09

데이터분석방법론(1)

Unusual and Influential Data

통계·데이터과학과 장영재 교수

한국방송통신대학교 대학원

학습목차

01

Outliers, Leverage, and Influence

1. Introduction

- Linear statistical models make strong assumptions about the structure of data, which sometimes do not hold in real world.
- The least squares method is very sensitive to the structure of the data and can be greatly influenced by one or a few unusual observations.
- We can abandon linear models and least squares estimation in favor of nonparametric regression and robust estimation.
- We can adapt and extend methods for examining and transforming data to diagnose problems with a linear model, and to suggest solutions.

2. Unusual Observations

- Unusual data are problematic in fitting linear models by least squares because they can unduly influence the results of the analysis and their presence can be a signal that the model does not capture important characteristics of the data.
- Some central distinctions are illustrated in Figure 1 (see next page) for the simple regression model

$$Y = \alpha + \beta X + \varepsilon.$$

  - In simple regression, an outlier is an observation whose response-variable value is conditionally unusual given the value of the explanatory variable.

2. Unusual Observations

◆ Figure 1. Unusual data in regression: (a) a low-leverage and hence un-influential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least-squares line for all of the data; the dashed line is the least-squares line with the unusual observation removed.

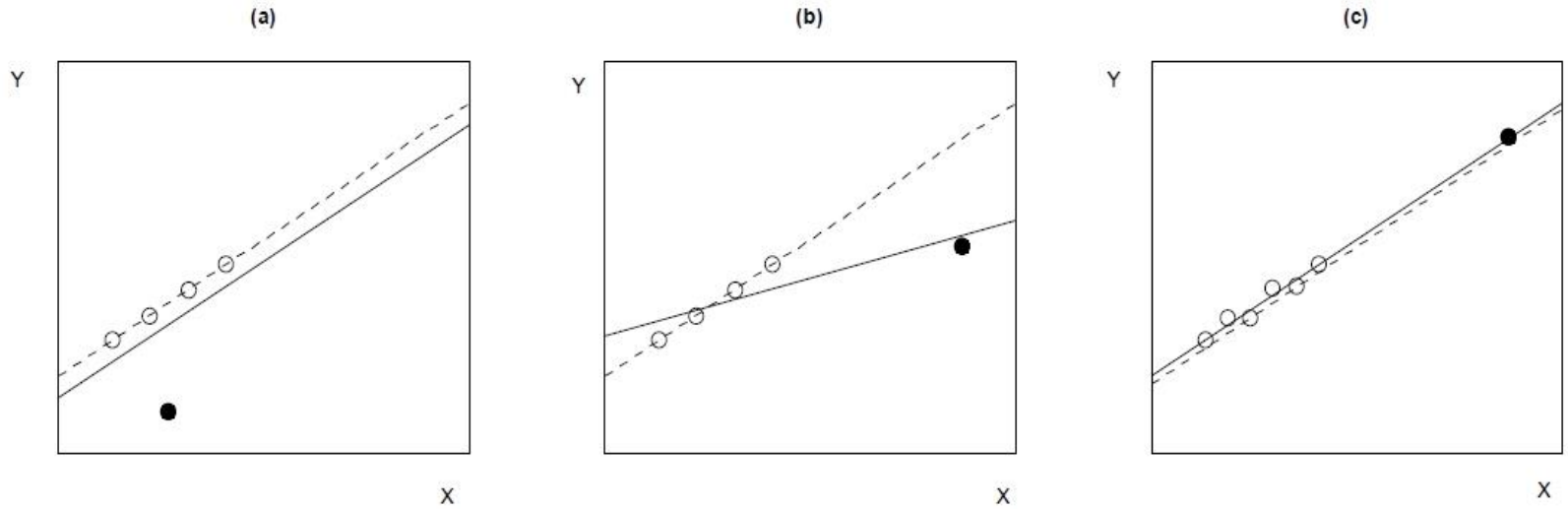


Figure 1. Outliers, Leverage, and Influence

2. Unusual Observations

- Regression outliers appear in (a) and (b).
- In (a), the outlying observation has an X -value that is at the center of the X distribution; deleting the outlier has little impact on the least-squares fit.
- In (b), the outlier has an unusual X -value; deletion has a large effect on both slope and intercept. Due to its unusual X -value, the abnormal last observation in (b) has a strong leverage on the regression coefficients, while the abnormal middle observation in (a) is at a low-leverage point. The combination of high leverage and regression outliers has a significant impact on the regression coefficients.

2. Unusual Observations

- In (c), the last observation does not affect the regression coefficients even at high leverage points. This is because this observation is consistent with the rest of the data.
- The following heuristic formula helps distinguish between the three concepts of influence, influence, and discrepancy. ('outlyingness'):

  Influence on Coefficients = Leverage × Discrepancy

3. Regression Analysis with unusual observations

- A simple example using real data from Davis (1990) is shown in Figure 2. The data record the measured and reported weight of her 183 male and female subjects participating in a regular physical exercise program. Davis' data can be processed in her two ways.

(1) We could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F, coded 1 for women and 0 for men), and an interaction regressor (formed as the product MW × F):

$$RW = 1.36 + 0.990MW + 40.0F - 0.725(MW \times F)$$
$$(3.28)(0.043)(3.9)(0.056)$$
$$R^2 = 0.89,\ S_E = 4.66$$

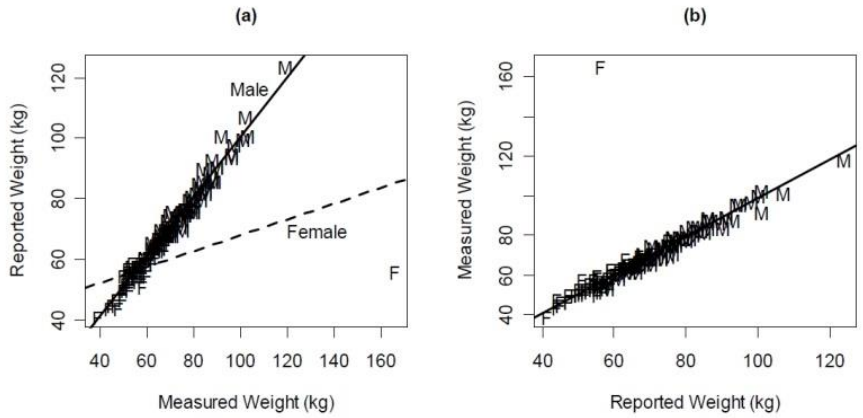
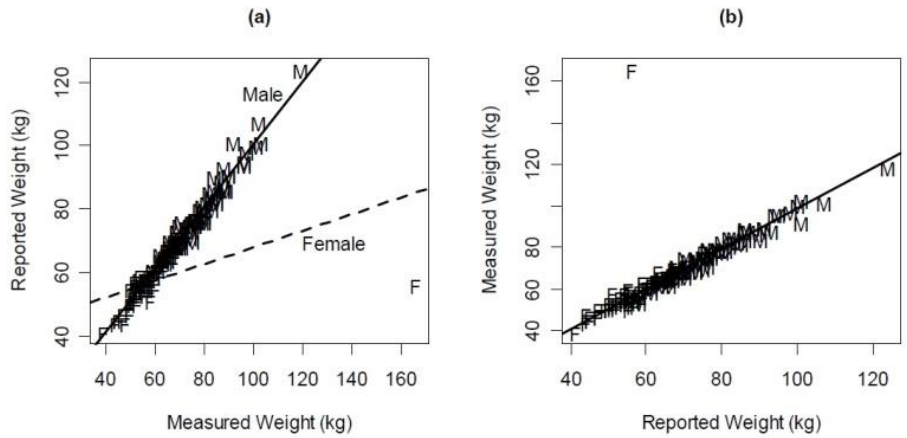


Figure 2. Fitted lines

3. Regression Analysis with unusual observations

- If we take these results seriously, then we can conclude that men report their weight fairly, while women tend to over-report their weights if they are relatively light and under-report if they are relatively heavy.
- The figure makes it clear that the differential results for women and men are due to one erroneous data point.
- Correcting the data produces the regression

$$RW = 1.36 + 0.990MW + 1.98F - 0.0567(MW \times F)$$
$$(1.58) \quad (0.021) \; (2.45) \quad (0.0385)$$
$$R^2 = 0.97, \; S_E = 2.24$$



3. Regression Analysis with unusual observations

(2) We could regress measured weight on reported weight, sex, and their interaction:

$$MW = 1.79 + 0.969RW + 2.07F - 0.00953(RW \times F)$$
$$\quad (5.92) \quad (0.076) \quad (9.30) \quad (0.147)$$
$$R^2 = 0.70, \qquad S_E = 8.45$$

- The outlier does not have much impact on the regression coefficients because the value of RW for the abnormal outlying observation is near $\overline{RW}$ for women.
- There is, however, remarkable effects on multiple correlations and standard errors: For the corrected data, $R^2 = 0.97$ and $S_E = 2.25$.

02

Detecting and Testing Unusual Observations

1. Hat-Values

- The hat-value h_i is a common measure of leverage in regression. The name comes from the possibility to express the fitted values $\hat{Y}_j$ ('Y -hat') in terms of the observed values Y_i:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$$

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

  - Thus, the weight h_{ij} implies the contribution of observation h_i to the fitted value $\hat{Y}_j$: If h_{ij} is large, then the *i* th observation can have a considerable impact on the *j* th fitted value.

- Properties of the hat-values:

  - $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on all of the fitted values.

  - $1/n \leq h_i \leq 1$

1. Hat-Values

- The average hat-value is $\bar{h} = \frac{k+1}{n}$ $\quad (Note\ that \sum_{i=1}^{n} h_{ii} = tr(H) = tr(I_{k+1}) = k+1)$
- In simple-regression analysis, the hat-values measure distance from the mean of X:

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

- In multiple regression, h_i quantifies the distance from the centroid (mean point) of the independent variables X's, considering both the correlational and variational patterns within the X's, as illustrated for k = 2 in Figure 3. Multivariate outliers in the X -space are thus high-leverage observations. It's important to note that the determination of leverage is solely based on the independent variable values, and the values of the response variable play no role in this assessment.

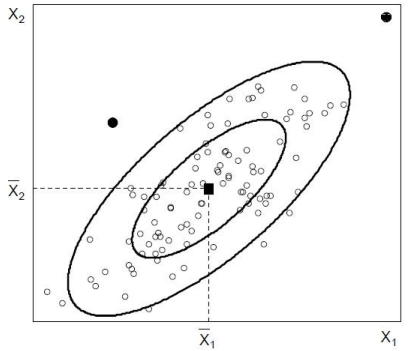


Figure 3. Contours of constant leverage in multiple regression with two explanatory variables, X_1 and X_2. The two observations marked with solid black dots have equal hat-values.

1. Hat-Values

- R code : hat value

```
> ginf = influence(g)
> names(ginf)
[1] "hat"          "coefficients" "sigma"        "wt.res"
> ginf$hat[1:20]
          1           2           3           4           5           6           7
0.012282789 0.009906711 0.011611240 0.017509292 0.009930159 0.012195172 0.012195172
          8           9          10          11          12          13          14
0.016260229 0.014263359 0.022258990 0.015178251 0.714185650 0.013143212 0.011869174
         15          16          17          18          19          20
0.012316500 0.012621331 0.033647263 0.010729214 0.012195172 0.010341411
> mean(ginf$hat)
[1] 0.02185792
```

◆ In Davis's regression analysis, where reported weight is regressed on measured weight, the 12th subject stands out with the highest hat-value. This is primarily due to an error in recording the measured weight for this subject, which was mistakenly documented as 166 kg : $h_{12} = 0.714$. This quantity is larger than the average hatvalue, $\bar{h} = (3+1)/183 = 0.0219$.

1. Hat-Values

- (Example) Duncan's regression of occupational prestige on income and education for 45 U.S. occupations in 1950: (https://search.r-project.org/CRAN/refmans/carData/html/Duncan.html)

$$\underset{}{\text{Prestige}} = \underset{(4.27)}{-6.06} + \underset{(0.120)}{0.599} \times \text{Income} + \underset{(0.098)}{0.546} \times \text{Education}$$

  - An index plot of hat-values for the observations in Duncan's regression is shown in Figure 4 (a), with a scatterplot for the explanatory variables in

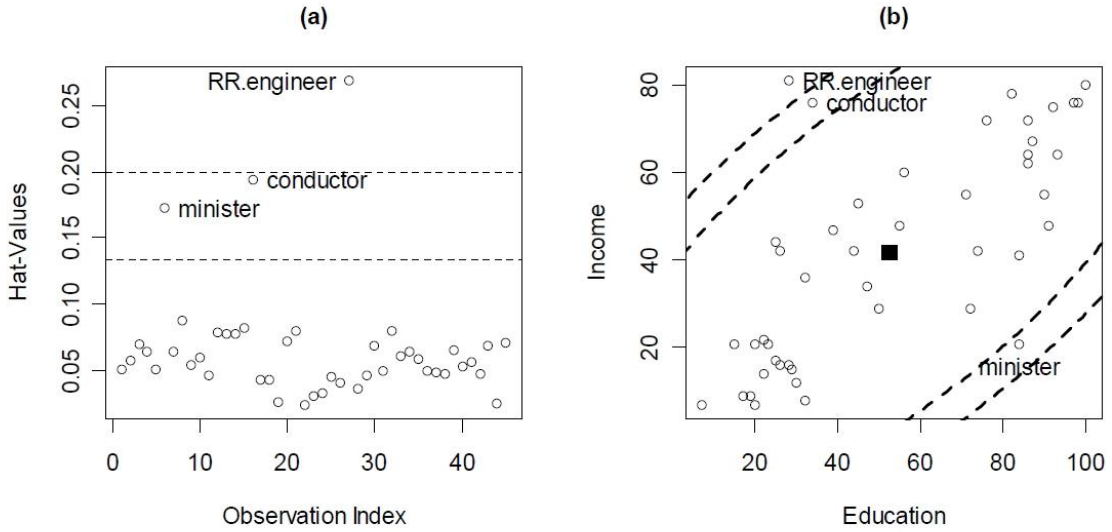


Figure 4. Duncan's occupational prestige regression: (a) hat-values; (b) scatterplot for education and income, showing contours of constant leverage At $2 \times \bar{h}$ and $3 \times \bar{h}$.

2. Detecting Outliers: Studentized Residuals

- Observations that deviate typically exhibit large residuals, but even if the errors ε_i have equal variances (as assumed in the general linear model), the residuals E_i do not:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

  - Observations with high leverage typically exhibit small residuals, as these observations have the ability to strongly influence the regression surface, compelling it to be in close proximity to them.

- Although we can form a standardized residual(or internally studentized residual)

$$r_i = \frac{E_i}{S_E\sqrt{1 - h_i}}$$

this measure is slightly inconvenient because its numerator and denominator are not independent, preventing $|r_i|$ from following a t-distribution: When r_i is large, $S_E = \sqrt{\sum E_i^2/(n - k - 1)}$, which contains E_i^2, tends to be large as well.

2. Detecting Outliers: Studentized Residuals

- Suppose that we refit the model by excluding the i-th observation. We obtain an estimate $S_{E(-i)}$ of σ_ε relying on the remaining $n-1$ observations.
  - Then we get the studentized residual (or externally studentized residual, jackknife residual) $t_i = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$

    Note that numerator and denominator are independent, and t_i follows a t-distribution with $n-k-2$ degrees of freedom.
  - An equivalent procedure for finding the studentized residuals employs a 'mean-shift' outlier model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \gamma D + \varepsilon$$

where D is a dummy variable

$$D = \begin{cases} 1 & \text{for observation } i \\ 0 & \text{o.w. (ohter observations)} \end{cases}$$

2. Detecting Outliers: Studentized Residuals

- Thus

$$E(Yi) = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma$$
$$E(Yj) = \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} \quad for \; j \neq i$$

- If we had a suspicion that observation i differed from the others before analyzing the data, it would be appropriate to explicitly define this model.
- Then to test $H_0 : \gamma = 0$, we can calculate $t_0 = \hat{\gamma}/\mathrm{SE}(\hat{\gamma})$. This test statistic is distributed as t_{n-k-2} under H_0, and is the studentized residual.

3. Testing for Outliers

- In typical applications, our aim is to identify any outliers that might emerge in the data. To achieve this, we can iteratively reapply the mean-shift model, generating studentized residuals $t_1, t_2, \ldots, t_n$.
  - Usually, our interest then focuses on the largest absolute t_i, denoted by t_{max}.
  - Because we have picked the biggest of n test statistics, it is not acceptable to solely utilize t_{n-k-2} to find a p-value for t_{max}.

3. Testing for Outliers

- One solution : simultaneous inference is to perform a Bonferroni adjustment to the *p* -value for the largest absolute t_1;

  Let $p' = \Pr(t_{n-k-2} > t_{max})$, then the Bonferroni *p*-value for testing the statistical

  significance of t_{max} is $p = 2np'$. (two-sided p-value x number of observations)

  - Note that achieving statistical significance requires a substantially larger t_{max} than an ordinary individual t-test (very conservative).

- Another approach is to construct a quantile-comparison plot for the studentized residuals, plotting against either the t or normal distribution.

4. 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:

$$D_{ij} = B_j - B_{j(-i)} \quad for \;\; i = 1, \ldots, n \;\; and \;\; j = 0, 1, \ldots, k$$

where the B_j are the least-squares coefficients calculated for all of the data, and the $B_{j(-i)}$ are the least-squares coefficients calculated with the ith observation removed.

- One problem associated with using the D_{ij} is their large number n(k + 1).
  - It is useful to have a single summary index of the influence of each observation on the least-squares fit.

4. Measuring Influence

- Cook (1977) has proposed measuring the 'distance' between the B_j and the corresponding $B_{j(-i)}$ by calculating the F -statistic for the 'hypothesis' that $\beta_j = B_{j(-i)}$ for $j = 0, 1, \cdots, k$.
  - This statistic is recalculated for each observation $i = 1, \ldots, n$.
  - The resulting values should not literally be interpreted as F tests, but rather as a distance measure which does not depend on the scales of the X's.
  - Cook's statistic can be written (and simply calculated) as

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{(k+1)\hat{\sigma}^2}$$
$$= \frac{r_i^2}{k+1} \times \frac{h_i}{1 - h_i}$$

  - 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 are substantially larger than the rest.

4. Measuring Influence

- Since the deletion statistics depend on hat-values and residuals, an alternative approach is to create a plot the t_i against h_i and identify instances where both values are big. A slightly more sophisticated version of this plot, which integrates Cook's D, is presented below.
- For Duncan's regression, the largest Cook's D is for ministers, D6 = 0.566:
- Figure 5 displays a plot of studentized residuals versus hat-values,

with the areas of the plotted circles proportional to values of Cook's D.

The lines on the plot are at t = ±2 (on the vertical axis), and

at $h = 2\bar{h}$ and $3\bar{h}$ (on the horizontal axis) ;

reporters have a relatively large residual but are at a low-leverage point,

while railroad engineers have high leverage but

a small studentized residual

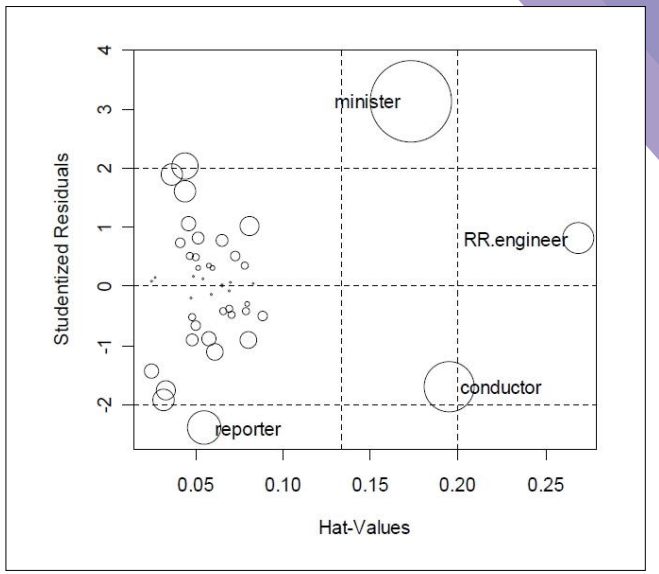


Figure 5. Studentized Residuals vs Hat-Values

5. Numerical Cutoffs for Diagnostic Statistics

- We have chosen not to propose explicit numerical criteria for identifying significant observations based on measures of leverage and influence. Instead, we believe that directly inspecting the distributions of these measures is generally a more effective approach for identifying unusual values.
  - While outlier-testing establishes a numerical cutoff for studentized residuals, it is essential to emphasize that even this should not replace the importance of visually examining the residuals.
- However, numeric thresholds can be useful, provided they are not overly emphasized, particularly when utilized to improve visual representations.
  - A numerical cutoff value can be set on a graph, and any individual observation surpassing this cutoff can be identified.

5. Numerical Cutoffs for Diagnostic Statistics

- The cutoffs for a diagnostic statistic can be determined either through statistical theory or by examination of the sample distribution of the statistic.
- Cutoffs may be absolute, or they may be adjusted for sample size.
  - Certain diagnostic statistics, such as measures of influence, are less likely to pinpoint noteworthy observations in large samples when using absolute cutoffs.
  - This is partly due to the capacity of large samples to assimilate divergent data without substantially altering the outcomes. However, there is still a frequent desire to identify relatively influential points, even if no observation has strong absolute influence.
  - The cutoffs presented afterwards are derived from statistical theory.

5. Numerical Cutoffs for Diagnostic Statistics

Hat-Values

- Belsley, Kuh, and Welsch suggest that hat-values exceeding about twice the average $\bar{h} = (k+1)/n$ are noteworthy.
- In small samples, using $2 \times \bar{h}$ tends to nominate too many points for examination, and $3 \times \bar{h}$ can be used instead.

5. Numerical Cutoffs for Diagnostic Statistics

Studentized Residuals

- Moving beyond the concern of 'statistical significance,' it can be beneficial to highlight residuals that exhibit considerable deviation.
- In optimal scenarios, approximately five percent of studentized residuals fall beyond the range $|t_i| \leq 2$. Consequently, it is justifiable to bring notice to observations lying outside this specified range.

5. Numerical Cutoffs for Diagnostic Statistics

Measures of Influence

- Various thresholds have been proposed for different influence measures, one of which is the size-adjusted cutoff for Cook's D, due to Chatterjee and Hadi:

$$D_i > \frac{4}{n-k-1}$$

- Establishing absolute cutoffs for D, such as setting $D_i > 1$, poses a potential risk of overlooking data points that may have a relatively significant impact.

03

Discussion

1. Should Unusual Data Be Discarded?

- While it's important not to overlook problematic data, automatic and unreflective deletion of such data should be avoided.

- Examining the reasons behind an unusual observation is crucial.
  - If the data is genuinely flawed, corrective measures can be applied or the data can be discarded.
  - On the other hand, when an anomalous data point is accurate, exploring the reasons for its uniqueness may provide valuable insights.

1. Should Unusual Data Be Discarded?

- Outliers or influential data may motivate model re-specification.
  - For instance, when observing patterns of outliers in the data, it might indicate the need to introduce supplementary explanatory variables.
  - In some instances, transformation of the response variable or of an explanatory variable bring apparent outliers closer to the rest of the data. This adjustment could involve making the error distribution more symmetric or removing nonlinearity.
    - It is crucial, however, to exercise caution and prevent 'over-fitting' the data—avoiding a scenario where a small portion of the data disproportionately influences the model's structure.

1. Should Unusual Data Be Discarded?

- Unless the situations are straightforward, it is understandable that we hesitate to remove observations or alter the model extensively to accommodate atypical data.
  - Certain researchers opt for alternative estimation approaches, like robust regression, which consistently assigns lower weights to outliers rather than outright including or excluding them.
  - While these methods give minimal or zero weight to highly deviant data points, the outcomes typically do not differ significantly from a meticulous use of least squares. In fact, the weights assigned in robust regression can serve as a tool for identifying outliers.

다음시간 안내

10

Qualitative Variables as Regressors