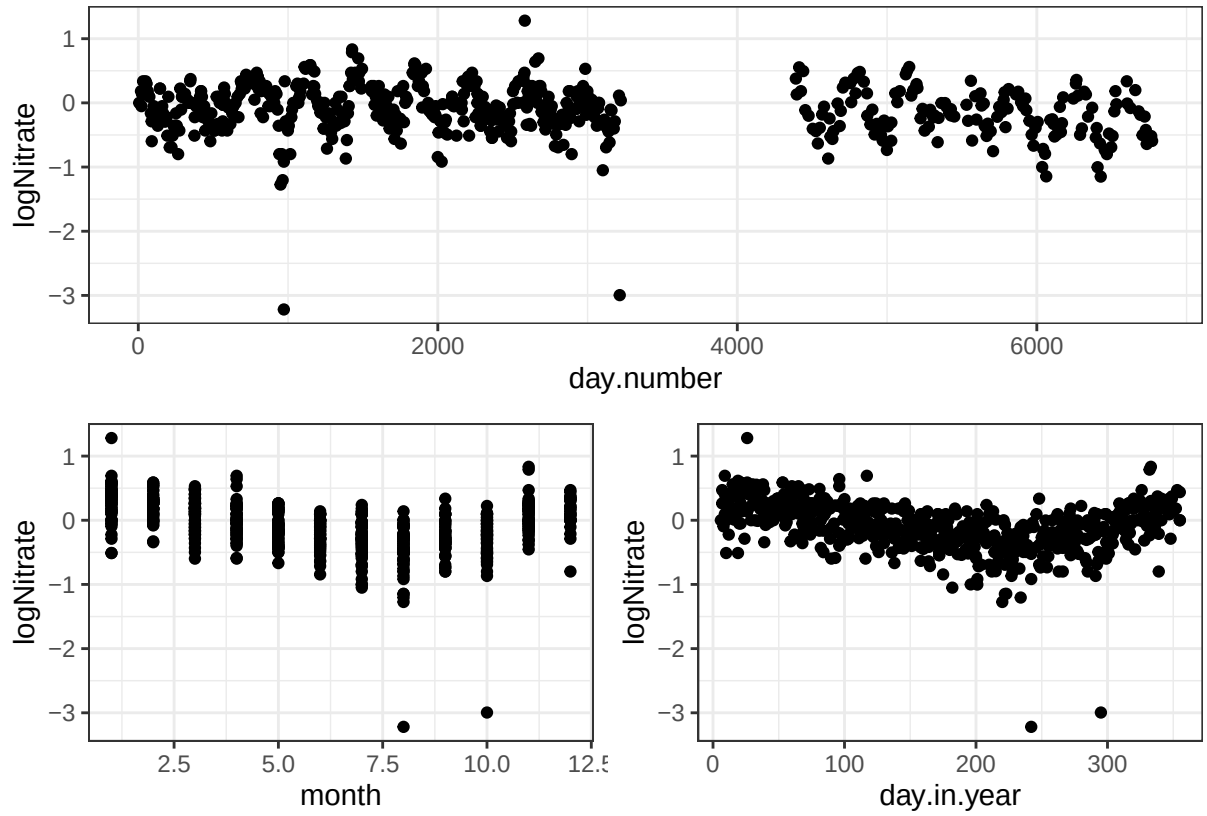
Week 10 solutions

Example 1

(i)

We want to produce plots of the nitrate concentration against each temporal independent variable.

```
tweed <- read.table("tweed.txt", header=TRUE)

ggplot(tweed) +
  geom_point(aes(day.number, logNitrate)) +
  theme_bw() ->
  p1

ggplot(tweed) +
  geom_point(aes(month, logNitrate)) +
  theme_bw() ->
  p2

ggplot(tweed) +
  geom_point(aes(day.in.year, logNitrate)) +
  theme_bw() ->
  p3

p1/(p2+p3)
```



It looks like there is some repetition over the course of a year, we should incorporate this as:

$$\beta_1 \sin\left(\frac{2\pi \texttt{day.in.year}}{365}\right) + \beta_2 \cos\left(\frac{2\pi \texttt{day.in.year}}{365}\right)$$

(ii) Fitting this model will then be:

```
mod_per <- lm(logNitrate ~ sin(2*pi*day.in.year/365) + cos(2*pi*day.in.year/365), data=tweed)
```

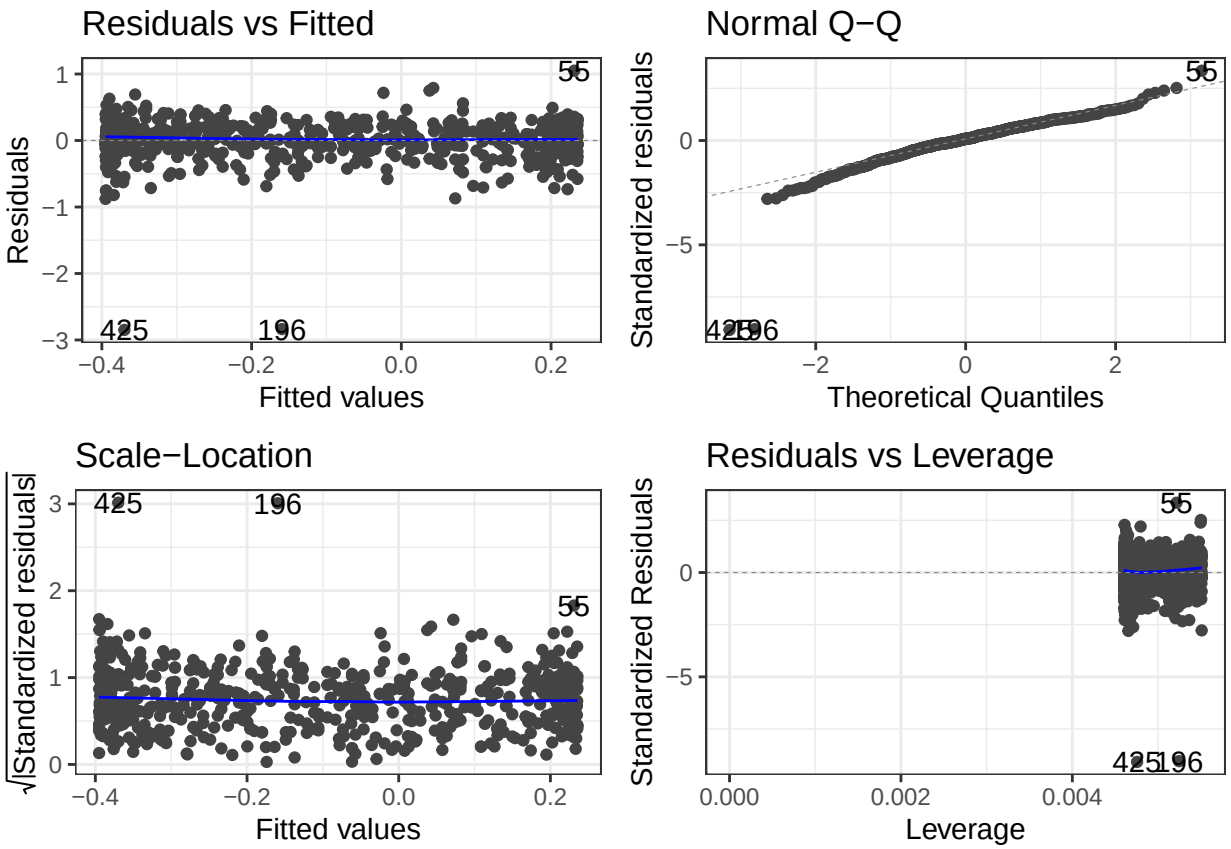
(ii) Checking the modelling assumptions:

```
require(ggfortify)
```

```
## Loading required package: ggfortify
```

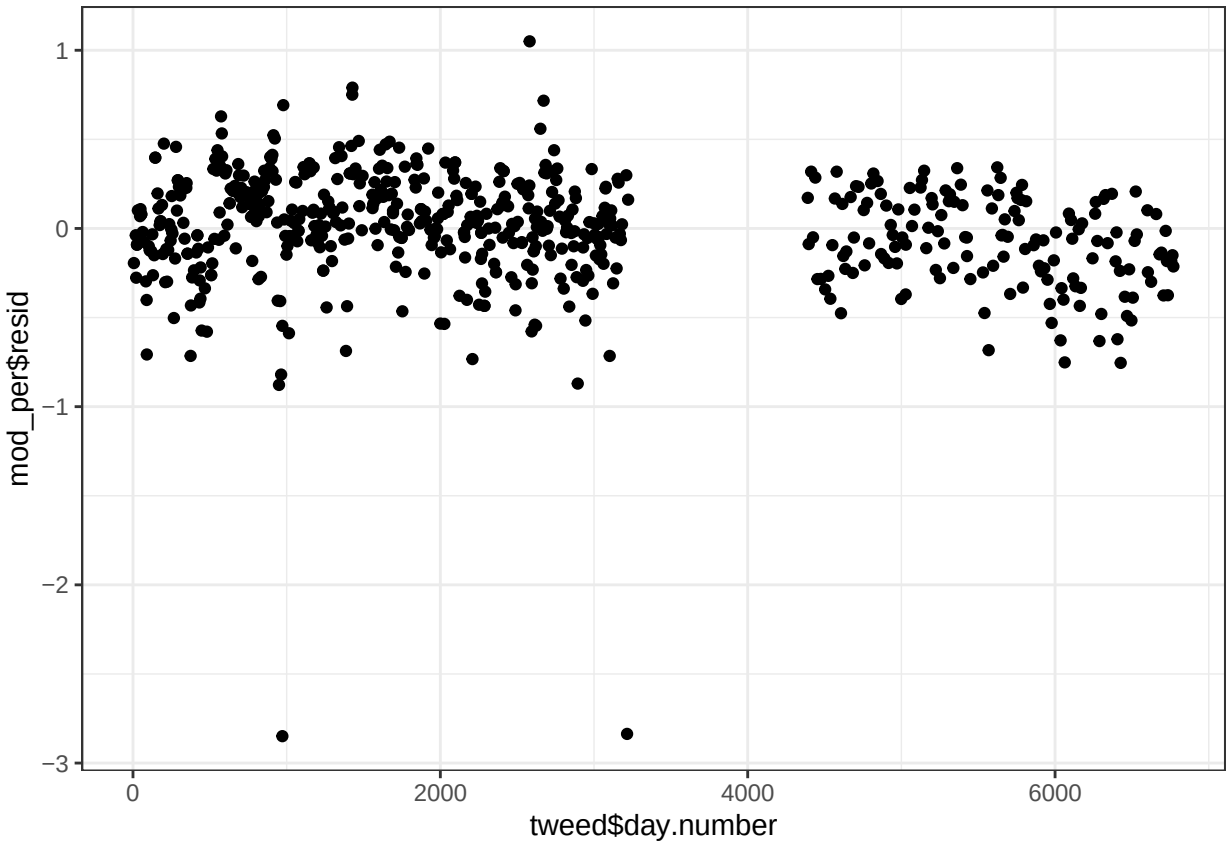
```
## Warning: package 'ggfortify' was built under R version 4.2.2
```

```
autoplot(mod_per) +
  theme_bw()
```



Residuals look alright based on the residual plots. Should also plot against the day number or month number to fully assess the independence assumption:

```
ggplot() +
  geom_point(aes(tweed$day.number, mod_per$resid)) +
  theme_bw()
```



It looks as though there is an underlying long term trend that has not been accounted for by the model. This is slightly curved so let's try incorporating a polynomial of degree 2:

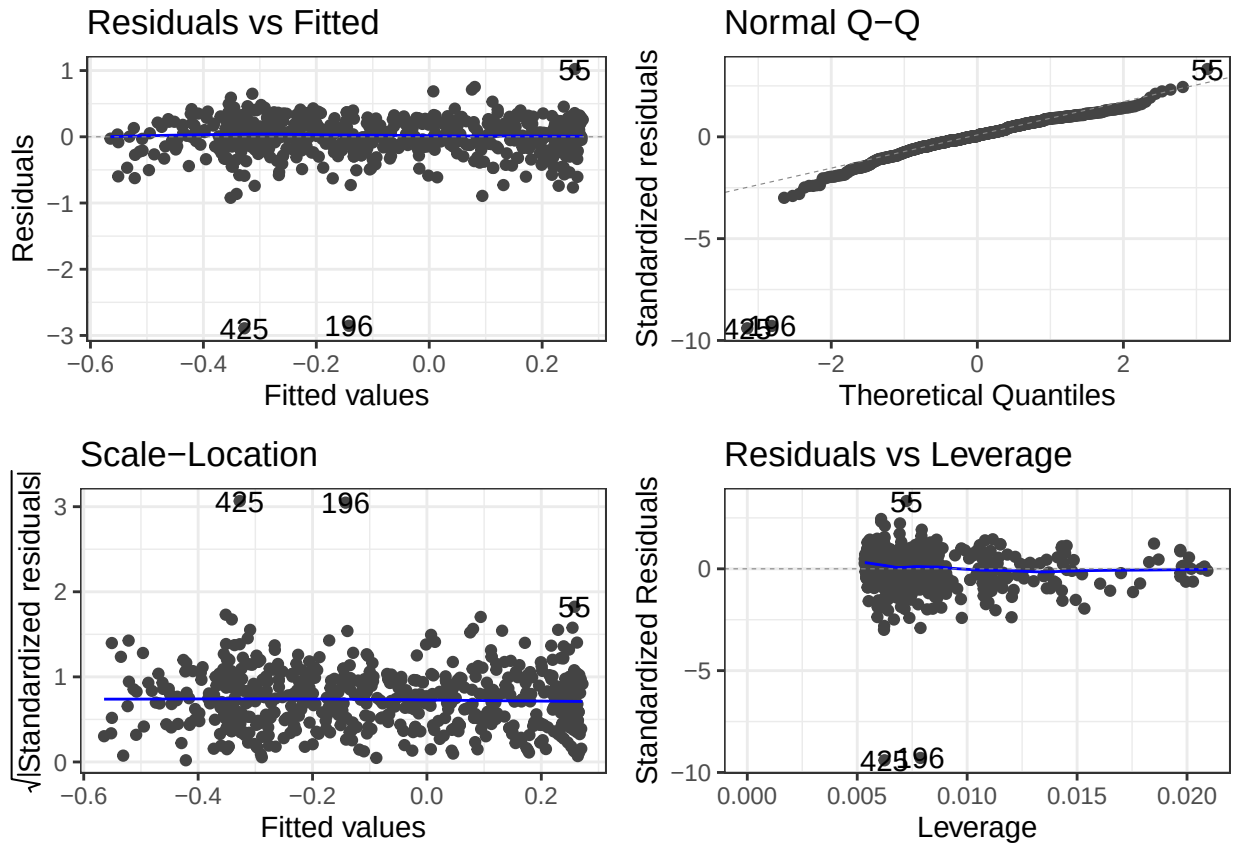
```
mod_per <- lm(logNitrate ~ sin(2*pi*day.in.year/365) +
                cos(2*pi*day.in.year/365) + poly(day.number, 2), data=tweed)

summary(mod_per)
```

```
##
## Call:
## lm(formula = logNitrate ~ sin(2 * pi * day.in.year/365) + cos(2 *
##     pi * day.in.year/365) + poly(day.number, 2), data = tweed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89157 -0.14006  0.01354  0.19868  1.02292
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.07995    0.01253  -6.382 3.47e-10 ***
## sin(2 * pi * day.in.year/365)  0.18205    0.01746  10.428  < 2e-16 ***
## cos(2 * pi * day.in.year/365)  0.25160    0.01795  14.017  < 2e-16 ***
## poly(day.number, 2)1          -1.40986    0.30902  -4.562 6.12e-06 ***
## poly(day.number, 2)2          -0.70127    0.30862  -2.272   0.0234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

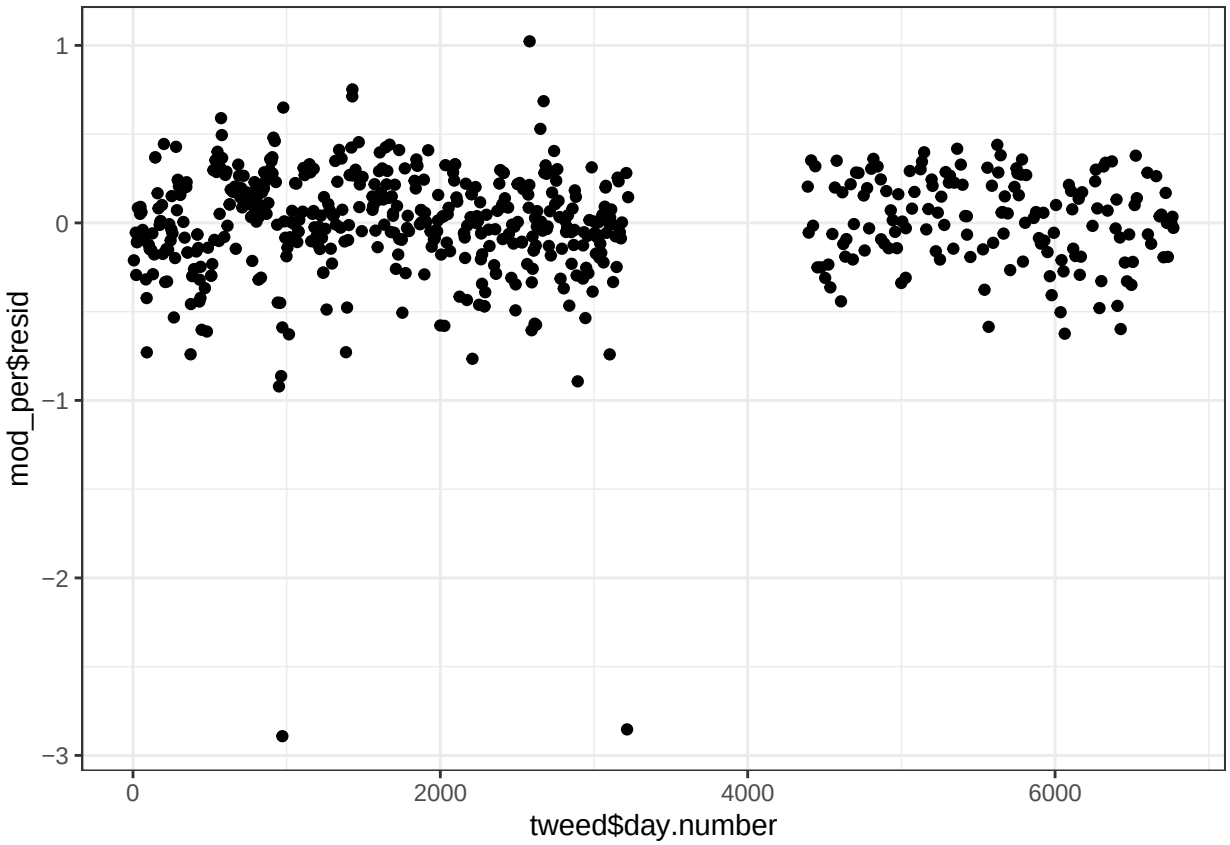
```
##
## Residual standard error: 0.3085 on 607 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3614
## F-statistic: 87.43 on 4 and 607 DF,  p-value: < 2.2e-16
```

The polynomial terms are significant - residual checks again:

```
autoplot(mod_per) +
  theme_bw()
```

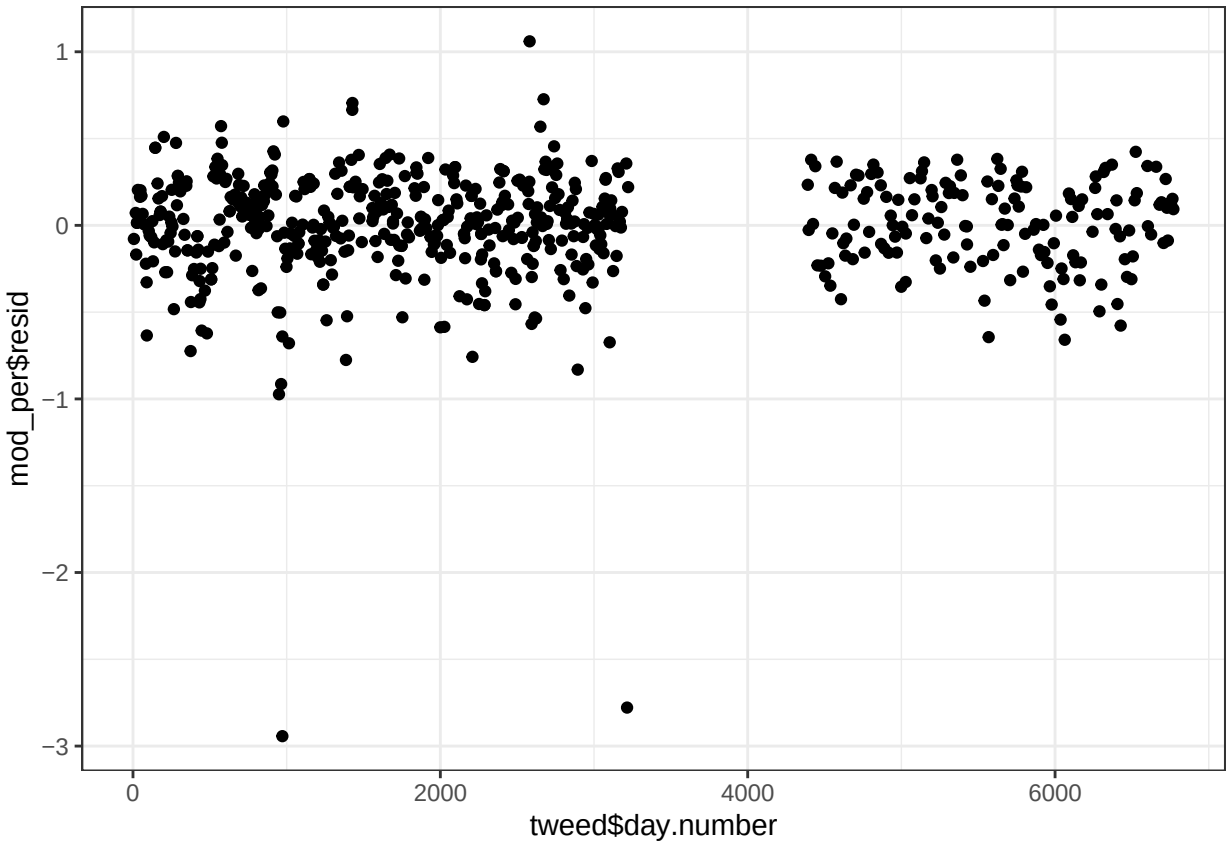



```
ggplot() +
  geom_point(aes(tweed$day.number, mod_per$resid)) +
  theme_bw()
```



This is a slight improvement on the residual trend. We can improve this even more by incorporating a degree 4 polynomial (though we should be wary of this if it's just to fit better to the fitting data - this doesn't appear to be overfitting as we were in the finance data).

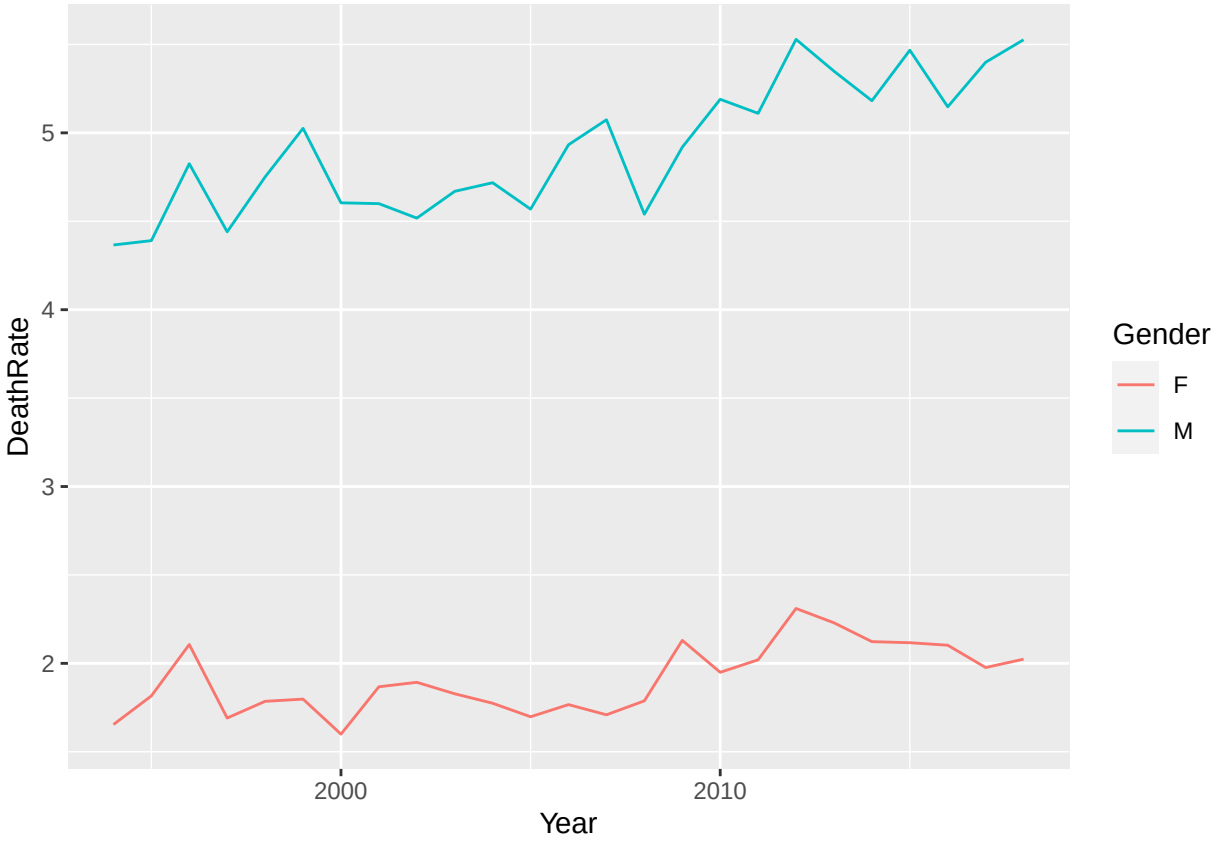
```
mod_per <- lm(logNitrate ~ sin(2*pi*day.in.year/365) +
                  cos(2*pi*day.in.year/365) + poly(day.number, 4), data=tweed)
ggplot() +
   geom_point(aes(tweed$day.number, mod_per$resid)) +
   theme_bw()
```



Example 2

(i)

```
load("Cancer.RData")

# Line plot preferably, but scatter plot will work
CancerG %>%
  ggplot()+
  geom_line(aes(Year, DeathRate, colour=Gender))
```



The trend here looks slightly different, and they are at different levels. I would expect either The full model, or the model with same slope but different intercepts to be the final selected model.

(ii)

```
# start with interaction model

mod1 <- lm(DeathRate~Gender + Year + Gender*Year, data=CancerG)

# use anova to get the f-test (can also use a t-test)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: DeathRate
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Gender       1 112.766 112.766 3414.119 < 2.2e-16 ***
## Year         1   2.269   2.269   68.682 1.119e-10 ***
## Gender:Year  1   0.418   0.418   12.661 0.0008793 ***
## Residuals   46   1.519   0.033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model with different slopes and different intercepts is most appropriate. Let's look at the model coefficients:

```
summary(mod1)
```

```
##
## Call:
## lm(formula = DeathRate ~ Gender + Year + Gender * Year, data = CancerG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45816 -0.10911 -0.02132  0.10243  0.40722
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -31.902439  10.111407  -3.155 0.002828 **
## GenderM      -47.878367  14.299689  -3.348 0.001630 **
## Year           0.016856   0.005041   3.344 0.001649 **
## GenderM:Year   0.025365   0.007128   3.558 0.000879 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1817 on 46 degrees of freedom
## Multiple R-squared:  0.987,  Adjusted R-squared:  0.9862
## F-statistic:  1165 on 3 and 46 DF,  p-value: < 2.2e-16
```

This suggests that the increase in cancer related deaths is increasing faster in males than in females. For each additional year, the growth in the rate of cancer deaths is 0.0169 in females and is nearly triple that in males at 0.042. Interpreting the intercepts doesn't really make sense here because we are extrapolating to the year 0BC - it also suggests that males have a lower rate than females, which does not match the plot and would confuse the interpretation. An option here would be to subtract either the smallest year or the median year to give this term an interpretation:

```
mod2 <- lm(DeathRate~Gender + I(Year-1994) + Gender*I(Year-1994), data=CancerG)

summary(mod2)
```

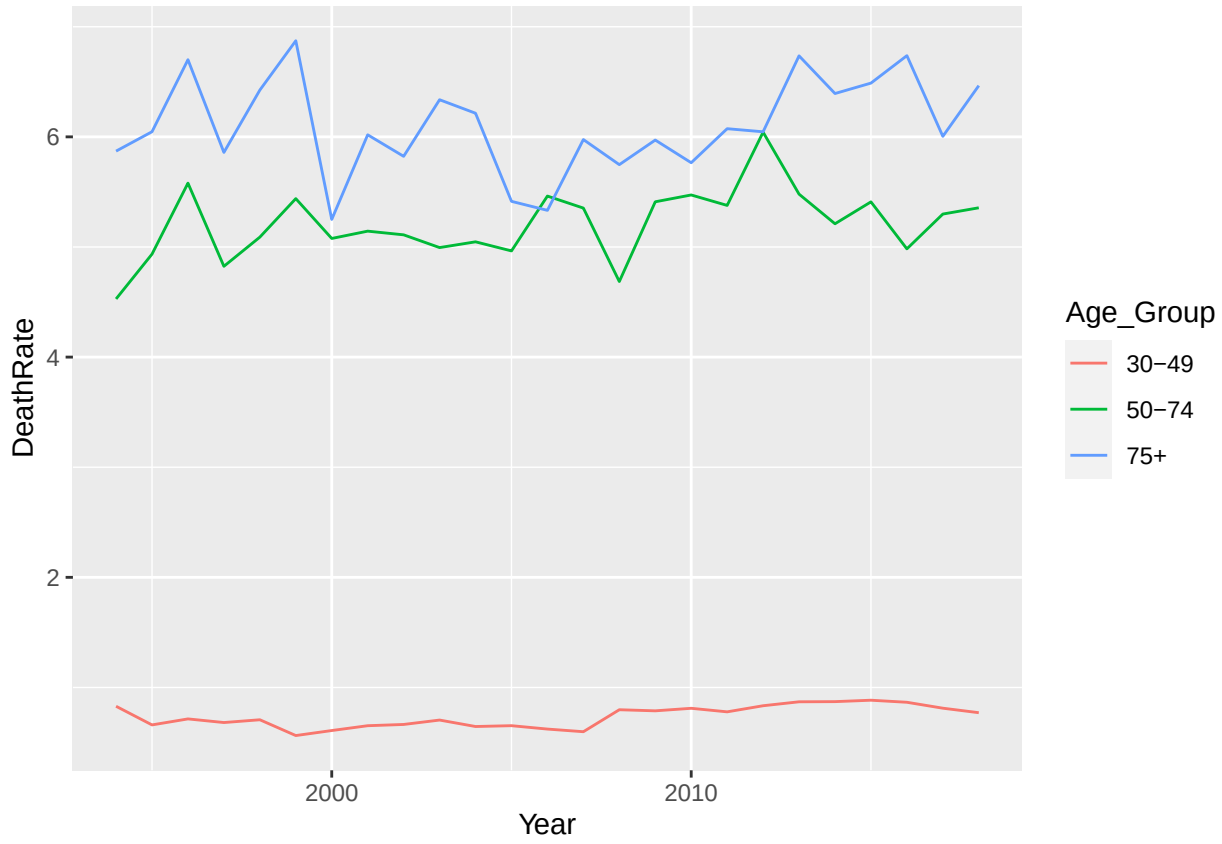
```
##
## Call:
## lm(formula = DeathRate ~ Gender + I(Year - 1994) + Gender * I(Year -
##     1994), data = CancerG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45816 -0.10911 -0.02132  0.10243  0.40722
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.707884   0.070568  24.202  < 2e-16 ***
## GenderM                 2.699165   0.099798  27.046  < 2e-16 ***
## I(Year - 1994)          0.016856   0.005041   3.344 0.001649 **
## GenderM:I(Year - 1994) 0.025365   0.007128   3.558 0.000879 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1817 on 46 degrees of freedom
## Multiple R-squared:  0.987,  Adjusted R-squared:  0.9862
## F-statistic:  1165 on 3 and 46 DF,  p-value: < 2.2e-16
```

Ive used `I()` to allow me to do a calculation within `lm()`. Now the slope coefficients are identical, but the intercept terms relate to the year 1994 (when the data began) and we can see that Males have had a higher death rate due to head and neck cancer since the beginning of the data.

(iii)

```
ggplot(CancerA) +
  geom_line(aes(Year, DeathRate, colour=Age_Group))
```



It's difficult to see here, but again I would expect a model that indicates that the intercepts are different but the trends are similar (they may be just different enough that the interaction is significant)

(iv)

```
# start with interaction model

mod1A <- lm(DeathRate~Age_Group + I(Year-1994) + Age_Group*I(Year-1994), data=CancerA)

# use anova to get the f-test (can also use a t-test)
anova(mod1A)
```

```
## Analysis of Variance Table
##
## Response: DeathRate
##                           Df Sum Sq Mean Sq   F value  Pr(>F)
## Age_Group                  2 413.63 206.814 2179.8787 < 2e-16 ***
## I(Year - 1994)             1   0.63   0.628    6.6224 0.01222 *
## Age_Group:I(Year - 1994)  2   0.07   0.034    0.3579 0.70043
## Residuals                 69   6.55   0.095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This indicates that the interaction can be removed:

```
# start with interaction model

mod2A <- lm(DeathRate~Age_Group + I(Year-1994), data=CancerA)

# use anova to get the f-test (can also use a t-test)
anova(mod2A)
```

```
## Analysis of Variance Table
##
## Response: DeathRate
##                Df Sum Sq Mean Sq   F value  Pr(>F)
## Age_Group       2 413.63 206.814 2220.0327 < 2e-16 ***
## I(Year - 1994)  1   0.63   0.628    6.7444 0.01142 *
## Residuals      71   6.61   0.093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both variables are still significant, so this suggests that the trend is similar across the three age groups, but the overall levels are different (i.e. three parallel lines)

looking at the model summary for more interpretation:

```
summary(mod2A)
```

```
##
## Call:
## lm(formula = DeathRate ~ Age_Group + I(Year - 1994), data = CancerA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77648 -0.13132 -0.01933  0.11762  0.85894
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.583139   0.084652   6.889 1.88e-09 ***
## Age_Group50-74 4.475674   0.086329  51.845  < 2e-16 ***
## Age_Group75+   5.367370   0.086329  62.174  < 2e-16 ***
## I(Year - 1994) 0.012693   0.004887   2.597   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3052 on 71 degrees of freedom
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9836
## F-statistic:  1482 on 3 and 71 DF,  p-value: < 2.2e-16
```

Within each age group there is only a small increase in the death rate due to head and neck cancer of 0.0127 per 10,000 in the population per year. Both 50-74 and 75+ have a significantly higher overall death rate compared to the 30-49 age group, though we can't say for certain whether these two groups significantly differ from each other. An easy way to check this is to change the reference level:

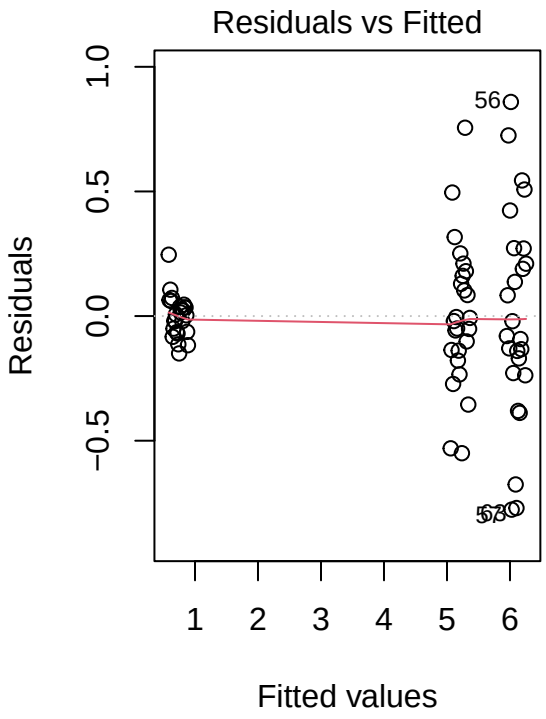
```
CancerA$Age_Group <- factor(CancerA$Age_Group, levels=unique(CancerA$Age_Group)[c(2,1,3)])

mod2A <- lm(DeathRate~Age_Group + I(Year-1994), data=CancerA)

summary(mod2A)
```

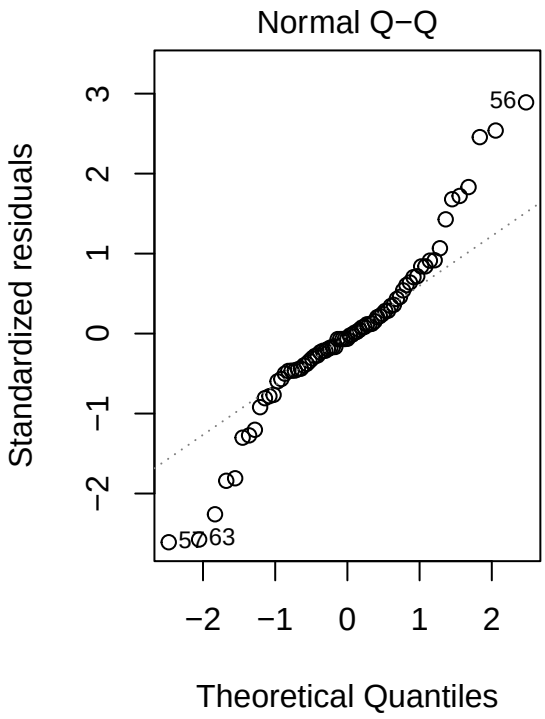
```
##
## Call:
## lm(formula = DeathRate ~ Age_Group + I(Year - 1994), data = CancerA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77648 -0.13132 -0.01933  0.11762  0.85894
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.058813   0.084652  59.760  < 2e-16 ***
## Age_Group30-49 -4.475674   0.086329 -51.845  < 2e-16 ***
## Age_Group75+     0.891696   0.086329  10.329 8.68e-16 ***
## I(Year - 1994)   0.012693   0.004887   2.597   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3052 on 71 degrees of freedom
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9836
## F-statistic:  1482 on 3 and 71 DF,  p-value: < 2.2e-16
```

From this we can see that there is a significant difference in the intercepts of the 50-75 age group and the 75+ age group.

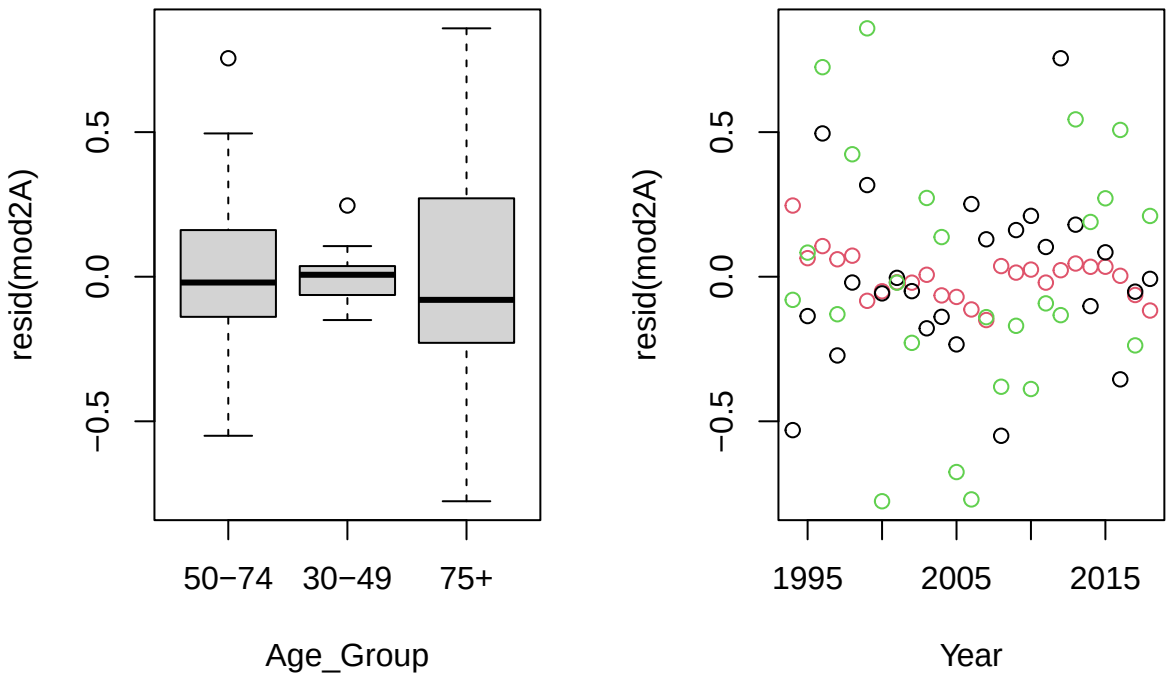
(iv)

```
par(mfrow=c(1,2))
plot(mod2A, which=c(1,2))
```






```
plot(resid(mod2A)~Age_Group, data=CancerA)

plot(resid(mod2A)~Year, data=CancerA, col=Age_Group)
```



This is a bit hidden in there, but there is some residual trend in the 75+ age group. It looks like a curve so perhaps adding in a squared term might be sensible.