Faraway, Chapter 6, question #1.
Here is documentation on the sat dataset:
`sat`
School expenditure and test scores from USA in 1994-95
Description
The
`sat`
data frame has 50 rows and 7 columns. Data were collected to study the relationship between expenditures on public education and test results.
Usage
`data(sat)`
Format
This data frame contains the following columns:
`expend`
Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)
`ratio`
Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
`salary`
Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
`takers`
Percentage of all eligible students taking the SAT, 1994-95
`verbal`
Average verbal SAT score, 1994-95
`math`
Average math SAT score, 1994-95
`total`
Average total score on the SAT, 1994-95
Source
"Getting What You Pay For: The Debate Over Equity in Public School Expenditures" D. Guber,
Journal of Statistics Education, 1999

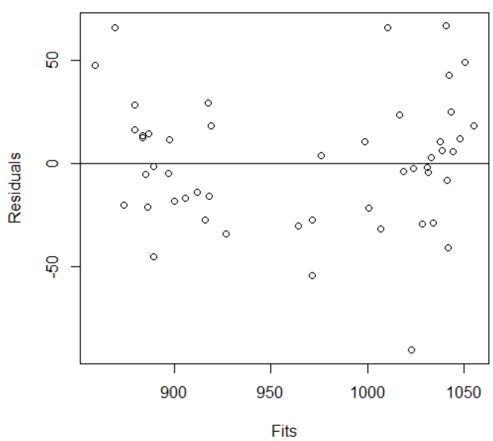
(a) Constant Variance Assumption

```
by.math <- sat[order(sat$math),]
by.verbal <- sat[order(sat$verbal),]
by.salary <- sat[order(sat$salary),]

attach(sat)
out <- lm(total ~ expend + ratio + salary + takers)
summary(out)
```

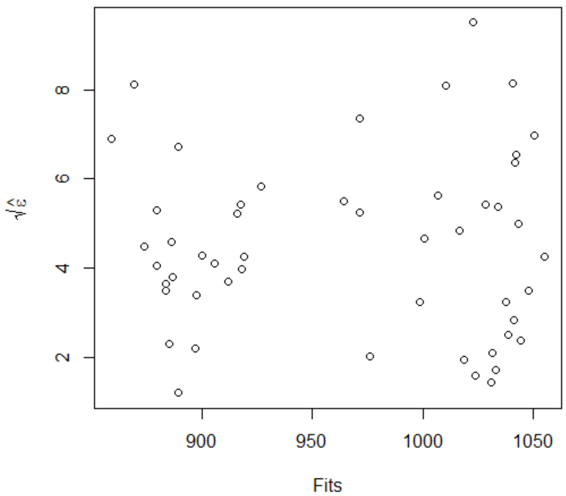
```
Call:
lm(formula = total ~ expend + ratio + salary + takers)

Residuals:
```

```
   Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
ratio         -3.6242     3.2154  -1.127    0.266
salary         1.6379     2.3872   0.686    0.496
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
plot(residuals(out) ~  fitted(out), xlab="Fits", ylab="Residuals")
abline(h=0)
```

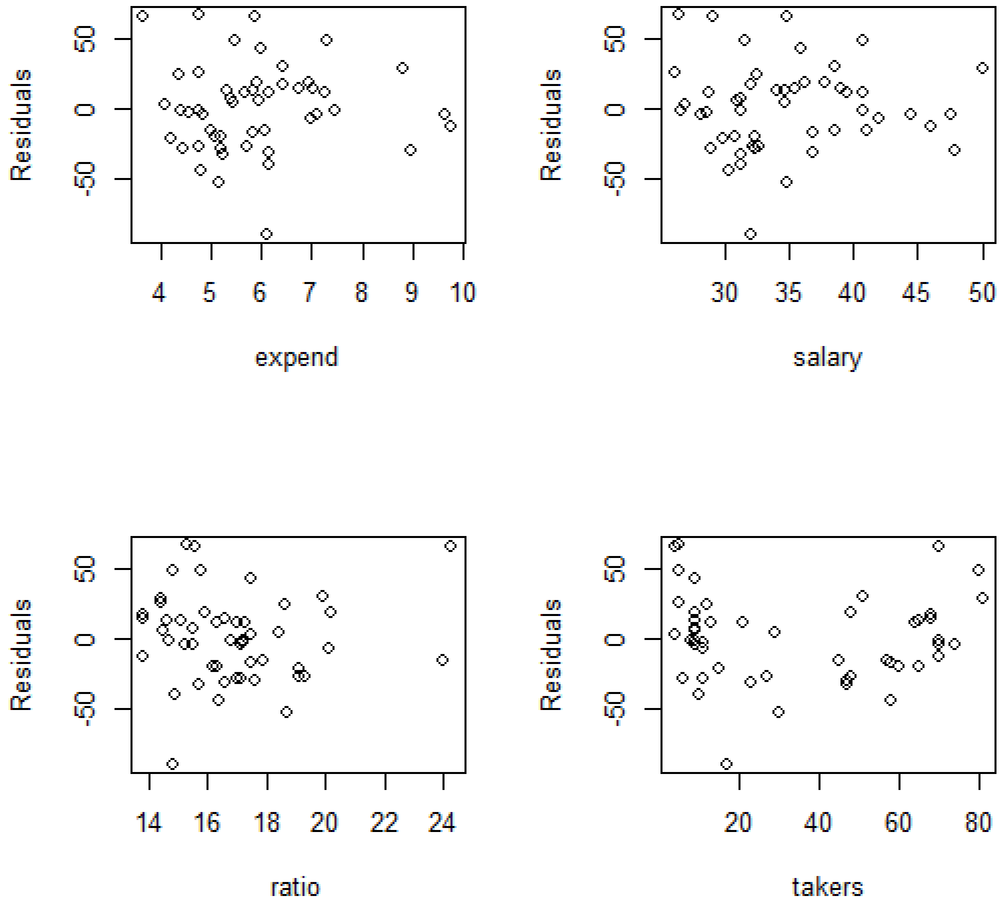



```
plot(sqrt(abs(residuals(out))) ~  fitted(out), xlab="Fits", ylab=expression(sqrt(hat(epsilon))))
```




```
sumary(lm(sqrt(abs(residuals(out)))~ fitted(out)))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.6484524  4.0660807  1.1432   0.2586
fitted(out) -0.0001637  0.0041994 -0.0390   0.9691

n = 50, p = 2, Residual SE = 1.99717, R-Squared = 0
```

```
windows()
par(mfrow=c(2,2))
plot(residuals(out) ~  sat$expend, xlab="expend", ylab="Residuals")
plot(residuals(out) ~  sat$salary, xlab="salary", ylab="Residuals")
plot(residuals(out) ~  sat$ratio, xlab="ratio", ylab="Residuals")
plot(residuals(out) ~  sat$takers, xlab="takers", ylab="Residuals")
```




```
var.test(residuals(out)[sat$expend<7], residuals(out)[sat$expend>7])

        F test to compare two variances

data:  residuals(out)[sat$expend < 7] and residuals(out)[sat$expend > 7]
F = 2.0212, num df = 40, denom df = 8, p-value = 0.2947
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5263949 5.1114419
sample estimates:
ratio of variances
          2.021241

var.test(residuals(out)[sat$salary<40], residuals(out)[sat$salary>40])

        F test to compare two variances
```

```
data:  residuals(out)[sat$salary < 40] and residuals(out)[sat$salary > 40]
F = 2.1487, num df = 39, denom df = 9, p-value = 0.2225
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6122021 5.2882878
sample estimates:
ratio of variances
         2.148672
```

```
var.test(residuals(out)[sat$ratio<18], residuals(out)[sat$ratio>18])
```

```
        F test to compare two variances

data:  residuals(out)[sat$ratio < 18] and residuals(out)[sat$ratio > 18]
F = 0.8576, num df = 38, denom df = 10, p-value = 0.6858
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2627366 2.0645530
sample estimates:
ratio of variances
        0.8576423
```

```
var.test(residuals(out)[sat$takers<40], residuals(out)[sat$takers>40])
```

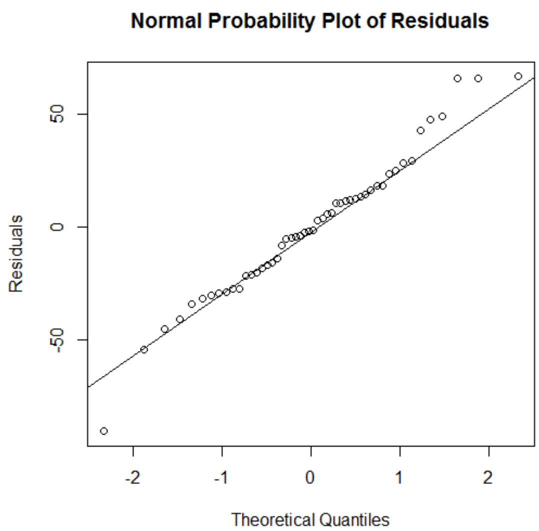
```
        F test to compare two variances

data:  residuals(out)[sat$takers < 40] and residuals(out)[sat$takers > 40]
F = 1.6451, num df = 26, denom df = 22, p-value = 0.2392
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.712512 3.691978
sample estimates:
ratio of variances
         1.645067
```

(b) Normality Assumption

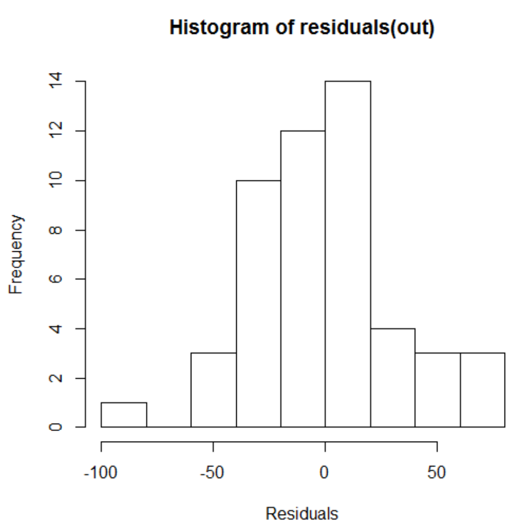
```
windows()
par(mfrow=c(1,1))
qqnorm(residuals(out), ylab="Residuals", main="Normal Probability Plot of Residuals")
qqline(residuals(out))
```

NOTE: Faraway writes that histograms and boxplots are not suitable for checking normality. While this may be true, people do it all the time!

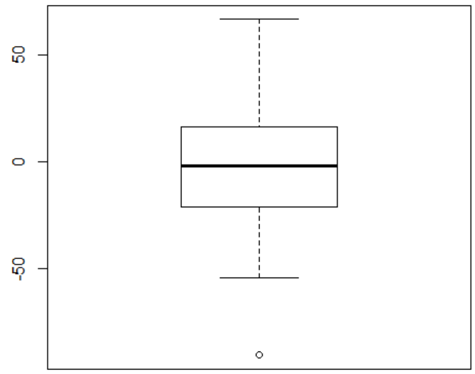
```
par(mfrow=c(1,1))
qqnorm(residuals(out), ylab="Residuals", main="Normal Probability Plot of Residuals")
qqline(residuals(out))
```




```
hist(residuals(out), xlab="Residuals")
```



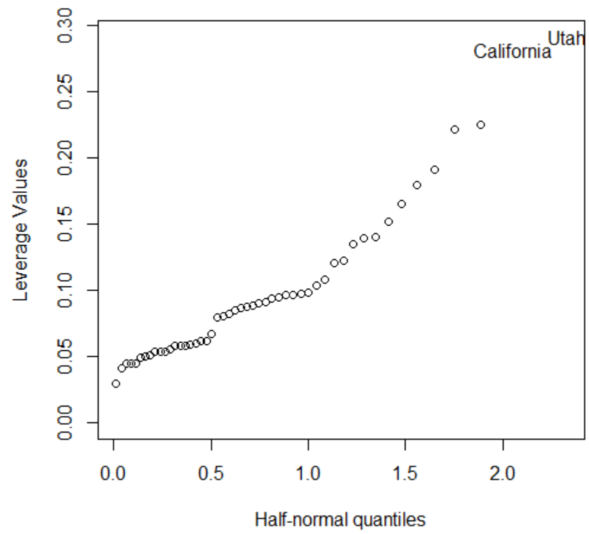

```
boxplot(residuals(out))
```



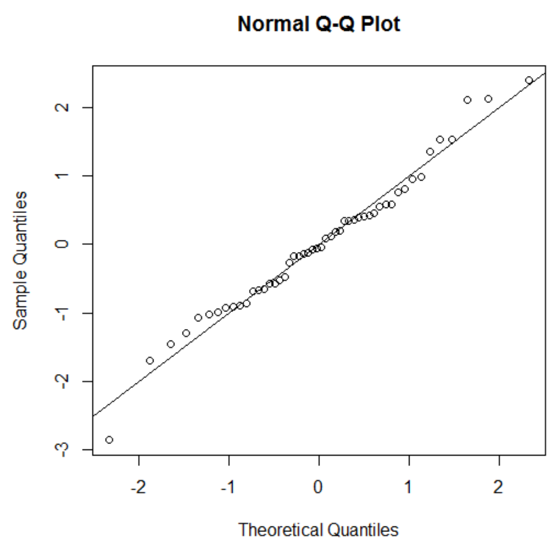

```
shapiro.test(residuals(out))

	Shapiro-Wilk normality test

data:  residuals(out)
W = 0.9769, p-value = 0.4304

library(nortest)
Warning message:
package 'nortest' was built under R version 3.1.0
ad.test(residuals(out))

	Anderson-Darling normality test

data:  residuals(out)
A = 0.3424, p-value = 0.4783
```

(c) Leverages

```
hatv <- hatvalues(out)
head(hatv)
         1          2          3          4          5          6 
0.09537668 0.18030612 0.04931612 0.05382878 0.28211791 0.03014533 
sum(hatv)
[1] 5
states <- row.names(sat)
halfnorm(hatv, labs=states, ylab="Leverage Values")
```



qqnorm(rstandard(out))
abline(0,1)



shapiro.test(rstandard(out))

Shapiro-Wilk normality test

data: rstandard(out)
W = 0.9802, p-value = 0.5607

(d) Outliers

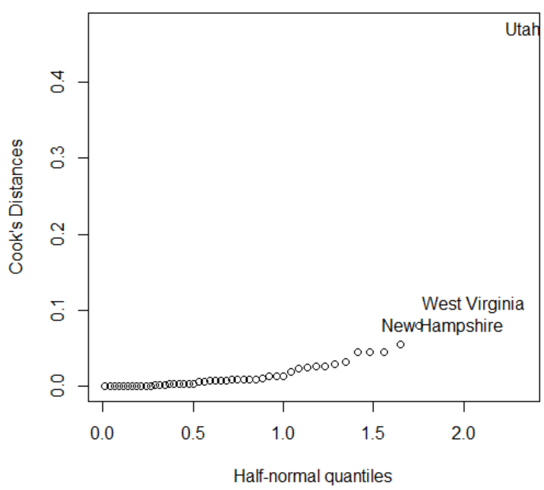
Use studentized residuals (also called jackknife residuals, or cross-validated residuals) here…

```
asdf <- rstudent(out)
asdf[which.max(abs(asdf))]
       48
-3.124428
qt(0.05/(50*2), 44)
#####  Bonferroni adjustment…  44 = n-p-1 = 50 -5-1 (n=# of records, p=#of model parameters)
[1] -3.525801
```

Conclude the 48th record (West Virginia) is not an outlier.

(e) Influential Points

```
cook <- cooks.distance(out)
halfnorm(cook, 3, labs=states, ylab="Cook's Distances")
```



Utah seems to have high leverage in the predictor space. Observe the effect of its exclusion on our model:

Original model output:

```
sumary(out)
              Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 1045.97154   52.86976  19.7839 < 2.2e-16
expend         4.46259   10.54653   0.4231    0.6742
ratio         -3.62423    3.21542  -1.1271    0.2657
```

```
salary        1.63792    2.38725   0.6861    0.4962
takers       -2.90448    0.23126 -12.5594 2.607e-16

n = 50, p = 5, Residual SE = 32.70199, R-Squared = 0.82
```

Model output when excluding Utah:

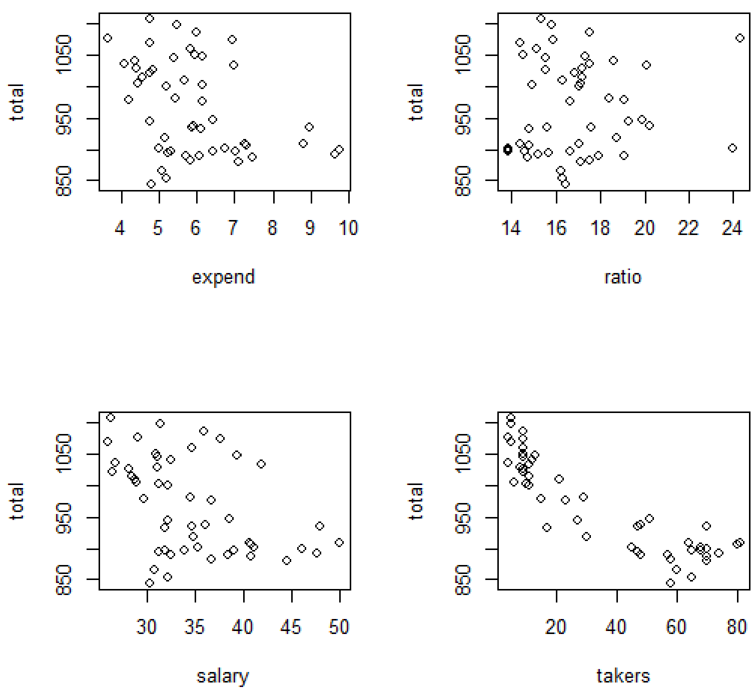
```
noUT.out <- lm(total ~ expend + ratio + salary + takers, subset=(cook < max(cook)))
sumary(noUT.out)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1093.84597   53.42255  20.4754   <2e-16
expend        -0.94274   10.19217  -0.0925   0.9267
ratio         -7.63914    3.42791  -2.2285   0.0310
salary         3.09643    2.32829   1.3299   0.1904
takers        -2.93080    0.21877 -13.3967   <2e-16

n = 49, p = 5, Residual SE = 30.90084, R-Squared = 0.84
```

Also compare using summary(). The sumary() command (with just one m) comes from the faraway package.

(f) Structure of Relationship Between Predictors and Response

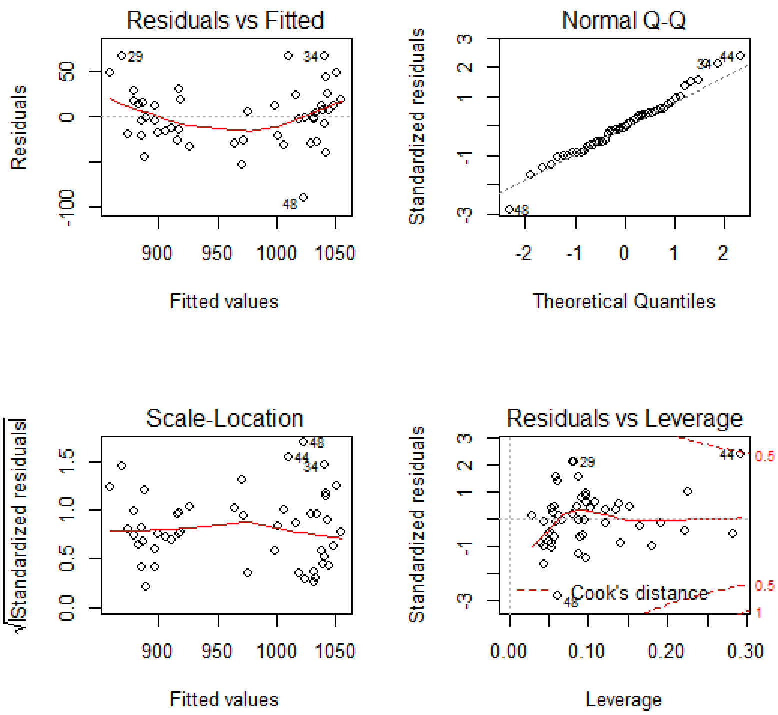
Definitely if you don't have too many predictor variables, you should try plotting the response vs each variable individually first- this is a good exploratory first step.

```
> windows()
> par(mfrow=c(2,2))
> plot(total ~ expend)
> plot(total ~ ratio)
> plot(total ~ salary)
```



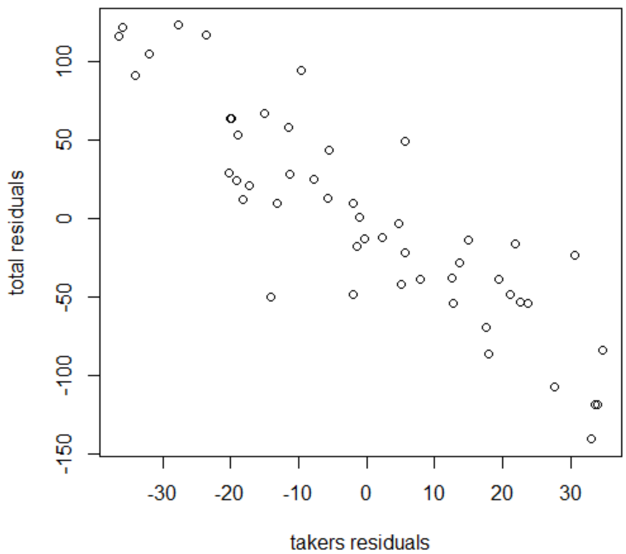
Also

```
windows()
par(mfrow=c(2,2))
plot(out)
```



Partial regression/added variable plots… check the takers variable…

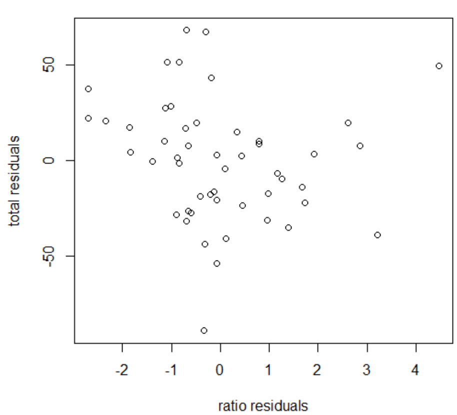
```
windows()
par(mfrow=c(1,1))
d <- residuals(lm(total ~ expend + ratio + salary))
m <- residuals(lm(takers ~ expend + ratio + salary))
plot(d ~ m, xlab = "takers residuals", ylab="total residuals")
```



No sign of non-linearity here. These partial regression/added variable plots allow us to examine the marginal effects of predictor variables on the response after the other predictor variables have been removed. These plots indicate how a predictor variable tracks with the response versus the other variables.

For instance, try examining the added value/marginal effects of “ratio”.

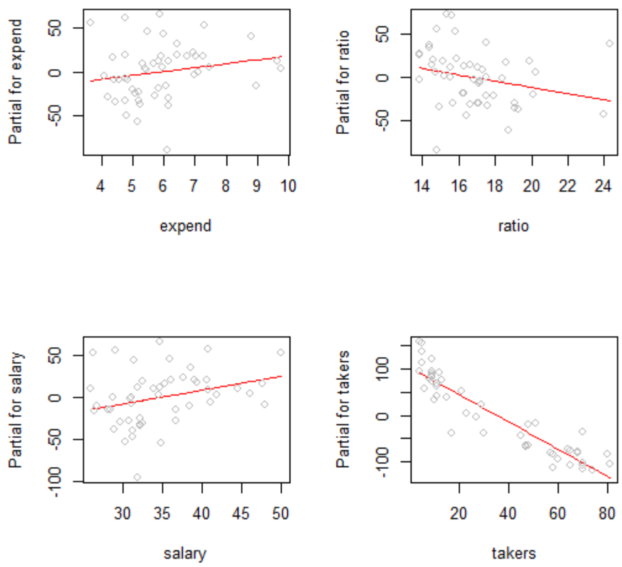
```
windows()
par(mfrow=c(1,1))
d2 <- residuals(lm(total ~ expend + salary + takers))
m2 <- residuals(lm(ratio ~ expend + salary + takers))
plot(d2 ~ m2, xlab = "ratio residuals", ylab="total residuals")
```



There is no real sign of linearity here, and so there is evidence that the ratio variable is not so important.

Partial residual plots are alternatives to added variable/partial regression plots.

```
windows()
par(mfrow=c(2,2))
termplot(out, partial.resid=TRUE, terms=1)
termplot(out, partial.resid=TRUE, terms=2)
termplot(out, partial.resid=TRUE, terms=3)
termplot(out, partial.resid=TRUE, terms=4)
```



Could be two groups in the takers variable…let's investigate…

```
mod1 <- lm(total ~ expend + ratio + salary + takers, subset=(takers < 40))
mod2 <- lm(total ~ expend + ratio + salary + takers, subset=(takers > 40))

sumary(mod1)
            Estimate Std. Error t value
(Intercept) 993.71775   84.50099 11.7598
expend        7.75814   16.43287  0.4721
ratio         1.42514    4.61109  0.3091
salary        1.02926    3.30577  0.3114
takers       -5.52423    0.87061 -6.3452
            Pr(>|t|)
(Intercept) 5.856e-11
```

```
expend        0.6415
ratio         0.7602
salary        0.7585
takers      2.194e-06

n = 27, p = 5, Residual SE = 30.95358, R-Squared = 0.66
sumary(mod2)
            Estimate Std. Error t value
(Intercept) 801.43294  105.67734  7.5838
expend      11.14438   10.83593  1.0285
ratio        3.91474    4.86273  0.8051
salary      -0.63544    2.71903 -0.2337
takers      -0.30035    0.88686 -0.3387
            Pr(>|t|)
(Intercept) 5.201e-07
expend        0.3174
ratio         0.4313
salary        0.8179
takers        0.7388

n = 23, p = 5, Residual SE = 23.74271, R-Squared = 0.26
```