西南大学 2022: STATS 201 Assignment 3

Runze liao 222020321102007

2022/5/31

Question 1

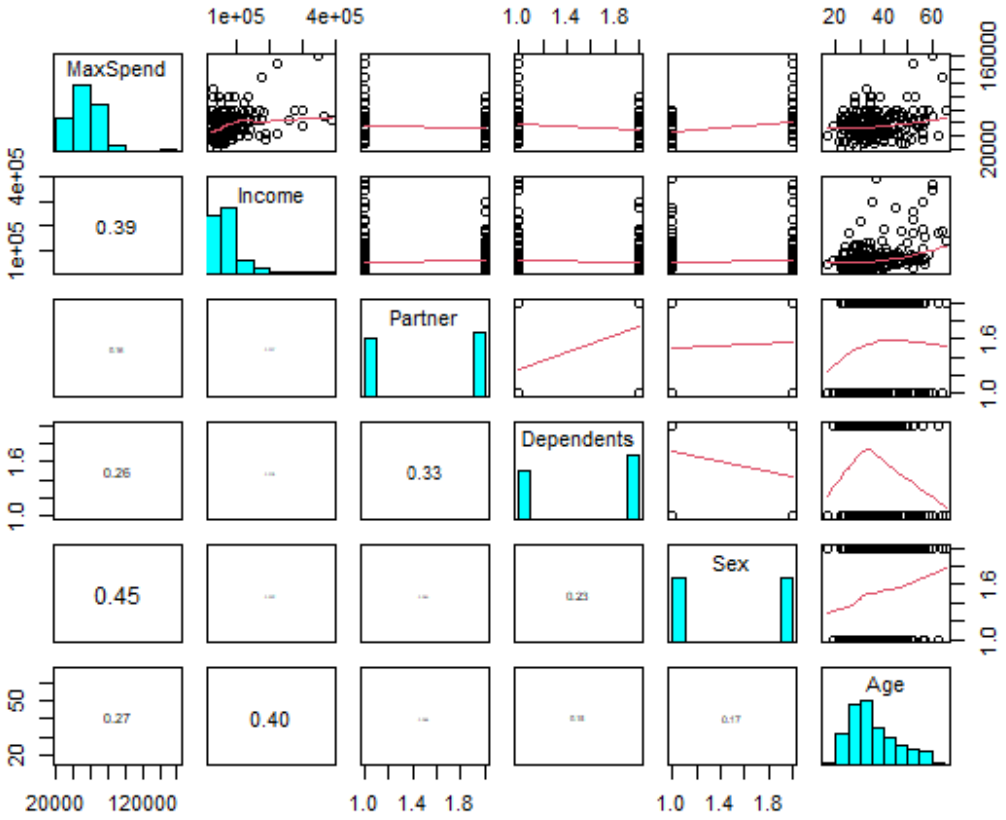
Question of interest/goal of the study

A leading car distributor invited visitors to its website to complete a survey to learn about how much they were willing to spend on a new car. It was of interest to see how this depended on the participant's annual income, marital status, dependents, gender and age. The variables in `CarSpend.txt` are:

- `MaxSpend`: Maximum participant will spend on a new car ($)
- `Income`: Annual income ($)
- `Partner`: 1=in a partnership, 0=single
- `Dependents`: 1=have financial dependents, 0=no financial dependents
- `Sex`: M or F
- `Age`: Age (years)

Read in and inspect the data using a pairs20x plot

```
pairs20x(CarSpend.df[,c(1,2,3,4,5,6)])
```



Comment on the pairs20x plot

From the pairs20x plot we can see that the relation between the Partner, Dependents and Sex is quiet linear, since it is the factor variable. However, with the Income and Age, they are not having strong relationship between them. And the explanatory variables seems have little interaction between each other.

Why log the responses?

In the analysis below you will log the response variable. Provide at least one reason why this is a sensible thing to do

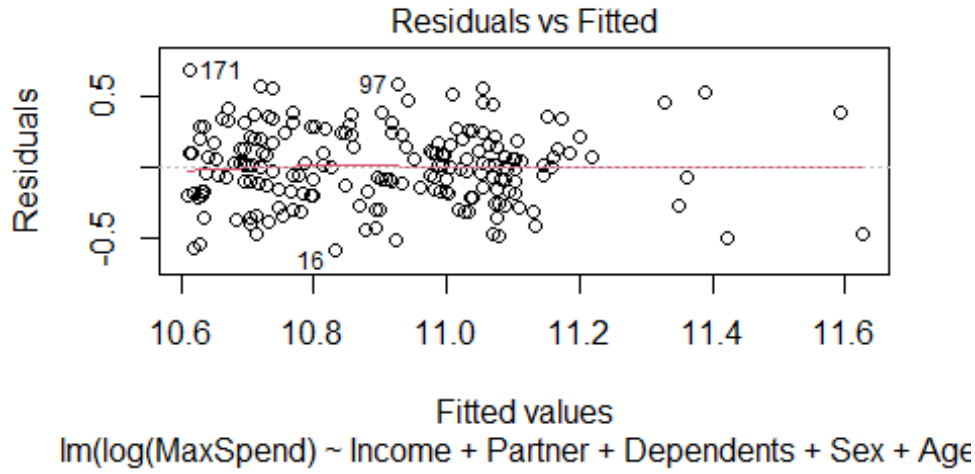
With the pairs20x plot, we can indicate that it is used multiplicative linear model to fit a nice model. As needs to fit the linear model, we need log transformation to turn the multiplicative effects in to additive effect. So we choose to log the responses variable. And the data is left hand side much more.

Fit model and check assumptions

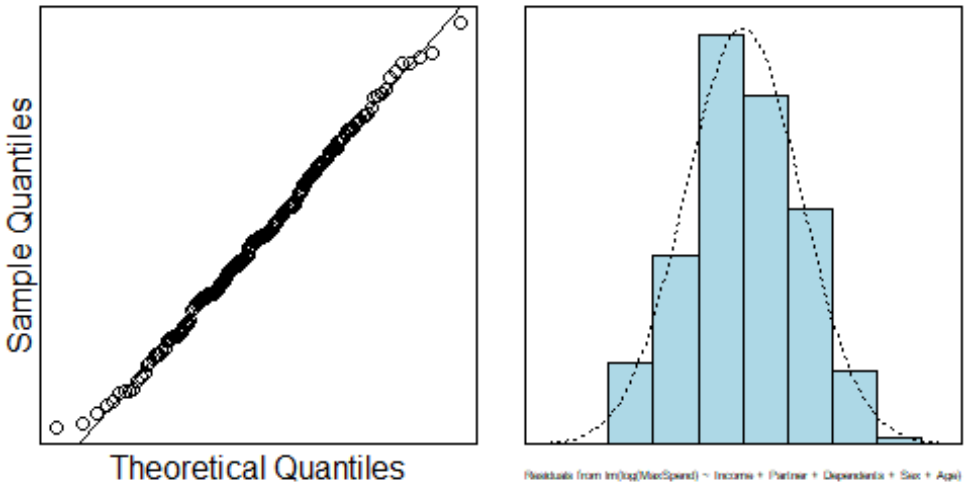
```
CarSpend.fit = lm(MaxSpend~Income+Partner+Dependents+Sex+Age, data = CarSpend.df)
```

we can see that the residual plot is quiet strange.

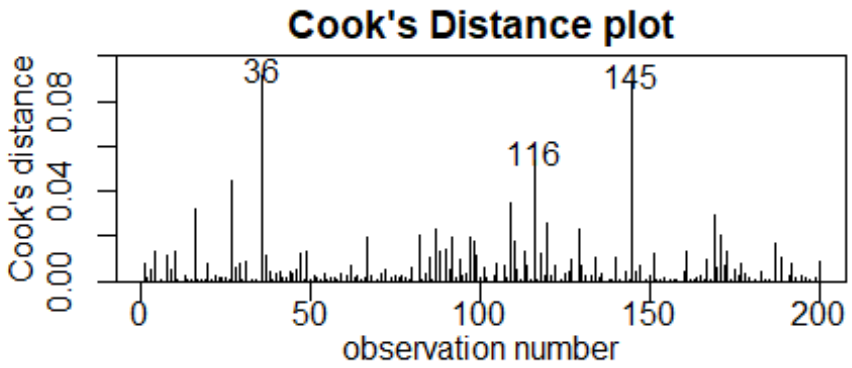
```
CarSpend.fit2 = lm(log(MaxSpend)~Income+Partner+Dependents+Sex+Age, dat
a = CarSpend.df)
plot(CarSpend.fit2 , which = 1)
```




```
normcheck(CarSpend.fit2)
```




```
cooks20x(CarSpend.fit2)
```



From the residual plot, we can see that we've been satisfied the eov assumption, the normcheck is fine, No too strong influence point, however, point 36, 145 seems strange, but we will keep it. All assumption were satisfied, let us see the summary part.

```
summary(CarSpend.fit2)

##
## Call:
## lm(formula = log(MaxSpend) ~ Income + Partner + Dependents +
##     Sex + Age, data = CarSpend.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59125 -0.17113 -0.00126  0.17831  0.67695
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.069e+01  7.581e-02 140.950  < 2e-16 ***
## Income       1.620e-06  3.357e-07   4.826 2.81e-06 ***
## Partner1    -8.225e-02  3.994e-02  -2.059   0.0408 *
## Dependents1 -7.246e-02  4.140e-02  -1.750   0.0817 .
## SexM         2.883e-01  3.865e-02   7.460 2.83e-12 ***
## Age          9.817e-04  2.010e-03   0.488   0.6258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.261 on 194 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:  0.3587
## F-statistic: 23.26 on 5 and 194 DF,  p-value: < 2.2e-16
```

By applying the Occam's Razor, we choose to remove the coefficients that are out of significant at the 5% level(those p-value > 0.05), which means we will remove the Dependents, Age. keeping the Income and Partner, Sex to fit a latest model.

```
CarSpend.fit3 = lm(log(MaxSpend)~Income+Partner+Sex, data = CarSpend.df)
summary(CarSpend.fit3)
```

```
##
## Call:
## lm(formula = log(MaxSpend) ~ Income + Partner + Sex, data = CarSpend.
df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59644 -0.17584 -0.00019  0.16949  0.66728
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.067e+01  3.970e-02 268.850  < 2e-16 ***
## Income       1.719e-06  3.086e-07   5.571 8.28e-08 ***
## Partner1    -1.049e-01  3.720e-02  -2.820  0.00529 **
## SexM         3.090e-01  3.719e-02   8.308 1.59e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2621 on 196 degrees of freedom
## Multiple R-squared:  0.3632, Adjusted R-squared:  0.3534
## F-statistic: 37.26 on 3 and 196 DF,  p-value: < 2.2e-16
```

All the assumptions seem to be satisfied, we have evidence to keep all the coefficients(p-value < 0.05), it is a good fit model, we can trust our final model. **Fit a linear regression model for log(MaxSpend) that contains the five explanatory terms log(Income), Partner, Dependents, Sex, Age. Then, apply Occam's Razor – that is, simplify the model by successively removing the least significant term until all are significant at the 5% level.**

```
exp(confint(CarSpend.fit3))
```

```
##                    2.5 %       97.5 %
## (Intercept) 3.996638e+04 4.674175e+04
## Income      1.000001e+00 1.000002e+00
## Partner1    8.367306e-01 9.689525e-01
## SexM        1.265721e+00 1.465719e+00
```

Method and Assumption Checks

By having looked at the Pair20 plot, we got that the Max Spend in car were related to serveral explanatory variables, So we construct a multiple linear regression model with a suitable selection of the explanatory variables, moreover, we choose to

log the responses variable by having the transformation turn the multiplicative effects in to additive effect.

Furthermore, we decide to keep the Income and Partner, Sex to fit a latest model, deleting the the Dependents, Age, because they were out of the 95% confidence interval.

So our final model is:

$$MaxSpend_i = \beta_0 + \beta_1 \times Income_i + \beta_2 \times Partner_i + \beta_3 \times Sex_i + \epsilon_i$$

where $\epsilon_i \sim iid.N(0, \sigma^2)$. Here our indicator variable takes value 1 if the Sex is Male.

Our model explains about 36.3% of the variability in people's max spend in cars.

Executive Summary

We wanted to have a model to explain how much visitors want to spend on a new car depending on the income, marital status, dependents, gender and age.
We have estimated that:

- The Male would choose to spend 1.26 to 1.46 than the Female on a new car.
- People who got married are more likely to spend 0.84 to 0.97 than those who not.
- For each one more in thier Income, we estmated that they would like spending incresing 1 than before.

Question 2

Question of interest/goal of the study

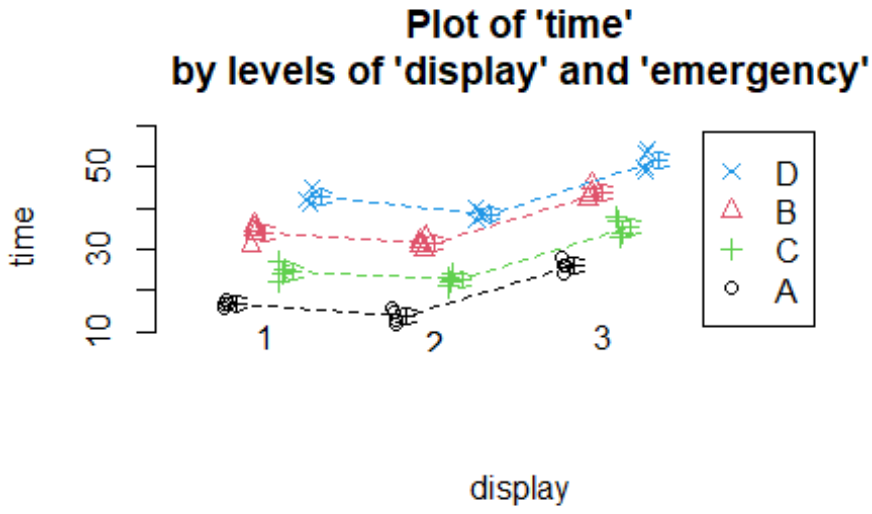
A company was interested in assessing 3 different display panels used by air traffic controllers. An experiment was conducted by simulating 4 different emergency conditions with 4 qualified air traffic controllers randomly assigned to each display/emergency combination. The time (in seconds) required to stabilise the emergency condition was recorded. The data are stored in the text file `airtraffic.txt`, which contains the variables:

- `time`: the time required to stabilise the emergency condition (seconds)
- `display`: the display panel: 1, 2 or 3
- `emergency`: the simulated emergency condition: A, B, C or D

We are **only interested in which display has the lowest time** to stabilise the emergencies, by how much lower it is than the other displays and whether answer this depends on the type of emergency simulated. You should *NOT* quantify extraneous information.

Read in and plot the data

```
airtraffic.df = read.table(file = "airtraffic.txt", header = TRUE)
airtraffic.df$display = as.factor(airtraffic.df$display)
airtraffic.df$emergency = as.factor(airtraffic.df$emergency)
interactionPlots(time~display+emergency, data = airtraffic.df)
```



Comment on the plot

By looking at the interaction plot of display and emergency, we see those parallel lines, which indicating that the two explanatory variables have no interaction. And

on average we can see that with the same emergency, time of display 3 > 1 > 2, and with the same display, time of emergency D > B > C > A.

Fit model, check assumptions and do inference (CIs etc)

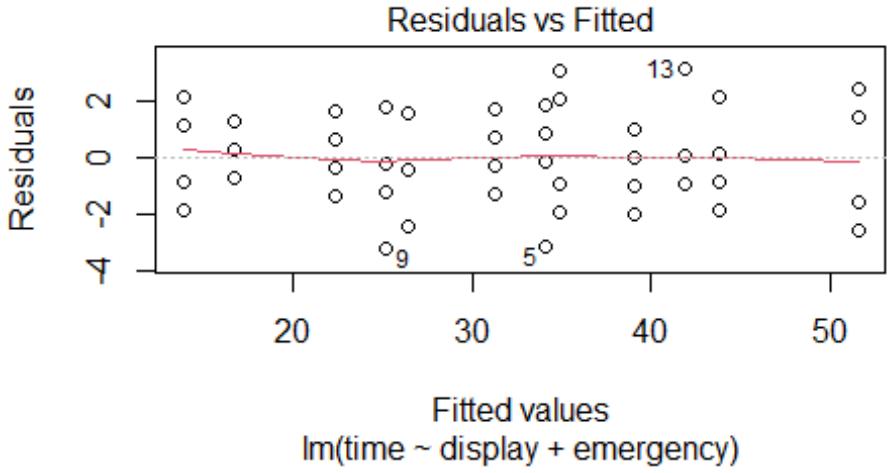
```
airtraffic.fit_inter = lm(time~display*emergency, data = airtraffic.df)
summary(airtraffic.fit_inter)
```

```
##
## Call:
## lm(formula = time ~ display * emergency, data = airtraffic.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.0   -1.5    0.0    1.5    2.5
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.700e+01  8.888e-01  19.127  < 2e-16 ***
## display2            -3.000e+00  1.257e+00  -2.387   0.0224 *
## display3             9.000e+00  1.257e+00   7.160 2.02e-08 ***
## emergencyB           1.700e+01  1.257e+00  13.525 1.11e-15 ***
## emergencyC           7.500e+00  1.257e+00   5.967 7.69e-07 ***
## emergencyD           2.550e+01  1.257e+00  20.288  < 2e-16 ***
## display2:emergencyB  5.000e-01  1.778e+00   0.281   0.7801
## display3:emergencyB  7.500e-01  1.778e+00   0.422   0.6756
## display2:emergencyC  1.000e+00  1.778e+00   0.563   0.5772
## display3:emergencyC  2.000e+00  1.778e+00   1.125   0.2680
## display2:emergencyD -1.000e+00  1.778e+00  -0.563   0.5772
## display3:emergencyD  1.188e-14  1.778e+00   0.000   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.778 on 36 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9743
## F-statistic: 163.2 on 11 and 36 DF,  p-value: < 2.2e-16
```

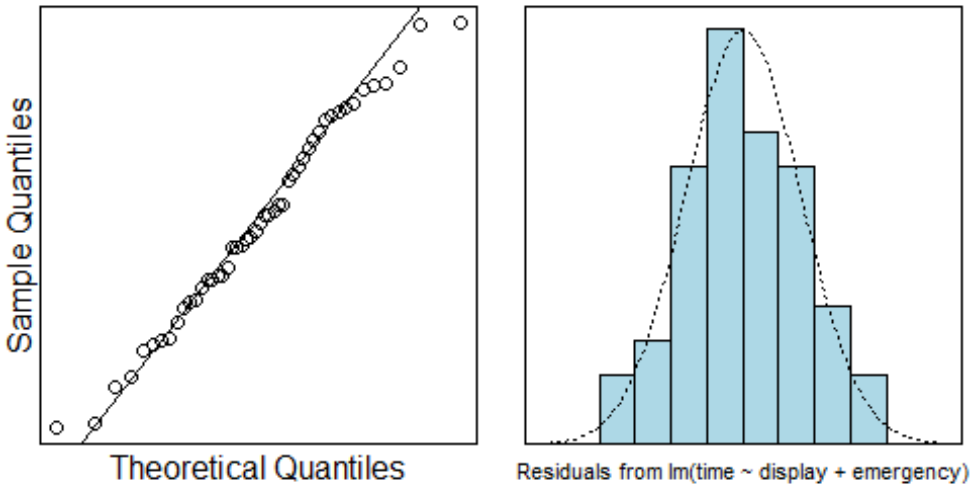
From the summary with interaction we can see that they have no interaction becasue the the Coefficients of interaction part are out of the 95% confidence interval.

So we choose to fit a model with no interaction and a two-ANOVA model.

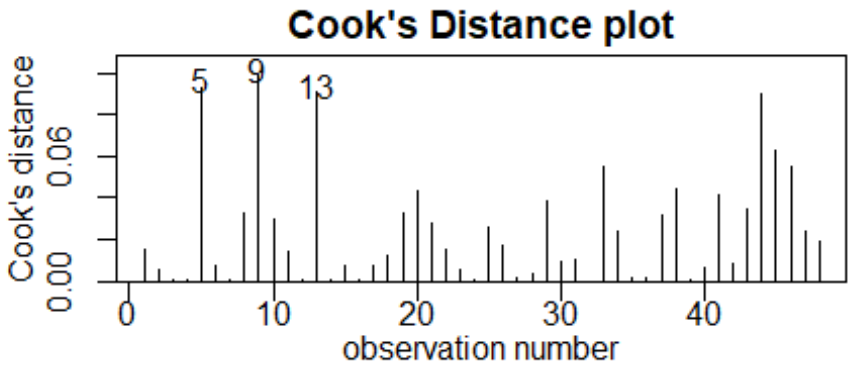
```
airtraffic.fit = lm(time~display+emergency, data = airtraffic.df)

plot(airtraffic.fit,which = 1)
```




```
normcheck(airtraffic.fit)
```




```
cooks20x(airtraffic.fit)
```



A little bit strange in the residual plot, and the normal check is strange too, however, we will tolerant it. From the cooks plot, no strong influence point, it seems it is a good model, and satisfy most of the assumptions. We can trust our model.

```
summary(airtraffic.fit)

##
## Call:
## lm(formula = time ~ display + emergency, data = airtraffic.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2292 -1.0729 -0.0833  1.3073  3.1042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.7292     0.6008  27.844  < 2e-16 ***
## display2     -2.8750     0.6008  -4.785 2.13e-05 ***
## display3      9.6875     0.6008  16.124  < 2e-16 ***
## emergencyB   17.4167     0.6938  25.104  < 2e-16 ***
## emergencyC    8.5000     0.6938  12.252 1.87e-15 ***
## emergencyD   25.1667     0.6938  36.275  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.699 on 42 degrees of freedom
## Multiple R-squared:  0.979,  Adjusted R-squared:  0.9765
## F-statistic: 392.3 on 5 and 42 DF,  p-value: < 2.2e-16
```

All Coefficients are in 95% confidence interval(p-value < 0.05), it seems that it is a nice model we can trust.

```
confint(airtraffic.fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 15.516659 17.941675
## display2    -4.087508 -1.662492
## display3     8.474992 10.900008
## emergencyB  16.016583 18.816750
## emergencyC   7.099916  9.900084
## emergencyD  23.766583 26.566750
```

```
#airtraffic.emmeans = emmeans(airtraffic.fit, specs = display~emergency)
summary2way(airtraffic.fit, page = "nointeraction")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fit)
##
## $display
##        diff       lwr       upr    p adj
## 2-1 -2.8750 -4.334693 -1.415307 6.24e-05
## 3-1  9.6875  8.227807 11.147193 0.00e+00
## 3-2 12.5625 11.102807 14.022193 0.00e+00
##
## $emergency
##          diff        lwr       upr p adj
## B-A 17.416667  15.560863 19.272470     0
## C-A  8.500000   6.644196 10.355804     0
## D-A 25.166667  23.310863 27.022470     0
## C-B -8.916667 -10.772470 -7.060863     0
## D-B  7.750000   5.894196  9.605804     0
## D-C 16.666667  14.810863 18.522470     0
```

Method and Assumption Checks

In this case, we have 2 explanatory factors variable, namely the display and emergency, the display can be 1 or 2 or 3, the emergency could be A, B and C and D, first we fit a two-way ANOVA model with interaction, however, the coefficient of the interaction part is out of the 95% CI, so finally we fitted a two-way ANOVA model with no interaction between the display and emergency.

Through the interaction plot we can see that the two variables have no interaction(it is parallel), so we build a linear model without interaction by having tow factors explanatory variables, the EOV check is not so good, the Normal Check is not so good, and no other influence points, nearly all the assumptions were satisfied by our final model.

Our model explains about 97.9% of the variability in the time required to stabilise the emergency condition.

Executive Summary

We were eager to have a model to explain the time required to stabilise the emergency condition influenced by its display and its emergency.

We found that the effect that the time required to stabilise the emergency condition depends on display and emergency, and they've got no interaction, so we can see them individually.

We estimate that:

- With the same emergency, the time of using display 2 is on average less 1.41 between 4.33 than using display 1.
- With the same emergency, the time of using display 2 is on average less 11.10 between 14.02 than using display 3.

So display 2 has the lowest time to stabilize the emergencies, and this does not depend on the type of emergency simulated, because it has no interaction between the each other(the interaction part of coefficient are out of the 95% CI).