MLM diagnostic procedure Case study

Case Study: Surgical Unit Example

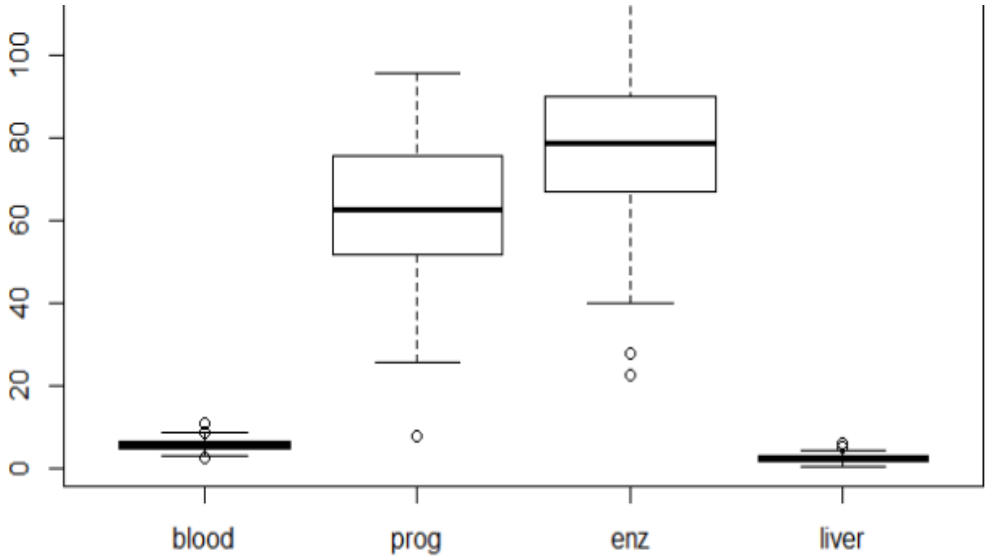
A hospital surgical unit was interested in predicting survival in patient undergoing a particular type of liver operation. A random number of 108 patients was available for analysis, but we only study (n=)54. For each patient record, the following information was extracted (data: surgery.csv):

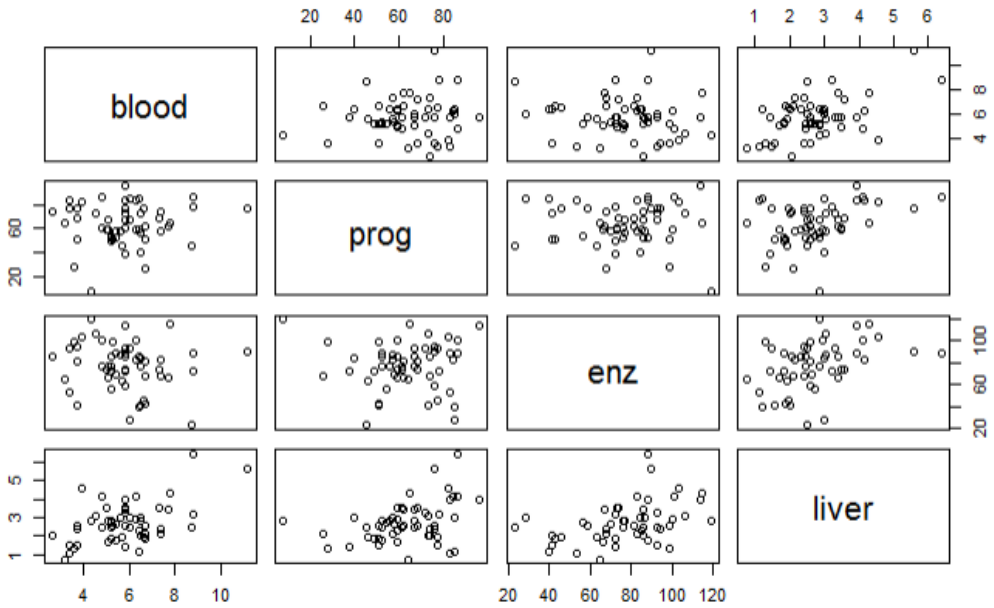
Potential predictors include,

- Blood clotting score (X_1, blood)
- A prognostic index (X_2, prog)
- Enzyme function test (X_3, enz)
- Liver function test (X_4, liver)

The response variable is survival time in days (Y, surv)

Model building process: explore the data





1. Judge by the boxplot and the scatter plot, do you think there are outliers ?_____

Is there multicollinearity among variables? _____

If the multicollinearity exists, it will probably be due to which variable? _____

Model building process: fit a first order $Y = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
blood      1 1005152 1005152 19.0470 6.567e-05 ***
prog       1 1278496 1278496 24.2267 1.010e-05 ***
enz        1 3442172 3442172 65.2269 1.461e-10 ***
liver      1   57862   57862  1.0964    0.3002
Residuals 49 2585839   52772
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1279.242    243.808  -5.247 3.30e-06 ***
blood          82.988     26.402   3.143  0.00284 **
prog            8.346      2.120   3.937  0.00026 ***
enz            10.870      1.923   5.652 8.01e-07 ***
liver          49.346     47.126   1.047  0.30018
```

2. Judge from the output of a first order MLR, which variables have a significant linear impact on Y?

_____

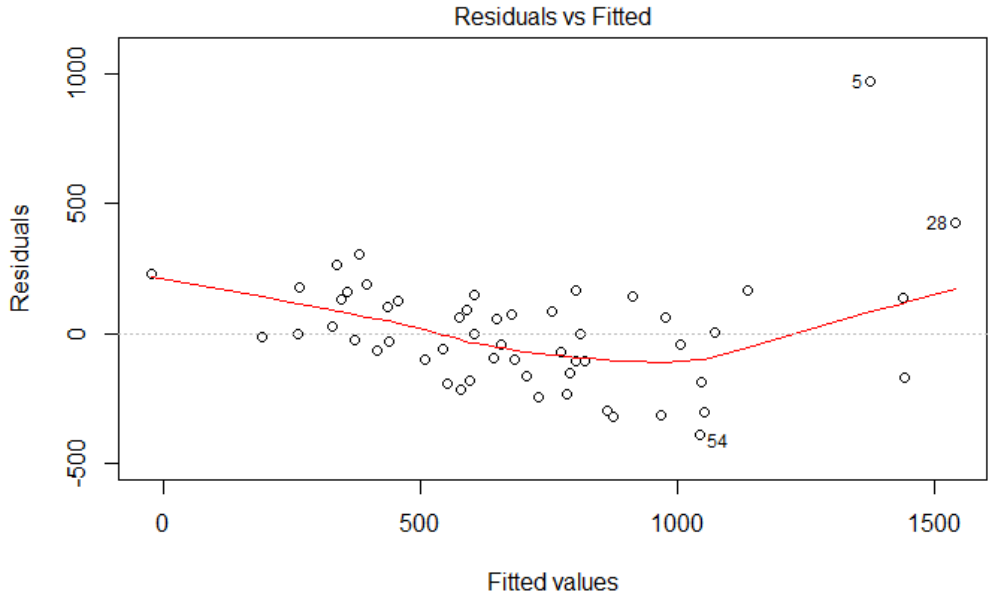
R^2=_____

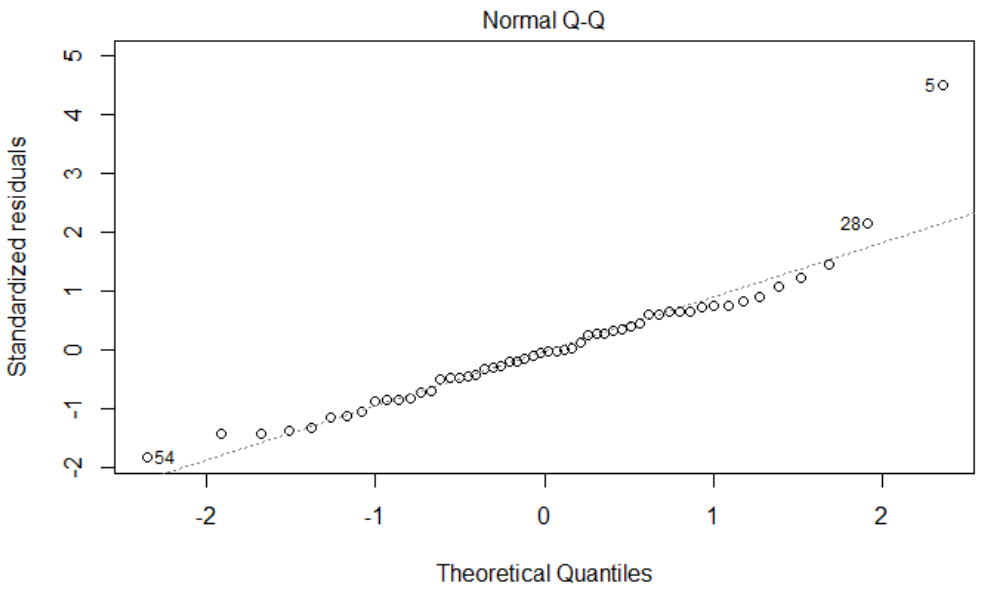
R^2_{adj} = _____

$s\{resid\}$ = _____

Model building process: diagnostic

$$Y = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$





3. Judge by the residual plot and Normality plot on the residuals, do you suspect any violation on the assumption?

Model building process: diagnostic

$$Y = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$$

```
  Brown-Forsythe Test
-------------------------------------------------------
  data : resid and group

  statistic  : 2.47243
  num df     : 4
  denom df   : 5.448805
  p.value    : 0.1643623

  Result     : Difference is not statistically significant.
-------------------------------------------------------
```

```
        Shapiro-Wilk normality test

data:  sur$resid
W = 0.90048, p-value = 0.0002946
```

4. From the available diagnostic plots and tests.

Is there violation on constant variance? ____(Yes/No/not sure)

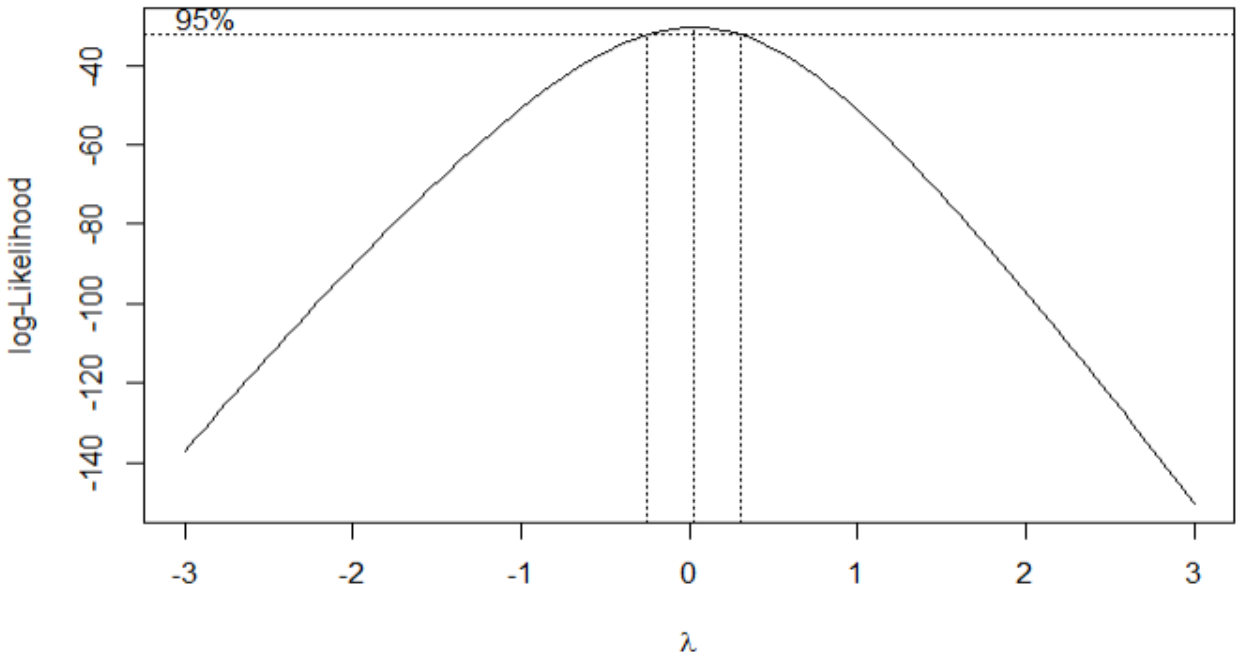
Is there violation on Normality? ____(Yes/No/not sure)

Is there violation on independence? ____(Yes/No/not sure)

5. Would you suggest transformation? _____

A) No. B) Yes, mainly on Y. C) Yes, mainly on X.

Model building process: box cox transformation



6. For simplicity, choose $\lambda = 0$

The transformation function is: $Y' = Y^{\lambda} =$ _____

The back transformation function is: $f^{-1}(Y') =$ _____

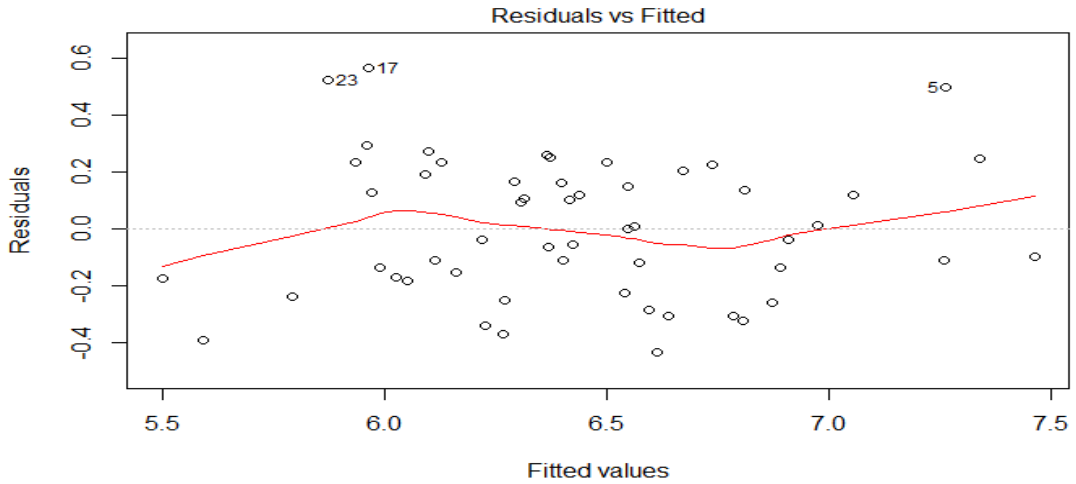
Model refit: $\ln(Y) = \beta_0 + \beta_1 blood + \beta_2 prog + \beta_3 enz + \beta_4 liver + \epsilon$

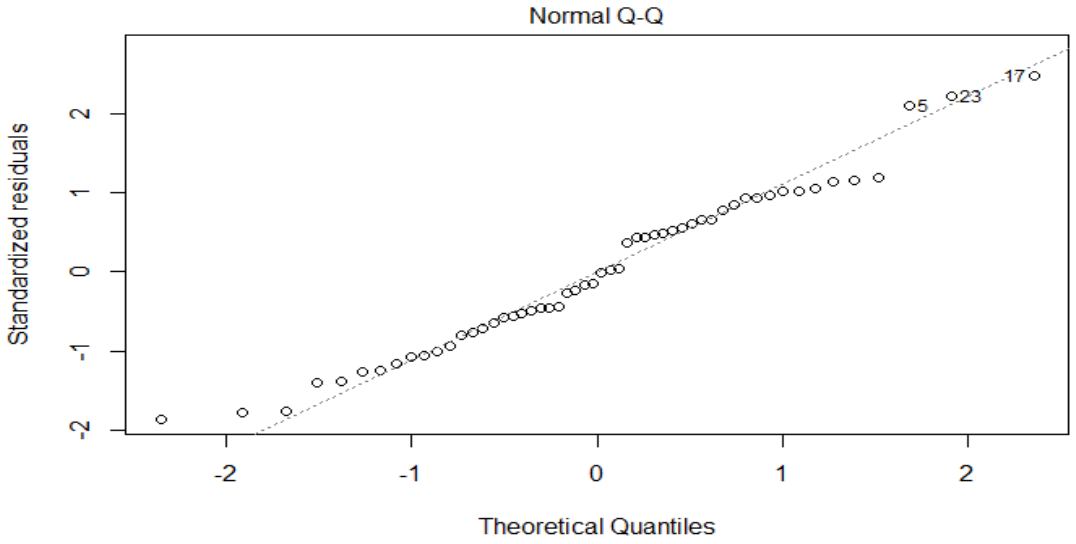
```
Analysis of Variance Table

Response: lny
          Df Sum Sq Mean Sq  F value    Pr(>F)
blood      1 0.7763  0.7763  12.3337 0.0009661 ***
prog       1 2.5888  2.5888  41.1325 5.377e-08 ***
enz        1 6.3341  6.3341 100.6408 1.810e-13 ***
liver      1 0.0246  0.0246   0.3905 0.5349320
Residuals 49 3.0840  0.0629

Residual standard error: 0.2509 on 49 degrees of freedom
Multiple R-squared:  0.7592,    Adjusted R-squared:  0.7396
F-statistic: 38.62 on 4 and 49 DF,  p-value: 1.388e-14
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.851948   0.266258  14.467  < 2e-16 ***
blood       0.083684   0.028833   2.902  0.00554 **
prog        0.012665   0.002315   5.471 1.51e-06 ***
enz         0.015632   0.002100   7.443 1.37e-09 ***
liver       0.032161   0.051465   0.625  0.53493
```





7. The refitted model is _____(better/worse) with a _____(higher/lower) multiple R-squared.

Model building process: diagnostic on the new model

```
  Brown-Forsythe Test
-------------------------------------------------------
  data : resid2 and group2

  statistic  : 1.324482
  num df     : 4
  denom df   : 17.42341
  p.value    : 0.3000292

  Result     : Difference is not statistically significant.
-------------------------------------------------------


        Shapiro-Wilk normality test

data:  sur$resid2
W = 0.96928, p-value = 0.1791
```

8. Any violation on the variance or the normality? _____

9. If the 95% confidence interval for mean response with the new model is (6.36, 6.53),
at a certain point (blood=6, prog=59, enz=81, and liver=2.5).
Then we are 95% confidence that the average survival time at this point is at least _____and at most _____days.