STAT 224 Autumn 2022 HW6

Matthew Zhao

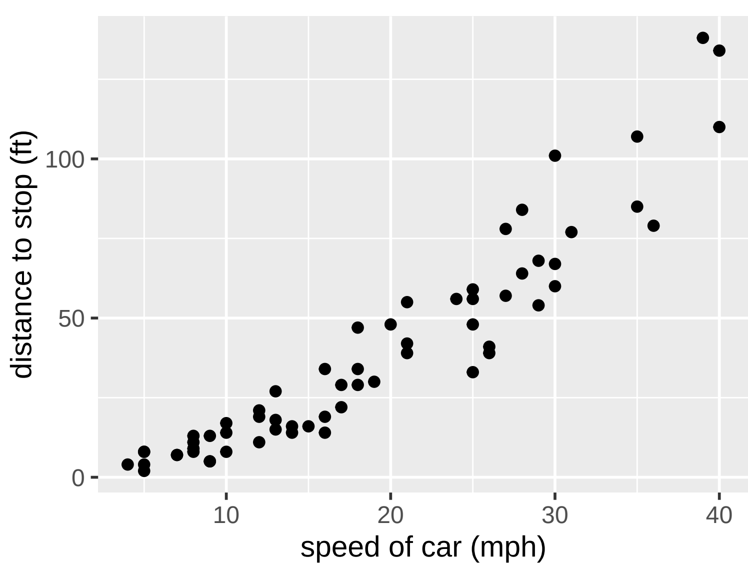
Question 1

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

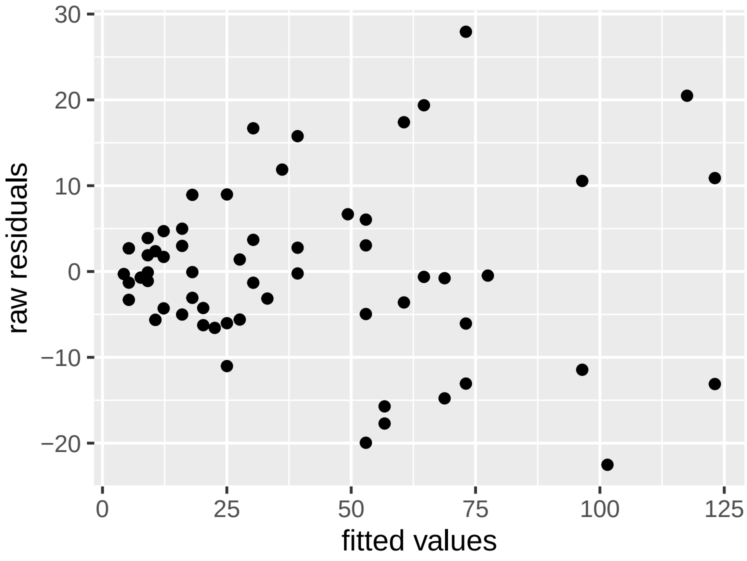
```
brake = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/brake.txt", header=T)
```

Q1a — 3 points

```
ols1 = lm(distance ~ speed + I(speed^2), data=brake)
ggplot(data=brake,aes(x=speed,y=distance)) +
  geom_point() + labs(x='speed of car (mph)',y='distance to stop (ft)')
```




```
ggplot(data=brake,aes(x=ols1$fitted.values,y=ols1$residuals)) +
  geom_point() + labs(x='fitted values',y='raw residuals')
```

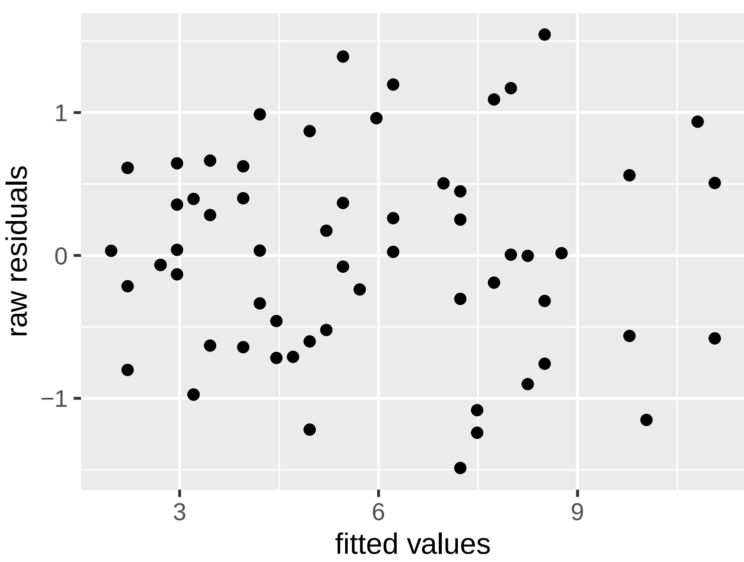


Yes, there is clearly nonconstant variance.

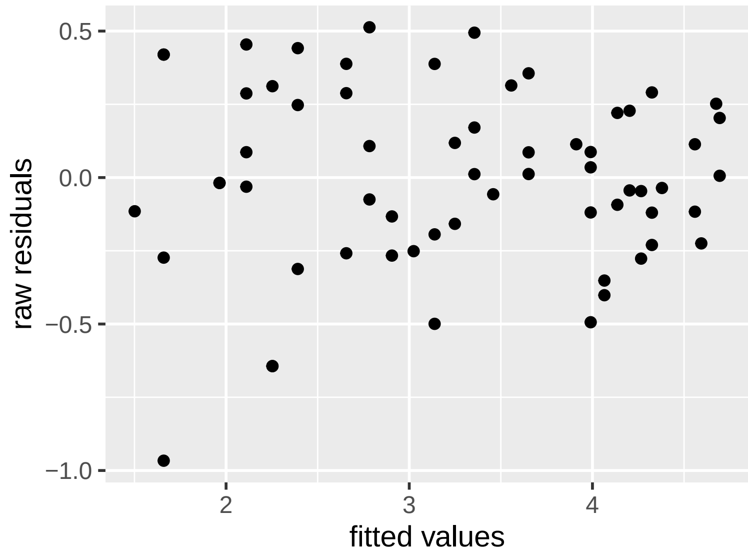
Q1b — 3 points

```
ols2 = lm(sqrt(distance) ~ speed + I(speed^2), data=brake)
ols3 = lm(log(distance) ~ speed + I(speed^2), data=brake)
```

```
ggplot(data=brake,aes(x=ols2$fitted.values,y=ols2$residuals)) +
  geom_point() + labs(x='fitted values',y='raw residuals')
```



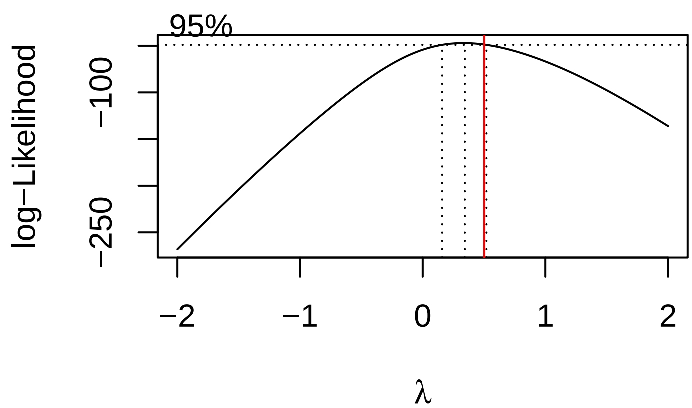

```
ggplot(data=brake,aes(x=ols3$fitted.values,y=ols3$residuals)) +
  geom_point() + labs(x='fitted values',y='raw residuals')
```



Square-root appears to be the most appropriate transformation since this transformation results in near constant variance based on the residual plots.

Q1c — 2 points

```
library(MASS)
boxcox(ols1)
abline(v=1/2, col="red")
```

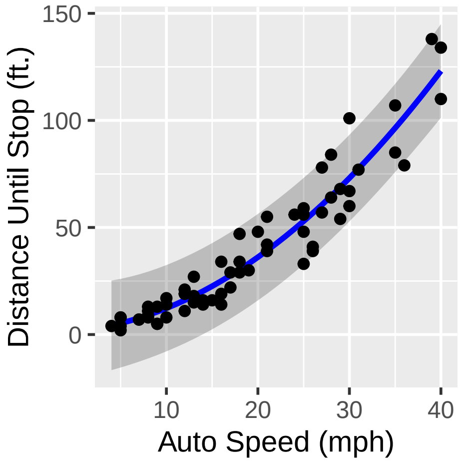


Since $\frac{1}{2}$ falls within the 95% CI, squart root is the most appropriate transformation.

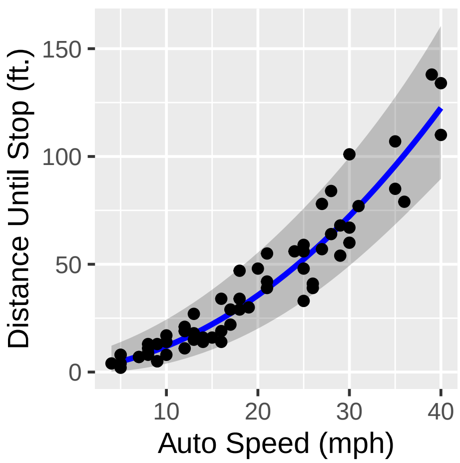
Q1d — 2 points

```
predCI = predict(ols1, data.frame(speed = 4:40), interval="prediction")
predCI = data.frame(x=4:40,predCI)
```

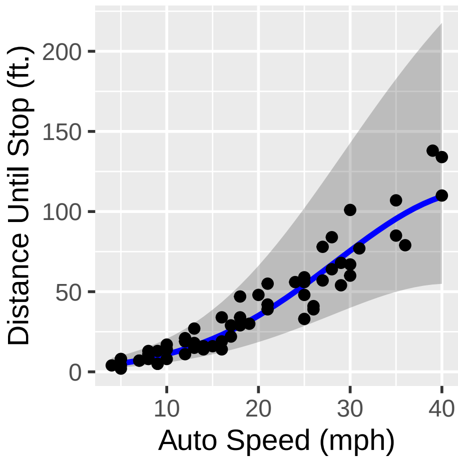
```
ggplot() +
  geom_ribbon(data=predCI, aes(x=x, ymin=lwr, ymax = upr), alpha=0.25) +
  geom_line(data=predCI, aes(x=x,y = fit),col="blue",lwd=1) +
  geom_point(data=brake, aes(x=speed, y=distance)) +
  labs(x="Auto Speed (mph)", y="Distance Until Stop (ft.)")
```




```
predCI = predict(ols2, data.frame(speed = 4:40), interval="prediction")^2
predCI = data.frame(x=4:40,predCI)
ggplot() +
  geom_ribbon(data=predCI, aes(x=x, ymin=lwr, ymax = upr), alpha=0.25) +
  geom_line(data=predCI, aes(x=x,y = fit),col="blue",lwd=1) +
  geom_point(data=brake, aes(x=speed, y=distance)) +
  labs(x="Auto Speed (mph)", y="Distance Until Stop (ft.)")
```




```
predCI = exp(predict(ols3, data.frame(speed = 4:40), interval="prediction"))
predCI = data.frame(x=4:40,predCI)
ggplot() +
  geom_ribbon(data=predCI, aes(x=x, ymin=lwr, ymax = upr), alpha=0.25) +
  geom_line(data=predCI, aes(x=x,y = fit),col="blue",lwd=1) +
  geom_point(data=brake, aes(x=speed, y=distance)) +
  labs(x="Auto Speed (mph)", y="Distance Until Stop (ft.)")
```



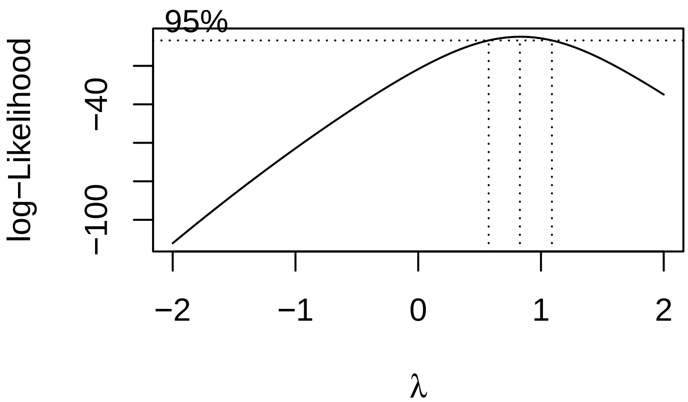
The prediction band for ols2 best matches the pattern of the data.

Q1e — 4 points

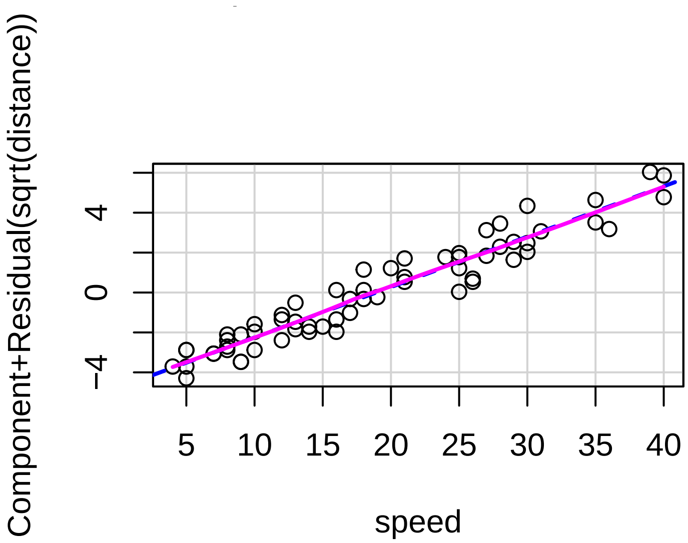
```
summary(ols2)
##
## Call:
## lm(formula = sqrt(distance) ~ speed + I(speed^2), data = brake)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4878 -0.5764  0.0106  0.5064  1.5458
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 0.979036   0.373620    2.62      0.011
## speed       0.246611   0.040741    6.05 0.00000011
## I(speed^2)  0.000141   0.000954    0.15      0.883
##
## Residual standard error: 0.727 on 59 degrees of freedom
## Multiple R-squared:  0.925,  Adjusted R-squared:  0.923
## F-statistic:  365 on 2 and 59 DF,  p-value: <2e-16
```

It is not significant.

```
ols2_nosquare = lm(sqrt(distance) ~ speed, data=brake)
boxcox(ols2_nosquare)
```




```
library(car)
crPlots(ols2_nosquare,'speed')
```



The 95% CI for box-cox contains 1 indicating that no further transformation is needed. Additionally we see no signs of nonlinearity via the residual plus component plot for our only covariate speed.

Q1f — 3 points

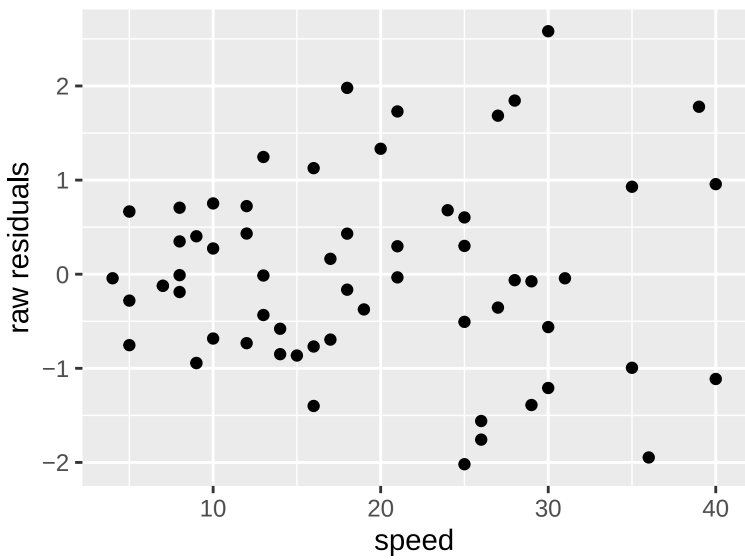
```
predCI = predict(ols2, data.frame(speed = seq(10,30,10)), interval="prediction")^2
data.frame(speed=seq(10,30,10),predCI)
##   speed   fit    lwr   upr
## 1    10 11.97  3.934 24.36
## 2    20 35.61 20.150 55.45
## 3    30 72.32 49.324 99.70
```

Question 2

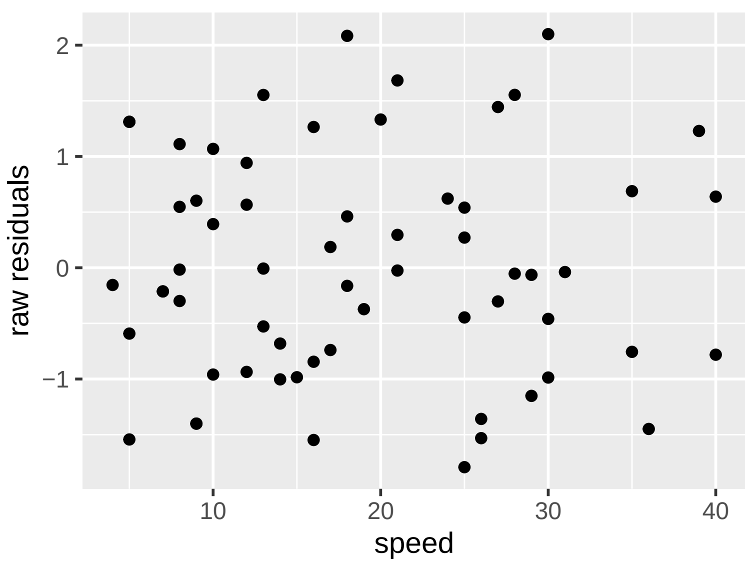
Q2a — 4 points

```
wls1 = lm(distance ~ speed + I(speed^2), data=brake, weight=1/speed)
wls2 = lm(distance ~ speed + I(speed^2), data=brake, weight=1/speed^2)
```

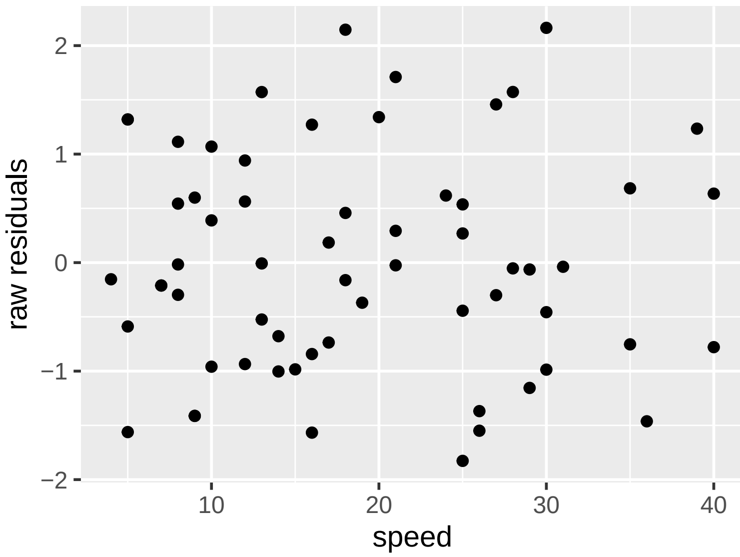
```
ggplot(data=brake,aes(x=speed,y=rstandard(wls1))) +
  geom_point() +
  labs(x='speed',y='raw residuals')
```




```
ggplot(data=brake,aes(x=speed,y=rstandard(wls2))) +
  geom_point() +
  labs(x='speed',y='raw residuals')
```



```
ggplot(data=brake,aes(x=speed,y=rstudent(wls2))) +
  geom_point() +
  labs(x='speed',y='raw residuals')
```



We should use the standardized (internally standardized) residuals for model 2 since model 2 resolves nonconstant variance.

Q2b — 2 points

```
predCI = predict(wls2, data.frame(speed = seq(10,30,10)), interval="prediction")
data.frame(speed=seq(10,30,10),predCI)
##   speed   fit   lwr   upr
## 1    10 12.26 10.38 14.14
## 2    20 36.13 33.14 39.12
## 3    30 73.11 67.80 78.41
```

Question 3

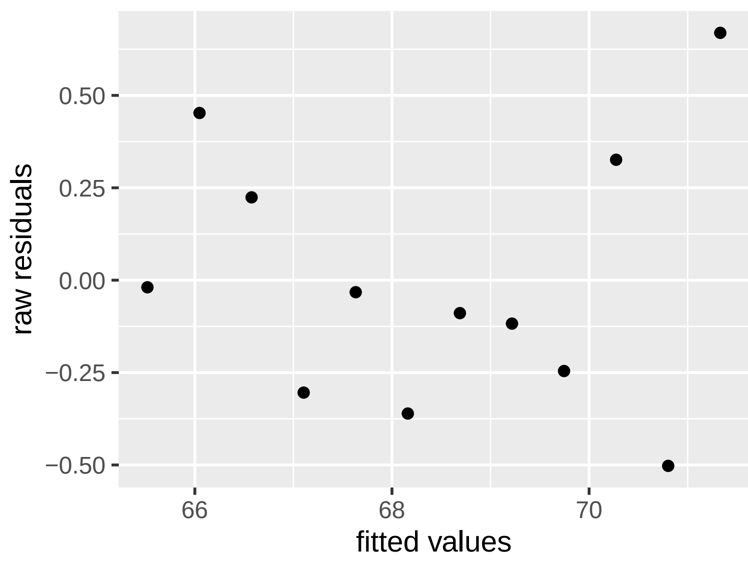
http://www.stat.uchicago.edu/~yibi/s224/data/fatherson.txt

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

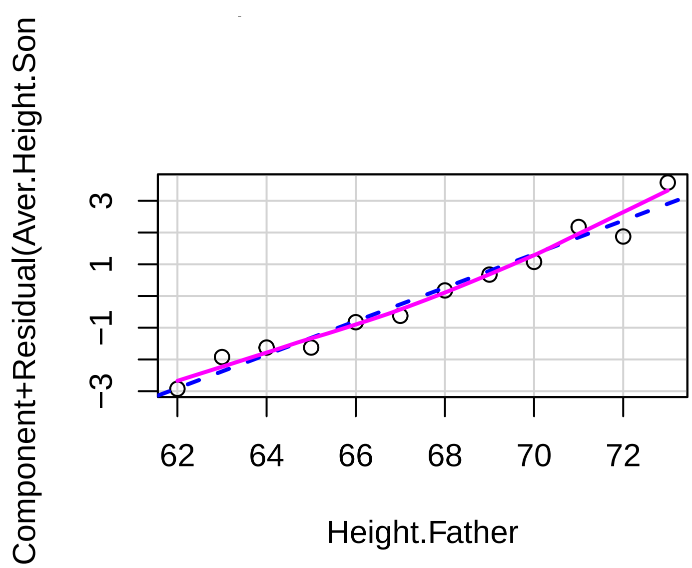
Q3a — 3 points

```
lm1 = lm(Aver.Height.Son ~ Height.Father, data=fatherson)
```

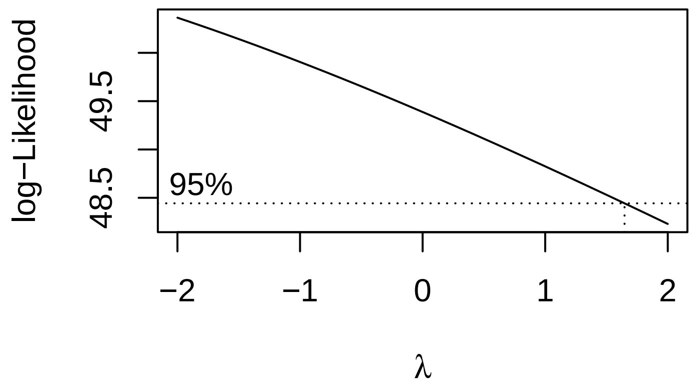
```
ggplot(data=fatherson,aes(x=lm1$fitted.values,y=lm1$residuals)) +
  geom_point() +
  labs(x='fitted values',y='raw residuals')
```




```
crPlots(lm1,'Height.Father')
```



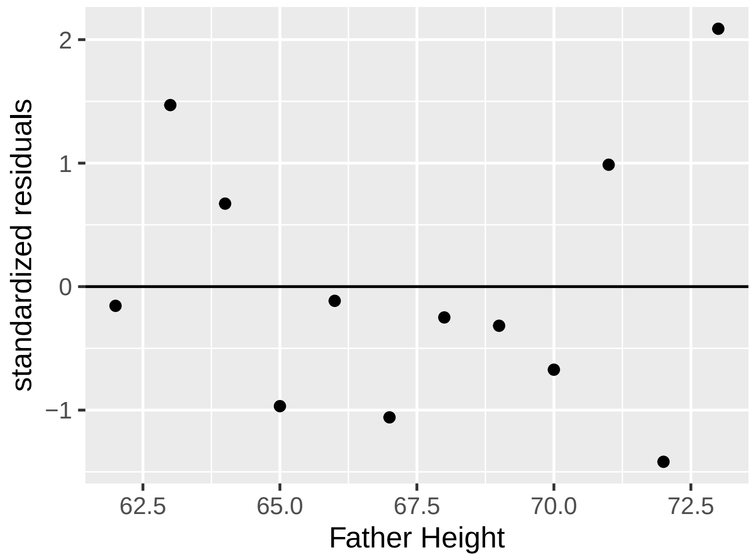

```
boxcox(lm1)
```



OLS is not appropriate since the residual plot reveals that nonconstant variance is violated. Additionally, boxcox shows that there are no ideal transformations of the response to solve this issue since many λs fall within the 95% CI.

Q3b — 6 points

```
wls3 = lm(Aver.Height.Son ~ Height.Father, data=fatherson,
          weight=1/(Height.Father^2))
summary(wls3)
##
## Call:
## lm(formula = Aver.Height.Son ~ Height.Father, data = fatherson,
##     weights = 1/(Height.Father^2))
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00670 -0.00372 -0.00098  0.00366  0.00950
##
## Coefficients:
##                Estimate Std. Error t value     Pr(>|t|)
## (Intercept)     33.0539     2.0234    16.3 0.0000000154
## Height.Father    0.5240     0.0301    17.4 0.0000000083
##
## Residual standard error: 0.00534 on 10 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.965
## F-statistic:  303 on 1 and 10 DF,  p-value: 0.00000000828
ggplot(data = fatherson,aes(x=Height.Father,y=rstandard(wls3))) +
  geom_point() + labs(x='Father Height',y='standardized residuals') +
  geom_hline(yintercept = 0)
```



Q3c — 2 points

```
predict(wls3, data.frame(Height.Father=70), weights=1/Height.Father^2,
interval="confidence")
##     fit   lwr   upr
## 1 69.73 69.43 70.03
```

Question 4

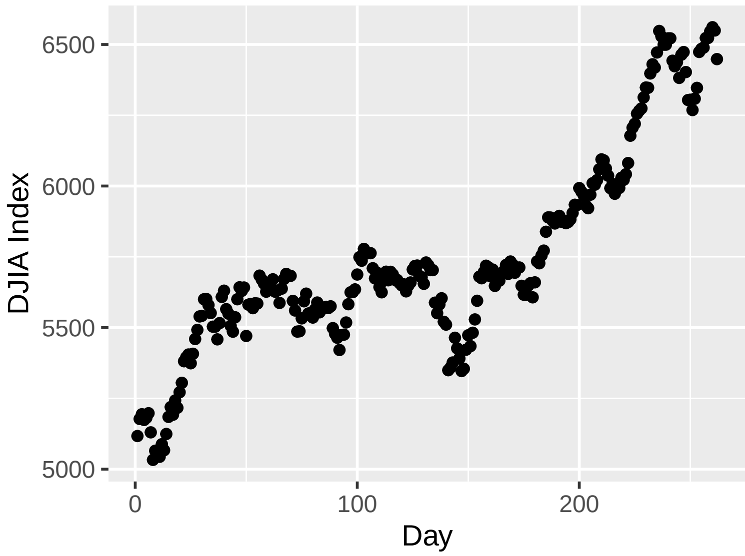
http://www.stat.uchicago.edu/~yibi/s224/data/P229-30.txt

```
stock = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/P229-30.txt", header=T)
```

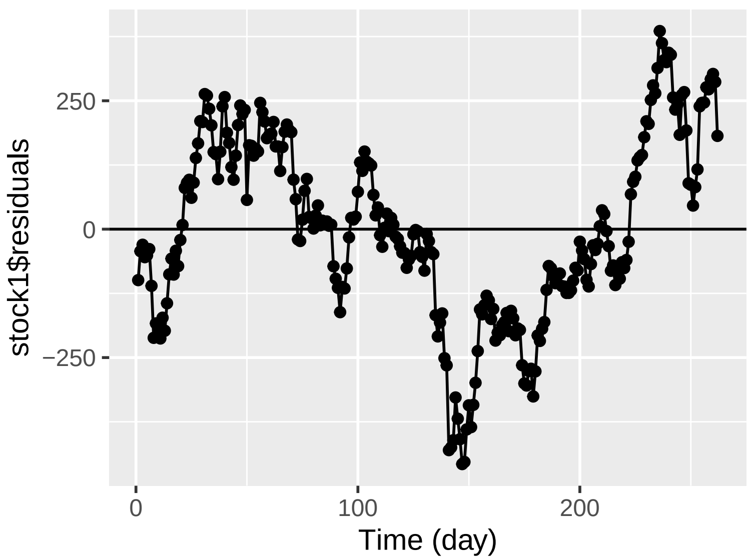
Q4a — 8 points

```
stock1 = lm(DJIA ~ Day, data=stock)
```

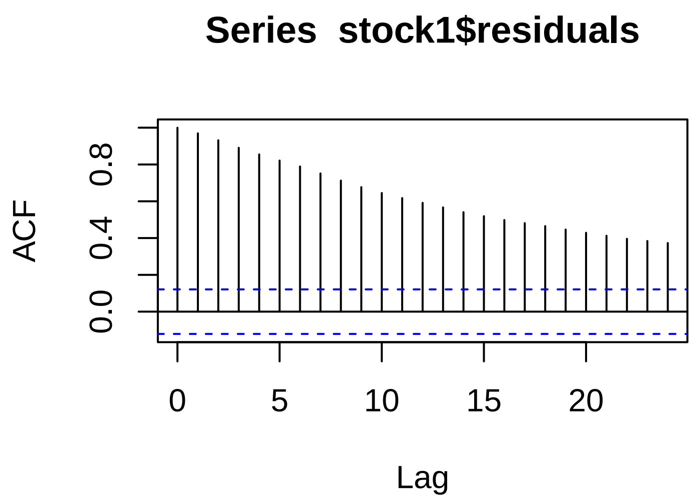
```
ggplot(data=stock,aes(x=Day,y=DJIA)) +
  geom_point() + labs(x='Day',y='DJIA Index')
```




```r
ggplot(data=stock,aes(x=Day,y=stock1$residuals)) +
  geom_line() + geom_point()+
  labs(x='Time (day)','Raw Residuals') +
  geom_hline(yintercept = 0)
```




```r
library(tseries)
runs.test(factor(stock1$residuals > 0))
##
##  Runs Test
##
## data:  factor(stock1$residuals > 0)
## Standard Normal = -15, p-value <2e-16
## alternative hypothesis: two.sided
durbinWatsonTest(stock1)
##  lag Autocorrelation D-W Statistic p-value
##    1          0.9695       0.05589       0
##  Alternative hypothesis: rho != 0
acf(stock1$residuals)
```

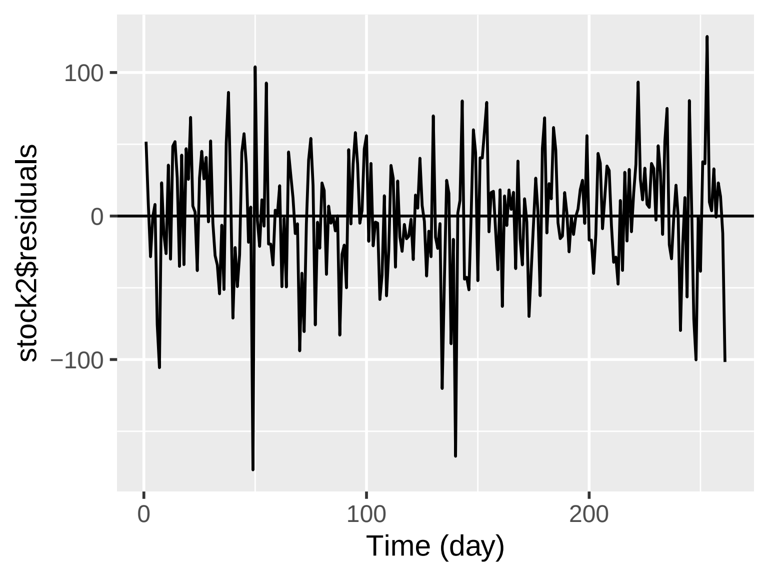


The model does exhibit time dependencies based on the number of runs and small p values for the runs test, durbin watson test, and autocorrelation plot.

Q4b — 8 points

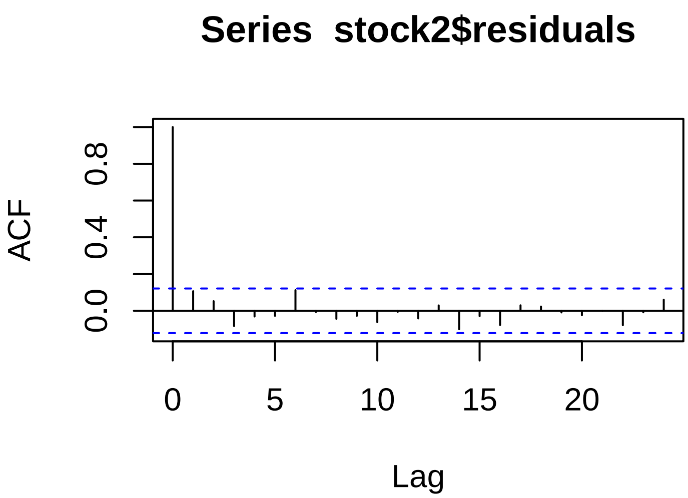
```
stock2 = lm(DJIA[2:262] ~ DJIA[1:261], data=stock)
```

```
ggplot(mapping=aes(x=stock$Day[1:261],y=stock2$residuals)) +
  geom_line() + labs(x='Time (day)','Raw Residuals') +
  geom_hline(yintercept = 0)
```




```
library(tseries)
runs.test(factor(stock2$residuals > 0))
##
##  Runs Test
```

```
##
## data:  factor(stock2$residuals > 0)
## Standard Normal = -1.7, p-value = 0.1
## alternative hypothesis: two.sided
durbinWatsonTest(stock2,alternative = 'negative')
##  lag Autocorrelation D-W Statistic p-value
##    1          0.1067         1.759   0.978
##  Alternative hypothesis: rho < 0
acf(stock2$residuals)
```



There is no longer any evidence of autocorrelation in the residuals.