# Checking Adequacy of Model

Assuming Regression model is adequate we CAN use it to make inferences, see previous pages. Before doing so, we need to check if the assumptions hold.

- Are $\varepsilon \sim N(0, \sigma^2)$ ?

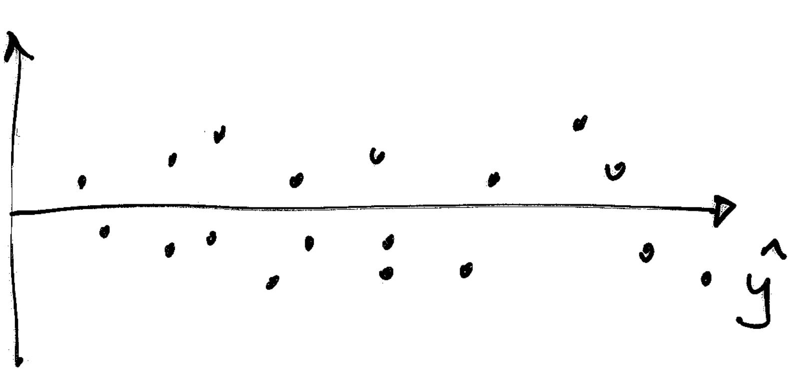
First we compute the residuals, after a, b the estimates of $\alpha$ and $\beta$ have been estimated.

**Then**

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

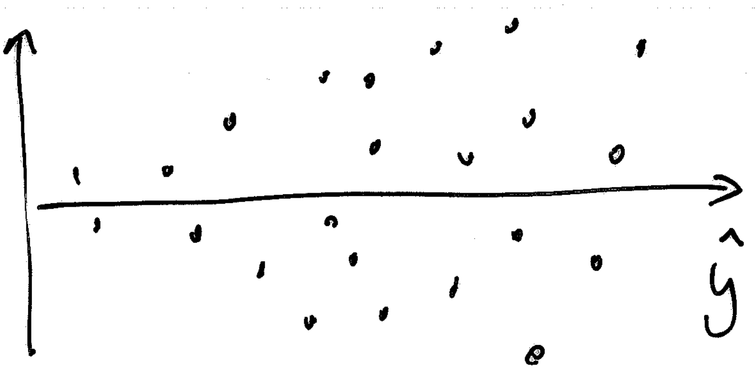
- Normplot the residuals $e_i$ to check if the normality assumption holds
- Plot residuals vs $\hat{y}_i$ (predicted values)

a



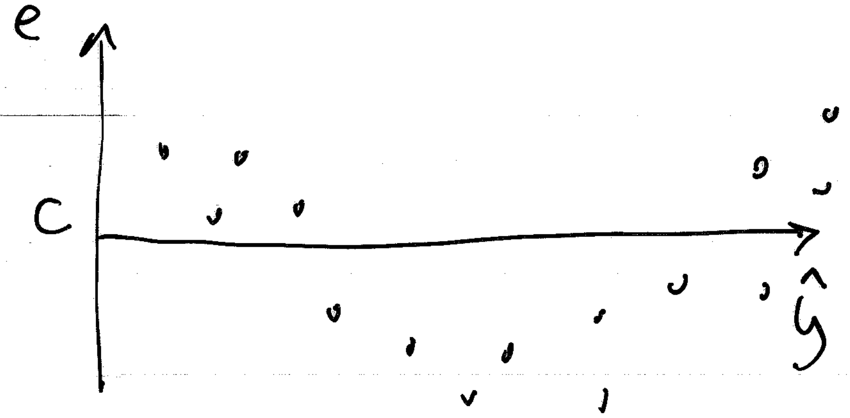

ideal "constant" band

b




decreases/
variance increases with the response and a transformation is needed

c




linear model is NOT the proper model

- Plot residuals vs time (or order when this applies) to detect any trends over time.

When variance of residuals is not constant, like in Figure 6, we need to do a transformation in order to "stabilize" the variance.

It is usual practise then to substitute in the regression

$$Y = \alpha + \beta X + \varepsilon \quad \text{or}$$

$$\hat{Y} = a + bX$$

the independent variable $X$ by a transformed one. Usually $X' = \log X$ or $X' = \sqrt{X}$ or $X' = X^k$ for some $k$, are the most common ones.

Be careful, $\log X$, $\sqrt{X}$ can be used when $X$ takes only positive values!

## Correlation

Up to now, we have assumed that the independent variable $\underline{\underline{x}}$ was assumed to be known without error. In some cases though both $x$ and $y$ have to be thought of as values of random vars $X$, $Y$.

**Example** Study relationship between input and output of a wastewater ~~the~~ treatment plant. / tensile strength and hardness of aluminum ...

In such cases data $(x_i, y_i)$, $i = 1, .., n$ are values of a pair of random variables $X$ and $Y$ having joint density $f_{X,Y}(x, y)$.

Scatterplot provides visual impression of the relation between x and y.

Closeness of the scatterplot to a straight line can be expressed in terms of correlation coefficient (numerically).

Sample correlation coefficient, $r$, is defined

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \cdot \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Where $s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$

$$s_y^2 = \sum \frac{(y_i - \bar{y})^2}{n-1} = \frac{S_{yy}}{n-1}$$

- When most of the pairs of obs are s.t. either both $(x_i, y_i)$ are above their sample means or both below their sample means the products tend to be large and positive so $r > 0$.

- When one value of the pair tends to be large when the other is small and vice versa, $r < 0$.

- $-1 \leq r \leq 1$.

1. <u>Magnitude of r</u> describes the <u>strength of a linear relation</u> and <u>sign of r</u> indicates the <u>direction</u>

   - $r = 1$ – all $(x_i, y_i)$ lie exactly on a straight line having positive slope.

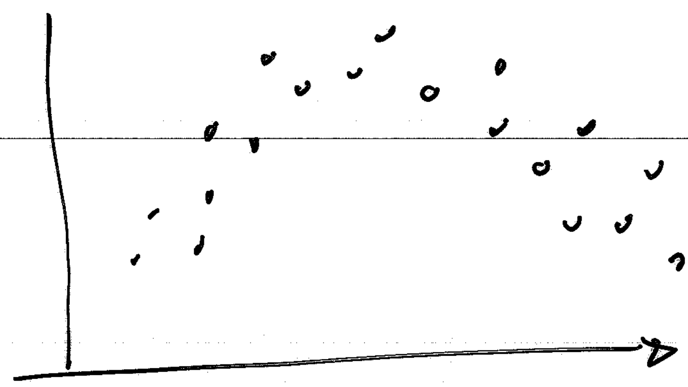
   - $r > 0$ – pattern of scatter plot runs from lower left to upper right

   - $r < 0$ – pattern of scattern plot runs from upper left to lower right.

   - $r = -1$ all pairs $(x_i, y_i)$ lie on a straight line having a negative slope.

• A value of $r$ near $+1$ or $-1$ describes strong linear relation.

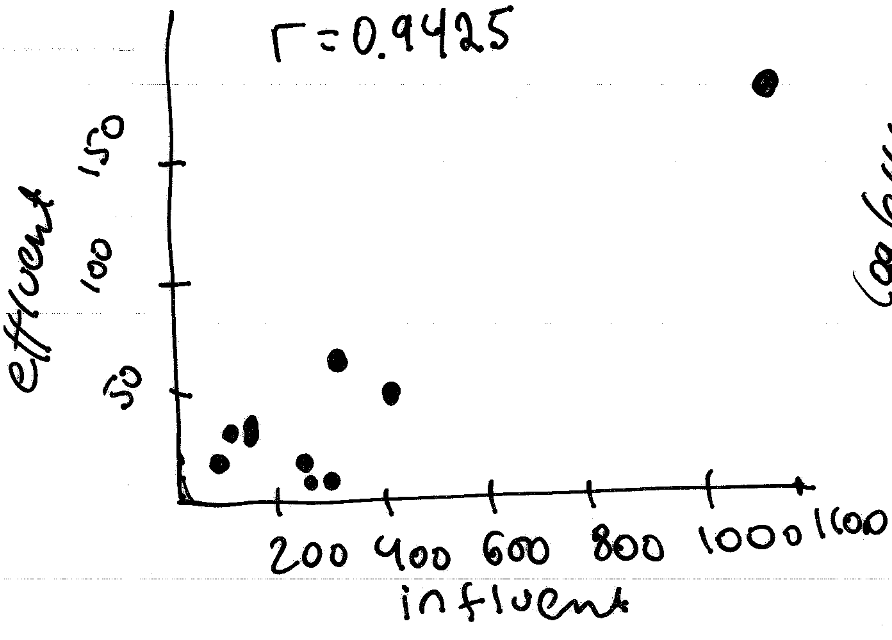
2) A value of $r$ close to 0 $\Rightarrow$ linear association is weak but there can be still strong association along for example, a curve

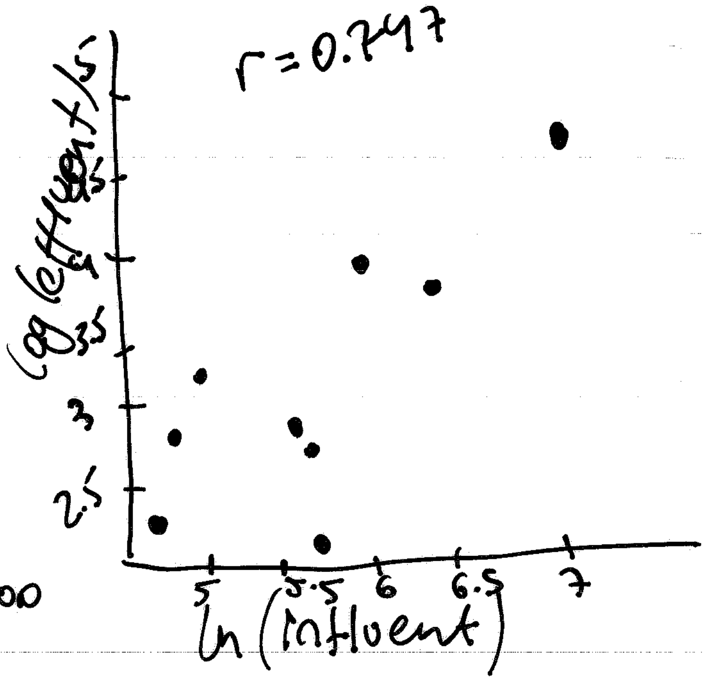


# Exploring and interpreting correlation

Example Heavy metals can inhibit the biological treatment of waste in municipal treatment plants. Monthy measurements were made at a state-of-the-art treatment plant of the amount of chronium ($\mu g/\ell$) in both the influent and effluent.

| Influent | 250 | 290 | 270 | 100 | 300 | 410 | 110 | 130 | 1100 |
|---|---|---|---|---|---|---|---|---|---|
| effluent | 19 | 10 | 17 | 11 | 70 | 60 | 18 | 30 | 180 |



Notice that $r$ is not really appropriate for the original data since the one large obs in the upper-right hand corner has too much influence. If the pair $(1100, 180)$ is dropped, $r$ drops to $0.578$.

## Correlation and Regression

There are two important relationships between $r$ and the least square fit of a straight line.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \frac{S_{xy}}{S_{xx}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \cdot b \Longrightarrow$$

$r$ and $b$ have the same sign!

total variation in y

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{xy}^2 / S_{xx} + \left( S_{\cancel{xy}yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

$$\Downarrow$$

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + \left( S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

total variability of y

Variability explained by linear relation

Residual or unexplained variability.

---

For the straight line to provide a good fit to the data, the sum of squares due to regression $\frac{S_{xy}^2}{S_{xx}}$, should be a substantial proportion of the total sum of squares $S_{yy}$.

$$\frac{\text{Sum of squares due to regression}}{\text{Total Sum of squares of } y} = \frac{\frac{S_{xy}^2}{S_{xx}}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = r^2$$

(sample correlation coefficient)$^2$.