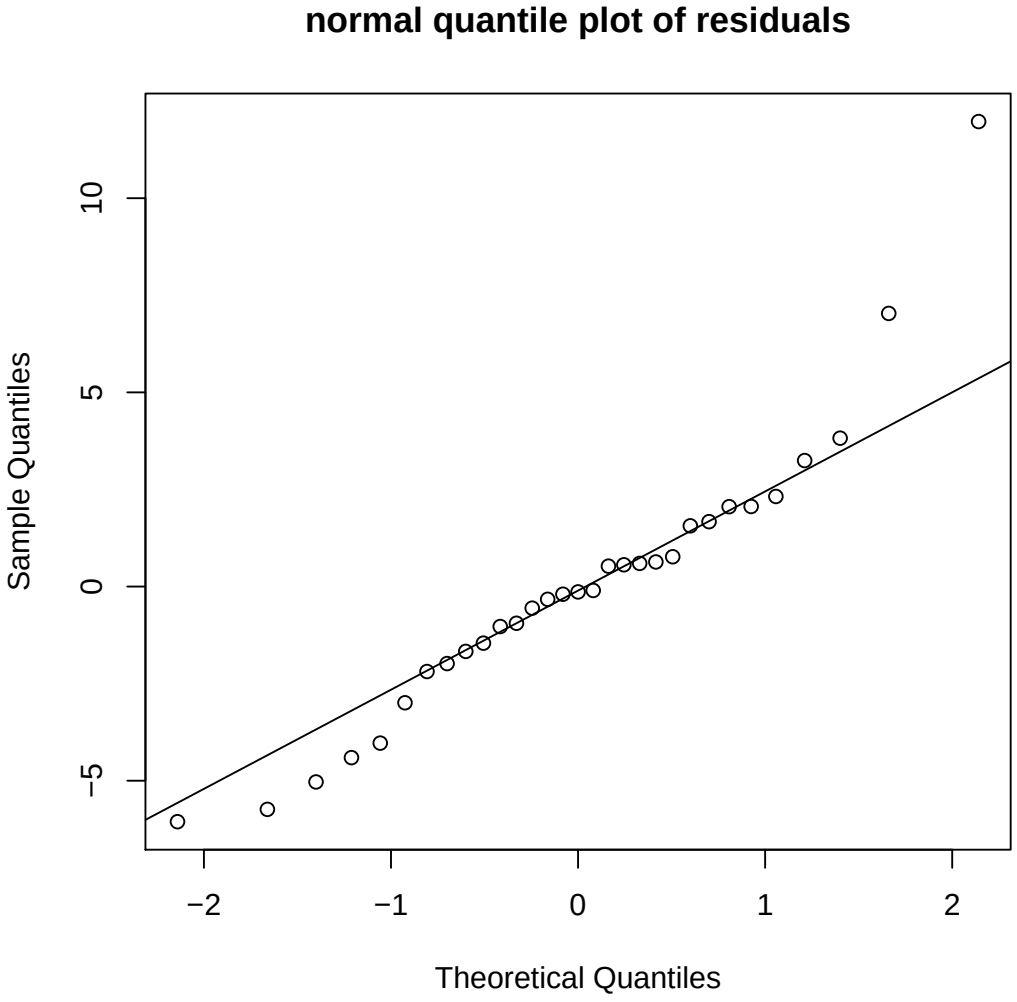
1 Residual analysis (section 4.2)

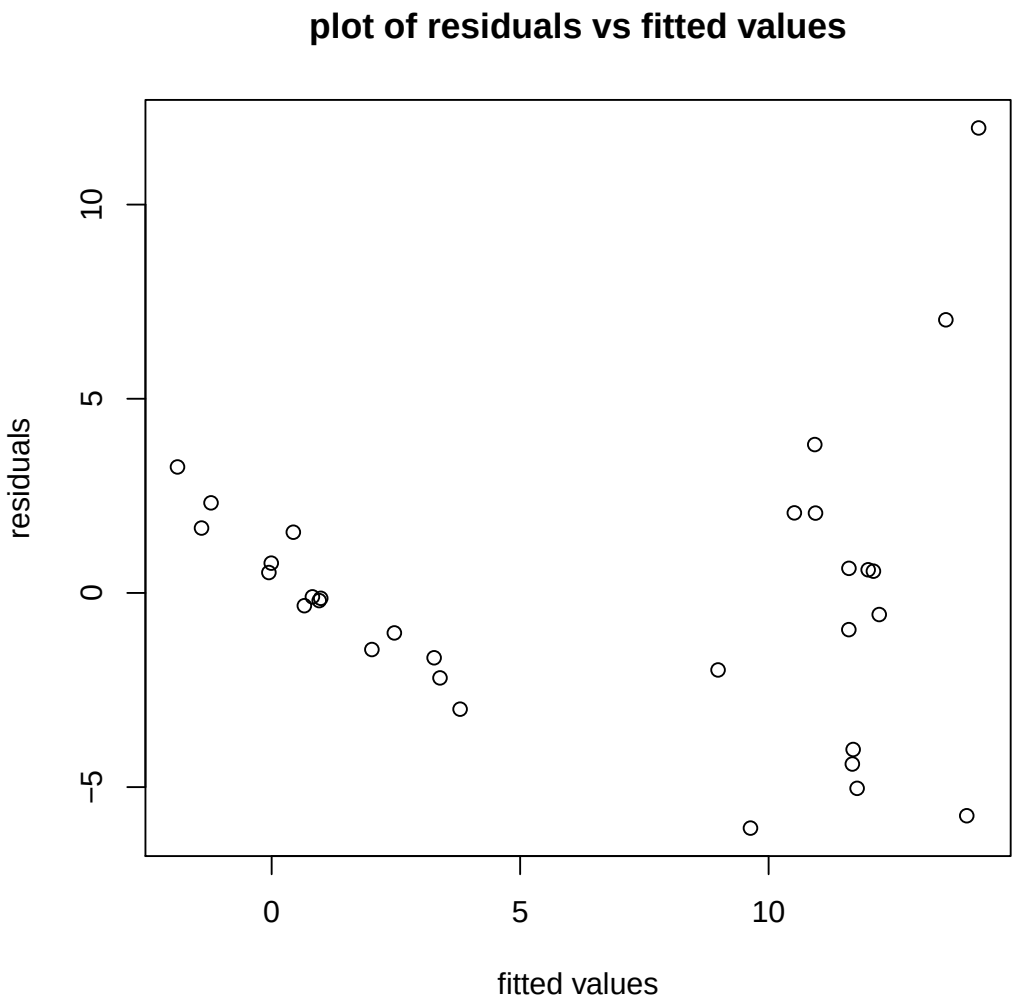
- residual analysis is used to assess the validy of model assumptions, and suggest corrective measures if model assumptions seem not to be me.
- The assumptions are:
  1. the regression model is correct. that is, it includes the correct set of predictor variables.
  2. the errors are i.i.d. $N(0, \sigma^2)$ random variables.
- There is no way to test for independence of the errors. Good experimental design is needed to ensure independence.
- A normal probability plot is useful to assess the assumption of normality of the residuals.
- To assess the assumption that the functional form of the model is correct, it is useful to plot the residuals against the fitted values $\hat{y}$, and individually against individual predictor variables. These plots may suggest that the variance of the errors in not constant (any may suggest a transformation to stabilize the variance [section 5.2]), or they may suggest that the functional form of the model is incorrect (for example, that we need to include a quadratic term x^2 in the regression model in additional to a linear term x [section 5.3, 5.4])
- A common way to model non-linear relationships is to fit a polynomial regression [chapter 7].
- An **added variable plot**, also known as a partial regression plot, is useful to assess the functional form of a regressor, for example, linear or quadratic, given that other regressors are included in the model.

```
> data=read.csv("asphalt.csv",header=T)
> data=data[,-c(1)]
> lm.out=lm(rutdepth~viscosity+surface+base+run+fines+voids,data=data)
> lm.resid=residuals(lm.out)
> lm.fits=fitted(lm.out)
> qqnorm(lm.resid,main="normal quantile plot of residuals")
> qqline(lm.resid)
```



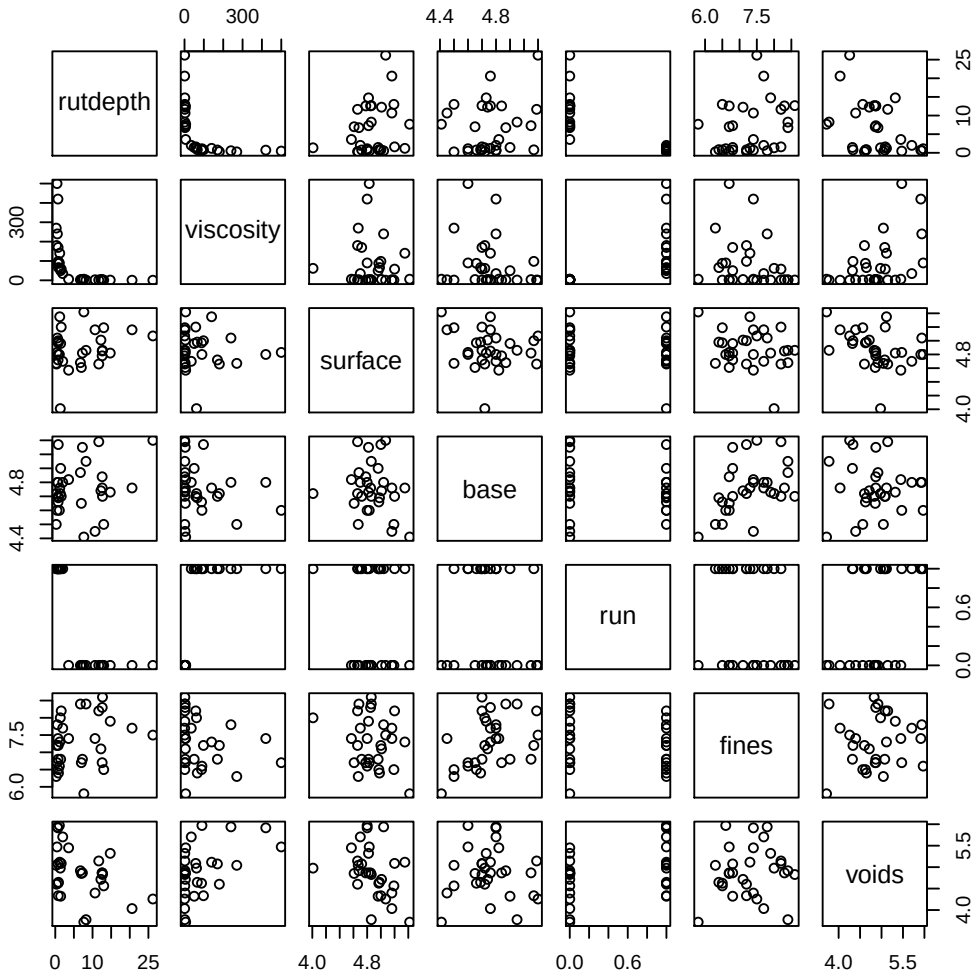
- The distribution of the residuals appears to be long tailed relative to the normal distribution

```
> plot(lm.fits,lm.resid,main="plot of residuals vs fitted values",
+     xlab="fitted values",ylab="residuals")
```

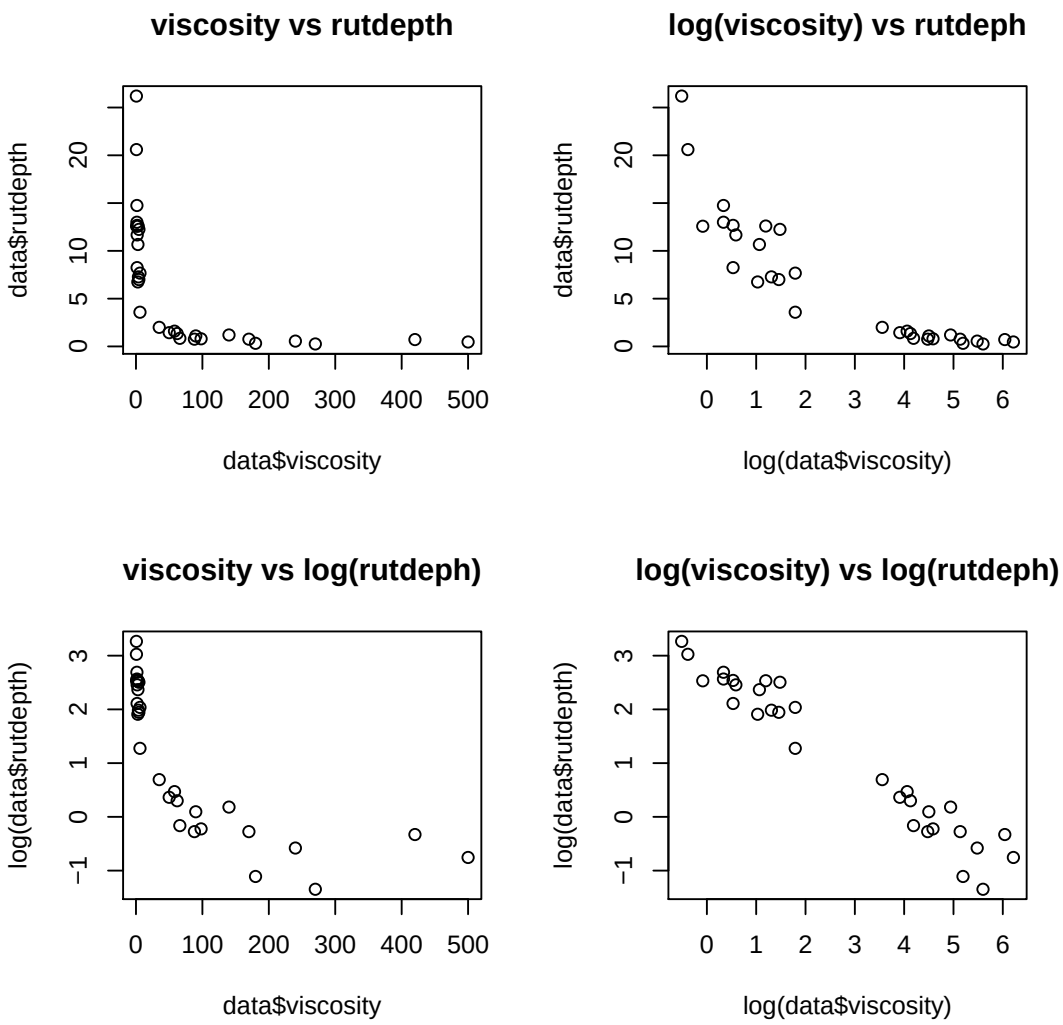


- there appears to be a nonlinear trend in the residual plot, suggesting the need for a transformation of one or more predictor variables or the dependent variable.

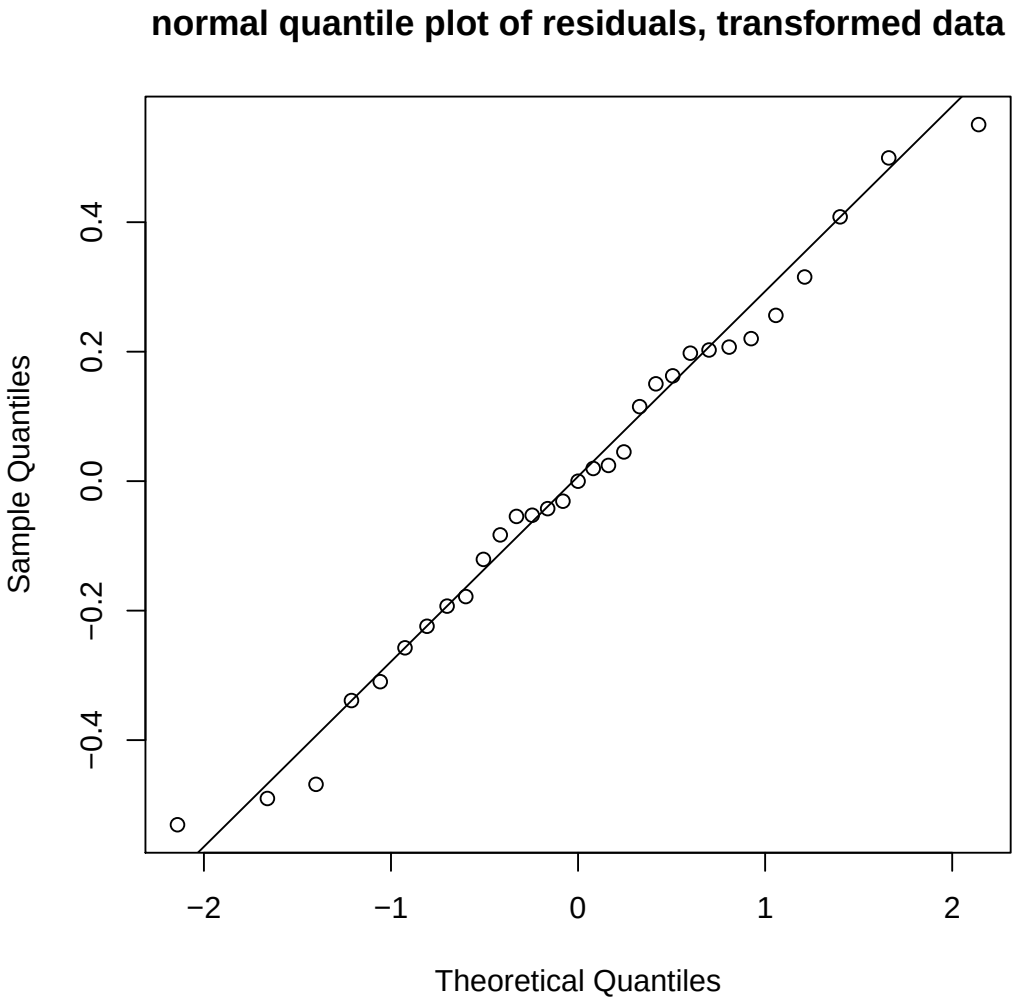
```
> pairs(data)
```


rutdepth
viscosity
surface
base
run
fines
voids
0
300
4.4
4.8
6.0
7.5
25
10
0.6
5.5
4.0

```
> par(mfrow=c(2,2))
> plot(data$viscosity,data$rutdepth,main="viscosity vs rutdepth")
> plot(log(data$viscosity),data$rutdepth,main="log(viscosity) vs rutdeph")
> plot(data$viscosity,log(data$rutdepth),main="viscosity vs log(rutdeph)")
> plot(log(data$viscosity),log(data$rutdepth),main="log(viscosity) vs log(rutdeph)")
```



```
> logvisc=log(data$viscosity)
> logdepth=log(data$rutdepth)
> data=cbind(data,logvisc,logdepth)
> lm.out2=lm(logdepth~logvisc+surface+base+run+fines+voids,data=data)
> lm.resid2=residuals(lm.out2)
> lm.fits2=fitted(lm.out2)
> par(mfrow=c(1,1))
> qqnorm(lm.resid2,main="normal quantile plot of residuals, transformed data")
> qqline(lm.resid2)
```



```
> plot(lm.fits2,lm.resid2,
+   main="plot of residuals vs fitted values,transformed data",
+   xlab="fitted values",ylab="residuals",col=data$run+1)
```