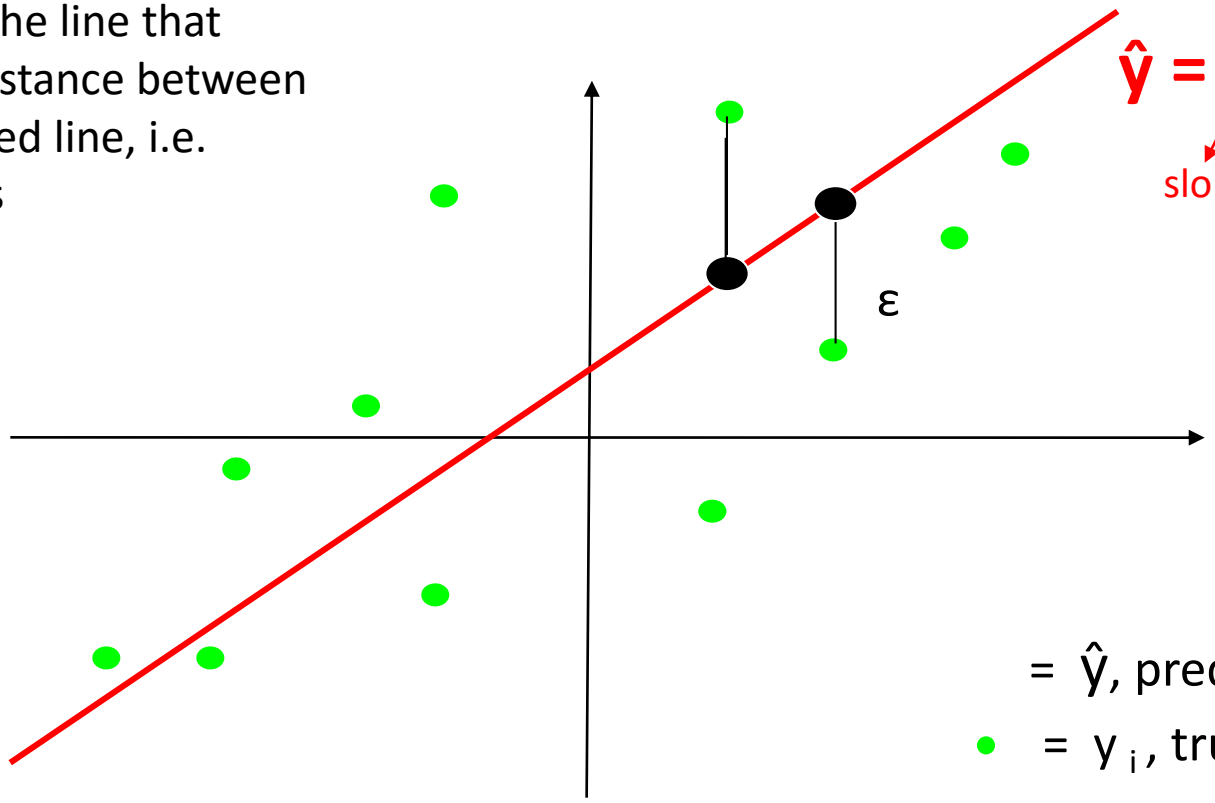
Describing Relationship between Two (and more) Quantities

Linear Regression

Best-fit Line

- The goal of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x

- This will be the line that minimises distance between data and fitted line, i.e. the residuals



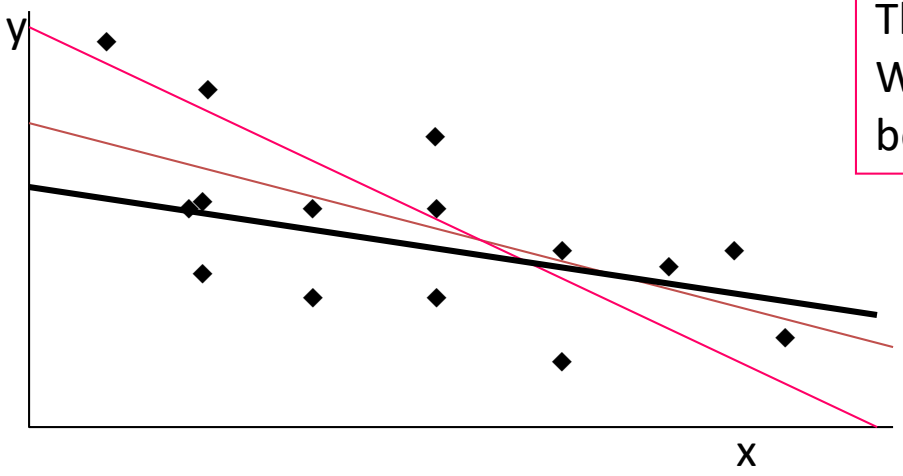
= $\hat{y}$, predicted value

= y_i, true value

ε = residual error

Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



The question is:
Which straight line fits best?

Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

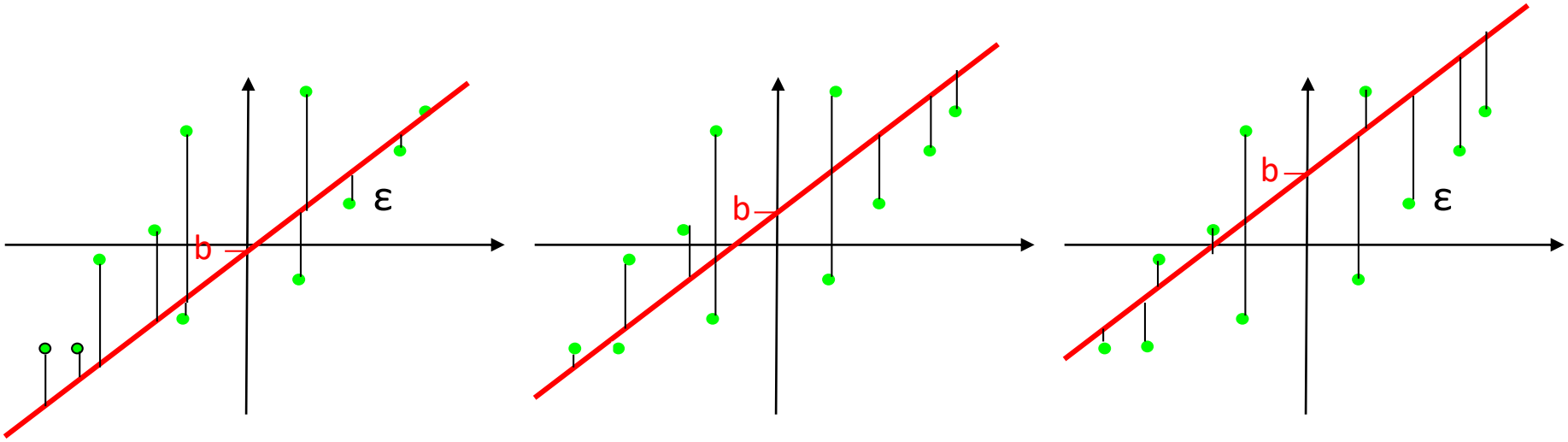
Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

we must find values of a and b that minimise

$$\Sigma (y - \hat{y})^2$$

Finding b

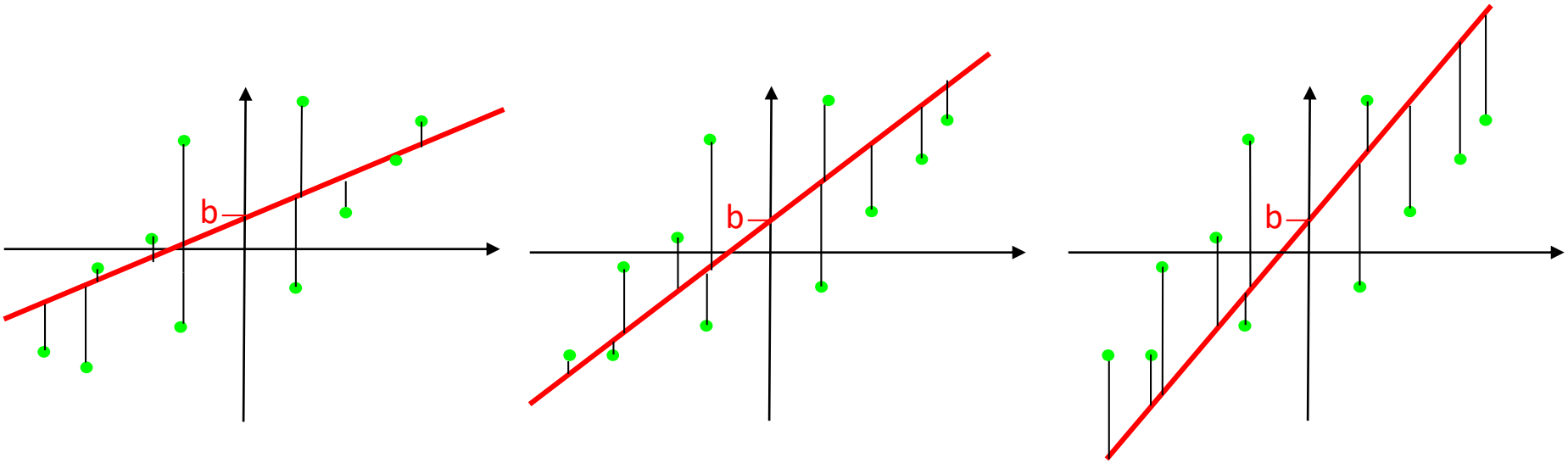
- First we find the value of *b* that gives the smallest sum of squares of deviations



Trying different values of *b* is equivalent to shifting the line up and down the scatter plot

Finding a

- Once again, we find the value of *a* that gives the smallest sum of squares of deviations



- Trying out different values of *a* is equivalent to changing the slope of the line, while b stays constant

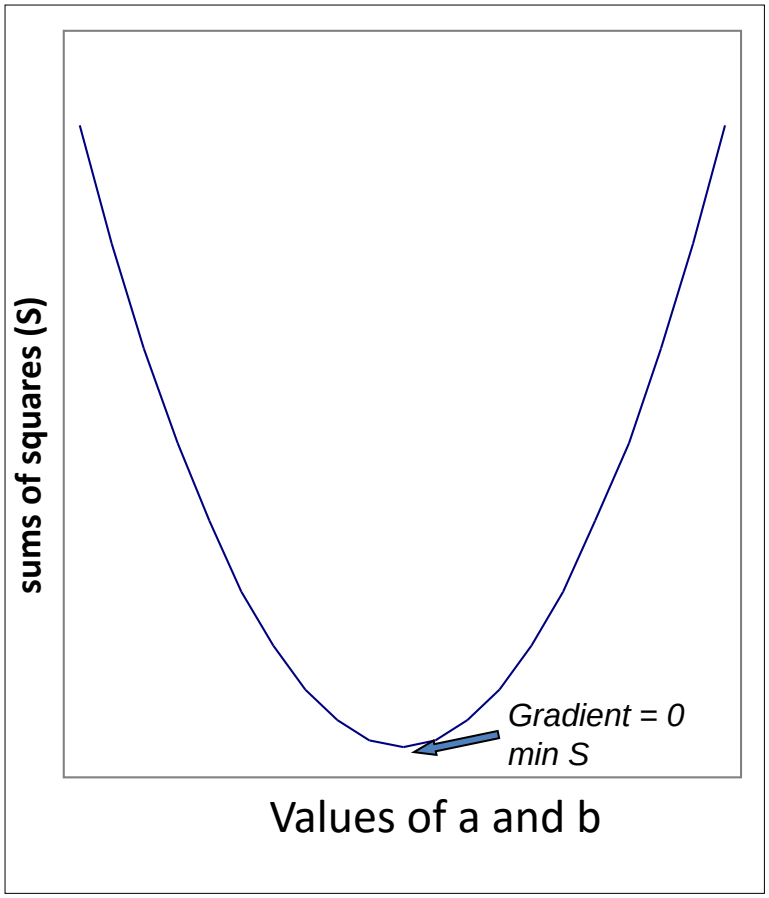
Minimising sums of squares

- Need to minimise $\Sigma(y-\hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimise:

$$\Sigma(y - ax - b)^2$$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term

- The smallest value for the sum of squares of deviations is at the bottom of the curve, where the gradient is zero.



The mathematics of it

- The minimum of sum of squares is at the bottom of the curve where the gradient = 0

- Thus, we can find *a* and *b* that give smallest sum of squares by taking partial derivatives of $\Sigma(y - ax - b)^2$ with respect to *a* and *b* separately

- Then we solve these derivatives for 0 to give us the values of *a* and *b* that give the smallest sum of squares

The solution for slope *a*

- Algebraic re-arrangement results in the following equations for *a*:

$$a = \frac{r\, s_y}{s_x}$$

r = correlation coefficient between x and y
s_y = standard deviation of y
s_x = standard deviation of x

Note:

A low correlation coefficient gives a flatter slope (small value of a)

Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)

Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)

The solution for intercept *b*

- The model equation is $\hat{y} = ax + b$
- This line must pass through the mean so:

$$\bar{y} = a\bar{x} + b \quad \Rightarrow \quad b = \bar{y} - a\bar{x}$$

We can substitute our equation for a into this:

$$b = \bar{y} - \frac{r\, s_y}{s_x}\, \bar{x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

The smaller the correlation, the closer the intercept is to the mean of y

Back to the model

$$\hat{y} = ax + b = \frac{r\,s_y}{s_x}\,x + \bar{y} - \frac{r\,s_y}{s_x}\,\bar{x}$$

(a: $\frac{r\,s_y}{s_x}$; b: $\bar{y} - \frac{r\,s_y}{s_x}\,\bar{x}$)

Rearranges to:

$$\hat{y} = \frac{r\,s_y}{s_x}\,(x - \bar{x}) + \bar{y}$$

- If the correlation is zero, we will simply predict the mean of y for every value of x, and our regression line is just a flat straight line crossing the x-axis at y. However, such a model is not very useful.

- Obviously, we can calculate the regression line for any data. However, the important question is how well does this model fit the data, or how good is the model at predicting y from x.

- This will be discussed as a separate topic – Measures of Accuracy

General Linear Model

- Linear regression is actually a form of the General Linear Model where the parameters are a, the slope of the line, and b, the intercept.

$$y = ax + b + \varepsilon$$

- A General Linear Model is just any model that describes the data in terms of a straight line

Multiple regression

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc, on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + b + \varepsilon$$

- The a parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.
- i.e. the amount of variance in y that is accounted for by each x variable after all the other x variables have been accounted for

Problems with Linear Regressions

- **Non-linearity of data**

  Linear regression assumes that there is a straight-line relationship between predictors and response. However, the underlying relationship may be highly non-linear

  In such cases, there are two possible courses of action:

  1. simple approach is to transform linear predictor (x) into something non-linear (for example: $z_1=\sin(x)$, $z_2=\log(x)$, $z_3=x^2$…) and perform linear regression of z_i
  2. use non-linear regression. We will introduce this technique later

Problems with Linear Regressions

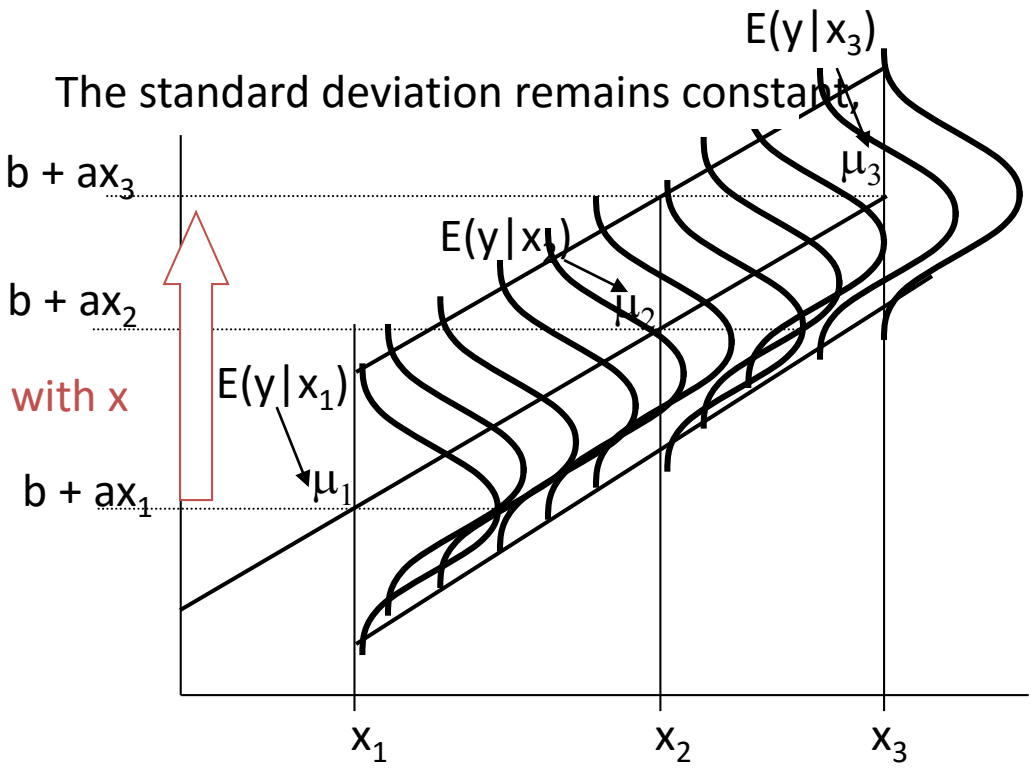
- **Colinearity:** the property that describes the situation when two independent (input) features are approximately in a linear relationship. This notion could be extended to several variables, in which case we call it **multi-colinearity**

Additional Problems with Linear Regression:
Required Conditions for error terms

- The error ε is a critical part of the regression model.
- Four requirements involving the distribution of ε must be satisfied.
  - The probability distribution of ε is normal.
  - The mean of ε is zero: $E(\varepsilon) = 0$.
  - The standard deviation of ε is σ_ε for all values of x.
  - The set of errors associated with different values of y are all independent.

From the first three assumptions we have:
y is normally distributed with mean
E(y) = b + ax, and a constant standard deviation
σ_ε

However the mean value changes with x



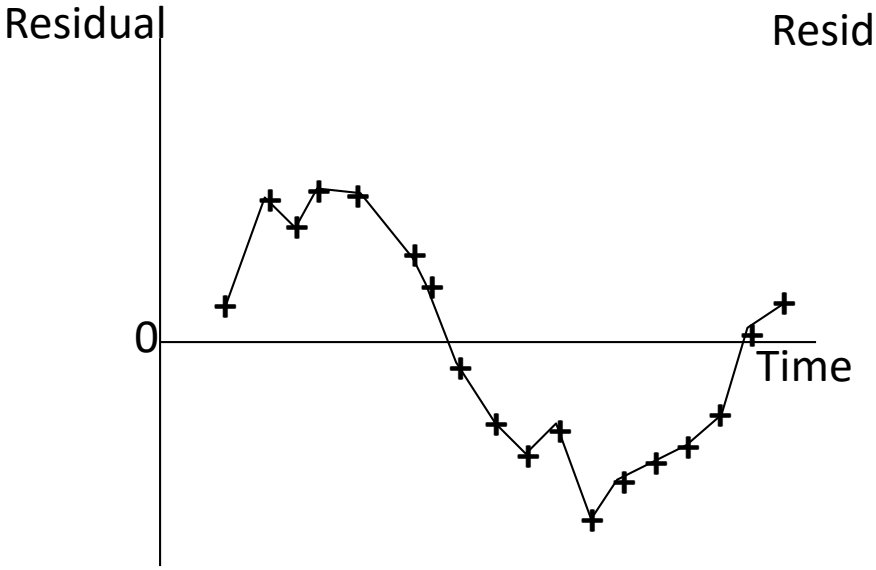
Problems with Linear Regressions

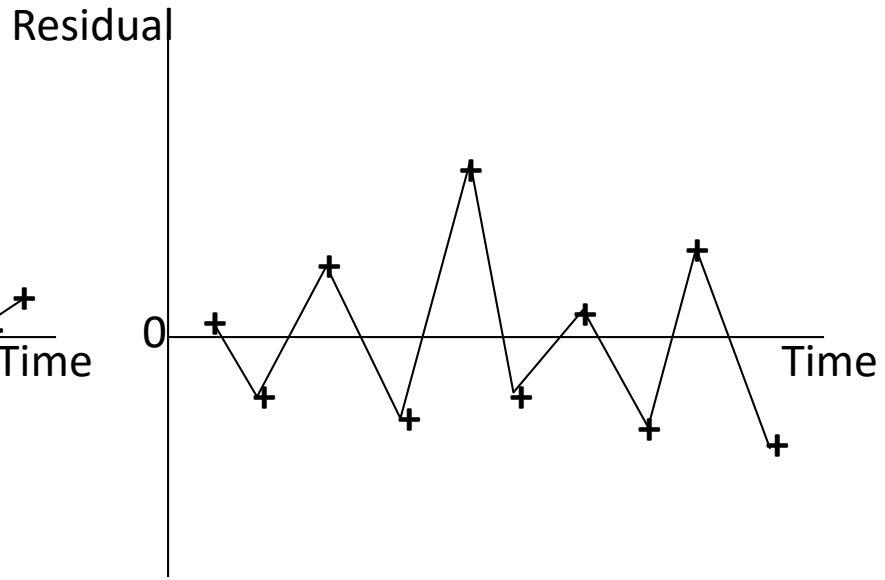
- **Correlation of error terms**

  An important assumption of the linear regression model is that the error terms ε_1, ε_2 , ε_3, … ε_n are un-correlated and completely independent of one another

  This means that knowledge about ε_i tells us nothing about ε_{i+1}
- However, particularly in time-series problems this may not be the case. There is separate class of time linear time series models to deal with the issue

Patterns in the appearance of the residuals over time indicates that autocorrelation exists.





Note the runs of positive residuals, replaced by runs of negative residuals

Note the oscillating behavior of the residuals around zero.

Problems with Linear Regressions

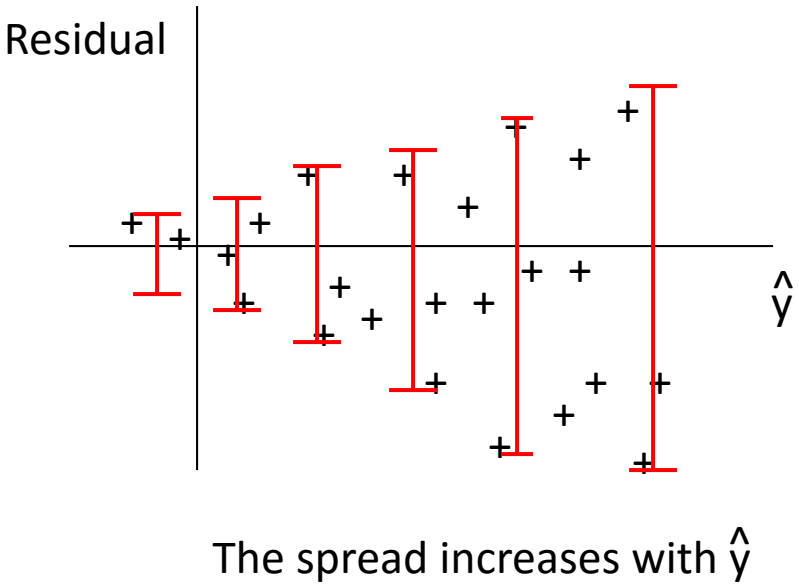
- **Non-constant variance of error terms**

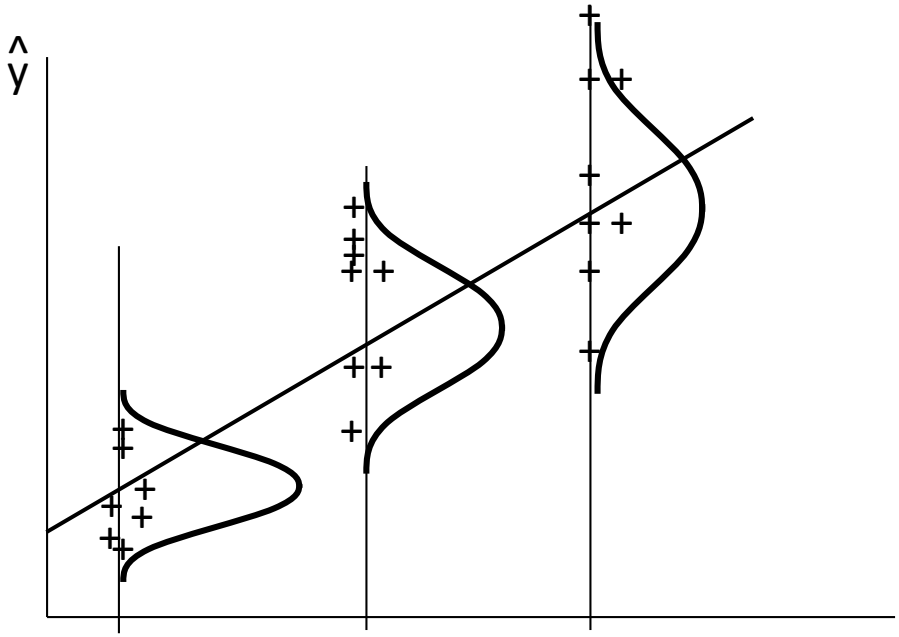
  The linear regression model assumes that that the error terms have a constant variance $Var(\varepsilon_1)=\sigma^2$

  Unfortunately, the variances of the error-terms are not constant. For example, the variances of the error terms may increase with the value of the response. This is referred to as *heteroscedasticity*

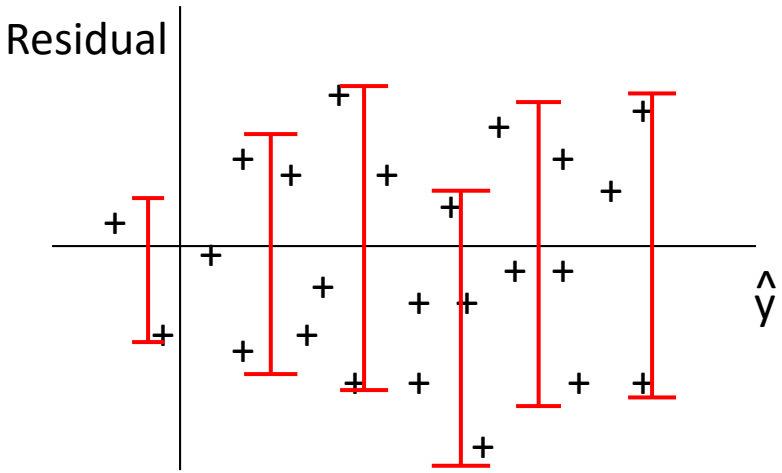
  A solution to this challenge may be to transform the response variable (Y) using concave functions such as log(Y) or SQRT(Y)

- Heteroscedasticity
  - When the requirement of a constant variance is violated we have heteroscedasticity.

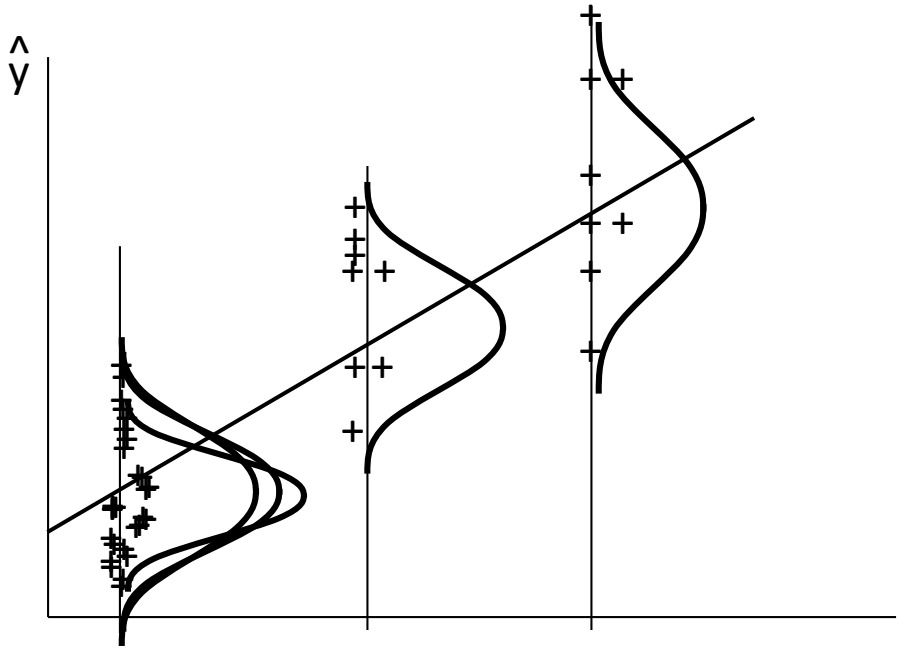




- In case when the requirement of a constant variance is not violated we have homoscedasticity.



The spread of the data points does not change much.



Problems with Linear Regressions

- **Outliers**

  This is quite obvious: fitting to outlier data results in erroneous estimates of regression terms