Regression Models
Lecture VI: Model Diagnostics
DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

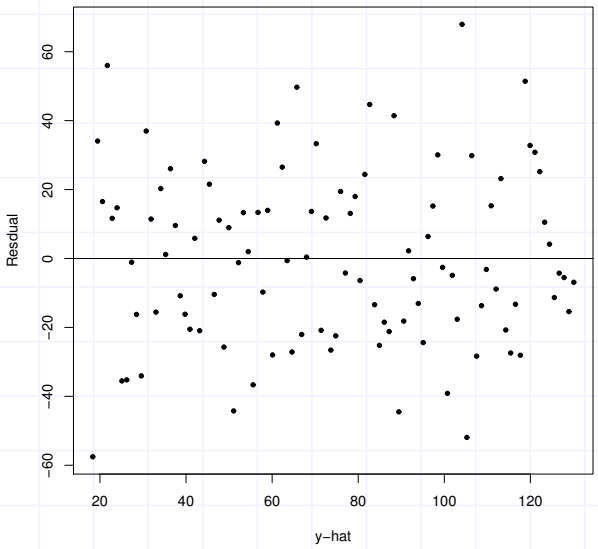
School of Mathematical Sciences
Technological University Dublin

Assessing Model fit - Residual Diagnostics, Influence and Leverage

- In fitting the above models and deciding what terms are significantly difference from zero etc., we have made some assumptions.
- In particular we have assumed that the residuals are iid normal with mean 0 and equal variance σ^2.
- Now we want to check that these assumptions are reasonable. We do this is validate what we are doing WRT model fitting. hypothesis testing and our conclusions.

Residual Diagnostics

If the fitted model and the model assumptions are correct then our residuals should be iid normal. If this is the case then if we plot the residuals against say their fitted values we should get a random scatter of points - a featureless cloud of points. For example, an ideal situation might look like the following,



We know from the model that the ε_i's are $N(0, \sigma^2)$ but what about their estimates, i.e. the e_i's?

We need to check that the e_i's follow the assumptions concerning the ε_i's.

Some properties of the e_i's are:

$$E[e_i] = 0$$

$$Var[e_i] = \sigma^2(1 - h_{ii}) \quad (1)$$

$$(2)$$

where h_{ii} is calculated from the 'hat matrix'.

The point is that the h_{ii} values are not the same - so the residuals do not have equal variance and tend to be correlated with each other.

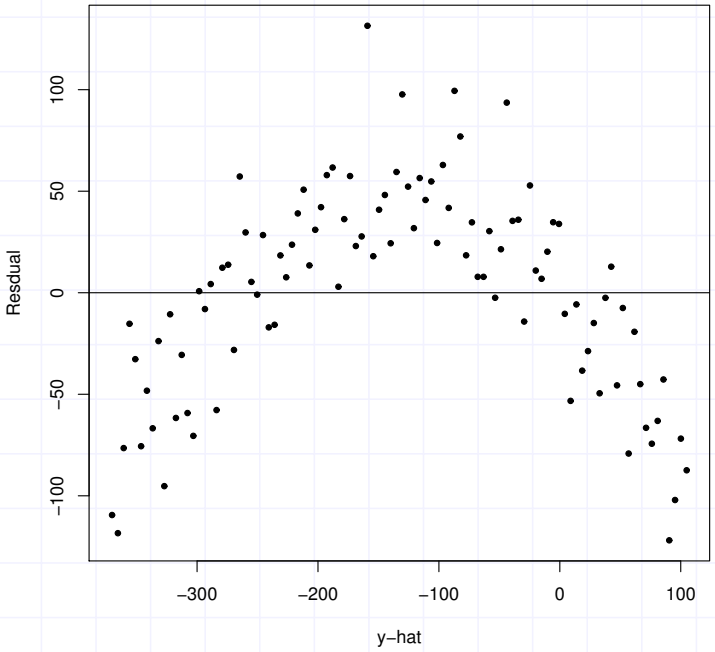
Despite this we can use the residuals in a simple way to show the following departures from the model assumptions;

- Model underspecification
- Departure from equal variation assumption (heterogeniety of variance)
- Existence of suspect data points (outliers)
- Departures from normality

Plotting of Raw Residuals

- The e's are the raw residuals. We can use them to detect model underspecification and non-constant variance.
- If the model is underspecified we will often see the structure of the underspecification in a plot of the residuals against the fitted values.
- So plot $e_i = (y_i - \hat{y}_i)$ against $\hat{y}_i$. Look for any structure in these plots which suggests problems.

The following plot suggests a missing higher degree polynomial term.

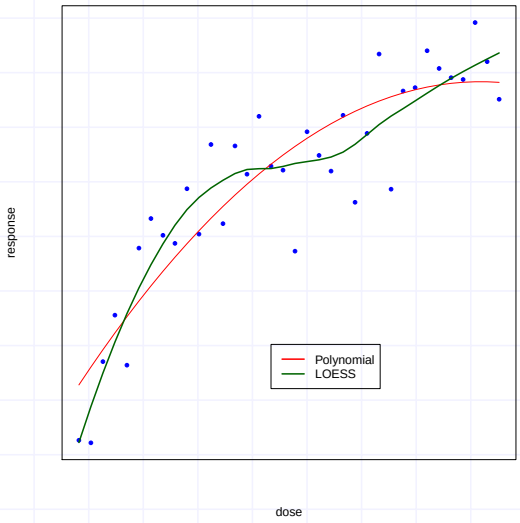


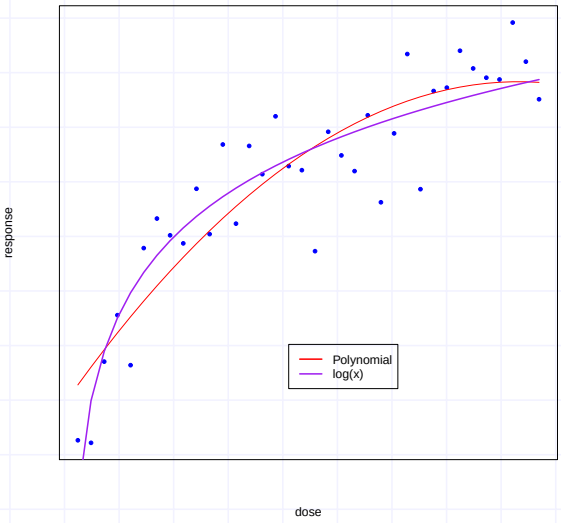
Remedy: Consider putting the appropriate polynomial term in the model.

Functional Form for Predictor

We have only reviewed polynomial functional forms for covariates.

What happens when they can't handle the nature of the relationship with the response?

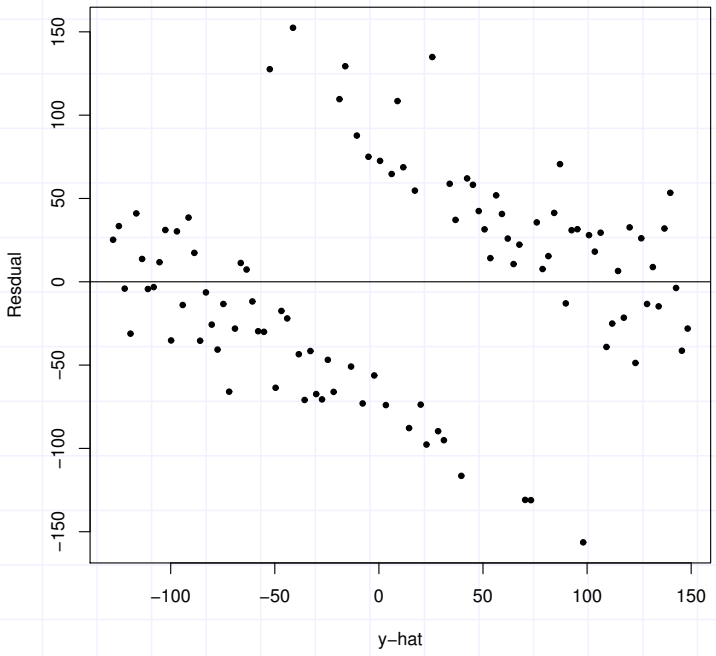




Some Methods for functional form

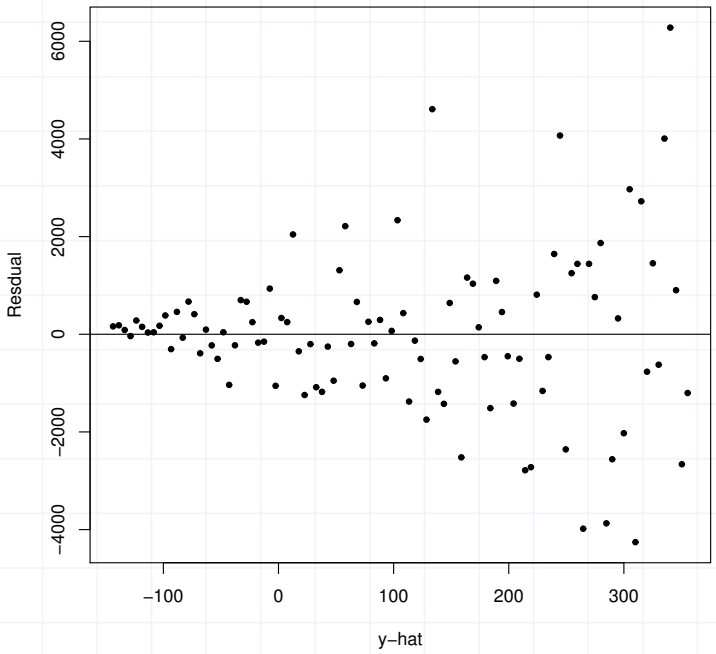
- Loess (Lowess): fit a least squares linear or quadratic to a small moving window of observations as the window moves left to right - join the fitted points together to give a flexible functional form.
- Use flexible functional forms such a splines or penalised splines - these are piecewise cubic which are forced to join together in a smooth way. Penalised splines impose a penalty on the complexity of the spline (not unlike the LASSO model).
- Transform the response (e.g. use $\log y_i$). Identify such transforms using plotting (learn what to look for), theoretical concerns or the Box-Cox method (e.g. `boxcox(.)` from the MASS library). Similarly transform the predictors perhaps by using the Box-Tidwell method (e.g. the `boxTidwell(.)` function from the car package).

The following suggests a missing grouping term (categorical term).



Remedy: Identify the missing categorical variable and include it in the model.

The following plot indicates that the variances are not constant.



Remedy: Use Generalised Least Squares (aka weighted regression)? Or transform your response (box-cox?). Or fit non-linear regressions?

Outlier Detection

- To identify a potential outlier we need to find a point that yields a residual that is big in absolute value compared to the rest.
- This suggests that this point(s) are poorly fitted by the model and therefore corrective action may be needed. The question is how big is too big?
- It would be helpful if the e_i's were standardised - i.e. had zero mean and unit variance.
- It can be shown that the e_i's will have zero mean when an intercept is included. But we have seen that their individual variances are given by $\sigma^2(1 - h_{ii})$ where h_{ii} is the i^{th} diagonal of the hat matrix. These will not be the same in general.

It can be shown that the variance of a residual for an observation which is far from the 'data centre' (i.e. has values that are at the extremes for predictors) will be small.

So, it is helpful to standardise the residuals to give them unit variance - so they are compared on the same scale.

One way of doing this is by **studentised residuals** (also called standardised residuals).

Studentised Residuals: $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$ (3)

This residual follows a t-like distribution (it is not exactly a t-distribution).

The denominator is the estimated standard error or the residual.

Example: Here are various residuals for the forestry data.

obs.	h_{ii}	residual	studentised	r-student
1	0.241	-0.005	-0.020	-0.019
2	0.218	-0.290	-1.118	-1.127
3	0.177	0.401	1.506	1.574
4	0.253	-0.351	-1.384	-1.428
5	0.135	0.047	0.172	0.166
6	0.246	0.128	0.503	0.491
7	0.075	-0.149	-0.527	-0.515
8	0.128	0.189	0.689	0.677
9	0.430	0.055	0.248	0.240
10	0.171	-0.571	-2.135	-2.444
11	0.165	0.125	0.468	0.456
12	0.182	-0.099	-0.374	-0.364
13	0.191	-0.230	-0.869	-0.862
14	0.103	-0.125	-0.450	-0.438
15	0.108	-0.194	-0.701	-0.689
16	0.163	0.185	0.687	0.676
17	0.175	-0.120	-0.448	-0.437
18	0.085	0.498	1.774	1.916
19	0.266	0.095	0.379	0.369
20	0.489	0.410	1.955	2.169

Compare observation (10) and (20) using (a) the raw residuals and (b) the studentised residuals.

What do we do with large residuals?

There are three main causes of large residuals:

1. true random variation - so do nothing,
2. a mistake has been made in the data collection - if this can be established to have happened then the observation can be removed from the analysis or corrected,
3. a model breakdown at some point in the data. In the case of (3) we have a particular problem - if the model breaks down at a given point then this is the same thing as an inadequate model for the data concerned.

There is a potential weakness of the studentised residual.

If an observation(s) is a true outlier, i.e. an observation that does not follow the model, then the s^2 will be inflated [recall what happens when we underfit].

If this is the case then is may be better to compute an s^2 estimate without the suspect observation included - denote this as s^2_{-i}.

This is an example of a 'leave one out analysis'. This is estimate is used instead of s^2 to give an **externally studentised** residual (hence the residual given by equation (3) is called internally studentised).

So the formula for the externally studentised (R-student) residual becomes,

$$\text{R-student: } t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} \qquad (4)$$

The R-student residuals follow a t distribution when testing certain model breakdown hypotheses. The model breakdown could be,

1. The model breakdown is caused by a location shift, i.e. $E[\varepsilon_i] \neq 0$. This is the mean shift outlier model. In this case we can use the R-student to test the null hypothesis $H_0 : \varepsilon_i = 0$.
2. The model breakdown is results in $Var[\varepsilon_i]$ being bigger than at the other data locations, i.e.$Var[\varepsilon_i] = \sigma^2 + \sigma_i^2$. In this case we can use the R-student to test the null hypothesis, $H_0 : \sigma_i^2 = 0$.

In both cases we use the critical values of the t distribution with $n - p - 1$ degrees of freedom. For example for the Forestry data, (n=20, p=3) we can compare with the critical value of the t-distribution with 16 DF (critical value at $\alpha = 0.05$ is 2.12).

But, if we have no *a priori* suspicion of any points and so are investigating them all simultaneously, then we need to correct for the Type I error rate.

Therefore, use the Bonferroni correction, so if α is the type I error rate use α/n where n is the number of residuals being investigated.

This will result in a Type I error rate across all the residuals that is no larger than α - so it is conservative.

For the Forestry data therefore, the correct critical value is 3.58, (i.e. value for $t_{.05/20,df=16}$).

It is probably better to use the critical values only as an approximate yardstick which points at suspect data points which should get more attention.

In fact, any point for which $|t_i| > 2$ may be flagged as 'suspect' since it is more than 2 standard deviations away from the mean.

Influence

- Influence refers to the amount of influence a particular data point has on the regression statistics.
- These statistics can be the estimated parameters, s^2 and/or other performance related statistics (e.g. residuals). What we are looking for is the answer to the question; if the i^{th} observation was omitted from the model, how would things change?
- Ideally we would like all observations to exert about the same amount of influence - or at least that no one observation or small number of them exert disproportionate amounts of influence.

Measuring Influence

We want to measure the actual influence on:

1. the fitted values and
2. the parameter estimates.

DFFITS
What would the fitted value $\hat{y}_i$ be for a set of predictors if observation i was not included in estimating the model regression parameters?

Denote the the fitted value with the ith observation included in the estimation $\hat{y}_i$ and denote $\hat{y}_{-i}$ the fitted value with it excluded from the estimation.

The logic is, if these two value are close, then observation i exerts little influence on the model fit. But, if they are far apart then observation i does exert strong influence on the model fit.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{-i}}{s_{-i}\sqrt{h_{ii}}} \quad (5)$$

where h_{ii} and s_{-i} are used to standardise the difference.

DFBETAS

This looks at the influence an observation has on the parameter estimates directly.

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{s_{-i}\sqrt{c_{jj}}} \quad (6)$$

where $\hat{\beta}_{j,-i}$ is the j^{th} parameter estimate calculated without the i^{th} observation, and $s_{-i}c_{jj}$ is the estimated standard error of the parameter.

Note also that there will be p DFBETAS values for each observation.

Cook's D

An aggregate measure of influence on all the parameters simultaneously is given by Cook's D (Cook's Distance).

Cook's D is a normalised measure of distance from the set of parameter estimates when the i^{th} observation included and excluded from the model fit.

To identify what parameter values a large D_i is influencing, use DFBETAS.

But, it is possible that a single observation has high influence spread over a number of parameters - i.e. doesn't stand out on any one parameter on there own.

Some authors advocate the use of yardstick for what are and are not large values for DFFITS, DFBETAS and Cook's D. for example;

Measure	Critical Yardstick
DFFITS	$2\sqrt{p/n}$
DFBETAS	$2/\sqrt{n}$
Cook's D	Use $\approx > 1$

Better to use comparisons among these measures - which is the biggest and why?

Is a suspect observation exerting high influence?

Is a particular sub-group of observation exerting high influence, etc.

These measures will identify any observation that are having a disproportionately large influence - the implications for the model may then be discovered.

Checking the normality assumption

- We assume that the errors are uncorrelated with the same variance.
- The residuals from the fitted model are centred and scaled to be uncorrelated with unit variance using the studentised residual.
- If we order these we should get an ordered group from a standard normal distribution.
- We can then use a result from order statistics that the expected value of the r^{th} order statistics from the standard normal is well approximated by:

$$q_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

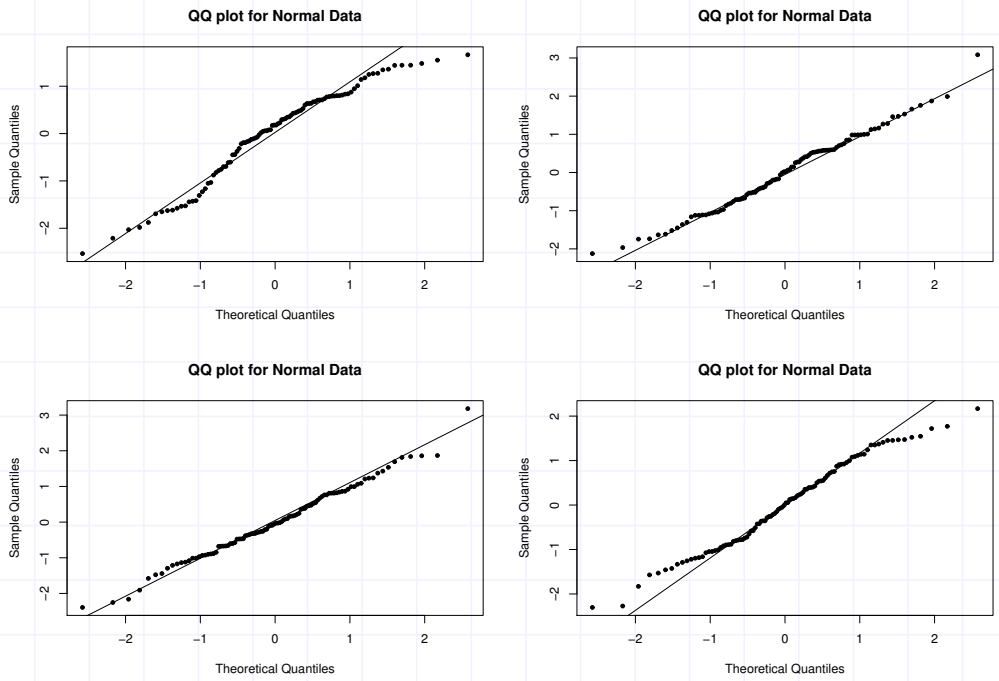
Where i is the order number of the q_i.

- An approach to check the normality assumption in linear models is to plot these expected **quantiles** against the observed quantiles of the studentised residuals.
- If the residuals are normally distributed then a straight line should be observed with an intercept at the origin. This is called *Normal probability Plot* or a *Quantile Quantile (QQ)* plot.
- In models involving real data we will never get a perfectly straight line, so it is important to learn to read a QQ plot to spot systematic departures from normality.

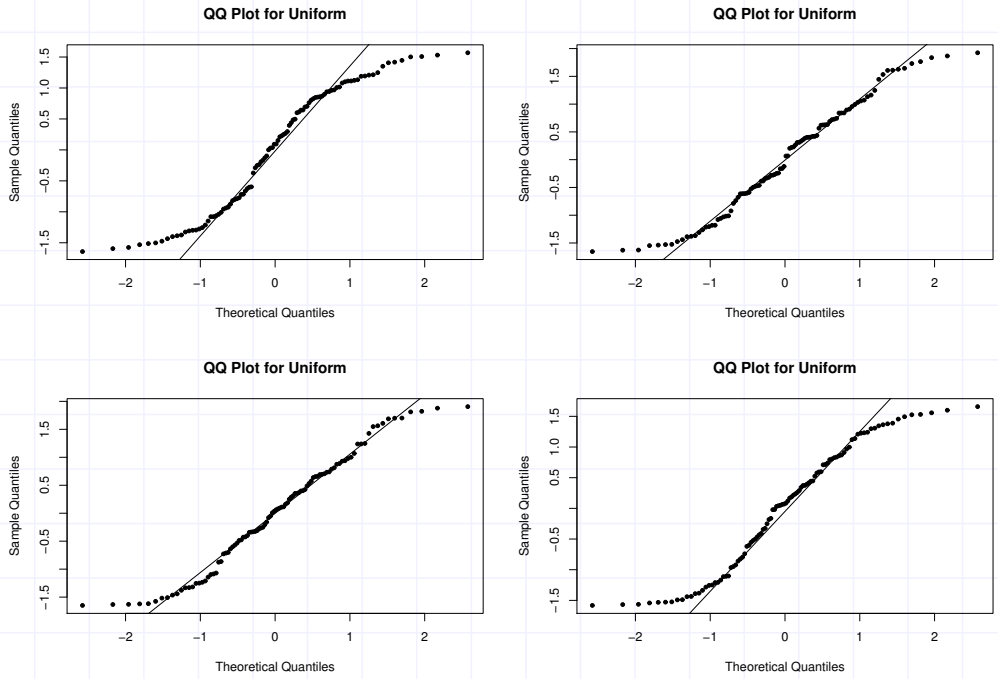
In the plots that follow the line is drawn though the 1^{st} and 3^{rd} quartiles of the data.

NB: You could also try formal goodness-of-fit tests such as Shaprio-Wilks or Anderson-Darling.

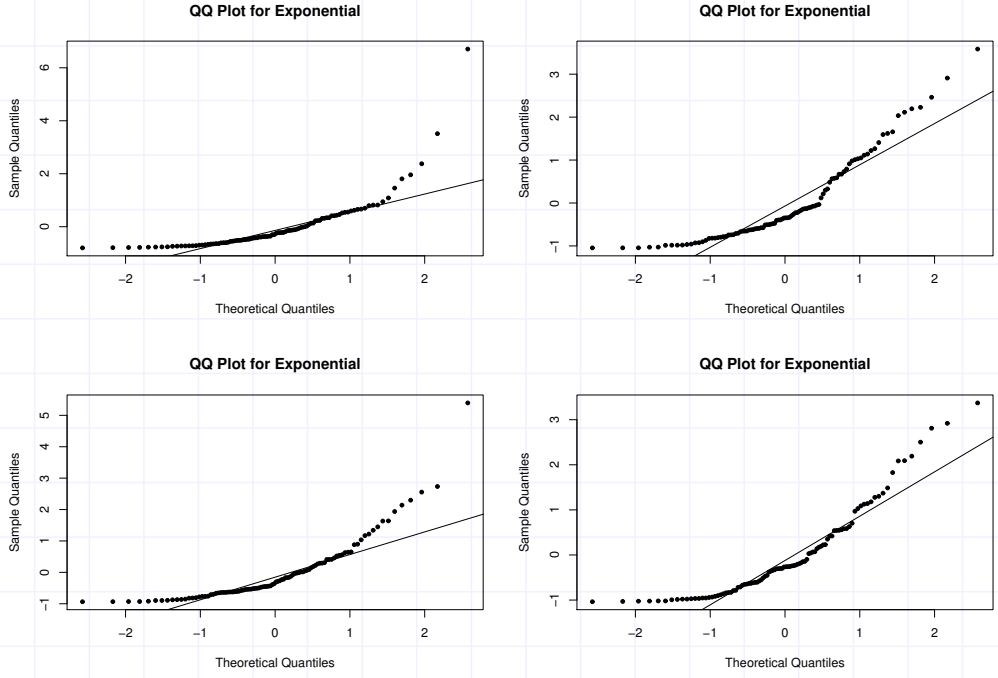
Normal Errors



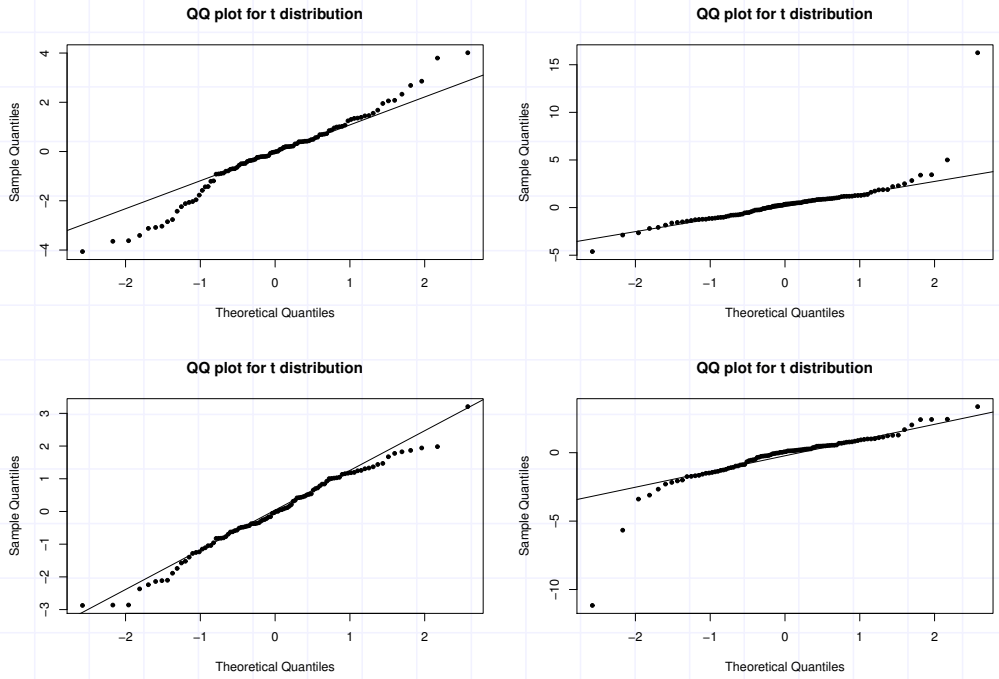
Light tailed Errors



Positive Skew, (i.e. light left tail, heavy right tail)



Heavy Tail Errors



Dealing with non-normality?

- Have you got the 'right' model - i.e. is there structure in the residuals due to an underfit?
- Are your responses normal - perhaps you should be fitting a non-normal response model such as Generalised Linear Models (GLMs)?
- Transformation of your response might help (this is what the Box-Cox method was developed for).
- Use non-parametric regression methods such as Loess with inference done using computer intensive methods like bootstrapping.

Multicollinearity

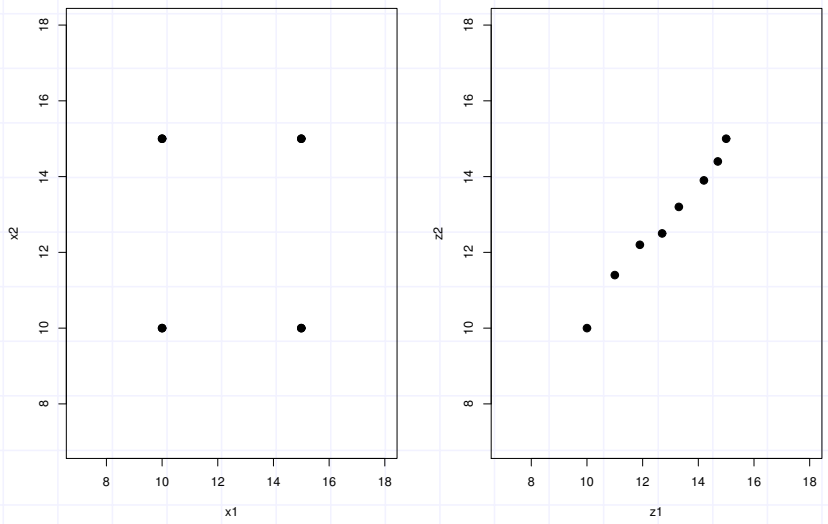
A highly desirable feature for linear models, is that the predictors in the model are uncorrelated with each other (orthogonal).

This results in the most efficient estimates of variances of the parameters and makes model building much easier.

Look at two extremes;

data 1:
x1 = 10, 10, 10, 10, 15, 15, 15, 15 $\sum = 100$
x2 = 10, 10, 15, 15, 10, 10, 15, 15 $\sum = 100$

data 2:
z1=10, 11.0, 11.9, 12.7, 13.3, 14.2, 14.7, 15 $\sum = 102.8$
z2=10, 11.4, 12.2, 12.5, 13.2, 13.9, 14.4, 15 $\sum = 102.6$



The set of data are virtually the same - but in data 2 the two variables are correlated.

Imagine these are predictor variables in a regression - what will the variance-covariance matrix for the β's look like in each case?

Since the numbers are virtually the same, you might expect a similar results, but it turns out:

$$Var[\beta_{(x_1,x_2)}] = s^2 \begin{bmatrix} 6.375 & -0.250 & -0.250 \\ -0.250 & 0.020 & 0.000 \\ -0.250 & 0.000 & 0.020 \end{bmatrix}$$

$$Var[\beta_{(z_1,z_2)}] = s^2 \begin{bmatrix} 11.631 & 2.845 & -3.748 \\ 2.845 & 2.835 & -3.063 \\ -3.748 & -3.063 & 3.361 \end{bmatrix}$$

So, multicollinearity leads to inflation of the variances of the coefficients.

Also note that the covariances have also increased in absolute value (this will be an issue for using general linear hypotheses).

Symptoms of Multicollinearity

Multicollinearity between predictors is often characterised by,

- Parameters that were highly significant suddenly become insignificant in the presence of additional variables. Or, why two variables which are individually significant, will both be insignificant when used in the model together.
- Several apparently different models fitting equally well.
- Adding or removing a predictor, causes a large change in the estimates of one or more other parameters (although it should be noted that this happens naturally in polynomial models).
- Parameter estimates which have unexpected sign or unexpected size.

```
fit1=lm(body_fat~Age+Height+
    Neck,data=bf_small)
cbind(round(fit1$coeff,3))
```

```
(Intercept) -21.828
Age           0.107
Height       -0.443
Neck          1.772
```

```
fit2=lm(body_fat~Age+Height+
    Neck+Chest,data=bf_small)
cbind(round(fit2$coeff,3))
```

```
(Intercept) -28.554
Age           0.088
Height       -0.311
Neck         -0.228
Chest         0.735
```

```
fit3=lm(body_fat~Age+Height+
    Neck+Chest+Weight,data=bf_
    small)
cbind(round(fit3$coeff,3)) %$
```

```
(Intercept) -12.728
Age           0.113
Height       -0.344
Neck         -0.470
Chest         0.555
Weight        0.072
```



```
> summary(fit1);summary(fit2);summary(fit3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.82765   12.39738  -1.761  0.08093 .
Age           0.10712    0.05141   2.083  0.03941 *
Height       -0.44251    0.14775  -2.995  0.00336 **
Neck          1.77156    0.25556   6.932 2.49e-10 ***

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.55369   10.47112  -2.727   0.0074 **
Age           0.08803    0.04333   2.032   0.0445 *
Height       -0.31079    0.12568  -2.473   0.0149 *
Neck         -0.22843    0.35757  -0.639   0.5242
Chest         0.73522    0.10505   6.999 1.83e-10 ***

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.72783   15.07424  -0.844  0.40025
Age           0.11266    0.04633   2.432  0.01659 *
Height       -0.34446    0.12721  -2.708  0.00782 **
Neck         -0.47023    0.39284  -1.197  0.23379
Chest         0.55472    0.16236   3.417  0.00088 ***
Weight        0.07201    0.04956   1.453  0.14898
```

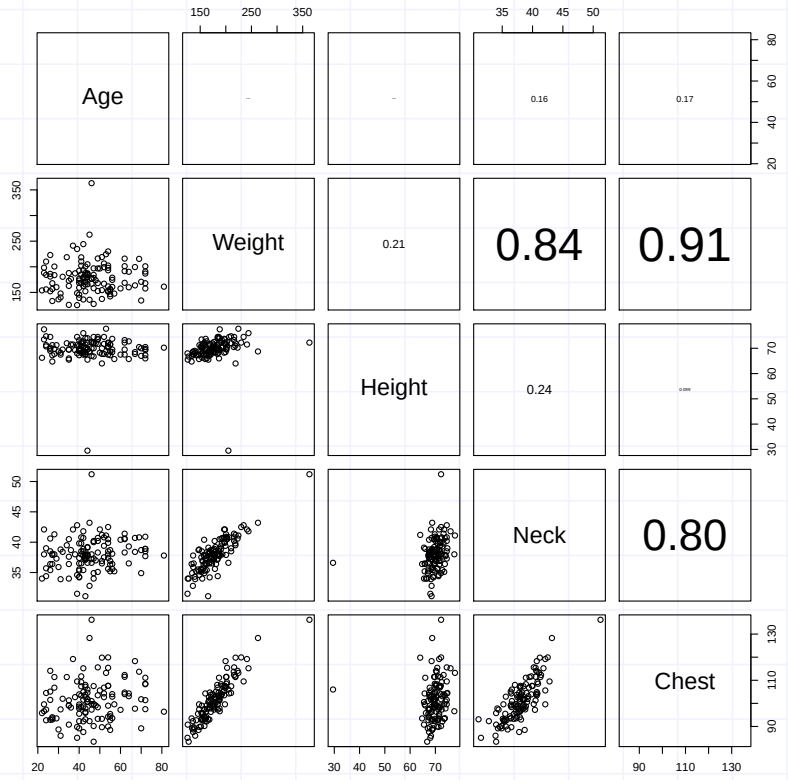
Spotting Multicollinearity

Your should compute the correlation matrix among your predictors.

```
> bf_small=read.csv('bodyfat_small.csv',header=T)
> preds=bf_small[,3:7]
> round(cor(preds),2)
          Age Weight Height Neck Chest
Age      1.00   0.02  -0.05 0.16  0.17
Weight   0.02   1.00   0.21 0.84  0.91
Height -0.05   0.21   1.00 0.24  0.10
Neck     0.16   0.84   0.24 1.00  0.80
Chest    0.17   0.91   0.10 0.80  1.00
```


Plot correlation lattice

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor
    , ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(preds,upper.panel=panel.cor)
```



Variance Inflation Factor

The drawback of pairwise correlations is that multicollinearity may not be apparent in the pairwise correlations.

The problem can also be measured using variance inflation factors (VIFs).

VIF is the ratio of the variance of a parameter estimate compared to the orthogonal-model (uncorrelated) gold standard.

It can be calculated as :

$$VIF = \frac{1}{1 - R_i^2}$$

where R_i^2 is from the regression of the ith predictor on the other predictors in the model.

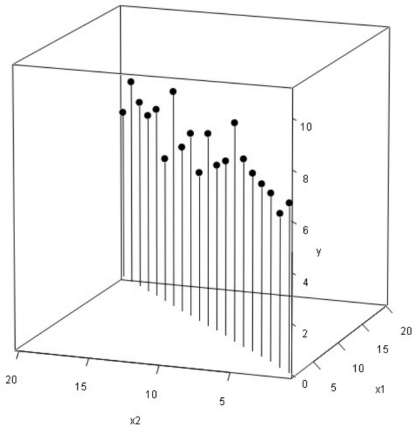
Thus the higher this linear relationship the lower the precision of the parameter estimate (i.e. higher variances).

As an approximate rule of thumb, a VIF approaching/exceeding 10 is cause for concern and further investigation.

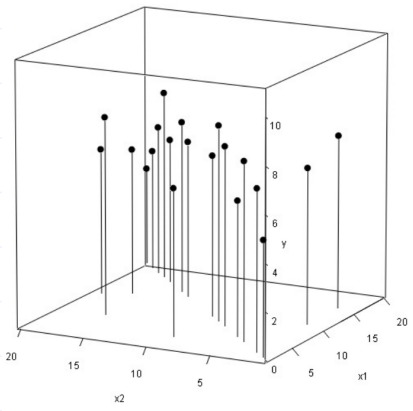
What are the implications of high VIFs?

- More likely to make a type II error on tests of parameters - i.e. not include important terms in the model.
- More difficult to identify a good candidate model using model building techniques
- More variance for predicted values - as these are functions of the parameter estimates.
- Numerical instability in the parameter estimates. look at the following two graphics;

Highly Collinear



Less Collinearity




```
> library(car)
> vif(fit3)
     Age   Height     Neck    Chest   Weight
1.197366 1.128901 3.682077 6.957515 8.344479
```

Countering Multicollinearity

There are a number of possibilities.

- Remove one or more predictors from the model which are causing the collinearity.
- Replace two correlated variables with a new variable which is a function of the two, e.g. $z = x_1 + x_2$.
- Use Ridge regression. This is a biased estimation method which makes the model behave more like an orthogonal model. It is not unlike the Lasso model in formulation.
- Use principal components regression. Here the basic idea can be considered as replacing the original predictors with the 'principal components' which are orthogonalised linear functions of the predictors - very like (can be identical) to factor scores from factor analysis.