5. Multicollinearity

A topic specific to multiple regression models is multicollinearity (collinearity). Recall the assumption that the rank of $\boldsymbol{X}$ is equal to k*. That is, the k columns of matrix $\boldsymbol{X}$ (independent variables) are linearly independent from each other: The violation of this assumption is called **perfect/exact** multicollinearity, i.e. existence of linear relationships among the x-variables: $c_1x_1 + c_2x_2 + ... + c_kx_k = 0$ for constant values $c_1, c_2, ..., c_k$ not all equal to zero. The implication is that $\boldsymbol{X}$ and similarly $\boldsymbol{X'X}$ are singular and $\boldsymbol{(X'X)^{-1}}$ does not exist. Hence, the LS estimator $\boldsymbol{b = (X'X)^{-1} X'y}$ is not defined.

Perfect multicollinearity can rarely happen by chance in a sample, but may be incorporated by a wrong design. On the other hand, in a multiple regression model two or more independent variables may be highly correlated with each other. This phenomenon which is referred to as **imperfect/near** multicollinearity is what we study here. Note that in designed experiments with multiple independent variables, researchers usually choose the variables so that there is no multicollinearity. In observational studies, however, it is nearly always the case that the independent variables will be correlated. The question is how much the model can **tolerate** it.

Sources of Multicollinearity

There are various sources for multicollinearity.

- Data collection (sampling) method. For example, in the data collection phase an investigator may have drawn the data from such a narrow subspace of the independent variables that multicollinearity appears.
- Physical constraints, such as design limits, may also impact the range of some of these independent variables.
- Model specification such as too many higher-ordered terms/interactions and outliers can lead to collinearity.

When there is no multicollinearity, the effects of the individual predictors can be estimated independently of each other. When multicollinearity is present, the estimated coefficients are correlated (confounding) with each other.

---

* Clearly, rank of $\boldsymbol{X}$ is not equal to k if $n \leq k$ (micronumerosity), but we can rule out this case.

Consequences of Multicollinearity

- The LS estimators of $\boldsymbol{\beta}$ remain **BLUE** as (near) multicollinearity does not violates the classical assumptions. But they have large variances and covariances, making the estimation imprecise.
- Inflated standard errors of the regression coefficient estimates, wider CI's, deflated t-tests for significance testing, leads to false non-significance of coefficients and degradation of model predictability.
- Despite insignificance of many coefficients, the R^2 can be very high.

Large Variance-Covariance

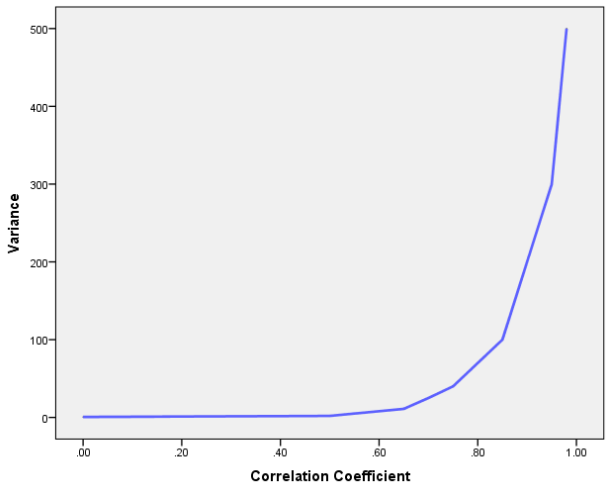
To show this ramification, consider a model with two regressors, $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, for which the variances and covariance of estimates are defined as:

$$Var(b_2) = \frac{\sigma^2}{(1 - r^2_{x_2 x_3}) \sum (x_2 - \bar{x})^2}$$

$$Var(b_3) = \frac{\sigma^2}{(1 - r^2_{x_2 x_3}) \sum (x_3 - \bar{x})^2}$$

$$Cov(b_2, b_3) = \frac{-r_{2,3}\sigma^2}{(1 - r^2_{x_2 x_3}) \sqrt{\sum (x_2 - \bar{x})^2 \sum (x_3 - \bar{x})^2}}$$

where $r_{x_2 x_3}$ is the Pearson's correlation coefficient between x_2 and x_3. The degree of inflated variances and collinearity can typically be demonstrated by the following graph.



Detecting Multicollinearity

We introduce three primary ways for detecting multicollinearity.

Method 1: Pairwise (Matrix) Scatterplots and Correlation

We can visually inspect the data by doing pairwise scatterplots of the independent variables. So if you have (k – 1) independent variables, then you should inspect all $\binom{k-1}{2}$ pairwise scatterplots together with the matrix of correlation coefficients, looking for any plots/values that seem to indicate a linear relationship between pairs of independent variables (*simple*/*zero-order*) correlation.

The problem with this criterion is that although high zero-order correlation may suggest collinearity the converse is not necessarily true. That is to say high zero-order correlations are a *sufficient* condition and not a *necessary* condition. For this reason, inspecting *partial* and *part* coefficients is also suggested.

A Note on Part and Partial Correlation

We define these concepts in the context of a two-regressor linear regression.

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

The generalisation to the higher cases is straightforward.

1. **Zero-order correlation:** These are (simple) correlation coefficients between the three variables: $r_{yx_2}, r_{yx_3}, r_{x_2x_3}$. Notice that r_{yx_2}, for example, is supposed to measure the linear association between y and x_2, but that is contaminated by the presence of x_3 in the model.
2. **Partial Correlation:** The correlation between the dependent variable and an independent variable when the linear effects of the other independent variables in the model have been removed from both:

$r_{yx_2.x_3} = \dfrac{r_{yx_2} - r_{yx_3} r_{x_2x_3}}{\sqrt{(1-r_{yx_3}^2)(1-r_{x_2x_3}^2)}}$, the partial correlation between y and x_2, holding x_3 constant.

$r_{yx_3.x_2} = \dfrac{r_{yx_3} - r_{yx_2} r_{x_2x_3}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_2x_3}^2)}}$, the partial correlation between y and x_3, holding x_2 constant.

$r_{x_2x_3.y} = \dfrac{r_{x_2x_3} - r_{yx_2} r_{yx_3}}{\sqrt{(1-r_{x_2x_3}^2)(1-r_{yx_3}^2)}}$, the partial correlation between x_2 and x_3, holding y constant.

These are also called the *first-order* correlation coefficients. (Therefore, $r_{yx_2.x_3x_4}$ is the *second-order* correlation coefficient, and so on.)

3. **Part Correlation:** Sometimes called the *semi-partial* correlation, is the correlation between the dependent variable and an independent variable when the linear effects of the other independent variables in the model have been removed from the independent variable. It is related to the change in R^2 when a variable is added to an equation.

$$sr_2 = \frac{r_{yx_2} - r_{yx_3} r_{x_2x_3}}{\sqrt{(1 - r_{yx_3}^2)}}$$

$$sr_3 = \frac{r_{yx_3} - r_{yx_2} r_{x_2x_3}}{\sqrt{(1 - r_{yx_2}^2)}}$$

- The more "tolerant" a variable is (i.e. the less highly correlated it is with the other regressors), the greater its unique contribution to R^2 will be.
- Once one variable is added or removed from an equation, all the other semi-partial correlations can change. The semi-partial correlations only tell you about changes to R^2 for one variable at a time.
- Semi-partial correlations are used in "Stepwise Procedures" which will be discussed later in the course

Method 2: Variance Inflation Factor

To use a measure of multicollinearity called the variance inflation factor (*VIF*). This is defined as:

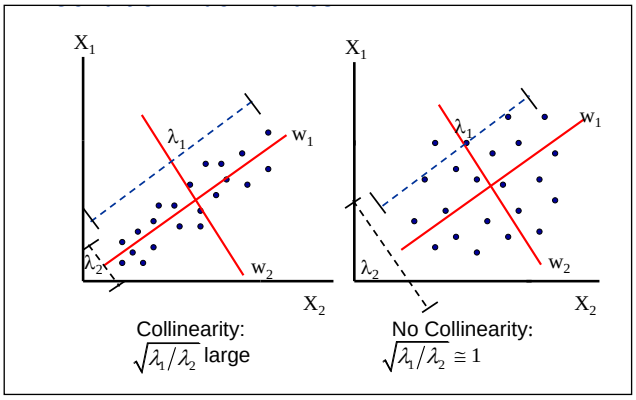
$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination obtained by regressing x_j on the remaining independent variables. A common rule of thumb is that if $VIF_j = 1$, then there is no multicollinearity, if $1 < VIF_j < 5$ then there is possibly some moderate multicollinearity and if $VIF_j \geq 5$ then there is a strong indication of a collinearity problem. Sometimes, the tolerance is also reported. The tolerance is simply the inverse of the *VIF*.

Method 3: Eigenvalue Methods

The third method for identifying potential multicollinearity concerns a variety of measures using eigenvalues and eigenvectors.

Recall that multicollinearity occurs when linear combinations of some of the columns in the ***X*** matrix equal zero, or nearly zero. Geometrically this occurs when at least one dimension of the ***X***-space has very little dispersion (shown in the left graph below). When an independent variable has limited dispersion, its column in the ***X*** matrix will almost be a multiple of a vector of ones, with the result that the variable will be nearly collinear with the column for the intercept.



The presence of collinearity is detected by *singular decomposition* of ***X*** or the *eigenanalysis* of $\boldsymbol{X'X}$.

A value λ is called the eigenvalue of the correlation matrix $\boldsymbol{X'X}$ if there is a nonzero vector ***z*** such that $(\boldsymbol{X'X})\boldsymbol{z} = \lambda \boldsymbol{z}$. The nonzero vector ***z*** is called the *eigenvector*.

A set of eigenvalues (λs) of relatively equal magnitudes indicates little multicollinearity, while a wide variation in magnitudes indicates severe multicollinearity. Therefore, the ratio of the eigenvalues can be useful for examining multicollinearity. More formally, a measure of the overall multicollinearity of the variables can be obtained by computing what is called the **condition index** of the correlation matrix and is defined as $\sqrt{\lambda_{\max} / \lambda_{\min}}$. Obviously this quantity is always greater than 1, so a large number is indicative of collinearity. Empirical evidence suggests that a value less than 30 typically means weak collinearity, values between 30 and 100 is evidence of moderate collinearity, while anything over 100 is evidence of strong collinearity. Condition numbers for the individual predictors can also be calculated by $\sqrt{\lambda_{\max} / \lambda_j}, (j = 2, ..., k)$.

Remedial Measures

As in the case of detection, there are no definite guides because multicollinearity is essentially a sample problem. Although the following measures may be applied, but the success depends on how serious multicollinearity is.

- Adding new data.
- Combining cross-sectional and time series data (pooling data).
- Removing violating variable(s) from the model, but be aware of *specification bias*!
- Using the deviation forms (centring) of regressors, in polynomial models.
- Using biased regression techniques such as *Ridge Regression*. The basic idea behind ridge regression is to reduce the variances of the parameter estimates by considering a substitute, non-singular matrix $\boldsymbol{X'X} + c\boldsymbol{I}$, where c is usually a small positive quantity sometimes referred to as a *shrinkage parameter*. The result is a biased LS estimator $\boldsymbol{b}^* = (\boldsymbol{X'X} + c\boldsymbol{I})^{-1}\boldsymbol{X'y}$. The choice of c is a compromise between decreasing variance and increasing bias.

Note: In general, multicollinearity may not be problematic if the objective is prediction.

Example 5.1

In a study to investigate the causes for high diastolic blood pressure, researchers observed the following variables based on 20 individuals.

Blood pressure (BP, mm Hg)

Age (years)

Weight (kg)

Body surface area (*BSA*, m^2)

Duration of hypertension (years)

Basal pulse (Pulse, beats per minute)

Stress index

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.998[a]	.996	.994	.407

a. Predictors: (Constant), Stress, BSA, Duration, Age, Pulse, Weight

ANOVA

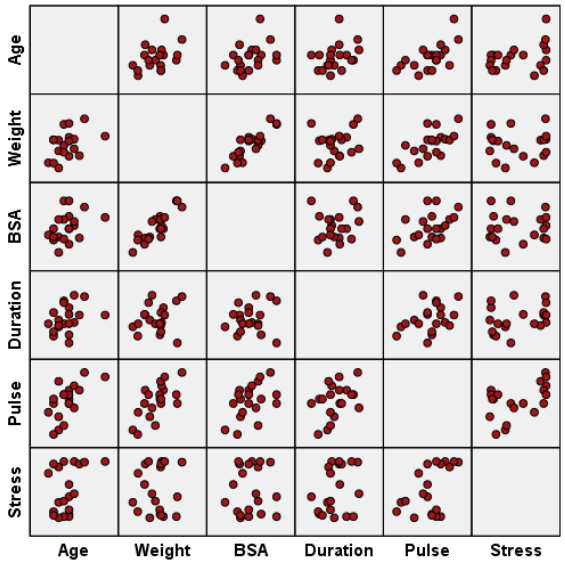
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	557.844	6	92.974	560.641	.000
	Residual	2.156	13	.166		
	Total	560.000	19			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.870	2.557		-5.034	.000
	Age	.703	.050	.324	14.177	.000
	Weight	.970	.063	.767	15.369	.000
	BSA	3.776	1.580	.095	2.390	.033
	Duration	.068	.048	.027	1.412	.182
	Pulse	-.084	.052	-.059	-1.637	.126
	Stress	.006	.003	.038	1.633	.126

a. Dependent Variable: BP

First impression: Very high R^2 but few significant regressors indicate multicollinearity.



Correlations							
	BP	Age	Weight	BSA	Duration	Pulse	Stress
Age	.659						
Weight	.950	.407					
BSA	.866	.378	.875				
Duration	.293	.344	.201	.131			
Pulse	.721	.619	.659	.465	.402		
Stress	.164	.368	.034	.018	.312	.506	

Comments: Excessive correlation between BSA and Weight, also Pulse and Weight.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-12.870	2.557		-5.034	.000					
	Age	.703	.050	.324	14.177	.000	.659	.969	.244	.567	1.763
	Weight	.970	.063	.767	15.369	.000	.950	.974	.264	.119	**8.417**
	BSA	3.776	1.580	.095	2.390	.033	.866	.552	.041	.188	**5.329**
	Duration	.068	.048	.027	1.412	.182	.293	.365	.024	.808	1.237
	Pulse	-.084	.052	-.059	-1.637	.126	.721	-.413	-.028	.227	**4.414**
	Stress	.006	.003	.038	1.633	.126	.164	.413	.028	.545	1.835

a. Dependent Variable: BP

Comments:

- VIF: greater than 5 for BSA and Weight indicate strong collinearity.
- Correlations: Simple correlation coefficients for BSA, Weight and Pulse are high, but part correlations are reduced drastically.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Age	Weight	BSA	Duration	Pulse	Stress
1	1	6.656	1.000	.00	.00	.00	.00	.00	.00	.00
	2	.268	4.984	.00	.00	.00	.00	.00	.00	.55
	3	.071	9.654	.00	.00	.00	.00	.93	.00	.05
	4	.003	50.067	.10	.09	.01	.17	.00	.00	.03
	5	.001	77.397	.37	.78	.01	.01	.03	.01	.04
	6	.001	**83.751**	.29	.04	.01	.07	.04	.44	.14
	7	.000	**201.496**	.23	.09	**.98**	**.74**	.00	**.55**	.19

a. Dependent Variable: BP

The "Variance Proportion" shows the proportion of the variation of each regressor's coefficient attributed to each eigenvalue. We look for regressors that have high VP on the same small eigenvalue (generally, bottom few rows), as this indicates that variances of coefficients are dependent.

Comments: Once again BSA, Weight and Pulse are flagged here.

Weight is dropped:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	6.212	9.429		.659	.521					
	Age	.563	.206	.259	2.737	.016	.659	.590	.199	.587	1.703
	BSA	24.554	3.452	.617	7.114	.000	.866	.885	.516	.700	1.428
	Duration	.077	.204	.030	.376	.713	.293	.100	.027	.808	1.237
	Pulse	.456	.159	.320	2.866	.012	.721	.608	.208	.424	2.361
	Stress	-.017	.013	-.114	-1.284	.220	.164	-.325	-.093	.665	1.503

a. Dependent Variable: BP

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Age	BSA	Duration	Pulse	Stress
1	1	5.669	1.000	.00	.00	.00	.00	.00	.01
	2	.258	4.691	.00	.00	.00	.00	.00	.69
	3	.069	9.046	.00	.00	.00	.93	.00	.05
	4	.002	47.783	.12	.06	.87	.00	.00	.02
	5	.001	72.964	.74	.72	.01	.04	.01	.09
	6	.001	78.856	.14	.22	.11	.03	.99	.15

a. Dependent Variable: BP

BSA is dropped:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-15.117	2.749		-5.499	.000					
	Age	.732	.056	.337	13.154	.000	.659	.962	.262	.603	1.660
	Weight	1.099	.038	.869	29.093	.000	.950	.992	.579	.443	2.256
	Duration	.064	.056	.025	1.145	.271	.293	.293	.023	.809	1.236
	Pulse	-.137	.054	-.096	-2.551	.023	.721	-.563	-.051	.278	3.600
	Stress	.007	.004	.051	1.934	.074	.164	.459	.038	.575	1.740

a. Dependent Variable: BP

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Age	Weight	Duration	Pulse	Stress
1	1	5.671	1.000	.00	.00	.00	.00	.00	.00
	2	.257	4.694	.00	.00	.00	.00	.00	.59
	3	.069	9.057	.00	.00	.00	.93	.00	.04
	4	.001	65.198	.11	.48	.21	.00	.10	.00
	5	.001	72.506	.84	.42	.00	.06	.06	.10
	6	.000	108.009	.05	.09	.79	.01	.85	.26

a. Dependent Variable: BP

Deleting either Weight or BSA reduces collinearity. Pulse seem to be also collinear with Weight but dropping it may lead to specification error.

Practical Week 5

The folder **cars.sav** contains data that was collected to examine the prices of cars (*Pace New Car and Truck 1993 Buying Guide*). The variables in the data set are

Manufacturer	name of the manufacturer
Model	name of the model
Type	type of vehicle (Compact, Large, Midsize, Small, Sporty, or Van)
Price	average price of the car
Citympg	average city miles per gallon
EngineSize	engine displacement size (in liters)
Horsepower	maximum horsepower
Weight	weight of the vehicle (in pounds)

Generate a polynomial linear regression model with **LogPrice** as the dependent variable and **Citympg**, **Citympg^2, EngineSize**, **Horsepower**, **Horsepower^2**, **Weight** as the regressors. Comment on your results and see if multicollinearity is a problem for this model?