Stress Free Stats

Regression Warnings

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Overview

- Model mis-specification
- Diagnostics
- Measurement error
- Remedies
- Influential data points
- Diagnostics
- Remedies

Model Mis-specification

- As mentioned, mis-specifying your model is the gravest mistake you can committ
- Missing relevant predictors should be added to the model in order to capture the true relationships.
- WE MUST RELY ON THEORY to know what is missing!
- But our theories are often too general
- Since we do not know what is missing, we put in everything
- This is very bad: overspecifying the model $->$ collinearity and inflated standard errors
- Can show relationships which are just random (not causal)

Incorrect Functional Form 1

True: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$
Estimated: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

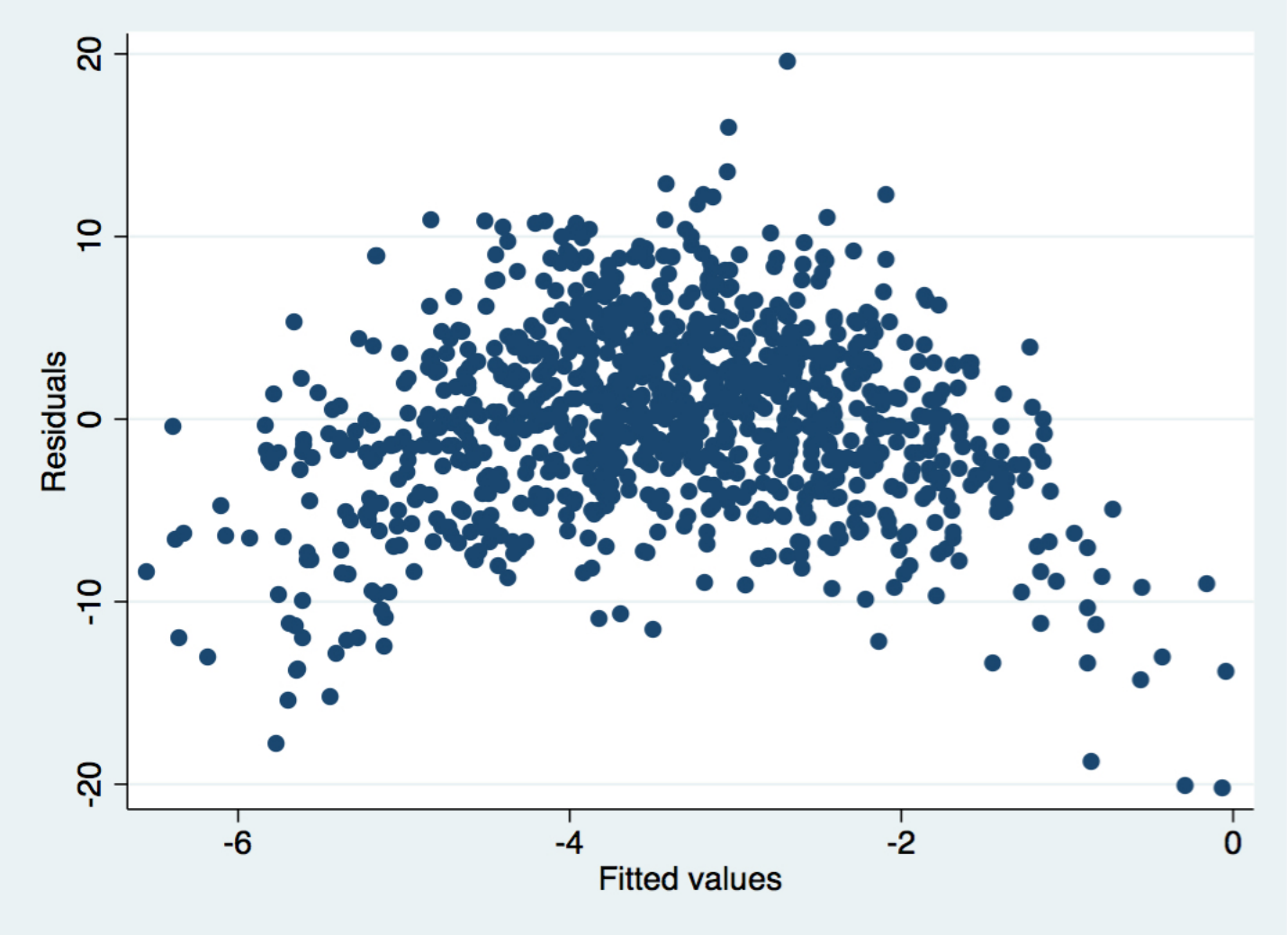
- Parameter estimates will be *biased* and *inconsistent* and standard errors may also be biased.
- Selecting a wrong functional form is like omitting a relevant predictor, which produces bias.
- The omission of z from the estimated model biases the estimate of $\hat{\beta}_1$ in the following way:

$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov[x,z]}{Var[x]}$
if $\beta_2 > 0$ and $Cov[x,z] > 0$ then positive bias
if $\beta_2 < 0$ and $Cov[x,z] > 0$ then negative bias
if $\beta_2 < 0$ and $Cov[x,z] < 0$ then positive bias
if $\beta_2 > 0$ and $Cov[x,z] < 0$ then negative bias

Incorrect Functional Form 2

True: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 r_i + \beta_4 z_i + \epsilon_i$

Estimated: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$



Diagnostics

- Basic logic is: if there is some systematic shift in residuals versus fitted values of our model, this model is misspecified.
- We can try to find predictors that can explain the residual variation – these predictors are the missing variables of the original model
- Plot residuals versus fitted values and look for an even band. If the band is curved, that suggests a missing quadratic term etc.

R: `plot(model$residuals, model$fitted)`

Measurement Error

- Implicit assumption of OLS: no measurement error. In reality, error exists.
- **Random error on Y**
  - The estimator will be *unbiased* since the randomness of error means that its mean remains =0. But *standard errors will be inflated* since error increases the estimator variance
- **Systematic error on Y**
  - Causes a bias in the intercept coefficient, but not in the slopes. It 'shifts' the function along the y axis.
- **Random error on X**
  - Error on X: values of X are correlated with the error term.
  - We get biased and inconsistent estimates of the coefficient.
  - The effect of the predictor with error is underestimated, which is called *attenuation bias*.

Remedies

- Instrumental Variables
  - Find a variable that is correlated with X but not with ϵ.
  - This is *not* a 'proxy', which is a variable used to measure latent – or unmeasurable, unobservable – variables.
- Structural Equations
  - A matter for another class

Influential Data Points

- Outliers and influential data – that is observations that 'stick out' – can significantly alter our statistical results.
- **Outlier** is an atypical value on y,
- **Leverage value** is an atypical value on x,
- **Influential value** is an atypical value in both x and y

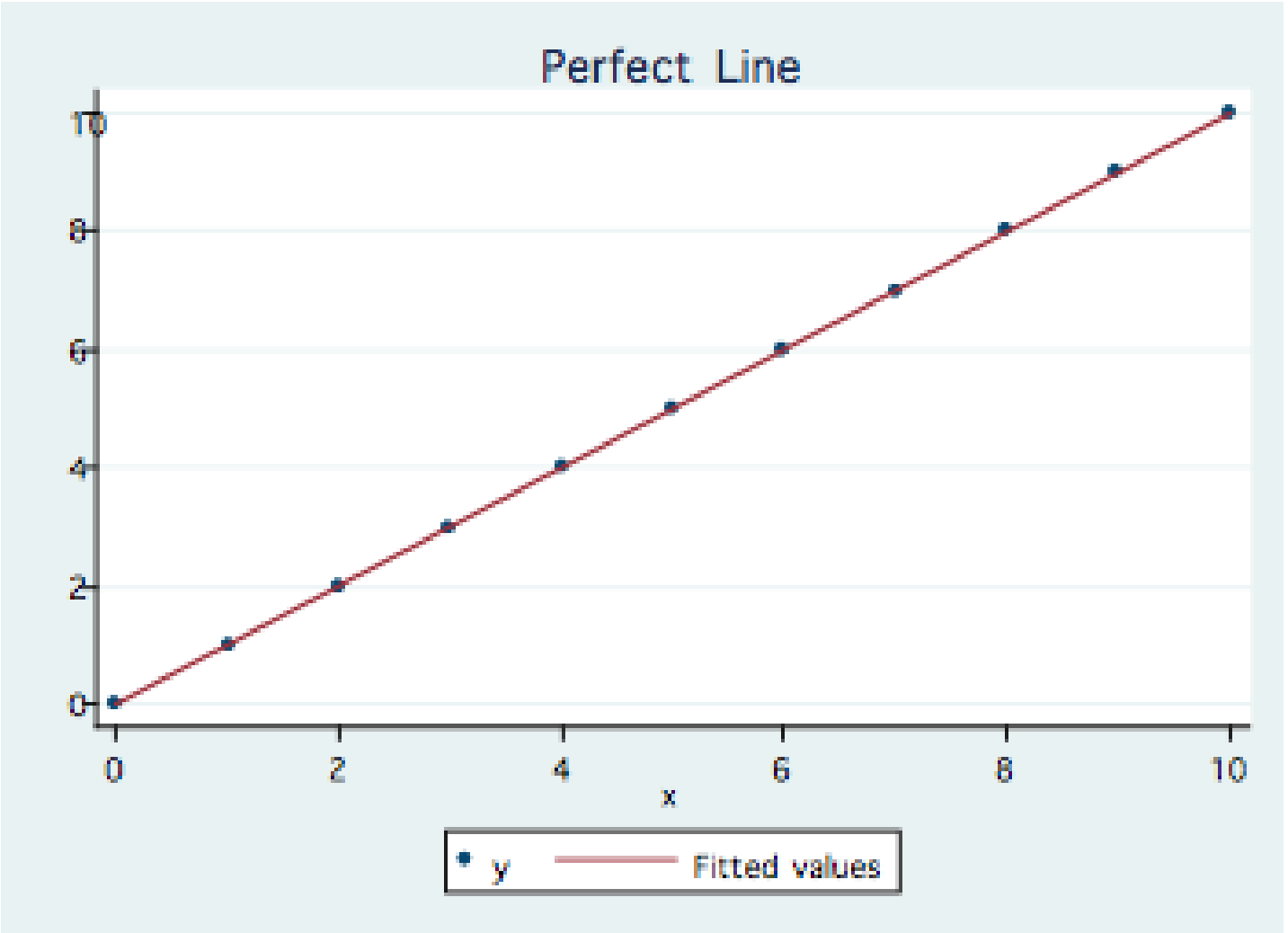
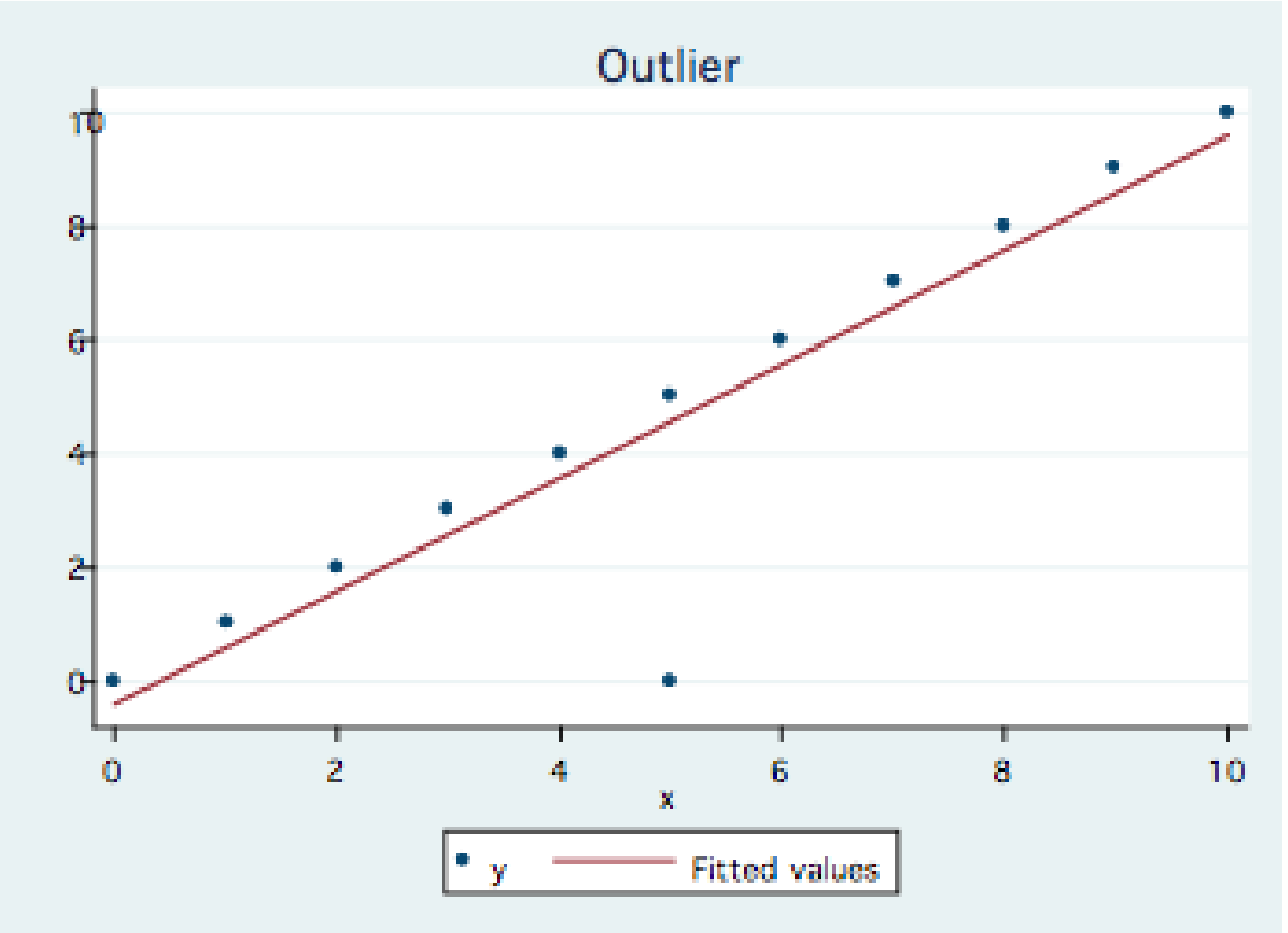
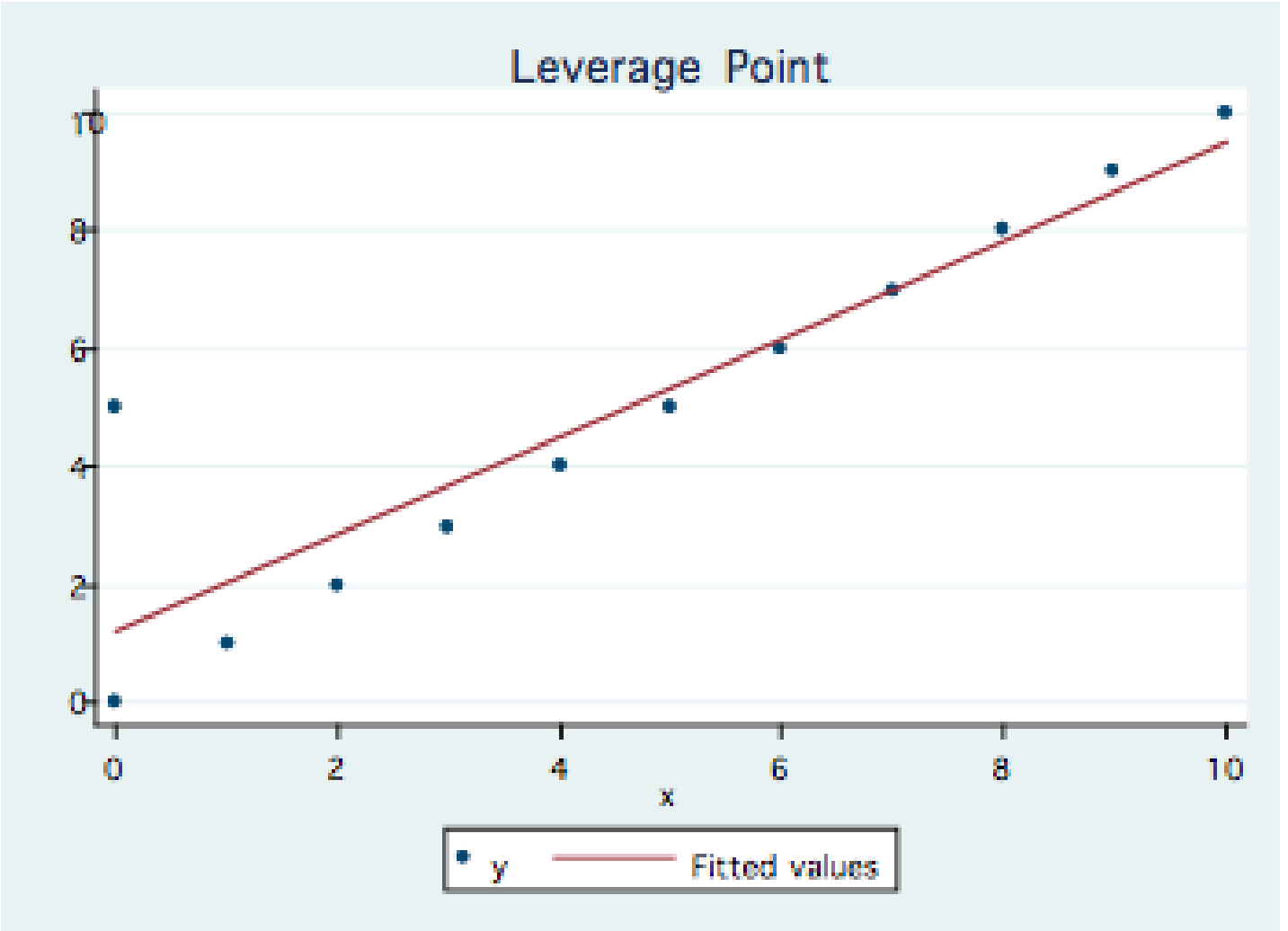


Figure: Perfect Line

Outlier is a point with a large residual. An outlier does not influence the slope of the regression, but it effects the intercept (it will affect the value of y when the regression line passes through the origin). Remember that: $u_i = (Y_i - \hat{Y})$



Leverage point is disproportionately distant from the bulk of the values of the predictor(s). Is capable of pulling the regression line towards itself, thus distorting the slope. In such a case, it is an influential data point.



Influential point is both an outlier and a leverage point. It modifies both the slope and the intercept of your regression line and may completely drive all the results.

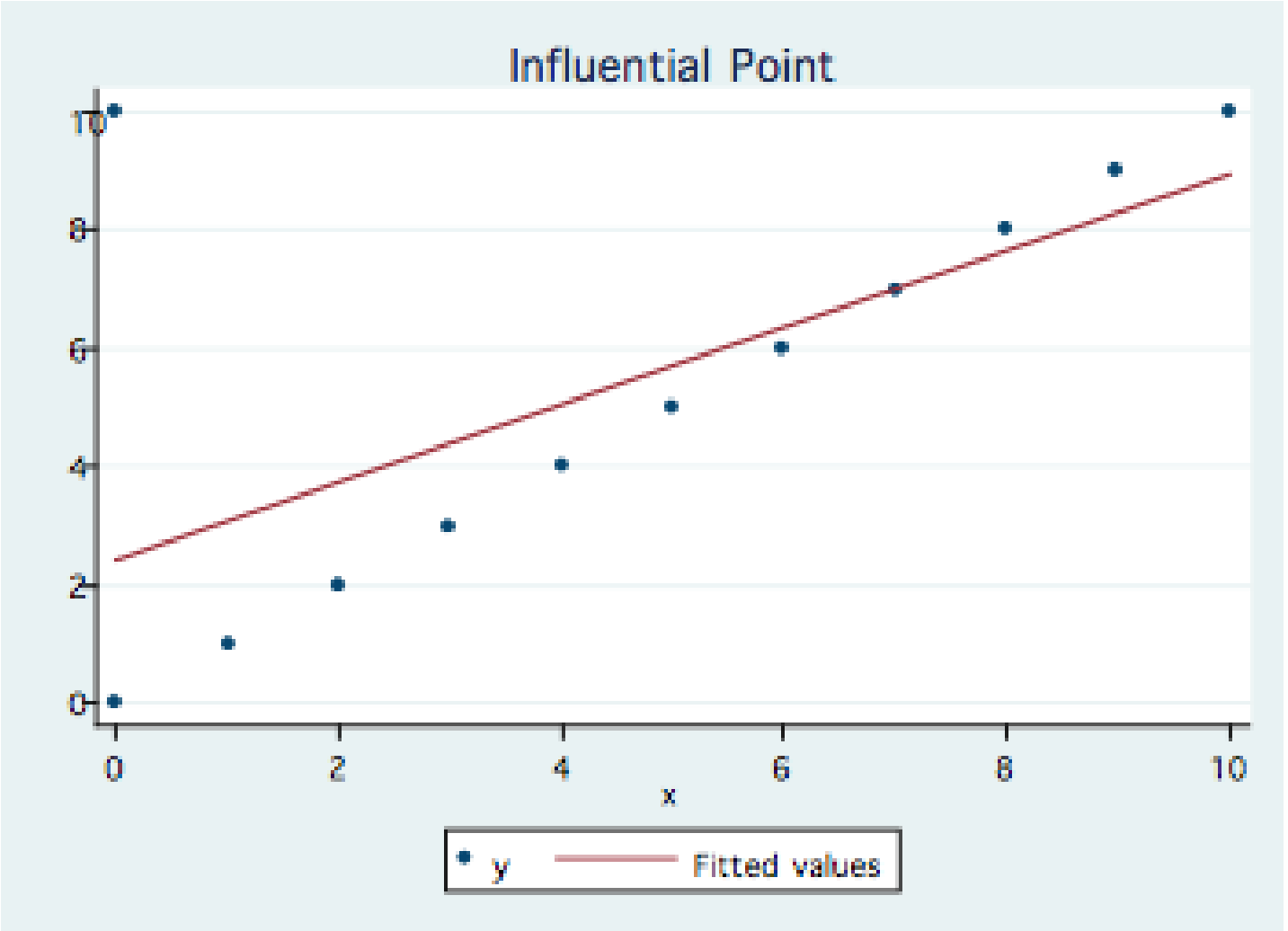


Figure: Influential Point

Diagnostics

DFITS

- Tells us the influence of an observation on predicted values (regression line)
- It runs regressions with and without each observation
- A highly influential observation is one where $|DFITS| > 2\sqrt{\frac{k}{N}}$
- **R:**
  - `dffits(modelname) # creates observations of influence`
  - `large<-dffits(modelname)>2*sqrt(k/n) # creates object`
  - `list(large)`
- This will tell you which particular observations are influencing your results

Diagnostics

DFBETAs

- Tells us the influence of an observation on each β coefficient
- A highly influential observation is one where $|DFBETA| > 2/\sqrt{n}$
- **R**
  - `dfbetasPlots(model) # produces a plot for each coefficient (car library)`
  - `dfbetas(model) # creates observations of influence on each coefficient`
  - `DF<-abs(dfbetas(model) # assign abs. dfbetas to matrix DF`
  - `list(DF>2/sqrt(n)) # identify influential observations on all coeffs.`
- This well tell you which particular observations are influencing β_1

Remedies

- The RULE is not to select on your dependent variable, so throwing the influential point out is not a solution!
  1. If the observation belongs to the target population and there is no coding error, throwing out the point carries worse problems with the estimation.
  2. Present the original results as main results and in the appendix show the results with the influential point removed, showing the influence of the point in your results.
  3. Find other estimator better than OLS, an alternative fit measure (paying the cost of being less efficient than OLS) – this is a matter for another class.