Multicollinearity and its Remediation

Objective:

- To understand how collinearity affects the Regression Model.
- Detection of Collinearity.
- Remediation of Collinearity.

Problem Related to Collinearity

- If the explanatory variables are correlated with each other then the least square coefficients are not uniquely defined.
- Coefficient standard errors are high which purports imprecision in the estimation of the least square coefficients.

Detecting Collinearity:

If there is perfect linear relationship between Xs then

$$c_1X_{i1} + c_2X_{i2} + \ldots\ldots c_kX_{ik} = c_0$$

Where all the $c_1, c_2 \ldots\ldots\ldots\ldots c_k$ are not all zeros. Since the columns of X are perfectly collinear therefore the regressor subspace is of deficient dimension therefore

1) The least square normal equations do not have a unique solution
2) Sampling variances of regression coefficients are infinite.

The sampling variance of the Least square slope coefficient B_j is

$$V(B_j) = \frac{1}{1-R_j^2} * \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$$

R_j^2 is the squared multiple correlation for the regression of X_j on the other X_s.

And

$$S_j^2 = \frac{\sum(X_{ij} - \overline{X_j})^2}{(n-1)}$$ is variance of X_j.

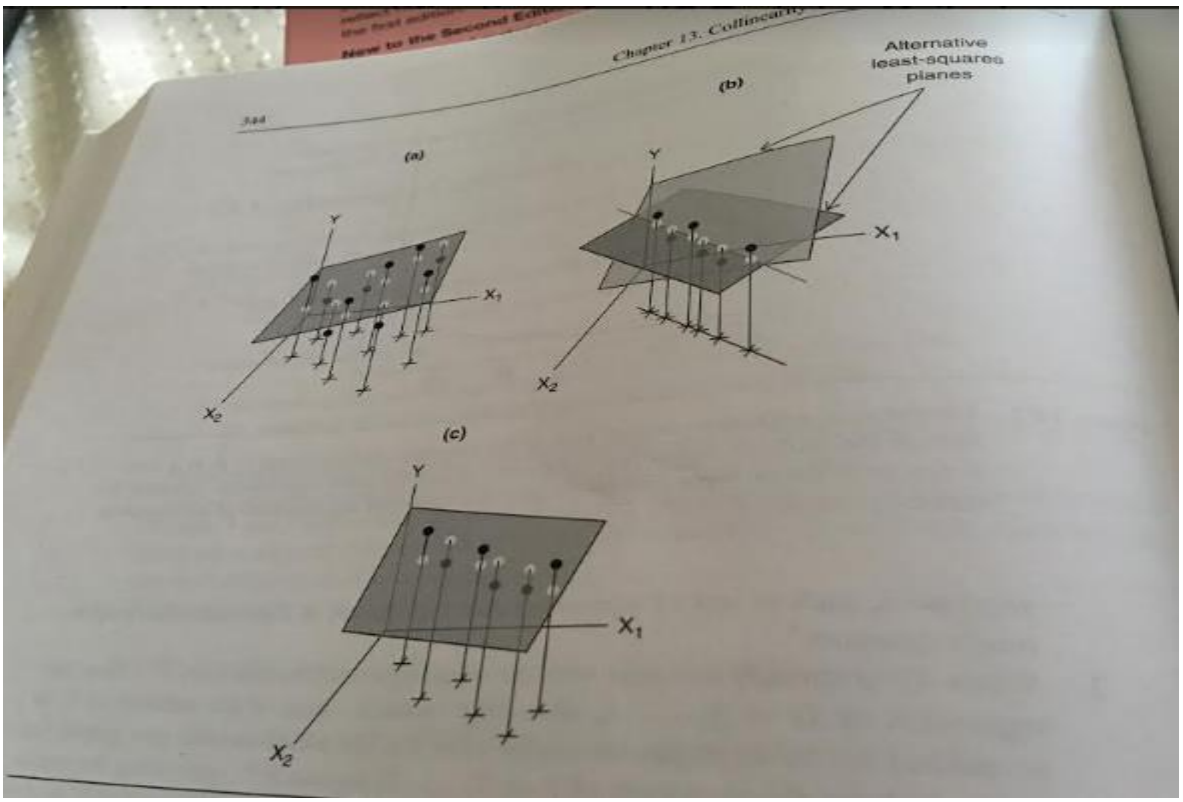
The term $\frac{1}{1-R_j^2}$ is called the variance inflation factor.

The VIF or its square root diagnoses the collinearity between variables.

VIF is not applicable to dummy variables or polynomial regressors.

The confidence intervals will also be wider if the sampling SD ($\sqrt{V(B_j)}$) is large.

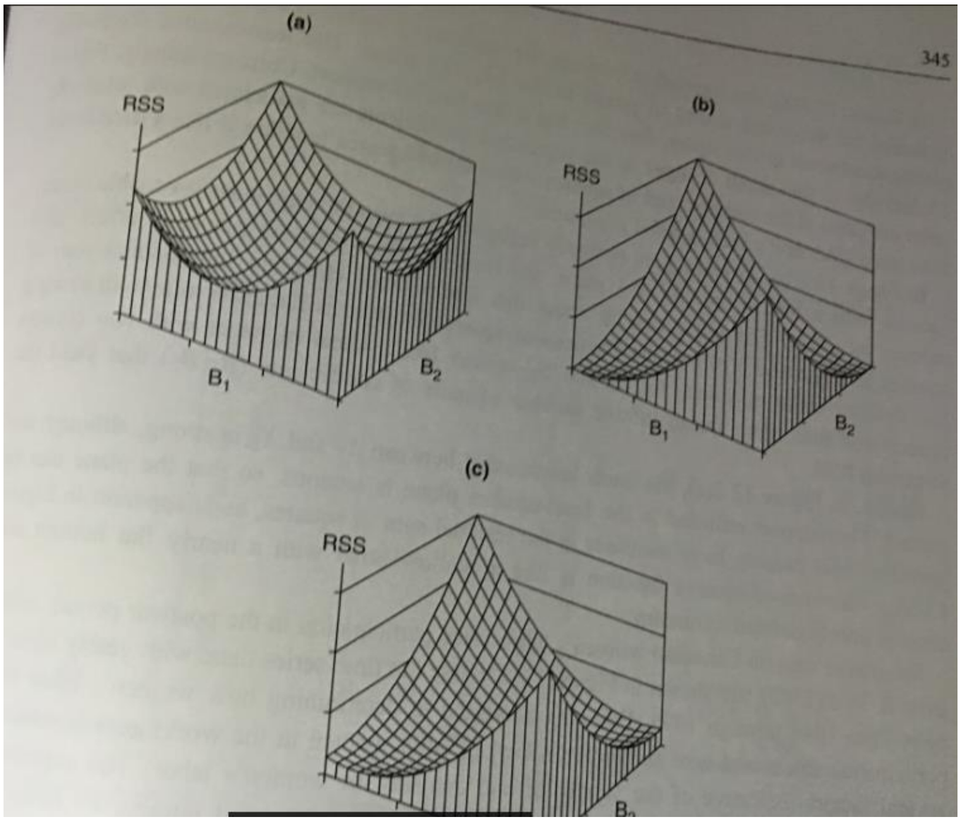
Due to collinearity or multicollinearity the least square coefficients are not useful estimator. The correlation relationships are not necessarily just pairs but can be in three or more.



If we use two explanatory variables X_1 and X_2 then in the above figure the gray dots and black dots represent the data. Gray are below the regression plane and black dots are above the regression plane. The white dots are fitted values on the plane. The x marks are projection of the data points onto the $\{X_1,X_2\}$ plane.

a) The correlation between X_1 and X_2 is very small. The regression plane is well supported by the data points.

b) The correlation between X_1 and X_2 is perfect. The correlation plane is not uniquely defined.
c) The correlation between X_1 and X_2 is significant. The regression plane is not well supported by the data points.



a) Graphing the residual sum of squares function RSS with coefficient B_1 and B_2. The correlation between X_1 and X_2 is very small. The Residual sum of squares has a well defined minimum.
b) The correlation between X_1 and X_2 is perfect. The minimum is flat above a line in the plane $\{B_1,B_2\}$. The coefficients B_1 and B_2 are not unique.
c) The correlation between X_1 and X_2 is significant. RSS is almost flat. The coefficients B_1 and B_2 are different from a)

Remedies of Collinearity:

1) PCA
2) Stepwise Regression Backwards and Forward using AIC Criterion.
3) Partial Least Squares
4) Ridge Regression

PCA or PLS

http://www.milanor.net/blog/performing-principal-components-regression-pcr-in-r/
https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Ridge Regression and Lasso

http://ricardoscr.github.io/how-to-use-ridge-and-lasso-in-r.html
http://www.thefactmachine.com/ridge-regression/
https://jamesmccammon.com/2014/04/20/lasso-and-ridge-regression-in-r/
http://www.milanor.net/blog/cross-validation-for-predictive-analytics-using-r/