Diagnostics

May 8, 2020

```
In [9]: library(MASS)
```

```
In [1]: df <- read.table('oil.txt', col.names = c('spirit','gravity','pressure','distil','endp
```

0.1 Fit a model with the three best explanatory variables

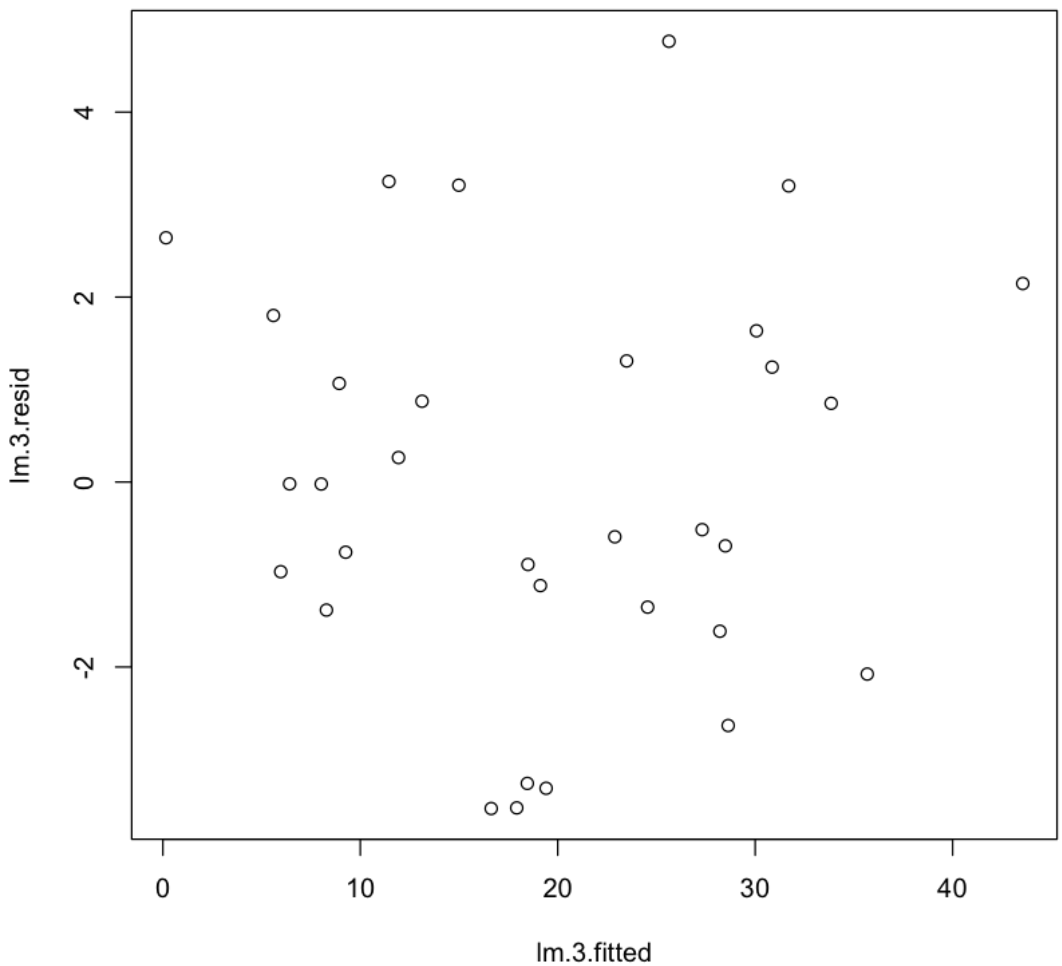
```
In [2]: lm.3 <- lm(spirit ~ gravity + distil + endpoint, data=df)

        summary(lm.3)
```

```
Call:
lm(formula = spirit ~ gravity + distil + endpoint, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5303 -1.3606 -0.2681  1.3911  4.7658

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032034   7.223341   0.558   0.5811
gravity      0.221727   0.102061   2.173   0.0384 *
distil      -0.186571   0.015922 -11.718 2.61e-12 ***
endpoint     0.156527   0.006462  24.224  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.283 on 28 degrees of freedom
Multiple R-squared:  0.959,Adjusted R-squared:  0.9546
F-statistic: 218.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

0.2 Manually extracting the fitted and residual values, and then plotting them

```
In [3]: lm.3.fitted <- fitted(lm.3)
        lm.3.resid <- residuals(lm.3)
```

```
In [4]: plot(lm.3.fitted, lm.3.resid)
```



0.3 Or using R's built in plot function with the model as the argument

0.4 Residual plots

1) Plot of standardised residuals Vs fitted values tests the assumption that the random error terms are mean zero and of constant variance

- e_i' against $\hat{y}$
- If there is a trend to the data, rather than being randomly dispersed around mean zero, then this indicates that the constant variance assumption does not hold in the model

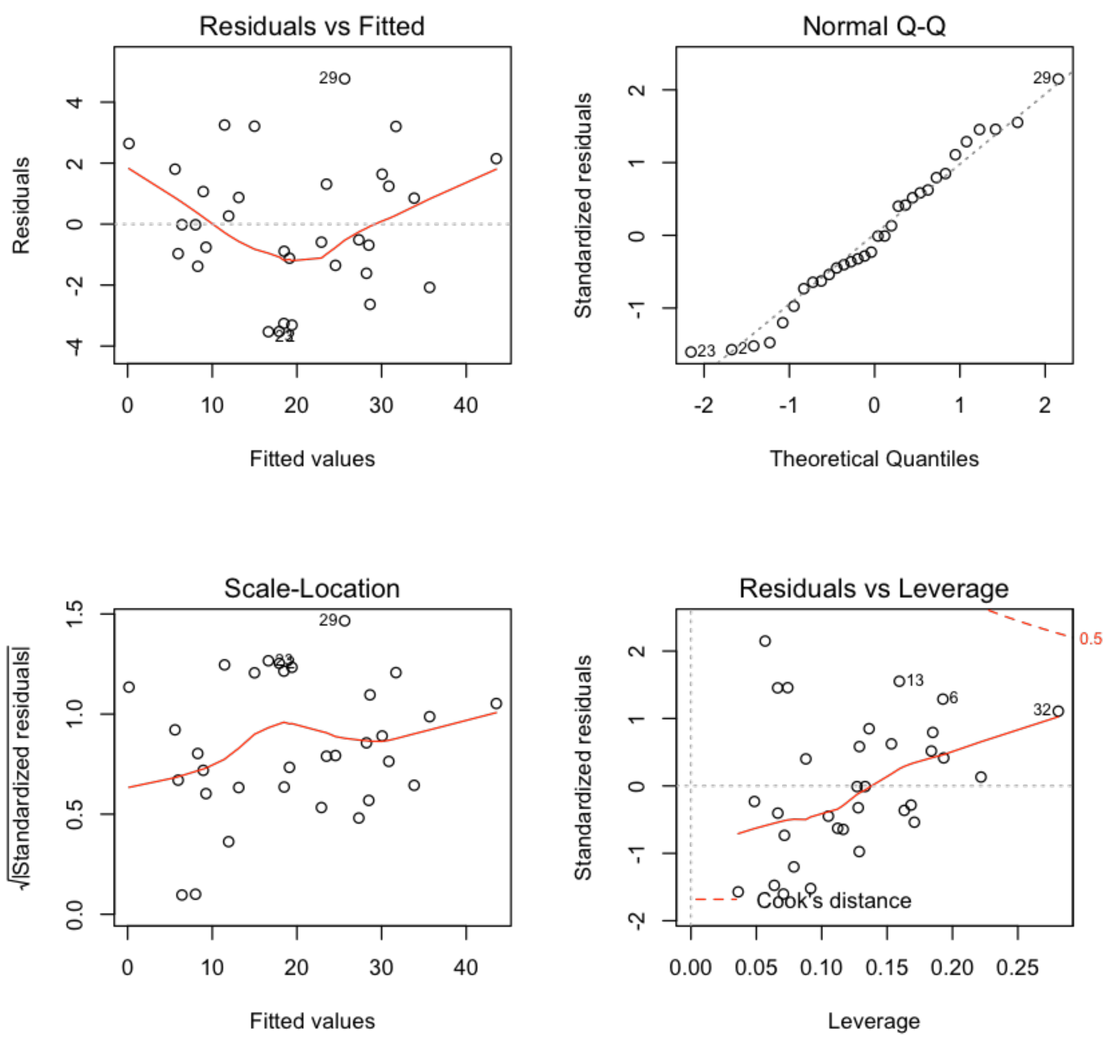
2) Plot the standardised residuals (which, as the name suggests, is approximately a standard normal distribution) Vs the theoretical quantiles of a standard normal

- This tests the assumption of normality of the random error term in the model (which also applies to the assumption of normality in the dependant variable, y.

3) Plot the root of the standardised residuals against the fitted values

- Line is a running average curve
- Shouldn't be a trend with the average line, otherwise this is problematic for our constant variance assumption

In [5]:
```
par(mfrow = c(2,2))
plot(lm.3)
```



0.5 Model adequecy

Plot the standardised residuals against **EACH EXPLANATORY VARIABLE**

Expect mean zero, random dispersion

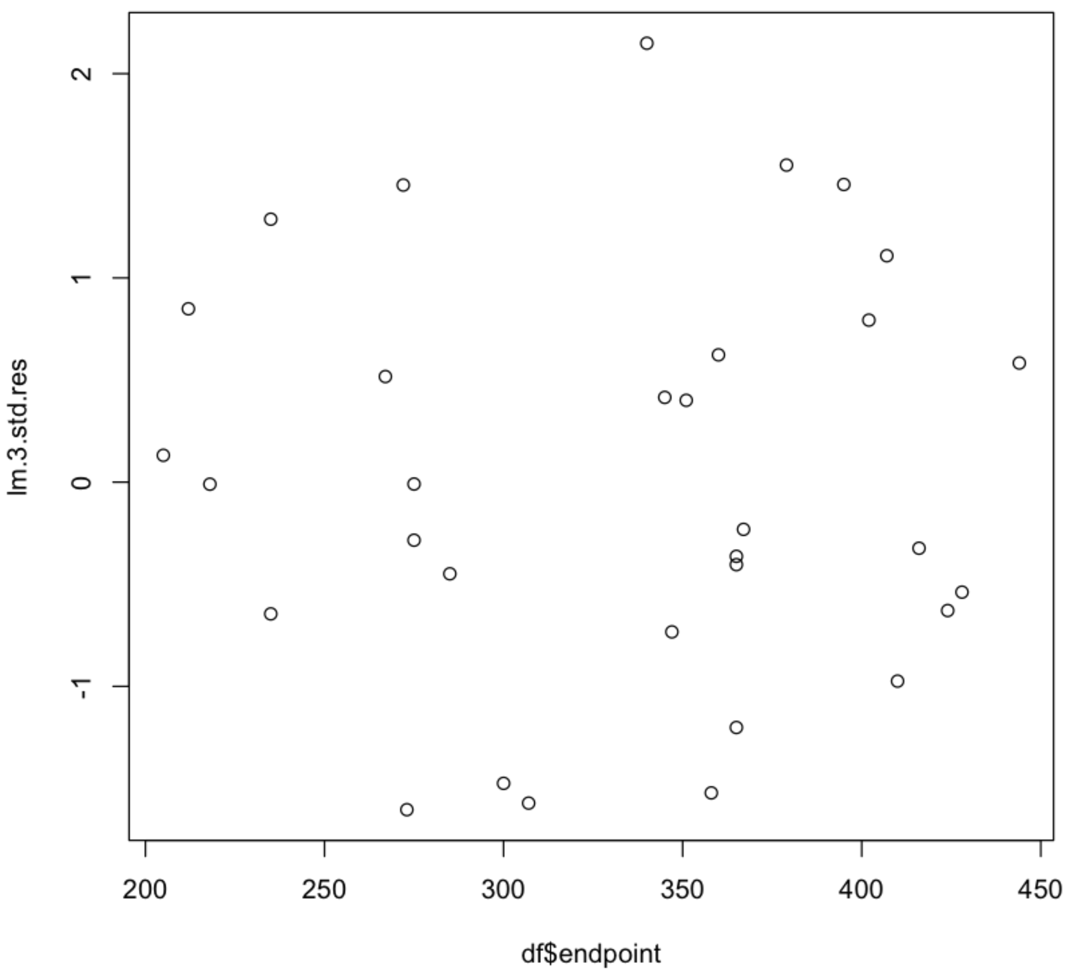
If there's trend:

- And the variable is already in the model, indicates a higher order variable term is needed (i.e. x^2)
- If the variable is **NOT** already in the model, then it indicates it should be.

```
In [11]: # can easily get the model residuals from:
         lm.3.std.res <- stdres(lm.3)
```

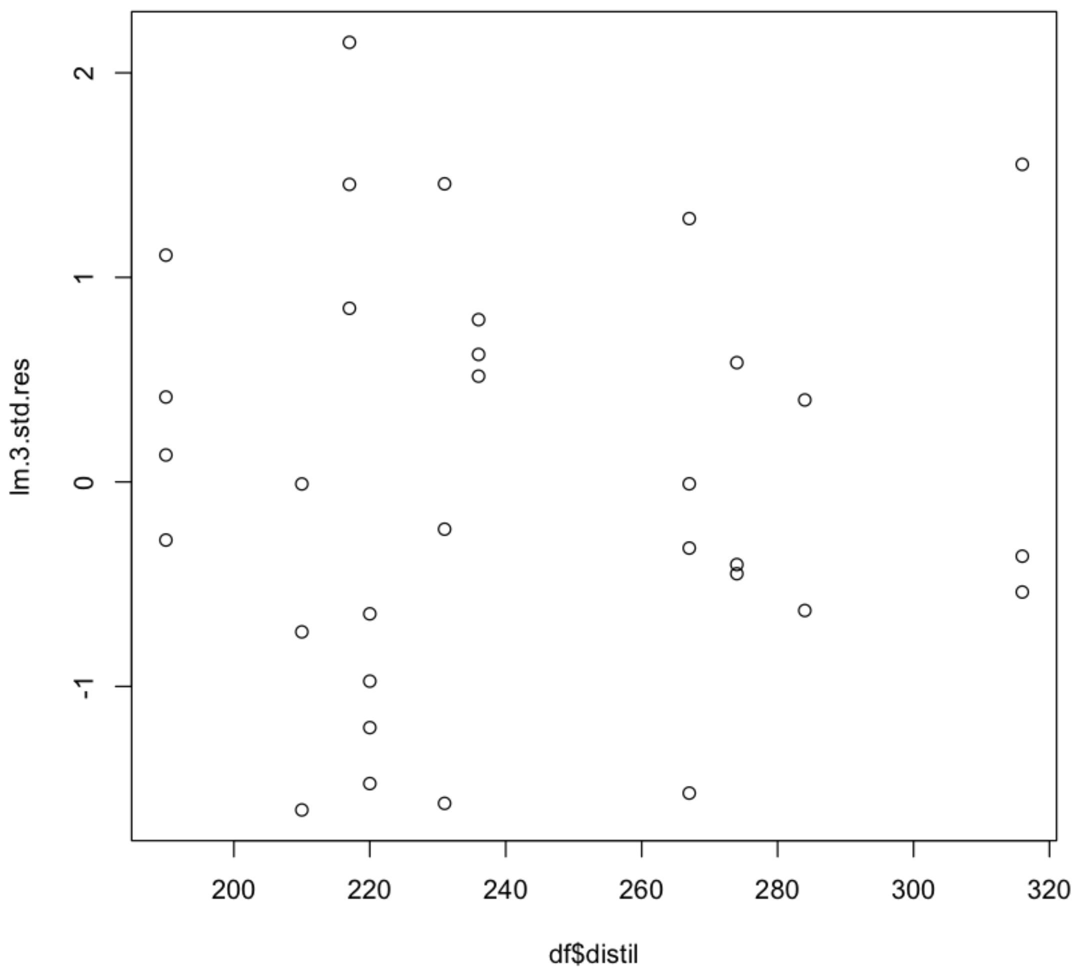
0.6 Plotting standardised residuals from the MASS library against endpoint

```
In [13]: plot(df$endpoint, lm.3.std.res)
```



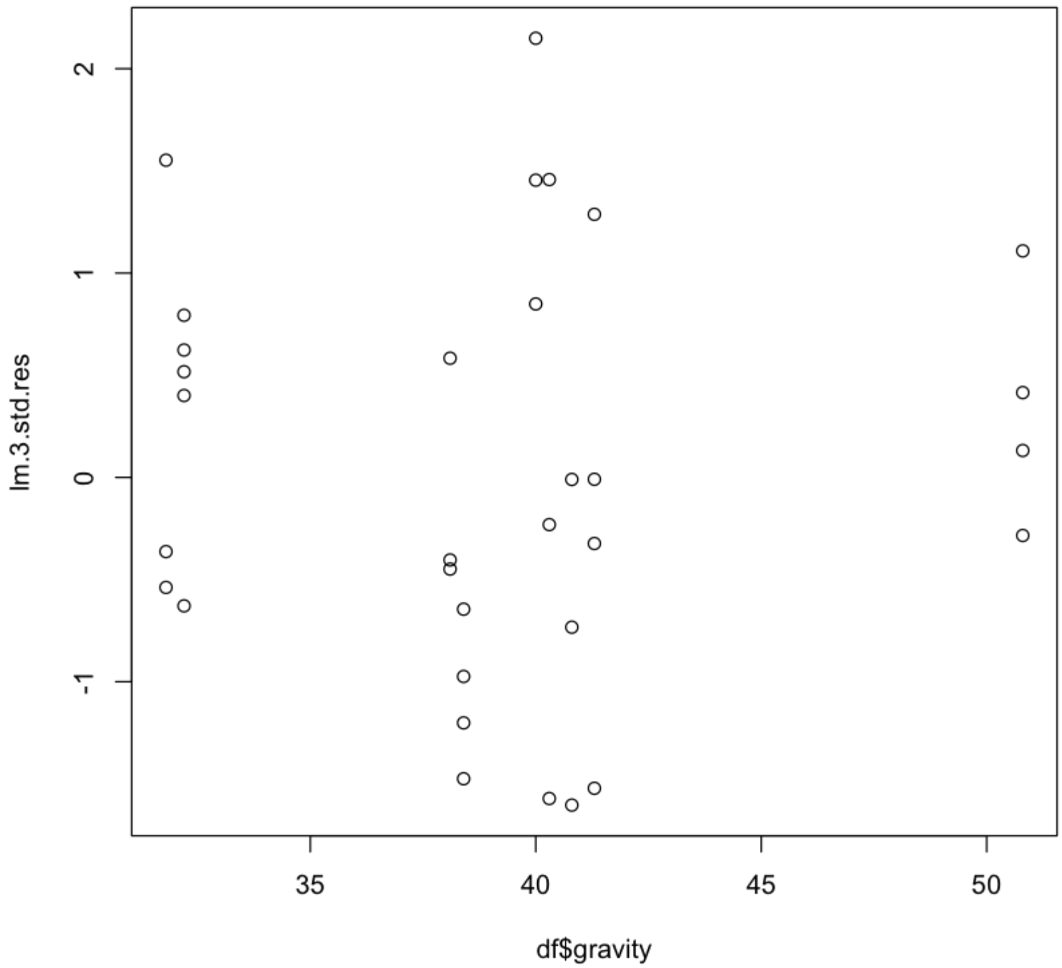
0.7 Plotting standardised residuals from the MASS library against distil

```
In [14]: plot(df$distil, lm.3.std.res)
```



0.8 Plotting standardised residuals from the MASS library against gravity

```
In [15]: plot(df$gravity, lm.3.std.res)
```

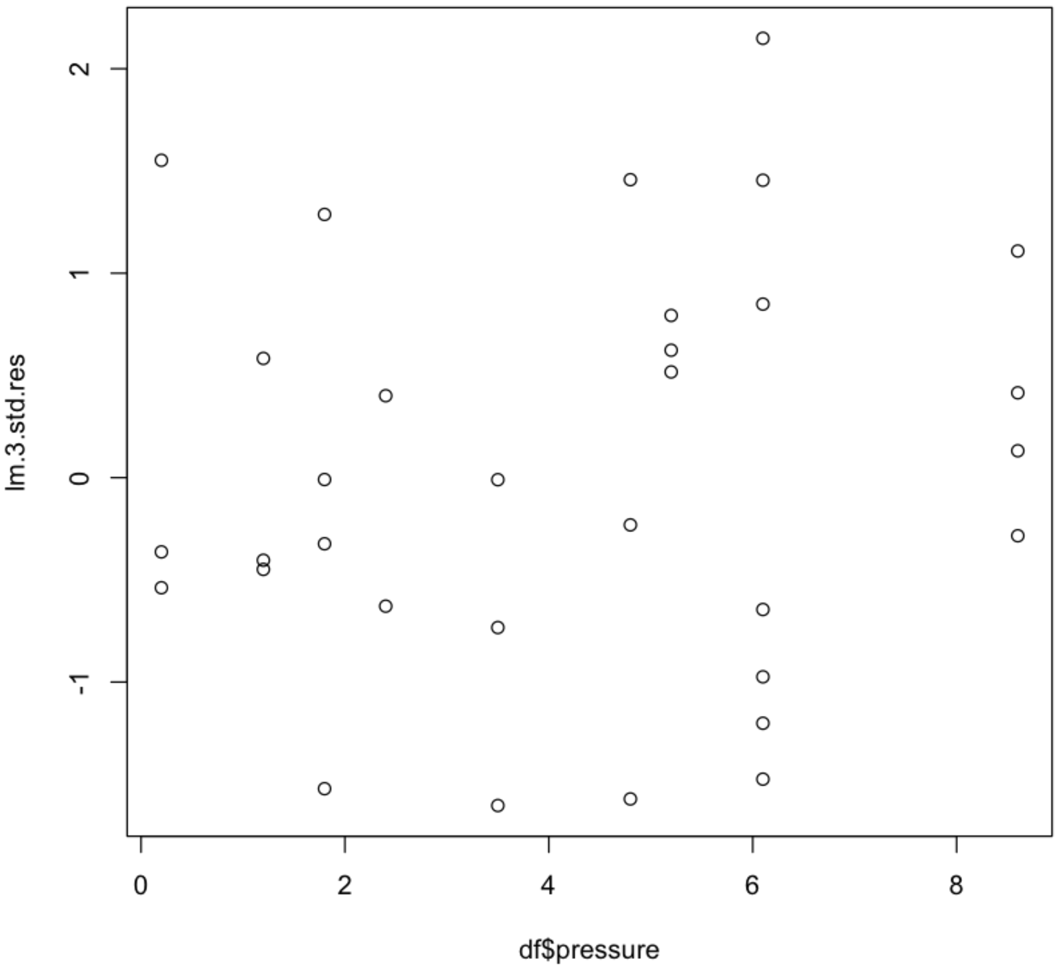


0.9 Plotting standardised residuals from the MASS library against pressure, which is NOT in the model.

If there was a significant trend here, we'd be tempted to add **pressure** back into the model, despite it not appearing as significant in our goodness-of-fit tests

Looks random, so we can be happy leaving it out.

```
In [16]: plot(df$pressure, lm.3.std.res)
```



1 Let's fit a model without the main predictor variable, endpoint, then plot the residuals of that model against endpoint to see if we see a trend.

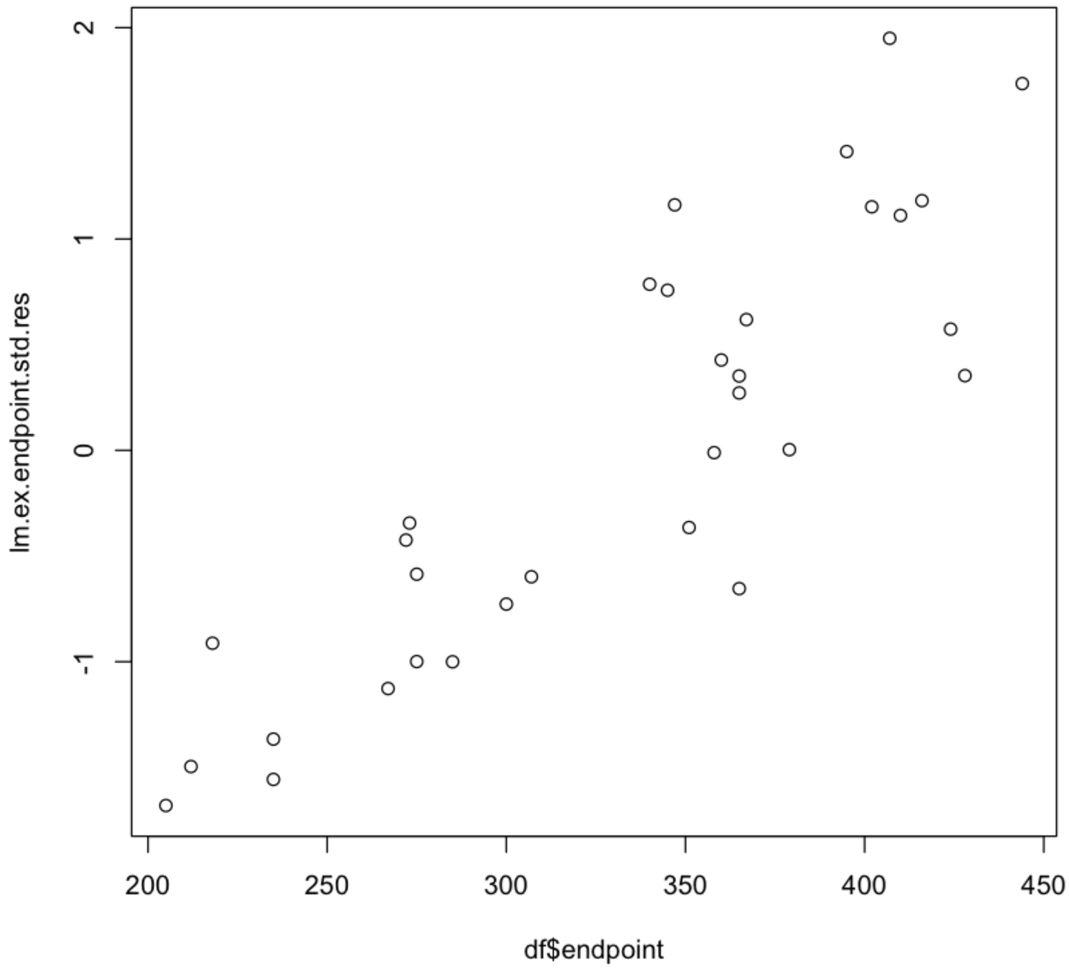
Note it doesn't pass the null hypothesis test so we wouldn't normally take this any further...

```
In [17]: lm.ex.endpoint <- lm(spirit ~ gravity + pressure + distil, data=df)

         summary(lm.ex.endpoint)
```

```
Call:
lm(formula = spirit ~ gravity + pressure + distil, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.5787  -7.5048  -0.0363   7.2252  17.9925 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.01312   46.95654  -0.235    0.816
gravity       0.12505    0.46321   0.270    0.789
pressure      2.27819    1.68264   1.354    0.187
distil        0.06724    0.12896   0.521    0.606

Residual standard error: 10.37 on 28 degrees of freedom
Multiple R-squared:  0.1558,Adjusted R-squared:  0.06536 
F-statistic: 1.723 on 3 and 28 DF,  p-value: 0.1851
```

```
In [18]: lm.ex.endpoint.std.res <- stdres(lm.ex.endpoint)
```

```
In [19]: plot(df$endpoint, lm.ex.endpoint.std.res)
```



1.1 We see clear, clear trend that endpoint should be added to the model!

```
In [ ]:
```