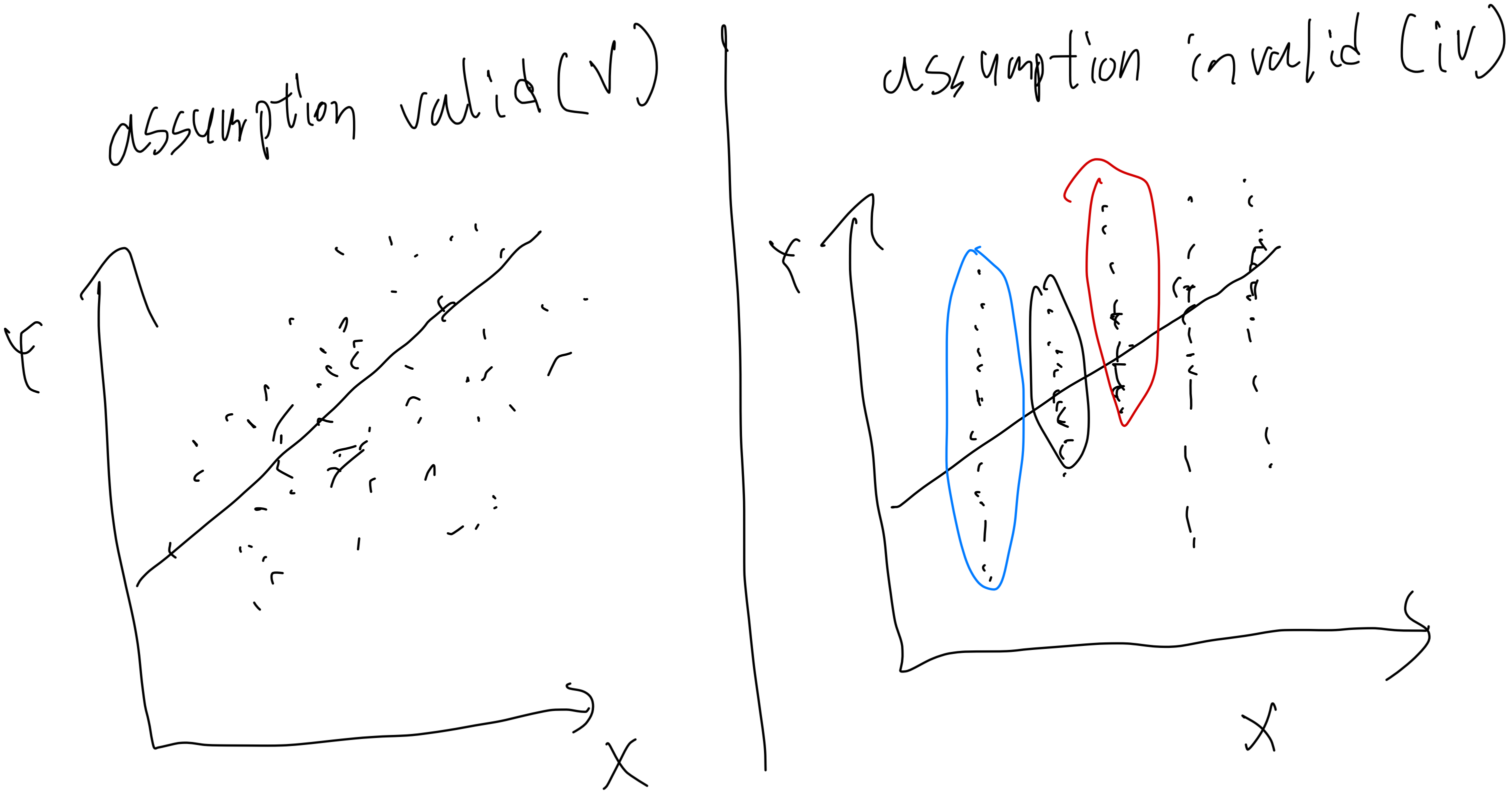
회귀분석HW3 20203374 이효웅

1. (b)

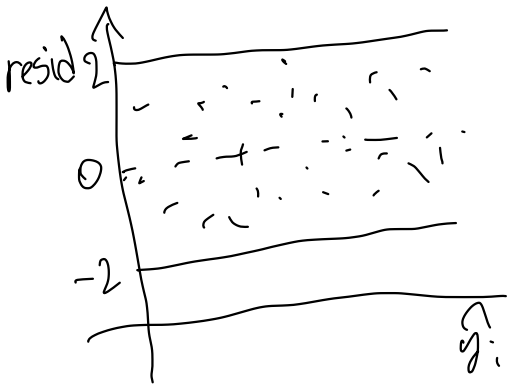


In graph (v) shows that observations are independent there's no specific pattern on graph. → that is valid case.

In graph (iv) shows that observations are not independent each others. They formed like some group. That kind of pattern can seen like medicine test with (treatment group with control group). → that is invalid case.
linear regression is conditional given observation. if independent assumption violated, that can be not affective.

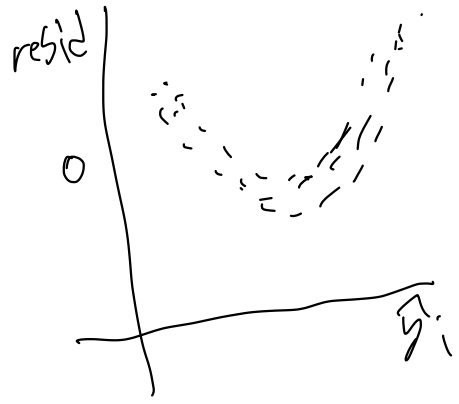
1. (c)

valid (v)

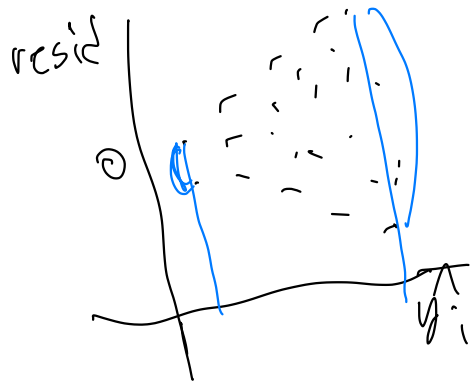


invalid

(iv I)



(iv II)



valid case

↓

in (v) residual ≈ standard normal dist, so it will plotted randomly around 0 and almost within [-2, 2] → that is constant variance case.

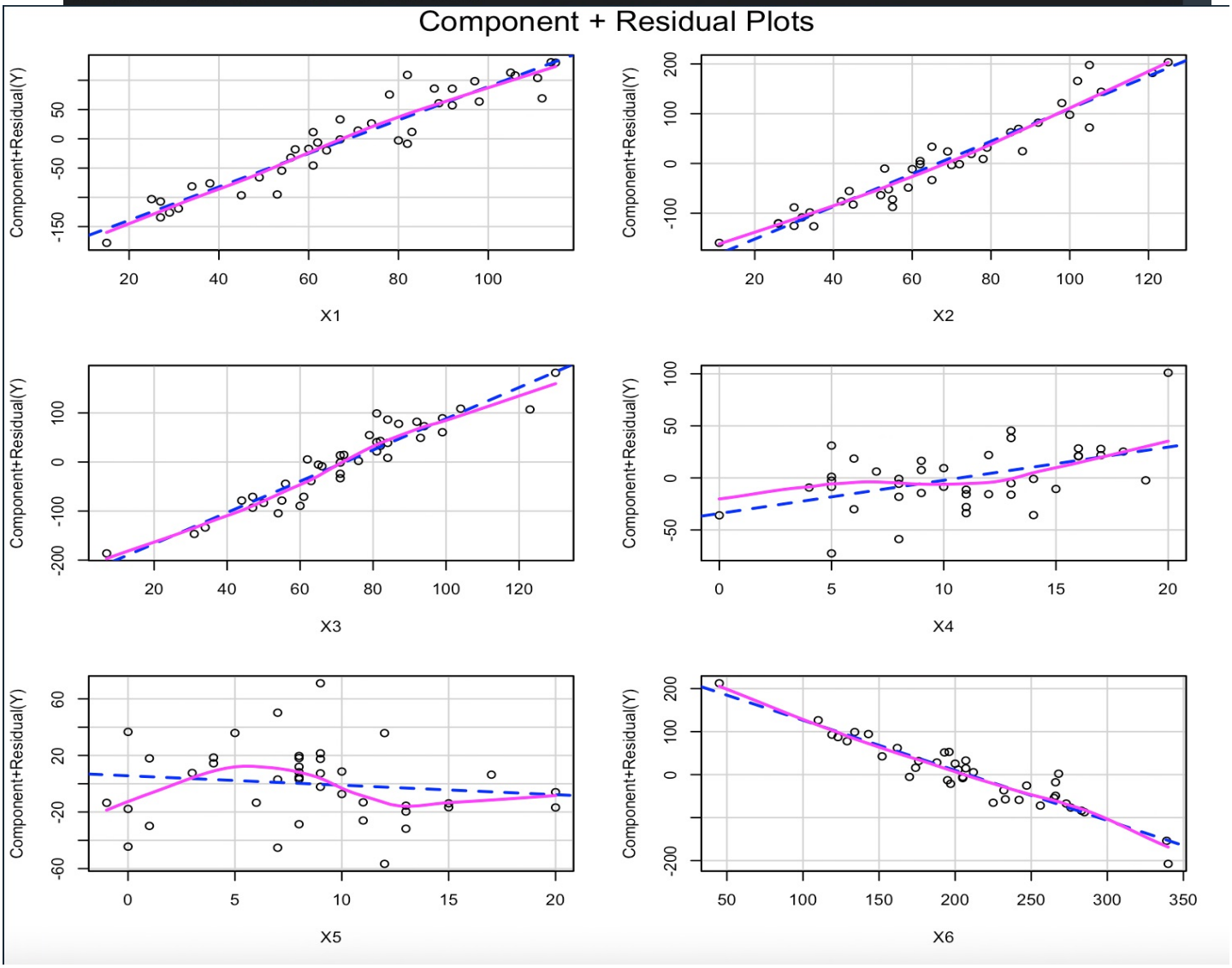
But invalid case like (iv II) that has hetero variance. And invalid case like (iv I) that shows non linear and not constant variance.

2.

(1)

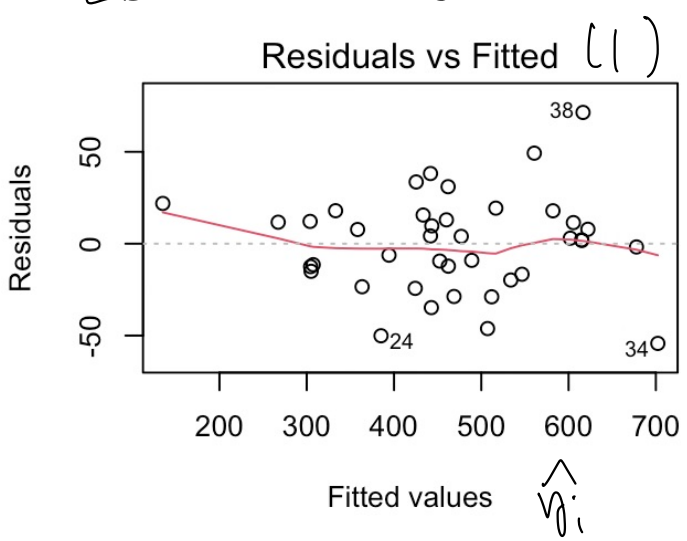
R 4.1.2 · ~/

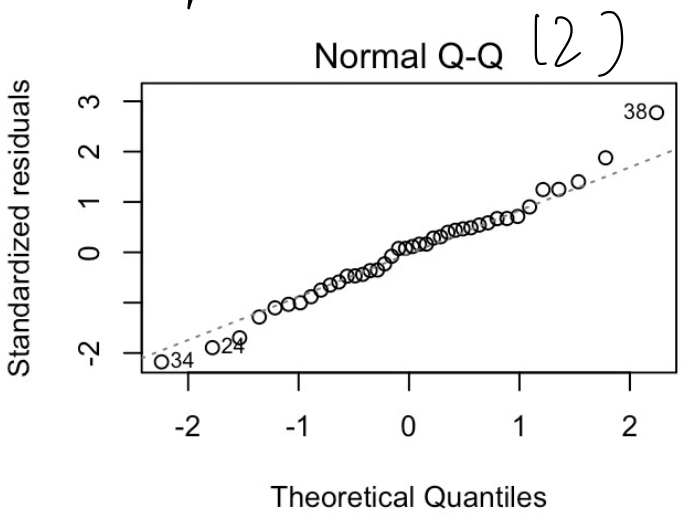
	Y	X1	X2	X3	X4	X5	X6
1	443	49	79	76	8	15	205
2	290	27	70	31	6	6	129
3	676	115	92	130	0	9	339
4	536	92	62	92	5	8	247
5	481	67	42	94	16	3	202
6	296	31	54	34	14	11	119

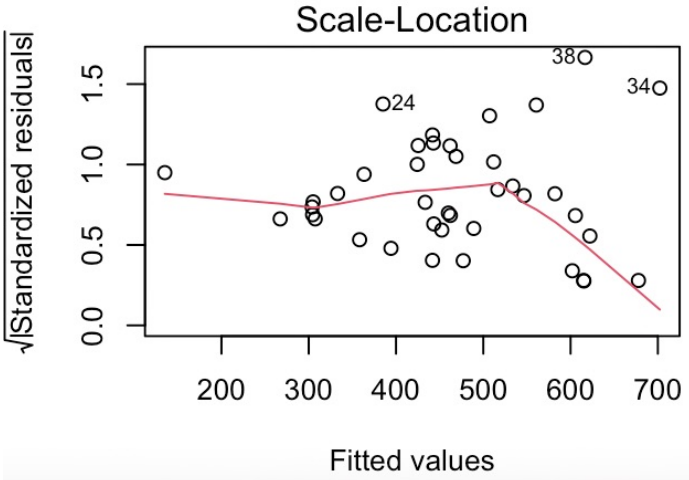


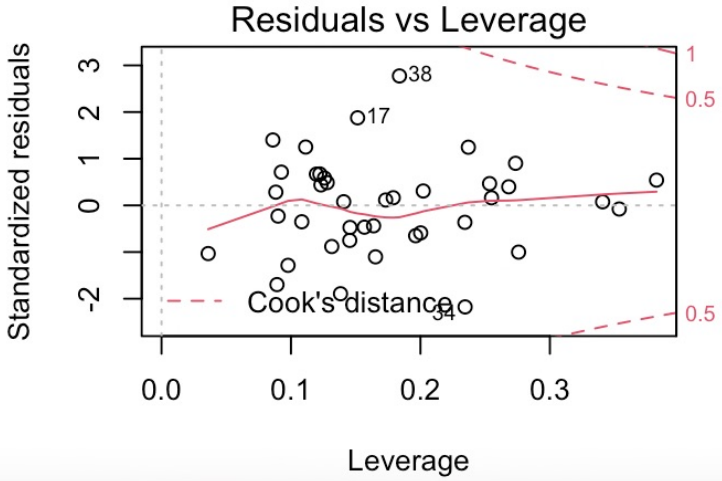
linearity assumption seems fine

Let's check error distribution assumption.









(2) and shapiro test shows normality assumption would fine.

```
> shapiro.test(lm$residuals)

        Shapiro-Wilk normality test

data:  lm$residuals
W = 0.98425, p-value = 0.8407
```

(1) shows some non constant variance

← non random pattern

constant variance assumption might break.

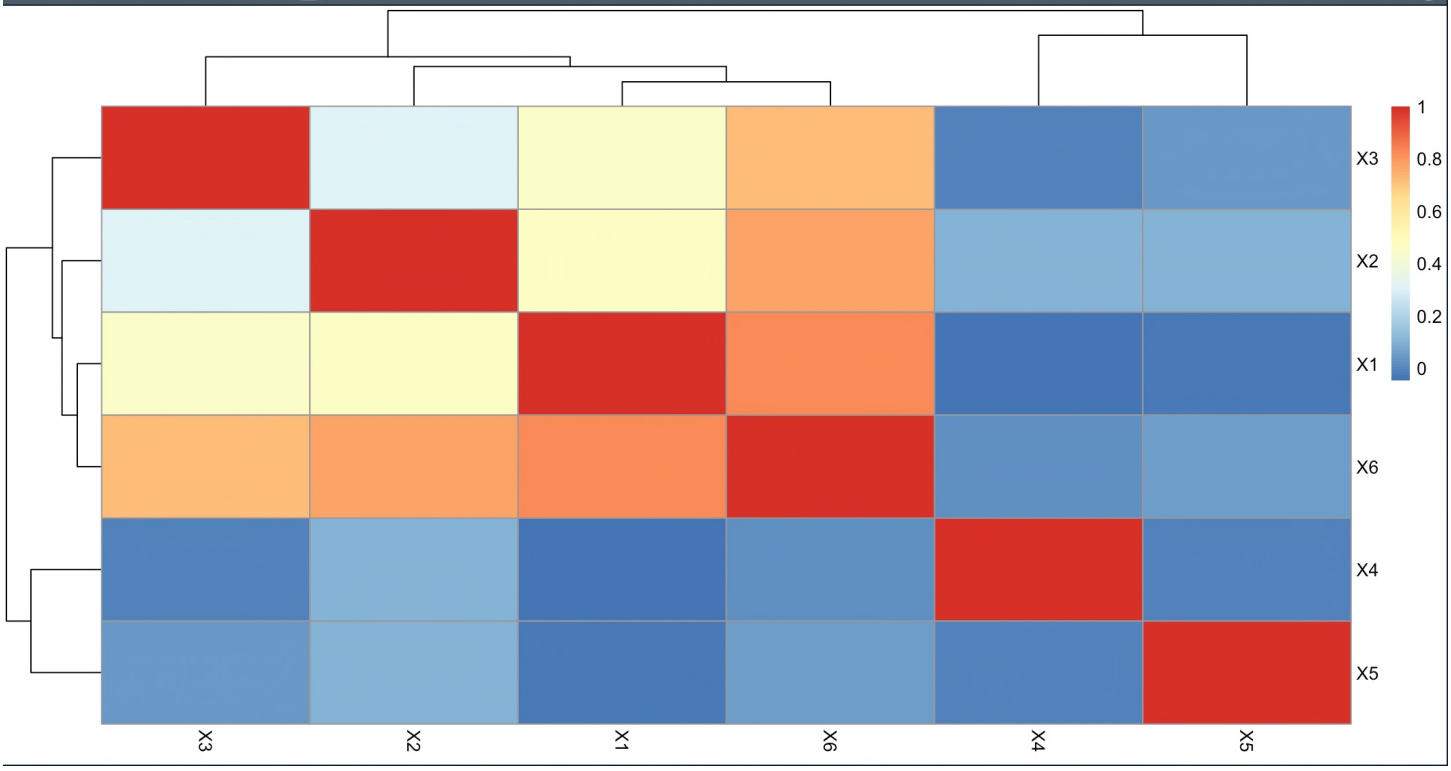
```
> durbinWatsonTest(residuals(lm))
[1] 2.305141
```

dw=2.30 thus over 2.0 and

(1) shows some kinds of non-independency.

Next page, I'll check predictor variables assumptions

```
> library(pheatmap)
> cor_matrix <- cor(df[,2:7])
> pheatmap(cor_matrix, legend = T)
>
```

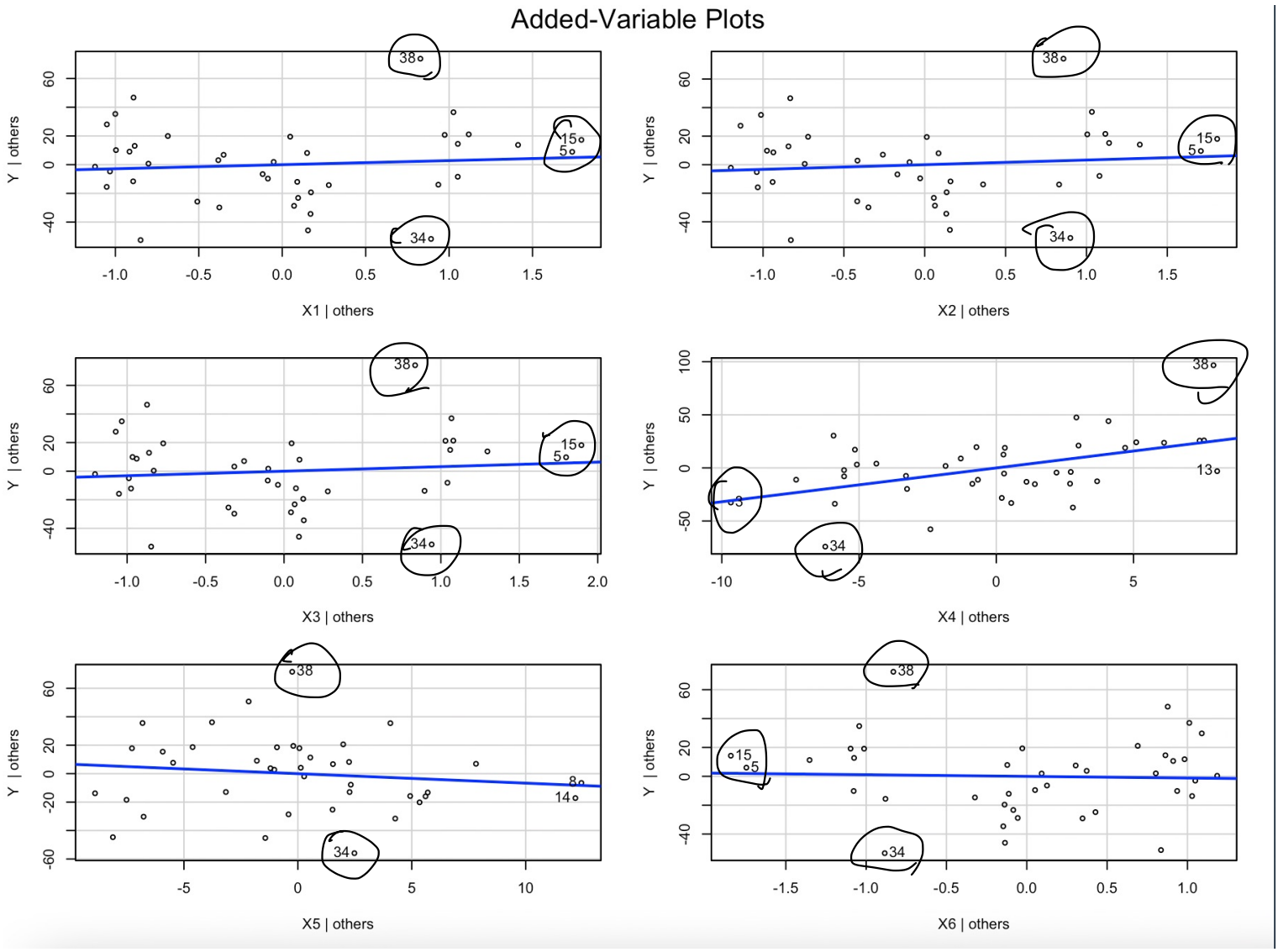


I'll check correlation matrix with independent variables.
And this plot shows some kinds of multicollinearity.

So, I think as error distribution assumption,
constant variance assumption were violated.
(independency

And about predictor variables, it shows some high correlation, so linearly independent assumption in predictor variables was violated. And it seems like have multicollinearity.

(2) unusual observations.



(3, 5, 15, 34, 34) points are suspicious.

(3, 15 → leverage points and outliers.
34, 34 → influential points and outliers.

And I'll check the details in next pages.

Check leverage point

We can use p_{ii} = i^{th} diagonal element of $P = X(X'X)^{-1}X'$ (leverage)

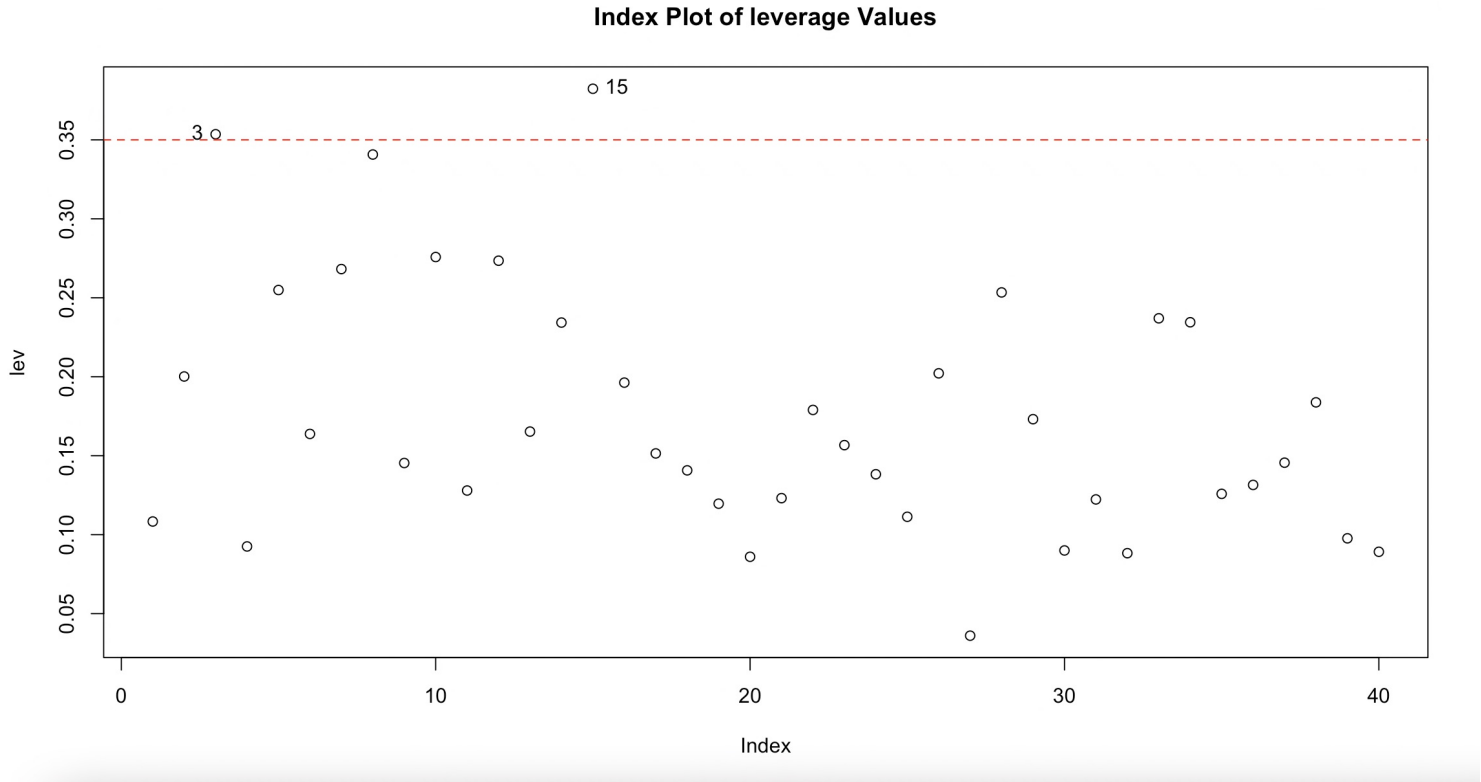
$n = 40, p = 6$ → and compare with $2\left(\frac{p+1}{n}\right)$

Standard_value $= \frac{2(p+1)}{n}$, we can compare $p_{ii} > 2\left(\frac{p+1}{n}\right)$ and if that's true → that points are high leverage points.

```
> lev<-influence(lm)$hat
> plot(lev, main = 'Index Plot of leverage Values')
> standard_value <- 2*(6+1)/40
> abline(h =  standard_value, lty = 2, col='red')
> identify(lev)
[1]  3 15
>
```

Plots | Packages | Help | Viewer

Zoom | Export | Publish



3, 15 points are $p_{ii} > \frac{2(p+1)}{n} = 0.35$ (above red line)

3, 15 points are leverage points.

check influential points → we can use cook's distance or DFITS but I'll use cook's.d

if $C_i \geq F(p+1, n-p-1; 0.5)$ or $C_i \geq 1$

then i^{th} is influential.

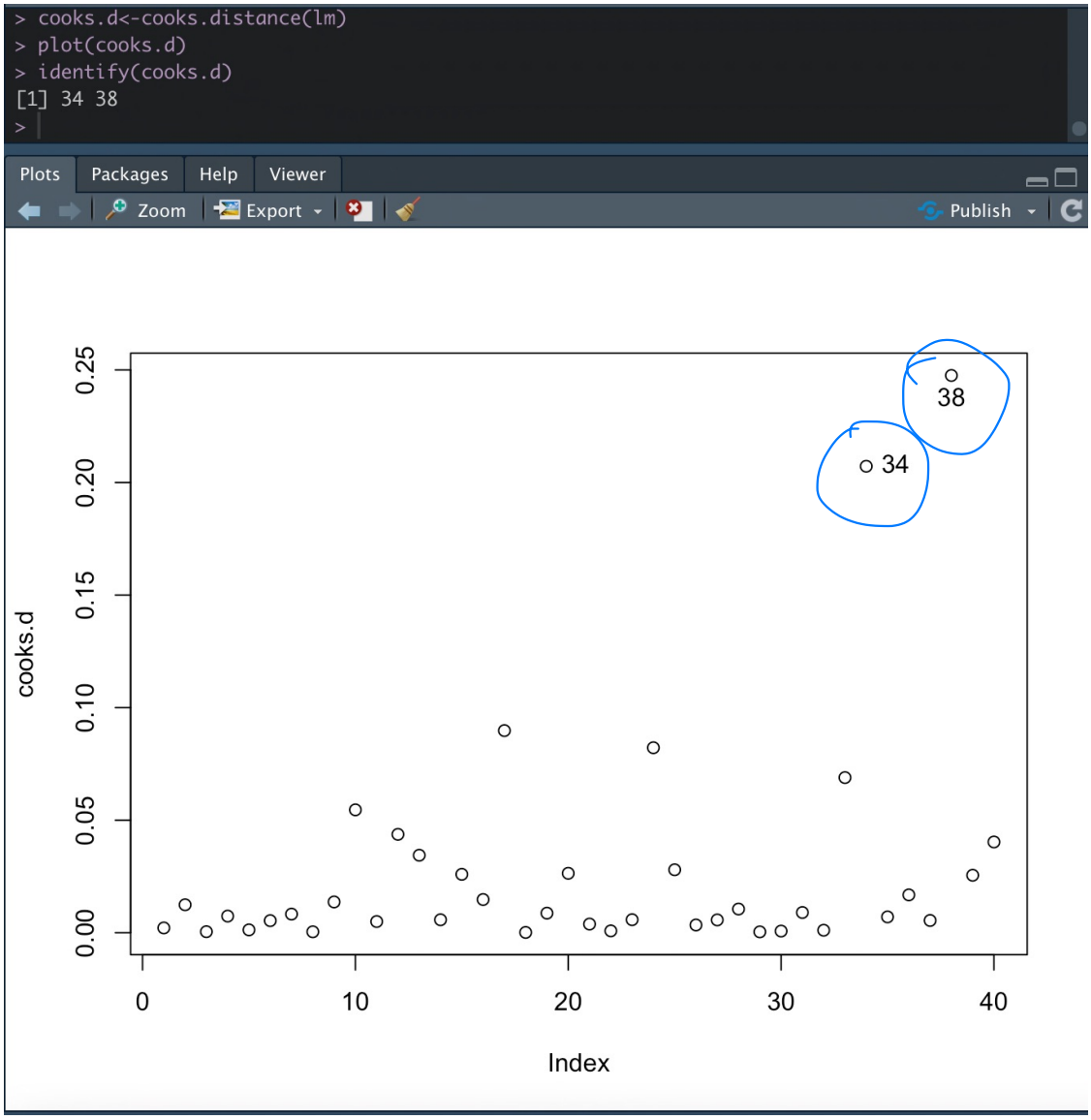
$p=6, n=40$

$F(7, 33, 0.5) =$

```
> qf(0.5, 7, 33)
[1] 0.9252857
```

→ we just need to check $C_i \geq 0.9252857$

```
> cooks.d<-cooks.distance(lm)
> plot(cooks.d)
> identify(cooks.d)
[1] 34 38
>
```



There's no points that $C_i \geq F(7, 33, 0.5)$

but we have to check 34, 38 points since it's far from other points. And It might be influential points.

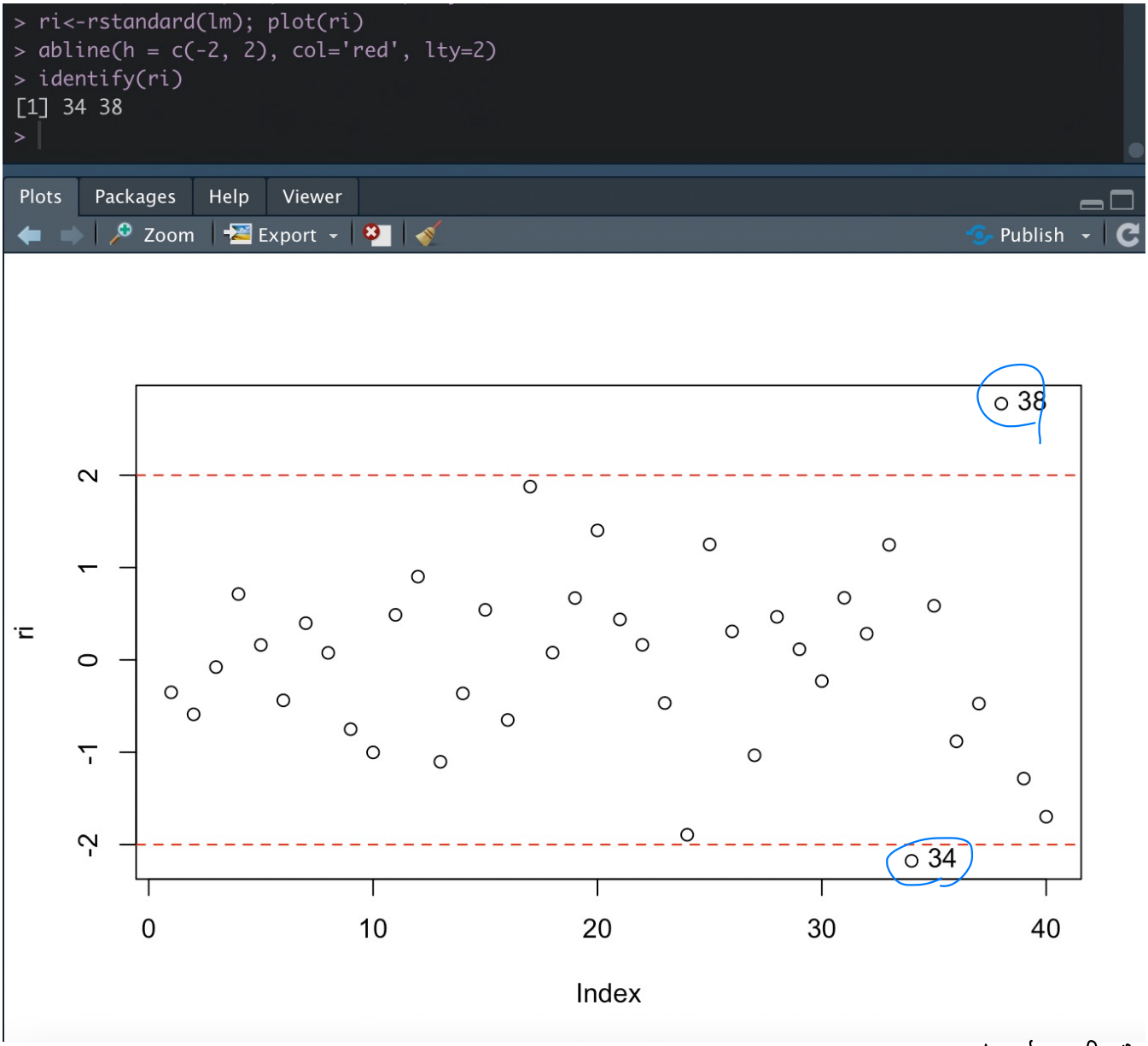
outliers

- x-axis → leverage_points.
- y-axis → standardized residual plot (studentized)

we can check outliers.

in (2) we found leverage points (3rd, 15th)

```
> ri<-rstandard(lm); plot(ri)
> abline(h = c(-2, 2), col='red', lty=2)
> identify(ri)
[1] 34 38
>
```



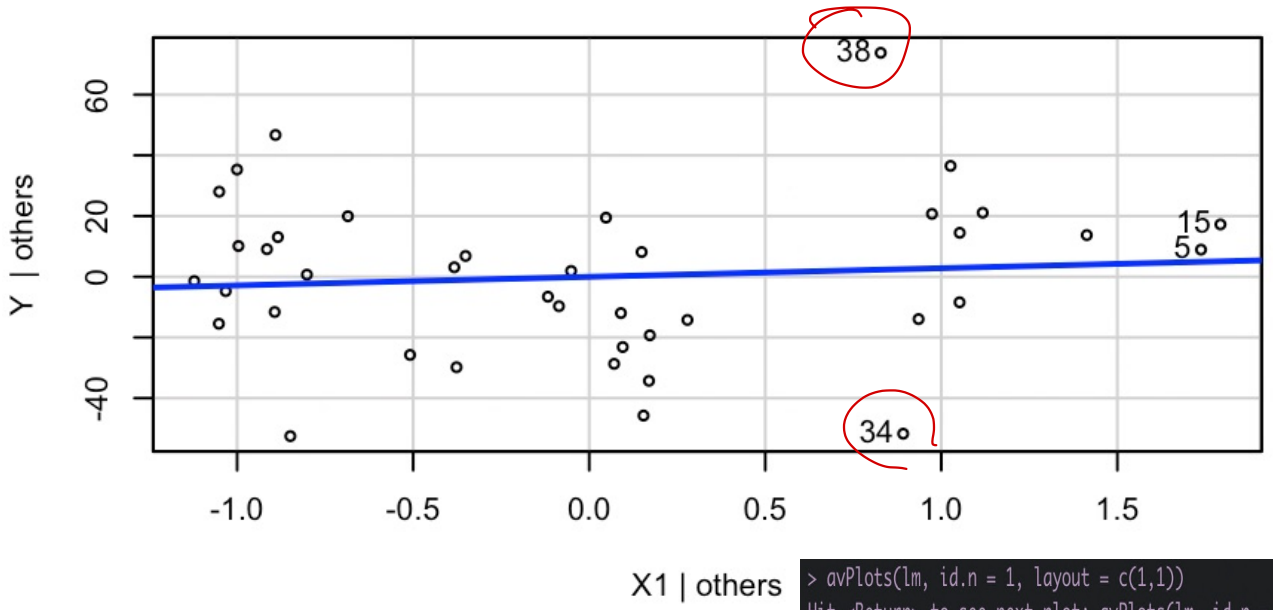
We can use standardized residual-index plot to find outliers. And 34, 38 seems out liers. Since residual $\hat{\sim} N(0,1)$, those points are out of (-2,2) (34, 38)

∴ (x-axis, 3, 15 points
y-axis, 34, 38 points)

we can conclude out liers (3, 15, 34, 38) points.

(3) A-V plot for X_1

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$

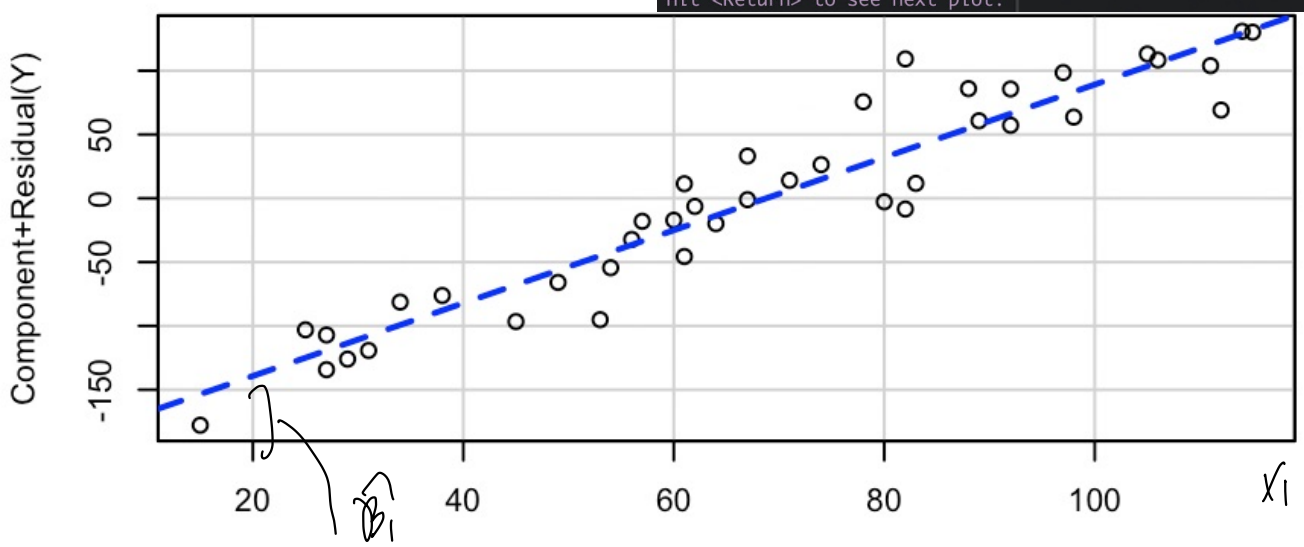



```
> avPlots(lm, id.n = 1, layout = c(1,1))
Hit <Return> to see next plot: avPlots(lm, id.n = 1, layout = c(1,1))
Hit <Return> to see next plot:
```

R+C plots for X_1

```
> crPlots(lm, smooth=F, layout =c(1,1))
Hit <Return> to see next plot: crPlots(lm, smooth=F, layout =c(1,1))
Hit <Return> to see next plot:
```

$e_i + \hat{\beta}_1 x_1$



A-V plot shows, remove $(X_2, X_3 \cdots X_6)$'s effects for X_1 and y and display linearity. And we can check X_1 can explain Y's linearity.

R+C Plot also shows linearity with X_1 and Y. Slope might be $\hat{\beta}_1$, and points seems good fitted with line. We can sure it's linearity.