Assessed Coursework 2 Solutions and full R code.

```
> Coursework2data <- read.csv("Coursework2data.csv")
> x1 <- Coursework2data$x1
> x2 <- Coursework2data$x2
> x3 <- Coursework2data$x3
> x4 <- Coursework2data$x4
> y <- Coursework2data$y
> model <- lm(y~x1+x2+x3+x4)
> summary(model)
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)
Residuals:
    Min      1Q  Median      3Q     Max
-50.762 -13.216  -1.495  11.683  55.030
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.886108   4.562664   1.071    0.286
x1           0.210453   0.044442   4.735 4.20e-06 ***
x2           0.307700   0.069538   4.425 1.60e-05 ***
x3           0.001395   0.115369   0.012    0.990
x4          10.541534   1.285194   8.202 3.12e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 19.25 on 195 degrees of freedom
Multiple R-squared:  0.5246,     Adjusted R-squared:  0.5149
F-statistic:  53.8 on 4 and 195 DF,  p-value: < 2.2e-16
```

The 4 plots shown are:

(a) Standardised Residuals versus Fitted y

(i) to check the constant variance assumption

(ii) we look for a random scatter plot. Here we seem to have a funnel shape with more variability in residuals at say fiited=40 than at fitted=100. So from this plot we have reason to doubt the constant variance assumption.

(iii) the code needed

```
> d_i = rstandard(model)
> fitted = fitted(model)
> plot(fitted, d_i, main="Standardised residuals v Fitted y")
```

(b) QQ Plot

(i) to check the Normal distribution assumption

(ii) we look for the plot to be close to the QQ line. Here there is some deviation at the two tails but not significantly so. From this plot we are likely to be content with the normal assumption.

(iii) the code needed

```
> qqnorm(d_i)
> qqline(d_i)
```

(c) Leverage

(i) to check for influential observations
(ii) we look for > 2x and >3x average leverage for high and very high leverage. Here there are at least two points we would like to investigate further.
(iii) the code needed

```
> v_i = hatvalues(model)
> i = 1:200
> plot(i, v_i, main = "Leverage values")
```

(d) Cook's Statistic

(i) a more formal check for influential observations
(ii) we look for large values > median of F(p,n-p). None of the higher leverage points exceed the median F line so we are content there are no influential observations according to Cook's Statistic.
(iii) the code needed [this one is a bit more challenging]

```
> D_i = cooks.distance(model)
> critical = rep(qf(0.5,4,196), times=200)
> plot(i, D_i, main = "Cooks Statistic and median of F", ylim
= c(0,1))
> lines(i, critical)
```