Midterm 2 STAT632

Brandon Keck (netID: qh9701)

```r
library(ggplot2)
library(dplyr)
library(readr)
library(car)
library(MASS)
library(tidyverse)
```

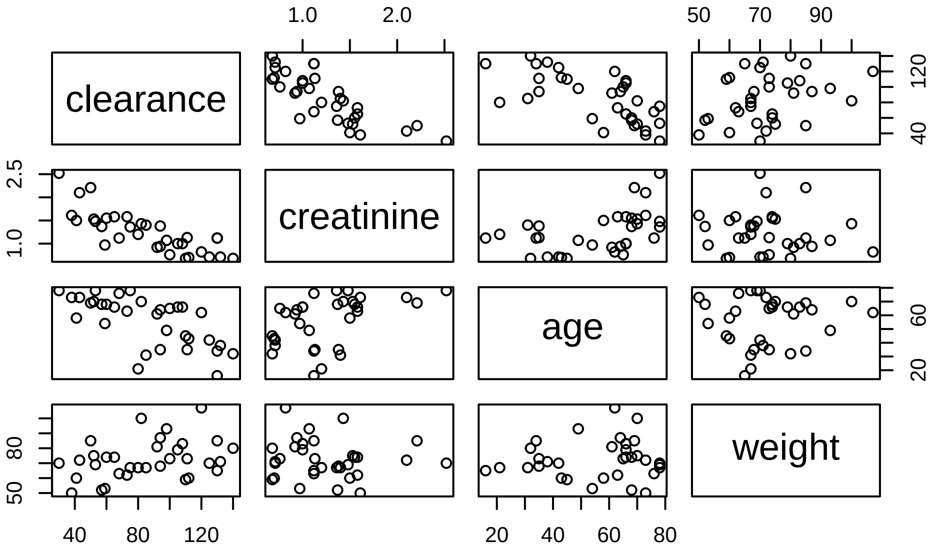
```r
kidney <- read.csv("Kidney.csv")
head(kidney)
```

```
  clearance creatinine age weight
1       132       0.71  38     71
2        53       1.48  78     69
3        50       2.21  69     85
4        82       1.43  70    100
5       110       0.68  45     59
6       100       0.76  65     73
```

1.

(a)

```r
pairs(clearance ~ creatinine + age + weight, data = kidney)
```



There appears to be some negative linearity between clearance and creatinine. However the rest of the variables do not appear to have any real signs of linearity between them. It does appear that a transformation may be needed of the response variable clearance but further investigation is needed.

(b)

```r
lm1 <- lm(clearance ~ creatinine + age + weight, data = kidney)
summary(lm1)
```

```
Call:
lm(formula = clearance ~ creatinine + age + weight, data = kidney)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.668  -7.002   1.518   9.905  16.006 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
creatinine  -39.9393     5.6000  -7.132 7.55e-08 ***
age          -0.7368     0.1414  -5.211 1.41e-05 ***
weight        0.7764     0.1719   4.517 9.69e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.46 on 29 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:  0.8398
F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12
```

```
lm3 <- lm(clearance ~ creatinine, data = kidney)
summary(lm3)$adj.r.squared
```

```
[1] 0.6313813
```

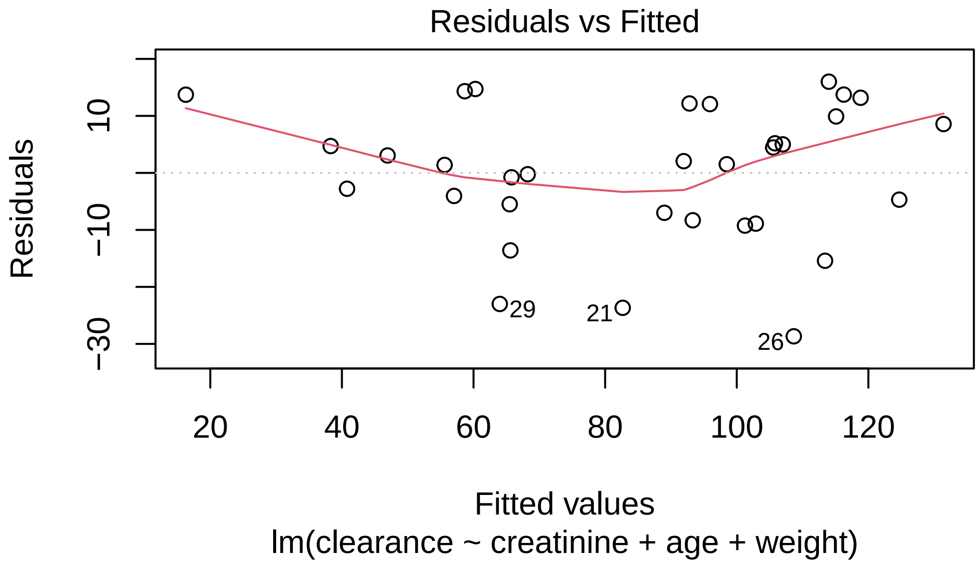
```
lm4 <- lm(clearance ~ creatinine + age + weight, data = kidney)
summary(lm4)$adj.r.squared
```

```
[1] 0.8397998
```

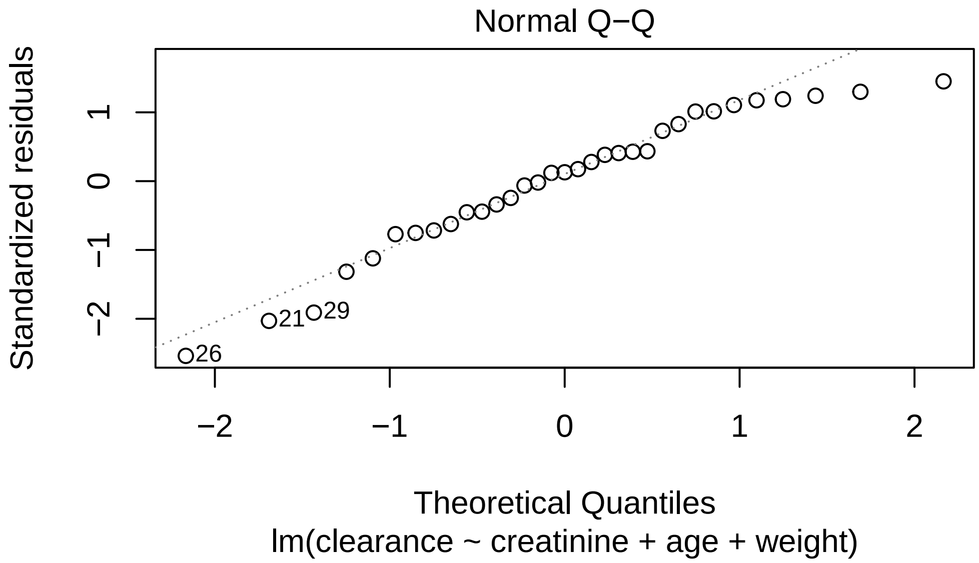
```
lm5 <- lm(clearance ~ creatinine + age + weight + age:weight, data = kidney)
summary(lm5)$adj.r.squared
```

```
[1] 0.8382065
```

```
plot(lm4, 1)
```


Residuals vs Fitted
Residuals
10
−10
−30
29
21
26
20
40
60
80
100
120
Fitted values
lm(clearance ~ creatinine + age + weight)

```
plot(lm4, 2)
```


Normal Q−Q
Standardized residuals
1
0
−1
−2
26
21
29
−2
−1
0
1
2
Theoretical Quantiles
lm(clearance ~ creatinine + age + weight)

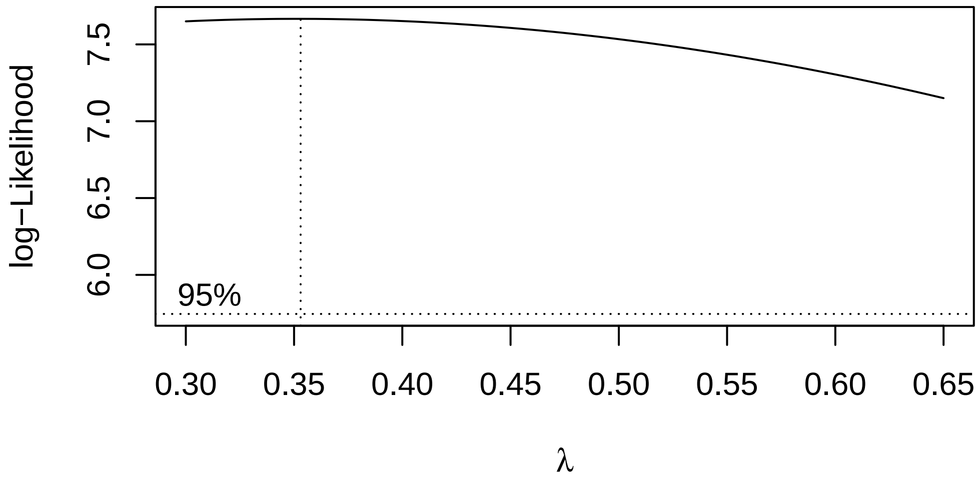
```r
shapiro.test(resid(lm1)) # p-value > 0.05 Normality good
```

```

	Shapiro-Wilk normality test

data:  resid(lm1)
W = 0.94004, p-value = 0.0681
```

From the Residuals vs Fitted values plot there appears to be some fanning pattern of the residuals. However, looking at the Normal QQ plot we do see some Normality in the residuals. There is some slight deviation at the top tail. There also appears to be some outliers but it doesn't appear to be too much cause for concern. Performing a Shapiro-Wilk test though we see that the normality assumption is ok and is satisfied.

(c)

```r
boxcox(lm1, lambda = seq(0.3, 0.65, by = 0.05))
```



```r
summary(powerTransform(lm1))
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.3513           0      -0.2065       0.9091

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df    pval
LR test, lambda = (0) 1.610087  1 0.20448

Likelihood ratio test that no transformation is needed
                           LRT df     pval
LR test, lambda = (1) 4.455553  1 0.034788
```

Using the Box-Cox procedure, the estimated value of the parameter is $\lambda = 0$ The 95% confidence interval for λ is between -0.2042 and 0.9639. Thus, the regression model with the transformed response is $log(\widehat{clearance}) = \beta_0 + \beta_1 creatinine + \beta_2 age + \beta_3 weight$

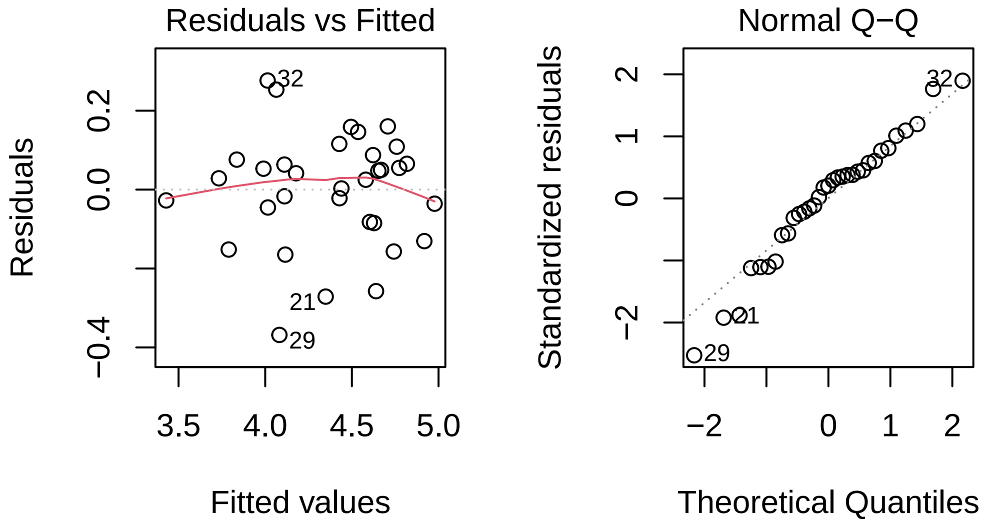
(d)

```
lm2 <- lm(log(clearance) ~ creatinine + age + weight, data = kidney)
summary(lm2)
```

```
Call:
lm(formula = log(clearance) ~ creatinine + age + weight, data = kidney)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36835 -0.08204  0.02860  0.07602  0.27647 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.822976   0.178890  26.961  < 2e-16 ***
creatinine  -0.574258   0.067808  -8.469 2.48e-09 ***
age         -0.008481   0.001712  -4.953 2.89e-05 ***
weight       0.010204   0.002081   4.903 3.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1508 on 29 degrees of freedom
```

```
Multiple R-squared:  0.8765,    Adjusted R-squared:  0.8638
F-statistic: 68.63 on 3 and 29 DF,  p-value: 2.782e-13
```

```
par(mfrow = c(1,2))
plot(lm2, 1:2)
```



```
glm1 <- glm(log(clearance) ~ creatinine + age + weight, data = kidney)

# Perform backward stepwise using AIC
glm_sel <- step(glm1)
```

```
Start:  AIC=-25.46
log(clearance) ~ creatinine + age + weight

             Df Deviance      AIC
<none>           0.65979 -25.4574
- weight      1  1.20662  -7.5366
- age         1  1.21804  -7.2259
- creatinine  1  2.29155  13.6297
```

```r
summary(glm_sel)
```

```

Call:
glm(formula = log(clearance) ~ creatinine + age + weight, data = kidney)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.36835  -0.08204   0.02860   0.07602   0.27647

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.822976   0.178890  26.961  < 2e-16 ***
creatinine  -0.574258   0.067808  -8.469 2.48e-09 ***
age         -0.008481   0.001712  -4.953 2.89e-05 ***
weight       0.010204   0.002081   4.903 3.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.0227513)

    Null deviance: 5.34438  on 32  degrees of freedom
Residual deviance: 0.65979  on 29  degrees of freedom
AIC: -25.457

Number of Fisher Scoring iterations: 2
```

```r
AIC(glm1, glm_sel)
```

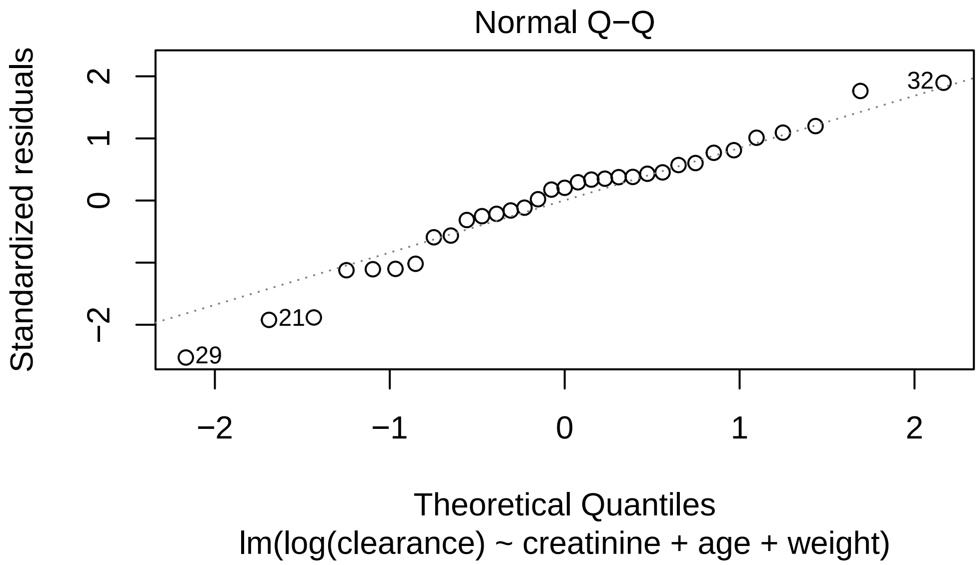
```
        df       AIC
glm1     5 -25.45743
glm_sel  5 -25.45743
```

I'm pretty sure I messed up on part (d).

I know that you are supposed to pick the model with the lowest AIC however both models yield the same results with an AIC of 265.86.

(e)

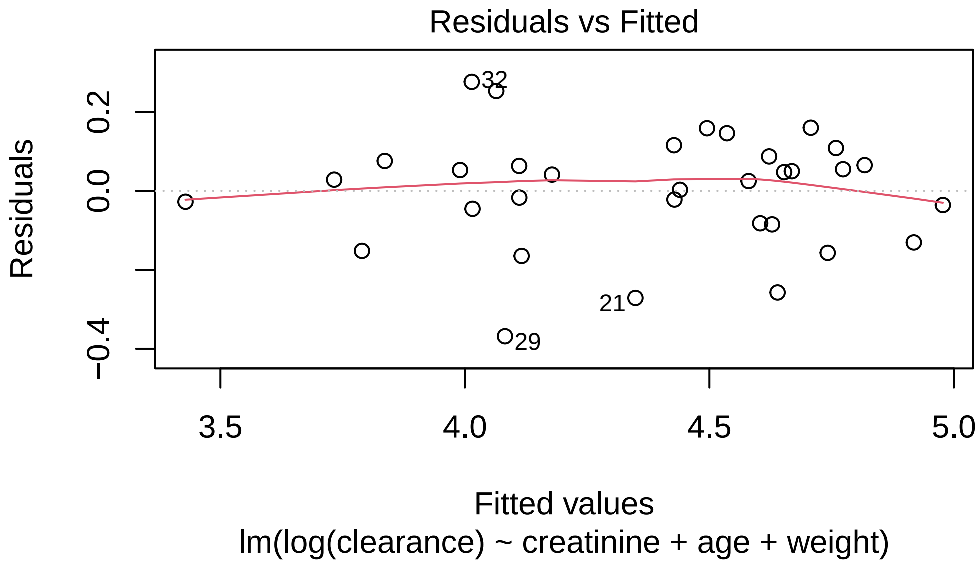
```
lm6 <- lm(log(clearance) ~ creatinine + age + weight, data = kidney)
plot(lm6, which = 2)
```



```
shapiro.test(resid(lm6))
```

```

    Shapiro-Wilk normality test

data:  resid(lm6)
W = 0.96792, p-value = 0.4249
```

```
plot(lm6, which = 1)
```



```r
summary(lm6)
```

```

Call:
lm(formula = log(clearance) ~ creatinine + age + weight, data = kidney)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36835 -0.08204  0.02860  0.07602  0.27647 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.822976   0.178890  26.961  < 2e-16 ***
creatinine  -0.574258   0.067808  -8.469 2.48e-09 ***
age         -0.008481   0.001712  -4.953 2.89e-05 ***
weight       0.010204   0.002081   4.903 3.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1508 on 29 degrees of freedom
Multiple R-squared:  0.8765,	Adjusted R-squared:  0.8638 
F-statistic: 68.63 on 3 and 29 DF,  p-value: 2.782e-13
```

The appropriate model is $\widehat{log(clearance)} = \beta_0 + \beta_1 creatinine + \beta_2 age + \beta_3 weight$ This is validated through the Normal QQ plot of the residuals. We see that now that the response variable is transformed that the residuals follow the line better. We can test for this as well with the Shapiro-Wilk test and see that normality has improved significantly. The residuals vs fitted values plot also shows improvement of the model.The summary statistics also shows that all predictors of the model are useful in predicting log(clearance) and our adjusted r squared is 0.8638.

(f)

```
new_x <- data.frame(creatinine = 0.8, age = 30, weight = 85)
predict(lm6, newdata = new_x, interval = "prediction")
```

```
       fit      lwr      upr
1 4.976463 4.647963 5.304963
```

The estimated creatinine clearance for a male with creatinine 0.8, aged 30, and weighing 85 kilograms is 4.98. With a 95% confidence interval the creatinine clearance will be between 4.65 and 5.31.