Lecture 34

Residuals

Recall that the residual is the difference between the observed value of the response variable and the predicted value is

$$e_i = y_i - \hat{y}_i$$

It turns out that because residuals show how far the observed data falls from the regression line they are useful for assessing how well a fitted regression line models the observed data. Note also that the mean of the residuals is always zero.

Specifically, the tools we use are called a residual plots.

Residuals plot

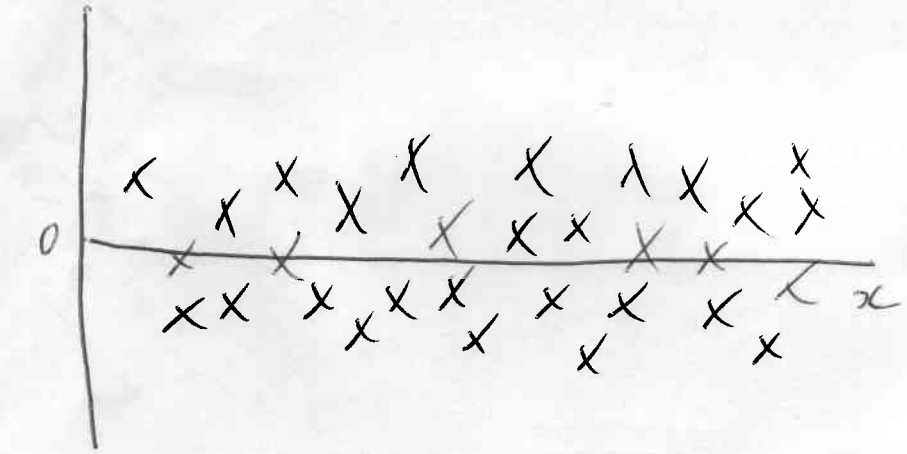
There are plots of residuals against ____

- residuals vs explanatory variable
- residuals vs predicted values

Note never plot residuals versus response variable.

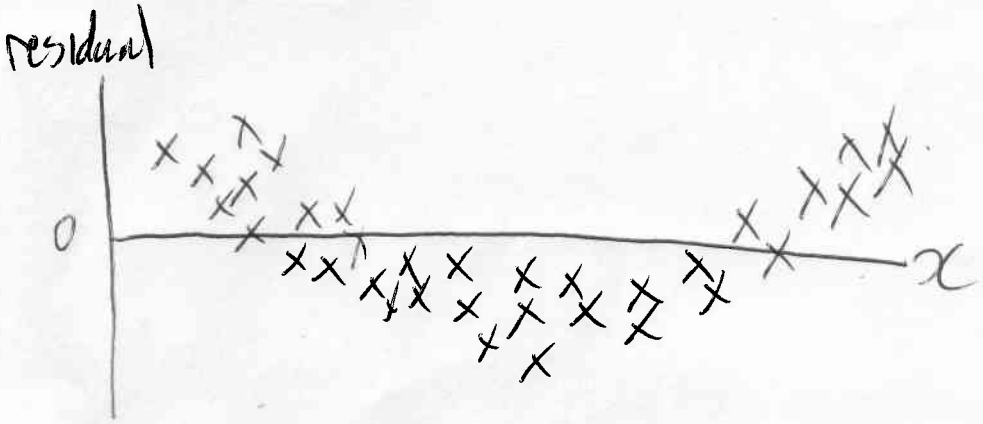
They are used to visually assess the fit of the regression line.

Ideal plot

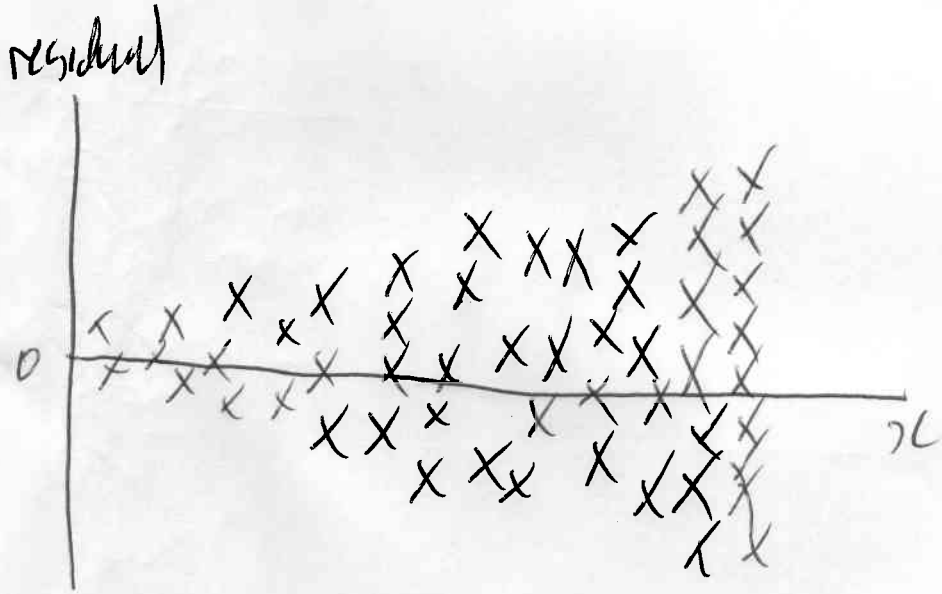


- Shows irregular pattern if no clear pattern.
- Even vertical spread across all values on x axis

Non ideal plots



- pattern (this particular pattern indicates a curved relationship exists between y and x that is not modelled by the linear regression line)



- Uneven vertical spread

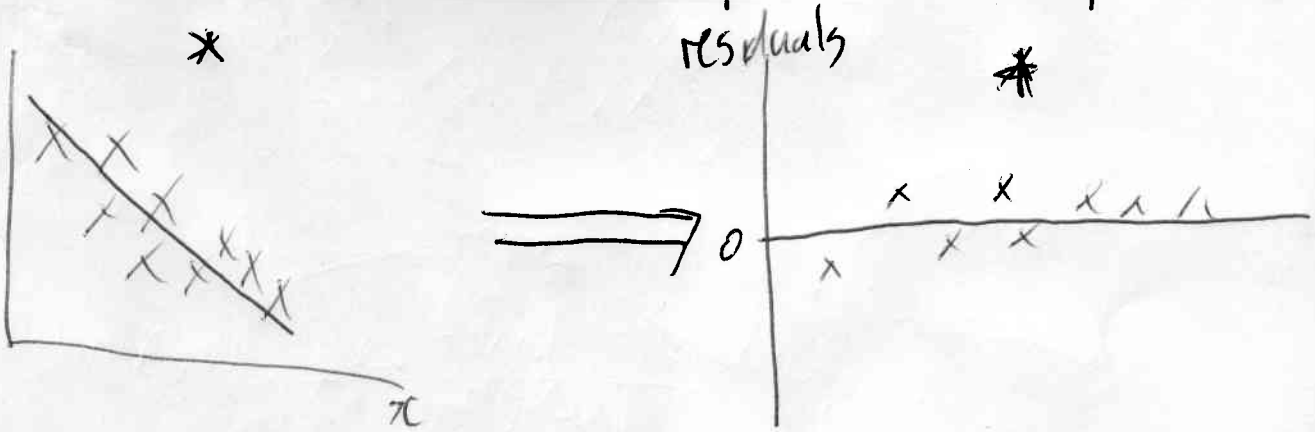
(this particular relationship indicates a transformation of y variable might be useful)

Outliers / Influential observations

An outlier is an observation that lies outside the overall pattern of the other observations.

An observation is influential if removing it would have large effect on the fitted regression line (perhaps even completely changing the equation of the regression line)

— Outliers in y tend to show up in residuals plots as



- outliers in x ~~are~~ often (but not always) influential

eg