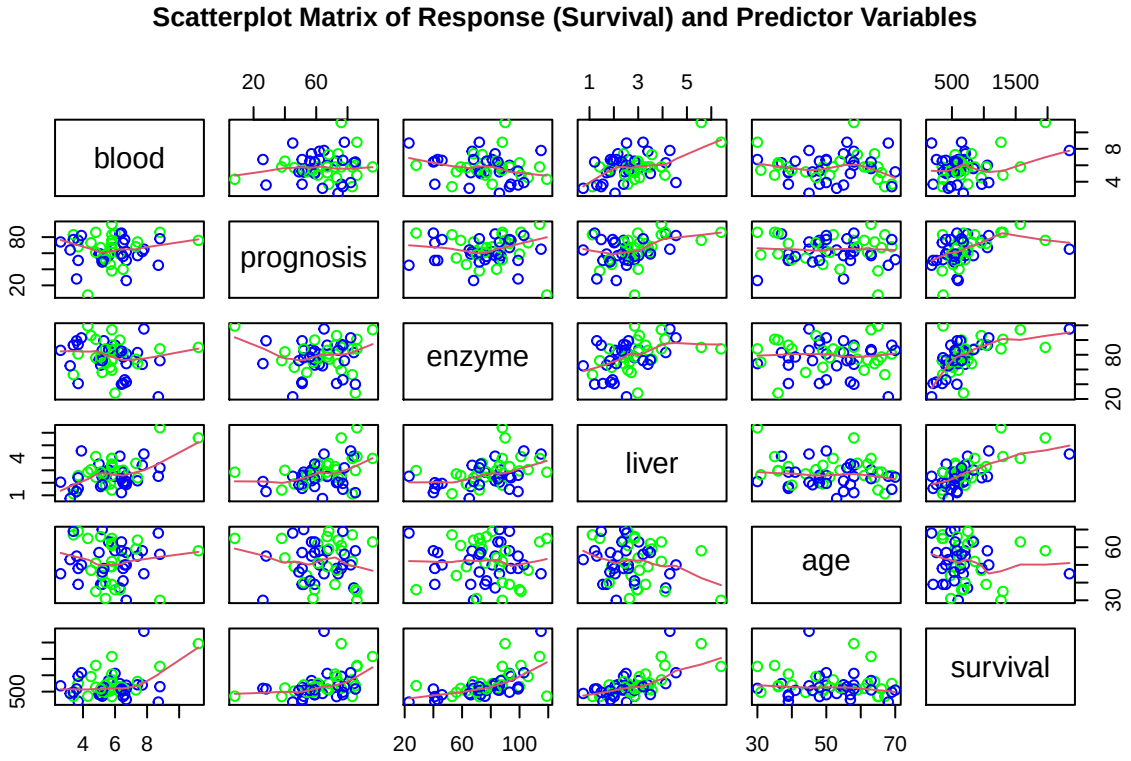
Assignment Montgomery Simes 46437819

Investigation of Survival Time (in days) Post Surgery



The scatterplot matrix above shows the relationships between the pairs of variables.

As the gender variable is categorical it has been removed and the gender of individual observations is displayed as green for female and blue for male.

It can be observed that overall females have a smaller blood clotting index and higher liver function index.

A positive linear relationship is evident between liver function index and survival time. A non-linear positive relationship is evident between the predictor variables enzyme function index and prognosis index and the response survival time. A possible small negative linear relationship could exist between age and survival.

The relationships between predictor variables is difficult to interpret due to statistical noise, however it is evident that there is no strong correlation between predictor variables.

Table 1: Correlation Matrix

	blood	prognosis	enzyme	liver	age	survival
blood	1.000	0.090	-0.150	0.502	-0.021	0.347
prognosis	0.090	1.000	-0.024	0.369	-0.048	0.420
enzyme	-0.150	-0.024	1.000	0.416	-0.013	0.578
liver	0.502	0.369	0.416	1.000	-0.207	0.674
age	-0.021	-0.048	-0.013	-0.207	1.000	-0.119
survival	0.347	0.420	0.578	0.674	-0.119	1.000

The correlation matrix above provides a summary of the possible linear relationships between pairs of variables.The correlations have a value between -1 (perfect negative correlation) and 1 (perfect positive correlation).
A medium positive correlation is evident between liver function index, enzyme function index and survival. A smaller correlation is evident between the response and prognosis index and blood clotting index. A slight negative correlation exists between age and survival time.

There is a small negative correlation between age and liver function index. A small to medium positive correlation exists between blood clotting index, prognosis index, enzyme index and liver function index. A small negative correlation exists between enzyme index and blood clotting index.

Multiple Linear Regression Analysis

Multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

β_0: Intercept

β_k: Partial regression coefficients

X_k: Predictor variables

Y: Response variable

Estimated regression model:

$$\hat{Survival} = b_0 + b_1 blood + b_2 prognosis + b_3 enzyme + b_4 liver + b_5 age + b_6 gender$$

$$\epsilon \sim N(0, \sigma^2)$$

b_0: Intercept

b_k: Partial regression coefficients

Overall ANOVA test of mulitple regression:

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_1 : \beta_i \neq 0$$

Table 2: Overall ANOVA table

Source	d.f.	S.S.	M.S.	Fvalue
Regression	6	5816714	969452.34	17.85
Residual	47	2552807	54315.03	
Total	53	8369521		

F statistic

$$F_{obs} = \frac{Regression\ M.S.}{Residual\ M.S.} = \frac{969452.34}{54315.03} = 17.84869$$

Null Distribution Under H_0, $F_{obs} \sim F_{6,47}$

P-Value $P(F_{6,47} \geq 17.84869) = 1.1902357 \times 10^{-10} < 0.05$

Conclusion As the P-value is less than the significance level of 0.05 there is evidence to reject H_0 for H_1. There is evidence suggesting a relationship exists between survival time and at least one of the predictor variables, subject to the validation of the model.

StepWise Backward Estimation

To improve the model a stepwise backward estimation was conducted. The table below shows a summary of the coefficients used in the above model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1179.189	283.823	-4.155	0.000
blood	86.644	27.492	3.152	0.003
prognosis	8.501	2.160	3.936	0.000
enzyme	11.125	1.982	5.613	0.000
liver	38.507	51.797	0.743	0.461
age	-2.341	3.014	-0.777	0.441
gender	-0.220	67.515	-0.003	0.997

As gender has the largest insignificant P-value of 0.997, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1179.367	275.619	-4.279	0.000
blood	86.630	26.905	3.220	0.002
prognosis	8.501	2.137	3.978	0.000
enzyme	11.124	1.958	5.683	0.000
liver	38.554	49.251	0.783	0.438
age	-2.340	2.969	-0.788	0.435

As Liver function index has the largest insignificant P-value of 0.438, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1246.655	260.835	-4.779	0.000
blood	100.660	19.987	5.036	0.000
prognosis	9.291	1.876	4.951	0.000
enzyme	12.101	1.502	8.058	0.000
age	-2.986	2.841	-1.051	0.298

As age has the largest insignificant P-value of 0.298, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1410.847	209.118	-6.747	0
blood	101.054	20.005	5.052	0
prognosis	9.382	1.876	5.000	0
enzyme	12.128	1.503	8.069	0

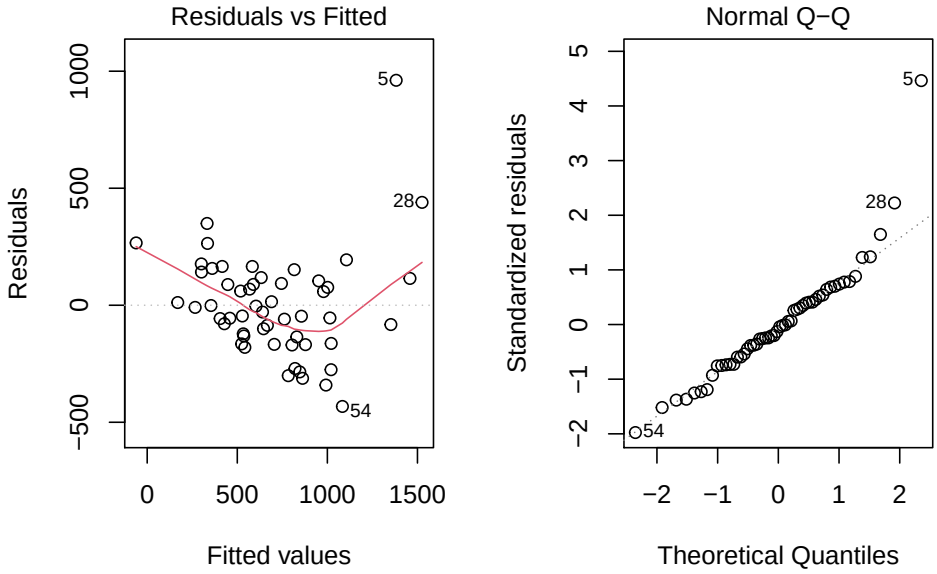
As the remaining coefficients were found to be significant the final modified model is:

$$\hat{Survival} = b_0 + b_1 blood + b_2 prognosis + b_3 enzyme$$

$$\hat{Survival} = -1410.85 + 101.05 blood + 9.38 prognosis + 12.13 enzyme$$

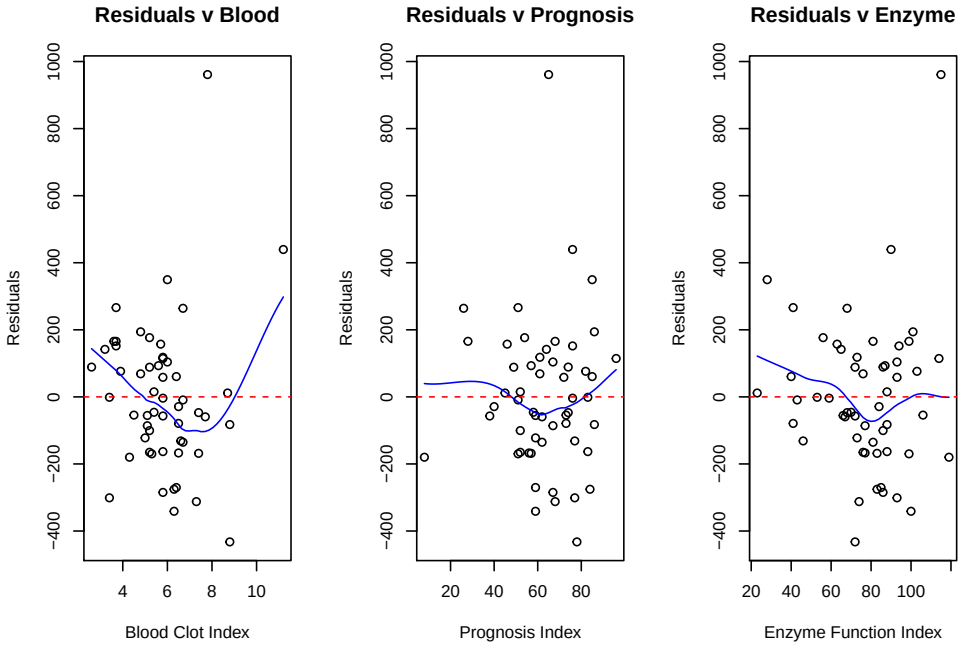
```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme, data = surgNum)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## 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
```

Model Validation



The Q-Q plot shows evidence of a right skew to the distribution of the residuals due to two outliers. The removal of these outliers may result in a more normal distribution of residuals.

The residuals v. fitted plot shows a concave up shape. This shows that the model does not accurately describe the changes in survival time based off changes to the predictor variables. As such the linearity assumption does not hold. There is also some evidence that the variance between residuals increases with larger fitted values.This suggests that the equal variance assumption does not hold.



An outlier is evident in all three plots. This data point should be investigated and possibly removed from the data set.

The residuals v blood clot index shows a curvature suggesting that the residuals do not show a random distribution around zero. The variance between the residuals could be consistent.

The residuals v prognosis index does show a slight curvature.The variance between residuals could be growing as the prognosis index increases.

The residuals v enzyme function index shows a downward trend in residuals as enzyme function increases. The variance between the residuals could be consistent.

Validation Summary The linearity assumption does not hold for the final model as evidenced by the pattern observed in the residuals v fitted values plot.

The equal variance assumption may not hold due to the changes in variance observed in the residual v fitted plot and residual v prognosis index plot.

As the model fails to meet the assumptions, it is not appropriate to explain survival time.

Overall ANOVA test of mulitple regression with response $log(survival)$:

Hypotheses

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$
$$H_1 : \beta_i \neq 0$$

Overall ANOVA Table

Source	d.f.	S.S.	M.S.	Fvalue
Regression	6	9.90	1.65	26.69
Residual	47	2.91	0.06	
Total	53	12.80		

F statistic

$$F_{obs} = \frac{Regression\ M.S.}{Residual\ M.S.} = \frac{1.649883}{0.06181303} = 26.69151$$

Null Distribution Under H_0, $F_{obs} \sim F_{6,47}$

P-Value $P(F_{6,47} \geq 26.69151) = 1.3899992 \times 10^{-13} < 0.05$

Conclusion As the P-value is less than the significance level of 0.05 there is evidence to reject H_0 for H_1. There is evidence suggesting a relationship exists between $log(survival)$ and at least one of the predictor variables, subject to the validation of the model.

StepWise Backward Estimation

To improve the model a stepwise backward estimation was conducted. the table below shows a summary of the coefficients used in the above model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.101	0.303	13.544	0.000
blood	0.095	0.029	3.234	0.002
prognosis	0.013	0.002	5.650	0.000
enzyme	0.016	0.002	7.683	0.000
liver	-0.003	0.055	-0.057	0.955
age	-0.005	0.003	-1.513	0.137
gender	-0.066	0.072	-0.918	0.363

As liver function index has the largest insignificant P-value of 0.955, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.105	0.291	14.117	0.000
blood	0.094	0.021	4.372	0.000
prognosis	0.013	0.002	6.398	0.000
enzyme	0.016	0.002	9.939	0.000
age	-0.005	0.003	-1.581	0.121
gender	-0.065	0.068	-0.949	0.347

As gender has the largest insignificant P-value of 0.347, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.029	0.279	14.434	0.000
blood	0.095	0.021	4.435	0.000
prognosis	0.013	0.002	6.574	0.000
enzyme	0.016	0.002	10.208	0.000
age	-0.005	0.003	-1.568	0.123

As age has the largest insignificant P-value of 0.123, it was removed from the regression model and the model was re-fit. The coefficients were recalculated to be:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.766	0.227	16.610	0
blood	0.095	0.022	4.401	0
prognosis	0.013	0.002	6.558	0
enzyme	0.016	0.002	10.089	0

As the remaining coefficients were found to be significant the final modified model is:

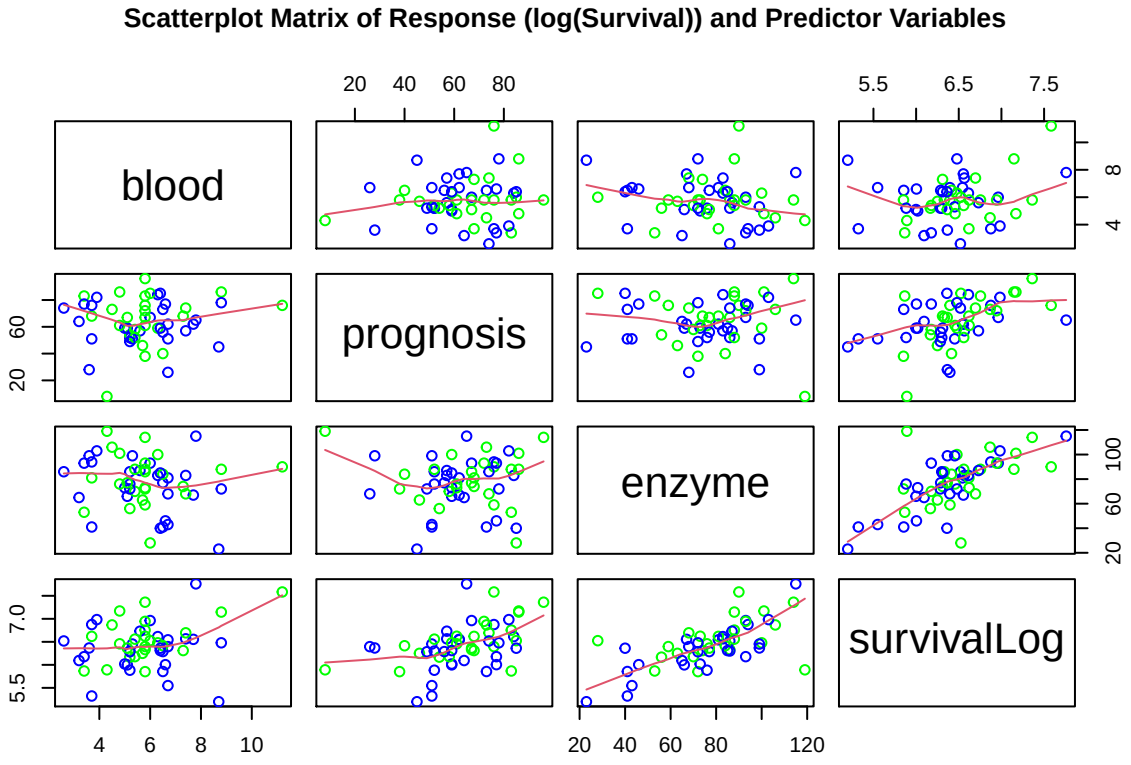
$$log(\hat{Survival}) = b_0 + b_1 blood + b_2 prognosis + b_3 enzyme$$

$$log(\hat{Survival}) = 3.77 + 0.095 blood + 0.013 prognosis + 0.016 enzyme$$

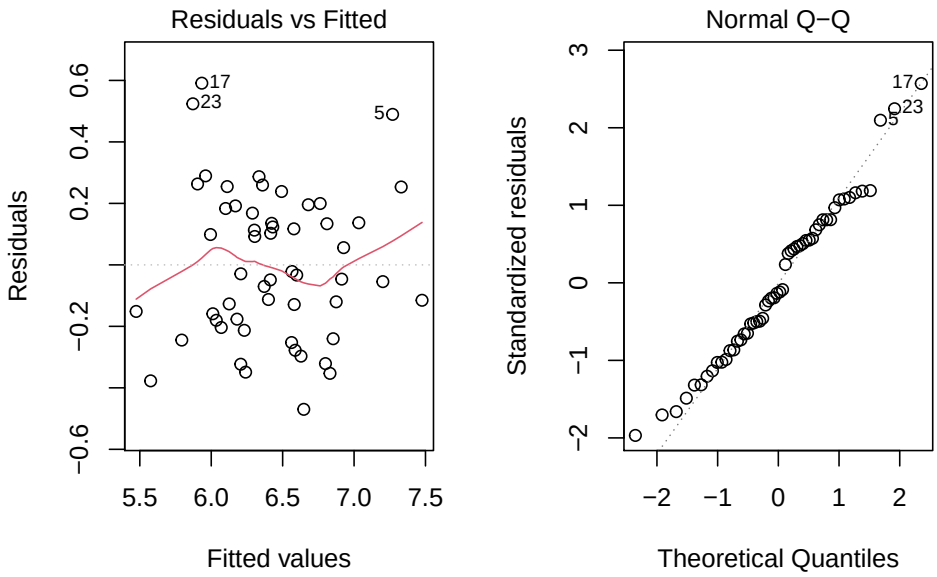
```
##
## Call:
## lm(formula = survivalLog ~ blood + prognosis + enzyme, data = surgNum)
```

```
##
## 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
```

Model Validation

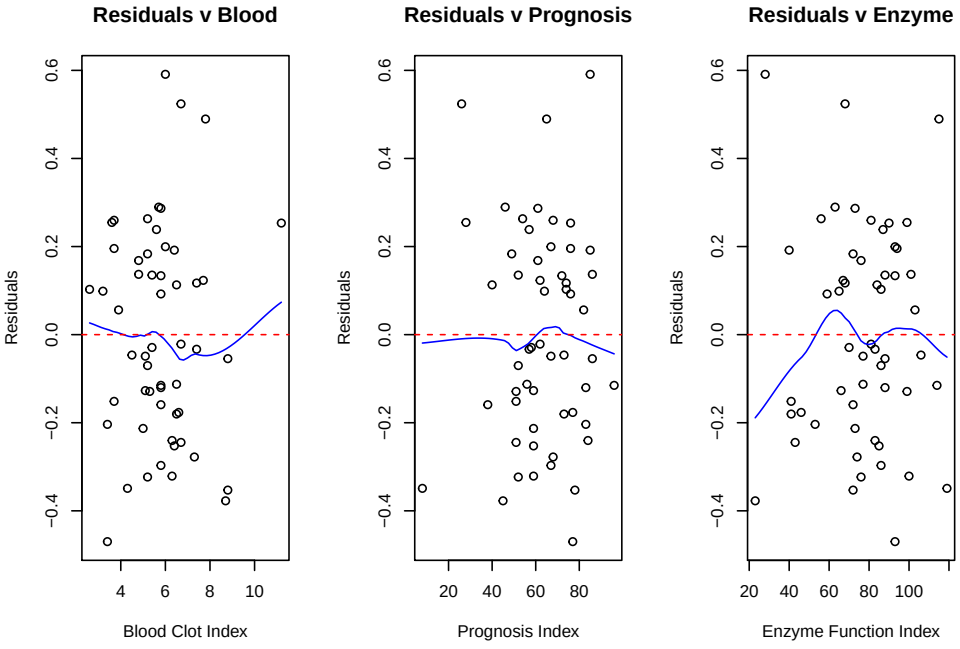


The scatterplot matrix above shows a clear positive linear relationship between enzyme function index and $log(survival)$ exists. A weak positive linear relationship between prognosis index and $log(survival)$ could exist. There is no strong correlation between predictor variables



The residuals v. fitted plot shows an even distibution around zero with no evident patterns.

The Q-Q plot has a slight 's' shape, however it would be reasonable to consider the residuals came from a normal distribution.



The residuals v blood clot index, residuals v prognosis index and residuals v enzyme function index appear relatively evenly distributed around zero and the variance appears constant.

Validation Summary

The model using $log(survival)$ as the response variable is a appropriate model to use to explain survival time, as the linearity, normality and equal variance assumptions have been satisfied.

Model Comparison

Both models found that the variables, blood clot index, prognosis index and enzyme function index were significant predictors of survival and $log(survival)$. The variables liver function index, age, and gender were not significant predictors of survival or $log(survival)$.

The adjusted R^2 for the first model is 0.6652. The adjusted R^2 for the second model is 0.7427. Based on the adjusted R^2 metric the second model is a more powerful model.

As the first model did not meet the required assumptions for a multiple regression model, it was found not to be suitable. As the second model did meet the assumptions, the second model is the superior model with regards to modeling survival time. However it should be noted that the final results of the $log(survival)$ will have to be re-transformed so that the results can be interpreted in survival time (in days).

Investigation of Fuel Efficiency of Car Engine

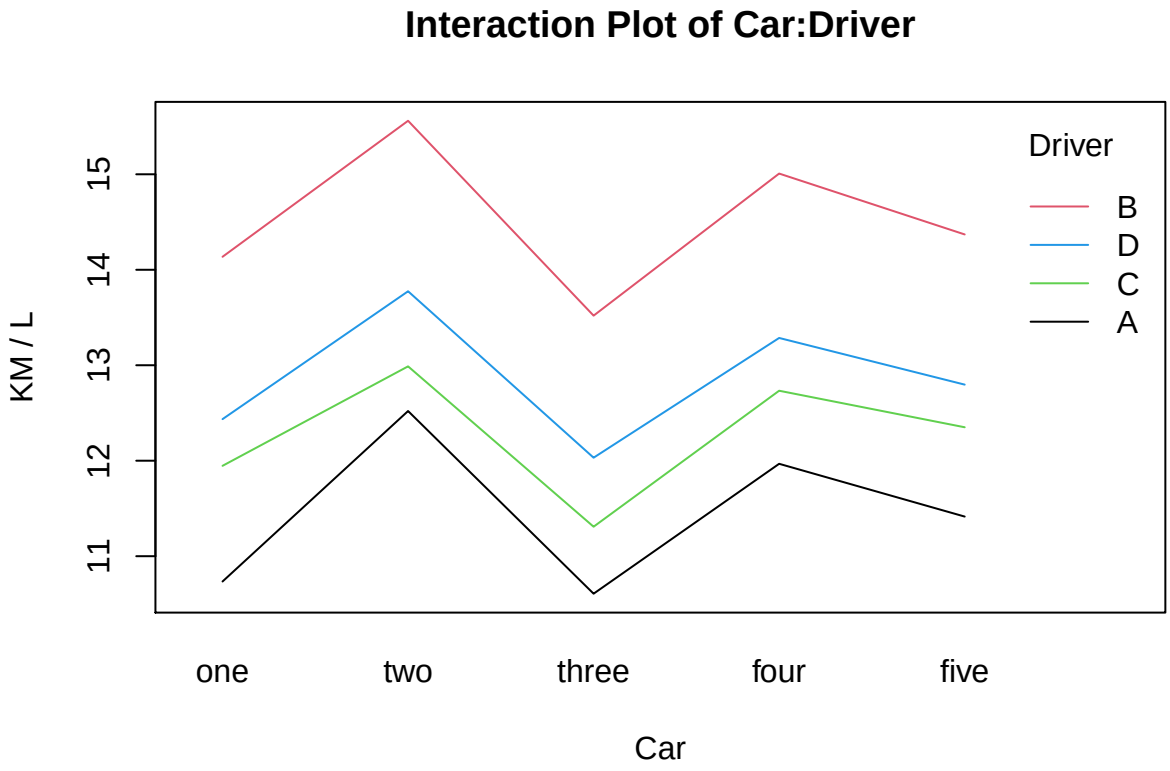
Balanced Design

A study with a balanced design has the same number of replicates for each combination of treatments. The table below shows the number of treatment combination observations within the study.

Table 12: Replicates per Car and Driver Combination

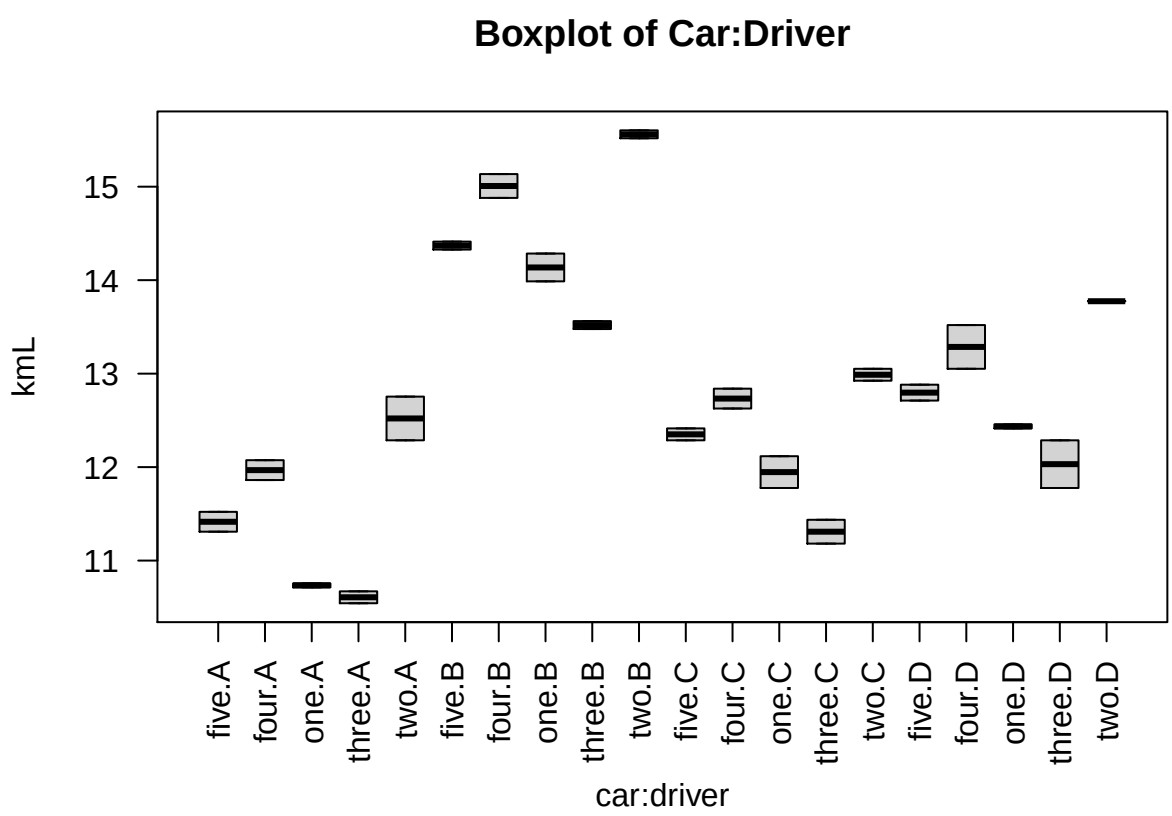
	one	two	three	four	five
A	2	2	2	2	2
B	2	2	2	2	2
C	2	2	2	2	2
D	2	2	2	2	2

In this case each combination of car and driver have equal replicates therefore the study is balanced.



The interaction plot above displays the interactions between changes in the response as a result of the combination of levels from the car factor and the driver factor. When the lines are parallel, no interaction is evident, that is one factor has a constant effect on the response regardless of the remaining factor. When the lines are non-parallel, an interaction is evident with the larger the difference in slope between the lines the greater the strength of the interaction.

The Interaction plot for this study shows little difference in the slopes and the plots are predominantly parallel. There are slight differences between the slopes of driver c between cars one and two compared to the other three drivers. A two way ANOVA test should be conducted to see if this interaction effect is significant.



The boxplot shows some variety in the spread, however this may be due to low sample size. The large differences in groups by driver factor does show a strong main effect.The difference in the medians between car type with the same driver shows a strong main effect by the type of car driven. An interaction effect may be present however this is hard to observe from the boxplot.

Two-Way ANOVA

Hypothesis Interaction effect: $H_0 : \gamma_{ij} = 0$ for all combinations of i, j $H_1 : \gamma_{ij} \neq 0$

Table 13: Two-Way ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
car	4	17.119	4.280	134.728	0.000
driver	3	50.661	16.887	531.597	0.000
car:driver	12	0.442	0.037	1.160	0.371
Residuals	20	0.635	0.032		

P-Value for car:driver $P(F_{12,20} \geq 1.1600284) = 0.3714839 > 0.05$

Summary The Pvalue for the car:driver interaction effect of 0.3714839 is statistically insignificant (> 0.05) therefore we cannot reject H_0. The car:driver interaction effect is not significantly different from zero.

The interaction effect was removed from the model to assess the main effects of the driver and car.

Hypothesis Main effect Car: $H_0 : \alpha_{one} = \alpha_{two} = \alpha_{three} = \alpha_{four} = \alpha_{five}$
$H_1 : \alpha_i \neq 0$

Main effect Driver: $H_0 : \beta_A = \beta_B = \beta_C = \beta_D$
$H_1 : \beta_i \neq 0$

Table 14: Two-Way ANOVA Table

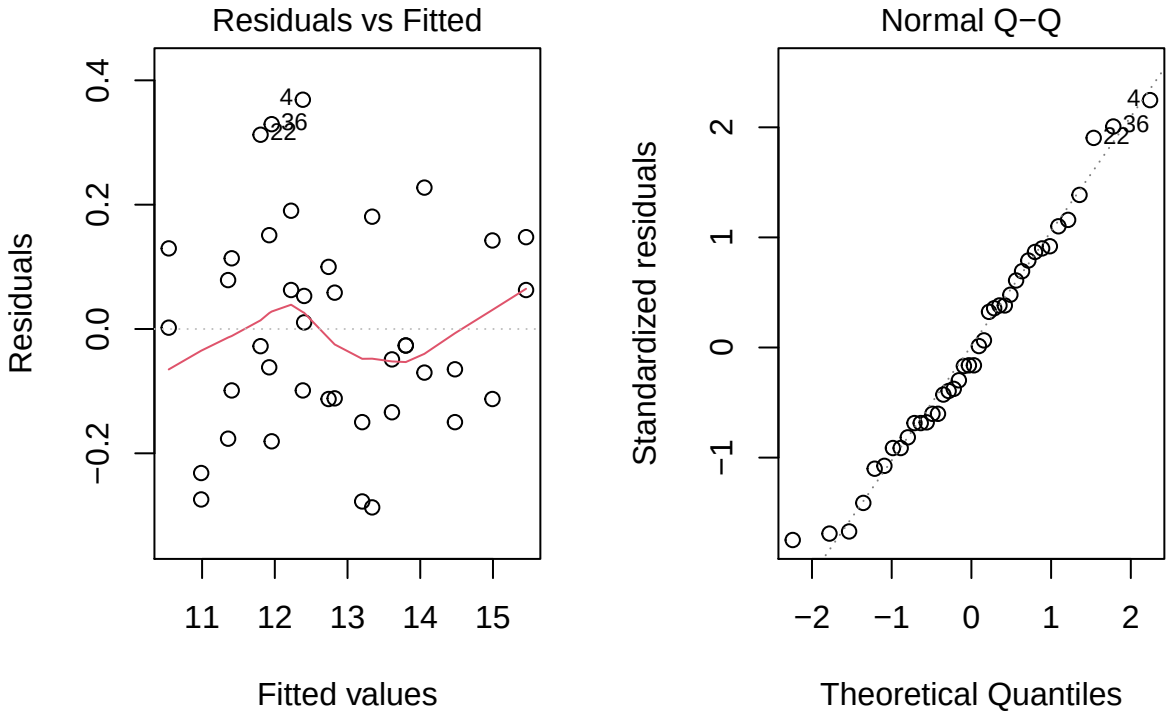
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
car	4	17.119	4.280	127.100	0
driver	3	50.661	16.887	501.502	0
Residuals	32	1.078	0.034		

P-Value for main effects

Main effect for car: $P(F_{4,32} \geq 127.1002265) = 3.6684849 \times 10^{-19} < 0.05$
Main effect for driver: $P(F_{3,32} \geq 501.501971) = 5.7284851 \times 10^{-27} < 0.05$

Summary There is sufficient evidence to reject the H_0 for both the main effects of driver and car as the pvalue is less than the significance level of 0.05.

Model Validation The boxpot above is used to check for equal variability across different groups. The boxplot does show some variability in spread. It must be considered that only 2 observations were recorded for each combination of factors.



The residuals v. fitted plot shows evenly distributed residuals around zero and variance appears constant. The normal QQ plot is linear suggesting that the residuals follow normal distribution. The assumptions for the model have been satisfied.

Conclusions

Upon investigating the interaction effect of car:driver it was found that there was insufficient evidence to suggest there is an interaction effect on fuel efficiency between the combination of factors. This was suggested by the primarily parallel configuration of the interaction plot and confirmed by the results of the two-way ANOVA test. It was found that the car factor and driver factor had a significant main effect on the fuel efficiency performance. This was evident by the differences evident in the boxplot plot and confirmed by the two-way ANOVA analysis.

Question 4 repository link

https://github.com/MQ-STAT2170-STAT6180/assignment-montysimes