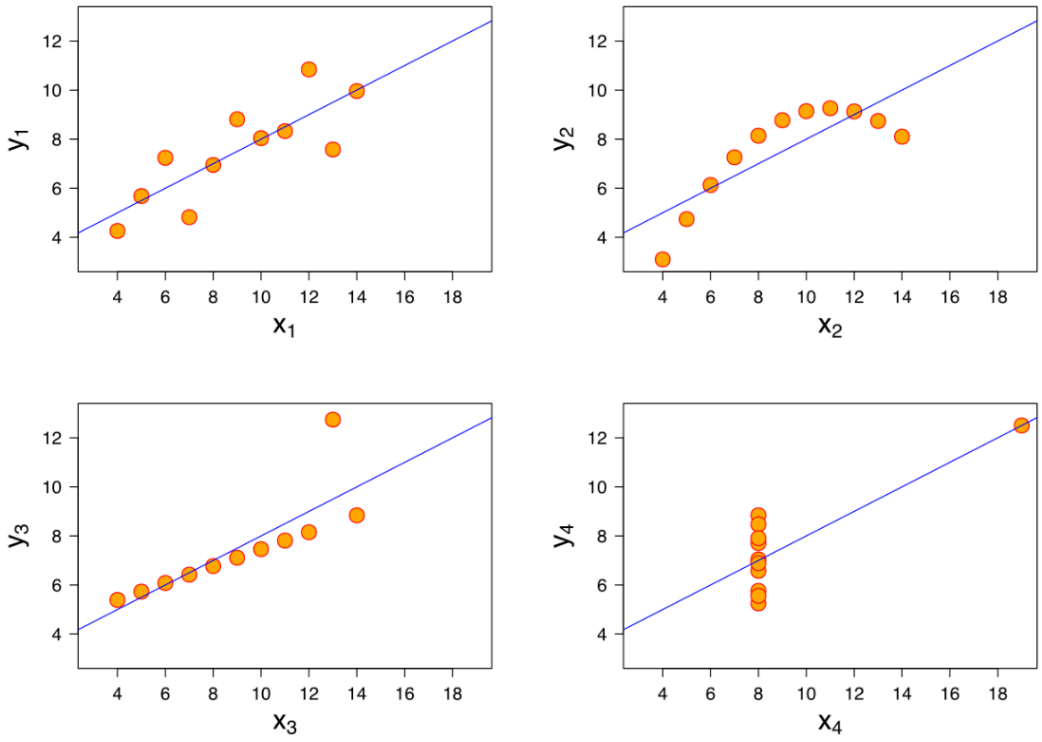
Chapter 4. Regression Diagnostics

4.1 Examining Residuals

Scatter Plot of Correct Model

To illustrate the importance of plotting data, four examples were developed by Anscombe(1973). In each example, the scatter plot of the data values is different. However, the regression equation and the R^2 statistics are the same.



In the first plot, the regression line adequately describes the data.

In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

In the third plot, there seems to be an outlying data value that is affecting the regression line. This outlier is an influential data value in that it is substantially changing the fit of the regression line.

In the fourth plot, the outlying data point dramatically changes the fit of the regression line. In fact, the slope would be undefined without the outlier.

➔ The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the R^2 statistics are the same even though the relationships between the two variables are different. Always produce a scatter plot before you conduct a regression analysis.

➔ 데이터의 대략적인 관계를 파악하기 위해 산점도를 통해 전체적인 분포를 확인해야 한다.

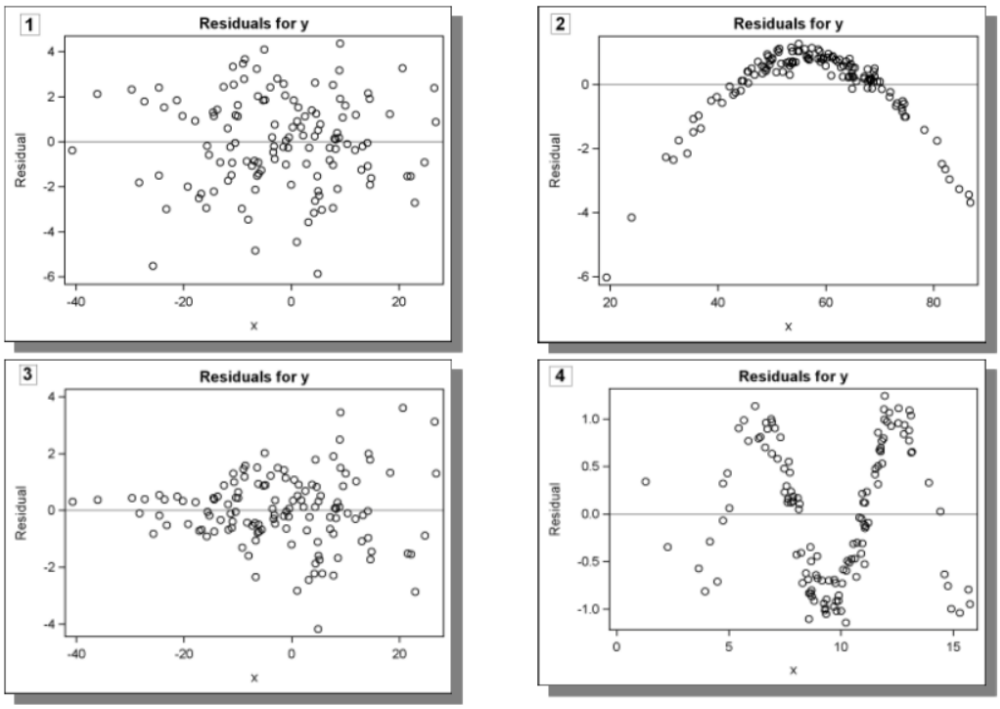
Verifying Assumptions

To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

$$r_i = Y_i - \hat{Y}_i$$

where $\hat{Y}_i$ is the predicted value for the i^{th} value of the dependent variable.

Examining Residual Plots



The graphs above are plots of residual values versus predicted values or predictor variable values for four models fit to different sets of data. If model assumptions are valid, then the residual values should be randomly scattered about a reference line at 0. Any patterns or trends in the residuals might indicate problems in the model.

1. The model form appears to be adequate because the residuals are randomly scattered about a reference line at 0 and no patterns appear in the residual values.

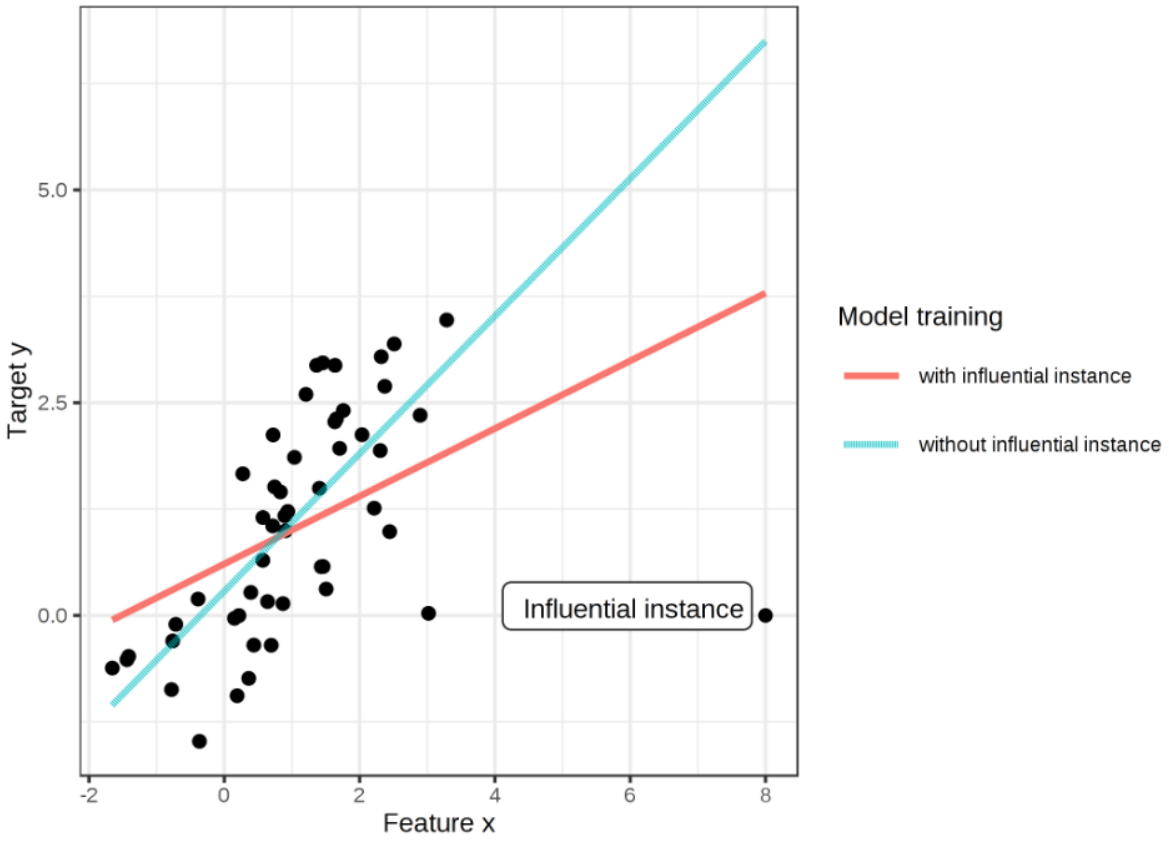
2. The model form is incorrect. The plot indicates that the model should take into account curvature in the data. One possible solution is to add a quadratic term as one of the predictor variables.

3. The variance is not constant. As you move from left to right, the variance increases. One possible solution is to transform your dependent variable.

4. The observations are not independent. For this graph, the residuals tend to be followed by residuals with the same sign, which is called *autocorrelation*. This problem can occur when you have observations that have been collected over time.

4.2 Influential Observations

Influential Observations (영향점)



An influential observation is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation. In particular, in regression analysis an influential point is one whose deletion has a large effect on the parameter estimates.

➔ 영향점은 회귀모형에 큰 영향을 끼치는 점이다. 각 관측점에 대해서 회귀분석에 포함된 경우와 그렇지 않은 경우를 비교해보고, 그 차이가 일정 수준을 넘을 때 영향점이라고 판단한다.

Diagnostic Statistics

Four statistics that help identify influential observations are

- STUDENT residual
- Cook's D
- RSTUDENT residual
- DFFITS

Cook's D Statistic (Cook's Distance)

Cook's distance measures the effect of deleting a given observation. Cook's distance D_i of observation i (for $\mathrm{i} = 1, \cdots, \mathrm{n}$) is defined as the sum of all the changes in the regression model when observation i is removed from it.

$$\mathrm{D}_i = \frac{\sum_{j=1}^{n}\left(\widehat{y_j} - \widehat{y_{j(i)}}\right)^2}{ps^2}$$

where $\widehat{y_{j(i)}}$ is the fitted response value obtained when excluding i, and $\mathrm{s}^2 = \frac{e^T e}{n-p}$ is the mean squared error of the regression model.

Equivalently, it can be expressed using the leverage (h_{ii}): $\mathrm{D}_i = \frac{e_i^2}{ps^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$.

A suggested cutoff is $\mathrm{D}_i > \frac{4}{n}$, where n is the sample size.

Studentized Residual (internally studentized)

Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.

$$\mathrm{SR} = \frac{\mathrm{e}_i}{\sqrt{MSE(1-h_{ii})}}$$

Suggested cutoffs are as follows:

- $|SR| > 2$ for data sets with a relatively small number of observations
- $|SR| > 3$ for data sets with a relatively large number of observations

RSTUDENT (externally studentized)

R-student residuals are similar to student residuals except that they are calculated after deleting the i^{th} observation. In other words, the R-student residual is the difference between the observed Y and the predicted value of Y excluding this observation from the regression.

$$\mathrm{t}_i = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_{ii})}}$$

where $\mathrm{s}_{(i)}^2 = \frac{(\mathrm{n}-\mathrm{p})\mathrm{MS}_{RES} - \frac{e_i^2}{(1-h_{ii})}}{n-p-1}$ (i 번째 자료를 제외한 MSE)

DFFITS

DIFFTS_i measures the impact that the i^{th} observation has on the predicted value.

$$\mathrm{DFFITS}_i = \frac{\widehat{Y_i} - \widehat{Y_{(i)}}}{s(\widehat{Y_i})}$$

where $\widehat{Y_i}$ is the i^{th} predicted value.

$\widehat{Y_{(i)}}$ is the i^{th} predicted value when the i^{th} observation is deleted.

$s(\widehat{Y_i})$ is the standard error of the i^{th} predicted value.

Suggested cutoff: $|DFFITS_i| > s\sqrt{\frac{p}{n}}$, where p is the number of terms in the current model, including the intercept, and n is the sample size.

DFBETAS (Difference in Betas)

DFBETAS measures how much the regression coefficient ($\hat{\beta}$) changes in standard deviation units if the i^{th} observation is removed.

$$\text{DFBETAS}_{j,i} = \frac{\widehat{\beta_j} - \widehat{\beta_{j(i)}}}{\sqrt{s^2_{(i)} C_{jj}}}$$

where $\widehat{\beta_{j(i)}}$ is an estimate of the j^{th} coefficient when the i^{th} observation is removed.

Suggested cutoff: $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$

How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors have occurred.

2. If the data is valid, one possible explanation is that the model is not adequate.

    - A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

If you had a larger sample size, there might be more observations like the unusual ones. In this case you might have to collect more data to confirm the relationship suggested by the influential observation. In general, do not exclude data. In many circumstances, some of the unusual observations contain important information. If you choose to exclude some observations, include a description of the types of observations you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.

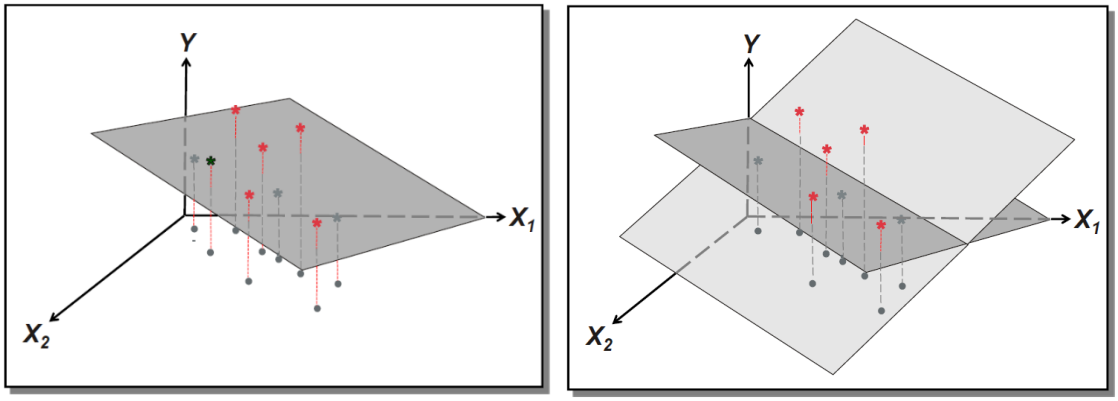
4.3 Collinearity

Illustration of Collinearity

- Collinearity is a condition in which some of the independent variables are highly correlated.
- Why is this a problem?

  – Neither might appear to be significant when both are in the model; however, either might be significant when only one is in the model. Thus, collinearity can hide significant effects. (The reverse can be true as well: collinearity can increase the apparent significance of effects.)

  – Collinearity also increases the variance of the parameter estimates and consequently increases prediction error.



➔ The removal of just one data point (or even just moving the data point) results in a very different prediction plane (as represented by the lighter plane). This illustrates variability of the parameter estimates when there is extreme collinearity.

➔ When collinearity is a problem, the estimates of the coefficients are unstable. Consequently, the true relationship between Y and the Xs might be quite different from that suggested by the magnitude and sign of the coefficients.

➔ Collinearity is **not** a violation of the assumptions of linear regression.

- When you have a highly significant Model F but no (or few) highly significant terms, collinearity is a likely problem.

Collinearity Diagnostics

PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- **VIF**: provides a measure of the magnitude of the collinearity.
- **COLLIN**: includes the intercept vector when analyzing the X'X matrix for collinearity.

- **COLLINOINT**: excludes the intercept vector.

Variance Inflation Factor (VIF)

The VIF is a relative measure of the increase in the variance because of collinearity. It can be thought of as the ratio:

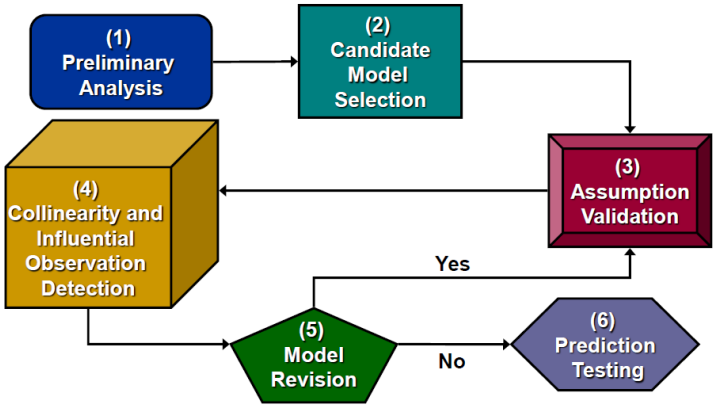
$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the R^2 of X_i, regressed on all the other Xs in the model.

For example, if the model is Y = X1 X2 X3 X4, i=1 to 4. To calculate the R^2 for X3, fit the model X3 = X1 X2 X4. Take the R^2 from the model with X3 as the dependent variable and replace it in the formula $\mathrm{VIF}_3 = 1/(1 - R_3^2)$. If the VIF_3 is greater than 10, X3 is possibly involved in collinearity.

A $\mathrm{VIF}_i > 10$ indicates that collinearity is a problem.

An Effective Modeling Cycle



(1) **Preliminary Analysis** includes the use of descriptive statistics, graphs, and correlation analysis.

(2) **Candidate Model Selection** uses the numerous selection options in PROC REG to identify one or more candidate models.

(3) **Assumption Validation** includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variance.

(4) **Collinearity and Influential Observation Detection** includes the use of the VIF statistic, condition indices, and variation proportions; the latter includes the examination of Rstudent residuals, Cook's D statistic, and DFFITS statistics.

(5) **Model Revision**. If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.

(6) **Prediction Testing**. If possible, validate the model with data not used to build the model.