 2.717   | 0.00732 ** |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5869 on 158 degrees of freedom  
Multiple R-squared: 0.2394, Adjusted R-squared: 0.2249  
F-statistic: 16.58 on 3 and 158 DF, p-value: 2.058e-09

```r
> vif(lm.out.2)
         EDUCL      AGE EDUCL:AGE
15.94743 49.31408 60.62104
```

Very high VIF values.
The additive model with centered variables:

\[ \text{cent.educ} = \text{EDUCL} - \text{mean(EDUCL)} \]
\[ \text{cent.age} = \text{AGE} - \text{mean(AGE)} \]

\[ \text{lm.out.3} = \text{lm(logHW} \sim \text{cent.educ} + \text{cent.age}) \]
\[ \text{summary(lm.out.3)} \]

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 2.352738 | 0.047030   | 50.026  | < 2e-16  *** |
| cent.educ      | 0.122022 | 0.021740   | 5.613   | 8.66e-08 *** |
| cent.age       | 0.014351 | 0.004213   | 3.406   | 0.000834 *** |

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5986 on 159 degrees of freedom
Multiple R-squared: 0.2038, Adjusted R-squared: 0.1938
F-statistic: 20.35 on 2 and 159 DF,  p-value: 1.347e-08

\[ \text{vif(lm.out.3)} \]

<table>
<thead>
<tr>
<th>cent.educ</th>
<th>cent.age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.004588</td>
<td>1.004588</td>
</tr>
</tbody>
</table>

The estimated partial regression coefficients for \textit{education} and \textit{age} are exactly the same as in the first additive model (as are the p-values). The VIFs are the same as well.
The interaction model with centered variables:

```r
> lm.out.4=lm(logHW~cent.educ+cent.age+cent.educ:cent.age)
> summary(lm.out.4)
```

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|)    |
|-----------|------------|---------|-------------|
| (Intercept) | 2.361884   | 0.046236 | 51.083 < 2e-16 *** |
| cent.educ  | 0.106834   | 0.022037 | 4.848 2.96e-06 *** |
| cent.age   | 0.015148   | 0.004141 | 3.658 0.000346 *** |
| cent.educ:cent.age | 0.005579 | 0.002054 | 2.717 0.007323 ** |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5869 on 158 degrees of freedom
Multiple R-squared: 0.2394, Adjusted R-squared: 0.2249
F-statistic: 16.58 on 3 and 158 DF, p-value: 2.058e-09

```r
> vif(lm.out.4)
```

<table>
<thead>
<tr>
<th>cent.educ</th>
<th>cent.age</th>
<th>cent.educ:cent.age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.073676</td>
<td>1.009647</td>
<td>1.076675</td>
</tr>
</tbody>
</table>

Much lower VIFs than in the interaction model fit with non-centered variables.
The model with $\text{AGE}^2$ (non-centered variables):

\begin{verbatim}
> lm.out.5=lm(logHW~EDUCL+AGE+AGESQ)
> summary(lm.out.5)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -1.7572186 | 0.5721705 | -3.071   | 0.00251 ** |
| EDUCL | 0.1195780 | 0.0207735 | 5.756 | 4.35e-08 *** |
| AGE | 0.1142559 | 0.0250803 | 4.556 | 1.04e-05 *** |
| AGESQ | -0.0012115 | 0.0003002 | -4.036 | 8.46e-05 *** |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5717 on 158 degrees of freedom
Multiple R-squared: 0.2782, Adjusted R-squared: 0.2645
F-statistic: 20.3 on 3 and 158 DF, p-value: 3.511e-11

> vif(lm.out.5)

    EDUCL    AGE    AGESQ
1.005442 39.023752 39.044383

Having $\text{AGE}$ and $\text{AGESQ}$ both in the model leads to high VIFs.
The model with $\text{AGE}^2$ (centered variables):

\begin{verbatim}
> lm.out.6=lm(logHW~cent.educ+cent.age+I(cent.age^2))
> summary(lm.out.6)

Coefficients:

Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  2.5043999  0.0585671  42.761  < 2e-16 ***
cent.educ    0.1195780  0.0207735   5.756   4.35e-08 ***
cent.age     0.0238467  0.0046614   5.116   8.95e-07 ***
I(cent.age^2) -0.0012115  0.0003002  -4.036   8.46e-05 ***  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.5717 on 158 degrees of freedom
Multiple R-squared: 0.2782, Adjusted R-squared: 0.2645
F-statistic: 20.3 on 3 and 158 DF,  p-value: 3.511e-11
\end{verbatim}

> vif(lm.out.6)

cent.educ  cent.age  I(cent.age^2)
1.005442  1.348028  1.346611

The centering of the variables greatly reduces the VIFs when we have a quadratic term entered into the model.
• If you have quantitative predictors, and are including interaction terms or polynomial terms (quadratics, cubics, etc.) in your model, it’s a good idea to center the predictor variables to reduce the effects of multicollinearity on your analysis.