HW4. 20203314 이호영

1. (1)

$F_1 : \begin{cases} 1 & \text{if } i^{th} \in \text{Fertilizer 1} \\ 0 & \text{o.w.} \end{cases}$

$F_2 : \begin{cases} 1 & \text{if } i^{th} \in \text{Fertilizer 2} \\ 0 & \text{o.w.} \end{cases}$

$F_3 : \begin{cases} 1 & \text{if } i^{th} \in \text{Fertilizer 3} \\ 0 & \text{o.w.} \end{cases}$

We can use factor function in R but I'll use another method. (since it said make indicator variables)

```
> rm(list=ls())
> load('/Users/hoyong/Downloads/RABE5.RData')
> df <- P158
> head(df)
  Yield Fertilizer
1    31          1
2    34          1
3    34          1
4    34          1
5    43          1
6    35          1
> str(df)
'data.frame':   40 obs. of  2 variables:
 $ Yield     : num  31 34 34 34 43 35 38 36 36 45 ...
 $ Fertilizer: int  1 1 1 1 1 1 1 1 1 1 ...
> df$Fertilizer <- factor(df$Fertilizer, levels = c(4,1,2,3)) ;str(df)
'data.frame':   40 obs. of  2 variables:
 $ Yield     : num  31 34 34 34 43 35 38 36 36 45 ...
 $ Fertilizer: Factor w/ 4 levels "4","1","2","3": 2 2 2 2 2 2 2 2 2 2 ...
```

↑ this is factor function methods

↓ I'll use this method

```
> rm(list=ls())
> load('C:/Users/Hoyong/Downloads/RABE5.RData')
> df <- P158
> head(df)
  Yield Fertilizer
1    31          1
2    34          1
3    34          1
4    34          1
5    43          1
6    35          1
> df$F1 <- ifelse(df$Fertilizer == 1, 1, 0)
> df$F2 <- ifelse(df$Fertilizer == 2, 1, 0)
> df$F3 <- ifelse(df$Fertilizer == 3, 1, 0)
> df$F4 <- ifelse(df$Fertilizer == 4, 1, 0)
```

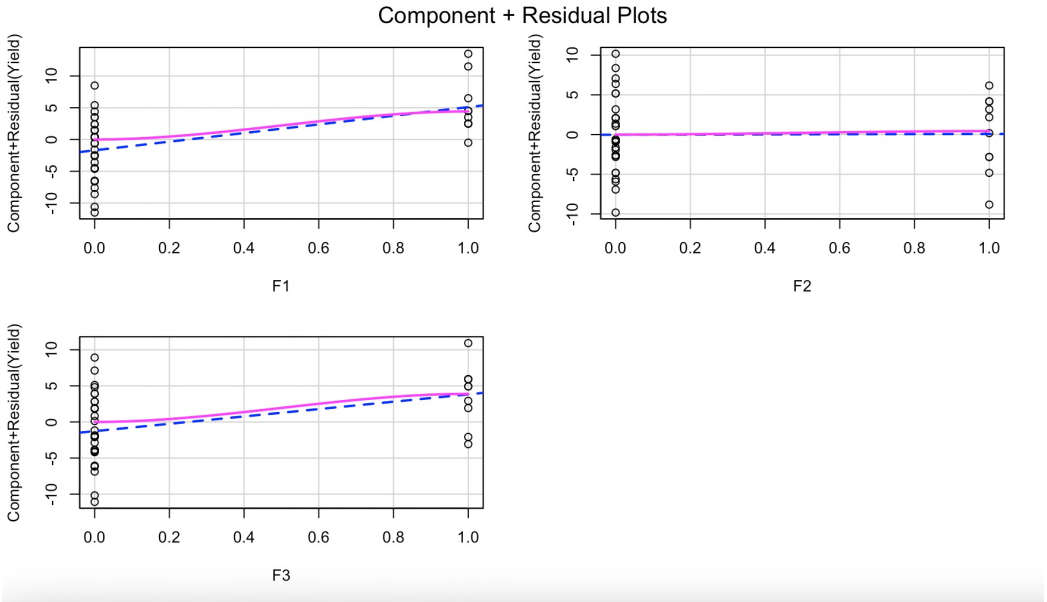
(2)

→ I'll use 2 methods. ← method I → anova with post-Hoc
method II → Joint hypothesis test.

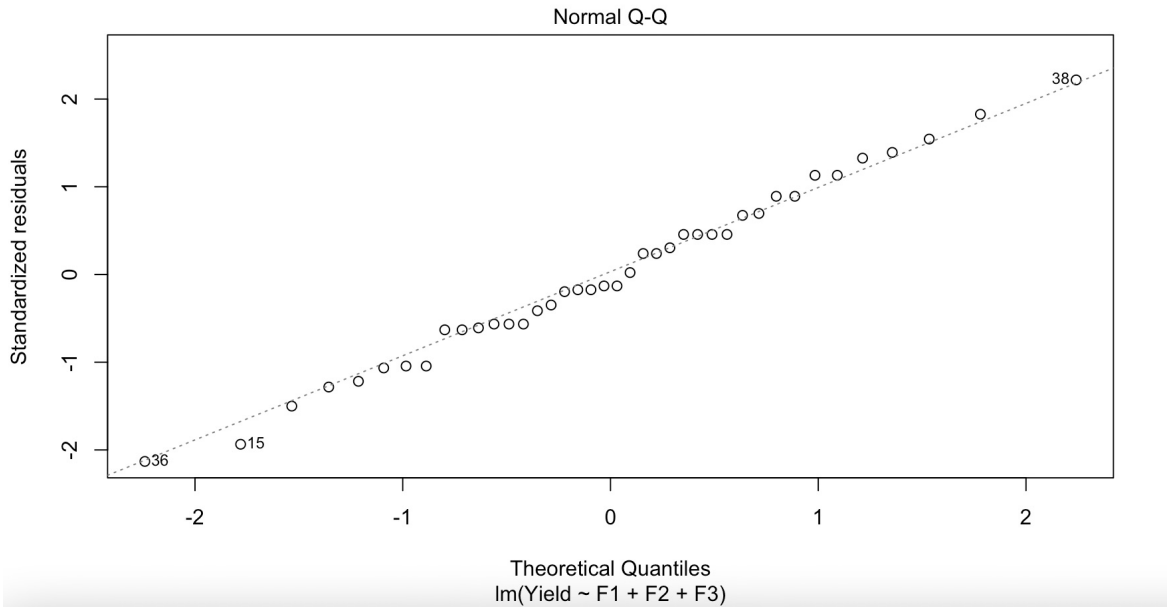
```
> anova(update(lm, ~1), lm)
Analysis of Variance Table

Model 1: Yield ~ 1
Model 2: Yield ~ F1 + F2 + F3
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     39 1208.4
2     36  845.8  3     362.6 5.1445 0.004605 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

→ more than 1 coefficients are significant.

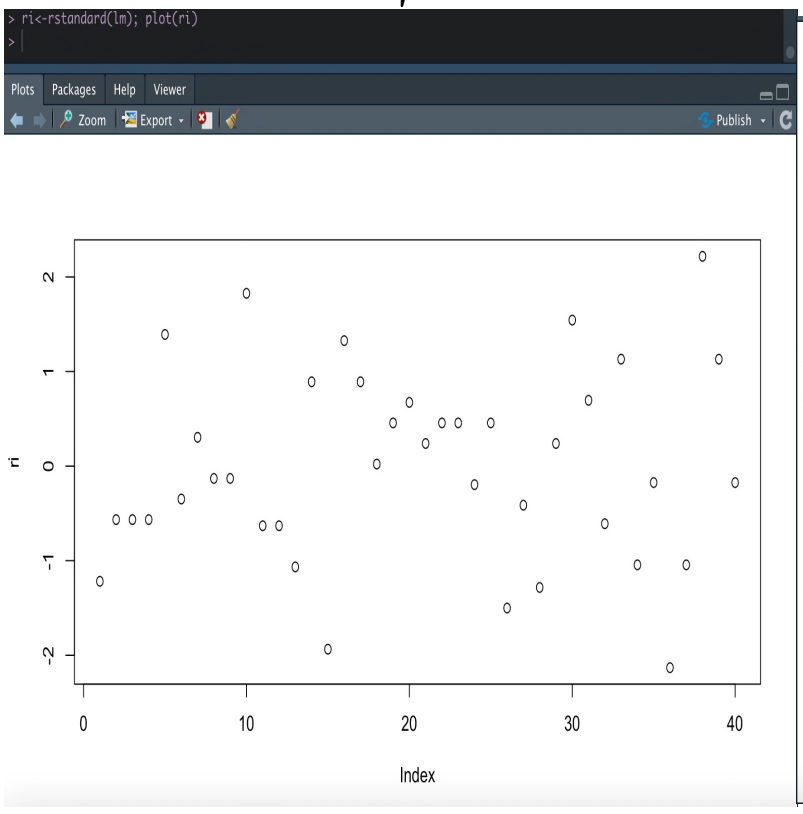
```
> summary(lm)

Call:
lm(formula = Yield ~ F1 + F2 + F3, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-9.800 -2.825 -0.600  3.125 10.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.800      1.533  19.442   <2e-16 ***
F1             6.800      2.168   3.137   0.0034 **
F2             0.100      2.168   0.046   0.9635
F3             5.100      2.168   2.353   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.847 on 36 degrees of freedom
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2417
F-statistic: 5.144 on 3 and 36 DF,  p-value: 0.004605
```

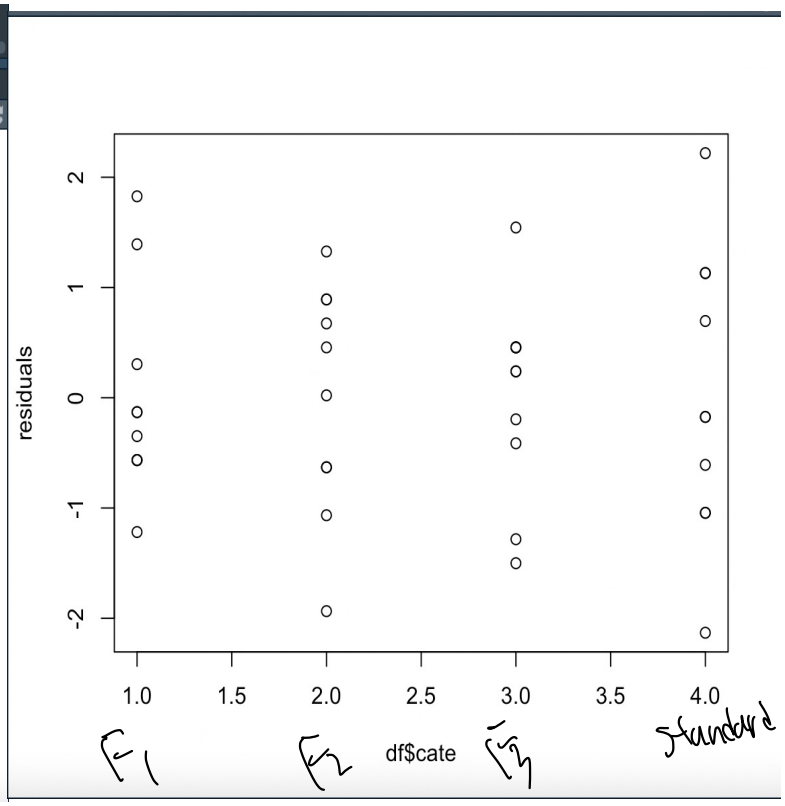


linearity seems fine.



normality seems fine





independency seems fine

constant variance seems fine.

Method(I) multiple mean comparison, we can use anova.

```
        Shapiro-Wilk normality test

data:  df$Yield[df$Fertilizer == "1"]
W = 0.87457, p-value = 0.113


        Shapiro-Wilk normality test

data:  df$Yield[df$Fertilizer == "2"]
W = 0.94075, p-value = 0.5614


        Shapiro-Wilk normality test

data:  df$Yield[df$Fertilizer == "3"]
W = 0.92453, p-value = 0.3963


        Shapiro-Wilk normality test

data:  df$Yield[df$Fertilizer == "4"]
W = 0.97409, p-value = 0.926
```

```
> car::leveneTest(Yield ~ Fertilizer, data=df, center=median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.6638 0.5797
      36
```

normality, homogeneity of variance assumption satisfied.

```
> summary(aov(df$Yield ~ df$Fertilizer, data=df))
             Df Sum Sq Mean Sq F value Pr(>F)
df$Fertilizer  3  362.6  120.87   5.144 0.0046 **
Residuals     36  845.8   23.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_0$

H_1: not H_0, p-value = 0.0046 < 0.05

Reject H_0.

post-hoc test

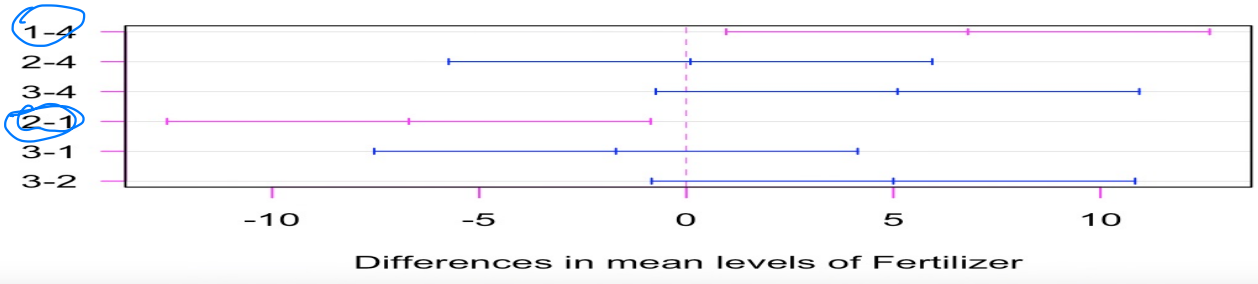
we have to check which pairs are different.

since number of each groups are same(10) and homogeneous variance assumption satisfied, we can use Tukey's LSD method to check which pairs are different.

```
> group_aov <- aov(Yield~Fertilizer, data=df)
> test <- TukeyHSD(group_aov)
> plot(test, las=1, col=ifelse(test$Fertilizer[,4] < 0.05,'magenta', 'blue'))
>
```

Plots | Packages | Help | Viewer

Zoom | Export | Publish



```
> group_aov <- aov(Yield~Fertilizer, data=df)
> TukeyHSD(group_aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Yield ~ Fertilizer, data = df)

$Fertilizer
     diff         lwr        upr     p adj
1-4   6.8   0.9619128 12.6380872 0.0170864
2-4   0.1  -5.7380872  5.9380872 0.9999640
3-4   5.1  -0.7380872 10.9380872 0.1050237
2-1  -6.7 -12.5380872 -0.8619128 0.0191978
3-1  -1.7  -7.5380872  4.1380872 0.8610988
3-2   5.0  -0.8380872 10.8380872 0.1154045
```

H_0: paired groups mean is same

H_1: not H_0

I'll use anova and tukey's multiple comparisons test.

$(F_1 - \text{control})$, $(F_2 - F_1)$ are not same and others are same. So three types of fertilizer does not have equal effects on corn corp.

Method II. (Joint hypothesis test)

Full model: $\mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \varepsilon_i$

H_0: $\mu_1 = \mu_2 = \mu_3$ → Reduced model

H_1: not H_0. → Full model

we can use F-test.

```
> df2$F1 <- ifelse(df2$Fertilizer == 1, 1, 0)
> df2$F2 <- ifelse(df2$Fertilizer == 2, 1, 0)
> df2$F3 <- ifelse(df2$Fertilizer == 3, 1, 0)
> df2$F4 <- ifelse(df2$Fertilizer == 4, 1, 0)
> lm_reduced <- lm(Yield ~ I(F1 + F2 + F3), data=df2)
> lm_full <- lm(Yield ~ F1 + F2 + F3, data=df2)
> anova(lm_reduced,lm_full)
Analysis of Variance Table

Model 1: Yield ~ I(F1 + F2 + F3)
Model 2: Yield ~ F1 + F2 + F3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     38 1088.4
2     36  845.8  2     242.6 5.1629 0.01068 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This F-test shows that p-value = 0.01064 < 0.05(α)
Reject H_0. Then we can't say that three types of fertilizers have same effect on corn corps.

$$F = \frac{(SSE(RM) - SSE(FM))/(3-1)}{SSE(FM)/(40-3-1)} = 5.1629$$

F-test statistic = 5.1629 > $F_{(2,36,0.05)}$

Reject H_0. i.e. with above result.

```
> sse_rm <- 1088.4
> sse_fm <- 845.80
> p <- 3 ; q <- 1 ; n=40
> numerator <- (sse_rm-sse_fm)/(p-q)
> denominator <- sse_fm/(n-p-1)
> f <- numerator/denominator
> f
[1] 5.162923
> qf(0.05, 2, 36, lower.tail = FALSE)
[1] 3.259446
```

(3) Base crop (control group) = 29.800 (Intercept of lm)

F_1 : $29.8 + 6.8 \pm 1.533 \times 2.024344$
$\hat{\mu}_0 + \hat{\mu}_1 \pm s.e.(\hat{\mu}_0) \times t(36, 0.025)$
critical value

F_2 : $29.8 + 0.1 \pm 1.533 \times 2.024344$
$\hat{\mu}_0 + \hat{\mu}_2 \pm s.e.(\hat{\mu}_0) \times t(36, 0.025)$

F_3 : $29.8 + 5.1 \pm 1.533 \times 2.024344$
$\hat{\mu}_0 + \hat{\mu}_3 \quad s.e.(\hat{\mu}_0) \times t(36, 0.025)$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.800      1.533  19.442   <2e-16 ***
F1             6.800      2.168   3.137   0.0034 **
F2             0.100      2.168   0.046   0.9635
F3             5.100      2.168   2.353   0.0242 *
---
```

μ_1 μ_2 μ_3

```
> critical_value <- qt(0.025, 40-2, lower.tail=FALSE)
> mu0_hat <- 29.8
> st_err <- 1.533
> mu <- c(6.8, 0.1, 5.1)
> f1_ci <- c(mu0_hat+mu[1] - st_err*critical_value, mu0_hat+mu[1] + st_err*critical_value)
> f2_ci <- c(mu0_hat+mu[2] - st_err*critical_value, mu0_hat+mu[2] + st_err*critical_value)
> f3_ci <- c(mu0_hat+mu[3] - st_err*critical_value, mu0_hat+mu[3] + st_err*critical_value)
> f1_ci ; f2_ci ; f3_ci
[1] 33.4966 39.7034
[1] 26.7966 33.0034
[1] 31.7966 38.0034
>
```

95% C.I for mean response corn crop

F_1 (33.4966, 39.7034)

F_2 (26.7966, 33.0034)

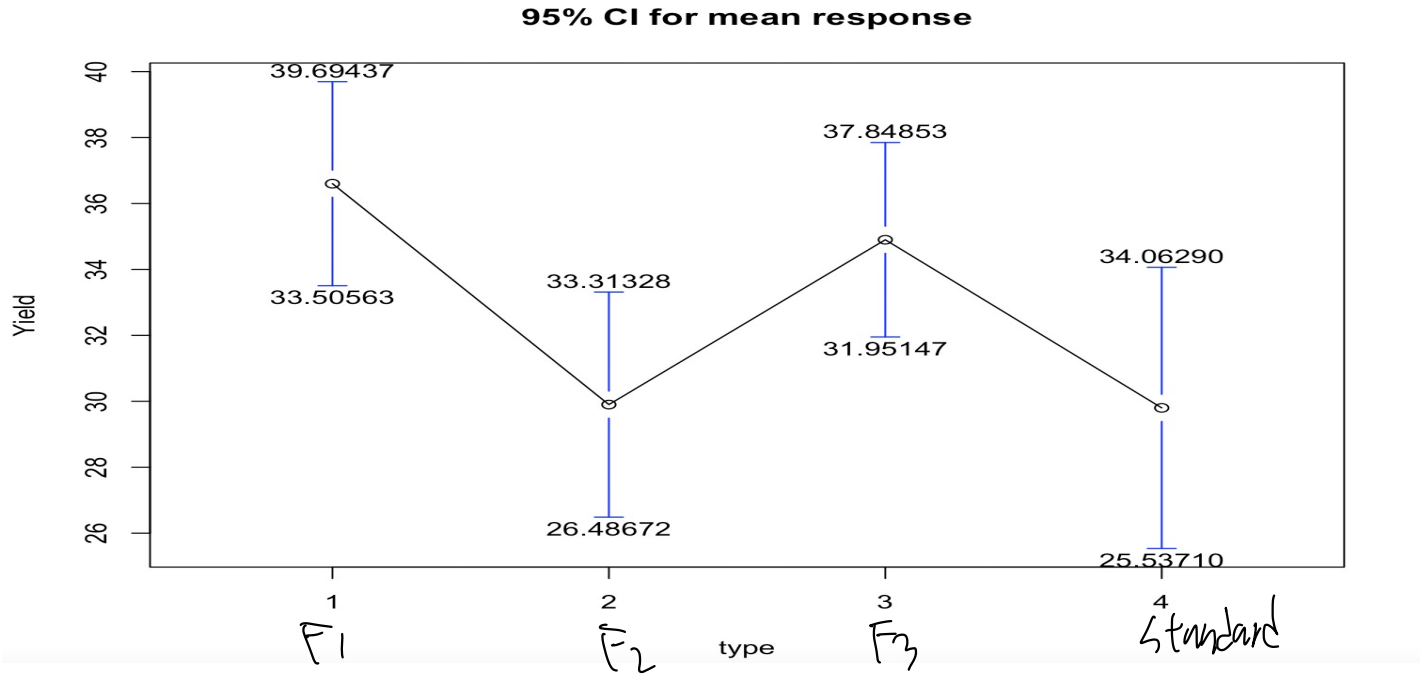
F_3 (31.7966, 38.0034)

```
> new.data <- data.frame(X=rep(0,3),
+                        F1=c('1','0','0'),
+                        F2=c('0','1','0'),
+                        F3=c('0','0','1'))
> predict(lm, new.data, se.fit=TRUE, interval = 'confidence')
$fit
   fit      lwr      upr
1 36.6 33.49136 39.70864
2 29.9 26.79136 33.00864
3 34.9 31.79136 38.00864

$se.fit
      1       2       3
1.53279 1.53279 1.53279

$df
[1] 36

$residual.scale
[1] 4.847107
```

```
> plotmeans(Yield ~ cate, xlab='type', data=df, p=.95, main = '95% CI for mean response', ci.label =
  TRUE, n.label=FALSE)
>
```

Plots | Packages | Help | Viewer

Zoom | Export | Publish



2.

(1)

```
> load('/Users/hoyong/Downloads/RABE5.RData')
> df <- P130
> head(df)
       S X E M
1 13876 1 1 1
2 11608 1 3 0
3 18701 1 3 1
4 11283 1 2 0
5 11767 1 3 0
6 20872 2 2 1
> df$E <- factor(df$E, levels=c(1,2,3))
> df$M <- factor(df$M)
> lm <- lm(S ~ X + E + M, data=df)
> summary(lm)

Call:
lm(formula = S ~ X + E + M, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1884.60  -653.60    22.23   844.85  1716.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8035.60     386.69  20.781  < 2e-16 ***
X             546.18      30.52  17.896  < 2e-16 ***
E2           3144.04     361.97   8.686 7.73e-11 ***
E3           2996.21     411.75   7.277 6.72e-09 ***
M1           6883.53     313.92  21.928  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1027 on 41 degrees of freedom
Multiple R-squared:  0.9568,    Adjusted R-squared:  0.952
5
F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

It's difference of standard variables. In lecture note E_3 is standard, above model E_1 is standard. (variable)

So X, M_1 are same and (Intercept, E_2) columns are different.

d.f, R^2, R_a^2 are same since it's just relative difference of standard variable.

'let's see lecture note's model E_2 vows. coef = 147.825 and this means E_3 have (147.825) less than E_2 kind of reverse effects.

above fitted our model E_2 coef = 3144.04 and this means perspective of E_1 (LHS), E_2 have (3144.04) more than E_1.

So coefficients are different. And in lecture note, E_2 is not significant but our fitted model, E_2 is significant since p-value < 0.05.

(2)

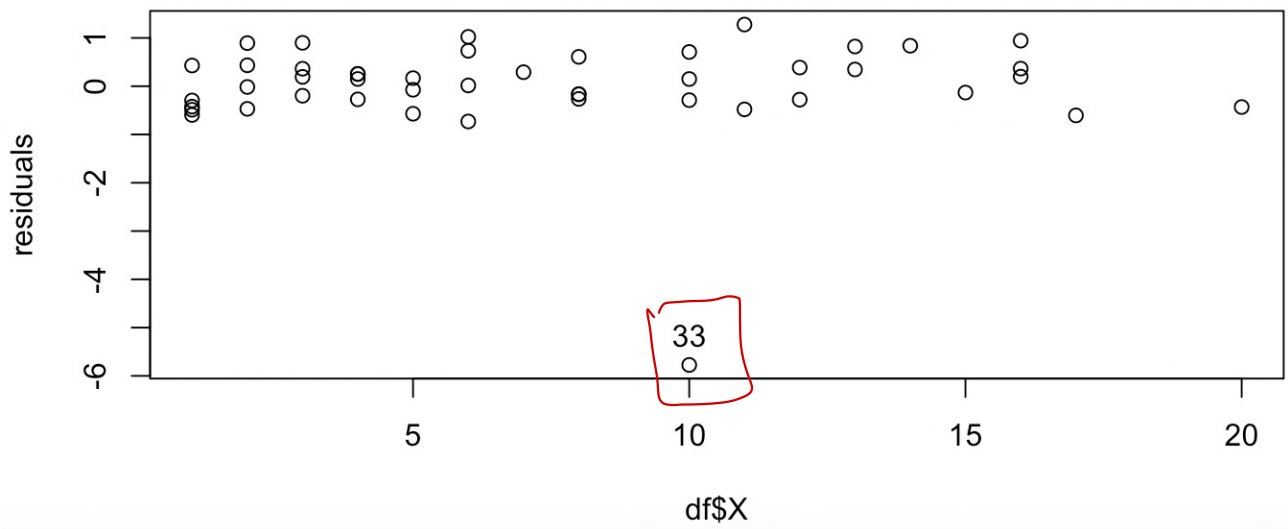
```
> lm2 <- lm(S ~ X + E + M + E*M, data=df)
> summary(lm2)

Call:
lm(formula = S ~ X + E + M + E * M, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-928.13  -46.21   24.33   65.88  204.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9472.685     80.344  117.90   <2e-16 ***
X            496.987      5.566   89.28   <2e-16 ***
E2          1381.671     77.319   17.87   <2e-16 ***
E3          1730.748    105.334   16.43   <2e-16 ***
M1          3981.377    101.175   39.35   <2e-16 ***
E2:M1       4902.523    131.359   37.32   <2e-16 ***
E3:M1       3066.035    149.330   20.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 173.8 on 39 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic:  5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

significant

```
> residuals <- rstandard(lm2)
> plot(df$X, residuals)
> identify(df$X, residuals)
[1] 33
>
```

Plots | Packages | Help | Viewer

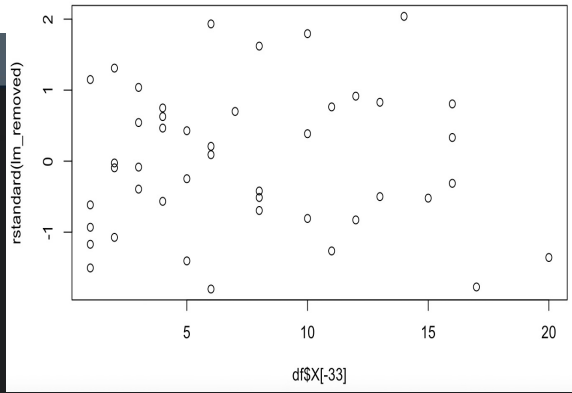
Zoom | Export | Publish



This plot shows that 33th point is outlier.

(3)

```
R 4.1.2 · ~/
> lm_removed <- lm(lm2, subset = -c(33))
> summary(lm_removed)

Call:
lm(formula = lm2, subset = -c(33))

Residuals:
     Min       1Q   Median       3Q      Max
-112.884  -43.636   -5.036   46.622  128.480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9458.378     31.041  304.71   <2e-16 ***
X            498.418      2.152  231.64   <2e-16 ***
E2          1384.294     29.858   46.36   <2e-16 ***
E3          1741.336     40.683   42.80   <2e-16 ***
M1          3988.817     39.073  102.08   <2e-16 ***
E2:M1       5049.294     51.668   97.73   <2e-16 ***
E3:M1       3051.763     57.674   52.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.12 on 38 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.543e+04 on 6 and 38 DF,  p-value: < 2.2e-16
```



Let lecture notes model = A1, our fitted model = A2.
A1, A2 both models all coefficients are significant.
So we can conclude 33th point is may not influential point.
And compare(A1, A2) X columns are same, E2, M, E2·M columns are different. Since difference of perspective. A1 set E3 as standard variable, A2 set E1 as standard variable. In conclusion, it's difference with absolute value, but relatively same (we can adjust) since it's difference of perspective. d.f, R^2, R_a^2, $\hat{\sigma}$ are same.

(4) $\beta_0 = 9456.378$, $\beta_1 = 1344.244$, $\beta_2 = 1741.336$
$\delta = 3986.817$, $\alpha_1 = 5649.244$, $\alpha_2 = 3051.763$

$y_0 \in \hat{y}_0 \pm t(n-2, \frac{\alpha}{2}) \times s.e.(y_0 - \hat{y}_0)$, where $s.e.(\hat{y}_0) = \hat{\sigma}\sqrt{1 + x_0'(X'X)^{-1}x_0}$

category	E	M	Coefficients	Estimate of Base salary	S.E.	95% prediction interval
1	1	0	β_0	9456.378	31.04072	9308.676, 9608.08
2	1	1	$\beta_0+\delta$	13447.195	31.74366	13296.890, 13597.50
3	2	0	$\beta_0+\gamma_1$	10842.672	26.15706	10696.843, 10988.50
4	2	1	$\beta_0+\gamma_1+\delta+\alpha_1$	19880.782	32.94428	19729.422, 20032.14
5	3	0	$\beta_0+\gamma_2$	11199.714	30.53338	11050.440, 11348.99
6	3	1	$\beta_0+\gamma_2+\delta+\alpha_2$	18240.294	28.54706	18092.640, 18387.95

```
> new.data <- data.frame(X=rep(0,6), E=c("1","1","2","2","3","3"), M=c
("0","1","0","1","0","1"))
> predict(lm_removed, new.data, se.fit=TRUE, interval="prediction")
$fit
        fit       lwr      upr
1  9458.378  9308.676  9608.08
2 13447.195 13296.890 13597.50
3 10842.672 10696.843 10988.50
4 19880.782 19729.422 20032.14
5 11199.714 11050.440 11348.99
6 18240.294 18092.640 18387.95

$se.fit
       1        2        3        4        5        6
31.04072 31.74366 26.15706 32.94428 30.53338 28.54706

$df
[1] 38

$residual.scale
[1] 67.11893
```