The F-test for deciding if

$\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ or not

in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i \quad (\dagger)$$

assumes that the ε_i are independent $N(0, \sigma^2)$.

To test these assumptions, in the skin cream example, we can plot

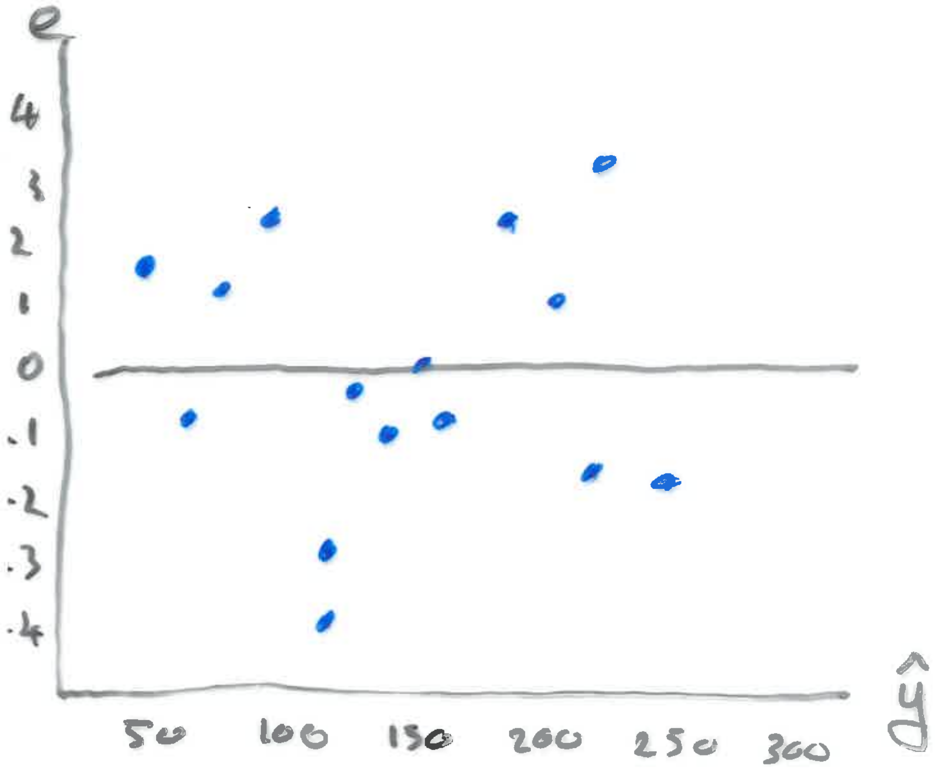
i) $\hat{y}_i$ against e_i

ii) x_{i1} against e_i

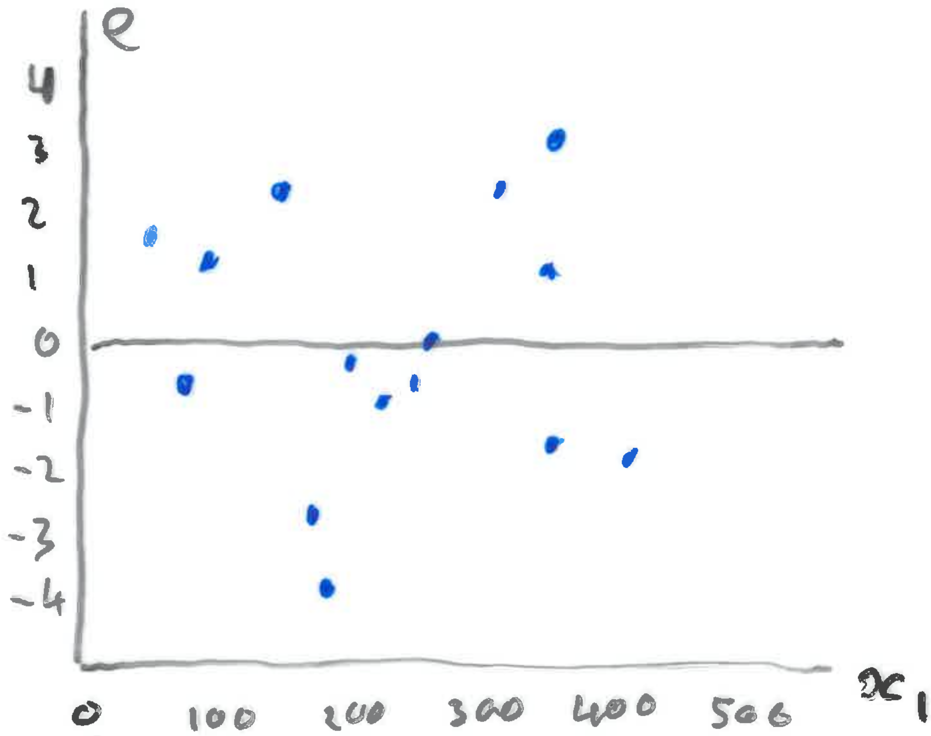
iii) x_{i2} against e_i.

See next slide

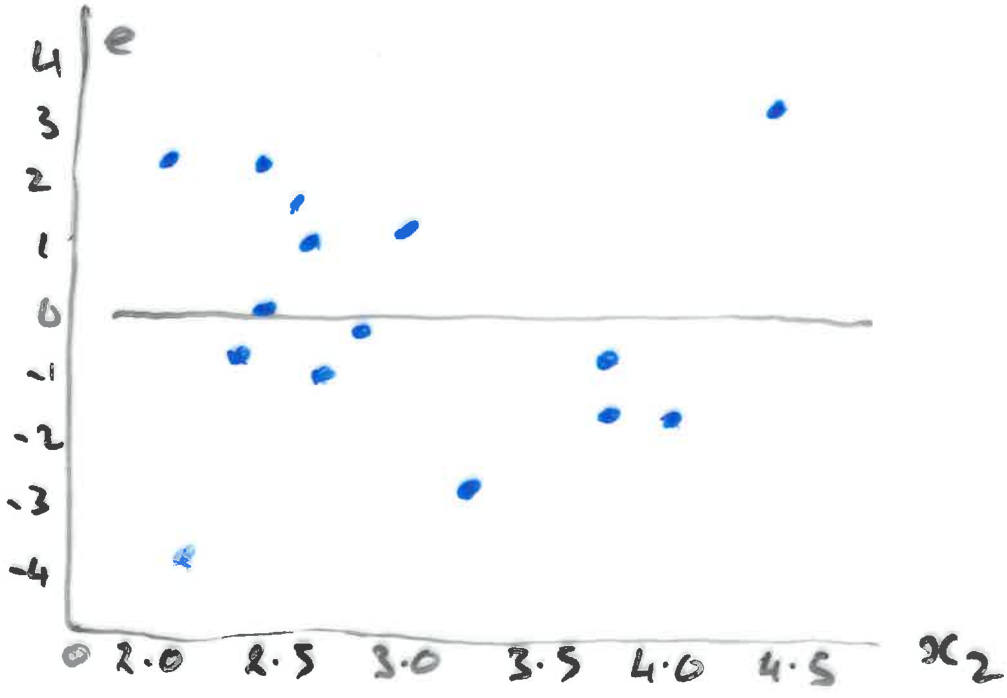
There appears to be no systematic deviation, the residuals seem to be independent and not depend on the level of $\hat{y}$ or the values of x_{i1}, x_{i2}. So it seems ok to accept that ε_i are indep and $N(0, \sigma^2)$.



y vs e



x_1 vs e



x_2 vs e

Defn The estimated covariance matrix for (T) is

$$s^2(B) = MSE\,(X^tX)^{-1}$$

$$= \begin{pmatrix} s^2(b_0) & s(b_0,b_1) & \cdots & s(b_0,b_{p-1}) \\ s(b_1,b_0) & s^2(b_1) & \cdots & \\ & & \ddots & \\ & & & s^2(b_{p-1}) \end{pmatrix}.$$

we only need $s^2(b_0), s^2(b_1), \cdots$.

Theorem Assume ε_i are independent $N(0,\sigma^2)$ the quantity

$$\frac{b_k - \beta_k}{s(b_k)}$$

follows a t-distribution with $n-p$ degrees of freedom.

So, if $q \leq p$ parameters β_k are to be estimated jointly, the confidence intervals with family coefficient $1-\alpha$ are:

$$b_k - T\, s(b_k) \leq \beta_k \leq b_k + T\, s(b_k)$$

where

$$T = t\left(1 - \frac{\alpha}{2q},\ n-p\right).$$

<u>Example</u> Continuing with skin cream sales, it is desired to estimate β_1 and β_2 jointly with a family confidence coefficient of 0.90.

$$s^2(B) = MSE(X^tX)^{-1} = \begin{pmatrix} 5.9081 & * & * \\ * & .00003656 & * \\ * & * & .00000937 \end{pmatrix}$$

$s^2(b_1) = .00003656$, $\quad s(b_1) = .006054$

$s^2(b_2) = .00000937$, $\quad s(b_2) = .0009681$

$$T = t\left(1 - \frac{0.10}{2 \times 2}, 12\right) = t(0.975, 12) = 2.179$$

So

$$0.4961 - (2.179)(.006054) \le \beta_1 \le 0.4961 + (2.179)(.006054)$$

or

$$0.483 \le \beta_1 \le 0.509$$

and similarly

$$0.0071 \le \beta_2 \le 0.0113$$

Principal Component Analysis

Consider a collection of data points $w_1, w_2, \dots, w_n \in \mathbb{R}^p$ where p may be large.

Example $w_1, \dots, w_n \in \mathbb{R}^{256^2}$ are vectors representing n grey-scale images of faces. An digital image is a 256×256 array of pixels

Each pixel's greyness is determined by an integer in $\mathbb{R}$, and the image is thus represented by a 256×256 real matrix. Concatenating rows yields a vector $w \in \mathbb{R}^{65536}$.

Define the mean of $w_1, \dots, w_n \in \mathbb{R}^p$ as

$$\bar{w} = \frac{1}{n}(w_1 + \dots + w_n).$$

Set

$$\boxed{v_i = w_i - \bar{w}}.$$

Then $v_1, v_2, \dots, v_n \in \mathbb{R}^p$ are data points with mean

$$\bar{v} = \frac{1}{n}(v_1 + \dots + v_n) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^p.$$

We'll use notation

$$v_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix}, v_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2p} \end{pmatrix}, \dots, v_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}$$

Define the covariance matrix

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & & \\ c_{p1} & \cdots & c_{pp} \end{pmatrix}$$

by

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$= \frac{1}{n} \sum_{k=1}^{n} x_{ki} \, x_{kj}$$

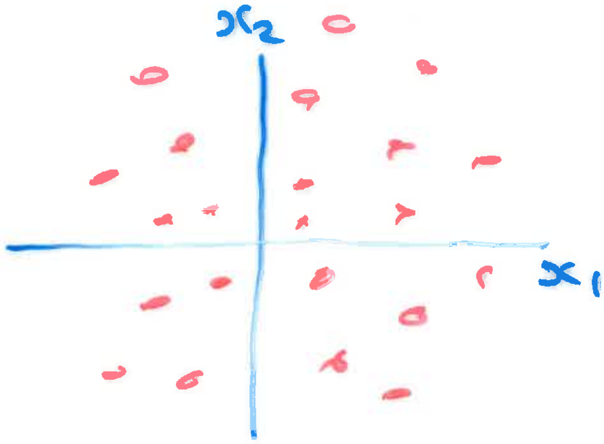
Defn x_{*i} and x_{*j} are uncorrelated

if $c_{ij} = 0 = c_{ji}$.

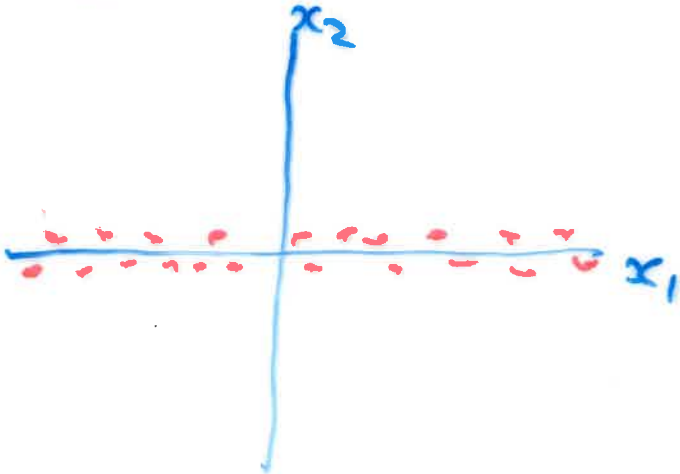
Example ($p=2$)

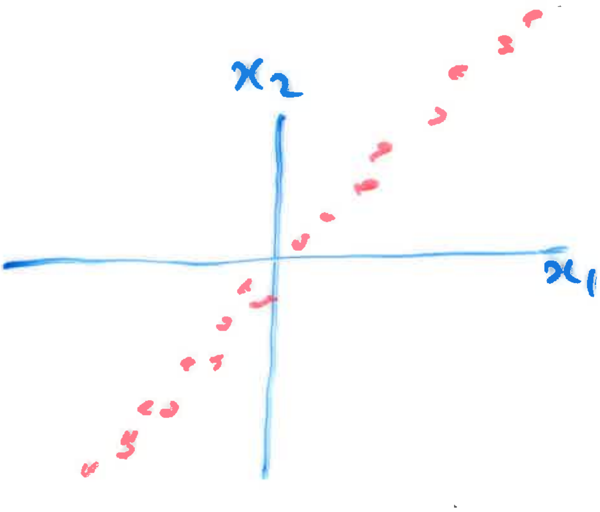
Consider four data sets $\{v_1, \ldots, v_n\} \subseteq \mathbb{R}^2$.

Data 1

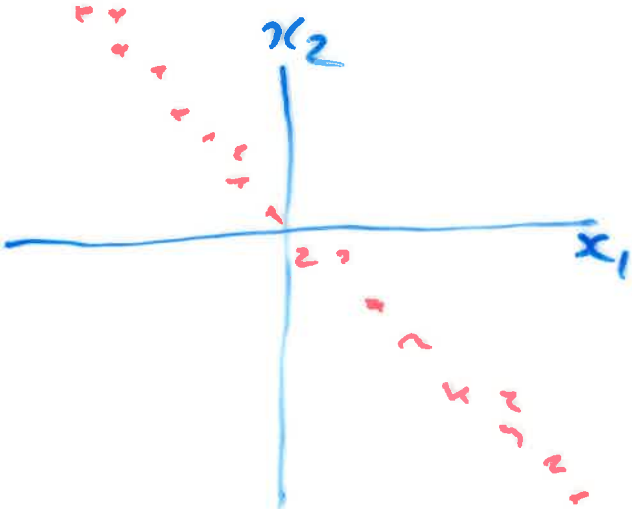


Data 2





Data 3



Data 4

For each of the 4 cases let's consider the covariance matrix

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

Data Set	c_{11}	c_{22}	$c_{12} = c_{21}$
1	large	large	0
2	large	small	0
3	large	large	large positive
4	large	large	large negative