STAT 224 Autumn 2022 HW5

Matthew Zhao

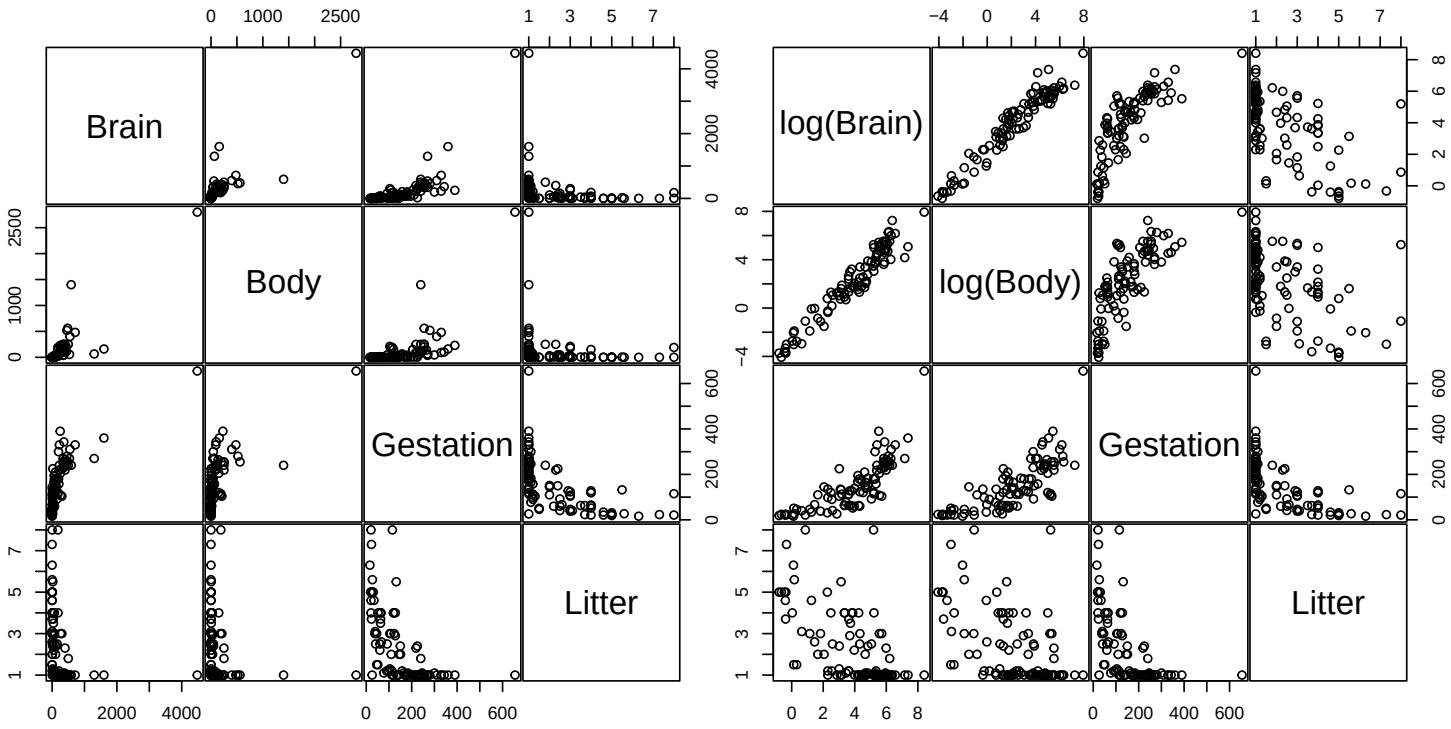
Question 1

http://www.stat.uchicago.edu/~yibi/s224/data/mammals.txt

```
mammals = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/mammals.txt", header=T)
```

Q1a — 4 points

```
pairs(~Brain+Body+Gestation+Litter, data=mammals, gap=1/10, oma=c(2,2,2,2))
pairs(~log(Brain)+log(Body)+Gestation+Litter, data=mammals, gap=1/10, oma=c(2,2,2,2))
# gap = 1/10 reduces the gaps between plots to 1/10 of the default gap width
# oma=c(2,2,2,2) reduces the margins of the plot
```

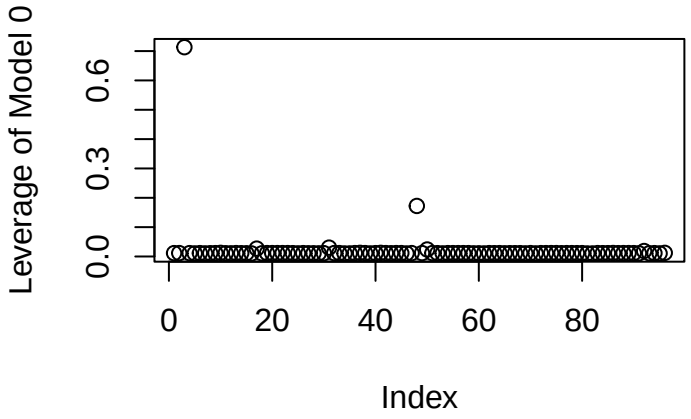


Through the scatterplots, we see evidence of nonlinearity in Brain and Body, suggesting that log transformation may be appropriate to solve this issue.

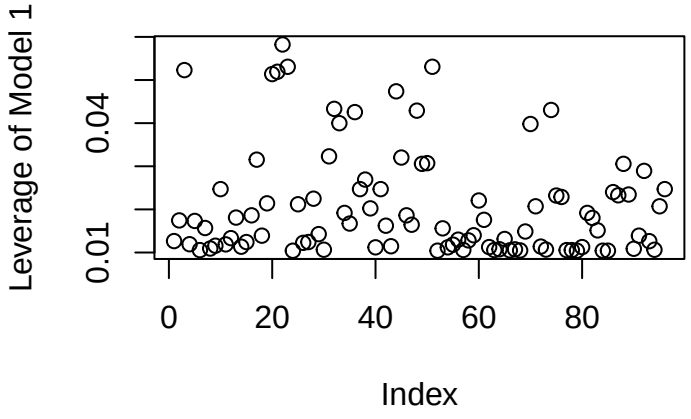
Q1b — 4 points

```
# Model 0
mod0 = lm(Brain ~ Body, data=mammals)
# Model 1
mod1 = lm(log(Brain)~log(Body), data=mammals)
```

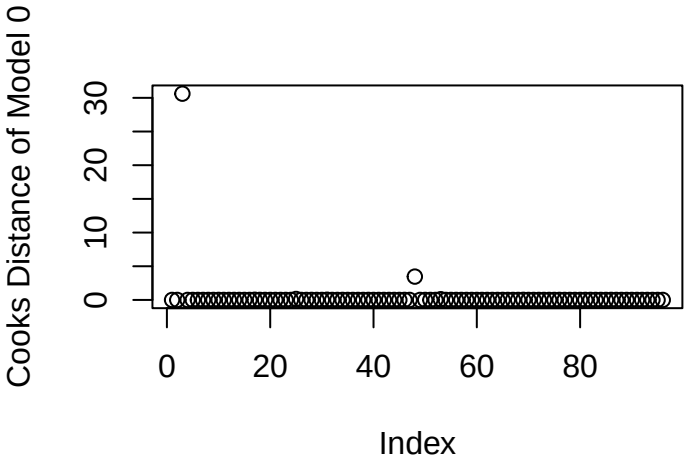
```
plot(hatvalues(mod0), ylab="Leverage of Model 0")
```



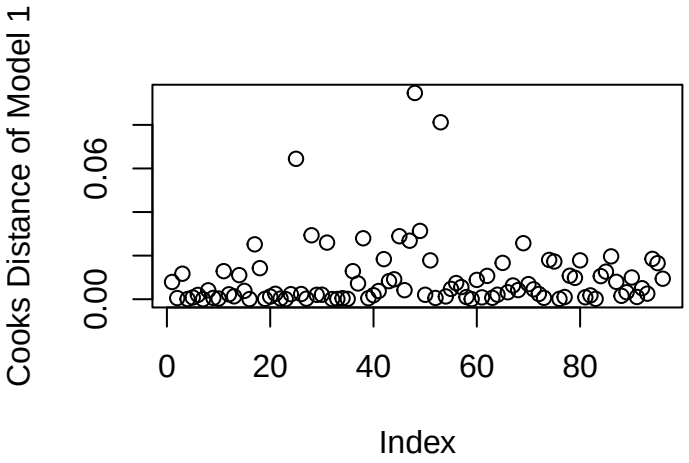

```
plot(hatvalues(mod1), ylab="Leverage of Model 1")
```




```
plot(cooks.distance(mod0), ylab="Cooks Distance of Model 0")
```



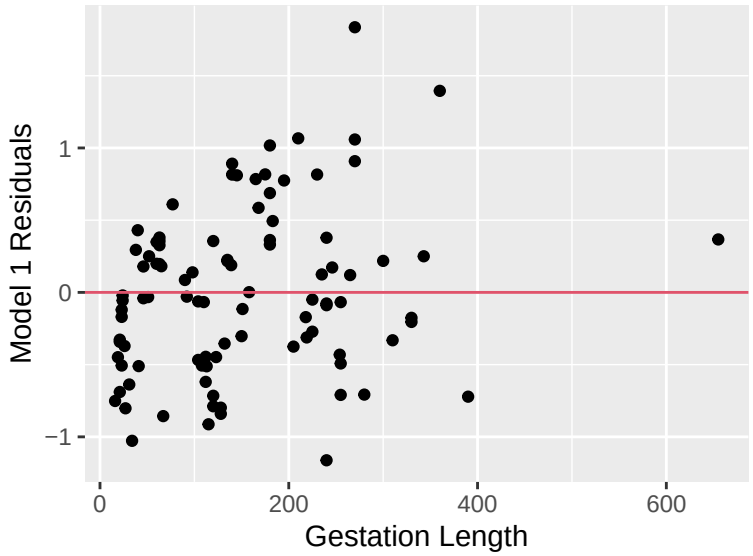

```
plot(cooks.distance(mod1), ylab="Cooks Distance of Model 1")
```



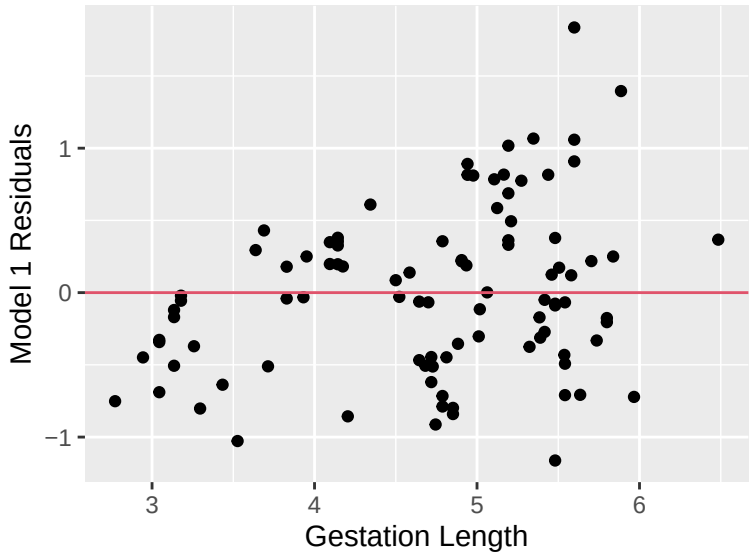
The African Elephant and the Hippopotamus have large leverages in model 0 at 0.7 and 0.1723 respectively and also have large cook's distances at 30.6 and 3.462. After log transformation, the magnitudes of these decrease to be roughly in line with those of other points.

Q1c — 4 points

```
ggplot(data=mammals) +
  geom_point(aes(x=Gestation,y=mod1$res)) +
  xlab('Gestation Length') + ylab('Model 1 Residuals') +
  geom_hline(yintercept = 0, col = 2)
```



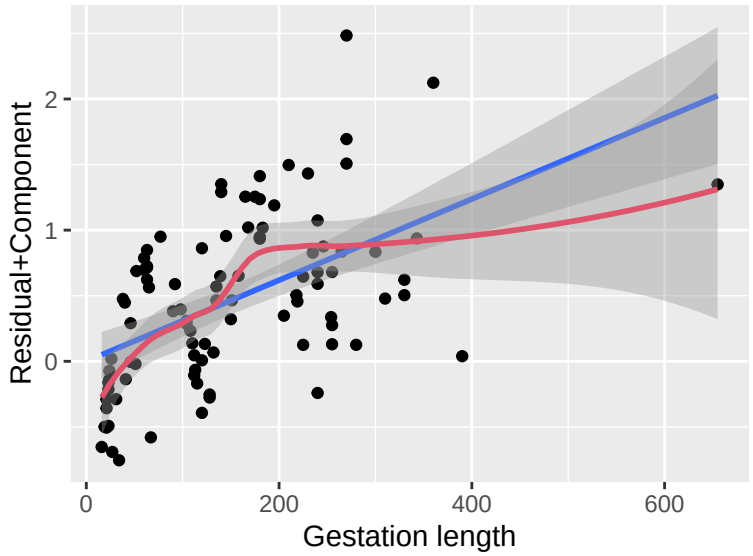

```
ggplot(data=mammals) +
  geom_point(aes(x=log(Gestation),y=mod1$res)) +
  xlab('Gestation Length') + ylab('Model 1 Residuals') +
  geom_hline(yintercept = 0, col = 2)
```



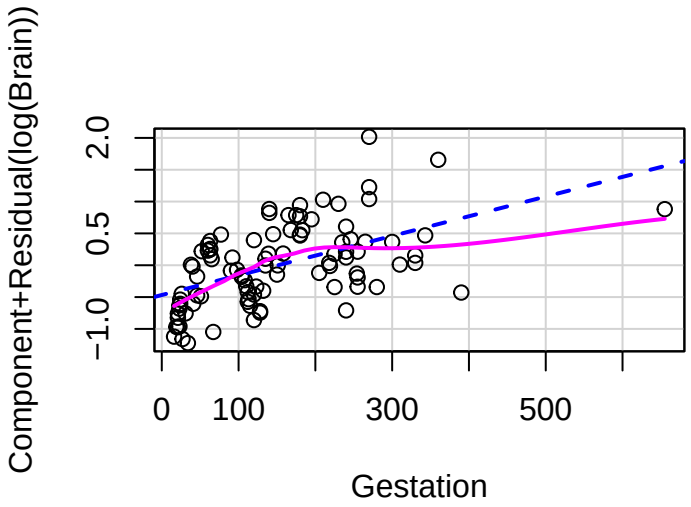
Yes gestation legnth appears to have an effect on mammal brain weight after accounting for mammal body size since the residuals of model 1 appear to increase with gestation and log gestation. It is better to use log gestation since the values are more spread out, improving the model.

Q1d — 6 points

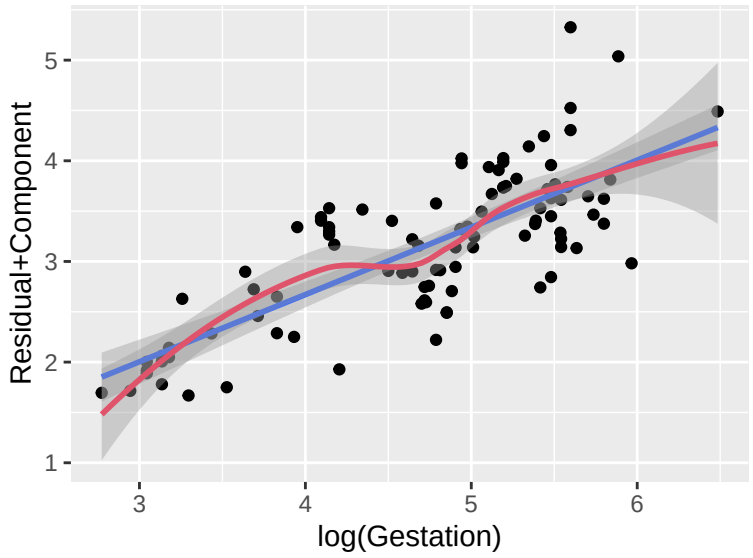
```
# Model 2
mod2 = lm(log(Brain) ~ log(Body) + Gestation, data=mammals)
ggplot(data=mammals,aes(x=Gestation,y=mod2$res+mod2$coefficients[3]*Gestation)) +
  geom_point() + geom_smooth(method='lm') +
  geom_smooth(method='loess',col=2)+
  labs(x="Gestation length",y="Residual+Component")
```



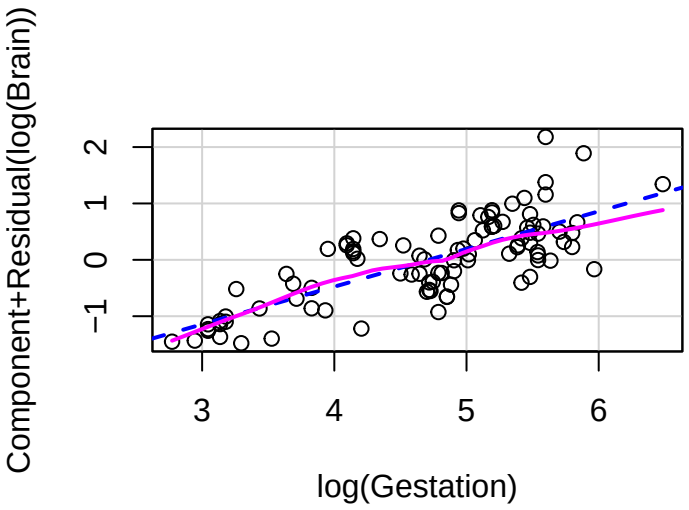

```
library(car)
crPlots(mod2, "Gestation")
```




```
# Model 3
mod3 = lm(log(Brain) ~ log(Body) + log(Gestation), data=mammals)
ggplot(data=mammals,aes(x=log(Gestation),y=mod3$res+mod3$coefficients[3]*log(Gestation))) +
  geom_point() + geom_smooth(method='lm') +
  geom_smooth(method='loess',col=2)+
  labs(x="log(Gestation)",y="Residual+Component")
```




```
crPlots(mod3, "log(Gestation)")
```

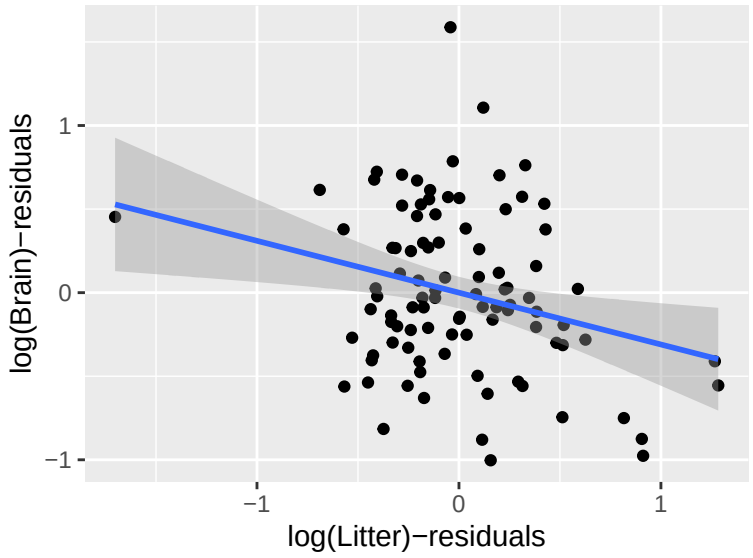


log(Gestation) since Gestation appears to be nonlinear.

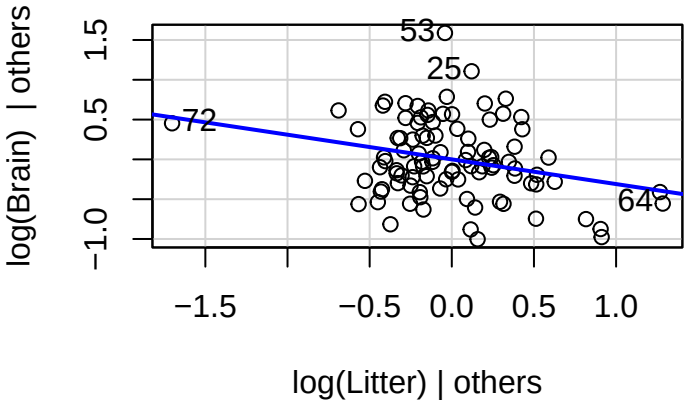
Q1e — 6 points

```
# Model 4
mod4 = lm(log(Brain) ~ log(Body) + log(Gestation) + log(Litter), data=mammals)
```

```
r1 = lm(log(Brain) ~ log(Body) + log(Gestation), data=mammals)$res
r2 = lm(log(Litter) ~ log(Body) + log(Gestation), data=mammals)$res
ggplot(data.frame(r1,r2),aes(x=r2,y=r1)) +
  geom_point() + geom_smooth(method='lm') +
  labs(x='log(Litter)-residuals',y='log(Brain)-residuals')
```

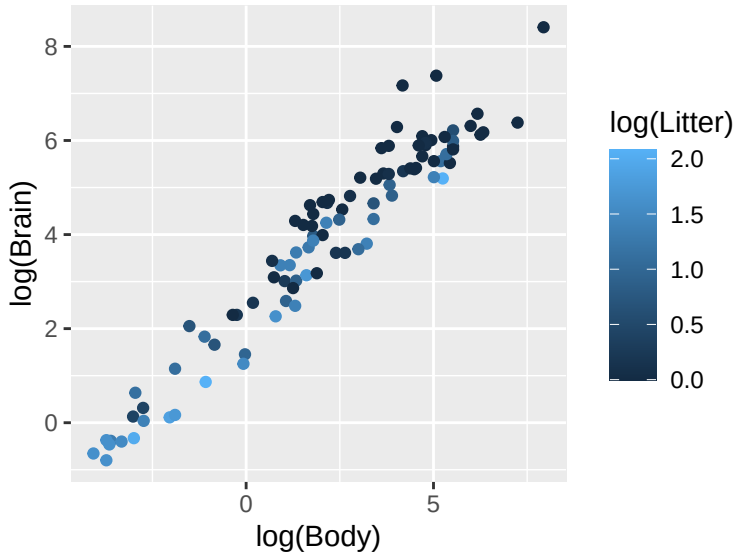



```
avPlots(mod4,"log(Litter)")
```

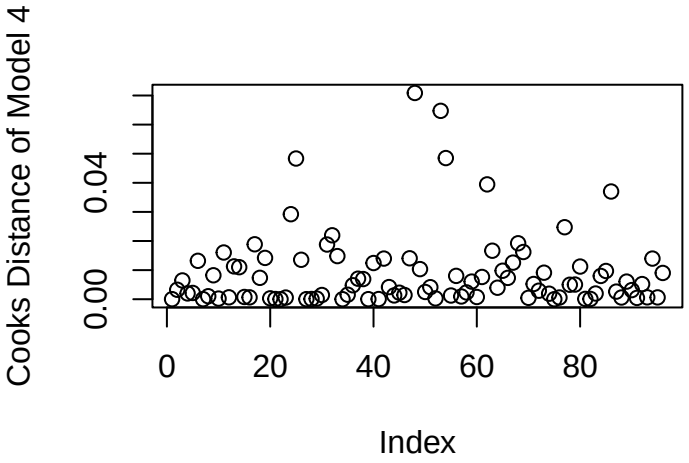


Nutria and Quokka have high leverage while Dolphin and Human being have high residuals.

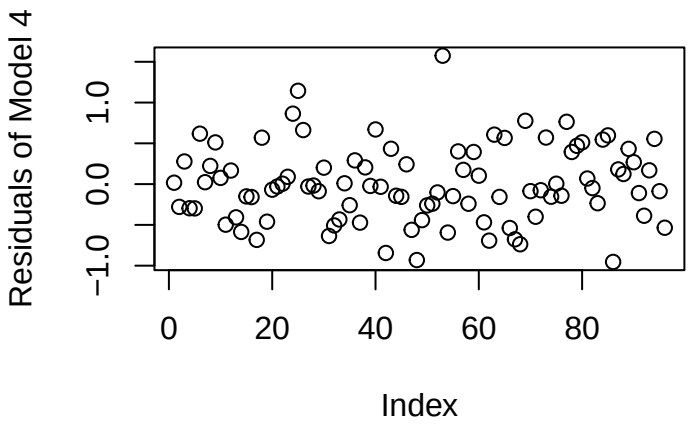
```
ggplot(data=mammals,aes(x=log(Body),y=log(Brain),color=log(Litter))) +
   geom_point()
```




```
plot(cooks.distance(mod4), ylab="Cooks Distance of Model 4")
```




```
plot(mod4$residuals, ylab="Residuals of Model 4")
```



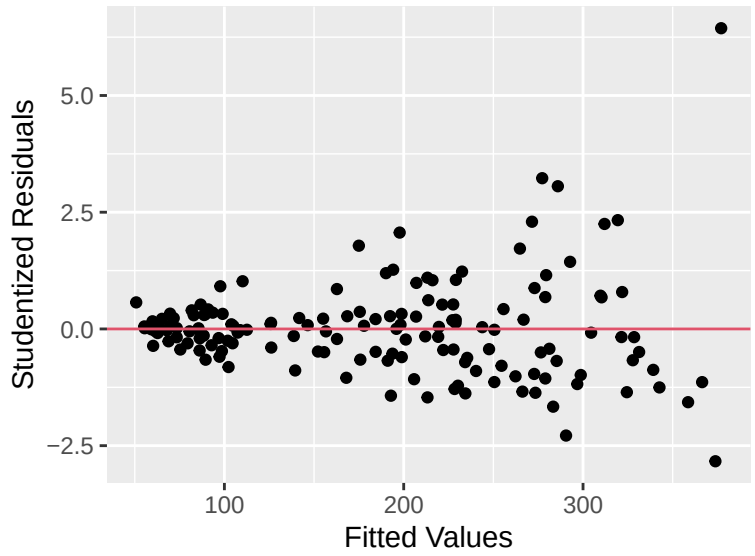
Question 2

http://www.stat.uchicago.edu/~yibi/s224/data/P151-153.txt

```
eduexp = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/P151-153.txt", header=T)
eduexp$region = factor(eduexp$region, labels=c("Northeast","Central","South","West"))
eduexp$year = as.factor(eduexp$year)
lm1 = lm(y ~ x1*(year+region), data=eduexp)
```

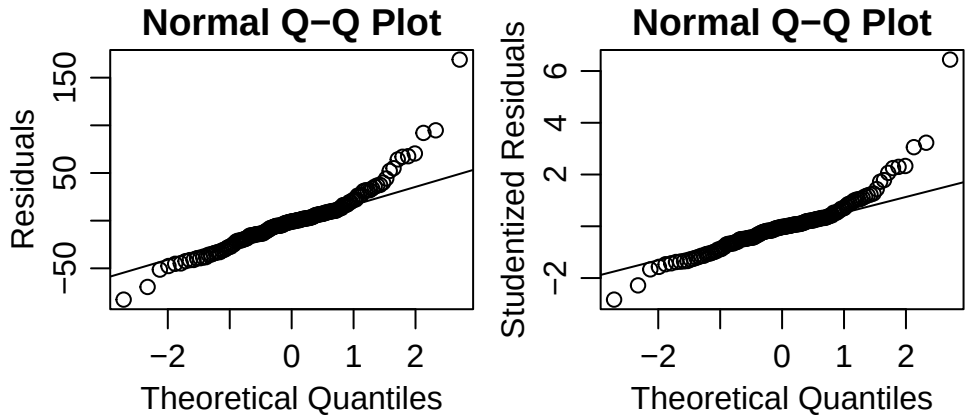
Q2a — 4 points

```
ggplot(data=eduexp,aes(x=lm1$fitted.values,y=rstudent(lm1))) +
  geom_point() +
  xlab('Fitted Values') + ylab('Studentized Residuals') +
  geom_hline(yintercept = 0, col = 2)
```



It appears that the errors are correlated with the fitted values, specifically residuals increase with fitted values. This means nonconstant variance in the errors.

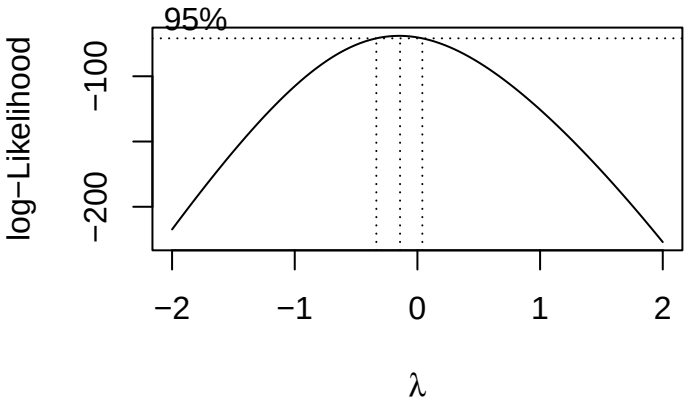
Q2b — 3 points



The residuals are somewhat right skewed.

Q2c — 3 points

```
library(MASS)
boxcox(lm1)
```

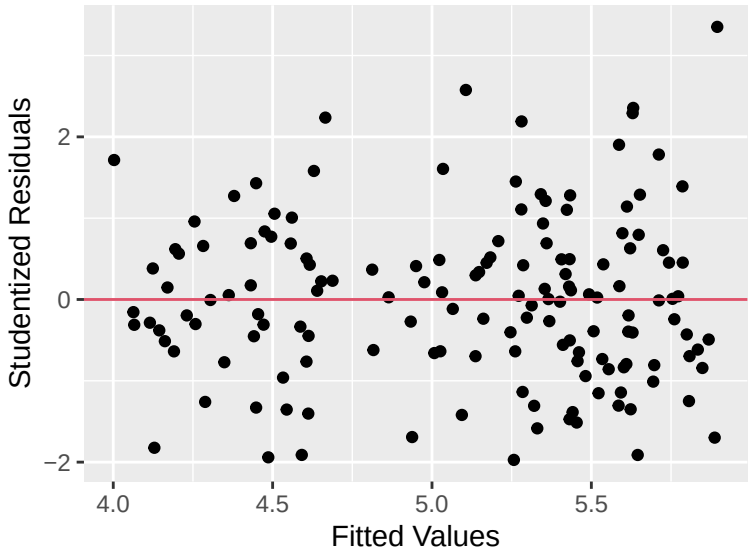


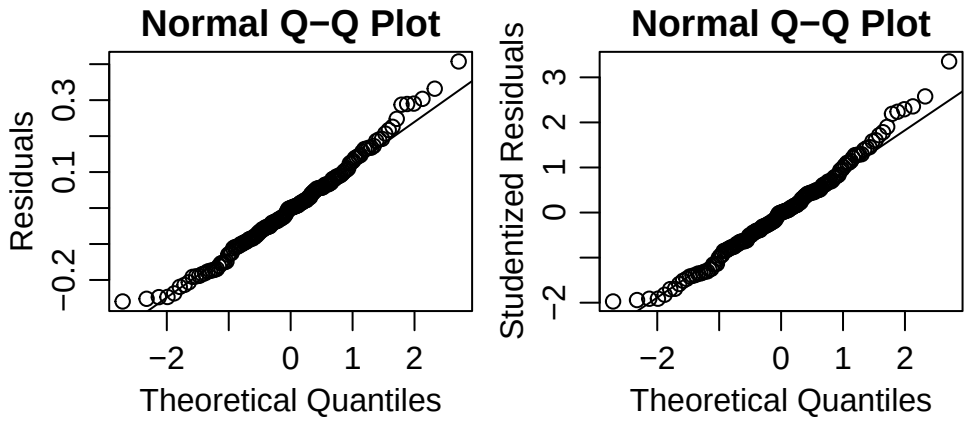
Since 0 falls within the 95% confidence interval for λ, log transformation is most appropriate.

Q2d — 4 points

```
lm2 = lm(log(y) ~ x1*(year+region), data=eduexp)
```

```
ggplot(data=eduexp,aes(x=lm2$fitted.values,y=rstudent(lm2))) +
  geom_point() +
  xlab('Fitted Values') + ylab('Studentized Residuals') +
  geom_hline(yintercept = 0, col = 2)
```





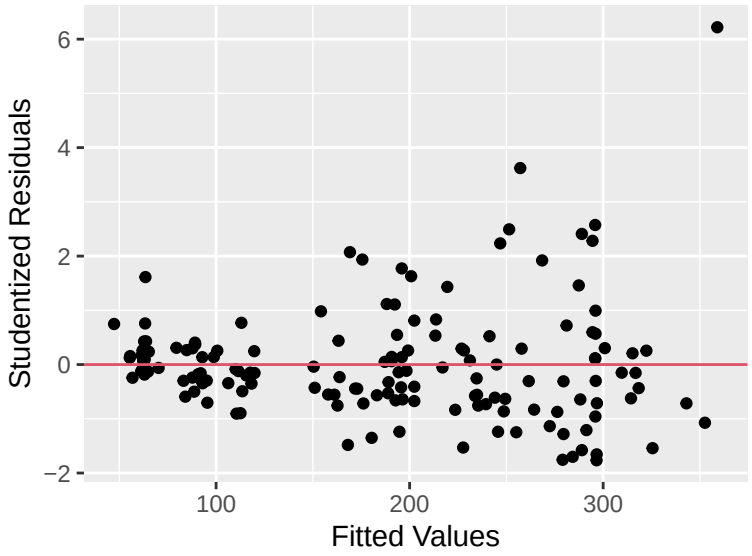
Model assumptions are no longer violated.

Question 3

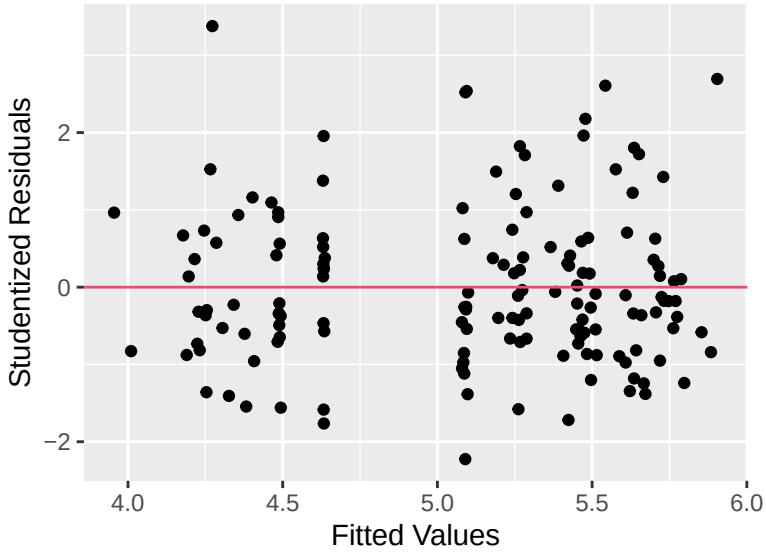
Q3a — 4 points

```
lm3 = lm(y ~ x2*(year+region), data=eduexp)
lm4 = lm(log(y) ~ x2*(year+region), data=eduexp)
```

```
ggplot(data=eduexp,aes(x=lm3$fitted.values,y=rstudent(lm3))) +
  geom_point() +
  xlab('Fitted Values') + ylab('Studentized Residuals') +
  geom_hline(yintercept = 0, col = 2)
```



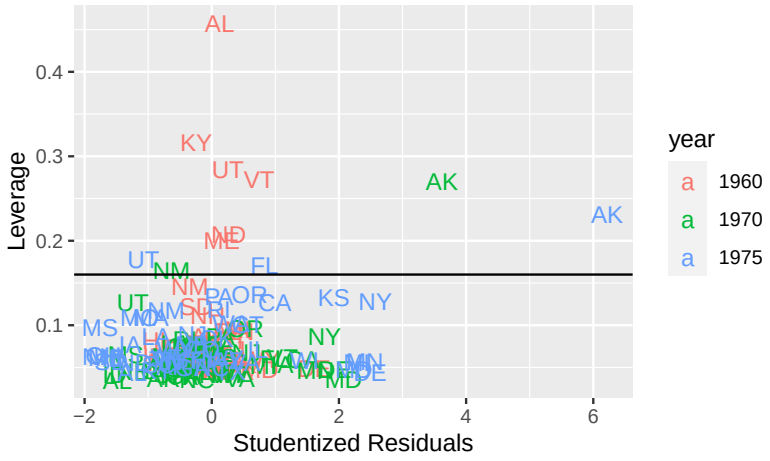

```
ggplot(data=eduexp,aes(x=lm4$fitted.values,y=rstudent(lm4))) +
  geom_point() +
  xlab('Fitted Values') + ylab('Studentized Residuals') +
  geom_hline(yintercept = 0, col = 2)
```



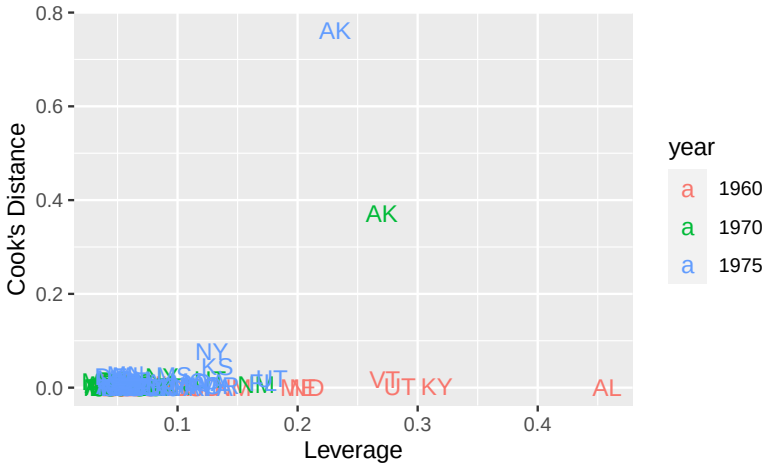
We use log transformation because we see nonconstant variance in the residuals vs fitted values plot with residuals increasing with fitted values.

Q3b — 4 points

```
library(ggplot2)
ggplot(eduexp, aes(x=rstudent(lm3), y=hatvalues(lm3), label=state, color=year)) +
  geom_text() +
  labs(x = "Studentized Residuals", y = "Leverage") +
  geom_hline(yintercept = 2*12/150)
```

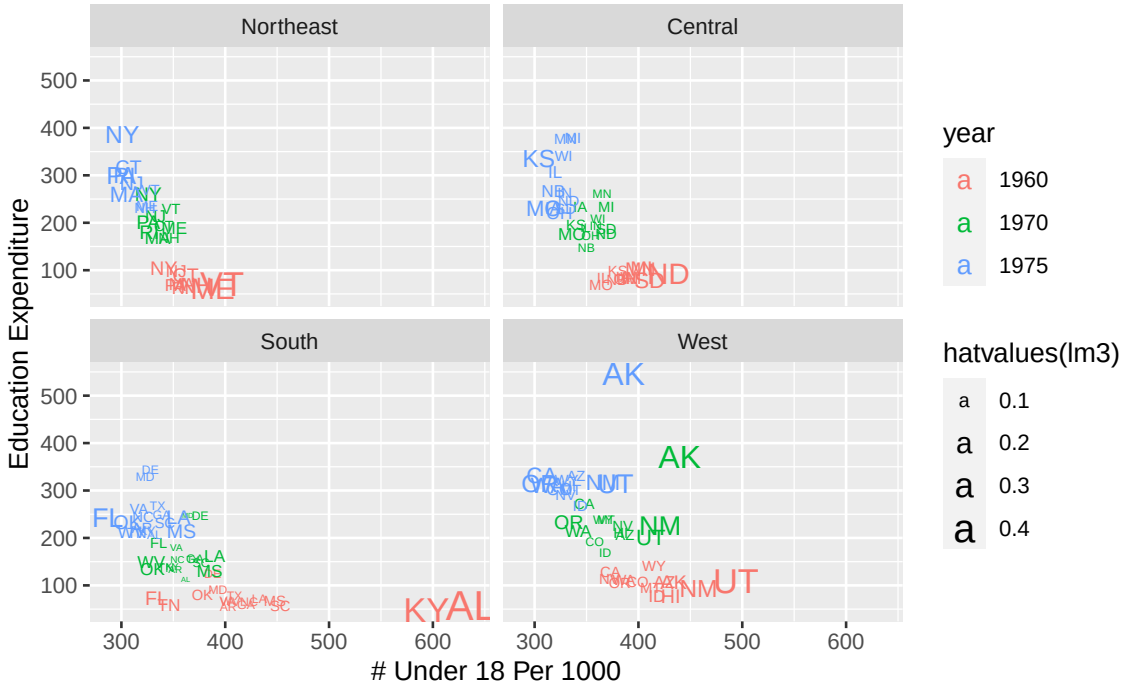



```
ggplot(eduexp, aes(x=hatvalues(lm3), y=cooks.distance(lm3), label=state, color=year)) +
  geom_text() +
  labs(x = "Leverage", y = "Cook's Distance")
```

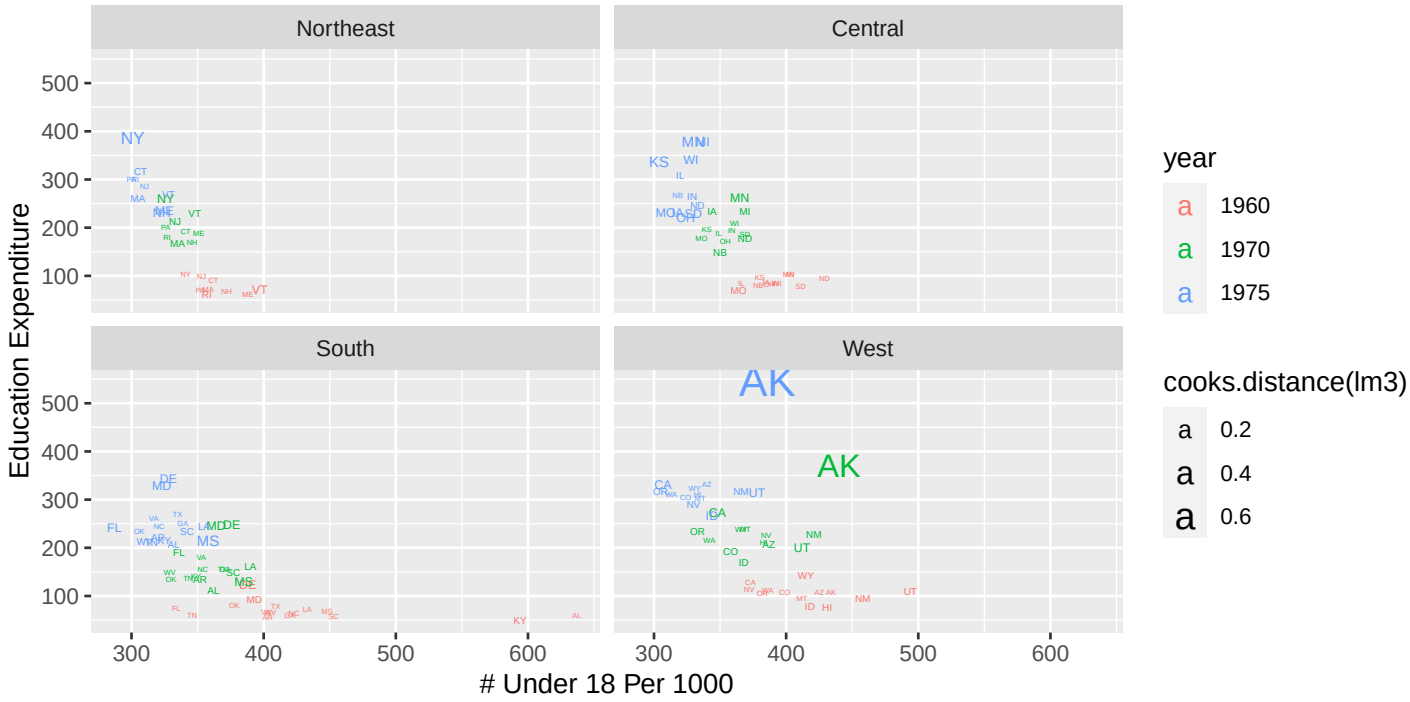


Alaska has the largest studentized residual in 1975. Alabama had the largest leverage in 1960.

```
ggplot(eduexp, aes(x = x2, y = y, color=year, label=state, size=hatvalues(lm3))) +
  geom_text() +
  facet_wrap(~region) +
  xlab("# Under 18 Per 1000") +
  ylab("Education Expenditure")
```



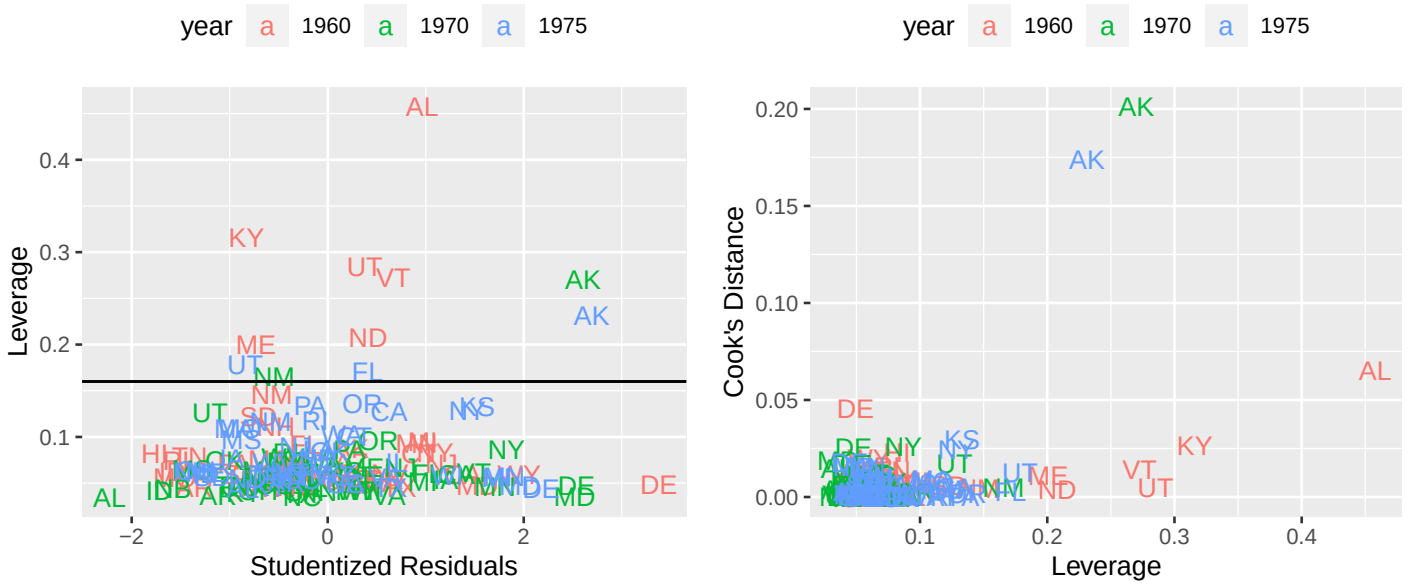

```
ggplot(eduexp, aes(x = x2, y = y, color=year, label=state, size=cooks.distance(lm3))) +
  geom_text() +
  facet_wrap(~region) +
  xlab("# Under 18 Per 1000") +
  ylab("Education Expenditure")
```



Yes, they are influential since they represent a value with high x value compared with the other observations. The influential points have large cooks distance.

Q3c — 4 points

```
ggplot(eduexp, aes(x=rstudent(lm4), y=hatvalues(lm4), label=state, color=year)) +
  geom_text() + geom_hline(yintercept = 2*12/150) +
  labs(x = "Studentized Residuals", y = "Leverage") +
  geom_hline(yintercept = 2*12/150) + theme(legend.position="top")
ggplot(eduexp, aes(x=hatvalues(lm4), y=cooks.distance(lm4), label=state, color=year)) +
  geom_text() + labs(x = "Leverage", y = "Cook's Distance") + theme(legend.position="top")
```



The leverage is roughly similar while the Cooks distance is significantly less. The points become less influential.