Residual plots

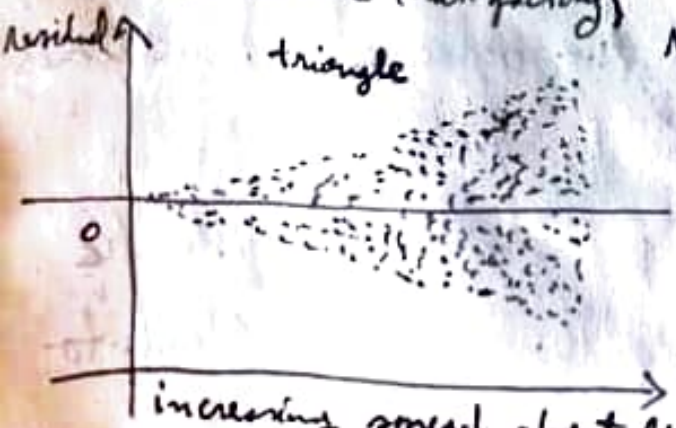
The residual $\hat{\varepsilon}_i$ is a measure of how closely a model agrees with the observation y_i. One can simply use the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$, or the <u>standardised residuals</u>, which have approximately a variance of 1. One should check for:
- isolated lack-of-fit (a few unusual observations);
- systematic lack-of-fit (the general behaviour of the data is different from the model).

The following plots may be useful to check the model assumptions:

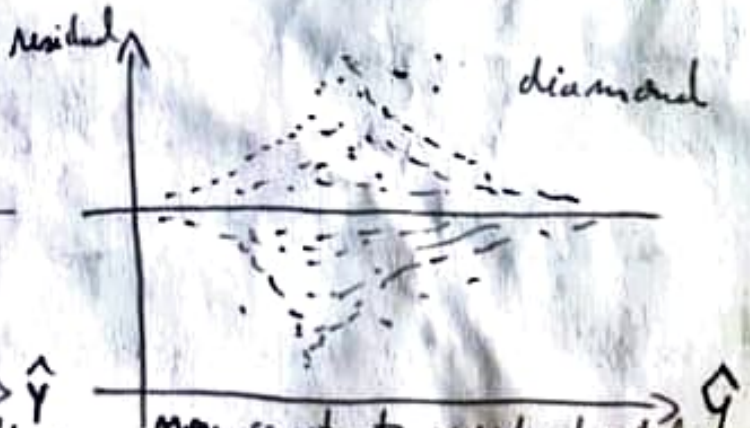
<u>Zero Mean and Constant Variance</u>

Plot the standardised residuals against the fitted values $\hat{y}_i$. The points should be scattered evenly around zero, with no systematic pattern.

random scatter (satisfactory)



increasing spread about line



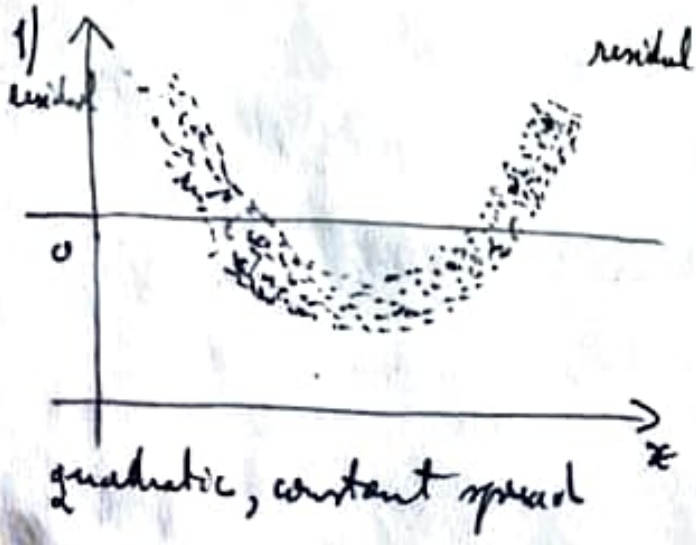
non-constant spread about line

In case of a triangle or diamond plot one can try to transform Y in order to obtain approximate constancy of error variance

<u>linearity</u>

The standardized residuals can be plotted against the individual explanatory variables (predictors). All such plots should indicate random scatter of equal width about zero. Some situations to consider

1)



quadratic, constant spread

non-linear, increased spread

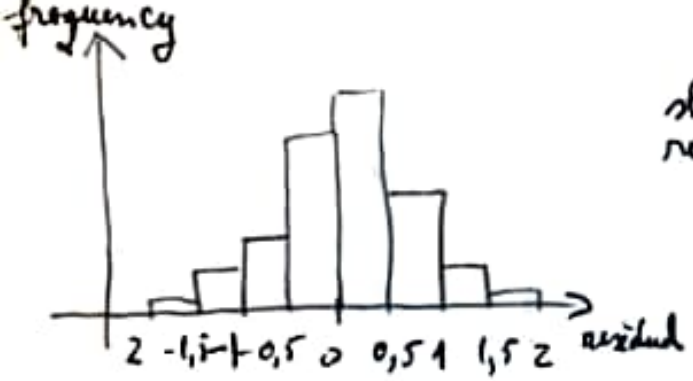
For predictors that are in the model non-linearity suggests that higher order terms involving these predictors should be added to the model. Transform x or/and y in order to achieve approximate linearity.

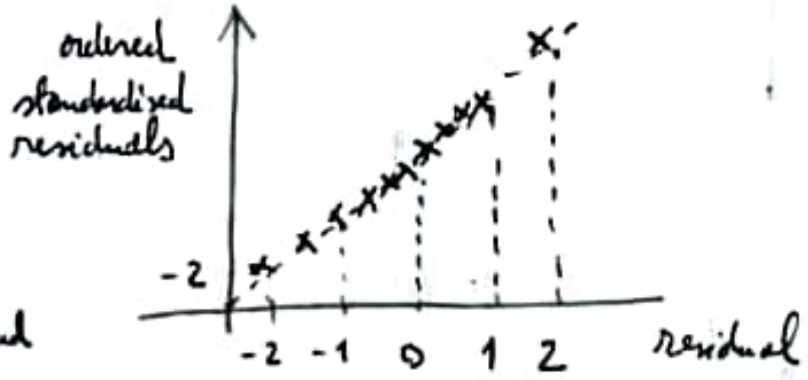
2) Two separate straight lines: Fit two separate regression lines, for example one for males and one for females.

3) For predictors that are not included in the model any systematic pattern with the residuals suggests that those should be added to the model.

Normality

One can look at a histogram of the standardized residuals.

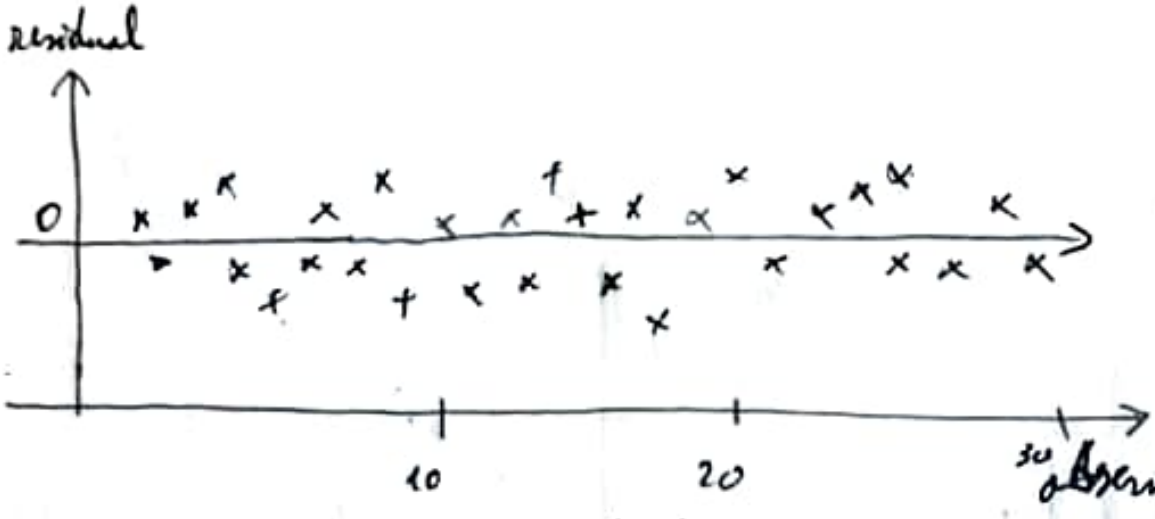




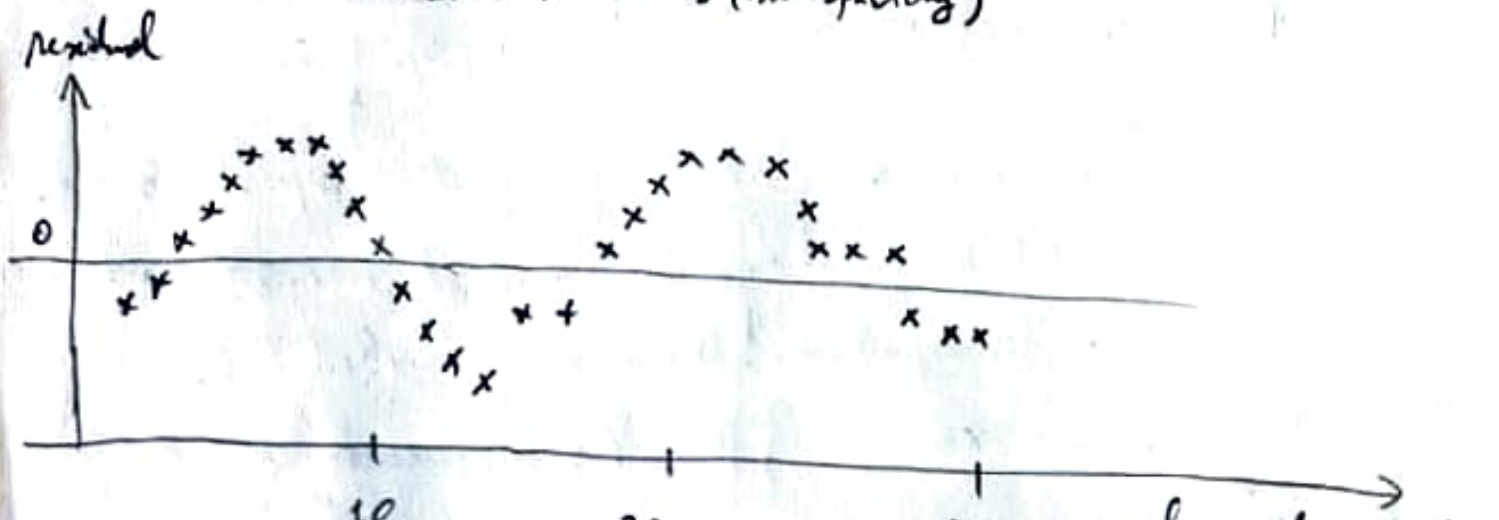
Alternatively, one can look at a normal QQ plot. Departures from normality are indicated by deviations from a straight line. There may be good reason for departures! Use a model with a different distribution for Y, for example if the Y's are counts one can try to use a Poisson distribution.

Independence

One can plot the standardized residuals against the serial order in which the observations were taken. If the model assumptions are correct a random scattering of points with no visible trend is expected. Correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time.
In many cases, observations that are obtained at adjacent points in time will have positively correlated errors. We may see tracking in the residuals – that is adjacent residuals may have similar values.



Uncorrelated residuals (satisfactory)



tracking in residuals

In the context of time series data many methods have been developed to properly take account of correlations of the residuals. See for example ARMA, ARIMA, ARCH, GARCH. Their presentation is beyond the scope of this course.

Outliers

Possible outliers may be indicated by points with large standardized residuals. Under normality assumption, approximately 95% of observations should have standardized residuals in the range (-2.0, 2.0).
A "rule of thumb" is to consider an observation with an absolute value of standardized residual >2.5 as an outlier and the accuracy of such an observed value should be investigated.

The Logistic Model

(The Logistic Regression Model)

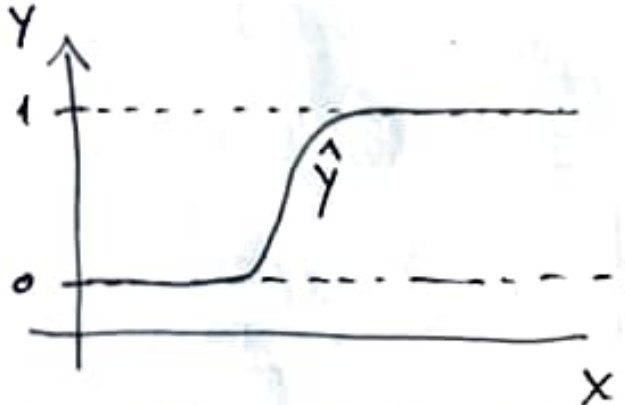
The linear model assumes that the response variable Y is quantitative. However, in many situations, the response variable is qualitative (categorical). For example, eye color: blue, brown, green.

Predicting qualitative responses is a process that is known as classification. Assume a binary (two level) qualitative response.

One approach is to code the response using a 0/1 dummy variable and predict 1 if $\hat{Y} > 0{,}5$ and 0 otherwise.

Remark The classification that we get if we use a linear model to predict a binary response coded as above is the same as for the linear Discriminant Analysis (LDA). The estimates that we get might be outside the $[0,1]$ interval, making them hard to interpret as probabilities.

Estimated Y using linear model



Predicted probabilities for $Y=1$ using the logistic model

The logistic model uses the logistic function

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

in order to model the relationship between the probability of $Y=1$ and X

$$\Leftrightarrow P(Y=1|X)\left(1+e^{\beta_0+\beta_1 X}\right) = e^{\beta_0+\beta_1 X}$$

$$\Leftrightarrow P(Y=1|X) = e^{\beta_0+\beta_1 X}\left(1-P(Y=1|X)\right)$$

$$\Leftrightarrow \frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0+\beta_1 X} \quad \Big| \log, \text{i.e. } \ln$$

$$\Leftrightarrow \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0+\beta_1 X$$

$\frac{P(Y=1|X)}{1-P(Y=1|X)}$ is called the <u>odds</u> and can take any value between 0 and ∞.

(Odds are traditionally used in horse-racing instead of probabilities, since they relate more naturally to the correct betting strategy)

$\ln \frac{P(Y=1|X)}{1-P(Y=1|X)}$ is called the <u>log-odds</u> or <u>logit</u>.

The logistic model has a logit that is linear in X!

<u>Multiple logistic model</u>

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}$$

$$\Leftrightarrow \ln \frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p$$

$X=(X_1, X_2 \ldots X_p)$