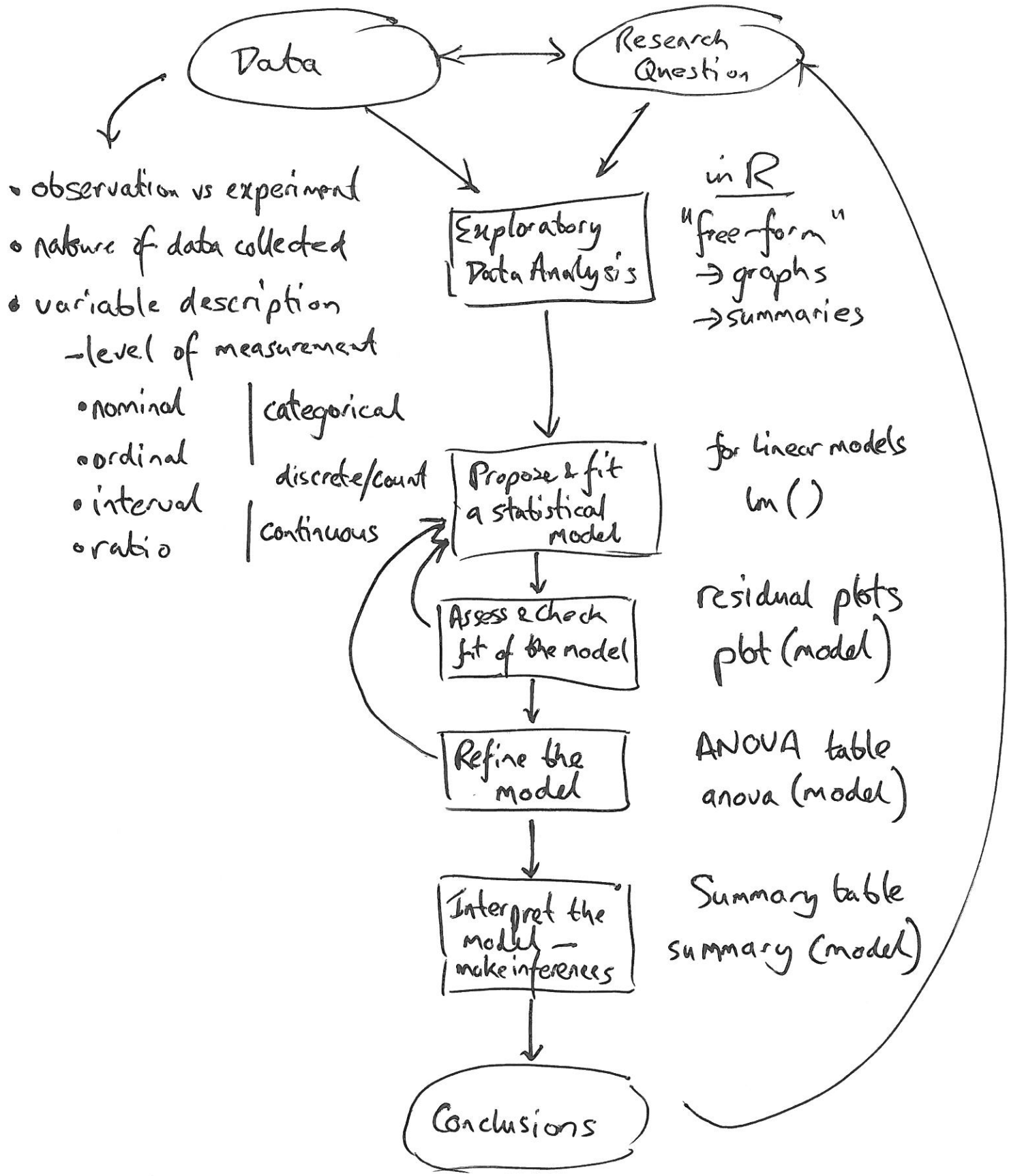
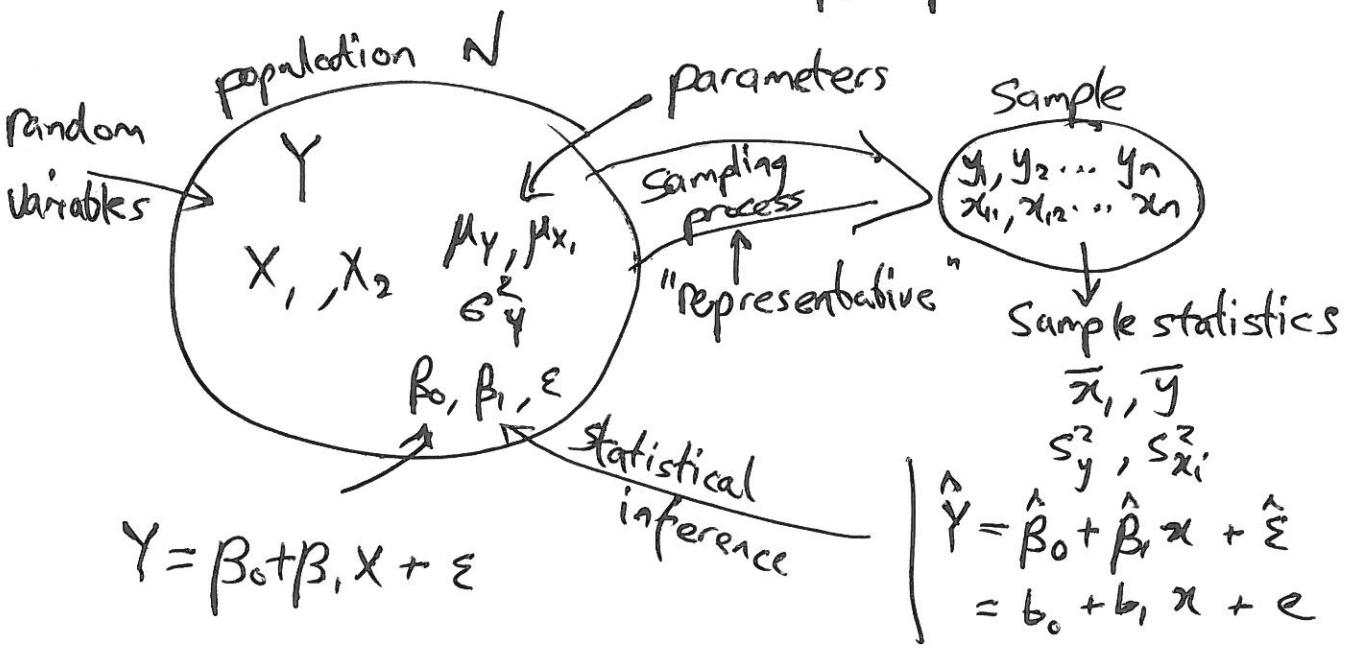
The modelling process (as part of a research process)



General Assumptions (underlying all statistical models)

- sample of data is representative of the population of interest (as defined by the research question)
- that all the error is in the direction of the response or unknown variable Y, ie that the X's are "known" (or measured with negligible error in comparison to Y)
- chosen model is appropriate



Notation capital letters Y, X_1, X_2 denote random variables

small letters $y, x_i \ldots$ denote sample observations

Greek letters μ, π, σ denote population parameters

Roman equivalents m, p, s denote sample statistics which are typically point estimates of the equiv. population parameters — though we can use $\hat{}$ for this purpose

$$\hat{\mu}_x = \underset{\text{not used}}{m_x} = \bar{x}, \quad \hat{\pi} = p, \quad \hat{\beta}_0 = b_0, \quad \hat{\beta}_1 = b_1$$

Assumptions underlying Regression Models:

Multiple Linear Regression Model or General Linear Model

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k}_{\text{deterministic part}} + \underbrace{\varepsilon}_{\text{stochastic part}}$$

$E[Y | X_1, X_2 \ldots X_k]$ — mean model

[probability model] — variance model

Model-specific assumptions are about the variance model

$$\varepsilon's \overset{iid}{\sim} N(0, \sigma^2)$$

The errors are independent & identically (normally) distributed with constant variance

→ these assumptions are all about the errors, but we assess them using the residuals, which are the observed or estimated errors

Key assumptions are:

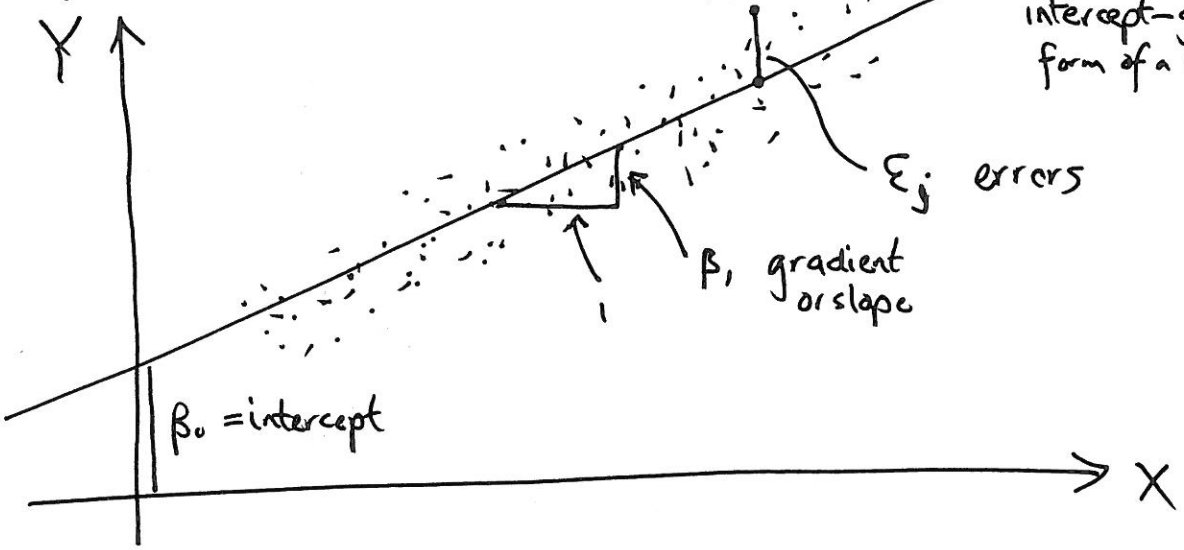
1. independence (no systematic patterns in the residuals)
2. constant variance (no systematic changes in the variance of the residuals)
3. normality (residual are normally distributed → needs a qq plot)

eg Simple Linear Regression

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $i = 1, \dots N$

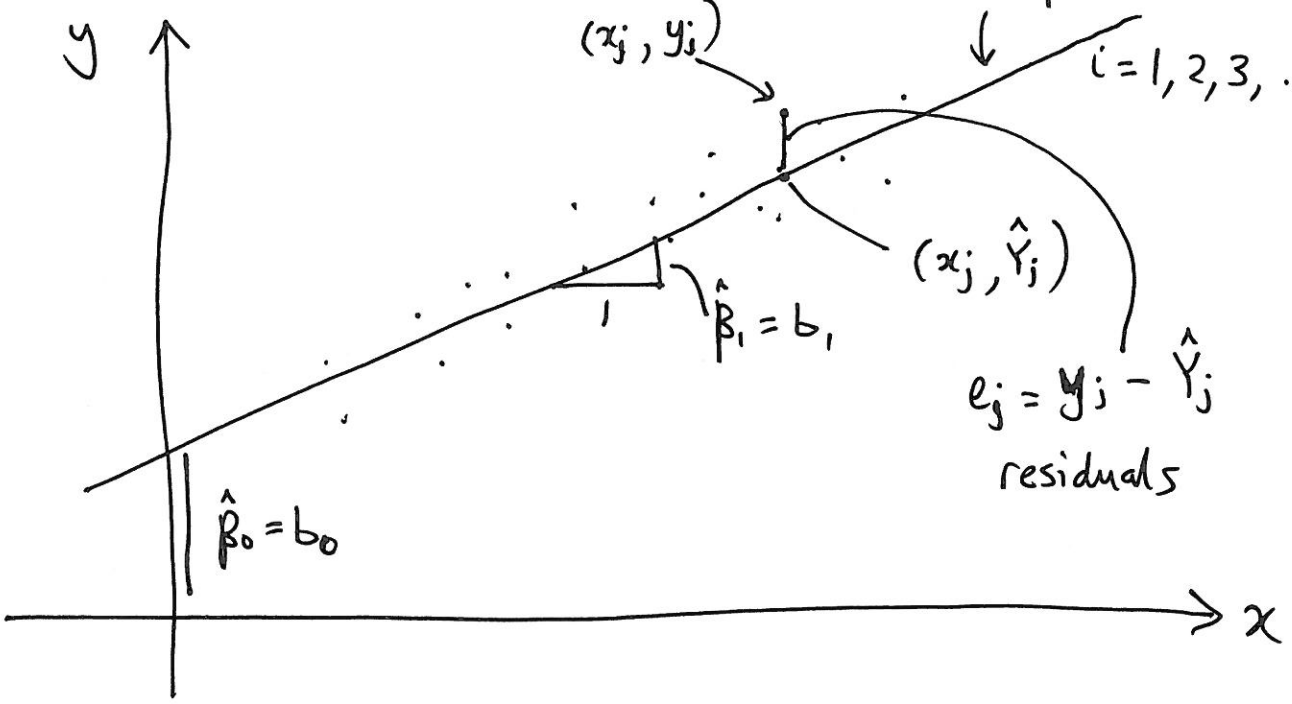
intercept–gradient form of a line

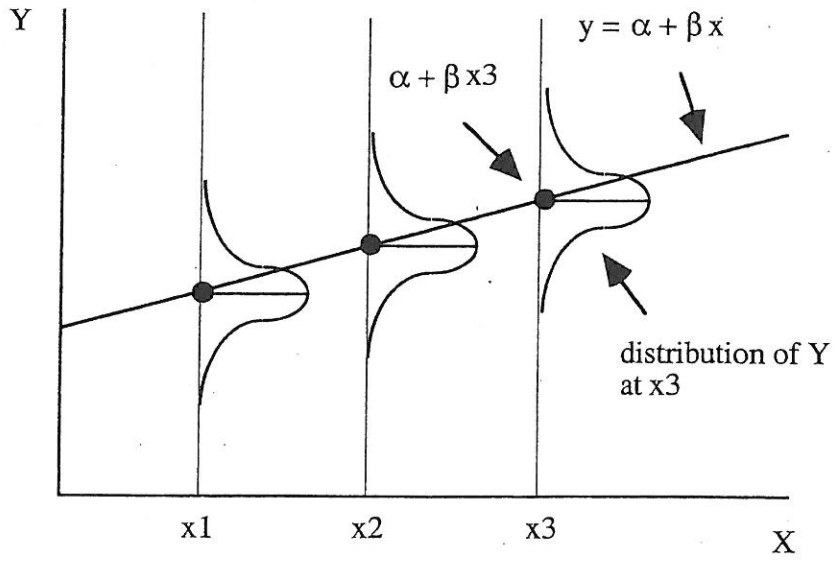
Population



Sample

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $i = 1, 2, 3, \dots n$

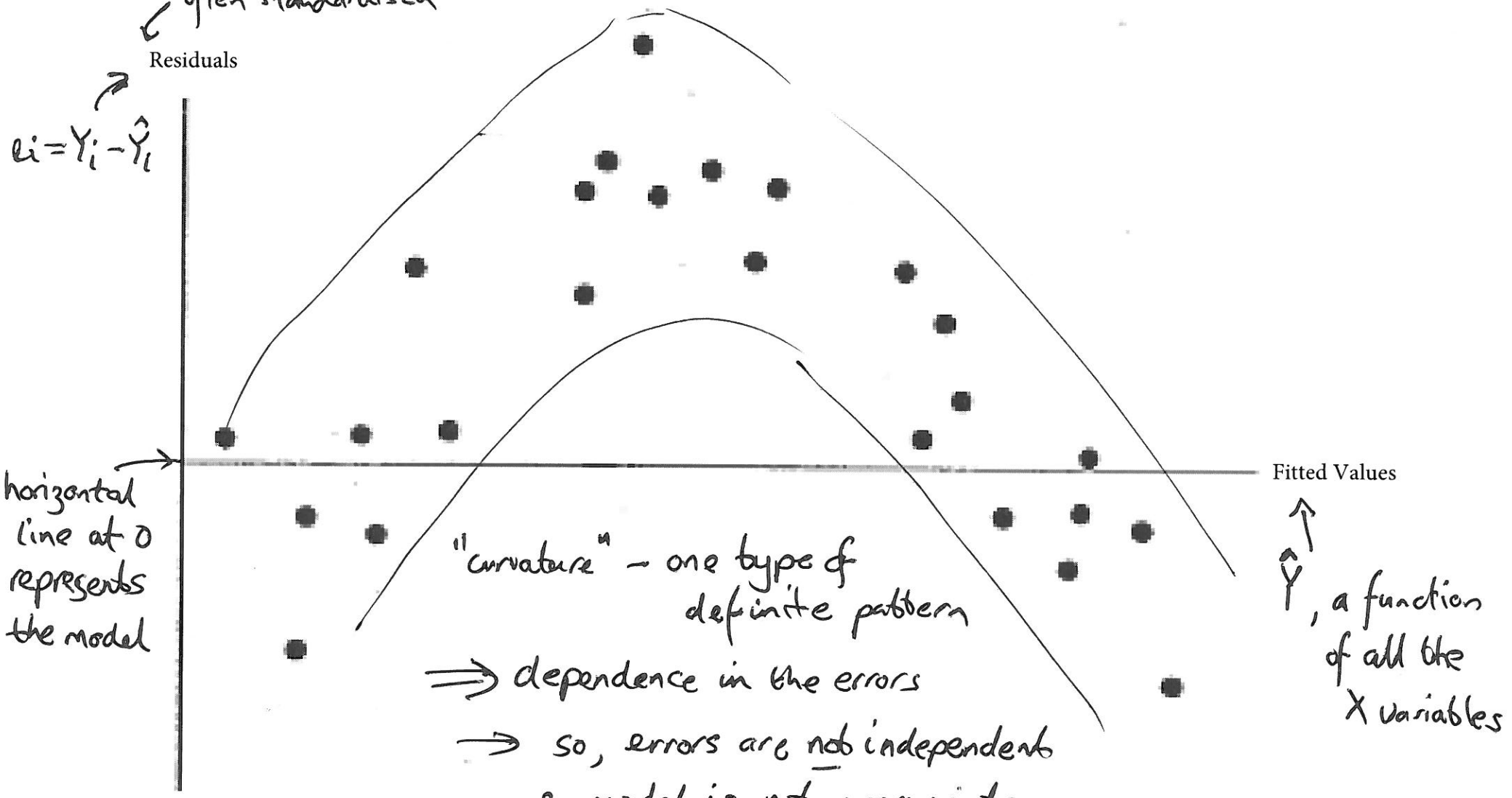


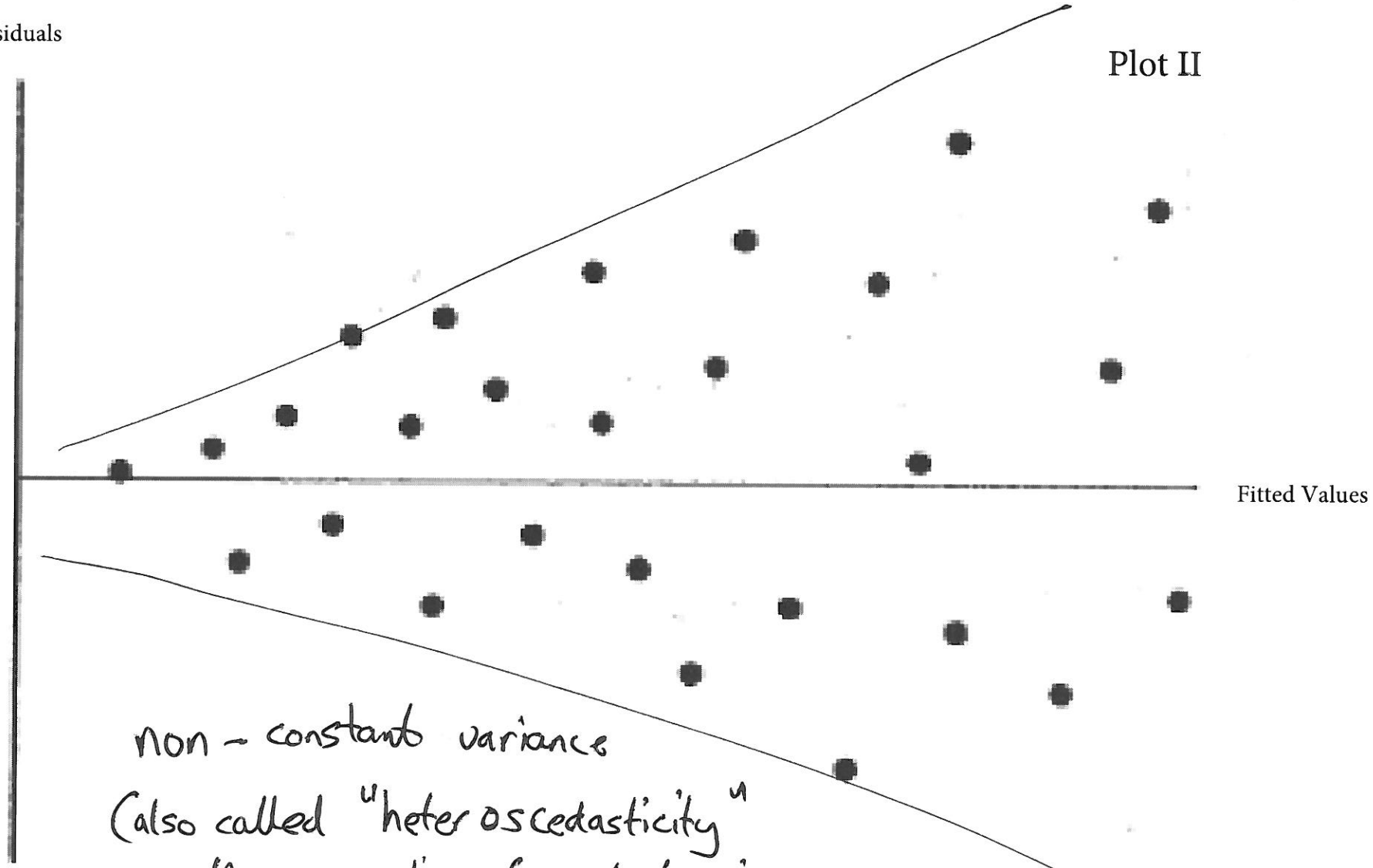


Assumption of constant variance of the errors (ie same distribution with same σ^2 for all values of X)

Main residual plot

Plot I





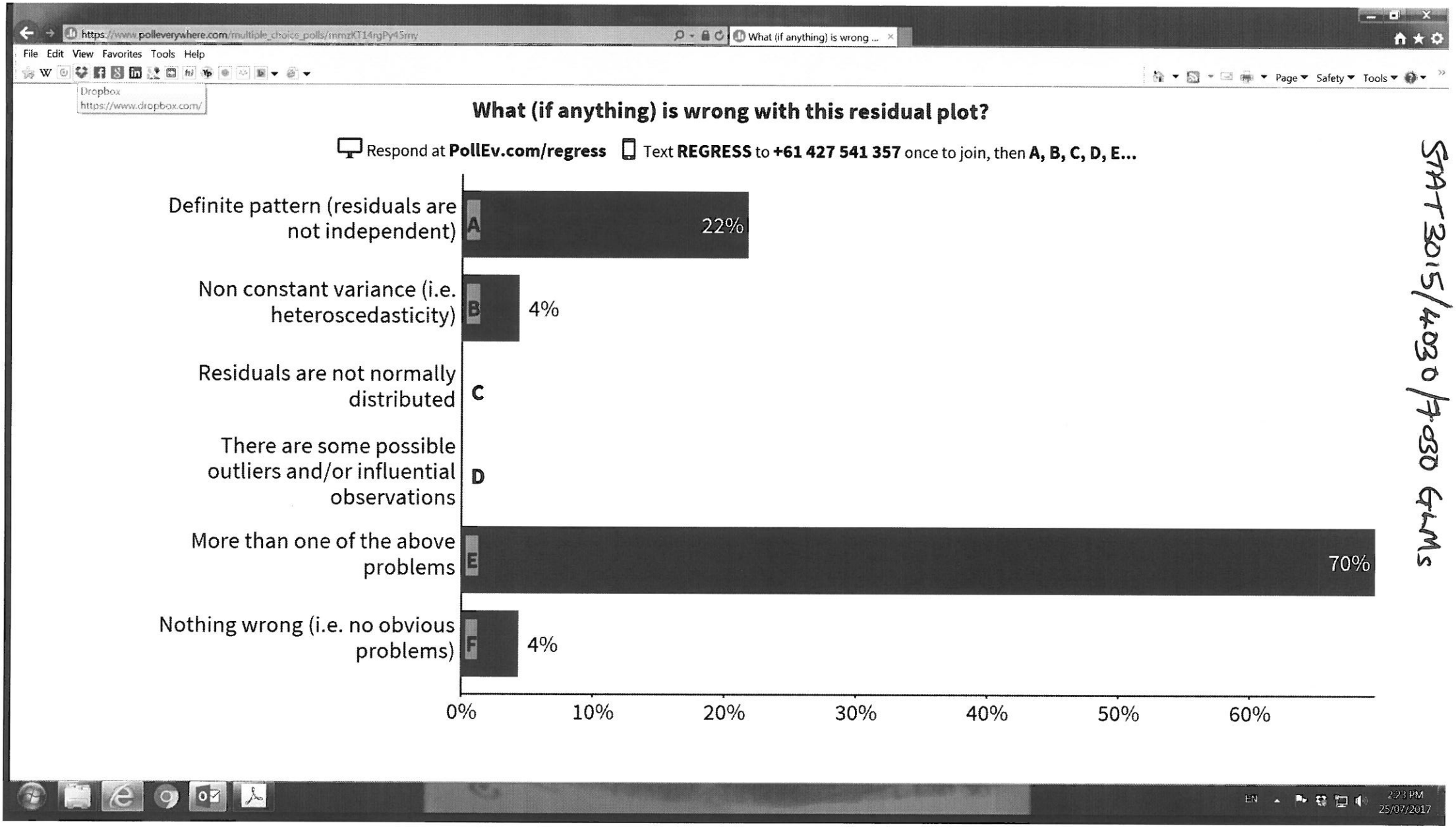
non-constant variance

(also called "heteroscedasticity"
as the assumption of constant variance
is assuming that the errors are "homoscedastic")

Plot III

Plot IV
Residuals
Fitted Values
not independent
potential outlier
variance not constant

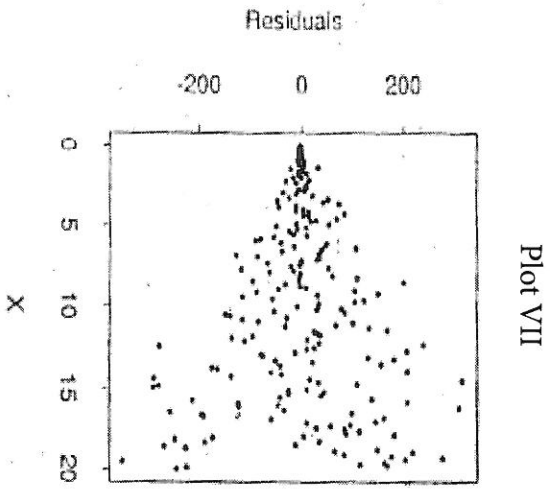
Results of Poll Everywhere poll for Plot IV



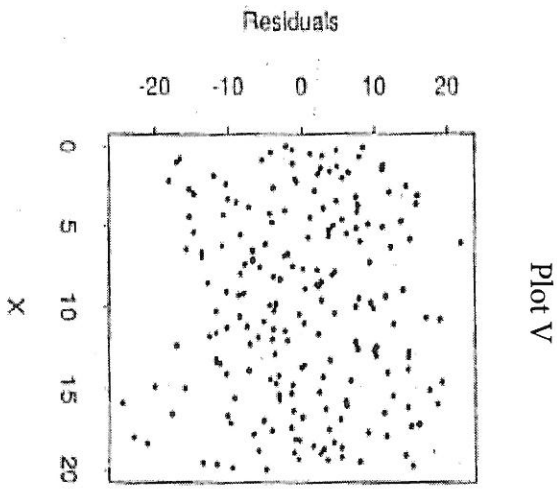
STAT3015/4030/7030 Gems

25/7
(10)

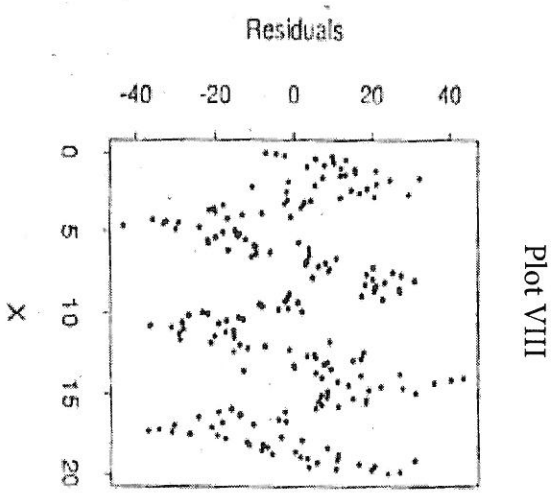
More examples for you to try:



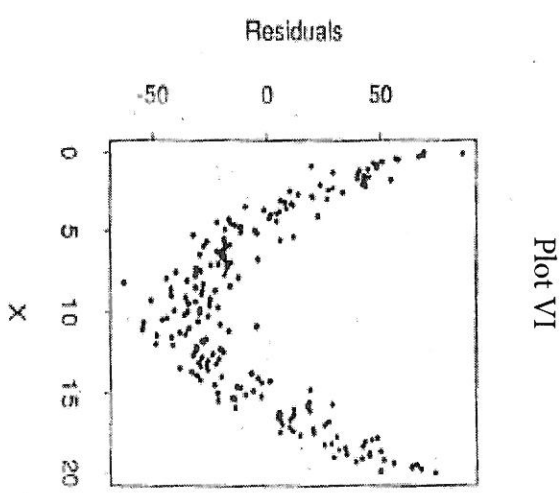
Plot VII



Plot V

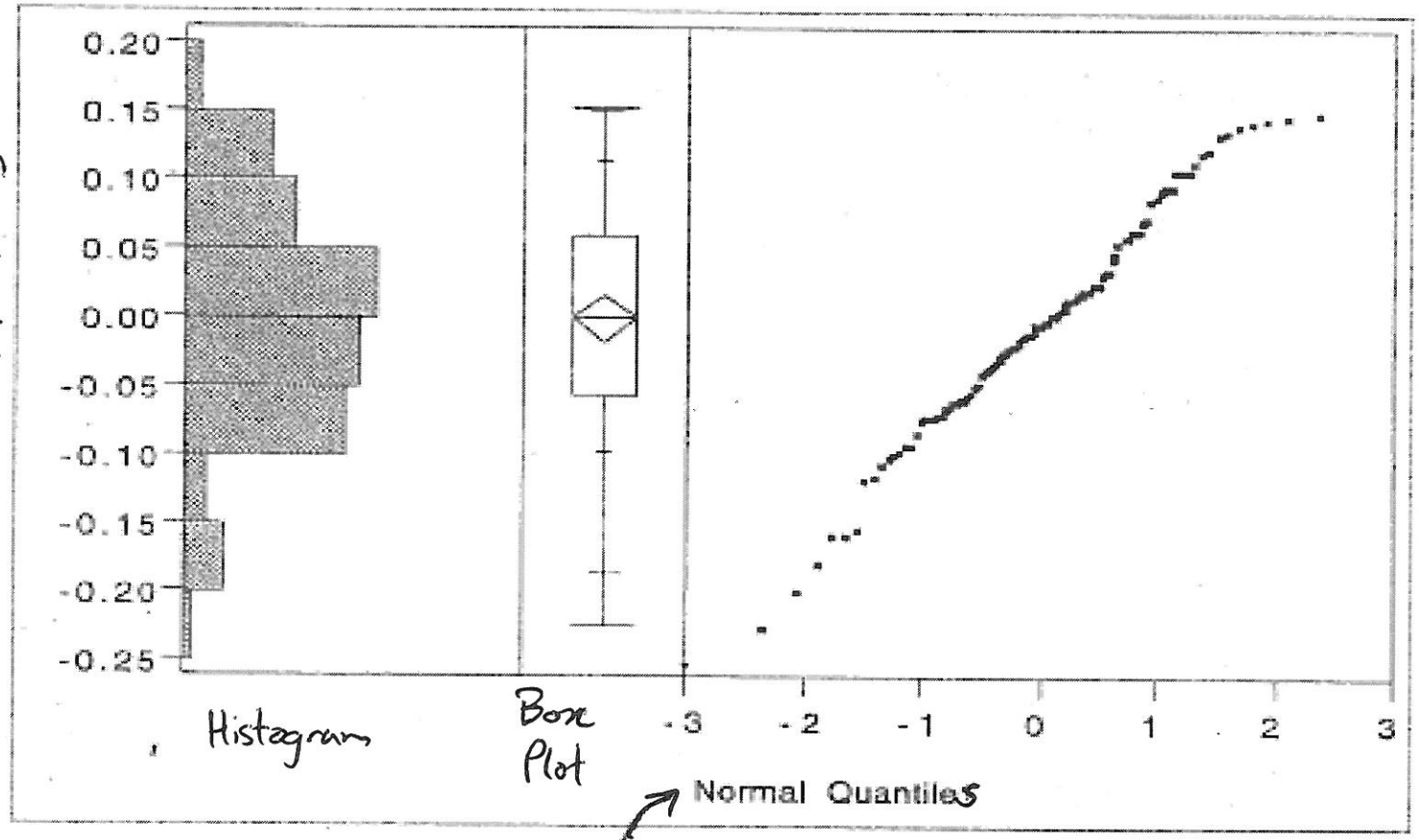


Plot VIII



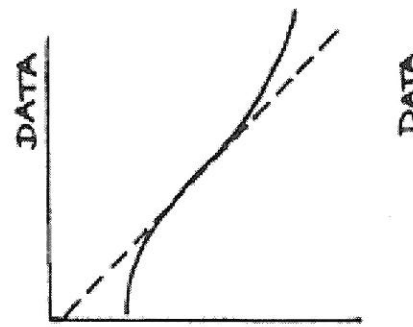
Plot VI

Assessing normality (only assumption not tested by main residual plot)

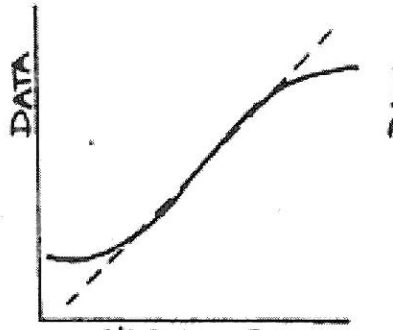


Normal quantile (qq) plot using equivalent quantiles from a standard normal distribution

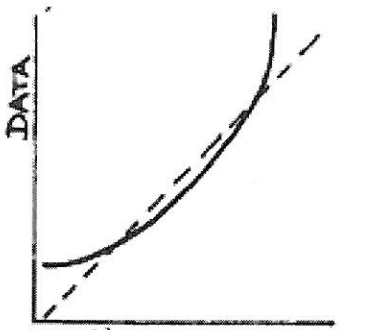
Features you might see on a normal quantile plot



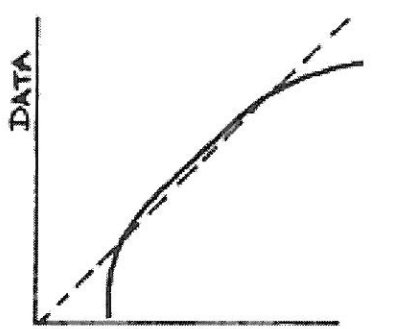
means...
HEAVY-TAILED
(compared to normal)



means...
LIGHT-TAILED
(compared to normal)



means...
SKEWED TO
THE RIGHT



means...
SKEWED TO
THE LEFT

Departures from normality