18 Lecture 18: March 10

Last time

- HW1 review
- Lab review

Today

- Dummy-Variable regression
- Interactions
- Unusual and influential data (JF chapter 11)

Hypothesis tests

We first test the null hypothesis that all population regression slopes are 0:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

The test statistics,

$$F = \frac{RegSS/p}{RSS/(n-p-1)}$$

follows an F-distribution with p and $n-p-1$ degrees of freedom.

We can also test a null hypothesis about a *subset* of the regression slopes, e.g.,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0.$$

Or more generally, test the null hypothesis

$$H_0 : \beta_{q_1} = \beta_{q_2} = \cdots = \beta_{q_k} = 0$$

where $0 \leqslant q_1 < q_2 < \cdots < q_k \leqslant p$ is a subset of k indices. To get the F-statistic for this case, we generally perform the following steps:

1. Fit the *full* ("unconstrained") model, in other words, model that provides context for H_0. Record SSR_{full} and the associated df_{full}
2. Fit the *reduced* ("constrained") model, in other words, full model constrained by H_0. Record SSR_{red} and the associated df_{red}
3. Calculate the F-statistic by

$$F = \frac{[SSR_{red} - SSR_{full}]/(df_{red} - df_{full})}{SSR_{full}/df_{full}}$$

4. Find p-value (the probability of observing an F-statistic that is at least as high as the value that we obtained) by consulting an F-distribution with numerator $df(ndf) = df_{red} - df_{full}$ and denominator $df(ddf) = df_{full}$. Notation: $F_{ndf,ddf}$, see Figure 18.1.

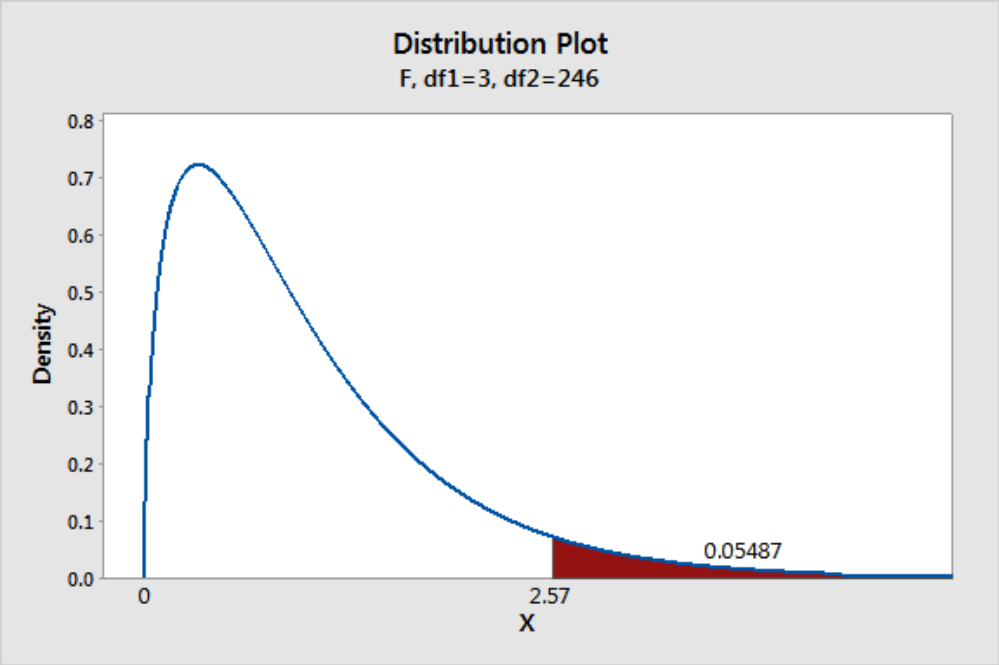

Figure 18.1: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution

Dummy-variable regression

For categorical data (factor), we use dummy variable regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i$$

where D, called a dummy variable regressor or an indicator variable, is coded 1 for one level and 0 for all others,

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases} .$$

Therefore, for women, the model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and for men

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i$$

For example, Figure 18.2 (a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the "effect" of gender is the vertical distance between the two regression lines, which, for parallel lines, is everywhere the same.

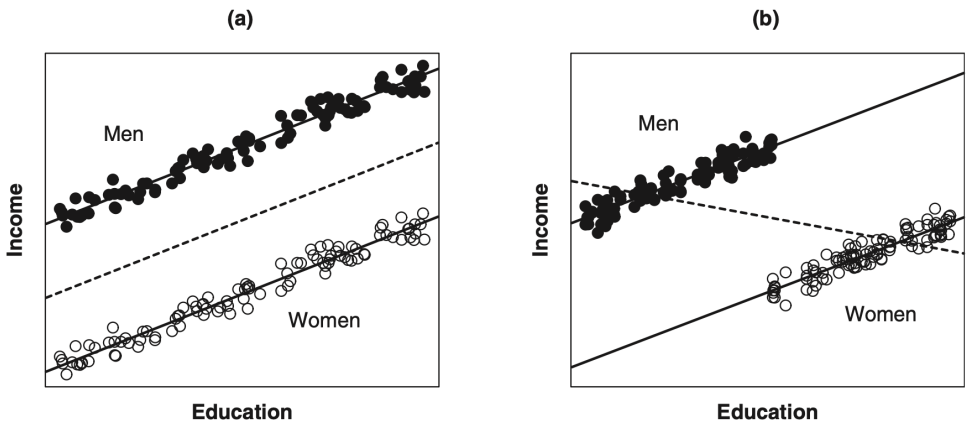

Figure 18.2: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e. marginal) regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.1.

Multi-level factor

We can model the effects of classification factors with m categories (levels) by using $m - 1$ indicator variables.

For example, the three-category occupational-type factor can be represented in the regression equation by introducing two dummy regressors:

Category	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{aligned}
\text{Professional:} \quad & Y_i = (\beta_0 + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\
\text{White collar:} \quad & Y_i = (\beta_0 + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\
\text{Blue collar:} \quad & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i
\end{aligned}$$

Therefore, the coeficient β_0 gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income).

In the above prestige example, we chose "blue collar" as the baseline category. Sometimes, it is natural to pick a particular category as the baseline category, for example, the "control group" in an experiment. However, in most applications, the choice of a baseline category is entirely arbitrary.

Matrix representation

For the above prestige model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

we have the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & D_{11} & D_{12} \\ 1 & X_{21} & X_{22} & D_{21} & D_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & D_{n1} & D_{n2} \end{bmatrix}$$

and the vector of coefficients β is

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

such that we have (again) the linear model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, in other words, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Interactions

Two explanatory variables are said to interact in determining a response variable when the partial effect of one depends on the value of the other. Consider the hypothetical data shown in Figure 18.3.

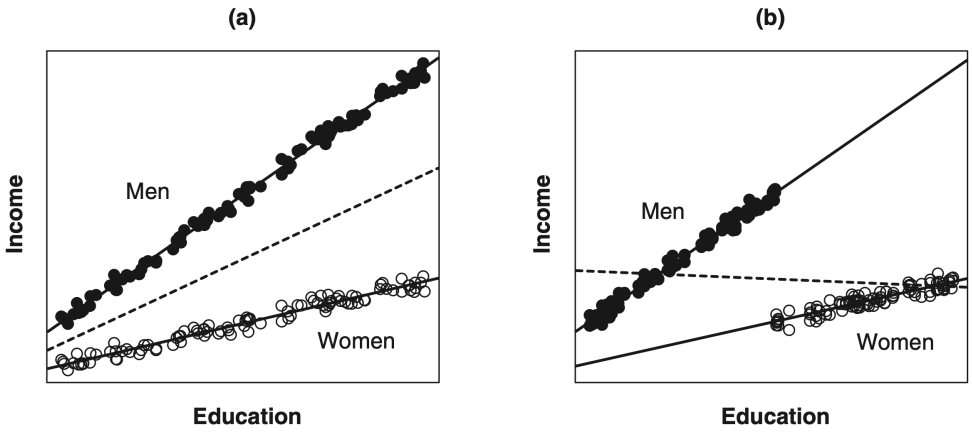

Figure 18.3: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are not parallel. The slope for men is greater than the slope for women, and consequently education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.7.

It is apparent in both Figure 18.3 (a) and (b) the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women.

Modeling interactions

We accommodate the interaction of education and gender by:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i D_i) + \epsilon_i$$

where we introduce the interaction regressor XD into the regression equation. For women, the model becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 0 + \beta_3 (X_i \cdot 0) + \epsilon_i \\ &= \beta_0 + \beta_1 X_i + \epsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 1 + \beta_3 (X_i \cdot 1) + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \epsilon_i \end{aligned}$$

The parameters β_0 and β_1 are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); β_2 gives the difference in intercepts between the male and female groups; and β_3 gives the difference in slopes between the two groups.

Usual guidance: Models that include an interaction between two predictors should also include the individual predictors by themselves regardless of the statistical significance of the associated β's.

Test for the interaction

We can simply test the hypothesis $H_0 : \beta_3 = 0$ and construct the test statistic $t = \frac{\hat{\beta}_i - 0}{\widehat{SE}(\hat{\beta}_i)} \sim t_{n-4}$ $(p = 3)$.

Interactions with multi-level factor

We can easily extend the method for modeling interactions by forming product regressors to multi-level factors, to several factors, and to several quantitative explanatory variables. Using the occupational prestige example, the occupational type could possibly interact both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \epsilon_i \end{aligned}$$

The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{array}{lrrrrr} \text{Professional:} & Y_i = & (\beta_0 + \gamma_1) + & (\beta_1 + \delta_{11}) X_{i1} + & (\beta_2 + \delta_{21}) X_{i2} + & \epsilon_i \\ \text{White collar:} & Y_i = & (\beta_0 + \gamma_2) + & (\beta_1 + \delta_{12}) X_{i1} + & (\beta_2 + \delta_{22}) X_{i2} + & \epsilon_i \\ \text{Blue collar:} & Y_i = & \beta_0 + & \beta_1 X_{i1} + & \beta_2 X_{i2} + & \epsilon_i \end{array}$$

Unusual and influential data

Linear models make strong assumptions about the structure of data, assumptions that often do not hold in applications. The method of least squares can be very sensitive to the structure of the data and may be markedly influenced by one or a few unusual observations.

Outliers

In simple regression analysis, an outlier is an observation whose response-variable value is *conditionally* unusual *given* the value of the explanatory variable: see Figure 18.4.

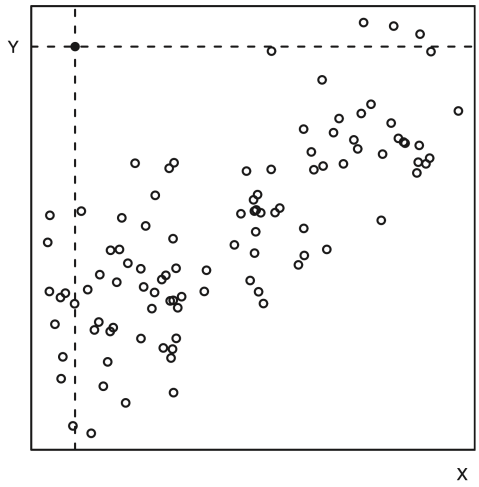

Figure 18.4: The black point is a regression outlier because it combines a relatively large value of Y with a relatively small value of X, even though neither its X-value nor its Y-value is unusual individually. Because of the positive relationship between Y and X, points with small X-values also tend to have small Y-values, and thus the black point is far from other points with similar X-values. JF Figure 11.1.

Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis. Their presence may be a signal that the model fails to capture important characteristics of the data.

Figure 18.5 illustrates some distinctions for the simple-regression model $Y = \beta_0 + \beta_1 X + \epsilon$.

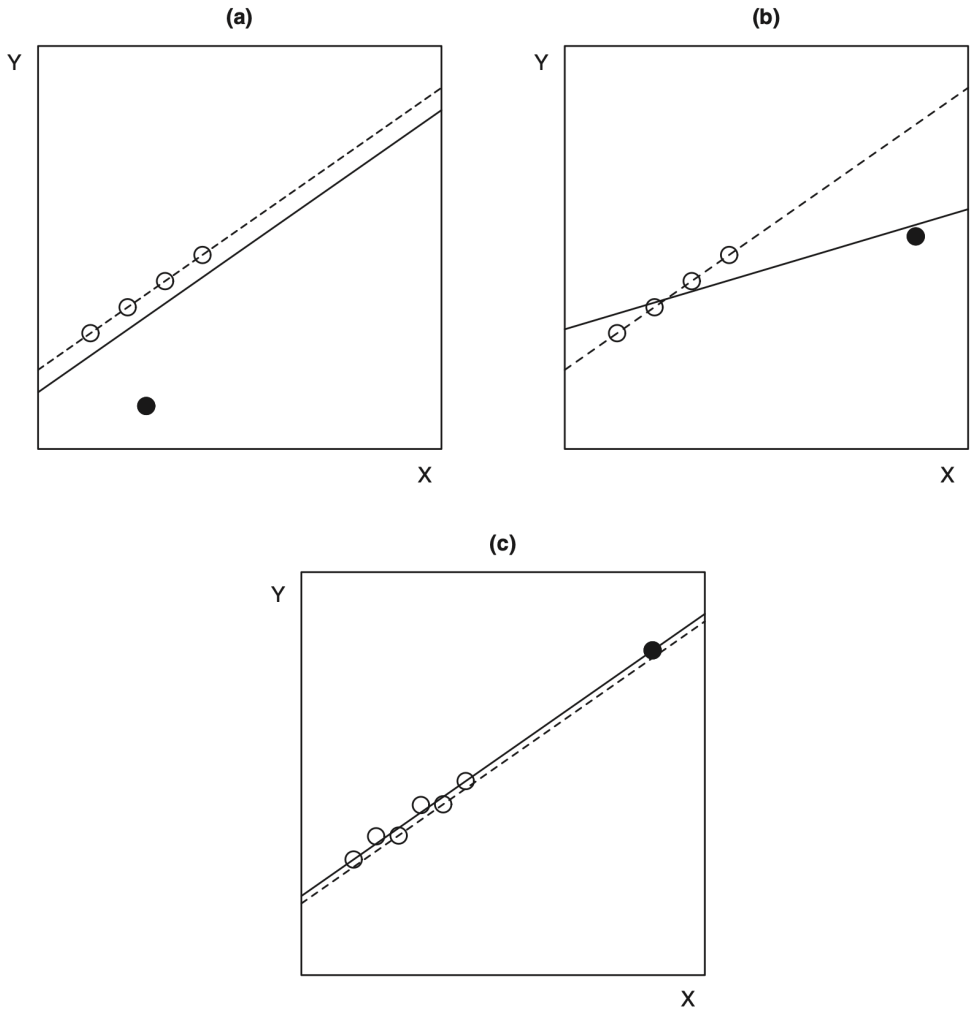

Figure 18.5: Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all the data, while the broken line gives the least-squares regression with the unusual data point (the black circle) omitted. (a) An outlier near the mean of X has low leverage and little influence on the regression coefficients. (b) An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients. (c) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect but are, in fact, coincident JF Figure 11.2.

Some qualitative distinctions between outliers and high leverage observations:

- An outlier is a data point whose response Y does not follow the general trend of the rest of the data.
- A data point has high leverage if it has "extreme" predictor X values:
  - With a single predictor, an extreme X value is simply one that is particularly high or low.

– With multiple predictors, extreme X values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values .

And the influence of a data point is the combination of leverage and discrepancy ("outlyingness") though the following heuristic formula:

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}.$$

Assessing leverage: hat-values

The hat-value h_i is a common measure of leverage in regression. They are named because it is possible to express the fitted values $\hat{Y}_j$ ("Y-hat") in terms of the observed values Y_i:

$$\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i.$$

The weight h_{ij} captures the contribution of observation Y_i to the fitted value $\hat{Y}_j$: If h_{ij} is large, then the ith observation can have a considerable impact on the jth fitted value. With the least square solutions, for the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

we (already) get the hat matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Properties:

- (idempotent) $\mathbf{H} = \mathbf{HH}$
- $h_i \equiv h_{ii} = \sum_{j=1}^{n} h_{ij}^2$
- $\frac{1}{n} \leqslant h_i \leqslant 1$ (a proof by Mohammad Mohammadi)
- $\bar{h} = (p+1)/n$

In the case of SLR, the hat-values are:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

Detecting outliers: studentized residuals

The variance of the residuals ($\hat{\epsilon}_i = Y_i - \hat{Y}_i$) do not have equal variances (even if the errors ϵ_i have equal variances):

$$\text{Var}(\hat{\epsilon}) = \text{Var}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \text{Var}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\text{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

so that for $\hat{\epsilon}_i$,

$$\text{Var}(\hat{\epsilon}_\text{i}) = \sigma^2(1 - \text{h}_\text{i}).$$

High-leverage observations tend to have small residuals (in other words, these observations can pull the regression surface toward them).

The standardized residual (sometimes called internally studentized residual)

$$\hat{\epsilon}_i' \equiv \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_i}},$$

however, does not follow a t-distribution, because the numerator and denominator are not independent.

Suppose, we refit the model deleting the ith observation, obtaining an estimate $\hat{\sigma}_{(-i)}$ of σ that is based on the remaining $n-1$ observations. Then the studentized residual (sometimes called externally studentized residual)

$$\hat{\epsilon}_i^* \equiv \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$$

has an independent numerator and denominator and follows a t-distribution with $n-p-2$ degrees of freedom.

The studentized and the standardized residuals have the following relationship (Beckman and Trussell, 1974):

$$\hat{\epsilon}_i^* = \hat{\epsilon}_i' \sqrt{\frac{n-p-2}{n-p-1-\hat{\epsilon}_i'^2}}$$

For large n,

$$\hat{\epsilon}_i^* \approx \hat{\epsilon}_i' \approx \frac{\hat{\epsilon}}{\hat{\sigma}}$$

Test for outlier

It is of our interest to pick the studentized residual $\hat{\epsilon}_{max}^*$ with the largest absolute value among $\hat{\epsilon}_1^*, \hat{\epsilon}_2^*, \ldots, \hat{\epsilon}_n^*$ to test for outlier. However, by doing so, we are effectively picking the biggest of n test statistics such that it is not legitimate simply to use t_{n-p-2} to find a p-value. We need a correction on the p-value because of multiple-comparisons.

Suppose that we have $p' = \Pr(t_{n-p-2} > |\hat{\epsilon}_{max}^*|)$, the p-value before correction. Then the Bonferroni adjusted p-value is $p = np'$.

Measuring influence

Influence on the regression coefficients combines leverage and discrepancy. The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

$$D_{ij} = \hat{\beta}_j - \tilde{\beta}_{j(-i)} \quad \text{for } i = 1, \ldots, n \text{ and } j = 0, 1, \ldots, p$$

where $\hat{\beta}_j$ are the least-squares coefficients calculated for all the data, and the $\tilde{\beta}_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. To assist in interpretation, it is useful to scale the D_{ij} by (deleted) coefficient standard errors:

$$D_{ij}^* = \frac{D_{ij}}{\widehat{SE}_{(-i)}(\tilde{\beta}_{j(-i)})}$$

Following Belsley, Kuh, and Welsh (1980), the D_{ij} are often termed DFBETA_{ij}, and D_{ij}^* are called DFBETAS_{ij}. One problem associated with using D_{ij} or D_{ij}^* is their large number: $n(p+1)$ of each.

Cook's distance calculated as

$$D_i = \frac{\sum_{j=1}^{n}(\tilde{y}_{j(-i)} - \hat{y}_j)^2}{(p+1)\hat{\sigma}^2} = \frac{\hat{\epsilon}_i'^2}{p+1} \times \frac{h_i}{1-h_i}$$

In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage. We look for values of D_i that stand out from the rest.

A similar measure suggested by Belsley et al. (1980)

$$\text{DFFITS}_i = \hat{\epsilon}_i^* \frac{h_i}{1-h_i}$$

Except for unusual data configurations, Cook's $D_i \approx \text{DFFITS}_i^2/(p+1)$.

Numerical cutoffs (suggested)

Diagnostic statistic	Cutoff value
h_i	$2\bar{h} = \frac{2(p+1)}{n}$, ($3\bar{h}$ for small sample)
D_{ij}^*	$\lvert D_{ij}^*\rvert > 1$ or 2 ($2/\sqrt{n}$ for large samples)
Cook's D_i	$D_i > \frac{4}{n-p-1}$
DFFITS	$\lvert\text{DFFITS}_i\rvert > 2\sqrt{\frac{p+1}{n-p-1}}$