General Linear Regression Model

Regression Diagnostics

Índice

General linear model setting

Data

- n cases, $i = 1, \ldots, n$
- 1 response (dependent) variable

$$y_i, \;\; i = 1, \ldots, n$$

- p explanatory (independent) variables

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})', \;\; i = 1, \ldots, n$$

General Goal

- Extract/exploit relationships between y_i and x_i.
- Prediction, explanation purposes.

General linear model setting

General Linear Model for each case i, the conditional distribution $[y_i|x_i]$ is given by

$$y_i = \hat{y}_i + \epsilon_i$$

where:

- $\hat{y}_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $\beta = (\beta_1, \ldots, \beta_p)'$ are p regresssion parameters (constant over all cases)
- ϵ_i residual (error) variable (varies over all cases)

They are a really broad class of models!

Steps for fitting a model

1. Propose a model in terms of
   - Response variable (Y)
   - Explanatory variables $X_1, \ldots, X_p$ (include different functions of explanatory variables if appropriate)
   - Assumptions about the distribution of ϵ over the cases
2. Specify/define a criterion for judging different estimators.
3. Characterize the best estimator and apply it to the given data.
4. Ckeck the assumptions in (1)
5. If necessary, modify model specifications and/or assumptions and go to (1).

A few words on ...

which important parts of model building we are skipping.

And also we lack the expertise in the field, the subject-matter knowledge, the possibility of at least discuss the process with experts in the field where the data comes from, which can be really important.

Assessing model Assumptions

Once the model is fitted ... I thought I was done. Unfortunately many things can go wrong. Any statistical model we fit is probably wrong or incomplete. All we hope for is a decent summary and valid inference.

Looking for evidence against the assumptions

Model adequacy checking based on residual analysis:

1. Basic graphical inspection:
   - Basic residuals scatterplots
   - Normal probability plots
   - residuals vs. fitted values
   - Influence plots
2. Statistical inspection
   - To check normality, any statistical test on the residuals (Anderson-Darling, Jarque-Bera, ...)
   - To check constant variance, for example, Breusch-Pagan test for Heteroscedasticity.
   - To check for independence, for example, Durbin-Watson test for autocorrelation of resiudals.
   - Influence observation measures.
   - Multicollinearity measures.

Residual Plots

- Residual plots show what has not yet been accounted for in the model.
- They offer an opportunity to learn something more.
- We need to know what patterns in residuals to look for.

Partial relations plots (vs. marginal relations plots):

- Added Variable Plot (Partial regression plot): they show the unique effects of one predictor, controlling (or adjusting) for others. Typically they are used to look for influential observation.
- Component + residual plot. Typically they are used to look for non linear relationships.

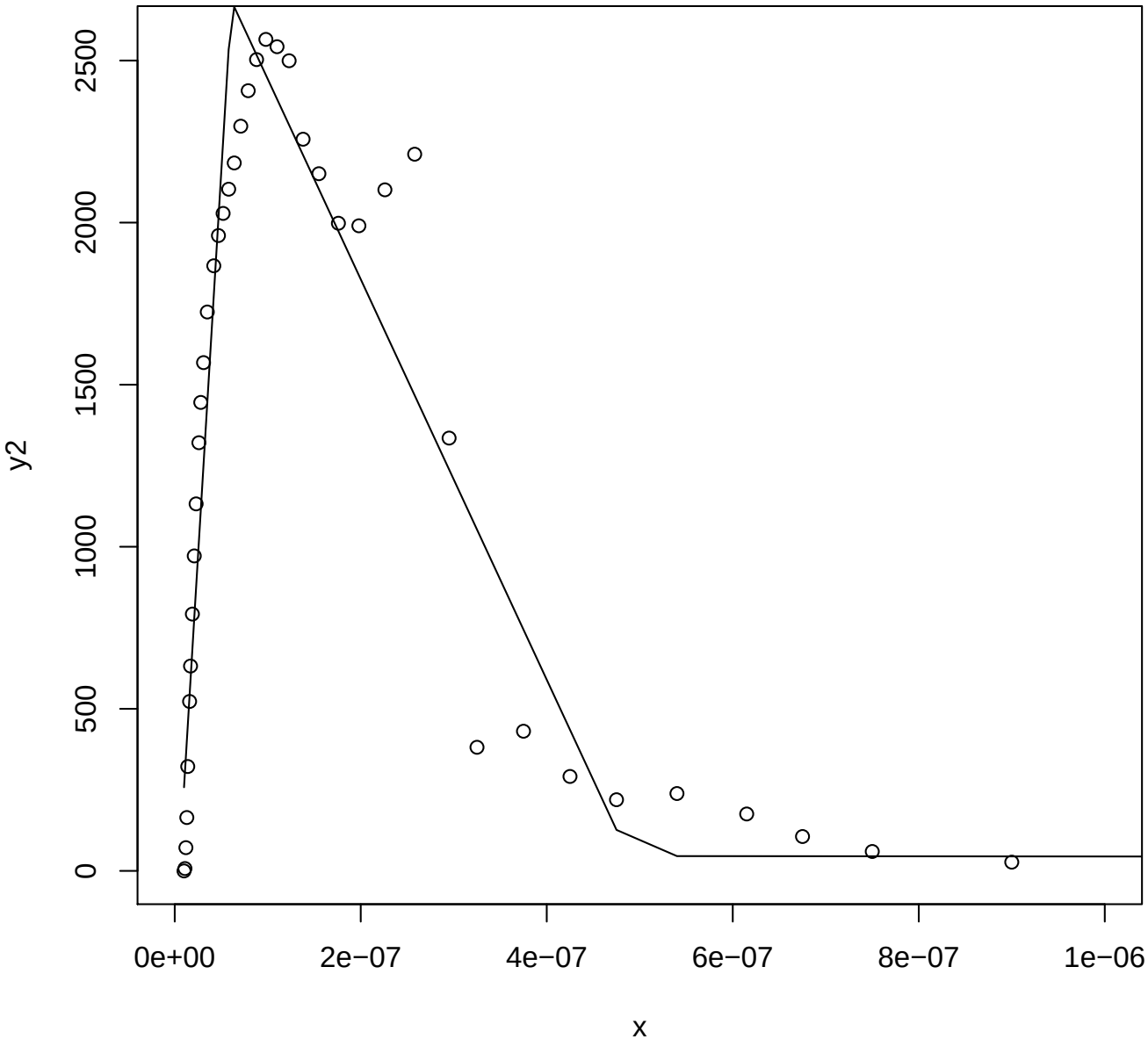
Consequences of model inadequacies

- If homocedasticity or independence are not fulfilled, least square estimators are still linear and unbiased, but are not the best. If autocorrelation is present in the residuals, the t and F tests can lead to wrong inferences.
- Although the linear model is fairly robust to the normality assumption, if normality is not adequate, the tests will only be approximately valid and it also poses problems of efficiency of estimates.
- Misspecification problems may be due to:
  - Inclusion of irrelevant variables.
  - Exclusion of relevant ones.
  - Incorrect functional form.

General ways to treat model inadequacies

1. Transformations to stabilize variance or to achieve a linear relationship.
   - Transformation on y: Box-Cox transformation, use the spread level plot to find a power transformation.
   - Tranformation on the regressor variables:
     - Using component+residual plots.
     - Multivariate Box-Cox transformation of the whole set of x's.
     - Picewise linear regression, broken-stick regression.
     - Polynomial regressors, regression splines, Fourier series.

Picewise linear regression example



General ways to treat model inadequacies

2. Model extensions (extensions of the estimation method):
   - Generalized and weighted least squares (non constant variance, autocorrelation)
   - Robust Regression (outliers, influential observations)
   - Ridge regression, partial least squares (multicollinearity)

Multicollinearity

Multiple regression attempts to separate out the effects of each independent variable holding the others constant.

If there are important linear relationships between the X variables, we have multicollinearity.

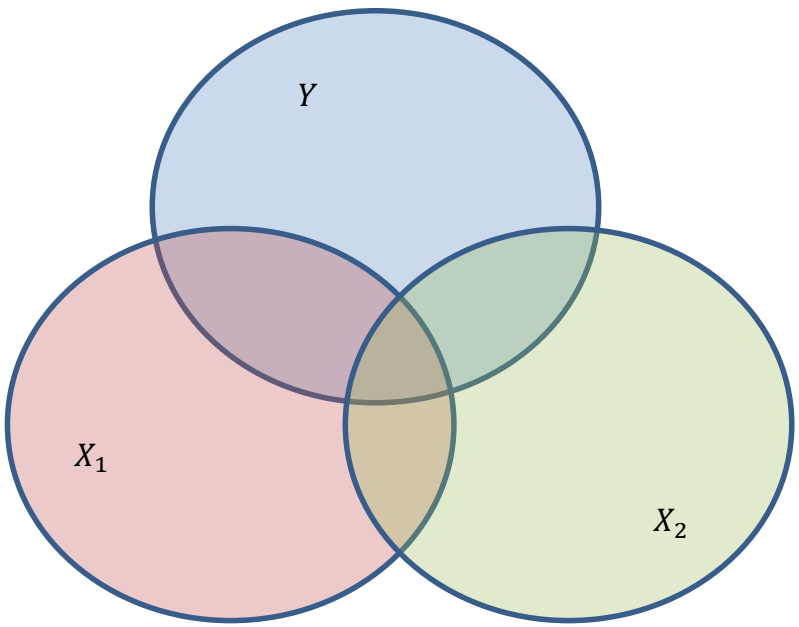
Facts about multicollinearity

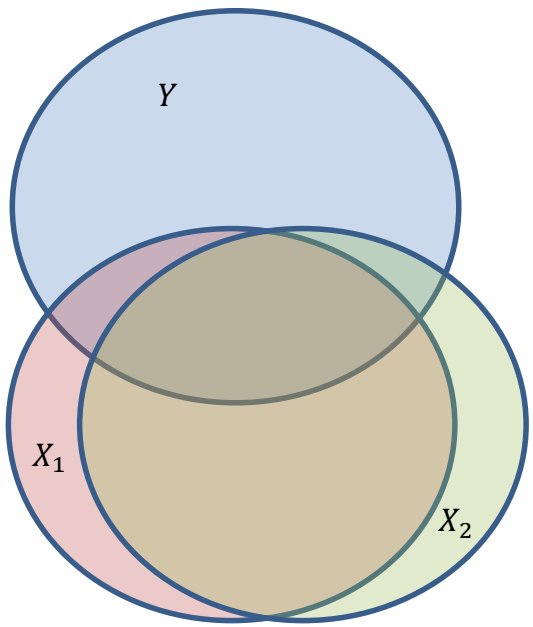
- A strong, but less than perfect, relationship among the X's causes unstable OLS coefficients, standard errors are large, reflecting the imprecision of the β's.
- If the goal is to use a regression model to predict Y, then multicollinearity is not problematic, as long as the model accounts for a high proportion in the variation in Y, the predictions should be quite accurate.
- If we wish to understand how each of the X's impacts Y, multicollinearity is potentially problematic.

Facts about multicollinearity

- Overall model may be highly significant while no (or few) individual predictors are.
- Signs of β coefficientes may lead to wrong conclusions.
- Even in the presence of collinearity, if it turns out that the standard errors are small, the degree of collinearity is irrelevant.
- Even if the standard errors are large, knowing that collinearity is the problem can only help if the study can be re-designed so that the correlations between the X's are reduced.

Multicollinearity





Detecting Multicollinearity

- Start by looking at the pairwise correlations between X's (0.8 or higher are of concern).
- Multiple correlations cofficients R^2, regressing each X on the others.
- Variance influence factors (VIFs)
- Eigenvalues from Principal Component Analysis. Perfect collinearity has an eigenvalue of 0.

Multicollinearity detection: *VIF's*

Variance Inflation Factor, for each variable X_i, como:

$$VIF_i = \frac{1}{1 - R^2_{i.others}}$$

where $R^2_{i.others}$ is the multiple correlation coefficient R^2, regressing each X on the others. $VIF's > 10$ indicate high multicollinearity.

The square root of the VIF tells the factor by which the standard error and confidence interval is inflated because of multicollinearity.

For instance, $\sqrt{VIF_i} = 20$ tells us that the standard error of coefficient β_i is 20 times higher than it would have been without collinearity. It indicates the impact of multicollinearity on the precision of β_i.

Condition Index or condition number

Based on eigenvalues of S or R (for the matrix of variables X's). Defined by:

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

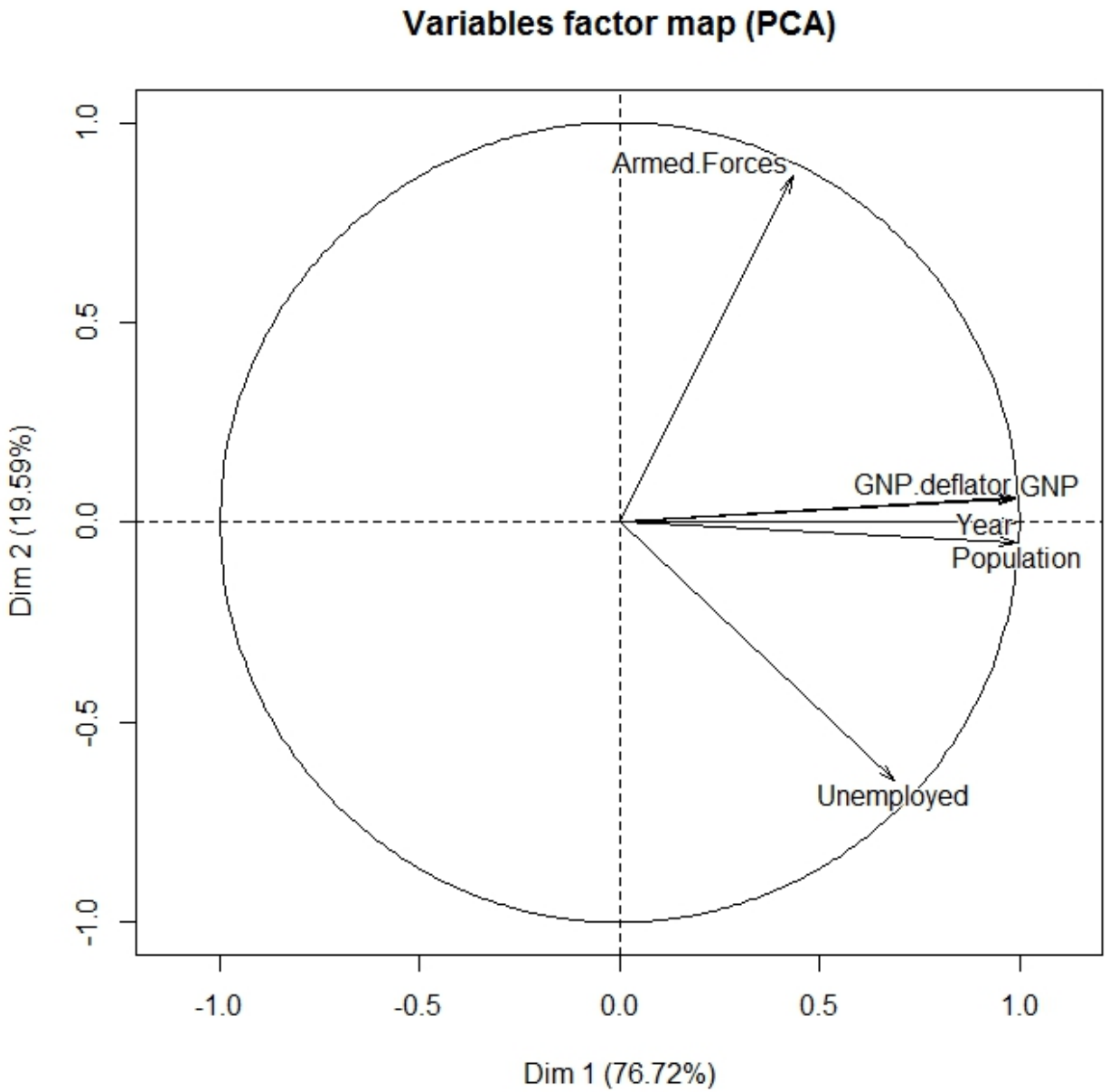
There can be defined one IC_i for each variable X_i:

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

According to *Belsley et al. (1980)*:

$$\begin{aligned} CI_i < 10 &\quad \text{Ok} \\ 10 < CI_i < 30 &\quad \text{watch out!} \\ CI_i > 30 &\quad \text{trouble!} \\ CI_i > 100 &\quad \text{disaster!} \end{aligned}$$

Detecting multicollinearity



Handling multicollinearity

- Eliminate some highly correlated variables.
- Principal Component Regression
- Ridge Regression

Variable Selection Methods (Model building)

1. There is no *best* regression model.
   - There are a number of ways we can choose the *best*, they will not all yield the same results.
   - There are other potential problems with the model that might have been ignored while selecting the *best* model.
2. All possible regressions (combinatorial).
3. Stepwise regression methods:
   - Forward selection.
   - Backward elimination.
   - Stepwise regression.

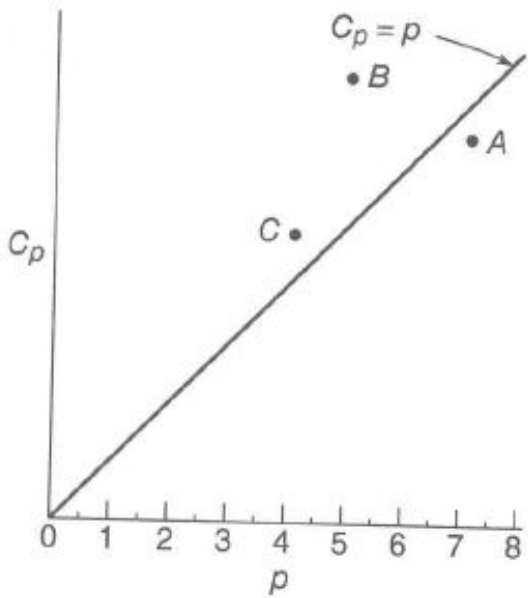
Variable Selection Methods

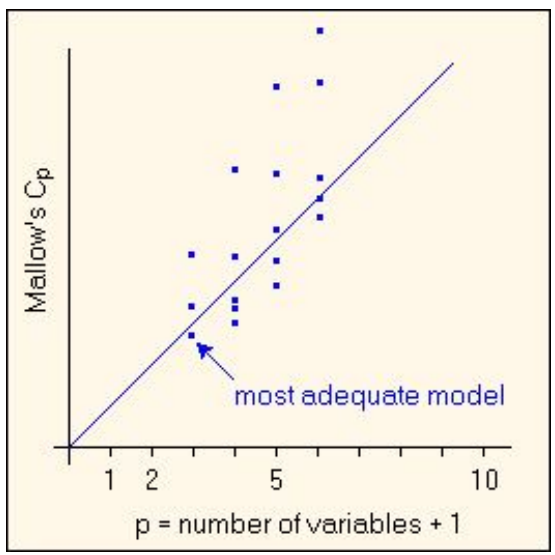
4. To evaluate and compare different candidate models we can use several information criteria:
   - Adjusted R_a^2 (largest)
   - Akaike's information criterion (AIC) (smallest, largest reduction)
   - Schwarz Bayesian information criterion (BIC) (smallest)
   - Mallow C_p criterion ($C_P \approx p$)

   For a fixed number of parameters, these criteria are monotonic in RSS (residual sum of squares, the *error* of the model).

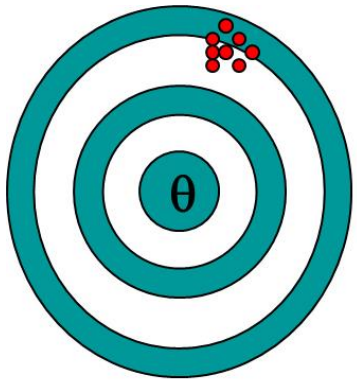
A note on Mallow's C_p criterion

It combines terms measuring **squared bias** and **variance**. It is helpful to visualize the plot of C_p as a function of p. Regression equations with little bias have values of C_p falling near the line $C_p = p$. Those with substantial bias will fall above this line. Generally, **small values of C_p are desirable**. In the figure, it may be preferable to accept some bias (point C) to reduce the average error in prediction.

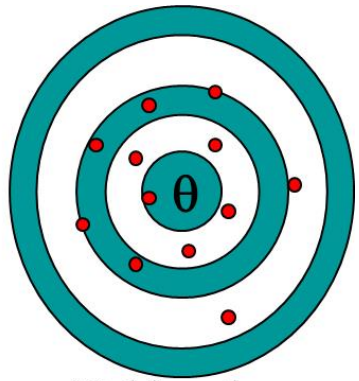




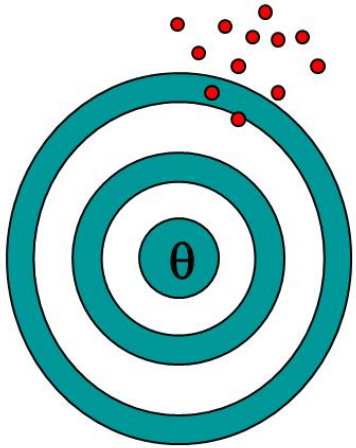
A note on bias and variance of an estimate



Biased
Low variance



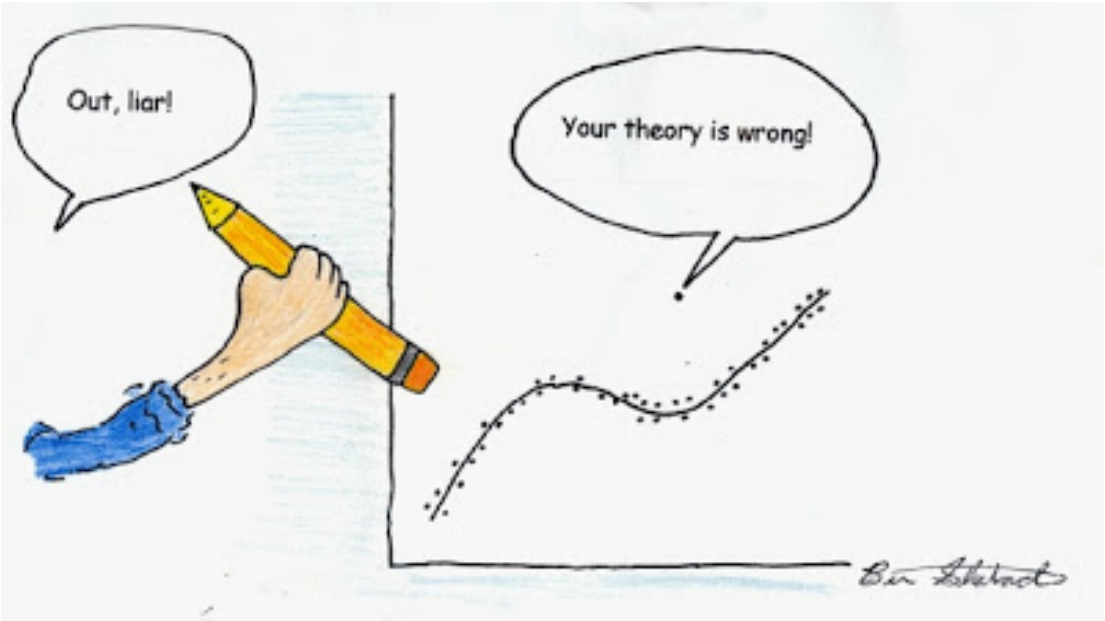
Unbiased
High variance



Biased
High variance

Unbiased
Low variance

Influential Observations



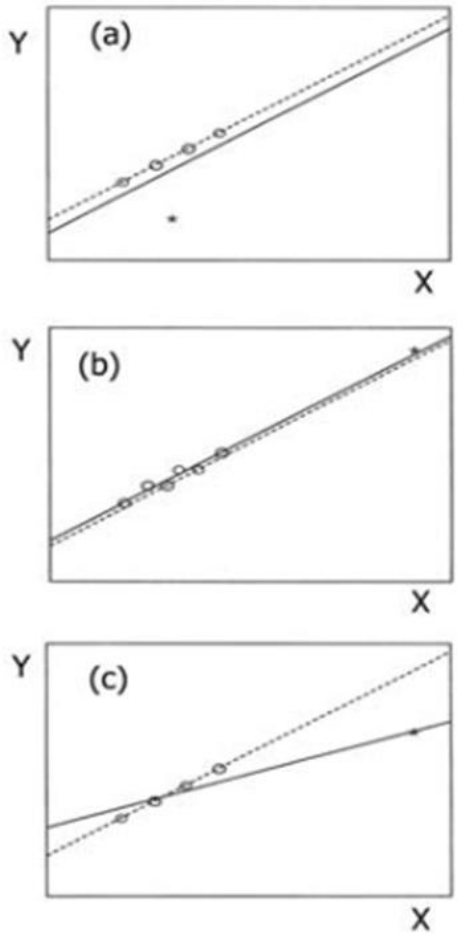
Types of unusual observations

- Regression Outlier (high residual): is an observation that has an usual value of Y, given its value of X. Neither the X nor the Y values are necessarily unusual on their own. They usually have large residuals but do not necessarily affect the regression slope coefficient.
- Leverage: an unsual X value, far from the mean of X's, has a leverage on the regression line. The further it sits from the mean of X, the more leverage it has. However, high leverage does not necessarily mean that it influences the regression coefficients (*good leverage points*).
- Influential observations: An observation with high leverage that is also a regression outlier will strongly influence the regression line. The line (plane, hyperplane) chases the observation.

$$\text{Influence} = X \text{ leverage} \times Y \text{ residual}$$

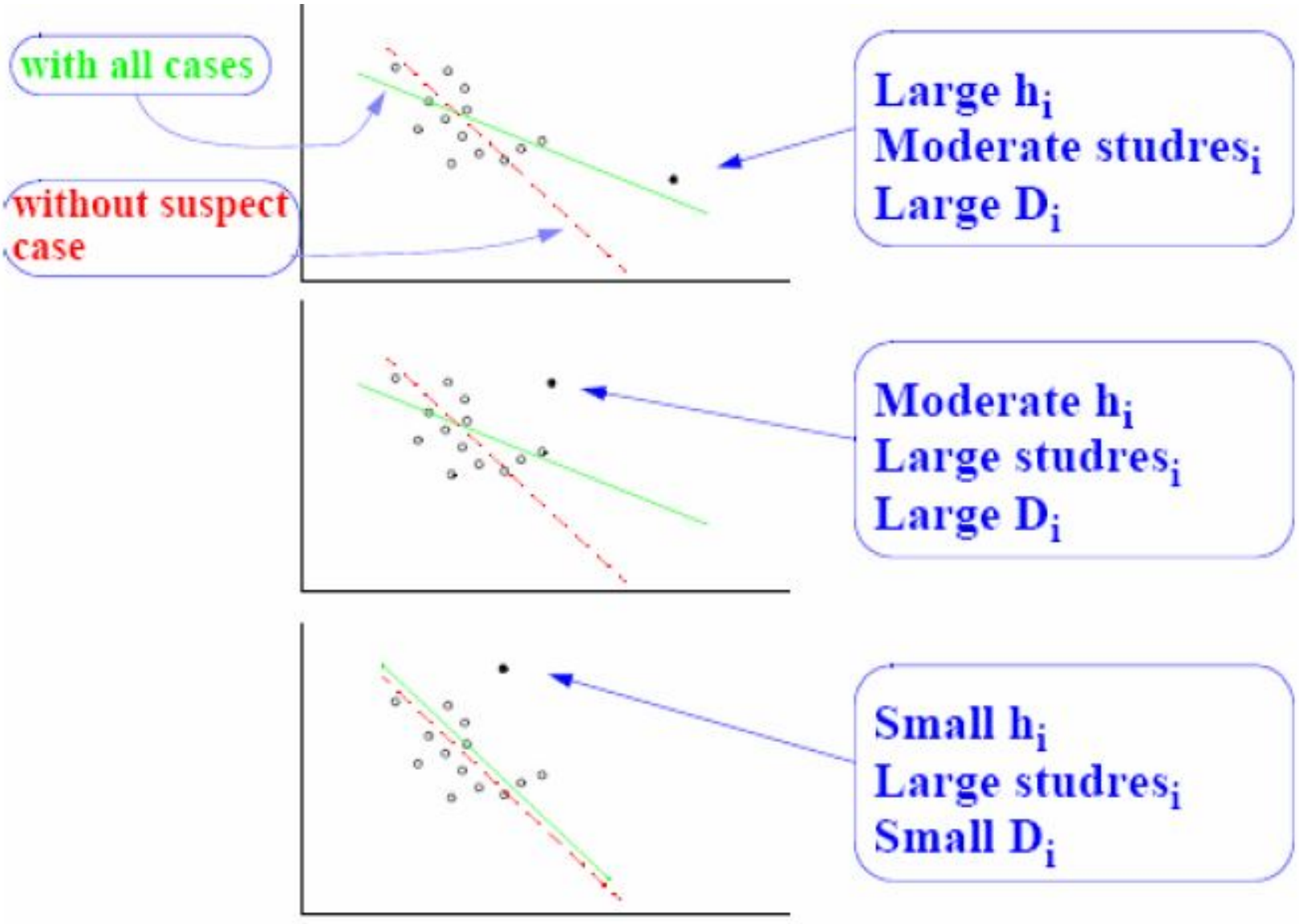
Example

- **Figure (a): Outlier without influence.** Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range
- **Figure (b) High leverage** because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has **no influence**
- **Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value)** results in strong influence. When this case is deleted both the slope and intercept change dramatically.

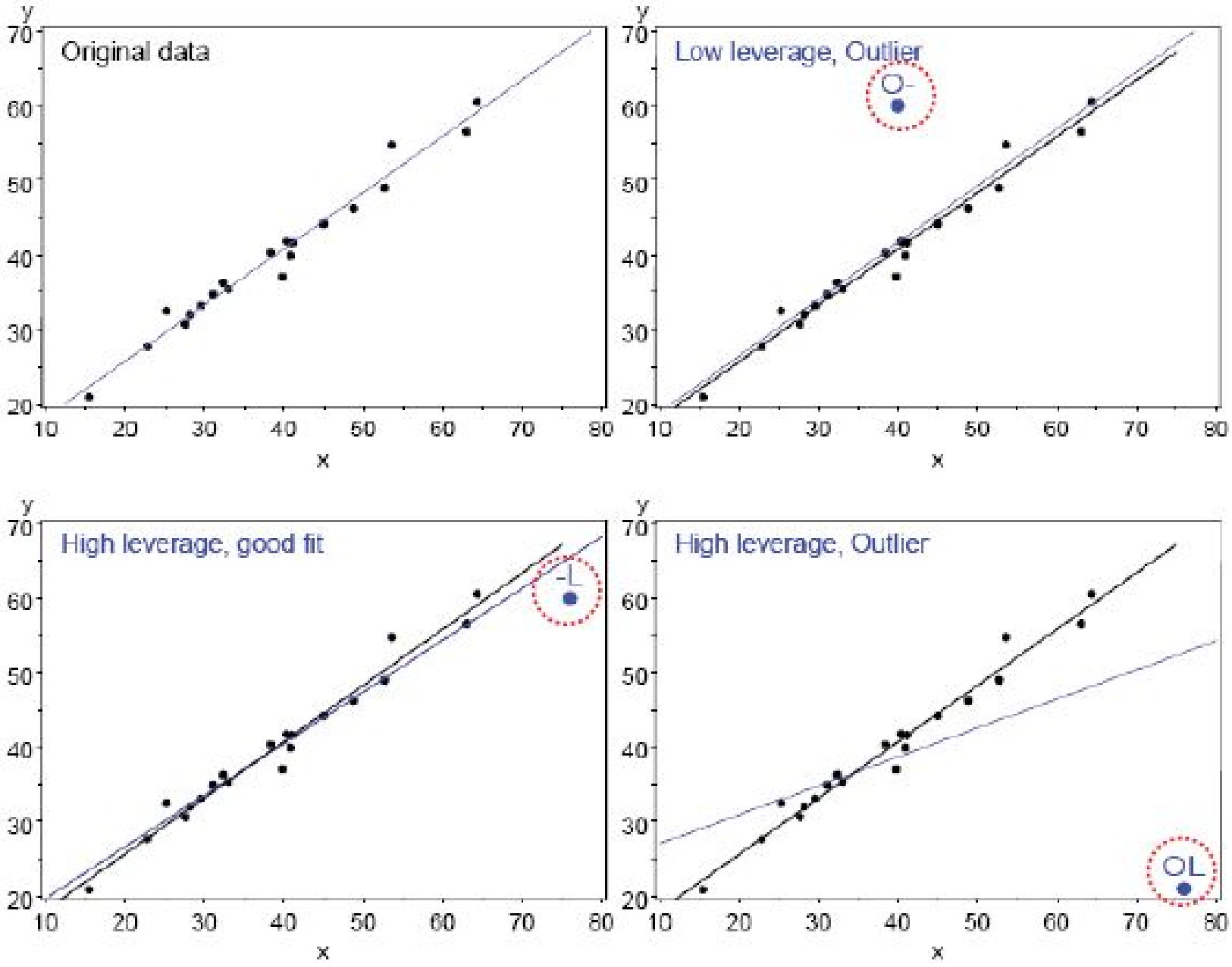


Adapted from Figure 11.1 (Fox, 1997)

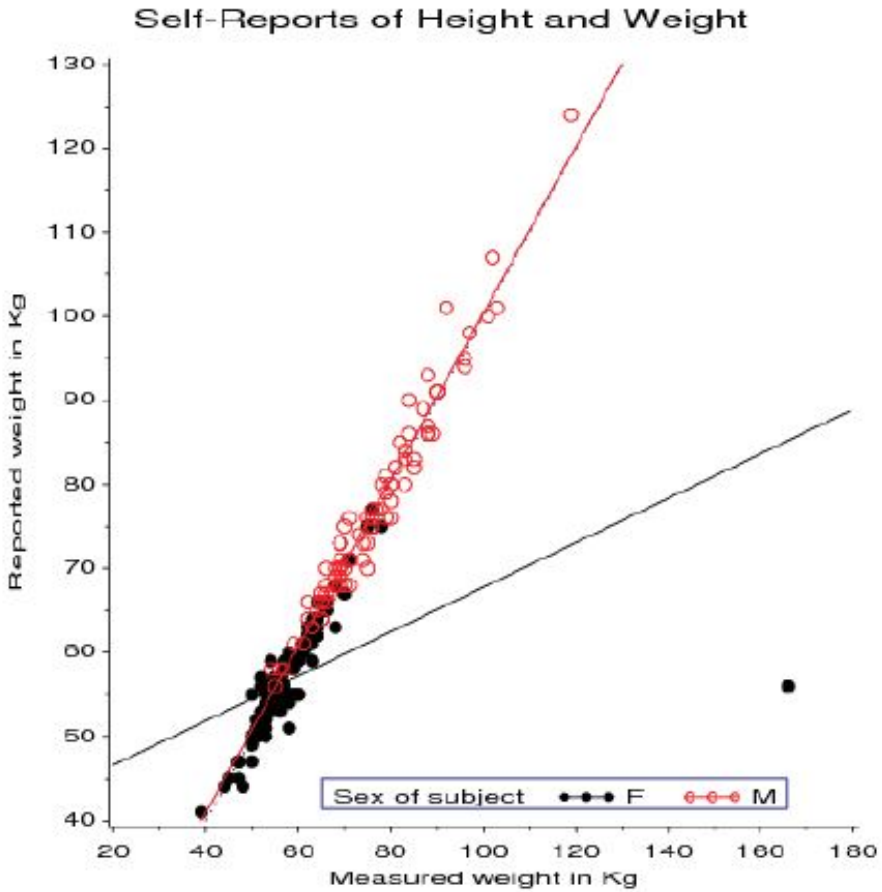
Example

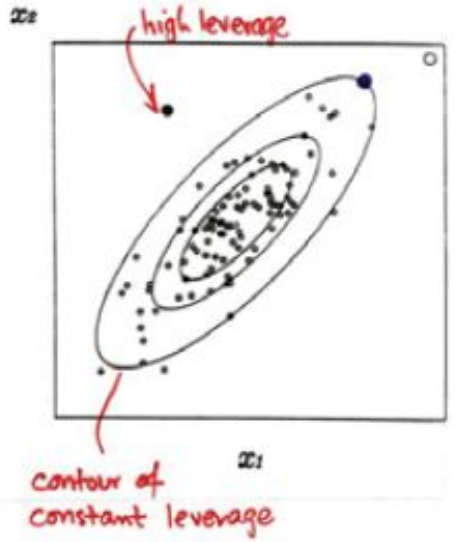

with all cases
without suspect case
Large h_i
Moderate studres$_i$
Large D_i
Moderate h_i
Large studres$_i$
Large D_i
Small h_i
Large studres$_i$
Small D_i

Example


Original data
Low leverage, Outlier
O-
High leverage, good fit
-L
High leverage, Outlier
OL
y
x
70
60
50
40
30
20
10
20
30
40
50
60
70
80

Example


Self-Reports of Height and Weight
Reported weight in Kg
Measured weight in Kg
Sex of subject
F
M
130
120
110
100
90
80
70
60
50
40
20
40
60
80
100
120
140
160
180


high leverage
contour of
constant leverage

Measuring unusualness

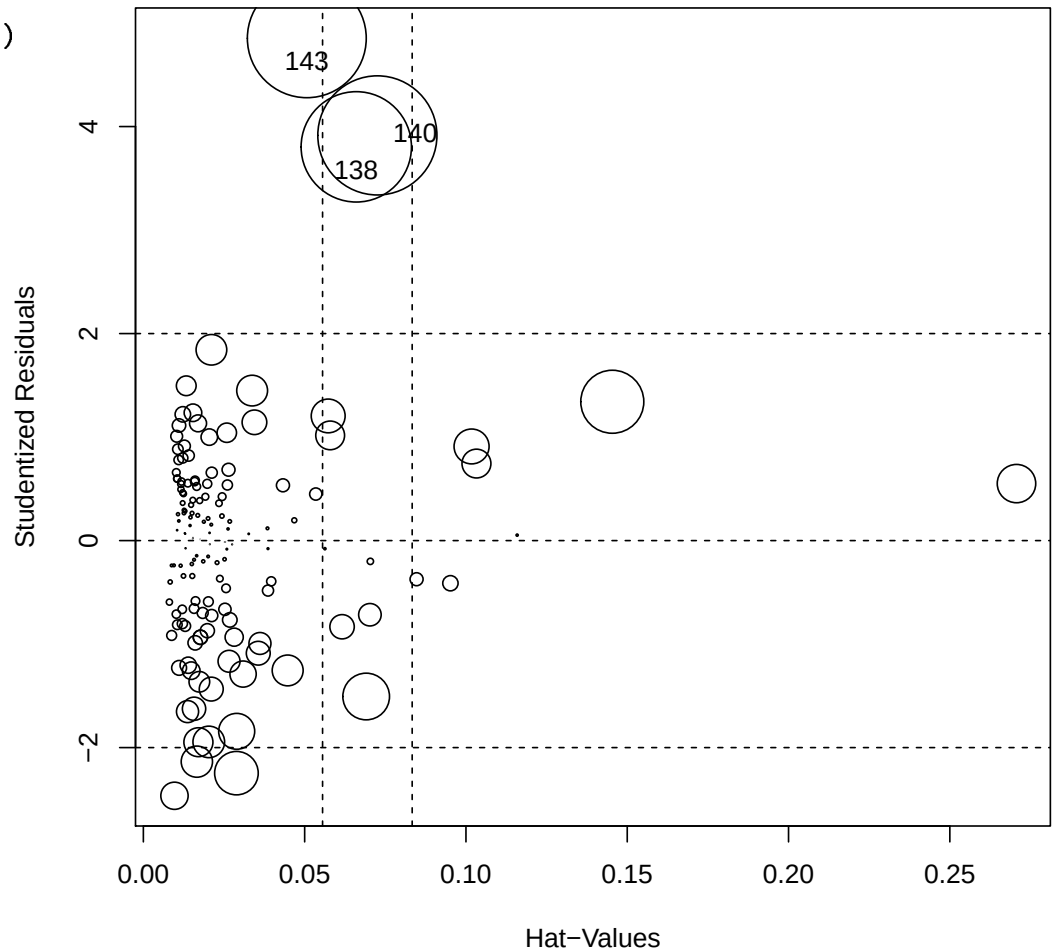
- Observations that have studentized residuals outside the range ± 2 are considered statistically significant at the 95 % level.
- Hat values (h_i) measure leverage. It measures the distance from the centroid point of all X's (point of means). We will consider observations with $h_i > 2\frac{(p+1)}{n}$. In small samples, $h_i > 3\frac{(p+1)}{n}$.
- Cook's Distance: combines a measure of discrepancy and leverage. Criterion $D_i > 1$ is risky, it can leave out noteworthy points. We will prefer:

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

Don't worry about criteria cutoffs, R is doing that for us

Example with R

```
> influencePlot(modelo3, id.method="identify")
       StudRes         Hat     CookD
85   0.5514224 0.270659854 0.1683773
88   0.7447086 0.103256779 0.1265533
106 -2.4647672 0.009647083 0.1194859
132  0.9101773 0.101745129 0.1532566
144  1.3413747 0.145365639 0.2758193
```



How to handle influential cases

- Unusual observations may reflect miscoding, then the observations can be rectified or deleted completely.
- Outliers are sometimes of substantive interest:
  - If only a few cases, we may decide to deal separately with them.
  - Several outliers may reflect model misspecification, i.e., an important variable that accounts for the subset of the data that are outliers has been neglected.
- Use methods that downweight outliers (robust methods). Often, these methods results to estimates similar to a model omitting influential cases, because they assign very low weight to highly influential cases.
- Sensitivity analysis: compare results with and without the observations and analyse the effect on conclusions.