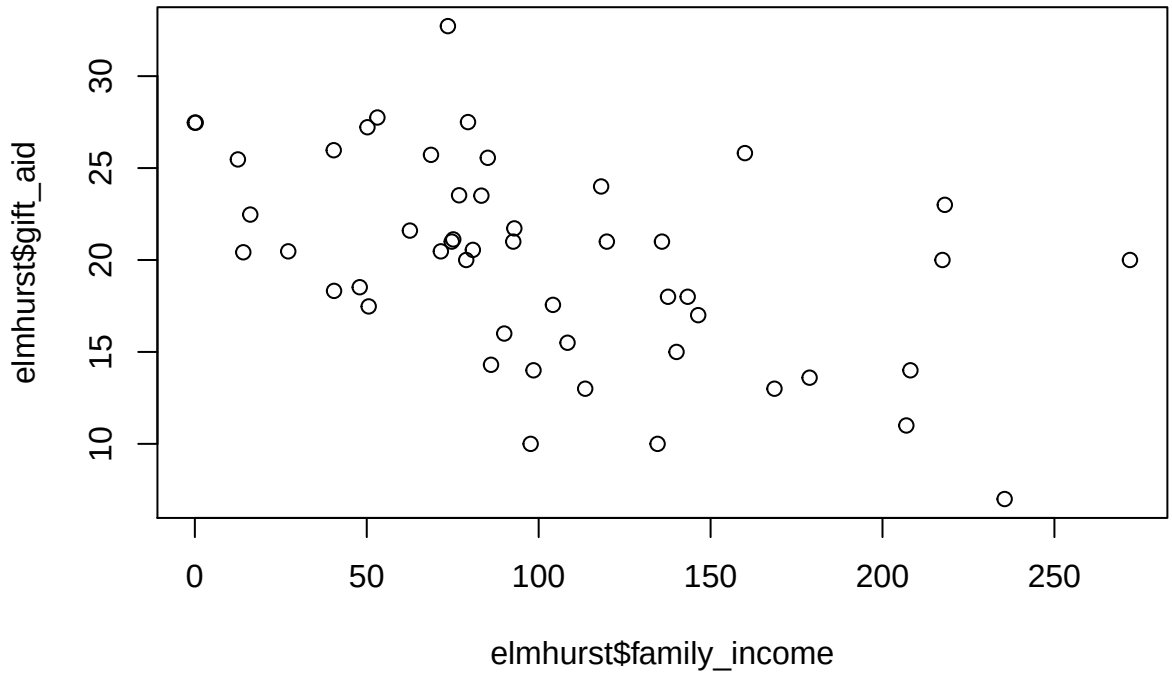
Lab 4

First, let's download some data. This data includes family income and amount of aid given to 50 students at Elmhurst college. What do you notice? What is the relationship between family income and amount of aid?

```
require(openintro)
data("elmhurst")
```

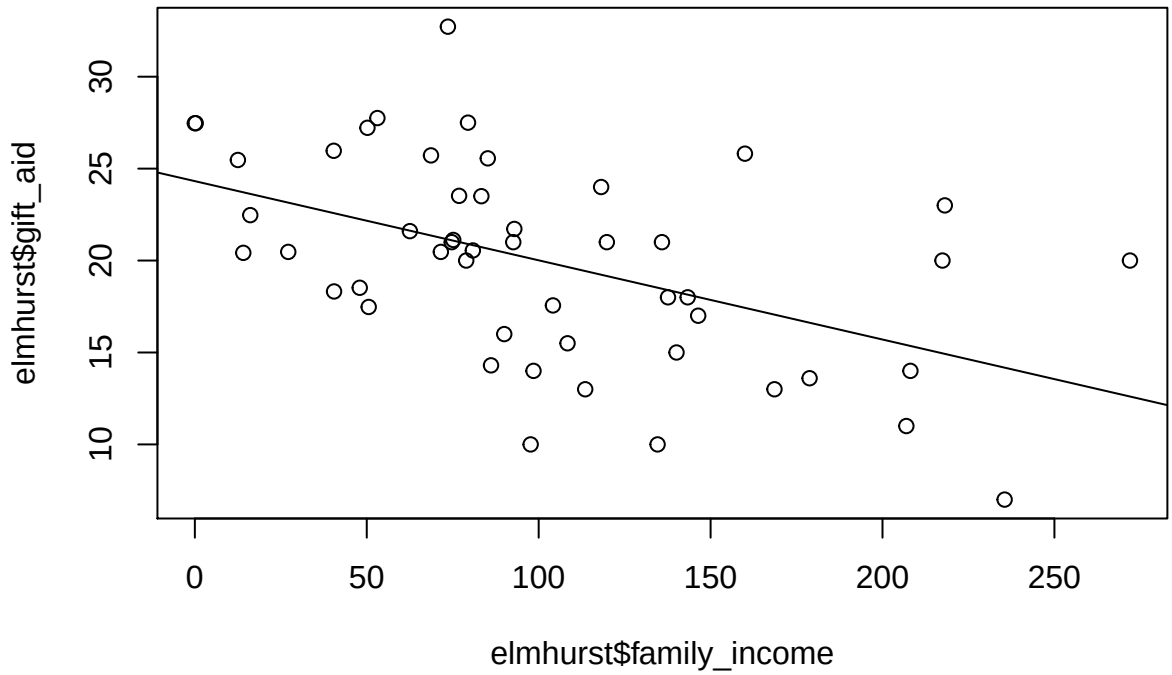
To investigate the relationship between income and aid, we can make a scatter plot of the data:

```
plot(elmhurst$family_income, elmhurst$gift_aid)
```



Now, we run a linear regression, with the aid amount as DV and family income as IV and add the regression line to the plot. Do you see many outliers along the x or y axis? Which points do you think have higher leverage? Are there any high influence points? Remember: outliers have large residuals, high leverage points are generally outliers along the x axis, high influence points are generally outliers along the x and y axis:

```
lm1 <- lm(gift_aid ~ family_income, data=elmhurst)
plot(elmhurst$family_income, elmhurst$gift_aid)
abline(lm1)
```



An aside: note that the R squared for simple regression is the same as the correlation between the IV and DV squared

```
# correlation:
cor(elmhurst$family_income, elmhurst$gift_aid)^2
```

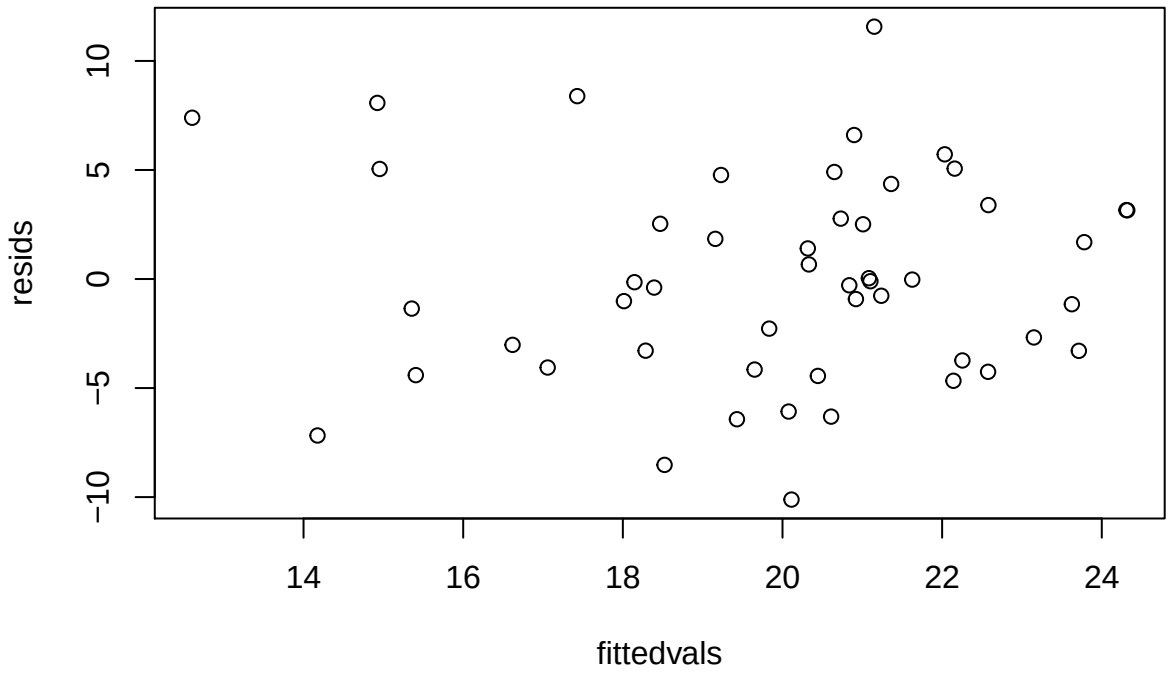
```
## [1] 0.2485582
```

```
#R squared:
summary(lm1)$r.squared
```

```
## [1] 0.2485582
```

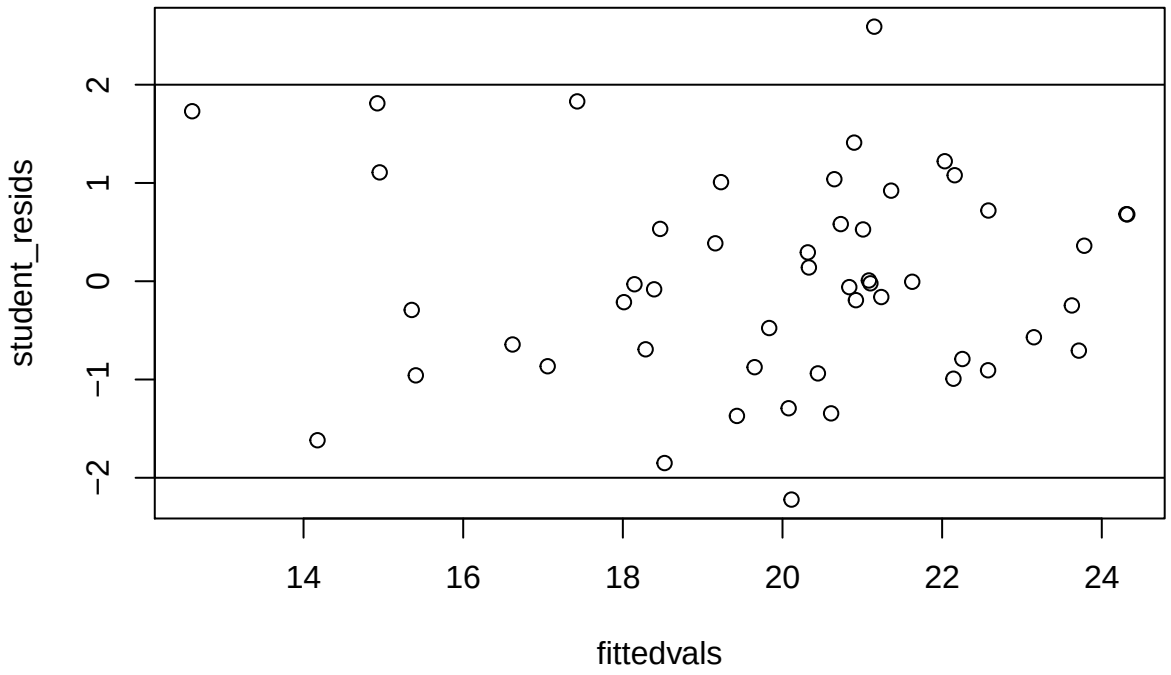
Now we can run some regression diagnoistics. One of the easiest things to do is look at a plot of the residuals against the fitted values. Something to think about: what are you looking for in this plot and what would indicate a problem? (note: remember the linearity assumption and equal variance assumption)

```
#get fitted values
fittedvals <- lm1$fitted.values

#get residuals
resids <- lm1$residuals

#plot fitted values against residuals
plot(fittedvals, resids)
```



Now, let's calculate studentized residuals and re-make the plot. Note: studentized residuals are similar to standardized residuals; however, we calculate the ith studentized residual by calculating it's standardized distance from a regression line that was fitted with all other points except the ith point. We do this because high influence points may "pull" the regression line closer to them, so a high influence point may not appear to be as much of an outlier as it really is if we use the regression line that was fit with that point included.

```
# Note: you may need to install.packages("MASS") in the Console
# if you have never done so
require(MASS)
student_resids <- studres(lm1)
plot(fittedvals, student_resids)
# Add horizontal lines at -2 and 2
abline(h=-2)
abline(h=2)
```

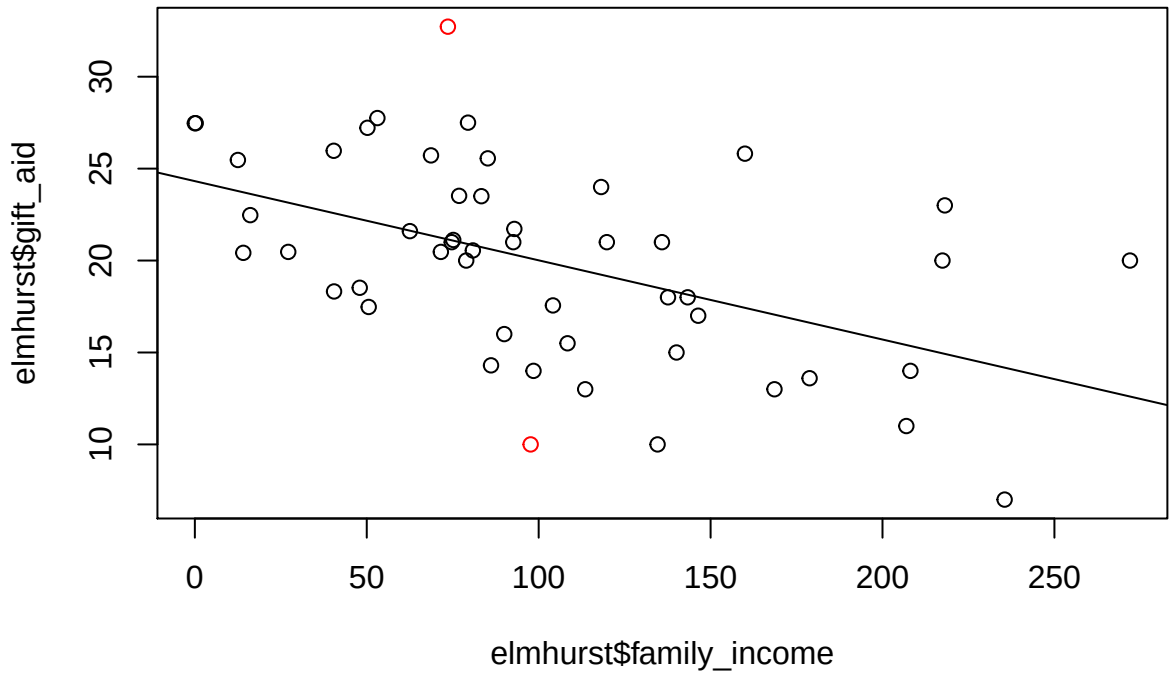


We expect about 5% of points to be outside 2 standard deviations, so this looks pretty good. If we wanted to know which points those two "outliers" are, we can use the which() function:

```
# Get indices
which(abs(student_resids)>2)
```

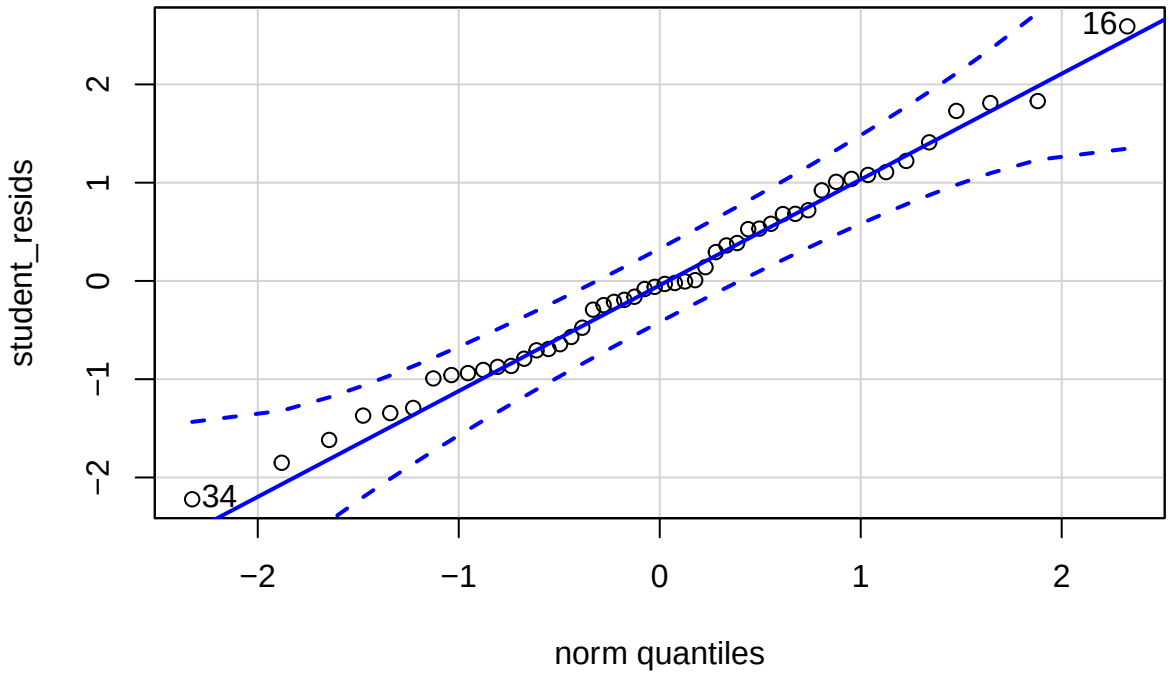
```
## 16 34
## 16 34
```

```
# Remove them from the data if we wanted, resave as elmhurst2
elmhurst2 <- elmhurst[-which(abs(student_resids)>2),]

# Or, we could color those values red in the original plot if we want
plot(elmhurst$family_income, elmhurst$gift_aid, col=(abs(student_resids)>2)+1)
abline(lm1)
```



We can also check the normality of errors assumption by making a QQ plot of the residuals. I like the qqPlot() function in the car package because it gives a confidence interval, but there are lots of others, like qqnorm() in the stats package. In this case, the errors look relatively normal.

```
# again, you might need install.packages("car")
require(car)
qqPlot(student_resids)
```



Measures of influence: high influence points are generally outliers with high leverage. In class you will (or possibly have already) discussed a few different measures of influence. Leverage is a measurement of how "unusual" an X value is compared to other X values (minimum is 1/n). Cook's distance is an overall measure

of how each point influences the regression coefficient estimate. Values greater than 4/n suggest high influence. DFFITS is essentially measuring the same thing as Cook's dstiance, just on a different scale: 0 means no influence, anything with absolute value larger tha 2*sqrt(p/n) is considered high influence (n is sample size and p is the number of parameters).

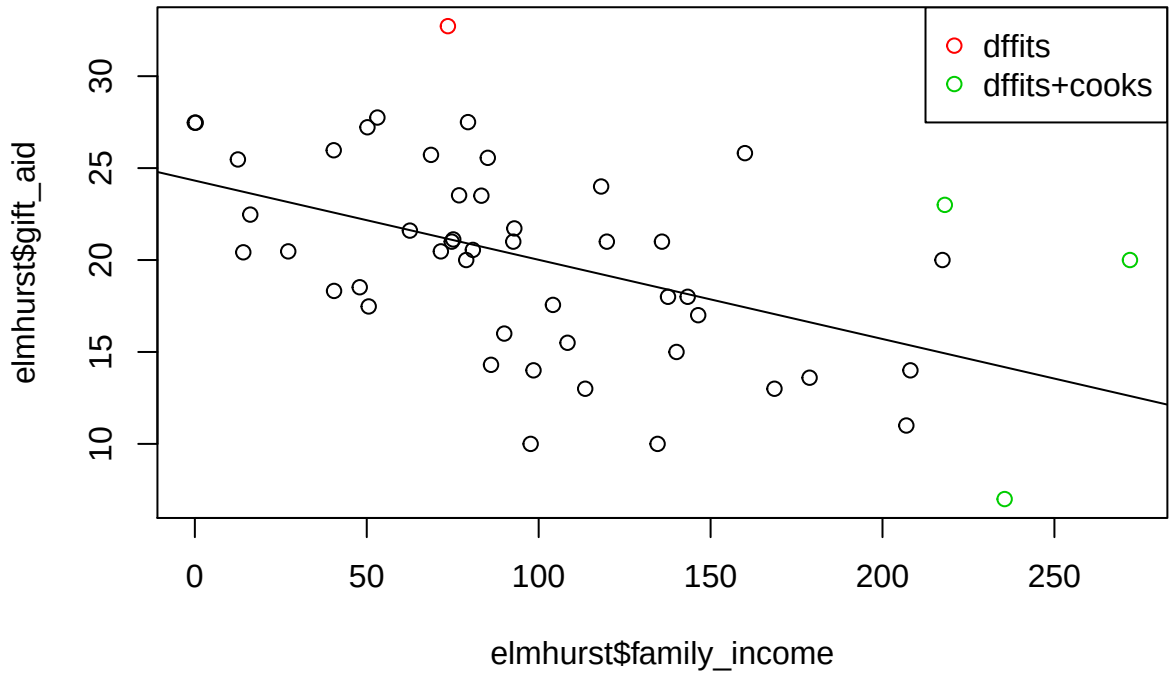
```
# Again may need to install.packages("stats")
require(stats)
leverage <-hatvalues(lm1)
cooksd <-cooks.distance(lm1)
diffits <- dffits(lm1)

#which values have high influence according to cooks distance?
which(cooksd>(4/nrow(elmhurst)))
```

```
## 17 19 22
## 17 19 22
```

```
#which values have high influence according to dffits?
p=2 #2 params in model: b0 and b1
which(abs(diffits)>(2*sqrt(p/nrow(elmhurst))))
```

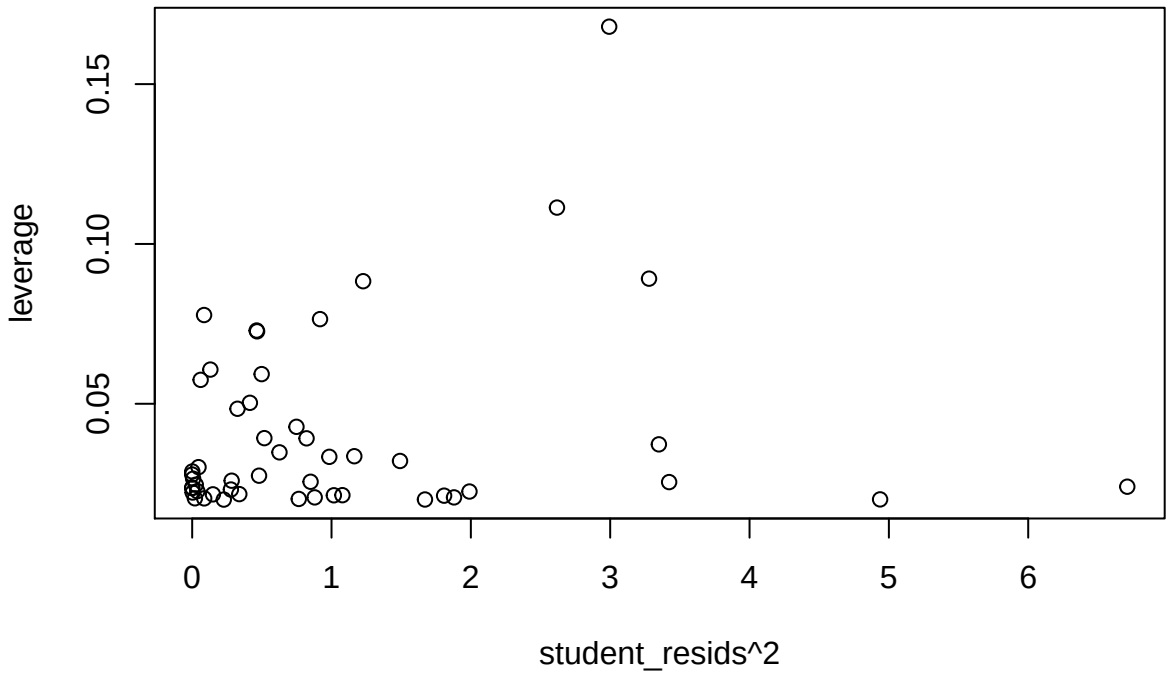
```
## 16 17 19 22
## 16 17 19 22
```

```
#plot them with colors representing flagged points
plot(elmhurst$family_income, elmhurst$gift_aid,
     col=(cooksd>(4/nrow(elmhurst)))+
       (abs(diffits)>(2*sqrt(p/nrow(elmhurst))))+1)
abline(lm1)
legend("topright", c("dffits", "dffits+cooks"), col=c(2,3), pch=1)
```



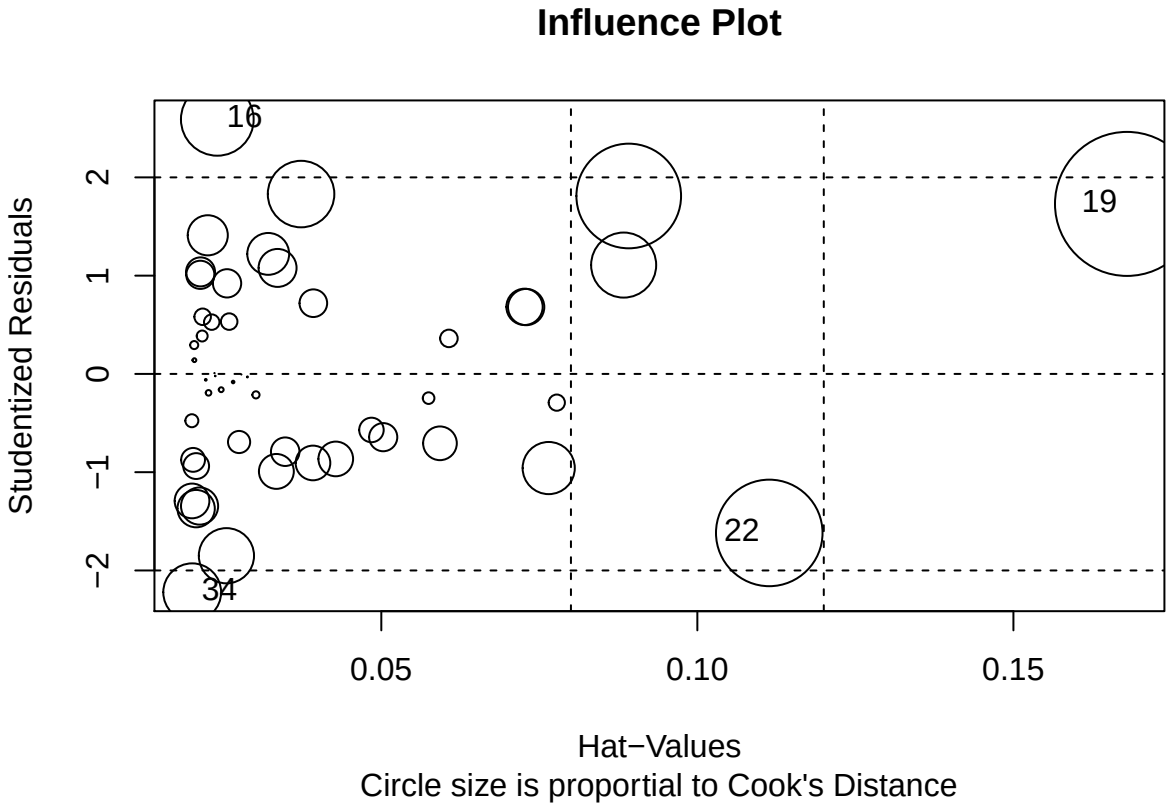
We can also plot leverage versus squared residuals to identify high influence points:

```
plot(student_resids^2, leverage)
```



Finally, the influencePlot function in the car package can also be used to plot studentized residuals against the leverage of each value, with circle size proportional to Cook's distance. The function identifies 4 outliers with respect to leverage, Cook's distance, and vertical distance from the regression line: at indices 16, 19, 22, 34. We had already flagged the first three above using dffits and/or Cook's distance. There are no hard and fast rules for when/if you should remove data – but it's definitely important to decide on criteria before receiving the data!

```
influencePlot(lm1,  main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance")
```




```
##      StudRes        Hat      CookD
## 16  2.590554 0.02405676 0.07391732
## 19  1.729978 0.16797118 0.29005560
## 22 -1.618076 0.11137478 0.15872173
## 34 -2.221930 0.02008648 0.04676398
```

A side note: Instead of labeling the X-axis as "Leverage", the default is hat-values. You may have also noticed that the function to calculate leverage is called hatvalues(). If you are wondering why this is, it is because the leverage values are the diagonal values of something called the "hat matrix". The hat matrix, $H = X(X^TX)^{-1}X^T$ is called the hat matrix because $HY = \hat{Y}$. In other words, the hat matrix, H, maps the observed Y values onto the predicted Y values, denoted $\hat{Y}$ (and pronounced "Y hat"). Thus, people say that H "puts a hat on Y".[1] To prove this, remember that the OLS beta coefficients can be calculated as $\beta = (X^tX)^{-1}X^tY$. Then, if we multiply both sides of that equation by X, we get $X\beta = X(X^tX)^{-1}X^tY$, which simplifies to $\hat{Y} = HY$ because $X\beta = \hat{Y}$ and $H = X(X^TX)^{-1}X^T$.

[1]Wikipedia contributors. "Projection matrix." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 3 Mar. 2019. Web. 12 Mar. 2019.

Practice

1. Download the "starbucks" dataset from the openintro package
2. Plot calories (the number of calories in each item) vs. carb (the amt of carbs in each item). Put calories on the y-axis and carb on the x-axis
3. Add the regression line for calories regressed on carbohydrates
4. Calculate the studentized residuals and then plot them against the fitted values. Comment on the linearity assumption and equal variance assumption.
5. Make a QQ plot of the residuals and comment on whether the normality assumption holds.
6. Identify any points for which absolute studentized residuals are >2 and replot the graph from question 2 with those points in a different color.
7. Calculate Cook's distance and dffits values for all the points and identify points that would be considered high influence based on each criterium. Create two more graphs with any points flagged by each criterium in a different color. Do they agree?
8. Use the influencePlot() function to plot the studentized residuals against leverage. Which values does this function flag and do they match the values flagged above?