Introduction to Statistical Modeling

Multicollinearity

Joris Vankerschaver

Multicollinearity

- There is **multicollinearity** when 2 or more predictors are correlated
- **Can possibly cause problems**: if there is strong correlation between 2 predictors X_1 and X_2, it becomes difficult to discern effect of X_1 of effect of X_2

Example: If $X_1 = X_2$, then

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + (\beta_1 + \beta_2) X_1$$

Consequences

- Numerically instable estimates
- Estimates with large standard errors
- Difficult interpretation of coefficients

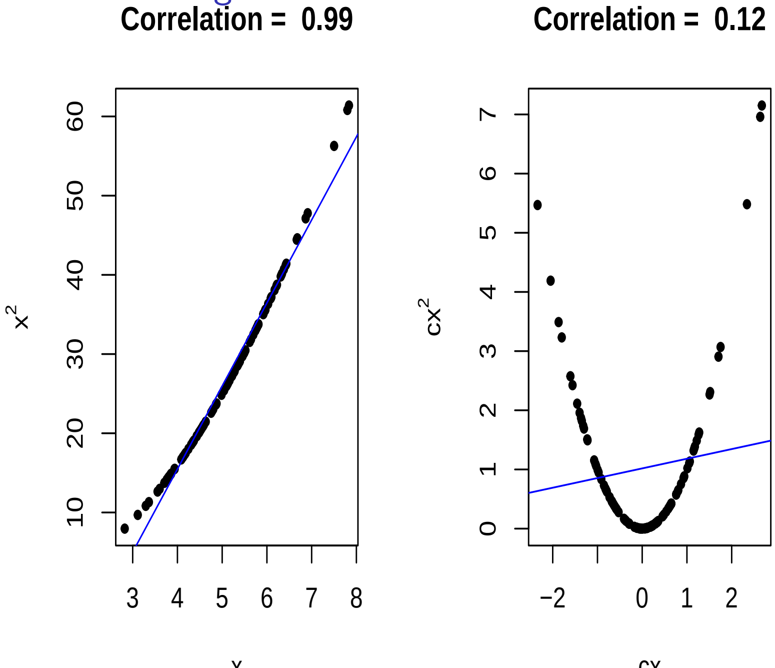
Diagnosing multicollinearity

Multicollinearity can be recognized through:

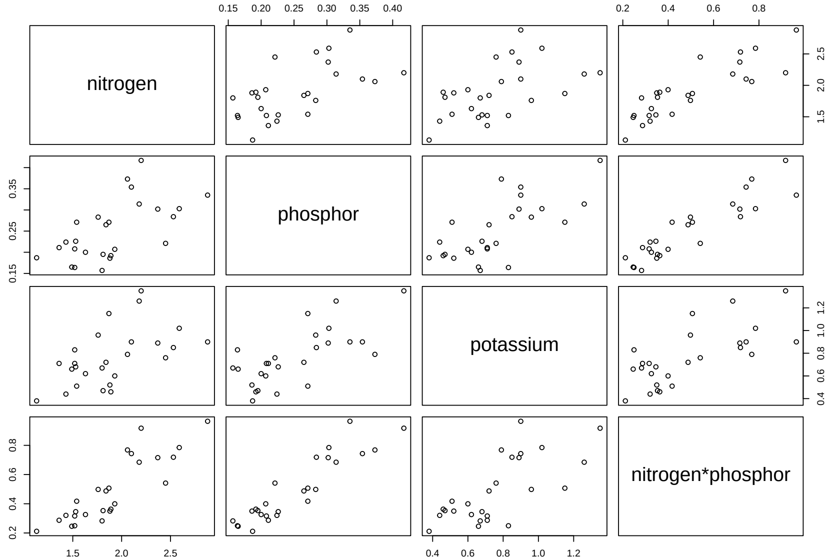
- **Instability**:
  - Large changes in coefficients after adding a predictor
  - Very wide confidence intervals
  - Unexpected results
- **Strong correlation** between predictors:
  - Example: usually strong correlation between X_f and $X_f X_s$
  - Can sometimes be eliminated by **centering** (subtracting the mean):

$$X \to X - \bar{X}.$$

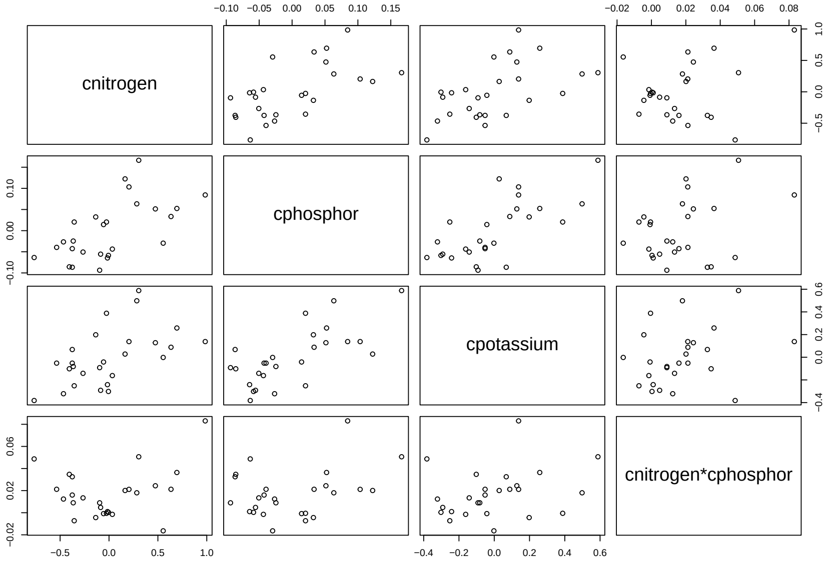
Impact of centering



Scatterplot matrix - before centering



Scatterplot matrix - after centering



Diagnosing multicollinearity

Previous diagnostics are limited:

- Even if pairwise correlations between predictors X_1, X_2, X_3 low, there can be strong multicollinearity.
- E.g., when strong correlation between X_1 and a linear combination of X_2 and X_3.

Variance inflation factor for k^{th} coefficient:

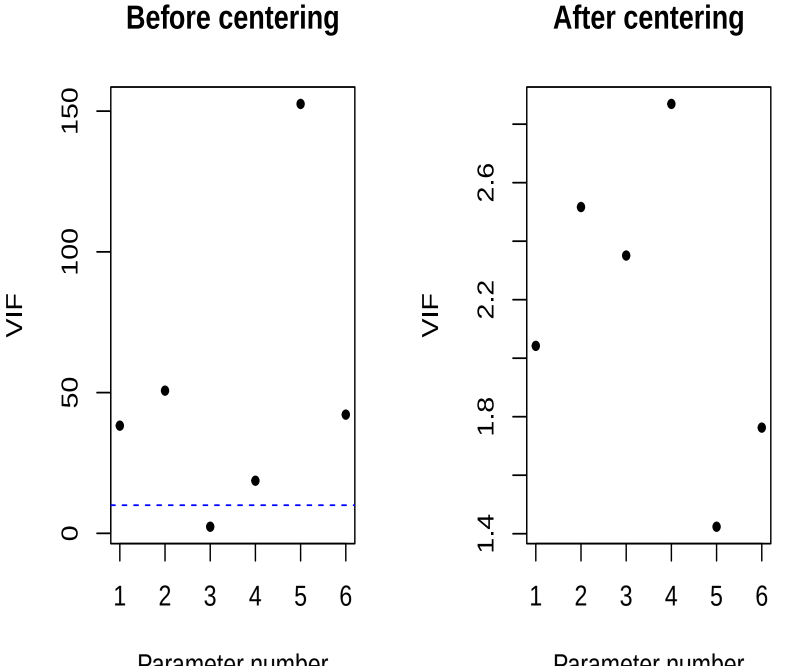
$$\mathrm{VIF}_k = \left(1 - R_k^2\right)^{-1}$$

with R_k^2 the R^2 of linear regression of k^{th} predictor on other predictors.

Interpretation VIF

- $\mathrm{VIF}_k \geq 1$; $\mathrm{VIF}_k = 1$ if k^{th} predictor not linearly associated with other predictors.
- Expresses how much larger variance on k^{th} coefficient is than when all predictors were independent.
- Average quadratic distance between estimated and true coefficients is proportionate with average VIF.
- Critical multicollinearity: maximum VIF of at least 10.

Variance inflation factors



Simpler interpretation of coefficients

Coefficients (without centering)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	160.66283	175.61424	0.9148622	0.370649894
nitrogen	-76.49677	92.34000	-0.8284250	0.416746264
phosphor	-1120.70470	711.42841	-1.5752881	0.130135986
potassium	138.06170	41.29966	3.3429260	0.003084272
nitrogen:phosphor	724.38231	353.05353	2.0517634	0.052870451

Coefficients (with centering)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	184.1200	9.244736	19.916194	4.079334e-15
cnitrogen	105.0167	21.458692	4.893901	7.703187e-05
cphosphor	252.5570	156.336392	1.615472	1.211339e-01
cpotassium	138.0617	41.299658	3.342926	3.084272e-03
cnitrogen:cphosphor	724.3823	353.053531	2.051763	5.287045e-02

Example: Prediction body fat

- Determining percentage body fat difficult and expensive
- Study investigates association between
  - Y: body fat
  - X_1: triceps skinfold thickness
  - X_2: thigh circumference
  - X_3: midarm circumference
- 20 healthy women between 25 and 34 years old

Analysis in R

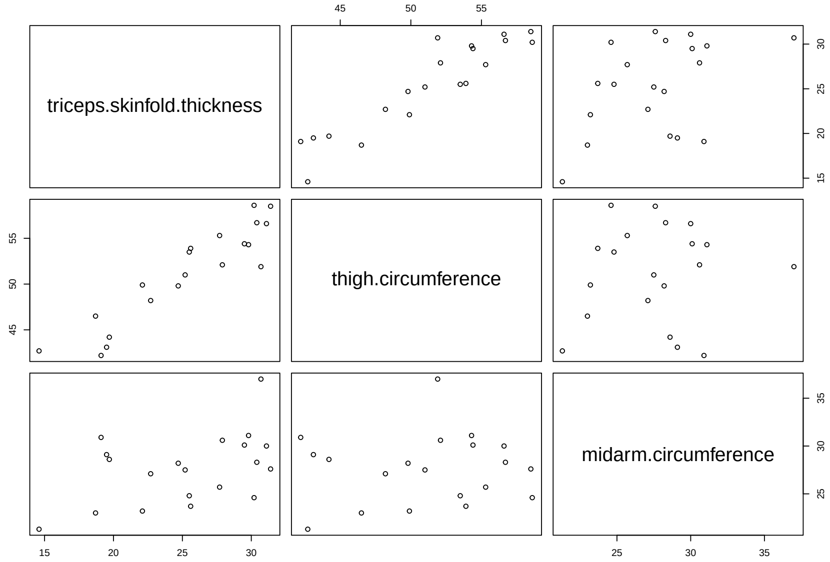
```
Call:
lm(formula = bodyfat ~ triceps.skinfold.thickness + thigh.circumference
    midarm.circumference)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 117.085     99.782   1.173    0.258
triceps.skinfold.thickness    4.334      3.016   1.437    0.170
thigh.circumference          -2.857      2.582  -1.106    0.285
midarm.circumference         -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared:  0.8014,    Adjusted R-squared:  0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Scatterplot matrix



Variance inflation factors

```
                              vif_bodyfat
triceps.skinfold.thickness       708.8429
thigh.circumference              564.3434
midarm.circumference             104.6060
```

- VIF on average 460.
- Large VIF for midarm circumference, although weakly correlated with other predictors.
- **How to correct for multicollinearity?**
  - Centering variables only valid option when higher order terms are in play.
  - Combine predictors, e.g., through principal component regression.
  - Ridge regression: allow some bias in exchange for increased precision and lower risk of overfitting.

Multicollinearity and confounding

- A lot of textbooks advise to remove predictors from model in case of multicollinearity
- However, multicollinearity can also indicate strong confounding!