Assignment

Nhu Mai Nguyen

5/16/2021

This is my first heading

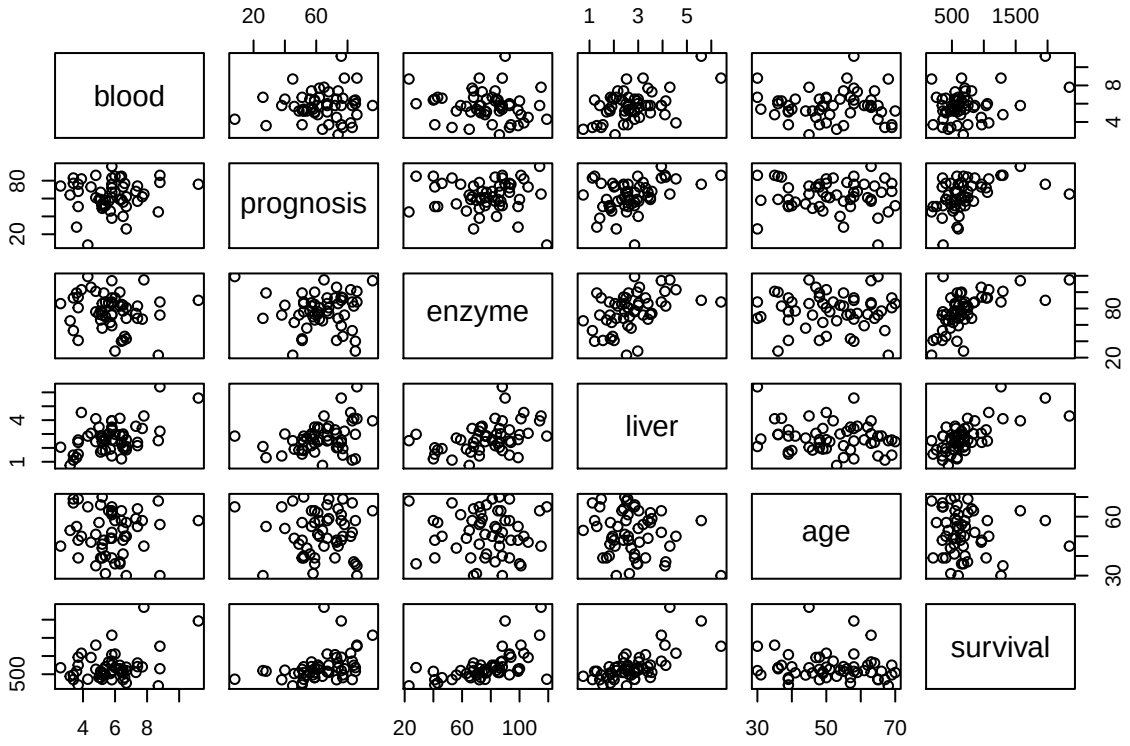
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Question 1

a. Produce a scatter plot

```
sur$gender <- NULL
plot(sur)
```



* Remove gender variable because it is dummy variable. It is IV's and has no role to impact DV's.

b. Compute the correlation matrix

```
cor(sur)
```

```
##                 blood   prognosis      enzyme      liver         age   survival
## blood      1.00000000  0.09011973 -0.14963411  0.5024157 -0.02068803  0.3465497
## prognosis  0.09011973  1.00000000 -0.02360544  0.3690256 -0.04766570  0.4204810
## enzyme    -0.14963411 -0.02360544  1.00000000  0.4164245 -0.01290325  0.5782260
## liver      0.50241567  0.36902563  0.41642451  1.0000000 -0.20737776  0.6741950
## age       -0.02068803 -0.04766570 -0.01290325 -0.2073778  1.00000000 -0.1191715
## survival   0.34654968  0.42048097  0.57822600  0.6741950 -0.11917146  1.0000000
```

c. Fit a model

- mathematical multiple regression model $Y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + \epsilon$

$Y = dependentvariable - survival$

$beta_0 = intercept$

$x_1 = thefirstIV's - bloodvariable$

$x_2 = thesecondIV's - prognosisvariable$

$x_3 = thethirdIV's - enzymevariable$

$x_4 = thefourthIV's - livervariable$

$x_5 = thefifthIV's - agevariable$

- the intercept is:

$b_0 = \bar{y} - \bar{x_1}b_1 - \bar{x_2}b_2 - ... - \bar{x_n}b_n$

- mathematical hypotheses for the overall ANOVA

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2$

```
sur.lm = lm(survival ~ blood + prognosis + enzyme + liver + age, data = sur)
anova(sur.lm)
```

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152 18.8997 7.133e-05 ***
## prognosis  1 1278496 1278496 24.0393 1.121e-05 ***
## enzyme     1 3442172 3442172 64.7226 1.883e-10 ***
## liver      1   57862   57862  1.0880    0.3021
## age        1   33032   33032  0.6211    0.4345
## Residuals 48 2552807   53183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Compute the F-test and P-value

```
sur.lm = lm(survival ~ blood + prognosis + enzyme + liver + age, data = sur)
summary(sur.lm)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + liver +
##     age, data = sur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.34 -147.74   11.74  124.67  954.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.367    275.619  -4.279 8.91e-05 ***
## blood          86.630     26.905   3.220 0.002302 **
## prognosis       8.501      2.137   3.978 0.000234 ***
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## liver          38.554     49.251   0.783 0.437595
## age            -2.340      2.969  -0.788 0.434514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.6 on 48 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6632
## F-statistic: 21.87 on 5 and 48 DF,  p-value: 2.386e-11
```

$F-statistic $= 21.87$ $ on 5 and $48 DF$

$p - value = 2.386e - 11$

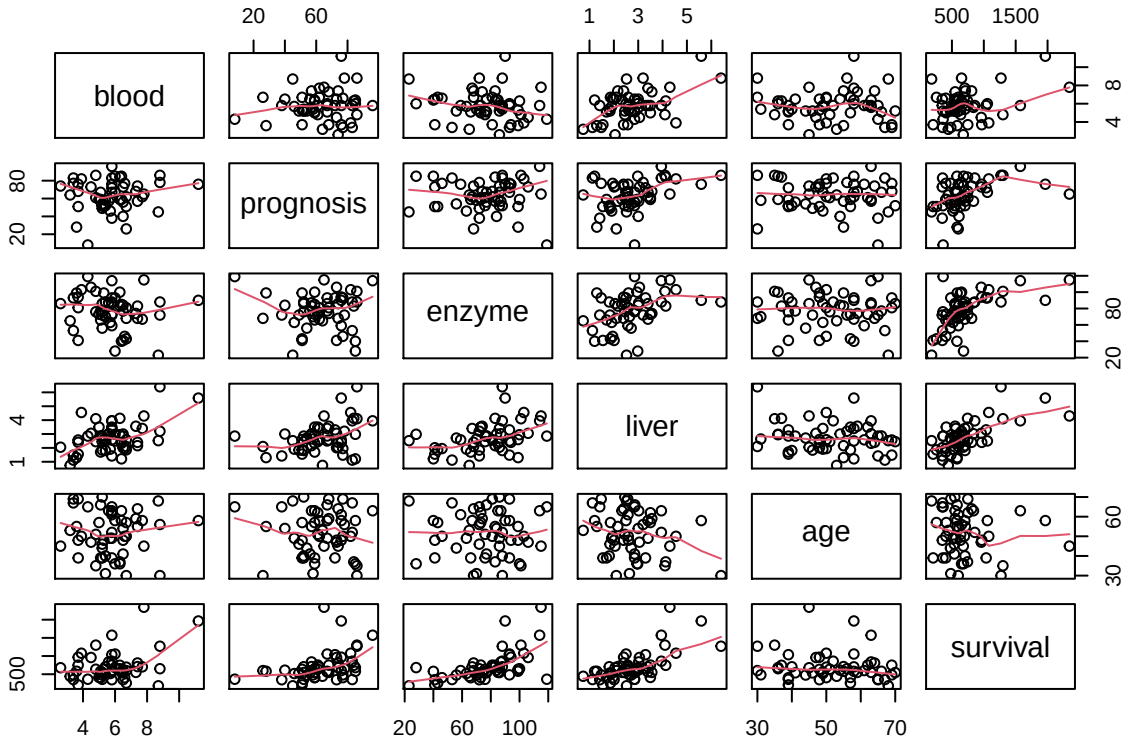
- State the null hypothesis:

$H_0 : \beta_0 + \beta_1 + \beta_2 + ... + \beta_k = 0$

d. Find the best linear regression

- First check the line of plot to see which variable has high correlated

```
pairs(sur, panel = panel.smooth)
```



- Start with all predictors

```
sur.1 = lm(survival ~ . , data = sur)
summary(sur.1)
```

```
##
## Call:
## lm(formula = survival ~ ., data = sur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.34 -147.74   11.74  124.67  954.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.367    275.619  -4.279 8.91e-05 ***
## blood          86.630     26.905   3.220 0.002302 **
## prognosis       8.501      2.137   3.978 0.000234 ***
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## liver          38.554     49.251   0.783 0.437595
## age            -2.340      2.969  -0.788 0.434514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.6 on 48 degrees of freedom
```

```
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6632
## F-statistic: 21.87 on 5 and 48 DF,  p-value: 2.386e-11
```

- After summary the data, we can see the liver with $p-value$ is larger than response, so it is insignificant variable so remove it.

```
sur.2 = lm(survival ~ blood + prognosis + enzyme + age, data = sur)
summary(sur.2)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + age, data = sur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.92 -142.56  -13.98  138.10  943.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1246.655    260.835  -4.779 1.64e-05 ***
## blood         100.660     19.987   5.036 6.83e-06 ***
## prognosis       9.291      1.876   4.951 9.14e-06 ***
## enzyme         12.101      1.502   8.058 1.56e-10 ***
## age            -2.986      2.841  -1.051    0.298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.7 on 49 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6659
## F-statistic: 27.41 on 4 and 49 DF,  p-value: 5.68e-12
```

- After removing the liver variable, we can see the age variable has same problem like liver, it shows up the $p-value$ larger than 0.05 then we need to remove it agains to get the best.

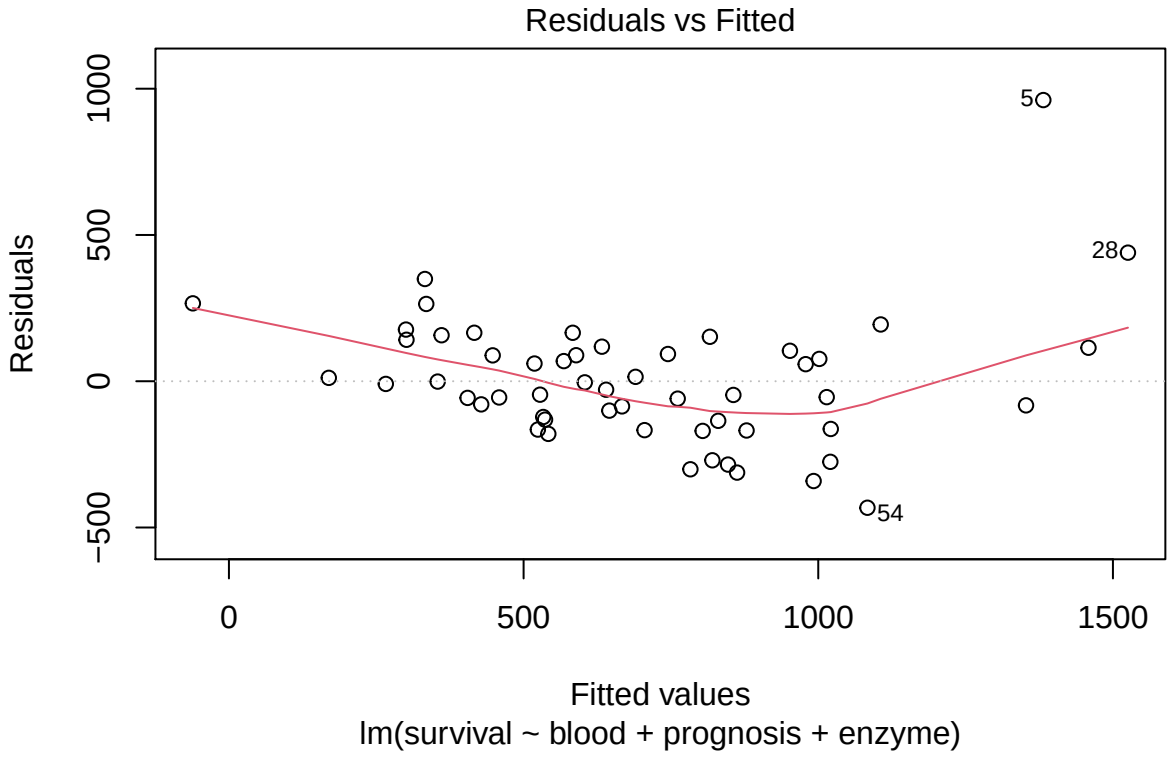
```
sur.3 = lm(survival ~ blood + prognosis + enzyme, data = sur)
summary(sur.3)
```

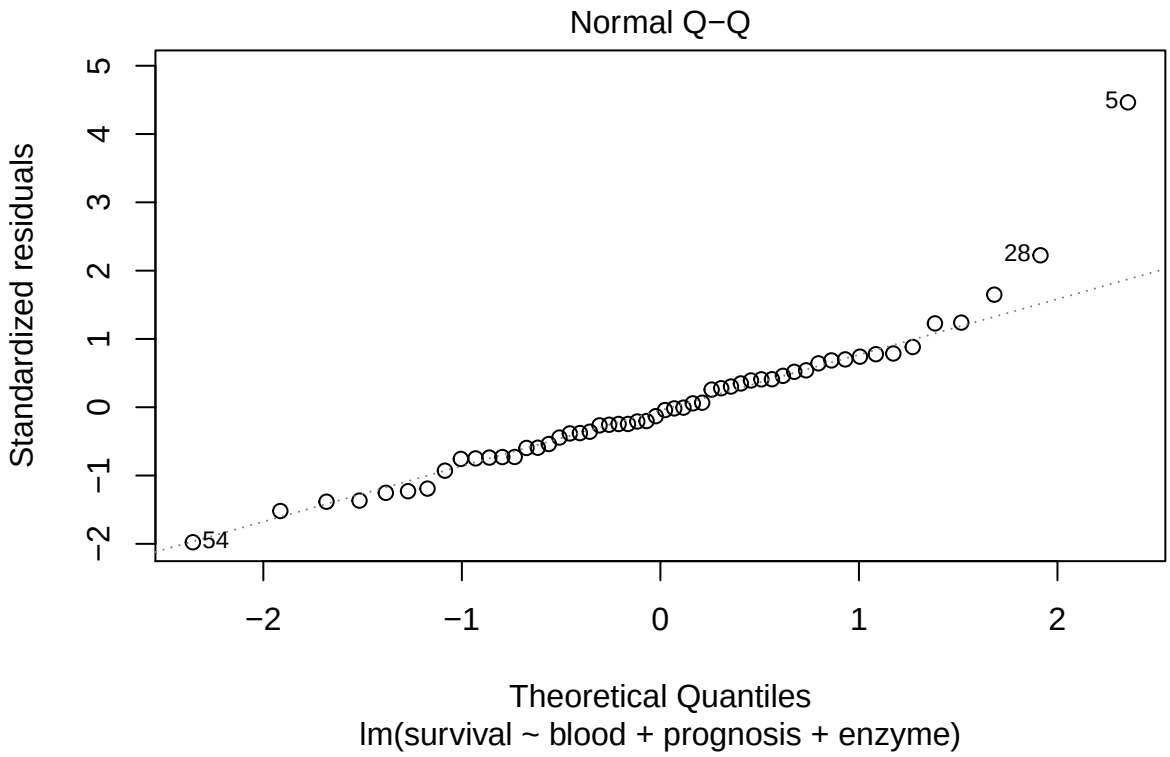
```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme, data = sur)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## blood         101.054     20.005   5.052 6.22e-06 ***
## prognosis       9.382      1.876   5.000 7.43e-06 ***
## enzyme         12.128      1.503   8.069 1.30e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 50 degrees of freedom
## Multiple R-squared:  0.6841, Adjusted R-squared:  0.6652
## F-statistic:  36.1 on 3 and 50 DF,  p-value: 1.469e-12
```

- Check diagnostics using plot

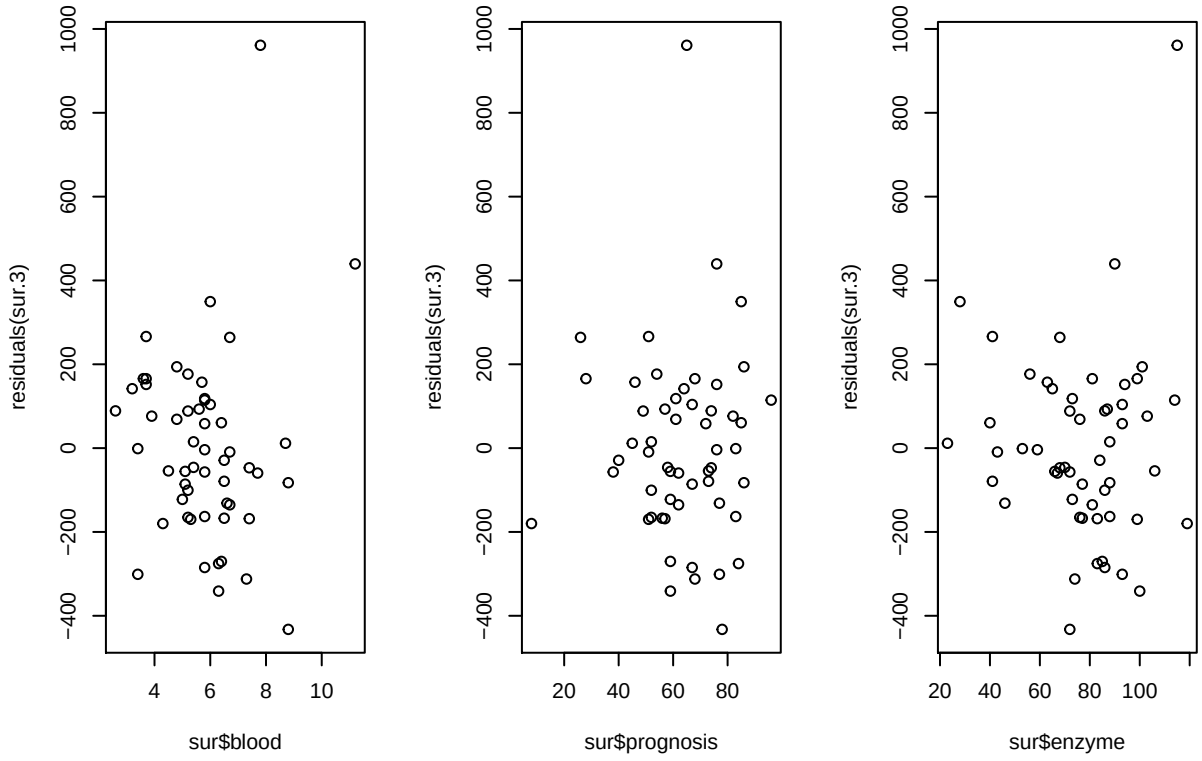
```
plot(sur.3, which = 1:2)
```





- Check residuals against predictors

```
par(mfrow = c(1,3))
plot(sur$blood, residuals(sur.3))
plot(sur$prognosis, residuals(sur.3))
plot(sur$enzyme, residuals(sur.3))
```



- Model interpretation - this part will help to find goodness of fit or the best fit linear

```
summary(sur.3)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1410.846901 209.117946 -6.746656 1.495123e-08
## blood         101.053887  20.004632  5.051525 6.220022e-06
## prognosis       9.381966   1.876399  4.999985 7.433593e-06
## enzyme         12.127807   1.503098  8.068542 1.303361e-10
```

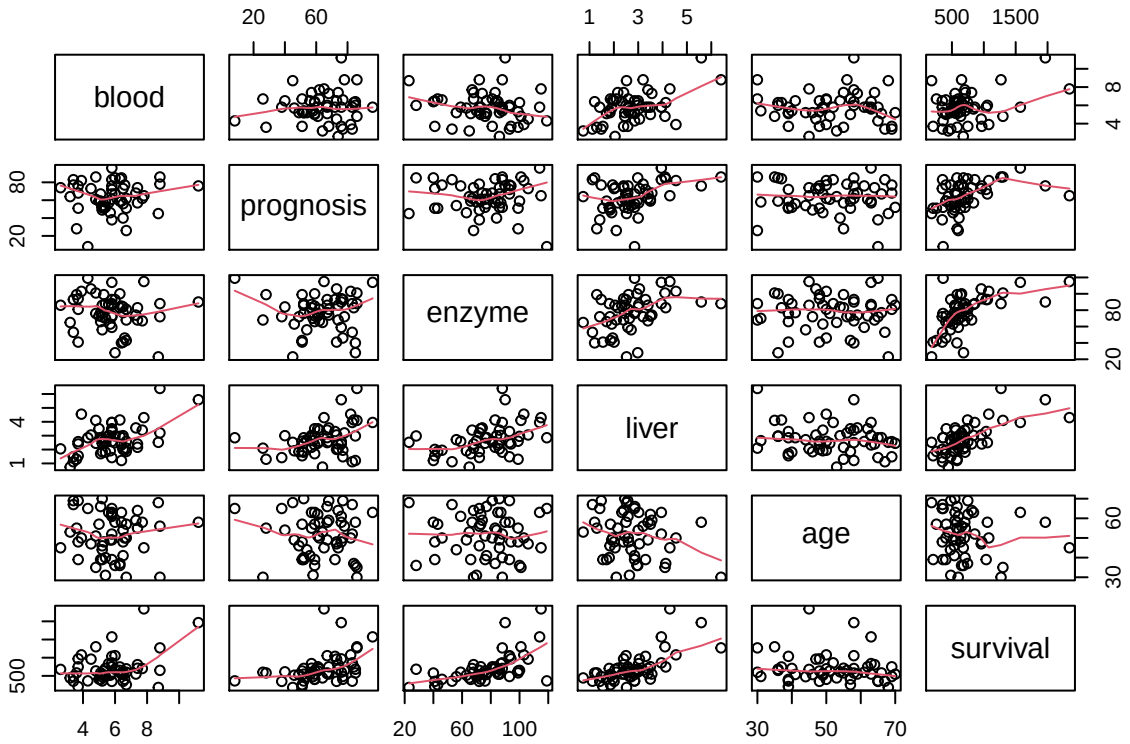
- For a unit increase in blood there is a 101.053 increase in survival time.
- For a unit increase in prognosis there is a 9.281 increase in survival time.
- For a unit increase in enzyme there is a 12.127 increase in survival time.

e.

- It is not appropriate to use the multiple regression model because it has a negative intercept. This means that the expected value on your dependent variable will be less than 0 when all predictors variables are set to 0.

f. Re-fit the model using log(survival)

```
pairs(sur, panel = panel.smooth)
```



- Start with all predictors

```
sur.1 = lm(log(survival) ~ . , data = sur)
summary(sur.1)
```

```
##
## Call:
## lm(formula = log(survival) ~ ., data = sur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3894 -0.1895  0.0045  0.1782  0.5103
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.047579   0.296655  13.644  < 2e-16 ***
## blood        0.090874   0.028958   3.138  0.00291 **
## prognosis    0.012975   0.002300   5.641 8.82e-07 ***
## enzyme       0.016126   0.002107   7.654 7.38e-10 ***
## liver        0.010914   0.053010   0.206  0.83775
## age         -0.004584   0.003196  -1.434  0.15796
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2482 on 48 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.745
## F-statistic: 31.97 on 5 and 48 DF,  p-value: 3.478e-14
```

- After summary the data, liver is insignificant variable so remove it.

```
sur.2 = lm(log(survival) ~ blood + prognosis + enzyme + age, data = sur)
summary(sur.2)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + age,
##     data = sur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39491 -0.18866 -0.00045  0.17491  0.51787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.028531   0.279090  14.434  < 2e-16 ***
## blood        0.094845   0.021386   4.435 5.20e-05 ***
## prognosis    0.013199   0.002008   6.574 3.04e-08 ***
## enzyme       0.016402   0.001607  10.208 1.01e-13 ***
## age         -0.004767   0.003040  -1.568    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2458 on 49 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:   0.75
## F-statistic: 40.74 on 4 and 49 DF,  p-value: 5.171e-15
```

- After removing the liver variable, need to remove age variable to get the best.

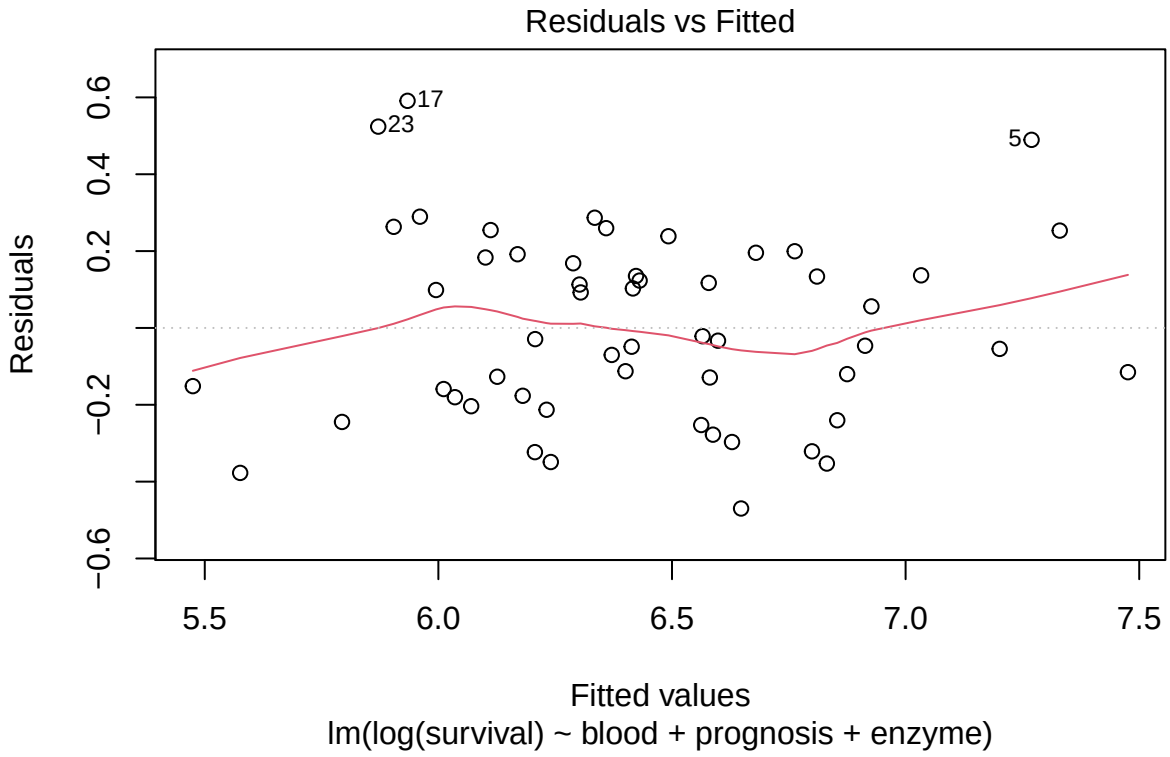
```
sur.3 = lm(log(survival) ~ blood + prognosis + enzyme, data = sur)
summary(sur.3)
```

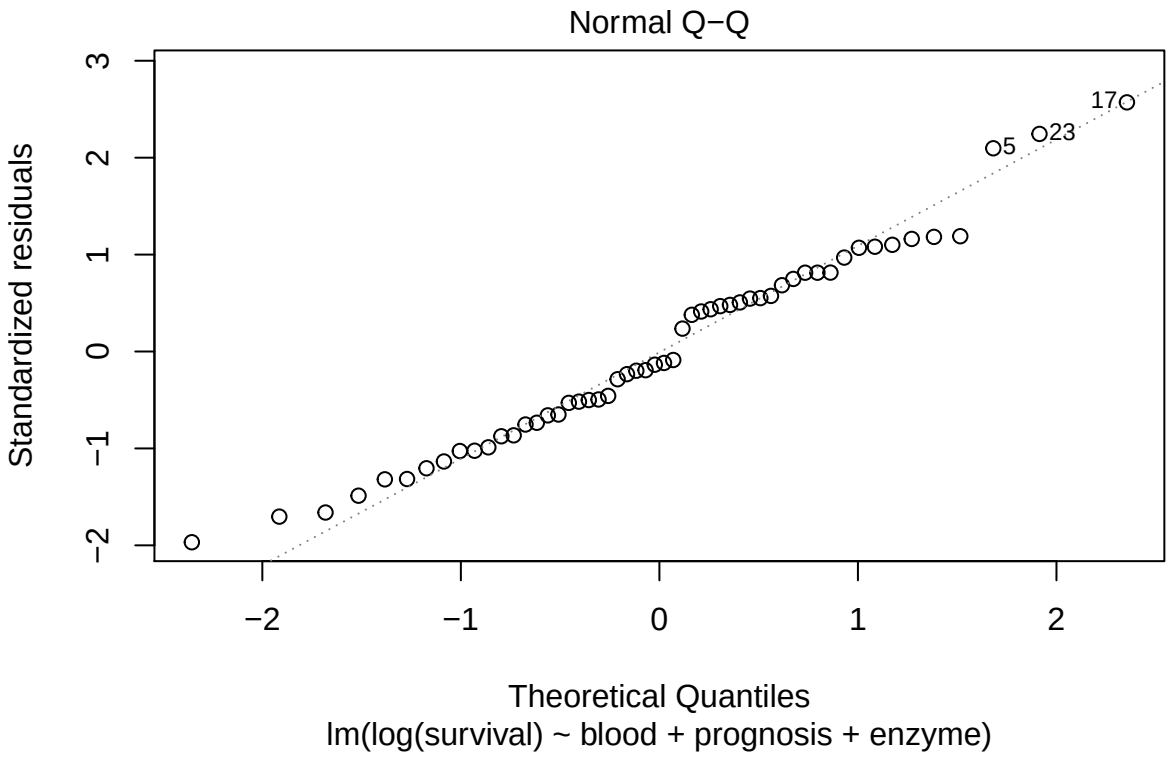
```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme, data = sur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46994 -0.17938 -0.03116  0.17959  0.59105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.766441   0.226757  16.610  < 2e-16 ***
## blood       0.095475   0.021692   4.401 5.66e-05 ***
## prognosis   0.013344   0.002035   6.558 2.95e-08 ***
```

```
## enzyme      0.016444   0.001630  10.089 1.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2493 on 50 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7427
## F-statistic: 51.99 on 3 and 50 DF,  p-value: 2.137e-15
```

- Check diagnostics using plot

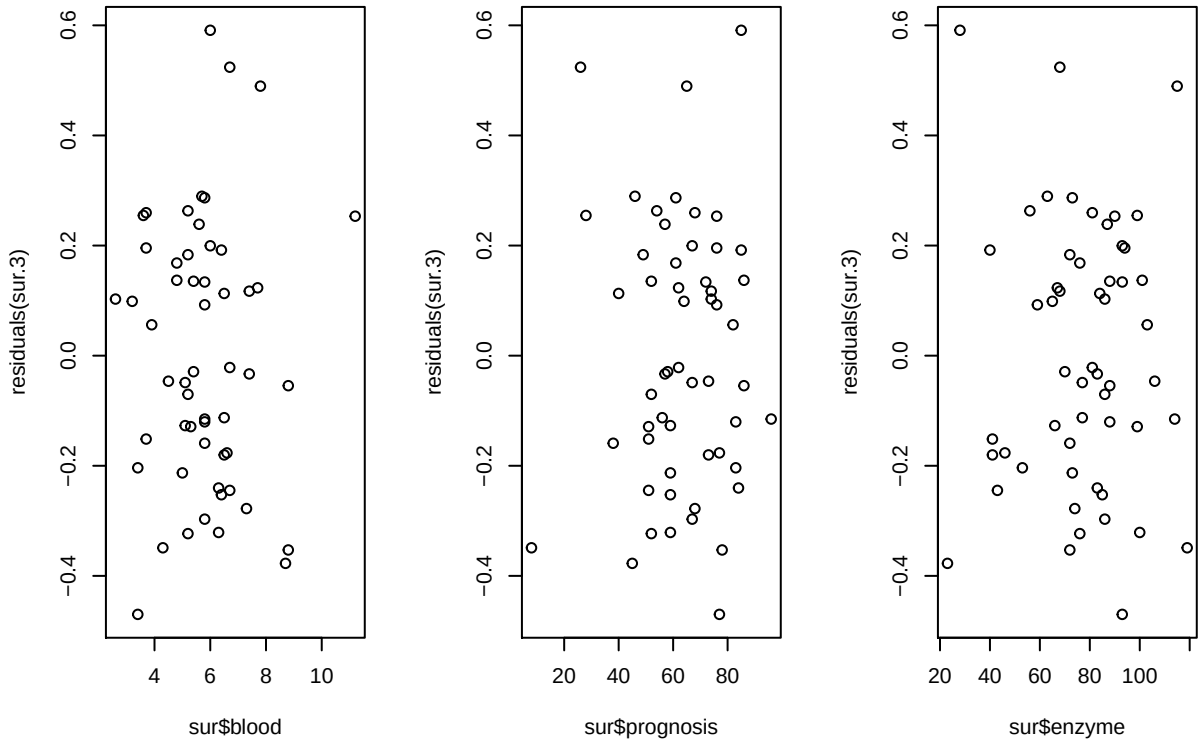
```
plot(sur.3, which = 1:2)
```





- Check residuals against predictors

```
par(mfrow = c(1, 3))
plot(sur$blood, residuals(sur.3))
plot(sur$prognosis, residuals(sur.3))
plot(sur$enzyme, residuals(sur.3))
```



- Model interpretation - this part will help to find goodness of fit or the best fit.

```
summary(sur.3)$coefficients
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 3.76644097 0.226757297 16.610010 5.399369e-22
## blood       0.09547451 0.021692046  4.401360 5.655790e-05
## prognosis   0.01334404 0.002034675  6.558313 2.946869e-08
## enzyme      0.01644450 0.001629886 10.089356 1.190806e-13
```

g. Explain the function of log to find the linear

- Log is a con venient means of transforming a highly skewed variable into a more normalized dataset.
- When modeling variables having non-linear realationships, the risks of making mistakes are increased.
- By changing the distribution of the feature to a more normally-shaped bell curve, using the logarithm of one or more variable enhances the model's fit.

###Question 2:

a.

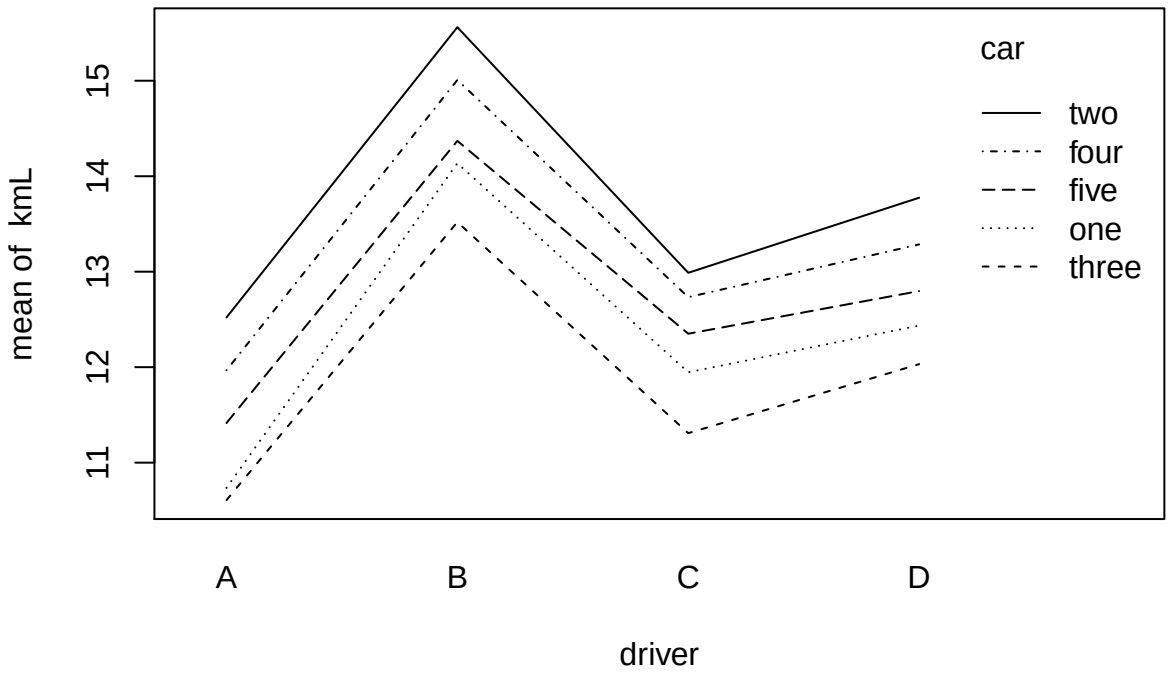
```
with(kml, table(driver, car))
```

```
##       car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

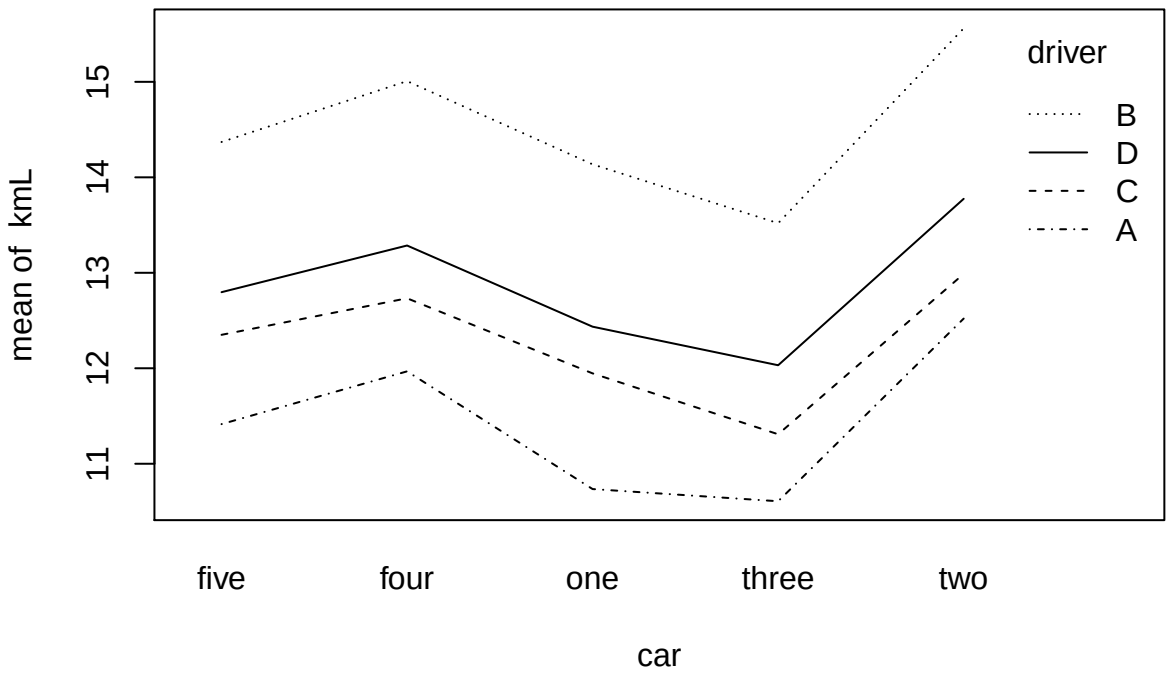
- The design of the study is balanced. Because the replicates are available for all factors' pairs

b. Prliminary investigation

```
with(kml, interaction.plot(driver, car, kmL))
```

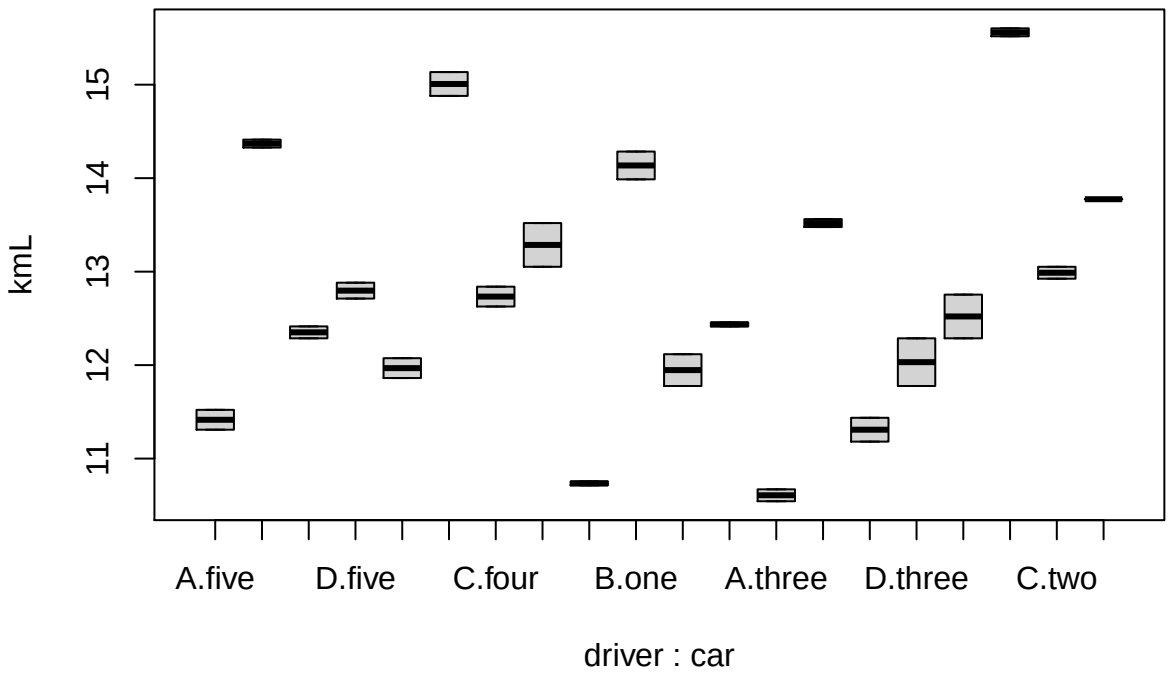



```
with(kml, interaction.plot(car, driver, kmL))
```



- Result: The lines are parallel, so there is no interaction between the two factors.
- Boxplot:

```
boxplot(kmL ~ driver + car, data = kml)
```



c. Group test

```
kml.int = lm(kmL ~ factor(car) * factor(driver), data = kml)
anova(kml.int)
```

```
## Analysis of Variance Table
##
## Response: kmL
##                            Df Sum Sq Mean Sq F value    Pr(>F)
## factor(car)                 4 17.119  4.2798  134.73 3.664e-14 ***
## factor(driver)              3 50.661 16.8869  531.60 < 2.2e-16 ***
## factor(car):factor(driver) 12  0.442  0.0368    1.16    0.3715
## Residuals                  20  0.635  0.0318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fit reduced model without interaction (only main effects)

```
kml.reduced = update(kml.int, . ~ . - car:driver)
```

- ANOVA table for the model with interaction

```
anova(kml.int)
```

```
## Analysis of Variance Table
##
## Response: kmL
##                             Df Sum Sq Mean Sq F value    Pr(>F)
## factor(car)                  4 17.119  4.2798  134.73 3.664e-14 ***
## factor(driver)               3 50.661 16.8869  531.60 < 2.2e-16 ***
## factor(car):factor(driver) 12  0.442  0.0368    1.16    0.3715
## Residuals                  20  0.635  0.0318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

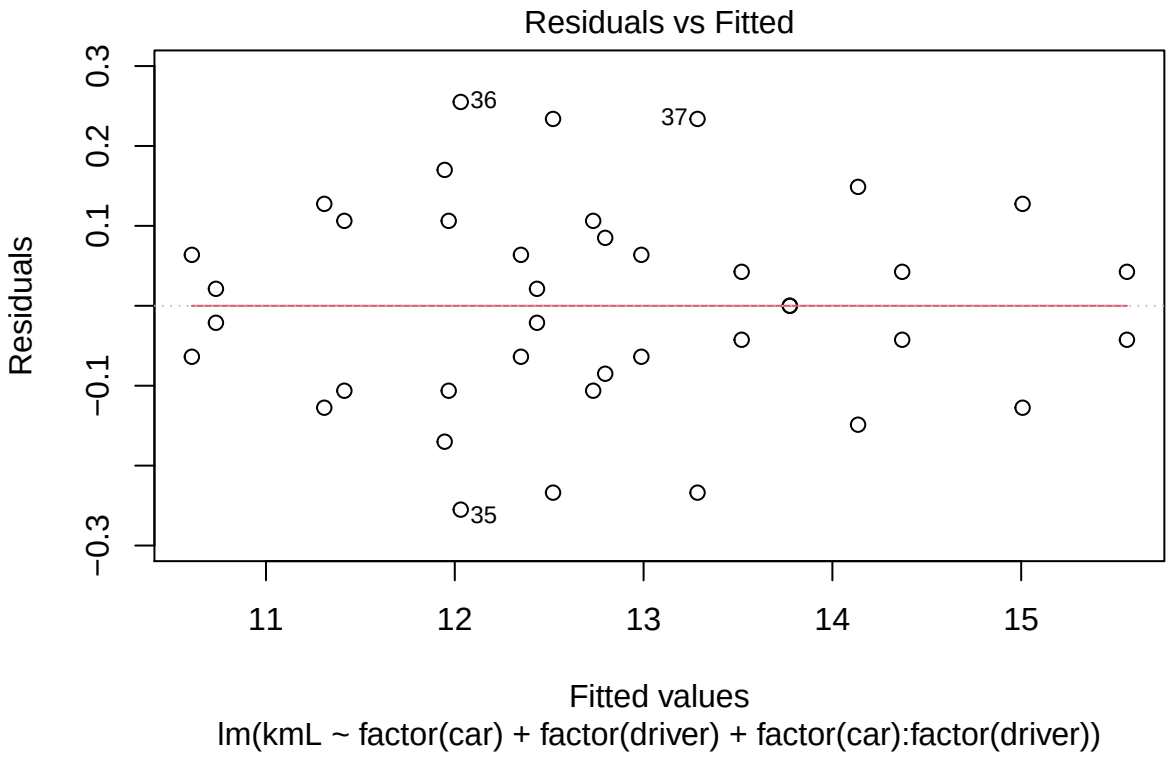
- Model $Y = \mu + \alpha_i + \beta_j + \gamma_i j + \epsilon$
  Hypotheses: $H0 : \gamma_i j = 0$, H1 : at least one $\gamma_i j$ non-zero
  $P - Value = 0.3715 > 0.05$
- The interaction is not significant fit reduced model with main effects only
- Fit model without interaction

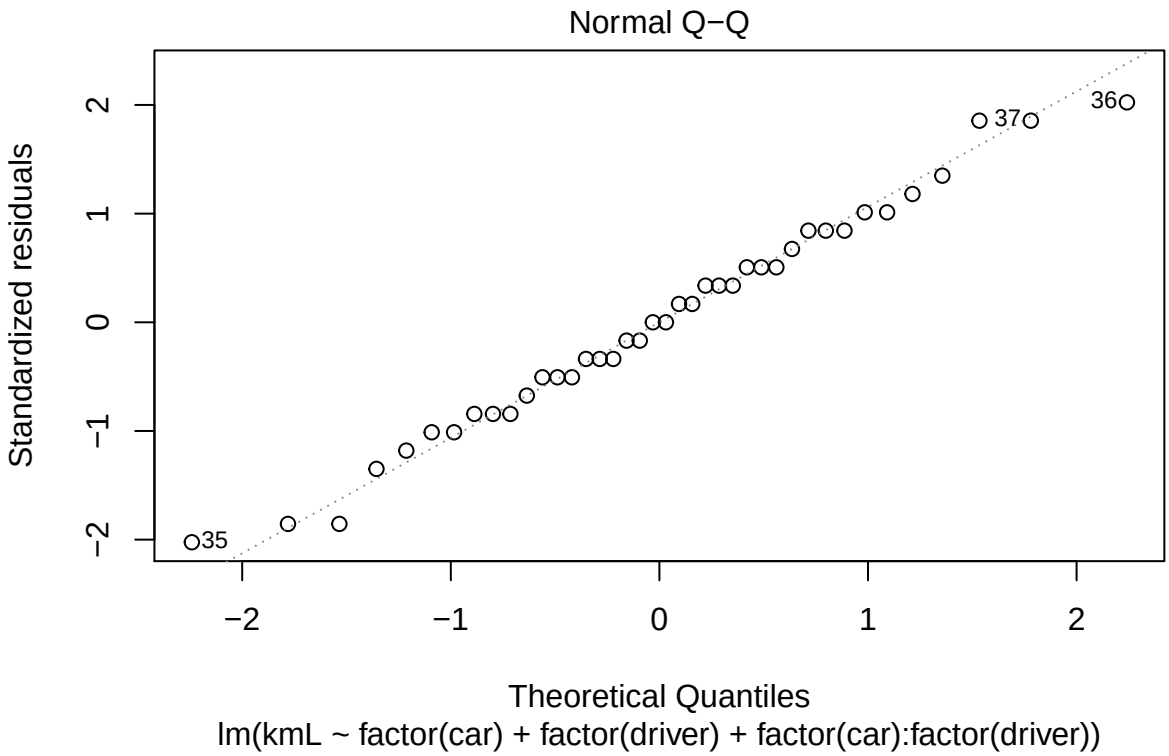
```
anova(kml.reduced)
```

```
## Analysis of Variance Table
##
## Response: kmL
##                             Df Sum Sq Mean Sq F value    Pr(>F)
## factor(car)                  4 17.119  4.2798  134.73 3.664e-14 ***
## factor(driver)               3 50.661 16.8869  531.60 < 2.2e-16 ***
## factor(car):factor(driver) 12  0.442  0.0368    1.16    0.3715
## Residuals                  20  0.635  0.0318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Main Effects: Driver
  Model $Y = \mu + \alpha_i + \beta_j + \epsilon$
  Hypotheses: $H0 : \beta_j = 0$ against H1 : at least one β_j non-zero
  $P - Value = 2.2e - 16 < 0.05$
  Driver is significant
- Main Effects: Car Model $Y = \mu + \alpha_i + \beta_j + \epsilon$
  Hypotheses: $H0 : \alpha_i = 0$ against H1 : at least one α_i non-zero
  $P - Value = 2.2e - 16 < 0.05$
- Car is significant
- Check Assumptions

```
plot(kml.reduced, which = 1:2)
```


Residuals vs Fitted
Residuals
0.3
0.2
0.1
−0.1
−0.3
36
37
35
11
12
13
14
15
Fitted values
lm(kmL ~ factor(car) + factor(driver) + factor(car):factor(driver))



* Normal QQ plot shows normality.

- The points is scattered randomly near the line.

d. Conclusion:

- Because the result for model with interaction, p-value is insignificant, so the effect on the mean outcome of a change in factor Driver has no interaction of the level of factor Car.