Case Study 7: Unusual and influential data

Richard Wilkinson

Davis's weight data

Lets look at real data on the measured and reported weight of 183 male and female subjects.

```
library(car) # the dataset is in this package as well as some plotting tools
str(Davis)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

```
Davis[1:10,]
```

```
##    sex weight height repwt repht
## 1    M     77    182    77   180
## 2    F     58    161    51   159
## 3    F     53    161    54   158
## 4    M     68    177    70   175
## 5    F     59    157    59   155
## 6    M     76    170    76   165
## 7    M     76    167    77   165
## 8    M     69    186    73   180
## 9    M     71    178    71   175
## 10   M     65    171    64   170
```

Suppose we are interested in whether there is a difference between the accuracy with which each sex reports their own weight. A sensible model to assess this might be

$$\text{repwt}_i = \begin{cases} a + b \times \text{weight} + \epsilon \text{ if Male} \\ (a+c) + (b+d) \times \text{weight} + \epsilon \text{ if Female} \end{cases}$$

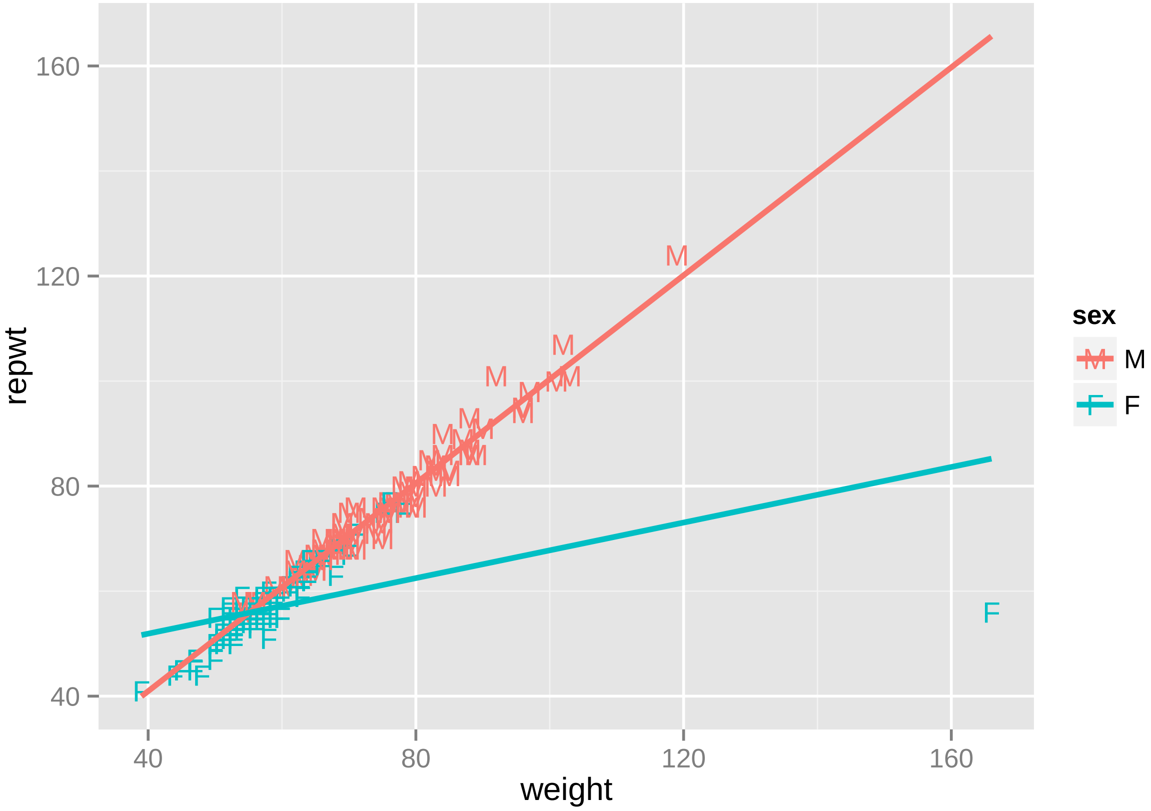
```
Davis <- within(Davis, sex <- relevel(sex, ref='M'))
fit1 <- lm(repwt ~ weight *sex, data = Davis)
summary(fit1)
```

```
##
## Call:
## lm(formula = repwt ~ weight * sex, data = Davis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.2230  -2.3247  -0.1325   2.0741  15.5783
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.35864    3.27719   0.415    0.679
## weight       0.98982    0.04260  23.236   <2e-16 ***
## sexF        39.96412    3.92932  10.171   <2e-16 ***
## weight:sexF -0.72536    0.05598 -12.957   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.661 on 179 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.8874, Adjusted R-squared:  0.8856
## F-statistic: 470.4 on 3 and 179 DF,  p-value: < 2.2e-16
```

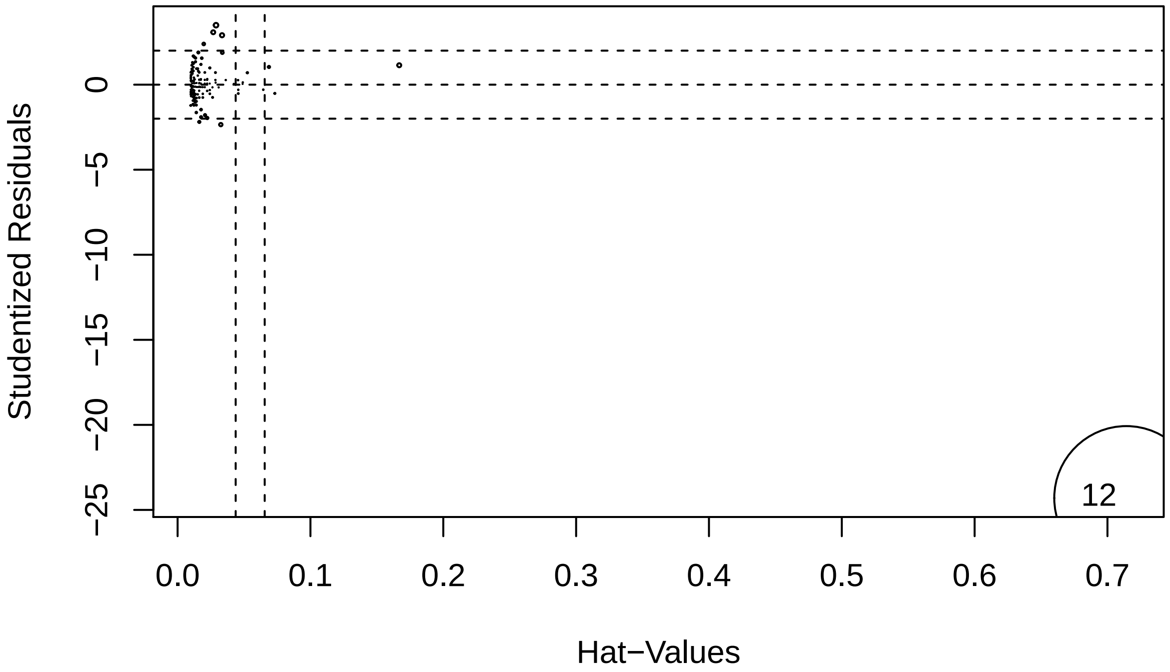
So this suggests males are unbiased estimators of their own weight (as $a \approx 0$ and $b \approx 1$) whereas females under-report their weight if they are relatively heavy (gradient is $0.99 - 0.73 = 0.26$), and over report if they are relatively light (intercept for women is $1.4 + 40.0 = 41.4$ kg). Note that the R^2 suggests a good fit.

However, if we plot the data we can see this is entirely due to a single female subject for whom the weight and height measurements were mis-labelled.



In this simple case we should have spotted the problem in our exploratory data analysis. However, lets use the diagnostic tools in the car package to see how easy it is to spot.

```
library(car)
influencePlot(fit1) # note these functions are in the car package,
```



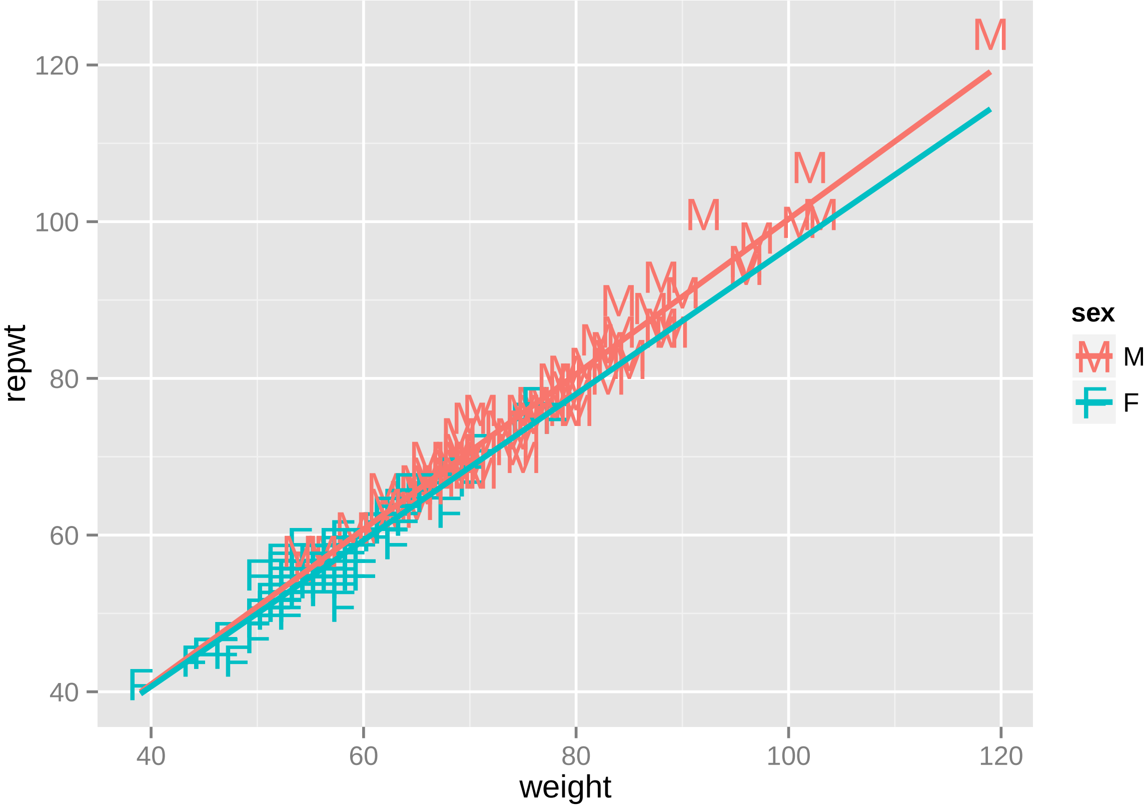
```
##      StudRes       Hat  CookD
## 12 -24.30446 0.7141856 9.2697
```

It is immediately obvious that there is one very influential point. The only sensible thing to do (if we cannot correct the observation) is to remove the offending data point and refit the model.

```
library(dplyr) # so that we can use the filter command
Davis2 <- filter(Davis, weight<160)
fit2 <-lm(repwt ~ weight *sex, data = Davis2)
summary(fit2)
```

```
##
## Call:
## lm(formula = repwt ~ weight * sex, data = Davis2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4685 -1.1332  0.0984  1.2342  8.5777
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.35864    1.58142   0.859    0.391
## weight       0.98982    0.02056  48.151   <2e-16 ***
## sexF         1.98929    2.45693   0.810    0.419
## weight:sexF -0.05671    0.03856  -1.471    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.249 on 178 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.9739, Adjusted R-squared:  0.9734
## F-statistic:  2211 on 3 and 178 DF,  p-value: < 2.2e-16
```

Having removed this data point, there is now no significant difference between the accuracy of men and women when reporting their weight.



Duncan's Occupational Prestige Data

Data on the prestige and other characteristics of 45 U.S. occupations in 1950. The data consists of the following measurements:

- type: Type of occupation. A factor with the following levels: prof, professional and managerial; wc, white-collar; bc, blue-collar.
- income: Percent of males in occupation earning $3500 or more in 1950.
- education: Percent of males in occupation in 1950 who were high-school graduates.
- prestige: Percent of raters in NORC study rating occupation as excellent or good in prestige.

Lets fit the model

$$\text{prestige} = a + b \times \text{education} + c \times \text{income} + \epsilon$$

and examine the data for influential data points.

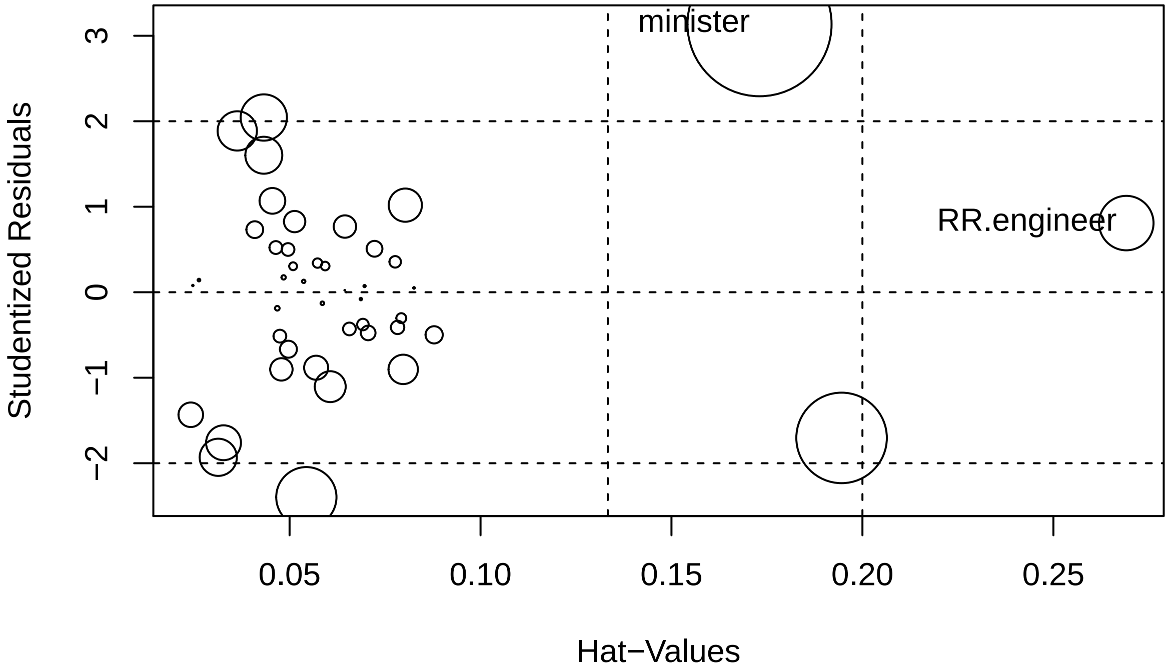
```
fit1 <- lm(prestige~education + income, data=Duncan)
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = prestige ~ education + income, data = Duncan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.538  -6.417   0.655   6.605  34.641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.06466    4.27194  -1.420    0.163
## education    0.54583    0.09825   5.555 1.73e-06 ***
## income       0.59873    0.11967   5.003 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 42 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:   0.82
## F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

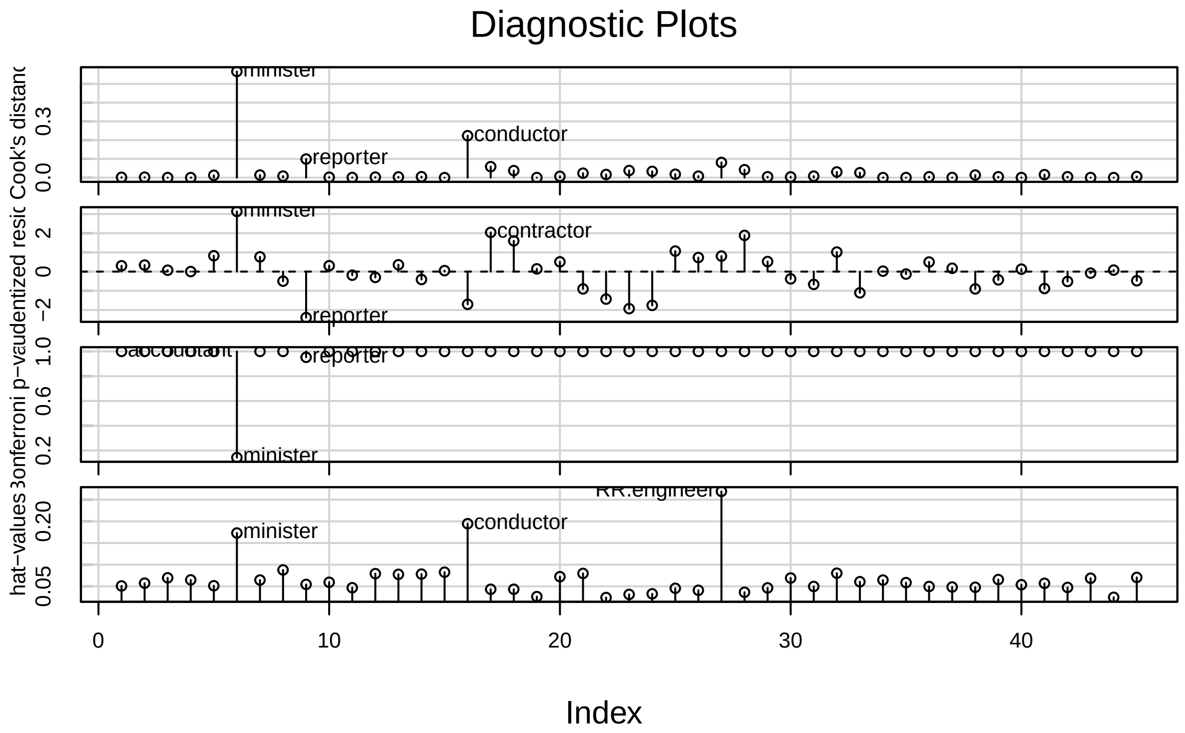
So we can see that both the education and income level of those in the industry are important indicators of prestige. Now lets looks for outliers and high-leverage points.

```
influencePlot(fit1)
```



```
##                 StudRes       Hat     CookD
## minister    3.1345186 0.1730582 0.7525820
## RR.engineer 0.8089221 0.2690896 0.2845489
```

```
influenceIndexPlot(fit1, id.n=3)
```



Both conductor and rail road engineer are high leverage points because of their relatively high salary but moderately low level of education level. Ministers are the most influential observation because they have a low income given their high level of education.

It is worth repeating the analysis removing these datapoints.

```
fit2 <- update(fit1,
       subset = rownames(Duncan) != "minister")
compareCoefs(fit1,fit2)
```

```
##
## Call:
## 1: lm(formula = prestige ~ education + income, data = Duncan)
## 2: lm(formula = prestige ~ education + income, data = Duncan, subset =
##   rownames(Duncan) != "minister")
##              Est. 1    SE 1  Est. 2    SE 2
## (Intercept) -6.0647  4.2719 -6.6275  3.8875
## education    0.5458  0.0983  0.4330  0.0963
## income       0.5987  0.1197  0.7316  0.1167
```

So removing minister has increased the income coefficient by about 20% and has decreased the education coefficient similarly.

```
fit3 <- update(fit1,
               subset = !(rownames(Duncan) %in% c('minister', 'conductor')) )
compareCoefs(fit1, fit2, fit3, se=FALSE)
```

```
##
## Call:
## 1: lm(formula = prestige ~ education + income, data = Duncan)
## 2: lm(formula = prestige ~ education + income, data = Duncan, subset =
##   rownames(Duncan) != "minister")
## 3: lm(formula = prestige ~ education + income, data = Duncan, subset =
##   !(rownames(Duncan) %in% c("minister", "conductor")))
##             Est. 1 Est. 2 Est. 3
## (Intercept) -6.065 -6.628 -6.409
## education    0.546  0.433  0.332
## income       0.599  0.732  0.867
```

So removing these two outliers does have a considerable effect on the estimated coefficients. Should we remove them before presenting our results? That is hard to say. I think the honest approach would be to highlight this sensitivity in the report, but on balance, I think there are clear reasons why these two professions buck the trend (Ministers tend to have a calling and accept a low salary and conductors are a little like footballers - they are a select bunch of very successful individuals who probably sacrificed school to get to where they are), and so the more honest trend is probably represented by the model that excludes these two professions. The case of rail road engineers is less clear cut, and so I would leave these in. Note that these are my subjective judgements, and would need careful explanation and justification in any report.