STAT 420 – Homework 6

1. Prostate Data (with R)

a. Check for observations with large leverage. Report the observations with large leverage and their leverage values.

```
> lev <- hatvalues(fit)
> lev_mean <- mean(lev)
> lev[lev > 2 * lev_mean]
    32     37     41     74     92
0.3305 0.2184 0.2410 0.1912 0.2092
```

b. Without any correction, six observations pop out as potential outliers. But a Bonferroni correction, the `numeric(0)` result means that there are none. Looking back at the studentized residuals

```
> sresid <- rstudent(fit)
> n <- length(sresid)
> p <- length(fit$coefficients)
> df <- n - p - 1
> alpha <- 0.05

> # without bonferroni
> crit <- qt(1 - alpha/2,df)
> sresid[abs(sresid) > crit]
    39     47     69     81     95     97
-2.617 -2.377  2.554  1.988  2.385  2.293

> # with bonferroni
> crit <- qt(1 - (alpha/2)/n,df)
> sresid[abs(sresid) > crit]
named numeric(0)
```

c. Check for any influential observations. Report the observations with large influence and their Cook's distance.

```
> cook <- cooks.distance(fit)
> cook[cook > 4/n]
     32      39      47      69      95      96      97
0.12270 0.05202 0.10574 0.10054 0.09874 0.05594 0.07378
```

d. It seems the largest differences occur with `lweight` and `gleason`.

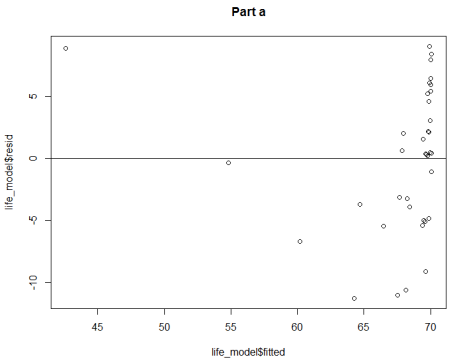
```
> fit_d <- lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,
              subset = cook < 4/n)
> coef(fit)
(Intercept)      lcavol     lweight         age        lbph         svi
     0.6693      0.5870      0.4545     -0.0196      0.1071      0.7662
         lcp     gleason       pgg45
    -0.1055      0.0451      0.0045

> coef(fit_d)
(Intercept)      lcavol     lweight         age        lbph         svi
    -0.2461      0.5650      0.5444     -0.0186      0.1333      0.7447
         lcp     gleason       pgg45
    -0.1554      0.1147      0.0066
```

```
> coef(fit_d) - coef(fit)
(Intercept)      lcavol     lweight         age        lbph         svi
    -0.9154     -0.0220      0.0900      0.0011      0.0262     -0.0214
         lcp     gleason       pgg45
    -0.0499      0.0696      0.0021
```

2. Life Expectancy Data (with R)

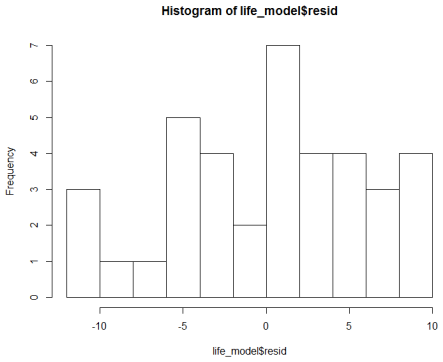
a. The residual plot does not appear random at all with many negative residuals leading up to a majority of the points being bunched together among larger fitted values. The Breusch-Pagan test does yield a *p*-value that would support the constant variance assumption, but the plot is too abnormal to ignore.

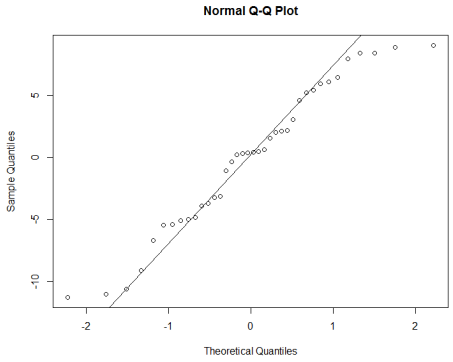
```
> bptest(life_model)

   studentized Breusch-Pagan test

data:  life_model
BP = 4.2061, df = 2, p-value = 0.1221
```



b. The histogram and jaggedness of the normal probability plot do not seem to support the normality assumption. Even though the Shapiro-Wilk test supports the normality of the error terms, we still have reservations based on the plots.

```
> shapiro.test(life_model$resid)

   Shapiro-Wilk normality test

data:  life_model$resid
W = 0.9587, p-value = 0.1725
```





c. There are three observations that have a leverage (i.e., hat value) that is more than twice the average.

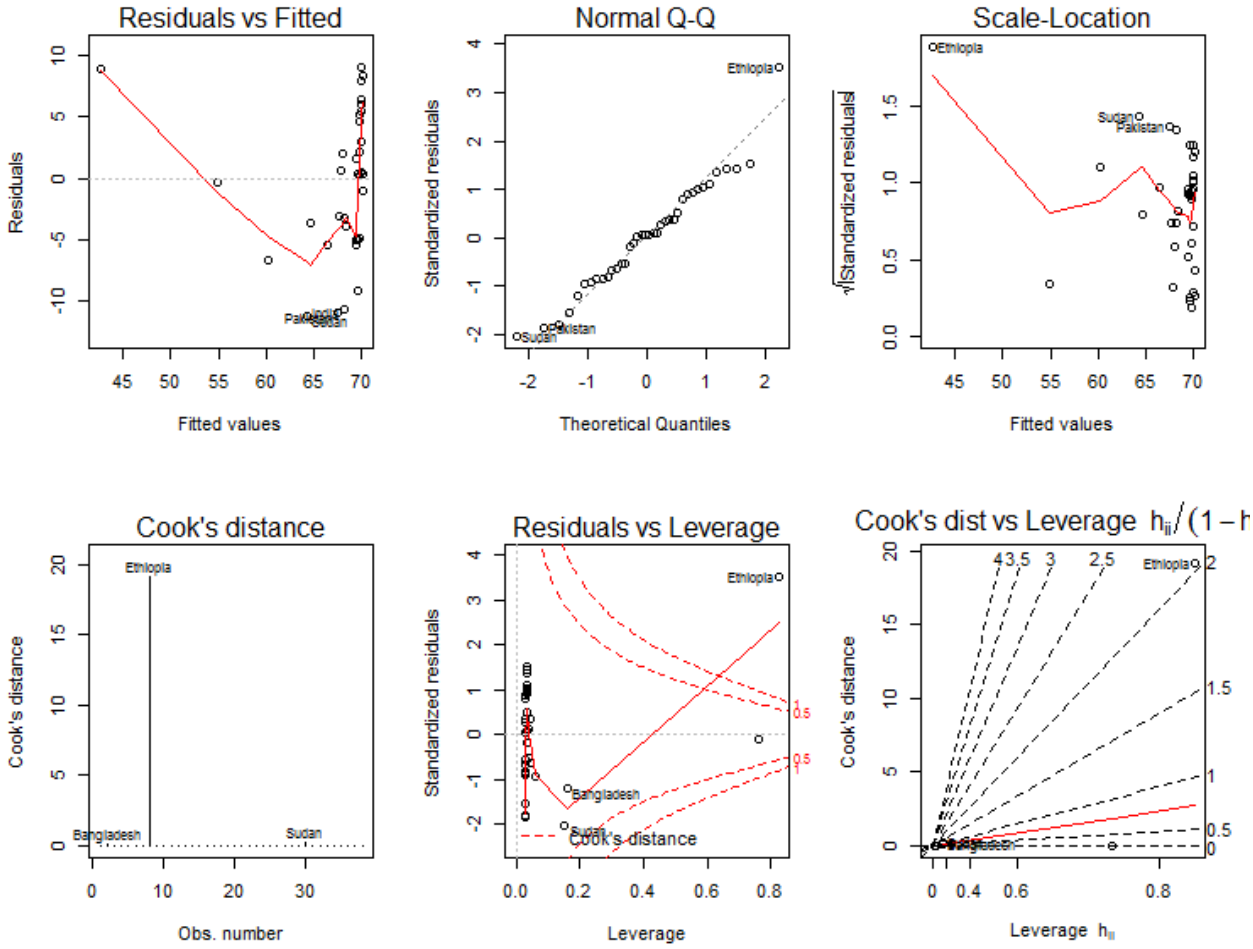
```
> lev <- hatvalues(life_model)
> lev_mean <- mean(lev)
> lev[lev > 2 * lev_mean]
Bangladesh   Ethiopia    Myanmar
 0.1597777  0.8222873  0.7598006
```

d. There are two outliers without any adjustment, but only one when we use the Bonferroni correction. Note that Ethiopia also appears as a leverage point.

```
> sresid <- rstudent(life_model)
> n <- length(sresid)
> p <- length(life_model$coefficients)
> df <- n - p - 1
> alpha <- 0.05

> crit <- qt(1 - alpha/2,df)
> sresid[abs(sresid) > crit]
 Ethiopia     Sudan
 4.314525 -2.144959

> # with bonferroni
> crit <- qt(1 - (alpha/2)/n,df)
> sresid[abs(sresid) > crit]
Ethiopia
4.314525
```

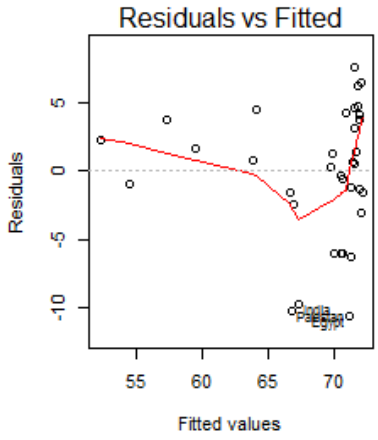
e. Again, Ethiopia shows up in our diagnostic analysis. It's Cook's distance is very large, but that's not surprising given it showed up as both a leverage and outlier point. The diagnostic plots also clearly show it's influence.

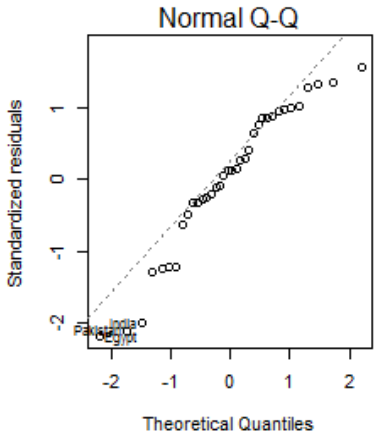
```
> cook <- cooks.distance(life_model)
> cook[cook > 4/n]
  Ethiopia      Sudan
19.0988509  0.2496147

> plot(life_model, 1:6)
```

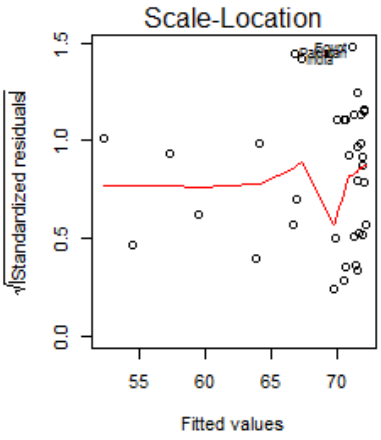


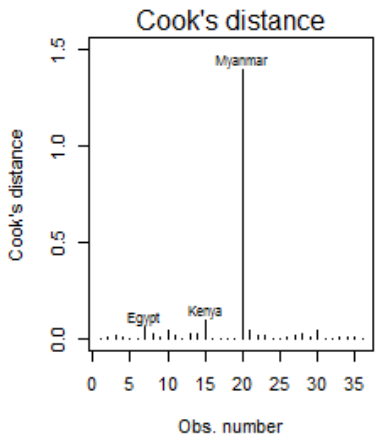
f. The most notable change in the coefficients is that the slope of the Doctor variable is now about four times greater. The diagnostic plots do not seem that much better. There's a slight improvement in the residual plot, but the normal probability plot is poor and it seems there's a new influential point (Myanmar). (Actually, the original residual plot and great disparity in the scale of values suggests a transformation, such as a log-transform, might be useful.)

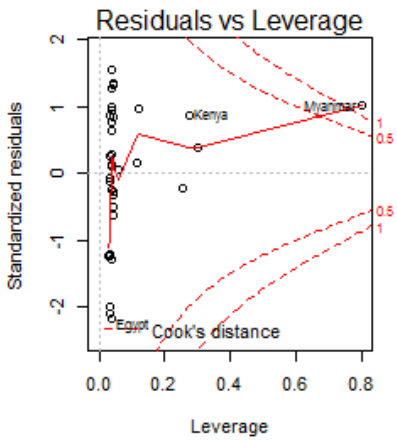
```
> life_f <- lm(life ~ ., data = tvdoctor, subset = cook < 4/n)
> coef(life_model)
 (Intercept)            tv        doctor
70.251957282 -0.023495365 -0.000432047

> coef(life_f)
 (Intercept)            tv        doctor
72.534110345 -0.024347007 -0.001693231

> coef(life_f) - coef(life_model)
  (Intercept)            tv        doctor
 2.2821530626 -0.0008516422 -0.0012611843

> plot(life_f, 1:6)
```

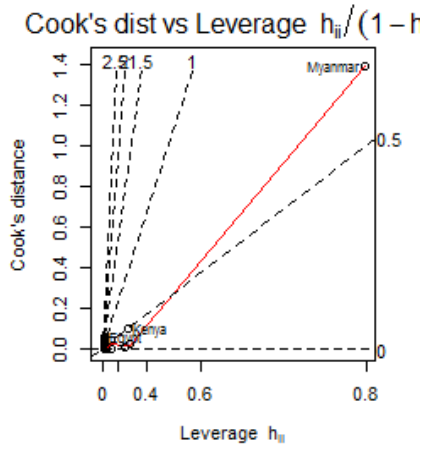

Residuals vs Fitted
Residuals
Fitted values
India
Pakistan
Egypt


Normal Q-Q
Standardized residuals
Theoretical Quantiles
India
Pakistan
Egypt


Scale-Location
√|Standardized residuals|
Fitted values


Cook's distance
Cook's distance
Obs. number
Myanmar
Egypt
Kenya


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Kenya
Myanmar
Egypt


Cook's dist vs Leverage hii/(1 – h
Cook's distance
Leverage hii
Myanmar
Kenya