Collinearity Among the Predictors

- We want the predictors to be highly correlated with the dependent variable.
- We do not want the predictors to be highly correlated with each other.
- Collinearity occurs when a predictor is "too" highly correlated with one or more of the other predictors.

$$\text{Variance of } B_j = \frac{1}{1-R_j^2}\left(\frac{MSE}{\sum\left(X_{ij}-\bar{X}_j\right)^2}\right) = VIF\left(\frac{MSE}{\text{Sum of Squares}_{jth\ \text{Predictor}}}\right)$$

$$SE(B_j) = \sqrt{\text{Variance of } B_j}$$

Impact of Collinearity

- The regression coefficients are very sensitive to minor changes in the data. Slight fluctuations in the data (e.g., due to sampling, measurement error, random error) may lead to substantial fluctuations in the sizes of the regression coefficients or even to changes in their signs.
- The regression coefficients have large standard errors, which lead to low power for the predictors.
- In the extreme case, singularity, you cannot calculate the regression equation. This is because with perfect collinearity, you would be trying to divide by zero.

Avoiding Accidental Collinearity in the Model

Do not put **all** subscale scores **and** the total score in the regression model. Omit either the total score or one of the subscale scores.

Total = Subscale1 + Subscale2 + Subscale3

Interaction terms and higher-order terms that are based on quantitative variables should be centered. The original variables should be untransformed.

$$X1$$
$$X2$$
$$X1X2 = \left(X1-\overline{X1}\right)\left(X2-\overline{X2}\right)$$
$$X1SQ = \left(X1-\overline{X1}\right)^2$$
$$X2SQ = \left(X2-\overline{X2}\right)^2$$

Detecting Collinearity

- **Tolerance (TOL)**

Tolerance tells us the amount of correlation between the predictor and all other remaining predictors. Tolerance values less than 0.10 are often considered to be an indication of collinearity. Low values of tolerance indicate the regression coefficients may be sensitive to rounding errors. Most statistical packages will not include variables into the analysis if their tolerance is lower than some pre-specified criterion. For example, variables with tolerance < .01 are not added to the model. You can override the default tolerance.

$$\begin{aligned}\text{Tolerance} &= 1 - R_X^2 \\ &= 1 - (\text{correlation between the predictor and all other predictors})^2\end{aligned}$$

- **Variance Inflation Factors (VIF)**

Variance inflation factors tell us the degree to which the standard error of the predictor is increased due to the predictor's correlation with the other predictors in the model. Values greater than 10 (or, Tolerance values less than 0.10) are often considered to be an indication of collinearity.

$$VIF = \frac{1}{TOLERANCE}$$

- The VIF will be large when the predictors are highly correlated among themselves.
- The larger the VIF, the larger the standard error of the regression coefficients in question. Large standard errors make the *t* ratio smaller and the confidence intervals wider, effectively decreasing power of the statistical test.
- While useful, the VIF are unable to determine the number of coexisting near dependencies.

- **Condition Indices and Variance Proportions**

Condition indices and variance proportions, simultaneously evaluated, tell us whether collinearity is a concern. If collinearity is a concern, they also tell us which predictors are "too" highly correlated.

Belsley (1991, p. 56) states:
'Weak Dependencies' have condition indices around 5-10.
'Strong Dependencies' have condition indices around 30-100.
Two or more variance proportions greater than 0.50 is 'large'.

If a high condition index is associated with two or more variance proportions that are above 0.50 then collinearity is indicated.

Handling Collinearity

Try one or more of the following:

- Convert all the predictors to *Z*-scores to minimize the effects of rounding errors. (This may not be sufficient.) Then, use the Z-scores in the linear regression analysis.

- Delete some of the predictors that are too highly correlated, but this may lead to model misspecification.

- Collect additional data…in the <u>hope</u> that additional data will reduce the collinearity.

- Use variable reduction techniques such as principal components analysis or factor analysis to consolidate the information contained in your predictors, but this may make interpreting the regression analysis more difficult. Another variable reduction technique would be to average some of the predictors.

- Use estimation methods other than OLS. For example, ridge regression or robust regression are resistant to collinearity, but these methods may lead to biased estimates.

Descriptive Statistics

	Mean	Std. Deviation	N
X1	10.6486	4.93362	74
X2	23.8649	14.20816	74
X3	51.4595	7.22664	74
X4	34.0676	3.52421	74
X5	29.9595	3.00429	74
X6	22.6351	2.56197	74
Y1	82.2162	16.10009	74

Correlations

		X1	X2	X3	X4	X5	X6	Y1
X1	Pearson Correlation	1	-.120	-.045	.486**	.101	.141	.288*
	Sig. (2-tailed)	.	.310	.704	.000	.393	.229	.013
	N	74	74	74	74	74	74	74
X2	Pearson Correlation	-.120	1	-.133	-.079	.084	.095	-.056
	Sig. (2-tailed)	.310	.	.257	.504	.475	.419	.638
	N	74	74	74	74	74	74	74
X3	Pearson Correlation	-.045	-.133	1	.750**	-.368**	-.394**	.696**
	Sig. (2-tailed)	.704	.257	.	.000	.001	.001	.000
	N	74	74	74	74	74	74	74
X4	Pearson Correlation	.486**	-.079	.750**	1	-.212	-.210	.797**
	Sig. (2-tailed)	.000	.504	.000	.	.070	.073	.000
	N	74	74	74	74	74	74	74
X5	Pearson Correlation	.101	.084	-.368**	-.212	1	.815**	-.430**
	Sig. (2-tailed)	.393	.475	.001	.070	.	.000	.000
	N	74	74	74	74	74	74	74
X6	Pearson Correlation	.141	.095	-.394**	-.210	.815**	1	-.473**
	Sig. (2-tailed)	.229	.419	.001	.073	.000	.	.000
	N	74	74	74	74	74	74	74
Y1	Pearson Correlation	.288*	-.056	.696**	.797**	-.430**	-.473**	1
	Sig. (2-tailed)	.013	.638	.000	.000	.000	.000	.
	N	74	74	74	74	74	74	74

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Zscore(X6), Zscore(X2), Zscore(X1), Zscore(X3), Zscore(X5), Zscore(X4)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Zscore(Y1)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.858[a]	.737	.713	.53529024

a. Predictors: (Constant), Zscore(X6), Zscore(X2), Zscore(X1), Zscore(X3), Zscore(X5), Zscore(X4)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	53.802	6	8.967	31.295	.000[a]
	Residual	19.198	67	.287		
	Total	73.000	73			

a. Predictors: (Constant), Zscore(X6), Zscore(X2), Zscore(X1), Zscore(X3), Zscore(X5), Zscore(X4)

b. Dependent Variable: Zscore(Y1)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	2.7E-15	.062		.000	1.000		
	Zscore(X1)	1.1E-02	.105	.011	.105	.917	.356	2.809
	Zscore(X2)	3.8E-02	.065	.038	.584	.561	.919	1.089
	Zscore(X3)	7.2E-02	.143	.072	.502	.617	.193	5.183
	Zscore(X4)	.675	.157	.675	4.290	.000	.159	6.307
	Zscore(X5)	-4.E-02	.109	-.039	-.358	.722	.333	3.003
	Zscore(X6)	-.277	.111	-.277	-2.508	.015	.321	3.111

a. Dependent Variable: Zscore(Y1)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Zscore(X1)	Zscore(X2)	Zscore(X3)	Zscore(X4)	Zscore(X5)	Zscore(X6)
1	1	2.415	1.000	.00	.00	.01	.02	.01	.03	.03
	2	1.545	1.250	.00	.09	.01	.00	.03	.03	.03
	3	1.000	1.554	1.00	.00	.00	.00	.00	.00	.00
	4	.980	1.570	.00	.01	.83	.01	.01	.00	.00
	5	.797	1.740	.00	.18	.08	.07	.00	.05	.03
	6	.183	3.630	.00	.00	.00	.00	.00	.88	.90
	7	7.868E-02	5.541	.00	.71	.07	.89	.95	.00	.00

a. Dependent Variable: Zscore(Y1)

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X6, X2, X1, X3, X5, X4[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.858[a]	.737	.713	8.61822

a. Predictors: (Constant), X6, X2, X1, X3, X5, X4

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13946.201	6	2324.367	31.295	.000[a]
	Residual	4976.340	67	74.274		
	Total	18922.541	73			

a. Predictors: (Constant), X6, X2, X1, X3, X5, X4

b. Dependent Variable: Y1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	13.184	17.225		.765	.447		
	X1	3.590E-02	.343	.011	.105	.917	.356	2.809
	X2	4.327E-02	.074	.038	.584	.561	.919	1.089
	X3	.160	.318	.072	.502	.617	.193	5.183
	X4	3.084	.719	.675	4.290	.000	.159	6.307
	X5	-.208	.582	-.039	-.358	.722	.333	3.003
	X6	-1.742	.694	-.277	-2.508	.015	.321	3.111

a. Dependent Variable: Y1

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	X1	X2	X3	X4	X5	X6
1	1	6.577	1.000	.00	.00	.00	.00	.00	.00	.00
	2	.259	5.039	.00	.05	.70	.00	.00	.00	.00
	3	.127	7.199	.00	.32	.21	.00	.00	.00	.00
	4	3.097E-02	14.573	.00	.00	.03	.04	.01	.02	.04
	5	3.046E-03	46.465	.78	.00	.01	.15	.00	.01	.22
	6	1.985E-03	57.557	.07	.01	.00	.01	.00	.97	.74
	7	1.149E-03	75.650	.15	.62	.04	.79	.99	.00	.00

a. Dependent Variable: Y1