Assignment 3: Multiple linear regression II

Matthew D. Sherman

Department of Mathematics and Statistics, University of Canterbury

STAT202: Regression Modelling

August 17, 2023

Question One:

I created a new folder for my work as well as a new RStudio project in that folder. I also downloaded the aquatic_toxicity.xlsx file and put that in the folder.

Question Two:

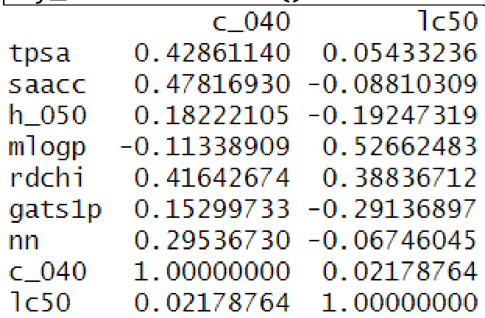
```
library(readxl)
toxic <- read_xlsx("aquatic_toxicity.xlsx")
```

Question Three:

```
set.seed(16645573)
my_toxic <- toxic %>% sample_n(525)
```

Question Four:

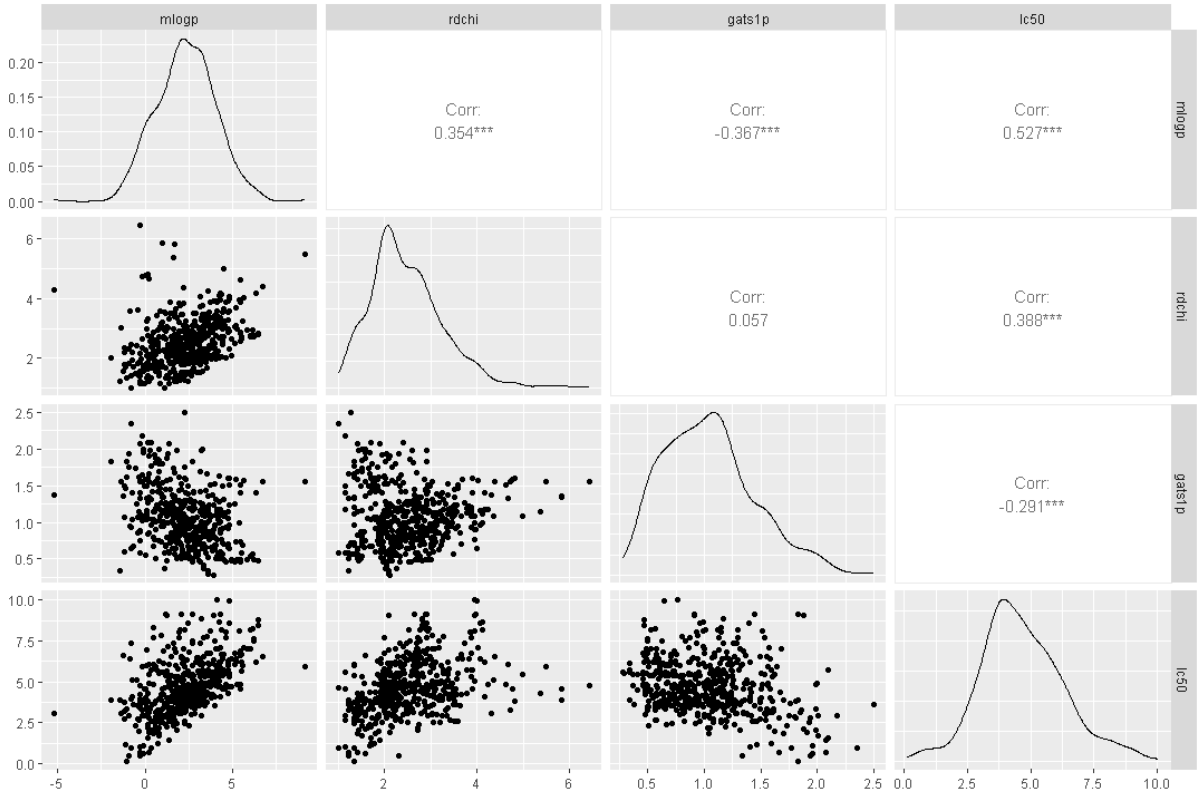
```
my_toxic %>% cor()
```



```
              c_040        lc50
tpsa     0.42861140  0.05433236
saacc    0.47816930 -0.08810309
h_050    0.18222105 -0.19247319
mlogp   -0.11338909  0.52662483
rdchi    0.41642674  0.38836712
gats1p   0.15299733 -0.29136897
nn       0.29536730 -0.06746045
c_040    1.00000000  0.02178764
lc50     0.02178764  1.00000000
```

mlogp, rdchi, and gats1p have the three strongest relationships with lc50 (0.52662483, 0.38836712, -0.29136897 respectively), hence these three variables will be used to predict lc50.

```
ggpairs(my_toxic, columns = c("mlogp", "rdchi", "gats1p", "lc50"))
```

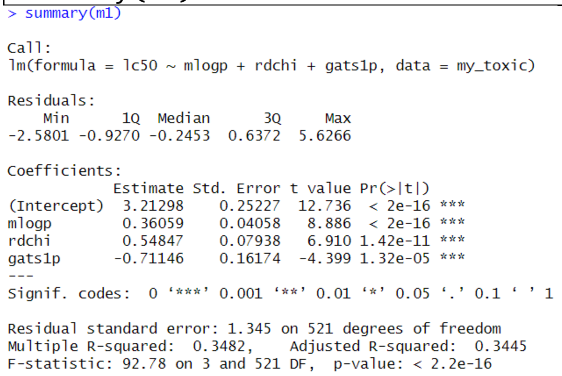


Explain in 50 words the relationships you observe in those plots.

The correlation strengths for the relationships range from very weak to moderate (0 to 0.6/-0.6 correlation). There are both positive and negative correlations, as well as essentially no correlation for rdchi vs gats1p (0.057 correlation).

Question Five:

```
m1 <- lm(lc50 ~ mlogp + rdchi + gats1p, data = my_toxic)
summary(m1)
```



```
> summary(m1)

Call:
lm(formula = lc50 ~ mlogp + rdchi + gats1p, data = my_toxic)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5801 -0.9270 -0.2453  0.6372  5.6266 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.21298    0.25227  12.736  < 2e-16 ***
mlogp        0.36059    0.04058   8.886  < 2e-16 ***
rdchi        0.54847    0.07938   6.910 1.42e-11 ***
gats1p      -0.71146    0.16174  -4.399 1.32e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.345 on 521 degrees of freedom
Multiple R-squared:  0.3482,	Adjusted R-squared:  0.3445 
F-statistic: 92.78 on 3 and 521 DF,  p-value: < 2.2e-16
```

Coefficients:

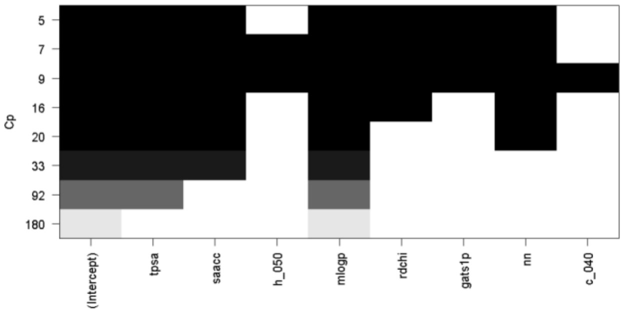
- Intercept: 3.21298
- mlogp: 0.36059
- rdchi: 0.54847
- gats1p: -0.71146

Adjusted R-squared: 0.3445

Residual Standard Error: 1.345

Question Six:

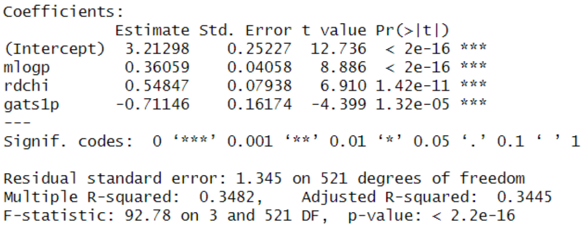
```
library(leaps)
all_mods <- regsubsets(lc50 ~ ., data = my_toxic)
plot(all_mods, scale = 'Cp')
```



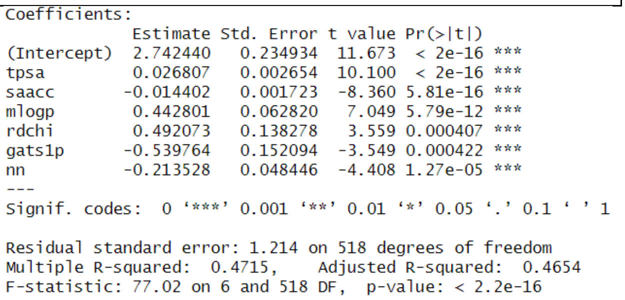
From the plot, the best model is the one at the lowest Cp value/the model at the top of the plot (which is at Cp 5). The predictors to include for this model are those which are represented as black at Cp value 5 – tpsa, saacc, mlogp, rdchi, gats1p, and nn.

Question Seven:

```
m2 <- lm(lc50 ~ tpsa + saacc + mlogp + rdchi + gats1p + nn, data = my_toxic)
summary(m2)
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.21298    0.25227  12.736  < 2e-16 ***
mlogp        0.36059    0.04058   8.886  < 2e-16 ***
rdchi        0.54847    0.07938   6.910 1.42e-11 ***
gats1p      -0.71146    0.16174  -4.399 1.32e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.345 on 521 degrees of freedom
Multiple R-squared:  0.3482,	Adjusted R-squared:  0.3445 
F-statistic: 92.78 on 3 and 521 DF,  p-value: < 2.2e-16
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.742440   0.234934  11.673  < 2e-16 ***
tpsa         0.026807   0.002654  10.100  < 2e-16 ***
saacc       -0.014402   0.001723  -8.360 5.81e-16 ***
mlogp        0.442801   0.062820   7.049 5.79e-12 ***
rdchi        0.492073   0.138278   3.559 0.000407 ***
gats1p      -0.539764   0.152094  -3.549 0.000422 ***
nn          -0.213528   0.048446  -4.408 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.214 on 518 degrees of freedom
Multiple R-squared:  0.4715,	Adjusted R-squared:  0.4654 
F-statistic: 77.02 on 6 and 518 DF,  p-value: < 2.2e-16
```

m1:

Adjusted R-squared: 0.3445

Residual Standard Error: 1.345

m2:

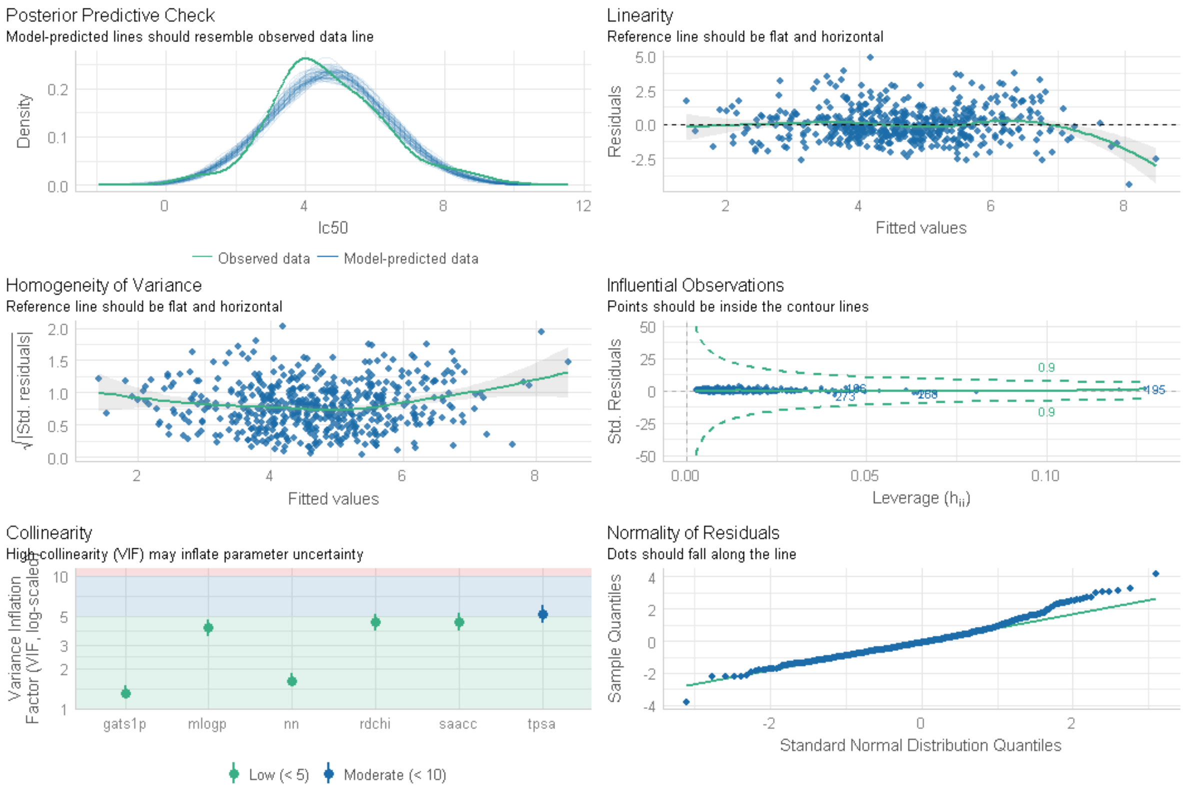
Adjusted R-squared: 0.4654

Residual Standard Error: 1.214

Compare m2 adjusted-r2 and residual standard error with m1. Discuss in 50 words the similarities and differences between the results of the 2 models.

M2 has a higher adjusted R-squared and lower RSE compared to m1, indicating that m2 has better explanatory power and a better fit compared to m1. Both models are predicting lc50 and share three of the same predictors. M2 has three additional predictors which could explain for the more favourable adjusted R-squared and RSE values.

Question Eight:

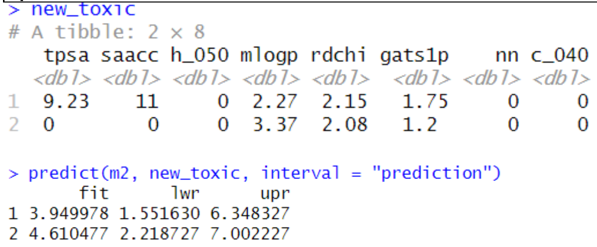


Check the model for assumptions for the residuals. Explain in no more than 70 words if there is anything unusual or wrong.

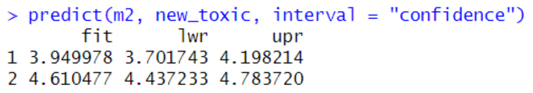
The HoV plot, indicates heteroscedasticity, meaning that the assumption of homoscedasticity may not be met. The reference line in the linearity plot curves downwards, suggesting that the assumption of linearity may not be met. In the NoR plot, the points fall above the line towards the end of the x-axis, meaning that the assumption of normality may not be met.

Question Nine:

```
new_toxic <- tibble(
  tpsa = c(9.23, 0),
  saacc = c(11, 0),
  h_050 = c(0, 0),
  mlogp = c(2.27, 3.37),
  rdchi = c(2.15, 2.08),
  gats1p = c(1.75, 1.20),
  nn = c(0, 0),
  c_040 = c(0, 0)
)
new_toxic

predict(m2, new_toxic, interval = "prediction")
predict(m2, new_toxic, interval = "confidence")
```



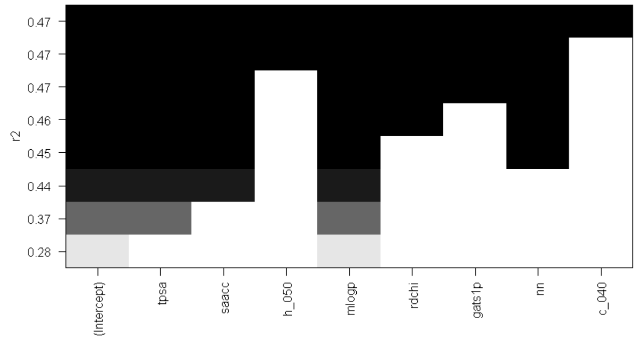
```
> new_toxic
# A tibble: 2 × 8
   tpsa saacc h_050 mlogp rdchi gats1p    nn c_040
  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
1  9.23    11     0  2.27  2.15   1.75     0     0
2  0        0     0  3.37  2.08   1.2      0     0

> predict(m2, new_toxic, interval = "prediction")
       fit      lwr      upr
1 3.949978 1.551630 6.348327
2 4.610477 2.218727 7.002227
```



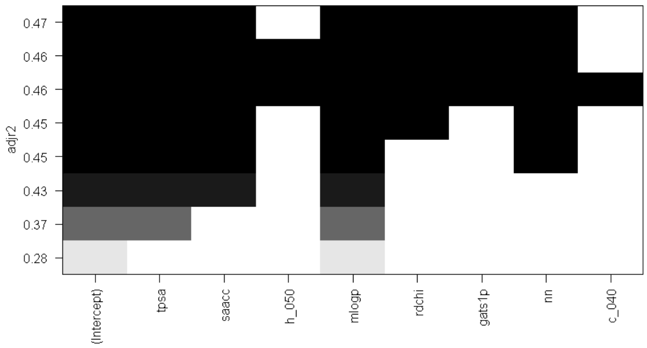
```
> predict(m2, new_toxic, interval = "confidence")
       fit      lwr      upr
1 3.949978 3.701743 4.198214
2 4.610477 4.437233 4.783720
```

Question Ten:

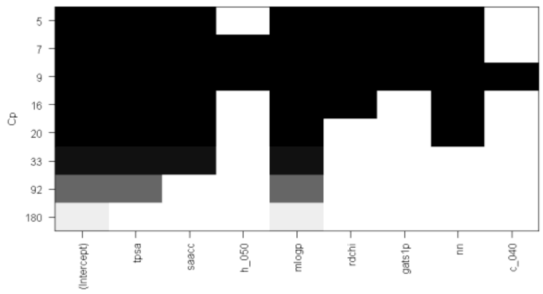
```
plot(all_mods, scale = 'r2')
```

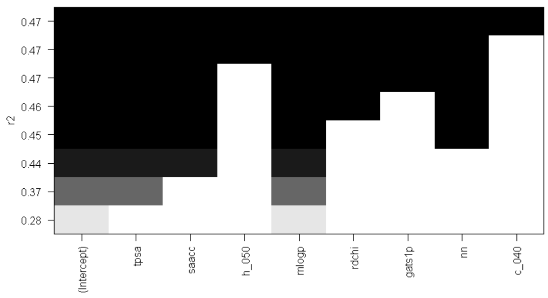


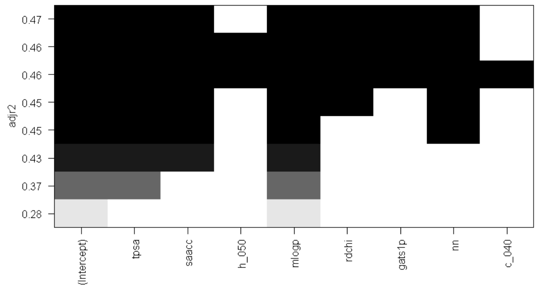
```
plot(all_mods, scale = 'adjr2')
```



Comparison:





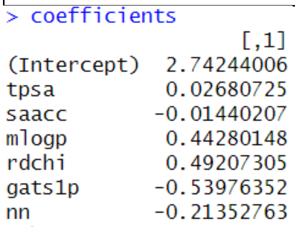


Compare the 3 plots and explain how/why they differ in 50 words

The r2 and adjr2 plots have the same y-axis values as they both assess goodness-of-fit but have different models as adjr2 accounts for model complexity. The Cp and adjr2 plots have the same models as they both prioritise model complexity but have different y-axis values as they assess the models differently.

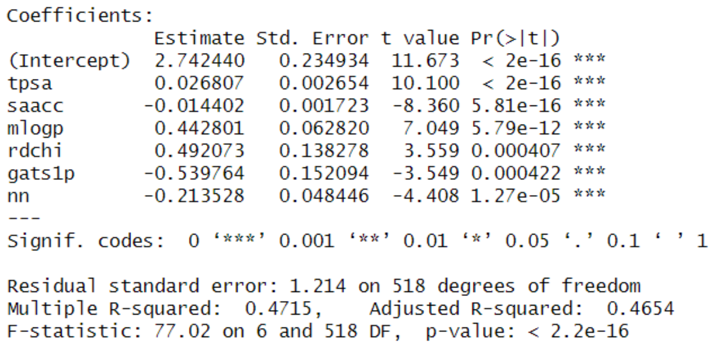
Question Eleven:

```
X <- model.matrix(m2)
y <- my_toxic$lc50
XtX <- t(X) %*% X
Xty <- t(X) %*% y
coefficients <- solve(XtX) %*% Xty
coefficients
```



```
> coefficients
                    [,1]
(Intercept)   2.74244006
tpsa          0.02680725
saacc        -0.01440207
mlogp         0.44280148
rdchi         0.49207305
gats1p       -0.53976352
nn           -0.21352763
```

These are the same coefficients as model 2:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.742440   0.234934  11.673  < 2e-16 ***
tpsa         0.026807   0.002654  10.100  < 2e-16 ***
saacc       -0.014402   0.001723  -8.360 5.81e-16 ***
mlogp        0.442801   0.062820   7.049 5.79e-12 ***
rdchi        0.492073   0.138278   3.559 0.000407 ***
gats1p      -0.539764   0.152094  -3.549 0.000422 ***
nn          -0.213528   0.048446  -4.408 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.214 on 518 degrees of freedom
Multiple R-squared:  0.4715,	Adjusted R-squared:  0.4654 
F-statistic: 77.02 on 6 and 518 DF,  p-value: < 2.2e-16
```

Question Twelve:

I have put all of my answer, figures and code in a PDF and have uploaded it to LEARN.

Full Code:

```
library(tidyverse)
library(performance)
library(GGally)
library(tibble)

library(readxl)
toxic <- read_xlsx("aquatic_toxicity.xlsx")

set.seed(16645573)
my_toxic <- toxic %>% sample_n(525)

my_toxic %>% cor()

ggpairs(my_toxic, columns = c("mlogp", "rdchi", "gats1p", "lc50"))

m1 <- lm(lc50 ~ mlogp + rdchi + gats1p, data = my_toxic)
summary(m1)

library(leaps)
all_mods <- regsubsets(lc50 ~ ., data = my_toxic)
plot(all_mods, scale = 'Cp')

m2 <- lm(lc50 ~ tpsa + saacc + mlogp + rdchi + gats1p + nn, data = my_toxic)
summary(m2)

check_model(m2)

library(tibble)
new_toxic <- tibble(
  tpsa = c(9.23, 0),
  saacc = c(11, 0),
  h_050 = c(0, 0),
  mlogp = c(2.27, 3.37),
  rdchi = c(2.15, 2.08),
  gats1p = c(1.75, 1.20),
  nn = c(0, 0),
  c_040 = c(0, 0)
)
new_toxic

predict(m2, new_toxic, interval = "prediction")
predict(m2, new_toxic, interval = "confidence")

plot(all_mods, scale = 'r2')
plot(all_mods, scale = 'adjr2')

X <- model.matrix(m2)
y <- my_toxic$lc50
XtX <- t(X) %*% X
```

```
Xty <- t(X) %*% y
coefficients <- solve(XtX) %*% Xty
coefficients
```