Linear Models, Problem 2

alexaoh

1 oktober, 2021

Simulated Data for Eight Regression Lines

```
data <- read.csv2("REG8.csv")
head(data)
```

```
#>   REG X        Y
#> 1   1 0 17.43409
#> 2   1 1 17.05458
#> 3   1 2 17.41801
#> 4   1 3 17.40713
#> 5   1 4 17.83299
#> 6   1 5 18.59063
```

```
(d <- dim(data))
```

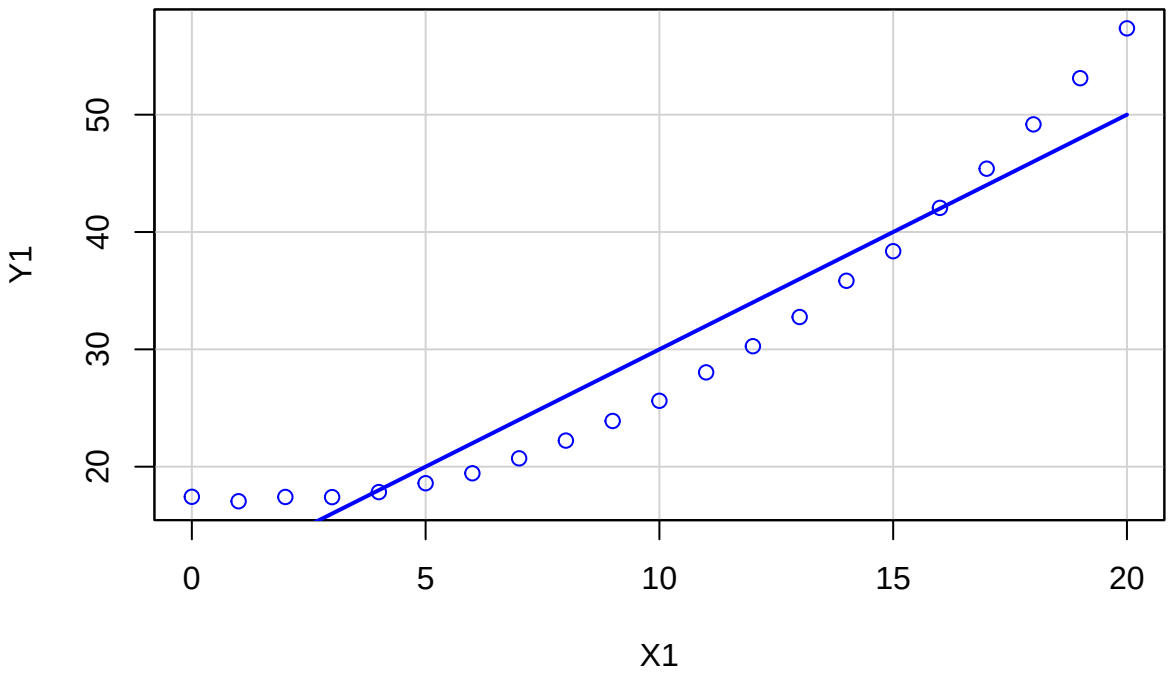
```
#> [1] 168   3
```

The task is to compute all 8 regression lines and compare them.

Regression Line 1

```
X1 <- data[data$REG == 1, "X"]
Y1 <- data[data$REG == 1, "Y"]

scatterplot(X1, Y1, smooth = F, boxplots = F)
```




```
lm1 <- lm(Y1~X1)
summary(lm1)
```

```
#>
#> Call:
#> lm(formula = Y1 ~ X1)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -4.384 -3.289 -1.409  3.176  7.434
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X1            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty high (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

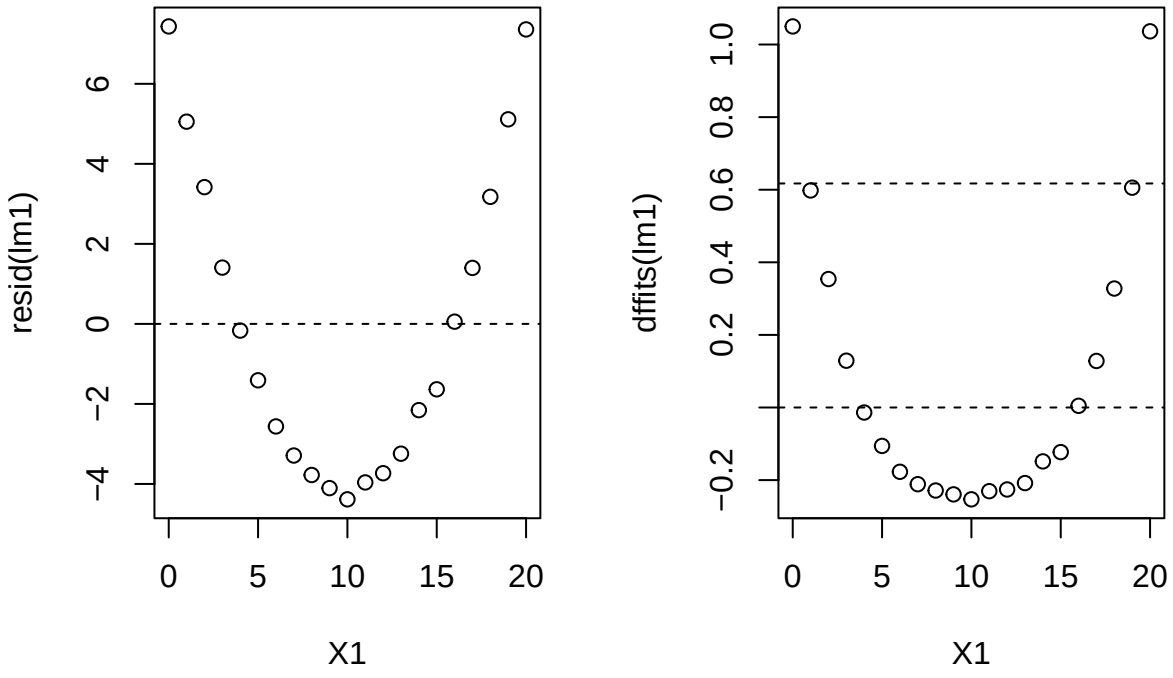
```
anova(lm1)
```

```
#> Analysis of Variance Table
#>
#> Response: Y1
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X1         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

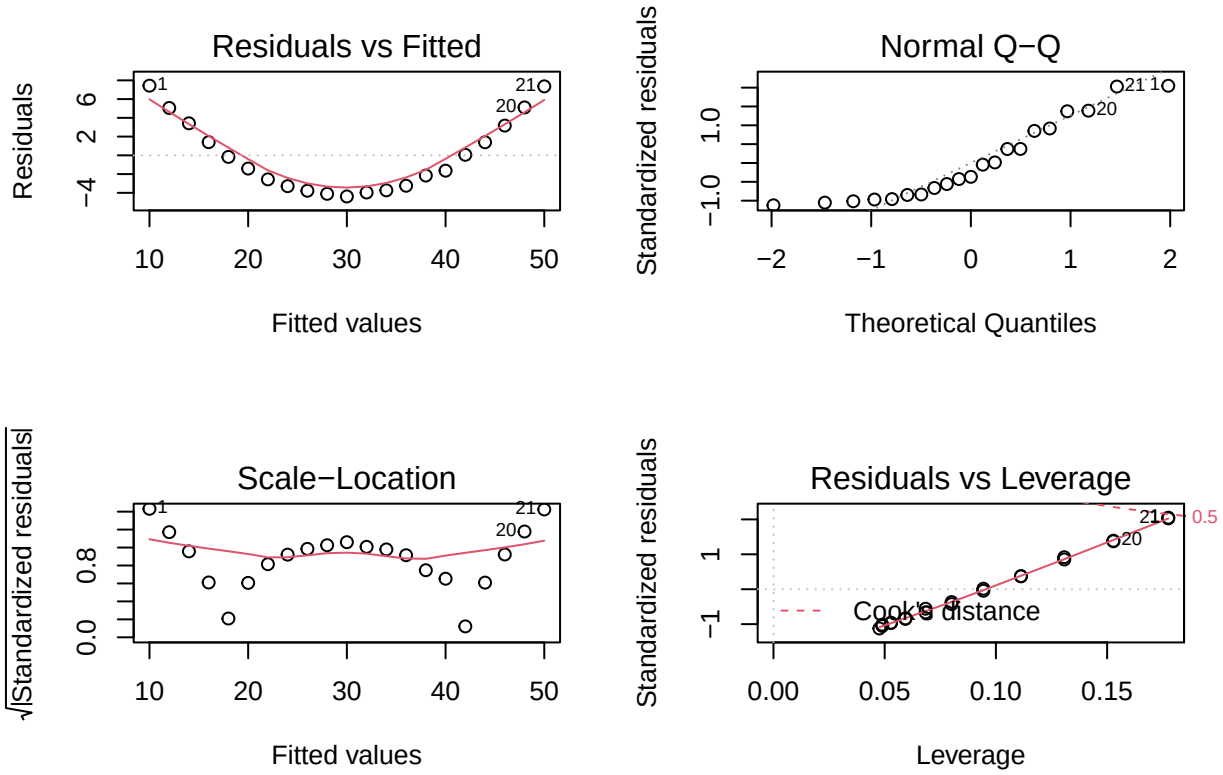
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X1,resid(lm1)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X1,dffits(lm1))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

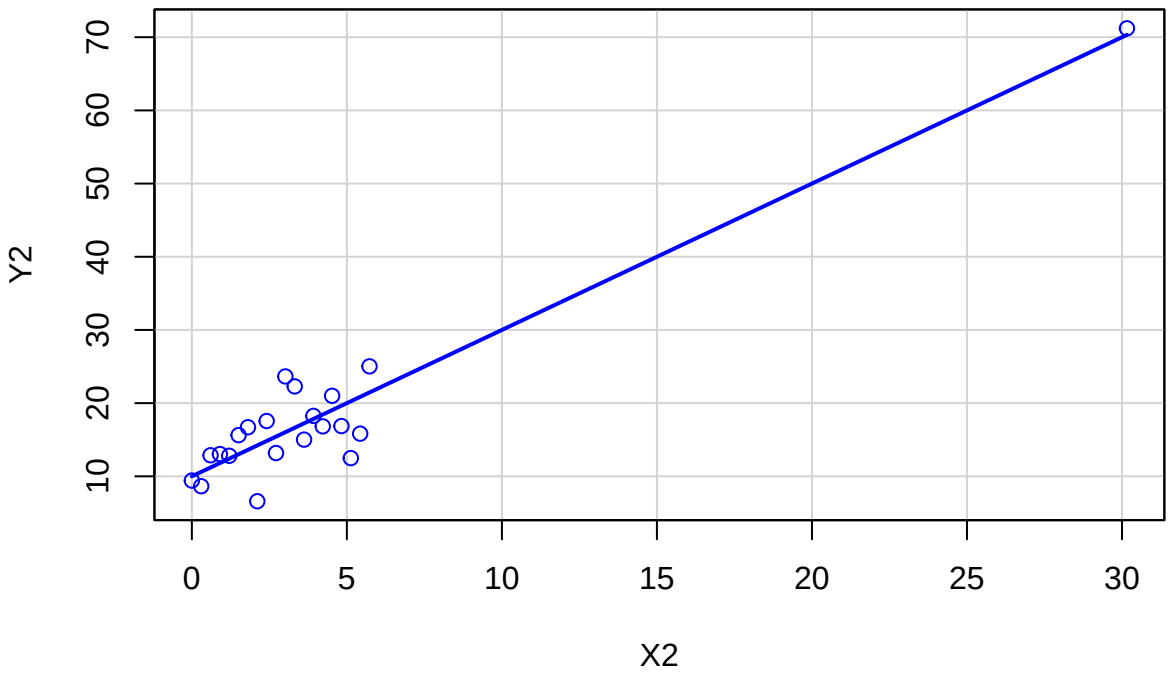
```
par(mfrow=c(2,2))
plot(lm1)
```



We see that the residuals have a pattern (quadratic), which shows that the homoscedasticity assumption (most likely) does not hold. Moreover, the QQ-plot shows that the standardized residuals do not resemble the quantiles of the normal distribution, which shows that the normality assumption of the errors (most likely) does not hold. Also, we see a pattern in the Scale-location plot, which is not a good sign for our linear model fit.

Regression Line 2

```
X2 <- data[data$REG == 2, "X"]
Y2 <- data[data$REG == 2, "Y"]
scatterplot(X2, Y2, smooth = F, boxplots = F)
```




```
lm2 <- lm(Y2~X2)
summary(lm2)
```

```
#>
#> Call:
#> lm(formula = Y2 ~ X2)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.7659 -2.2250  0.4169  2.6096  7.6157
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.0594   9.439 1.32e-08 ***
#> X2            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty high (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

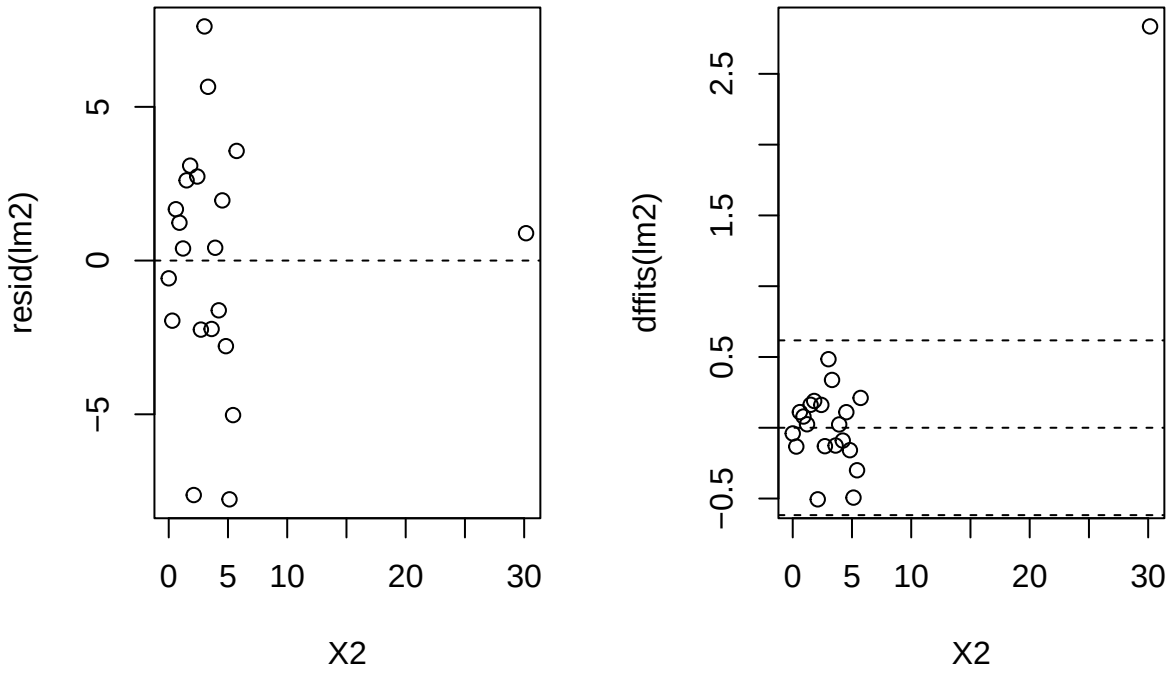
```
anova(lm2)
```

```
#> Analysis of Variance Table
#>
#> Response: Y2
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X2         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

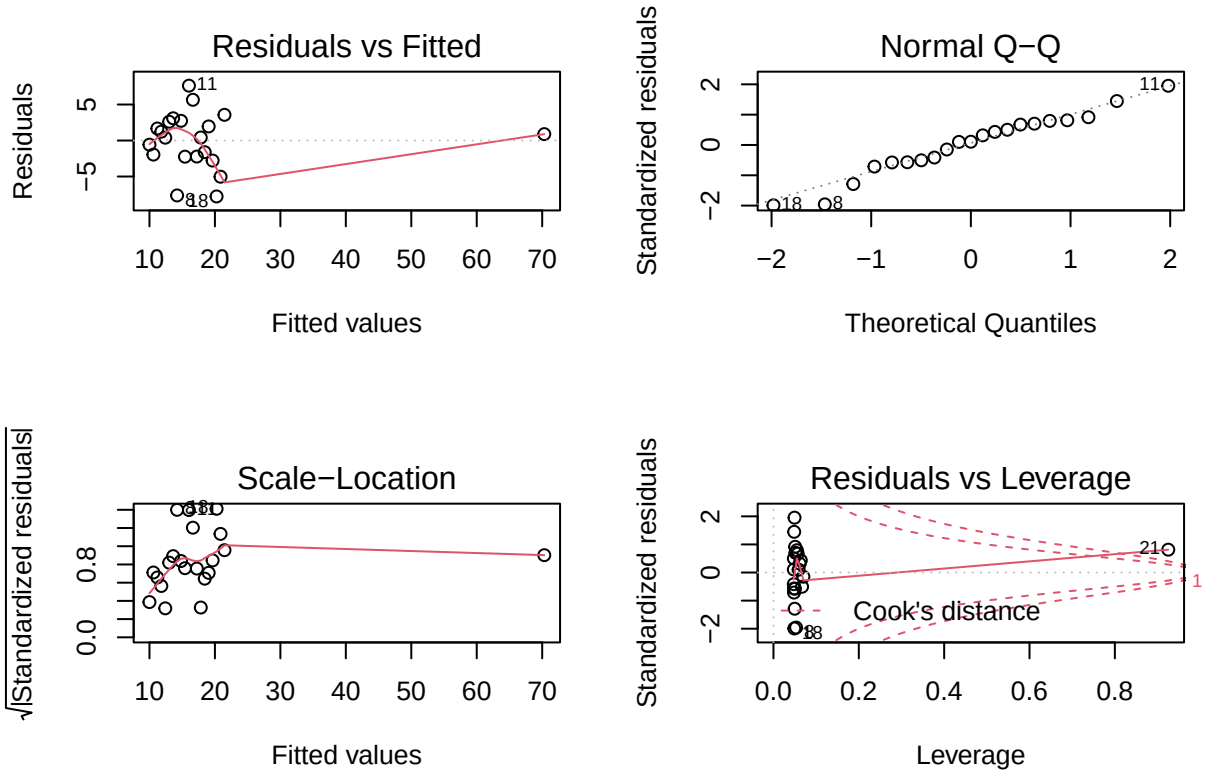
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X2,resid(lm2)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X2,dffits(lm2))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

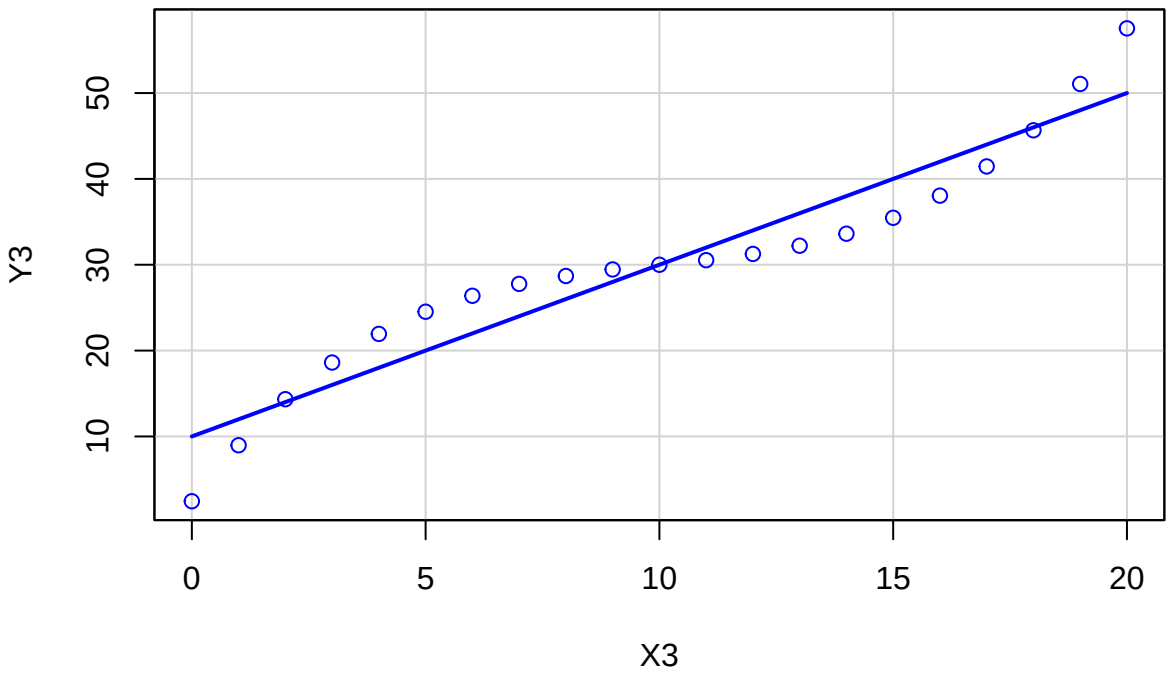
```
par(mfrow=c(2,2))
plot(lm2)
```



Looking past the outlier, we see that the residuals have a pattern (quadratic), which shows that the homoscedasticity assumption (most likely) does not hold. Moreover, the QQ-plot shows that the standardized residuals do not resemble the quantiles of the normal distribution, which shows that the normality assumption of the errors (most likely) does not hold. Also, we see a pattern in the Scale-location plot (could be approximately linearly increasing for all points except the outliers), which is not a good sign for our linear model fit. These patterns can be seen more closely by removing the outlier, but then the values of the estimations and the F-test change.

Regression Line 3

```
X3 <- data[data$REG == 3, "X"]
Y3 <- data[data$REG == 3, "Y"]
scatterplot(X3, Y3, smooth = F, boxplots = F)
```




```
lm3 <- lm(Y3~X3)
summary(lm3)
```

```
#>
#> Call:
#> lm(formula = Y3 ~ X3)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.5485 -3.0263 -0.0003  3.0596  7.5382
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X3            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

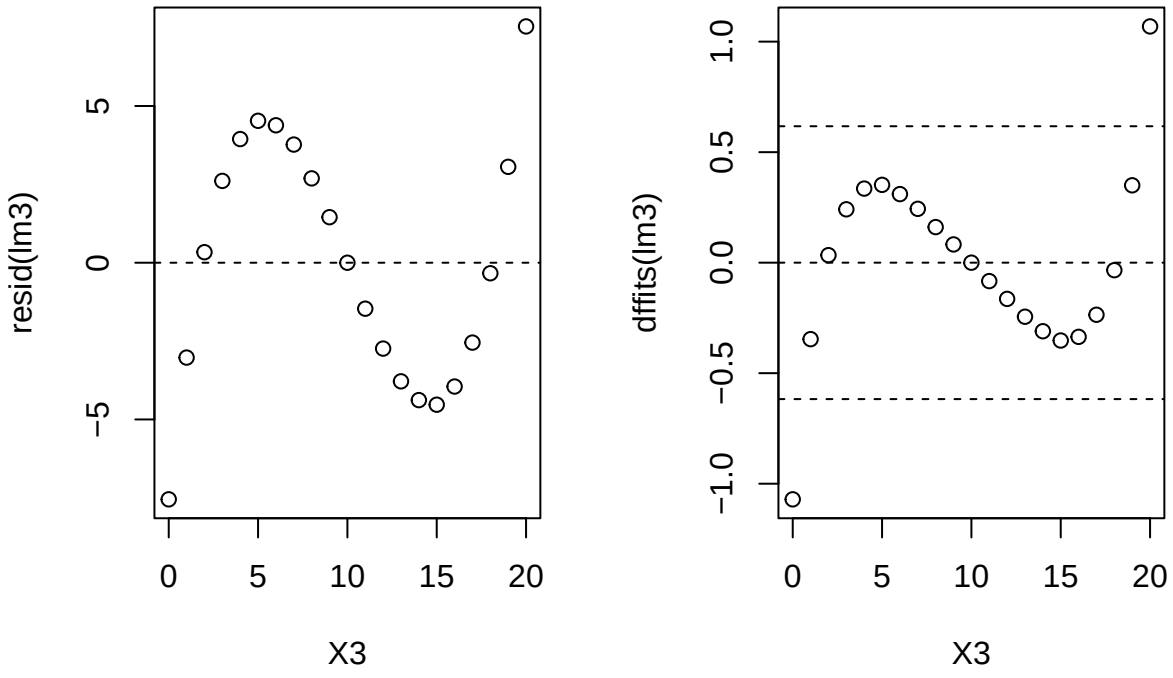
```
anova(lm3)
```

```
#> Analysis of Variance Table
#>
#> Response: Y3
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X3         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

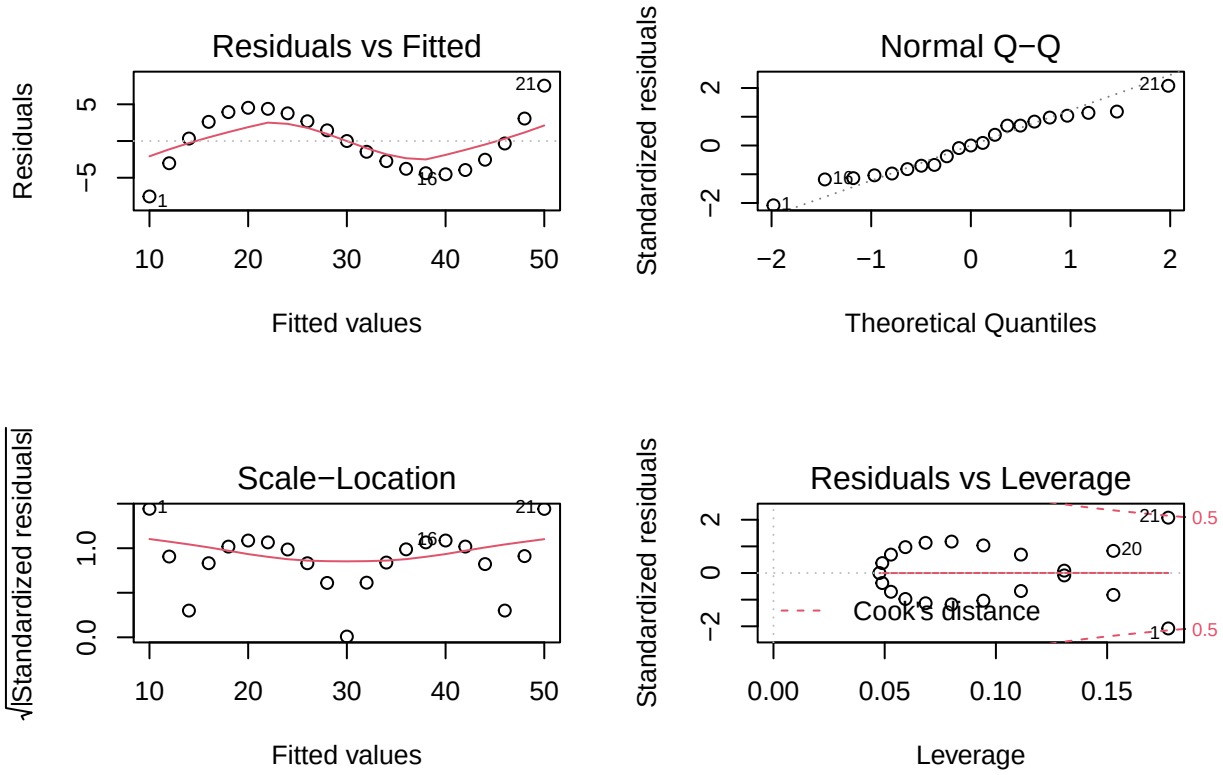
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X3,resid(lm3)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X3,dffits(lm3))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

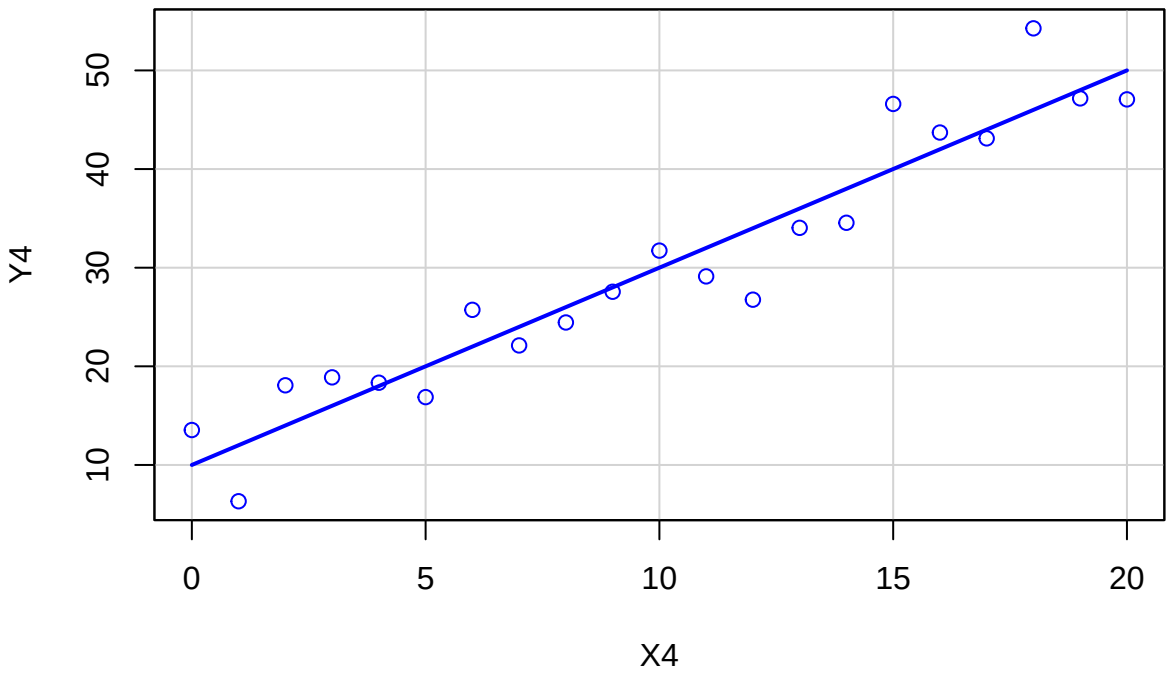
```
par(mfrow=c(2,2))
plot(lm3)
```



We see that the residuals have a pattern (sinusoidal), which shows that the homoscedasticity assumption (most likely) does not hold. Moreover, the QQ-plot shows that the standardized residuals do not resemble the quantiles of the normal distribution, which shows that the normality assumption of the errors (most likely) does not hold. Also, we see a pattern in the Scale-location plot (could be approximately quadratic, according to the red line), which is not a good sign for our linear model fit.

Regression Line 4

```
X4 <- data[data$REG == 4, "X"]
Y4 <- data[data$REG == 4, "Y"]
scatterplot(X4, Y4, smooth = F, boxplots = F)
```




```
lm4 <- lm(Y4~X4)
summary(lm4)
```

```
#>
#> Call:
#> lm(formula = Y4 ~ X4)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.2433 -2.8824 -0.8368  2.8820  8.2657
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X4            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

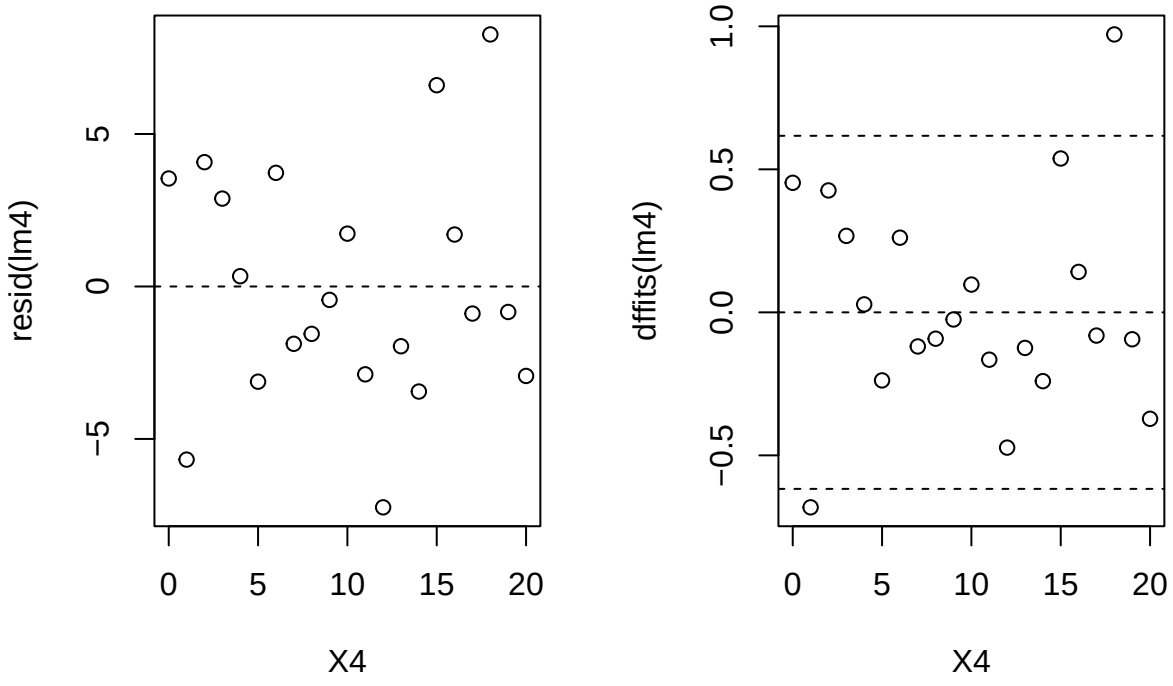
```
anova(lm4)
```

```
#> Analysis of Variance Table
#>
#> Response: Y4
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X4         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

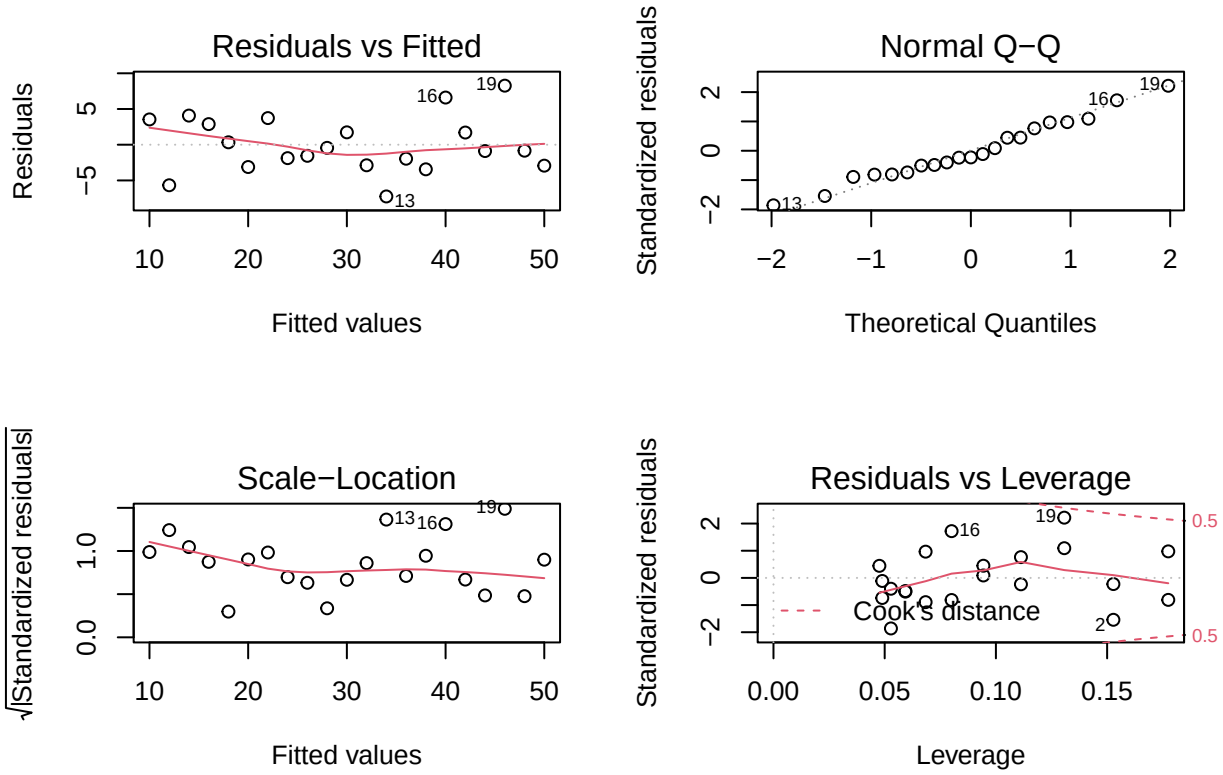
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X4,resid(lm4)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X4,dffits(lm4))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

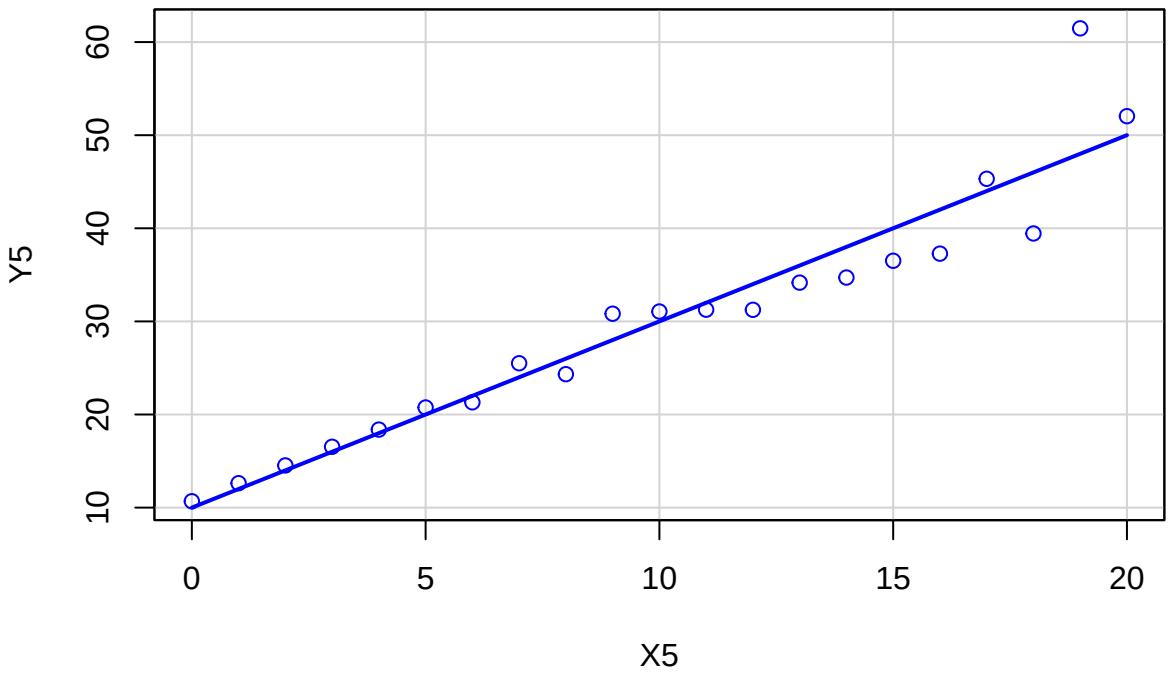
```
par(mfrow=c(2,2))
plot(lm4)
```



Now the pattern is less then in the other cases, but one can still argue that that the homoscedasticity assumption (most likely) could not hold. Moreover, the QQ-plot shows that the standardized residualsresemble the quantiles of the normal distribution more than earlier, but still one could conclude that the normality assumption of the errors (most likely) does not hold. The pattern in the Scale-location plot is not as pronounced as in earlier cases.

Regression Line 5

```
X5 <- data[data$REG == 5, "X"]
Y5 <- data[data$REG == 5, "Y"]
scatterplot(X5, Y5, smooth = F, boxplots = F)
```




```
lm5 <- lm(Y5~X5)
summary(lm5)
```

```
#>
#> Call:
#> lm(formula = Y5 ~ X5)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -6.5558 -1.8347  0.5321  1.0613 13.4747
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X5            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

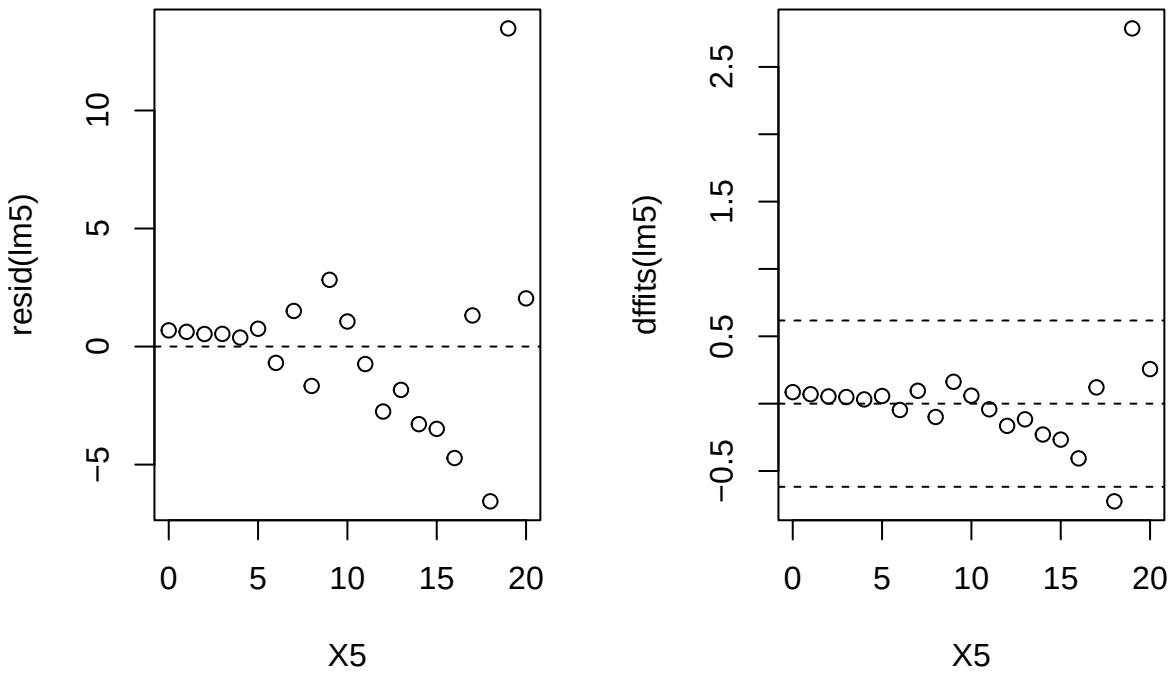
```
anova(lm5)
```

```
#> Analysis of Variance Table
#>
#> Response: Y5
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X5         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

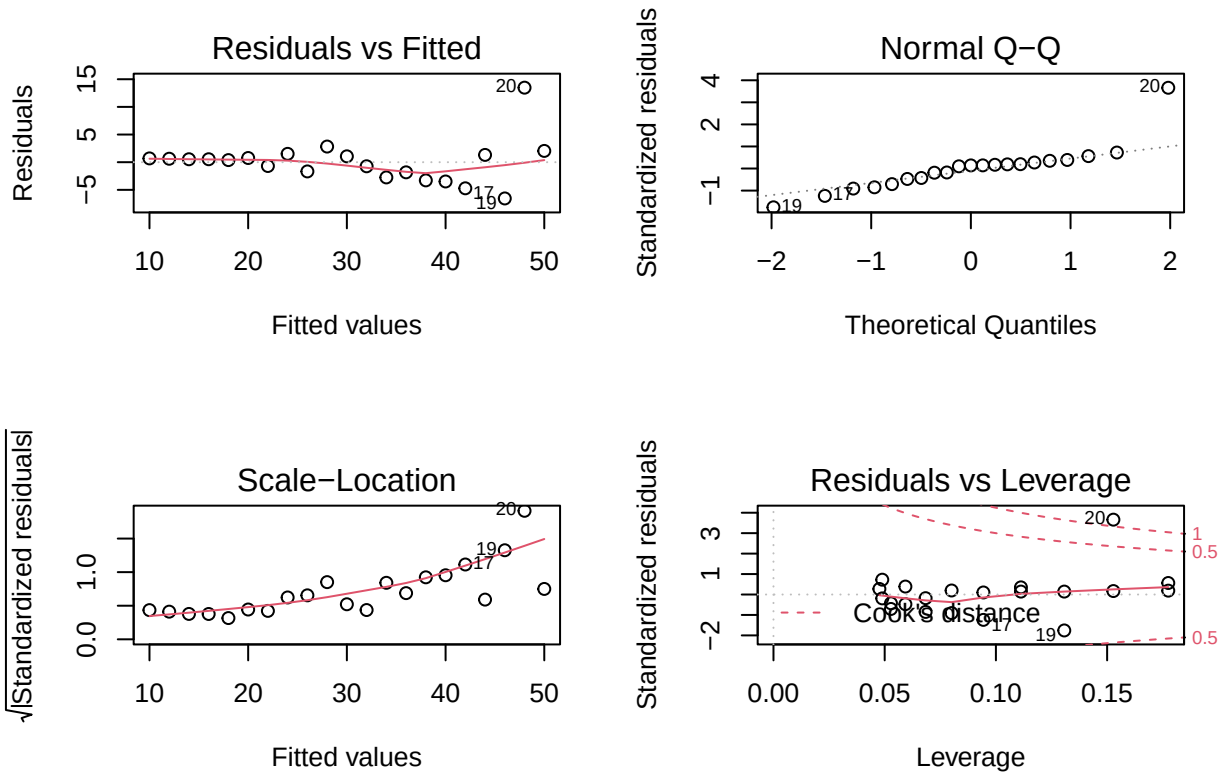
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X5,resid(lm5)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X5,dffits(lm5))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

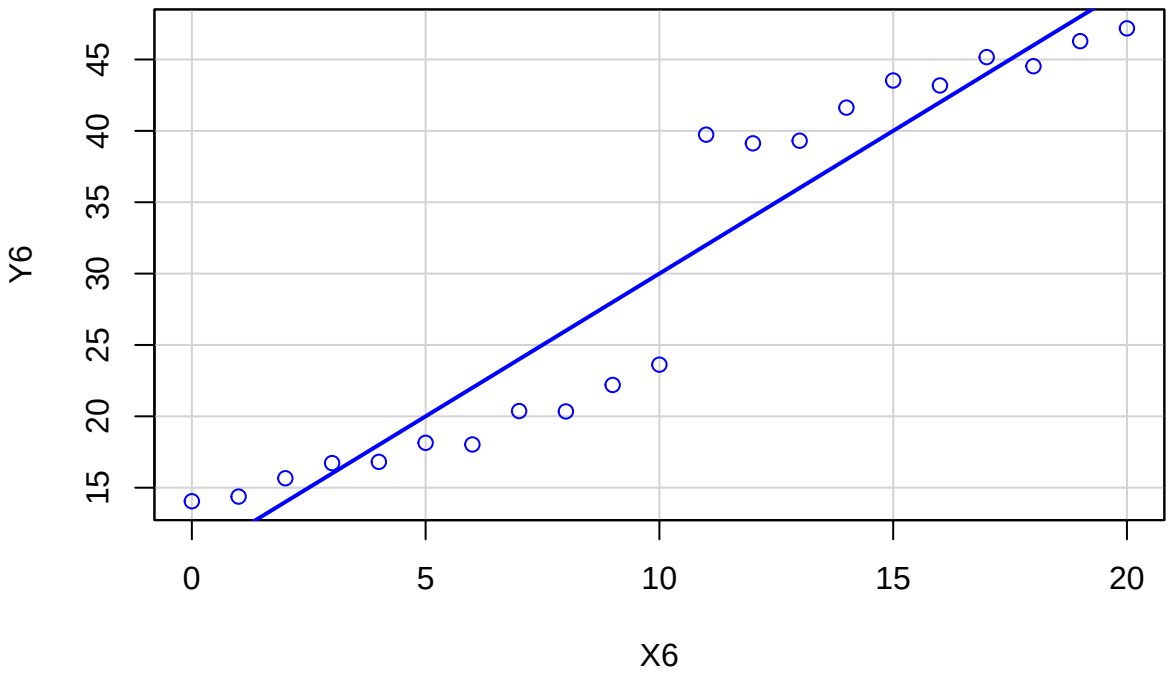
```
par(mfrow=c(2,2))
plot(lm5)
```



The patterns keep changing depending on the data, even though the estimates and the regression line stay the same.

Regression Line 6

```
X6 <- data[data$REG == 6, "X"]
Y6 <- data[data$REG == 6, "Y"]
scatterplot(X6, Y6, smooth = F, boxplots = F)
```




```
lm6 <- lm(Y6~X6)
summary(lm6)
```

```
#>
#> Call:
#> lm(formula = Y6 ~ X6)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -6.3797 -2.8178  0.7244  3.3092  7.7345
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X6            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

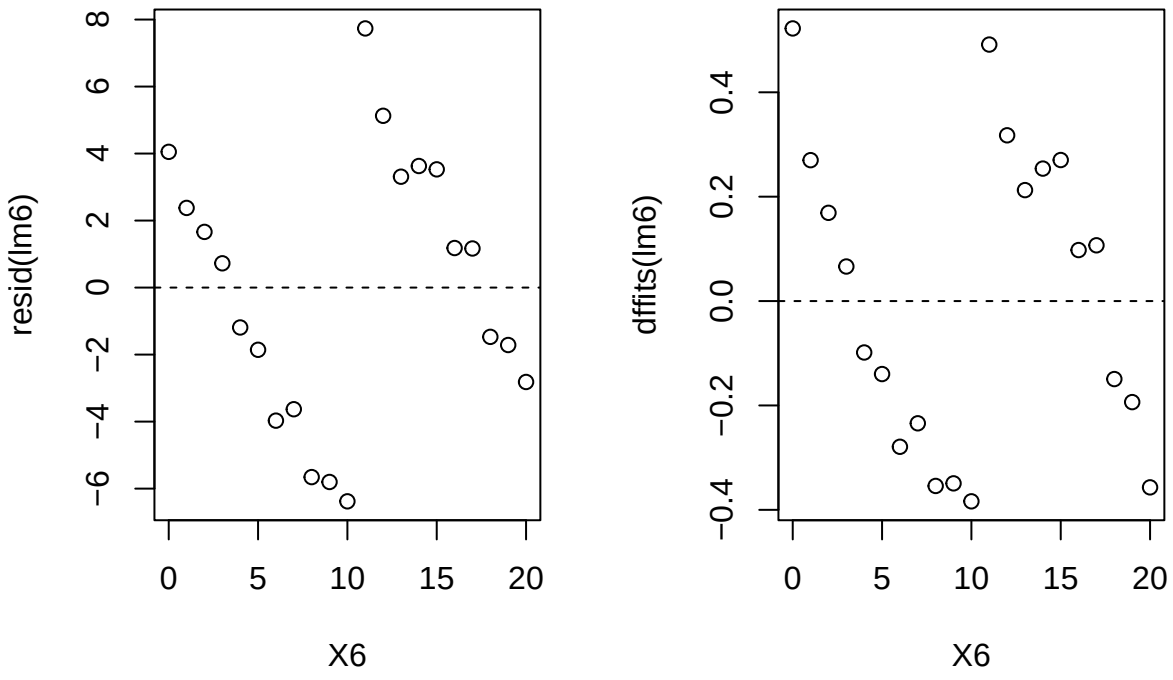
```
anova(lm6)
```

```
#> Analysis of Variance Table
#>
#> Response: Y6
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X6         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

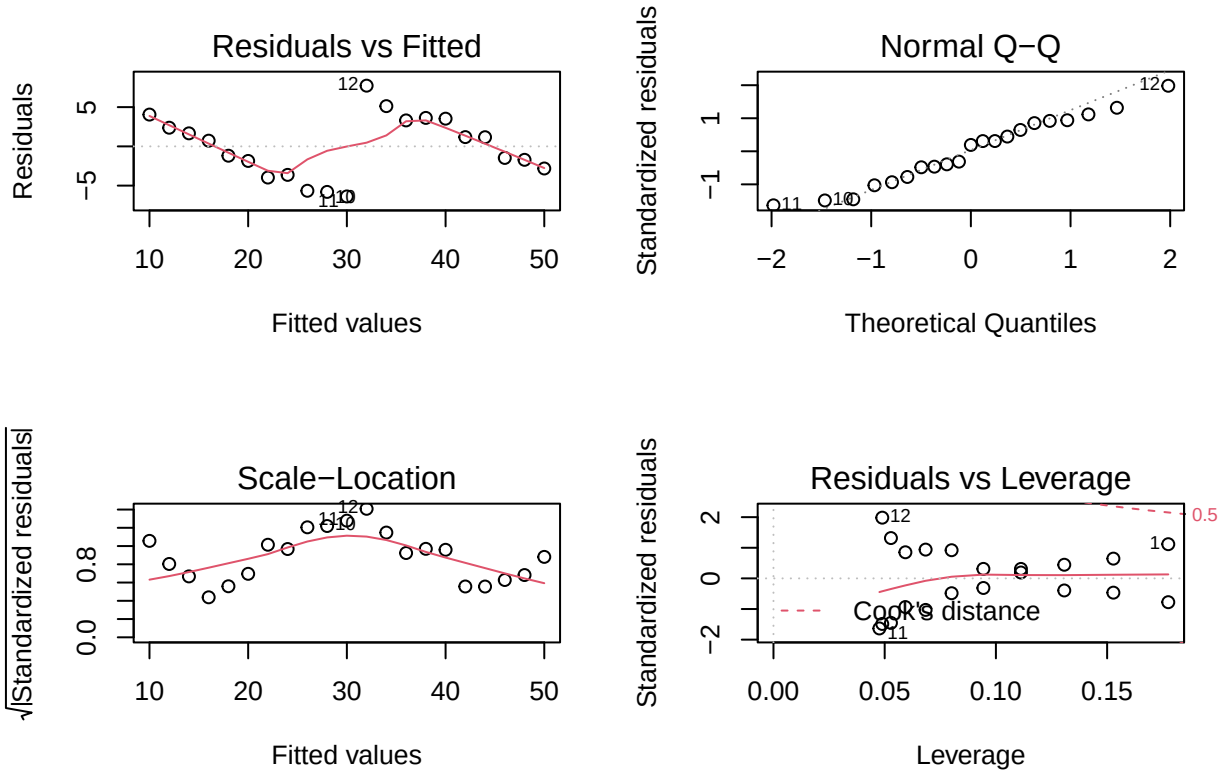
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X6,resid(lm6)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X6,dffits(lm6))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

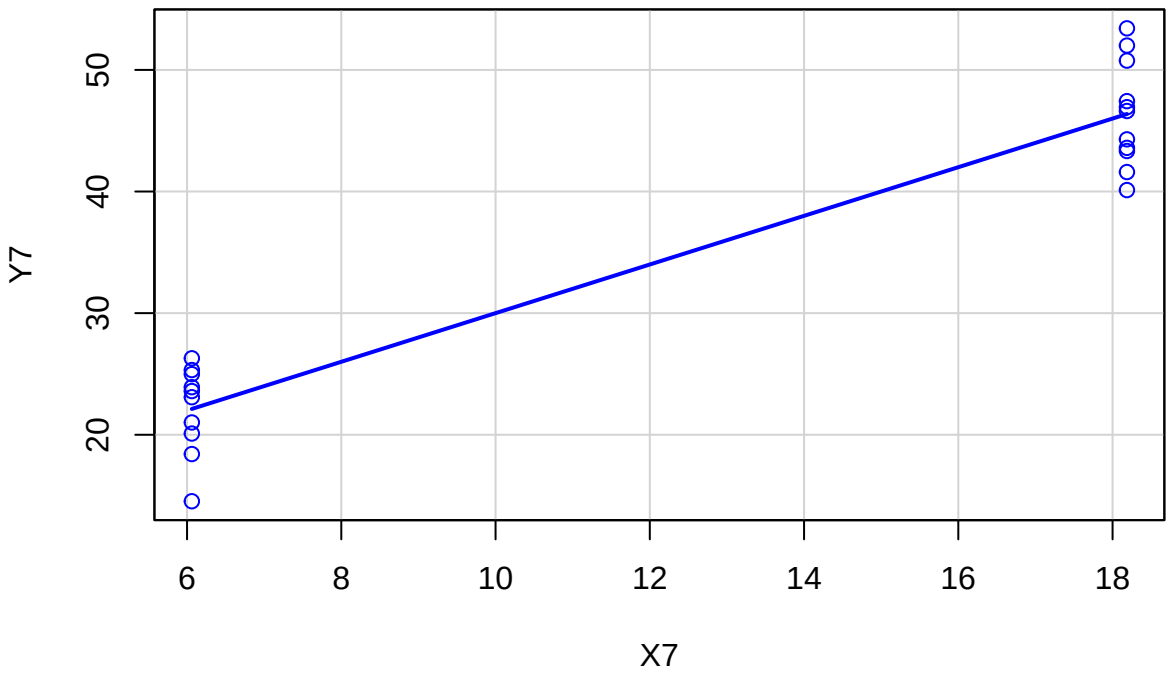
```
par(mfrow=c(2,2))
plot(lm6)
```



The patterns keep changing depending on the data, even though the estimates and the regression line stay the same.

Regression Line 7

```
X7 <- data[data$REG == 7, "X"]
Y7 <- data[data$REG == 7, "Y"]
scatterplot(X7, Y7, smooth = F, boxplots = F)
```




```
lm7 <- lm(Y7~X7)
summary(lm7)
```

```
#>
#> Call:
#> lm(formula = Y7 ~ X7)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -7.590 -2.782  0.563  2.858  7.044
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.9909   5.023 7.55e-05 ***
#> X7            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

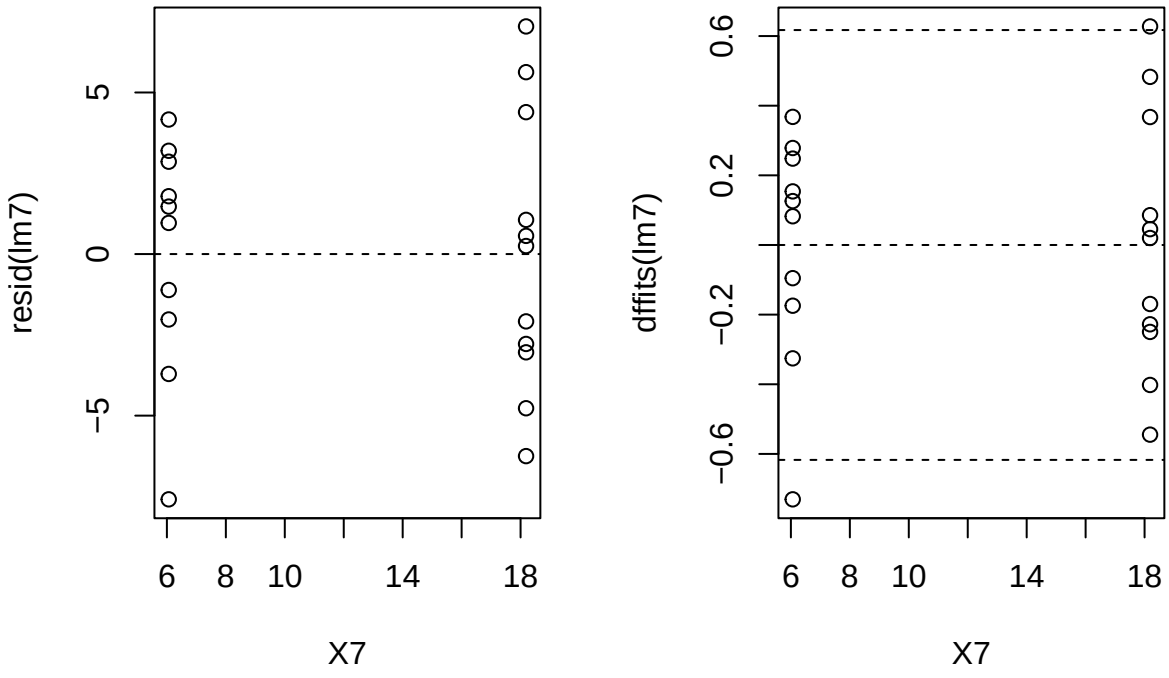
```
anova(lm7)
```

```
#> Analysis of Variance Table
#>
#> Response: Y7
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X7         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

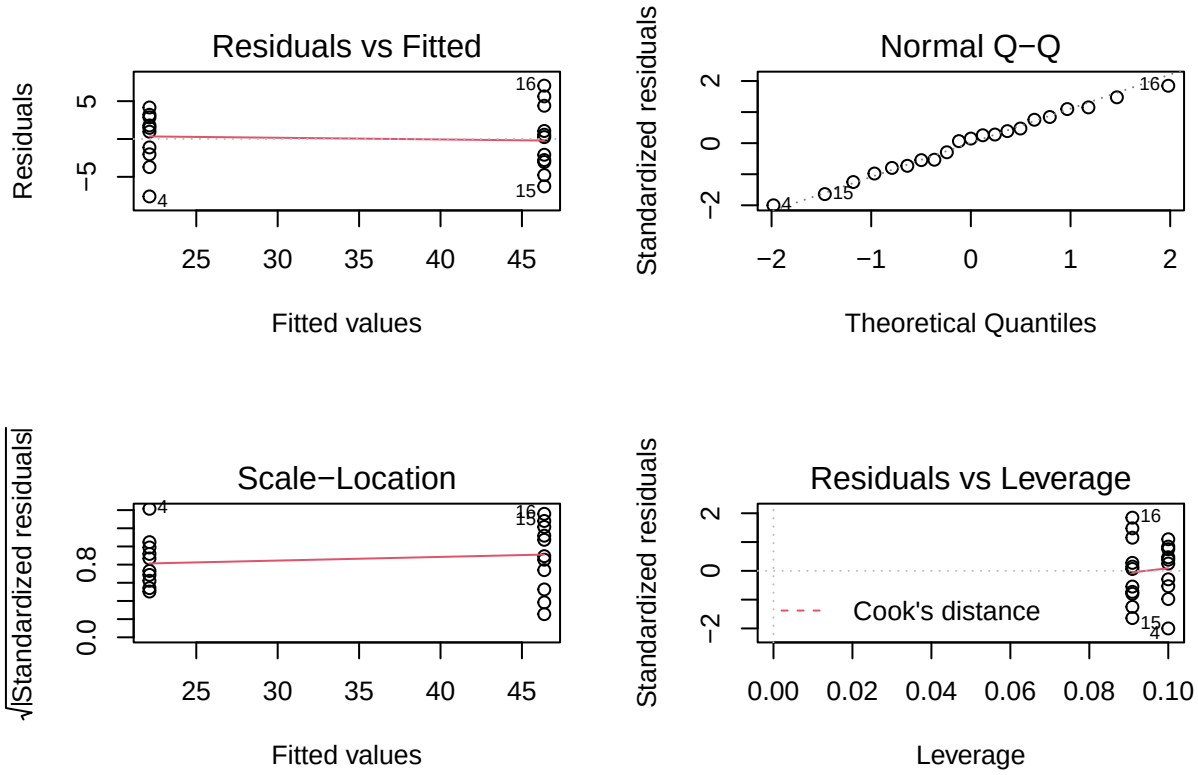
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X7,resid(lm7)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X7,dffits(lm7))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

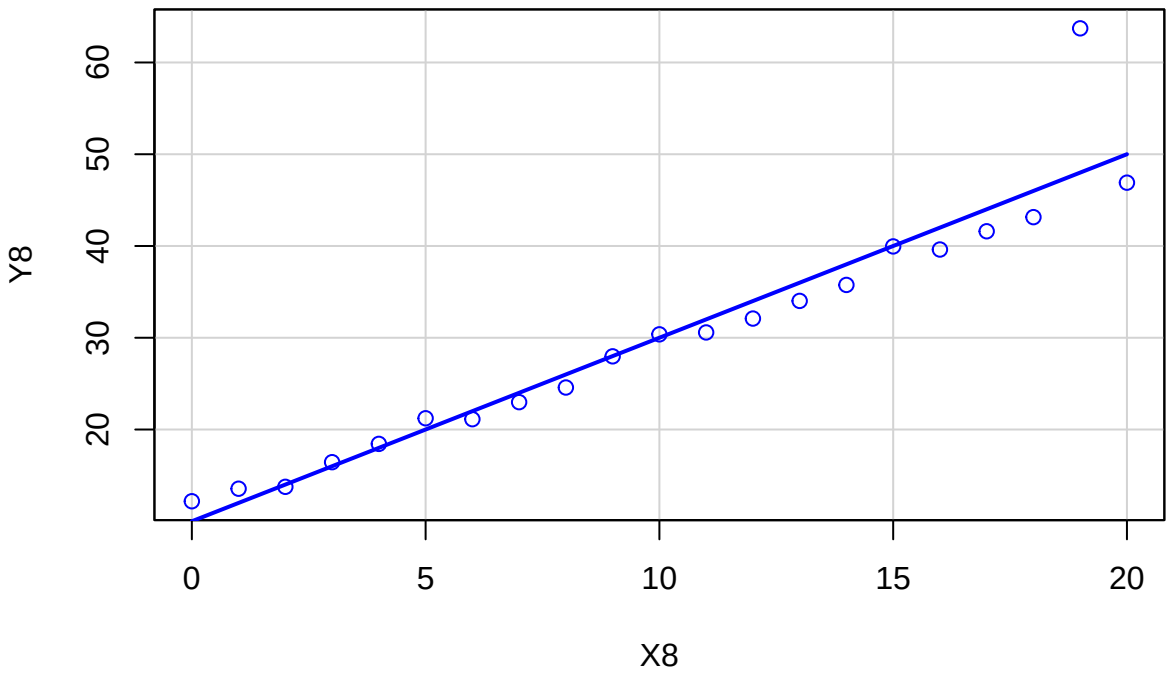
```
par(mfrow=c(2,2))
plot(lm7)
```



The patterns keep changing depending on the data, even though the estimates and the regression line stay the same.

Regression Line 8

```
X8 <- data[data$REG == 8, "X"]
Y8 <- data[data$REG == 8, "Y"]
scatterplot(X8, Y8, smooth = F, boxplots = F)
```




```
lm8 <- lm(Y8~X8)
summary(lm8)
```

```
#>
#> Call:
#> lm(formula = Y8 ~ X8)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -3.0982 -1.9827 -0.8758  0.4341 15.7201
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
#> X8            2.0000     0.1441  13.874 2.15e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4 on 19 degrees of freedom
#> Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
#> F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
p <- 2
n <- d[1]/8 # 21
```

R^2 is pretty large (close to 1). The points in the scatter plot look exponentially distributed (or polynomial of second degree).

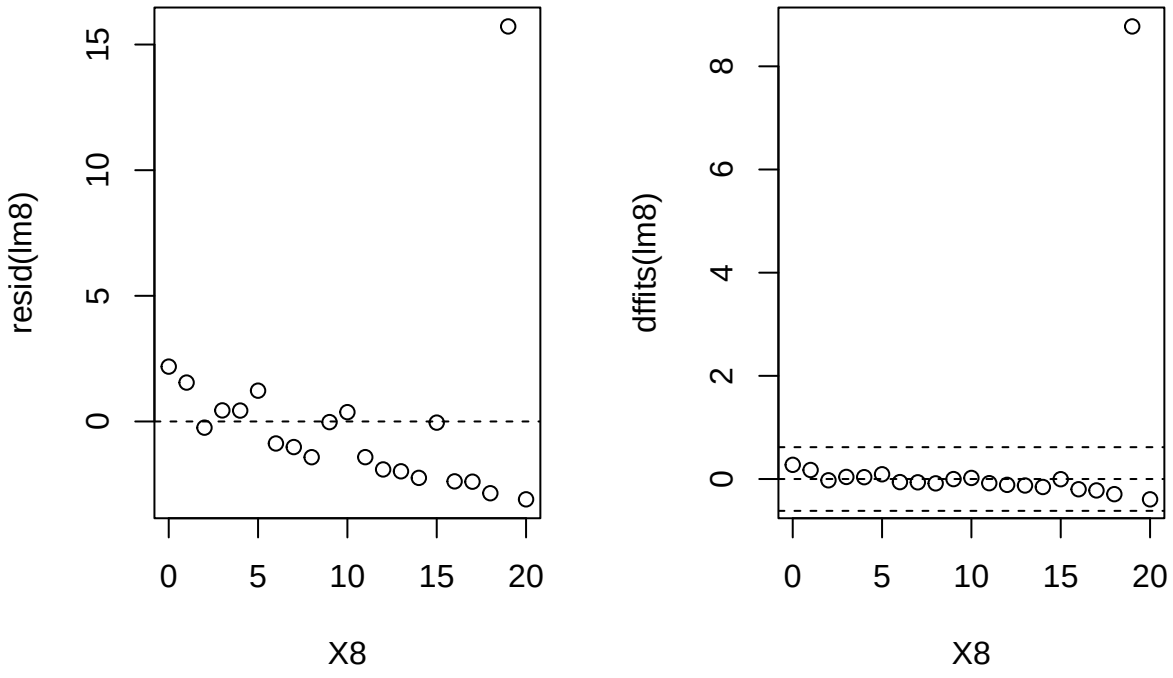
```
anova(lm8)
```

```
#> Analysis of Variance Table
#>
#> Response: Y8
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> X8         1   3080    3080   192.5 2.153e-11 ***
#> Residuals 19    304      16
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova-test shows that the model has some merit compared to the null model, i.e. that at least one of the coefficients is significantly different from zero.

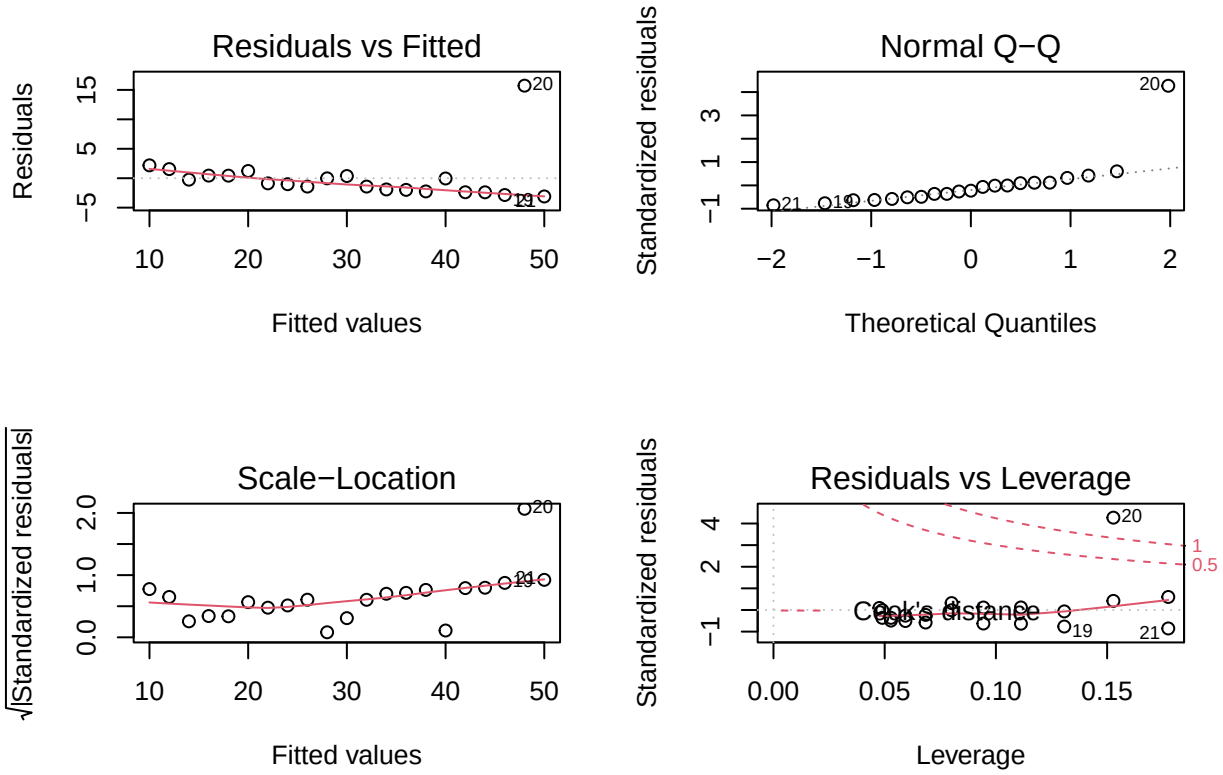
Some diagnostics:

```
par(mfrow=c(1,2))
plot(X8,resid(lm8)) #o rstudent(m) , h=c(-2,0,2)
abline(h=0,lty=2)
plot(X8,dffits(lm8))
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Still not sure what to conclude from the **dffits** and why those lines are plotted. What is the theory behind?

```
par(mfrow=c(2,2))
plot(lm8)
```



The patterns keep changing depending on the data, even though the estimates and the regression line stay the same.

Conclusions

All the lines have the same regression coefficients, where most of them have the same estimated standard errors and p-values. They also have the same value of R^2 and the anova/F-test gives the same conclusion. What changes is the validity of the linear model in each case, based on the distributions of the data. I would say that some of them could be valid (line 4 or 8 perhaps), i.e. the assumptions of the linear model seem to hold based on the diagnostic plots, while most of them do not hold because of patterns or trends in the data (which are clearly not linear). This could be fixed by using other regression models, or using GAMs (with splines, polynomials, exponentials, logs, etc.) if one insists on using linear models.