Revision of previous lecture.

Linear regression model

Our model for the response random variable Y when the explanatory variable is x comprises two components:

(1) - a mean response $\mathbb{E}(Y) = \beta_0 + \beta_1 x$; plus
- variability in the response

Linearity between ~~mea~~ response mean + explanatory variable (s)

where

(2) - this variability has a Normal distribution /

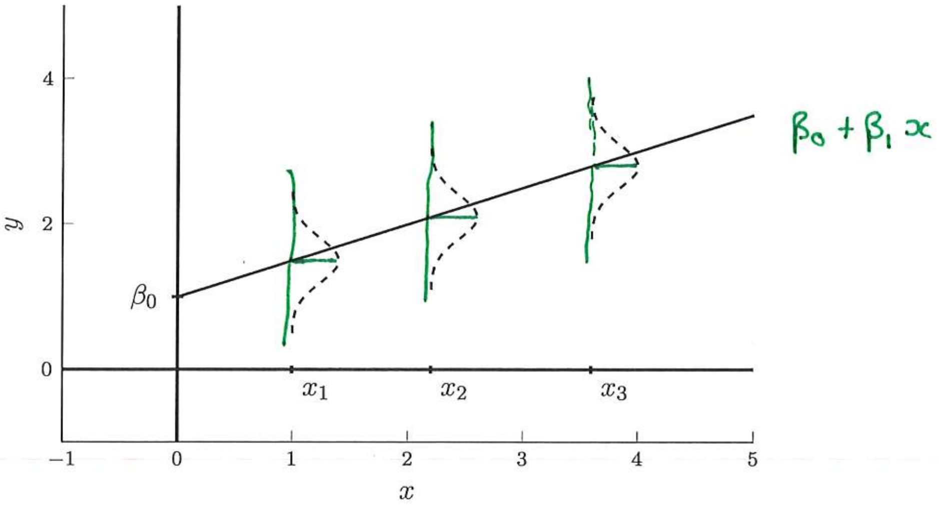
(3) - the amount of variability does not depend on x. — variance of $\mathbb{E}Y$ is the same regardless of the value of x.

+ independence of $Y_1, Y_2, \ldots, Y_n$

In other words, the response is

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where $\varepsilon \sim \mathsf{Normal}(0, \sigma^2)$, and σ^2 is constant for all x. The (unobservable) errors ε capture deviations from the general trend due to other factors that we did not take into account.



Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, where the $Y_1, Y_2, \ldots, Y_n$ are independent random variables and the distribution of Y_i depends on the explanatory variable $x = x_i$. Let $\mathbf{X}$ be the matrix such that $\mathbf{X}_{i,1} = 1$ and $\mathbf{X}_{i,2} = x_i$ and let $\beta = [\beta_0 \;\; \beta_1]^T$. We can write the distribution of Y_i and $\mathbf{Y}$ from this model as

$$Y_i \sim \mathsf{Normal}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

$$\mathbf{Y} \sim \mathsf{Normal}\left(X\beta, \sigma^2 I\right)$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{bmatrix} \qquad X\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \end{bmatrix}$$

to construct the confidence interval. Using the same kind of reasoning we used in Chapter 6, we arrive at the $100(1-\alpha)\%$ confidence interval

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} \pm t_{n-2;1-\alpha/2} \times s\sqrt{\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}^{T}}.$$

Returning to the facebook example, lets construct at 95% confidence interval for the mean grey matter density of a person who has 250 facebook friends. So $\mathbf{x}_{new} = [1 \quad 250]$ and

$$\mathbf{x}_{new}\widehat{\boldsymbol{\beta}} = [1 \quad 250]\begin{bmatrix} -1.3118 \\ 0.0028855 \end{bmatrix} = -0.59041.$$

estimated mean GM density for 250 facebook friends

This is our estimate of $\mathbf{x}_{new}\boldsymbol{\beta}$. We now need the standard error of this estimate. We can get the matrix $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ from the result of `fitlm` in MATLAB.

```
facebooklm.CoefficientCovariance

 0.2912788549    -5.896487e-04
-5.896487e-04     1.291815e-06
```

So

$$s^2\mathbf{x}_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}^{T} = [1 \quad 250]\begin{bmatrix} 0.2912788549 & -5.896487e-04 \\ -5.896487e-04 & 1.291815e-06 \end{bmatrix}\begin{bmatrix} 1 \\ 250 \end{bmatrix}$$

$$= 0.07719289$$

$$s.e.(\mathbf{x}_{new}\widehat{\boldsymbol{\beta}}) = \sqrt{0.07719289} = 0.2778361$$

The 95% confidence interval for $\mathbf{x}_{new}\boldsymbol{\beta}$ is

$$-0.59041 \pm t_{38,0.975} \times 0.2778361$$

estimate ± (critical value) × s.e (estimate)

which is

$$-0.59041 \pm 0.5624498$$

The function `predict` can be used to construct this confidence interval.

```
facebooklm = fitlm(facebook,'GMdensity~FacebookFriends');
[yhat,ci]=predict(facebooklm,250,'Alpha',0.05)

yhat =

    -0.5904


ci =

    -1.1529    -0.0280
```

fitted model

new explanatory variable value

95% confidence interval.

Diagnostics

Just because you can fit a linear regresssion model to your data doesn't mean should. The inferences we make are dependent on the model assumptions. It is necessary to employ some diagnostics to check that our data is consistent with those assumptions.

Recall. The assumptions of the linear regression model are:

- Linearity: The mean of the response variable is a linear function of the explanatory variable.
- Normality: The variability about the mean has a normal distribution.
- Constant variability: The variability about the mean has a constant variance.

The diagnostics we will use are graphical and require some interpretation – and hence experience to use correctly. The diagnostics will be based on the observed *residuals* from the model fit. The residuals are given by

$$\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i. \quad = (\text{observed}) - (\text{fitted value})$$

Residuals are not the same as the normal errors ε in the regression model. Note that

$$\widehat{\varepsilon} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (I - X(X^TX)^{-1}X^T)Y$$

$\hat{\beta} = (X^TX)^{-1}X^TY$

$Y = X\beta + \varepsilon$

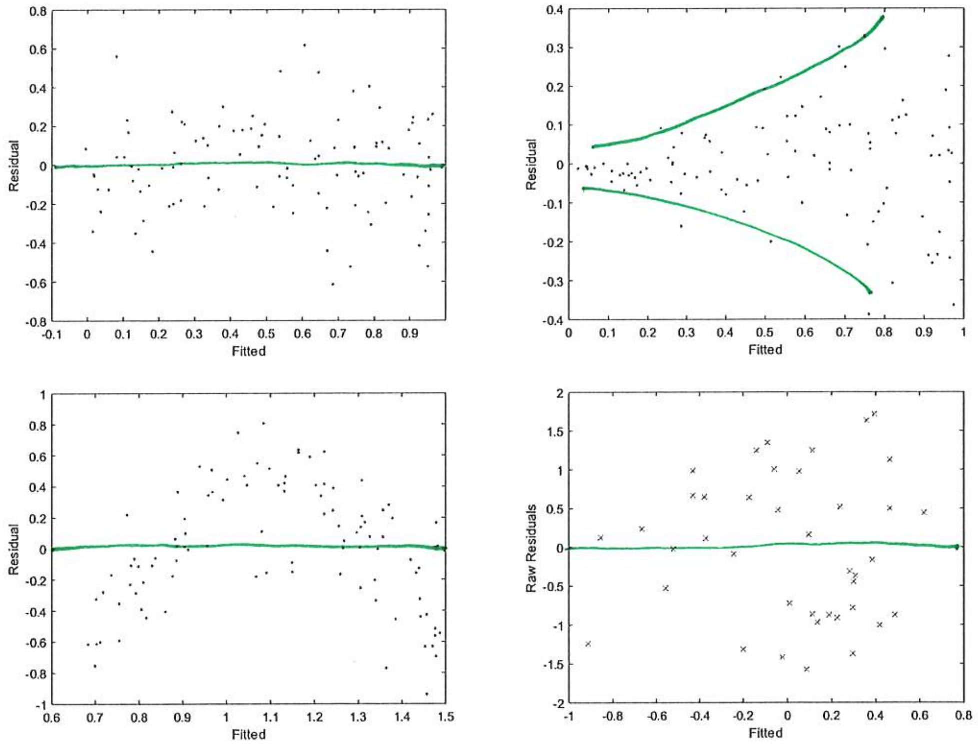
$$= (I - X(X^TX)^{-1}X^T)(X\beta + \varepsilon)$$

$$= \cancel{X\beta} + \varepsilon - \cancel{X(X^TX)^{-1}X^TX\beta} - X(X^TX)^{-1}X^T\varepsilon$$

So the residuals will not be independent and not necessarily have constant variance. However, this effect is usally small.

$$= (I - X(X^TX)^{-1}X^T)\,\varepsilon \simeq \varepsilon$$

One of the most useful diagnostic plots is the plot of the residuals $\widehat{\varepsilon}$ against the fitted values $\widehat{y} = \mathbf{x}\widehat{\beta}$. In this plot there should be constant symmetrical variation in the vertical direction.



These are plots of the residuals against the fitted values in four regression models. How would you describe these plots in relation to the assumptions of the linear regression model? Are they consistent with the assumptions of Linearity and Constant Variability? (We will use a different plot to check Normality.)

Top-Left: OK — no trend in residuals with fitted values + even spread

Top-Right: spread of residuals increased with fitted values - indicates variance not constant.

Bottom-Left: apparent trend in residuals - indicates mean is non linear in explanatory variable.

Bottom-Right (Facebook data):

no obvious departures from model assumptions - even spread (constant variance) and no trend (linear relationship between response + explanatory variable).

To check normality of the residuals we can use the *normal probability plot.* In general, a *probability plot* is used to graphically check that sample data has a specified distribution. The construction of this plot requires the *quantile function.* Recall (from Chapter 5) that the quantile function q for a (continuous) cumulative distribution function F is defined by

$$F(q(x)) = x.$$

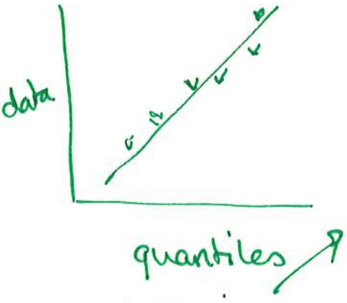
Suppose we have a sample of size n. Then the i-th smallest observation is plotted against the $(i - 0.5)/n$ quantile of the specified distribution. If the observations came from the specified distribution, then the points should lie close to a 45° line. Large departures from a 45° line would suggest that the data are not well modelled by the specified distribution.

The probability plot, as described, requires the theoretic distribution to be completely specified. This means to check that our observations come from a normal distribution using the probability plot we would need to know the mean and variance of the distribution. However, a nice property of the normal distribution means we can use the quantiles of the standard normal distribution in the probability plot. Let $q(x; \mu, \sigma^2)$ be the quantile function of the Normal(μ, σ^2) distribution, then

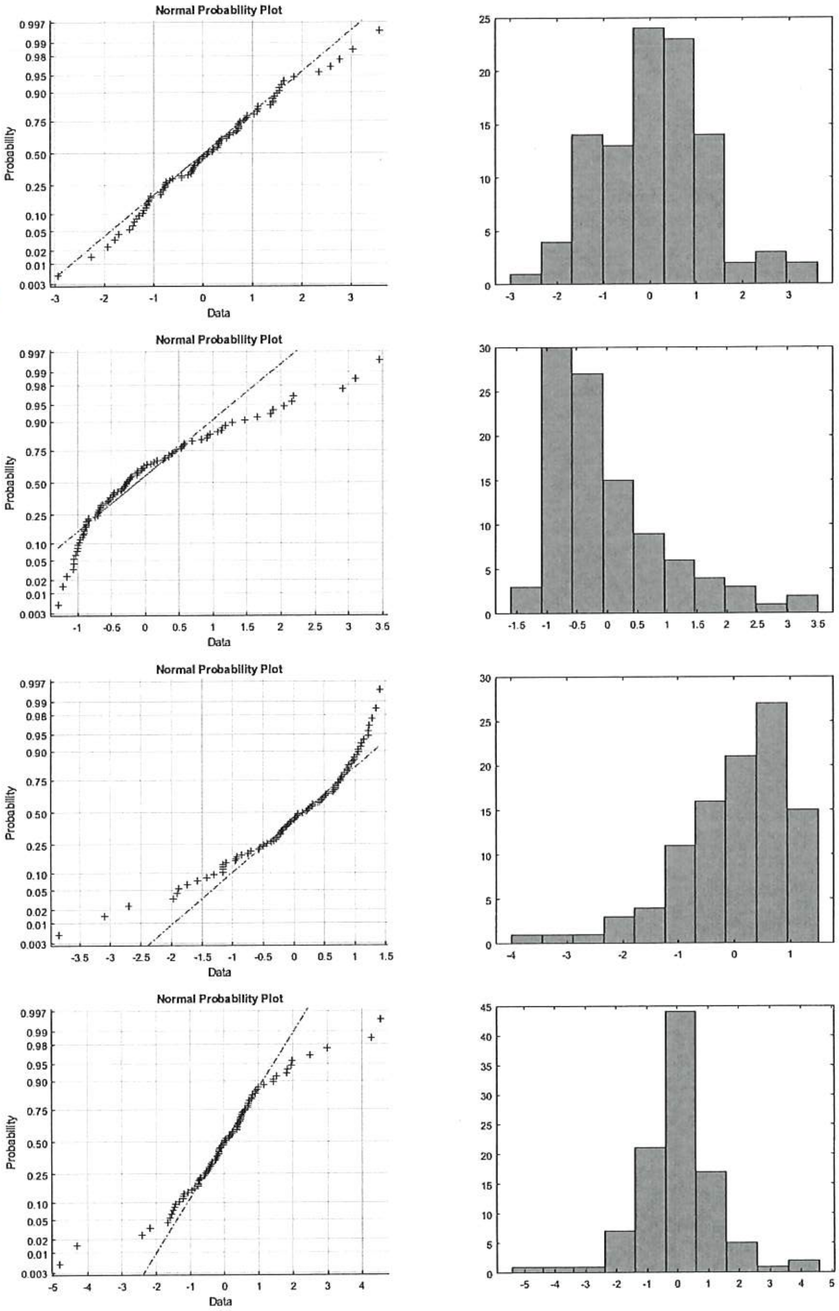
$$q(x; \mu, \sigma^2) = \mu + \sigma \times q(x; 0, 1).$$

So, if we plot the i-th smallest observation against the $(i-0.5)/n$ quantile of the standard normal distribution and those points lie on a straight line, this would indicate the observations can from a normal distribution. In MATLAB, the y-axis is marked with the probabilities $(i - 0.5)/n$ rather than values of the quantile function, though the spacing from the quantile function is used.

Quantile



consistent with normality.

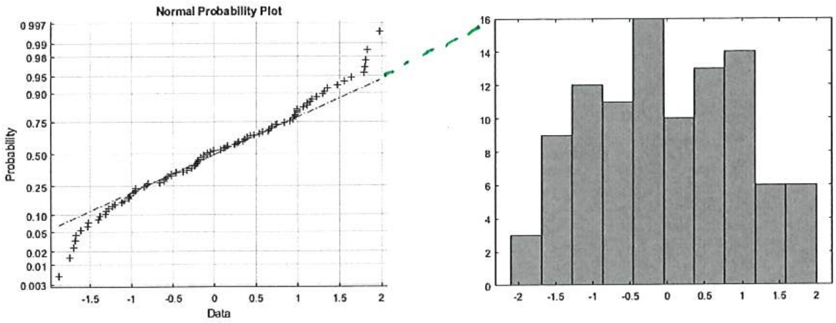


normal observations

right-skewed

left skewed

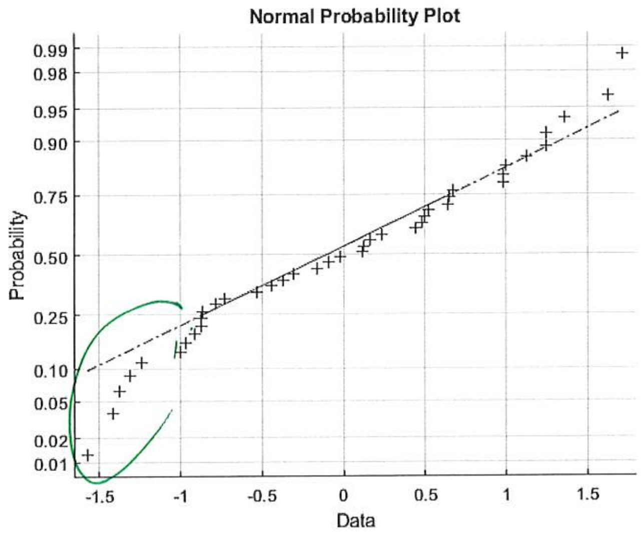
fat tails



short tails

Do the residuals from the facebook example appear normal?

```
facebooklm = fitlm(facebook,'GMDensity~Facebook');
normplot(facebooklm.Residuals.Raw)
```

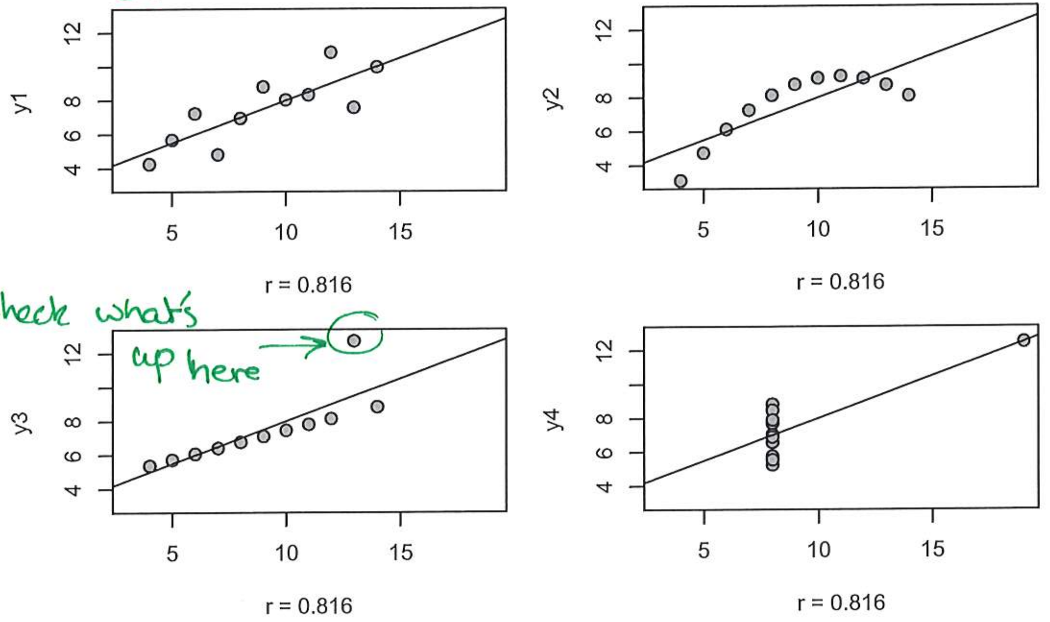


lef tail looks a little shorter than the normal distribution, otherwise ok.

Independence is another important assumption in our analysis. Unfortunately, there is no simple plot we can do which will tell use whether or not our observations are the realisation of independent random variables. The independence assumption is often called into question when our data has a temporal or spatial aspect.

Question – Anscombe's quartet. Four datasets are plotted below. The exact same linear regression can be fitted to each dataset, but for which, if any, is the linear regression model we have discussed appropriate?

ok



nonlinear

Check what's up here

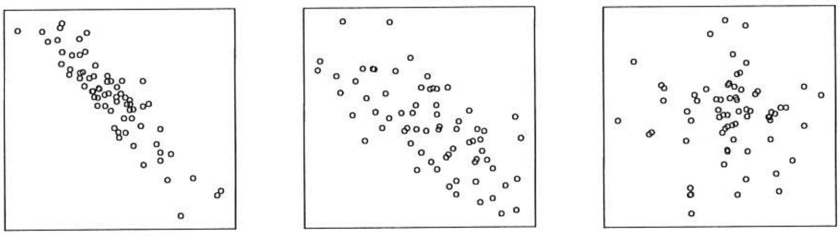
- two groups for explanatory var.
- single obs. in one group.

Correlation and the coefficient of determination

In this chapter we have taken one variable to be the response variable and the other to be the explanatory variable. Sometimes there is no reason to treat the two variables asymmetrically. Instead we might only be interested in the correlation between the two variables. Correlation can be estimated from the pairs of observation using the (Pearson) correlation coefficient, r. If the points in our scatter plot are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, then

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

where s_x and s_y are the sample standard deviations of the x and y values of the points. This estimate of correlation is always between -1 and 1.



$r = -0.9$ $r = -0.7$ $r = 0$