Why is multicollinearity a problem?

If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem as the regression coefficient estimators are still **unbiased**. The predictions will still be accurate, and the overall R^2 (or $\mathrm{R}^2_{\text{adjusted}}$ /$\mathrm{R}^2_{\text{predicted}}$) will quantify how well the model predicts the Y values and will be close to each other.

But, if the goal is how the various X variables impact Y, then multicollinearity is a big problem. One problem, as discussed earlier, is that multicollinearity increases the standard errors of the coefficients. Increased standard errors may lead to an important predictor to become insignificant, whereas without multicollinearity and with lower standard errors, these same coefficients would have been significant.

The other problem is that due to the presence of multicollinearity, confidence intervals on the regression coefficients becomes very wide. The confidence intervals may even include zero, which means one can't even be confident whether an increase in the X value is associated with an increase, or a decrease, in Y.

Detecting multicollinearity

Multicollinearity can be detected by looking at the correlations among pairs of predictor variables. If they are large, we can conclude that the variables are collinear.

Looking at correlations only among pairs of predictors, however, is limiting. It is possible that the pair wise correlations are small, and yet a linear dependence exists among three or even more variables. That's why many regression analysts often rely on what are called **variance inflation factors** (*VIF*) to help detect multicollinearity, which are basically the diagonal elements of C^*.

It can be shown that, if some of the predictors are correlated with the predictor x_k then the variance of b_k is inflated and the same is given by

$$\mathrm{Var}(\hat{b}_k) = \sigma^2 C_{kk}^* = \sigma^2 \times \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2-value of the model obtained by regressing the k^{th} predictor on the remaining (p-1) predictors. Above shows that the variance of b_k is inflated by the factor $1/(1-R_k^2)$ because of x_k's linear dependence with other predictors and hence the name. So, formally VIF is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

Note that, the greater the linear dependence of the predictor x_k with other predictors, the larger the R_k^2 value. And, as the above formula suggests, the larger the R_k^2 value, the larger will be the corresponding VIF. If $R_k^2 = 0$, then corresponding VIF will be 1, which is the minimum possible value of VIF. It may be noted that VIF exists for each of the predictor variables in a multiple regression model.

The general rule of thumb is that VIFs exceeding 4 (i.e. $R_k^2 > 0.75$) warrant further investigations, while VIF exceeding 10 (i.e. $R_k^2 > 0.9$) is sign of serious multicollinearity and taken as an indication that the multicollinearity may be unduly influencing the least squares estimates.

Dealing with Multicollinearity

There are multiple ways to overcome the problem of multicollinearity.

- ✓ One may use ridge regression or principal component regression or partial least squares regression.
- ✓ The alternate way could be to drop off variables which are resulting in multicollinearity. One may drop of variables which have VIF more than 10.
- ✓ If two or more variables have multicollinearity in the range of 10 or more, remove those variables that have least/low correlation/ impact with the response variable.

Influential Observations

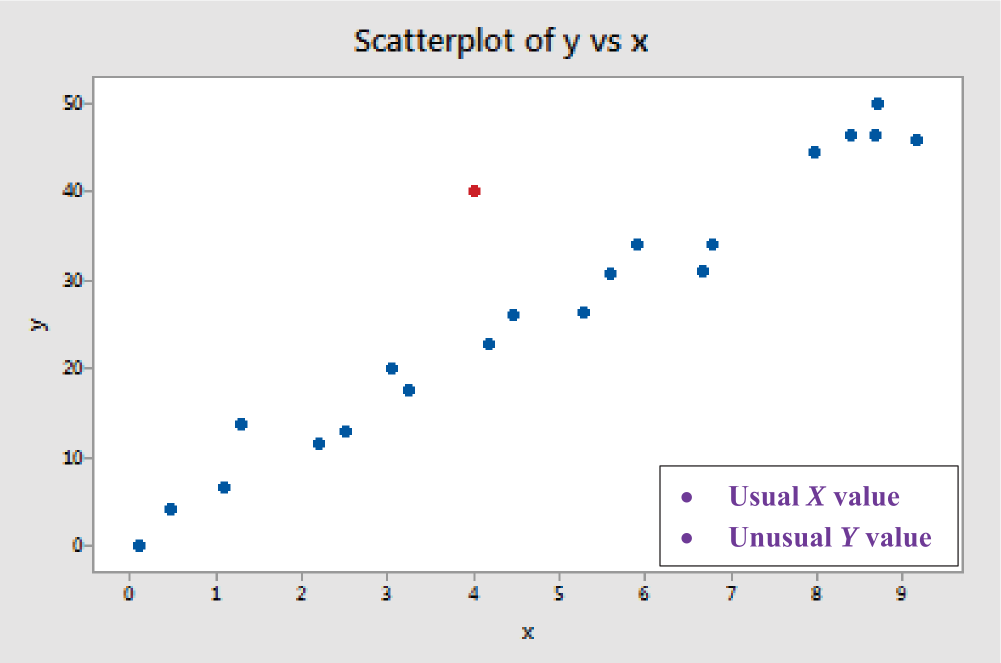
The influence of an observation can be thought of in terms of how much the predicted values for other observations would differ if the observation in question were not included. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Outliers

An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.

- ▪ An observation whose response value is unusual given its values on the predictor variables (X), resulting in large residual, or error in prediction.

- An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

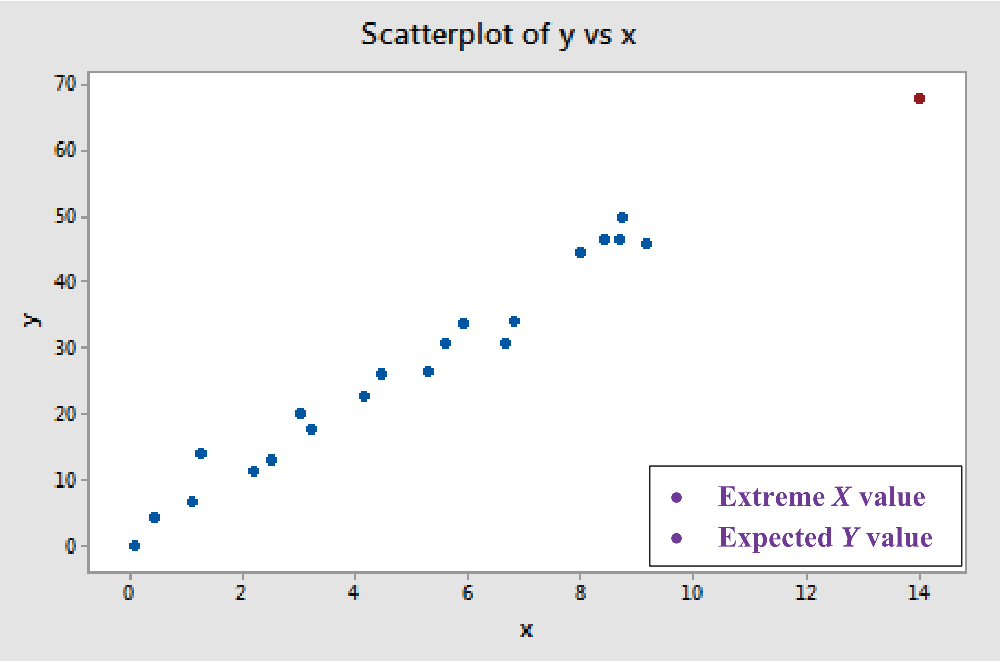


In this case, the red data point, though have a usual X value, but have an unusual Y value and hence will result in a large residual.

Leverage

A data point has high leverage if it has an extreme **predictor value, i.e. *X*-values**.

- Leverage is a measure of how far a predictor variable deviates from its mean value.
- These leverage points may or may not have an effect on the estimate of regression coefficients.



In this case, the red data point does follow the general trend of the rest of the data. Therefore, it is not deemed an outlier here. However, this point does have an extreme x value, so it does have high leverage.

- Regression Equation **excluding** the High Leverage point:

$$y = 0.0513 + 5.2624\,x$$

- Regression Equation **including** the High Leverage point:

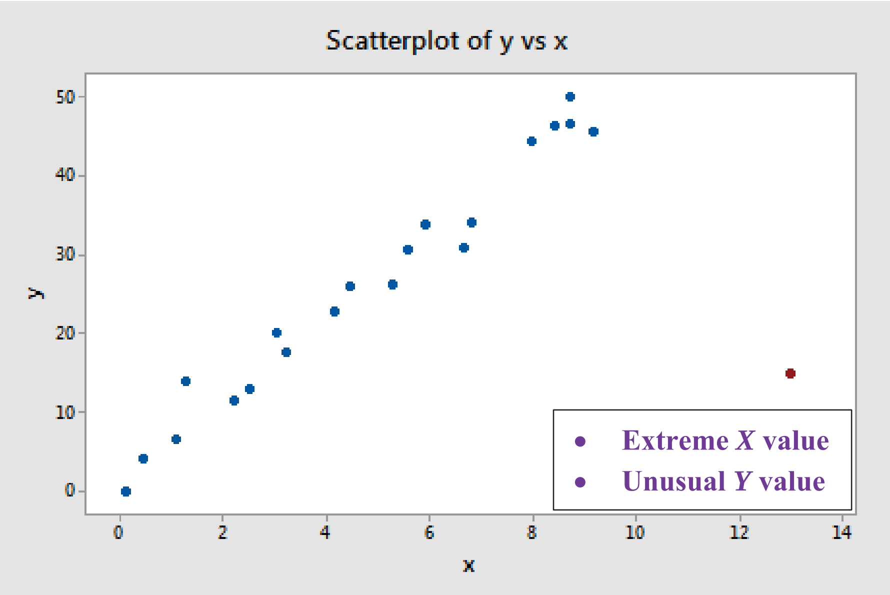
$$y = 0.9092 + 5.0523\,x$$

Influence

When an observation has ***high leverage*** (in terms of X-value) and is an ***outlier*** (in terms of Y-value) it will strongly influence the regression line.

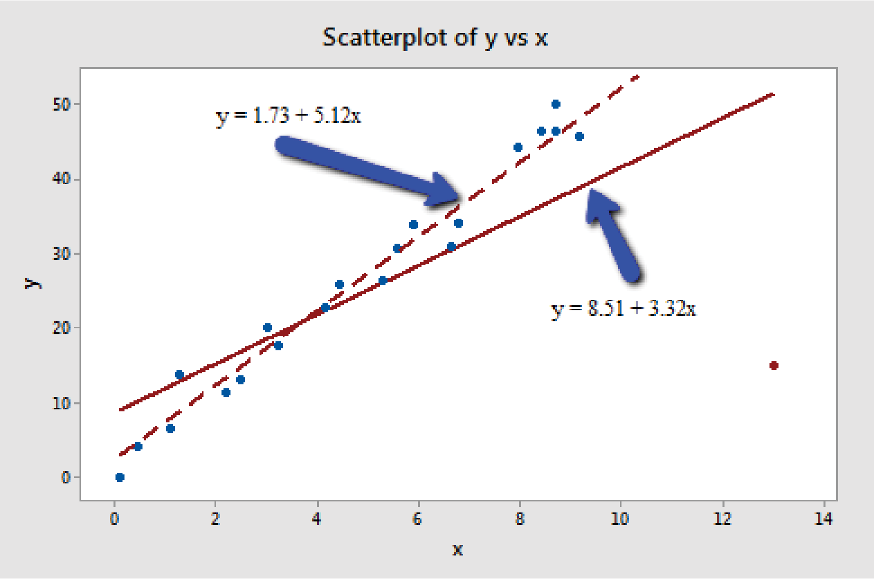
In other words, it must have an unusual X-value with an unusual Y-value *given* its X-value. In such cases both the intercept and slope are affected, as ***the line chases the observation***.

- Influence can be thought of as the product of leverage and error in prediction. **Influence = Leverage X Residual**.
- Removing the observation substantially changes the estimate of coefficients.



In this case, the red data point is most certainly an outlier and has high leverage! The **red** data point does not follow the general trend of the rest of the data and it also has an extreme x value. And, in this case the red data point is influential.

The two best fitting lines — one obtained when the red data point is included and the other obtained when the red data point is excluded:



Effect of Leverage

The greater an observation's leverage, the more potential it has to be an influential observation. For example, an observation with *X*-value equal to the mean on the predictor variable has no influence on the slope of the regression line. On the other hand, an observation that has an unusual *X* value has the potential to affect the slope greatly.

A data point that has an unusual *X* value is known as a **Leverage Point**. The **diagonal elements** $\boldsymbol{h_{ii}}$ of the hat matrix have some useful property: their values are always between 0 and 1, i.e. $0 \leq h_{ii} \leq 1$ and **their sum is** $\boldsymbol{P}$, the number of parameters estimated (including the intercept), i.e. $P = p+1$.

These H values are functions only of the dependent variable (X) values; h_{ii} measures the distance between the X values for the *i*-th data point, i.e. $\left(X_{i1}, X_{i2}, \cdots\cdots, X_{ip}\right)$ to the mean of all X values for all n data points, called the "centroid", i.e. $\left(\bar{X}_1, \bar{X}_2, \cdots\cdots, \bar{X}_p\right)$. Each is also called the "**leverage**"; the larger the leverage the point is further away from the centroid.

The fitted value $\hat{y} = Hy$ is linear combination of the observed Y values, where h_{ii} is the weight corresponding to the observation y_i . We can express $\hat{y}_i$ as

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$$

the **leverage**, h_{ii} quantifies the effect that the observed response y_i has on its predicted value $\hat{y}_i$. That is, if h_{ii} is small, then the observed response y_i plays only a small role in determining the predicted response $\hat{y}_i$. On the other hand, if h_{ii}, is large, then the observed response y_i plays a large role in determining the predicted response $\hat{y}_i$. It's for this reason that h_{ii} are called the "**leverages"**.

Also, since $\sigma^2(e_i) = (1 - h_{ii})\sigma^2$, large h_{ii} will result in small residual variation and will force the fitted value to be closer to the observed value. A leverage value is usually considered to be **large** if it is more than twice the mean leverage value (which is $2P/n$).

Data points with high leverage have the potential of moving the regression line up or down as the case may be. Recall that the regression line represents the regression equation in a graphic form, and is represented by the b coefficients. High leverage points make our estimation of b coefficients inaccurate. In such a situation, any conclusions drawn about the response variable could be misleading. Similarly, any predictions made on the basis of the regression model could be wrong.

Measure of Influence

Cook's Distance

If leverage gives us a warning about data points that have the potential of influencing the regression line, then Cook's Distance indicates how much actual influence each case has on the slope of the regression line.

Cook's Distance is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Cook's *Distance* is thus a way of identifying data points that actually do exert too big an influence.

$$D_i = \frac{\sum_{j=1}^{n}\left[\hat{y}_j - \hat{y}_{j(i)}\right]^2}{P \times MS_E}$$

where

- $\hat{y}_j$ = prediction for observation *j* from the full model,
- $\hat{y}_{j(i)}$ = prediction for *j*-th observation from the model in which i-th observation has been removed,
- $P = p + 1$ = number of coefficients in the full model, and

- MS_E = mean square error for the full model

Above expression can be algebraically simplified to

$$D_i = \frac{r_i^2}{P} \times \frac{h_{ii}}{1-h_{ii}}, \qquad i = 1, 2, \cdots, n \text{ and } r_i \text{ = Studentized residual.}$$

It may be noted that first component measures how well the model fits the i-th observation y_i (since smaller value of r_i implies better fit) whereas the second component gives the impact of the leverage of the i-th observation.

It may also be noted that D_i is large, if

i) Studentized residual is large, i.e. i-th observation is unusual w.r.t. y-values and

ii) the point is far from the centroid of the *X*-space, that is, if h_{ii} is large, or i-th observation is unusual w.r.t. x-values. In that case i-th data point will have substantial pull on the fit and the second term will be large.

Large values for Cook's Distance signify unusual observations. D_i > 1 require careful checking; whereas D_i > 4 would indicate that the point has a high influence.

[*Ref*: Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression", *Technometrics*, **19** (1), pp 15–18]

Difference in Fits (DFFITS)

We may also investigate the deletion influence of the ith observation on the predicted of fitted value. This leads to following diagnostic proposed by Belsley, Kuh and Welsch.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}, \qquad i = 1, 2, \cdots, n$$

where $\hat{y}_{(i)}$ is the fitted value of y_i and $S_{(i)}^2$ is the estimate of σ^2 based on a data set with the i-th observation removed. The denominator is just used for standardization, since $\mathrm{var}(\hat{\mathrm{y}}_{\mathrm{i}}) = \sigma^2 \mathrm{h}_{\mathrm{ii}}$.

Thus, $DFFITS_i$ is the change in the fitted value $\hat{y}_i$, in standard deviation term, if the observation i is removed. Belsley, Kuh and Welsch suggest that any observation for which $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ warrants attention.

Computationally we may find above, after simplification, as

$$\boldsymbol{DFFITS_i = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \times \frac{e_i}{\sqrt{S_{(i)}^2 (1 - h_{ii})}}, \qquad i = 1, 2, \cdots, n}$$

The second term in above expression is known as ***R-Student***, another scaled residual.

[*Ref*: D. A. Belsley, K. Kuh and R. E. Welsch. (1980) John Wiley & Sons, New York: *Regression diagnostics: Identifying influential data and sources of collinearity*]

PRESS Residuals

PRESS residuals are defined as

$$e_{(i)} = y_i - \hat{y}_{(i)} \qquad i = 1, 2, \cdots, n$$

where $\hat{y}_{(i)}$ is the estimated value of y_i based on a model where i-th observation is ignored. This prediction error calculation is repeated for all n observations. The i-th press residual can be simplified to

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \qquad i = 1, 2, \cdots, n$$

It may be noted that observations having large h_{ii} values will have large PRESS residuals. These observations will in general be **high influence** points.

Note that h_{ii} is always between 0 and 1. If h_{ii} is larger (close to 1), even a small value of residual e_i could result in a larger value of the PRESS residual. On the other hand, if h_{ii} is small (close to 0), even a large value of the ordinary residual e_i may result in a relatively small value of the PRESS residual. Thus, an influential observation is determined not only by the magnitude of residual but also by the corresponding value of leverage h_{ii}.

Example

Table 1 shows the leverage, Studentized residual, and influence for each of the five observations in a small dataset.

Table 1. Example Data.

ID	X	Y	h	r	D
A	1	2	0.39	-1.02	0.40
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.20	1.22	0.19
E	8	7	0.73	-1.68	8.86

h is the leverage, *r* is the Studentized residual, and *D* is Cook's measure of influence.

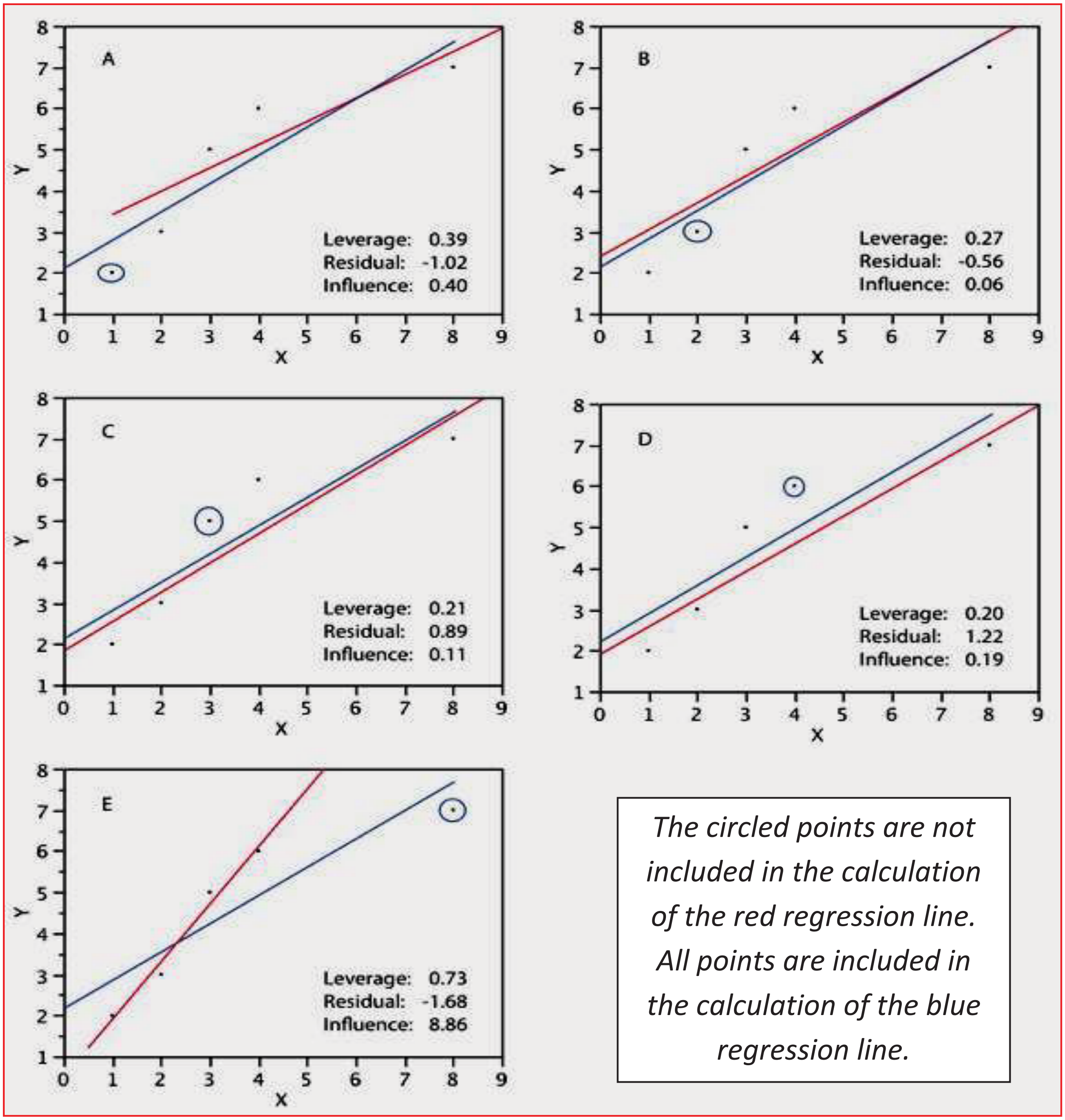
Observation A has fairly high leverage, a relatively high residual and moderately high influence.

Observation B has small leverage and a relatively small residual. It has very little influence.

Observation C has small leverage and a relatively high residual. The influence is relatively low.

Observation D has the lowest leverage and the second highest residual. Although its residual is much higher than Observation A, its influence is much less because of its low leverage.

Observation E has by far the largest leverage and the largest residual. This combination of high leverage and high residual makes this observation extremely influential.



The circled points are not included in the calculation of the red regression line. All points are included in the calculation of the blue regression line.

Selection of variables and Model building

An important problem in many application of regression analysis involves selecting the set of regressor variables to be used in the model. Sometimes, domain knowledge may help the analyst to specify the set of regressor variables to be used in a particular situation. Usually, however, the problem consists of selecting an appropriate set of regressor variables that adequately models the response variable and provides a reasonably good fit. In such a situation, we are interested in **variable selection** that is, screening the candidate variables to obtain a regression model that contains the "**best**" subset of regressor variables.

All Possible regression

This approach requires that the analyst fit all the regression equations involving one candidate variable, all regression equations involving two candidate variables, and so on. Then these equations are evaluated according to some suitable criteria to select the "best" regression model. If there are K candidate regressor, there are 2^K total equations to be examined. For example, if $K = 4$, there are $2^4 = 16$ possible regression equations; while if $K = 10$, there are $2^{10} = 1024$ possible regression equations. Hence, the number of equations to be examined increases rapidly as the number of candidate variables increases. However, there are some very efficient computing algorithms for all possible regressions available and they are widely implemented in statistical software, so it is a very practical procedure unless the number of candidate regressor is fairly large. Look for a menu choice such as "**Best Subsets**" regression.

Several criteria may be used for evaluating and comparing the different regression models obtained. A commonly used criterion is based on the value of R^2 or R^2_{adj}. Basically, the analyst continues to increase the number of variables in the model until the increase in R^2 or R^2_{adj} is small. Often, we will find that R^2_{adj} will stabilize and actually begin to decrease as the number of variables in the model increases. Usually, the model that maximizes R^2_{adj} is considered to be a good candidate for the best regression equation. Because we can write $R^2_{adj} = 1 - [MS_E/\{SS_T/(n-1)\}]$ and $SS_T/(n-1)$ is constant, the model that maximizes the R^2_{adj} value also minimizes the mean square error, so this is a very attractive criterion.

Another criterion used to evaluate regression models is the Mallow's C_p statistics that is related to the mean square error of a fitted value and is defined as

$$C_P = \frac{SS_E(P)}{MS_E} - n + 2P$$

where MS_E is the mean square error corresponding to the full $P = p + 1$ term model [see Montgomery, Peck and Vining or Myers]. Generally small values of C_P are desirable, i.e. a model with smaller value of C_P is considered to be better among the candidate regression models. For the full model involving P = p+1 coefficients, $C_P = P$.

The **PRESS** statistic can also be used to evaluate competing regression models. PRESS is an acronym for **Prediction Error Sum of Squares,** and it is defined as the sum of the squares of the differences between each observation y_i and the corresponding predicted value based on a model fit by ignoring ith observation, say $\hat{y}_{(i)}$(i.e. PRESS residuals). So PRESS provides a measure of how well the model is likely to perform when predicting *new* data, i.e. a data that was not used to fit the regression model.

The computing formula for PRESS is

$$\text{PRESS} = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)}\right]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right)^2$$

where $e_i = y_i - \hat{y}_i$ is the usual residual. Thus PRESS is easy to calculate from the standard least squares regression results.

A better regression model should be less sensitive to each individual observation. In other words, a better regression model should be less impacted by excluding one observation, that is should have a small value of $(y_i - \hat{y}_i)$ for all i. Therefore, a regression model with a smaller value of the PRESS statistic should be a preferred model.

The PRESS statistic can be used to compute an R^2-like statistic for prediction that would give the predictive capability of the model while predicting new observations.

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{\text{SS}_\text{T}}$$

$R^2_{\text{prediction}}$ value of, say, 0.9209 would mean that we expect the model to explain about 92.09% of the variability in predicting new observations.

Stepwise Regression

Stepwise Regression is probably the most widely used variable selection technique. The procedure iteratively constructs a sequence of regression models by adding or removing variables at each step. The criterion for adding or removing a variable at any step is usually expressed in terms of a partial F-test. Let f_{in} be the value of the F-random variable for adding a variable to the model, and let f_{out} be the value of the *F*-random variable for removing a variable from the model. We must have $f_{in} \geq f_{out}$, and usually $f_{in} = f_{out}$.

Stepwise regression begins by forming a one-variable model using the regressor variable that has the highest correlation with the response variable Y. This will also be the regressor producing the largest F-statistic. For example, suppose that at this step, x_1 is selected. At the second step, the remaining K - 1 candidate variables are examined, and the variable for which the partial F-statistic

$$F_j = \frac{SS_R\left(\beta_j|\beta_1,\beta_0\right)}{MS_E\left(x_j,x_1\right)} \qquad (1)$$

is a maximum is added to the equation, provided that $f_j > f_{in}$. In equation 1, $MS_E\left(x_j,x_1\right)$ denotes the mean square for error for the model containing both x_1 and x_j. Suppose that this procedure indicates that x_2 should be added to the model. Now the stepwise regression algorithm determines whether the variable x_1 added at the first step should be removed. This is done by calculating the F-statistic

$$F_1 = \frac{SS_R(\beta_1|\beta_2,\beta_0)}{MS_E(x_1,x_2)} \qquad (2)$$

If the calculated value $f_1 < f_{out}$, the variable x_1 is removed; otherwise it is retained, and we would attempt to add a regressor to the model containing both x_1 and x_2.

In general, at each step the set of remaining candidate regressor variables are examined, and the regressor with the largest partial F-statistic is entered, provided that the observed value of F exceeds f_{in}. Then the partial F-statistic for each regressor already in the model is calculated and the regressor, with the smallest observed value of F, is deleted if the observed $f < f_{out}$. The procedure continues until no other regressor variables can be added to or removed from the model.

Stepwise regression is almost always performed using a computer program. The analyst exercises control over the procedure by the choices of f_{in} and f_{out}. Some stepwise regression computer programs require that numerical values be specified for f_{in} and f_{out}. Since the number of degrees of freedom on MS_E depends on the number of variables in the model, which changes from step to step, a fixed value of f_{in} and f_{out} causes the type I and type II error rates to vary. Some computer programs allow the analyst to specify the type I error levels for f_{in} and f_{out}. Sometimes it is useful to experiment with different values of f_{in} and f_{out} (or different type I error levels) in several different runs to see if this substantially affects the choice of the final model.

Forward Selection

The **forward selection** procedure is a variation of stepwise regression and is based on the principle that regressor variables should be added to the model one at a time until there are no remaining candidate regressor variables that produce a significant increase in the regression sum of squares. That is, variables are added one at a time as long as their partial F-value exceeds f_{in}. Forward selection is a simplification of stepwise regression that omits the partial F-test for deleting variables from the model that have been added at previous steps. This is a potential weakness of forward selection; that is, the procedure does not explore the effect that adding a regressor at the current step has on regressor variables added at earlier steps. Notice that forward selection method