STAT350 Tutorial 6

13/10/2020

This week's tutorial will cover diagnostics for leverage and influence. This is an extension to what we covered in previous tutorials with regards to outliers (observations with unusual response values) and residual analysis, in that we will be looking at the data we have for points that effect our model.

We will be using a different dataset this week; the savings data from the `faraway` package which includes per capita financial information for 50 countries.

```
library(faraway)
data(savings)
head(savings)
```

```
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
```

In this dataset `sr` is the response and respresents the ratio of savings to disposable income, `dpi` is the disposable income, `ddpi` is the rate of change for diposable income, and `pop15` and `pop75` represent the percentage of the population which are below the age of 15 and above the age of 75 respectively.

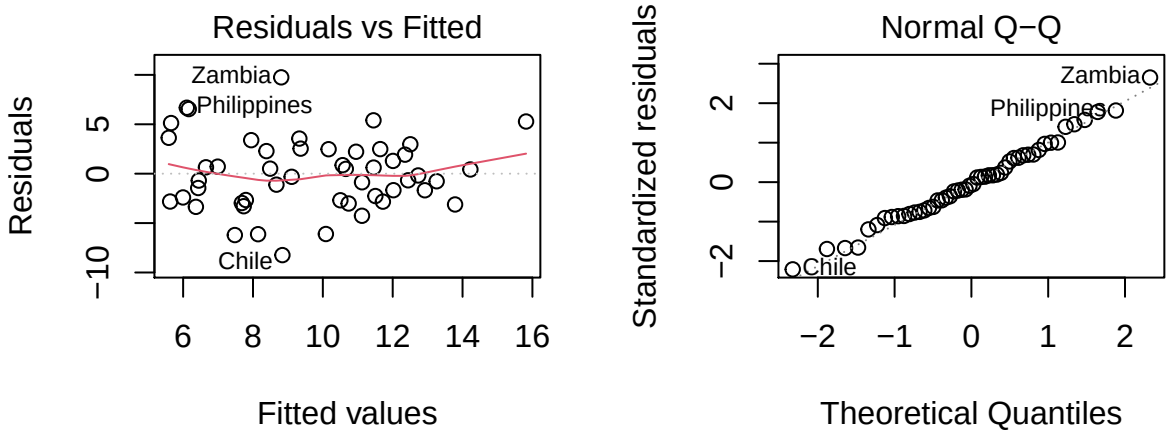
I'm now going to fit the multiple linear regression model with all four predictors included. Here using `.` to the right of the `~` tells R to use all other variables in the model.

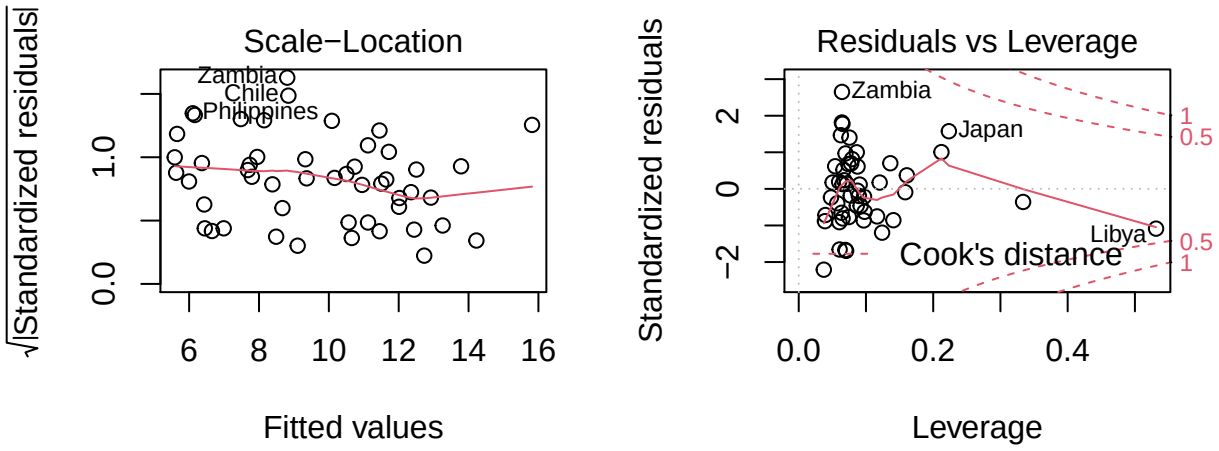
```
mdl <- lm(sr ~ ., data = savings)
(mdl_sum <- summary(mdl))
```

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

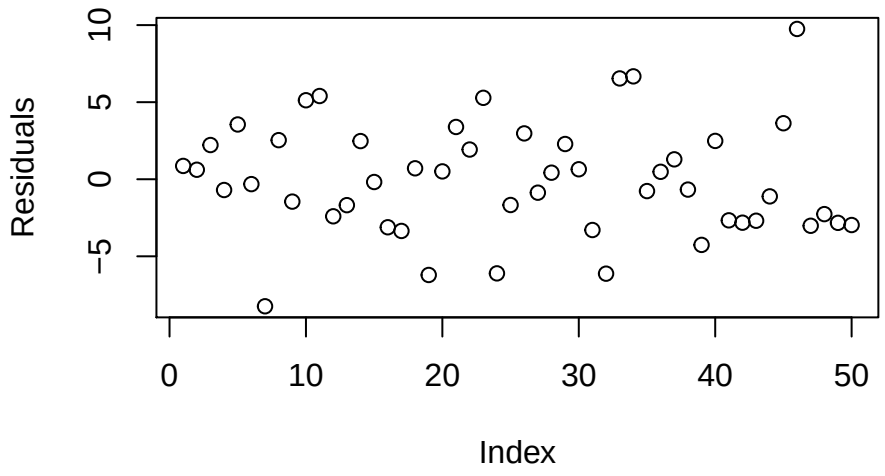
```
## lm(formula = sr ~ ., data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

So, everything looks good in the model summary – now we'll check assumptions





From the first plot we can see that the linear assumption is met. The second plot shows us that the normality of errors assumption is met. The third along with the first plot allow us to say that the constant variance assumption is met. Lastly, from the below residuals versus index plot we see that the uncorrelated errors assumption is met.



As for unusual points, Zambia and Chile should be checked due to their relatively large residuals. And, Libya should be checked due to it's high leverage (we'll go more into this in the next section).

Leverage

The leverage of a point is define by it's place in the design space; a point far from all others can have a great effect on the properties of a given regression model. Leverages, h_{ii}, are given by the diagonal elements of the hat matrix:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}.$$

A common rule given is that points who's leverage exceed $2p/n$ should be looked at more closely, where n is the number of data points and p is the number of coefficients in the model. For our dataset that value is:

```
2*ncol(savings)/nrow(savings)
```

```
## [1] 0.2
```

A quick way to get the influences of points in R is with the `influence()` function, using the model object as it's only argument. The `hat` component of the resulting object gives us the diagonal of the hat matrix.

```
mdl_inf <- influence(mdl)
sort(mdl_inf$hat, decreasing = TRUE)
```

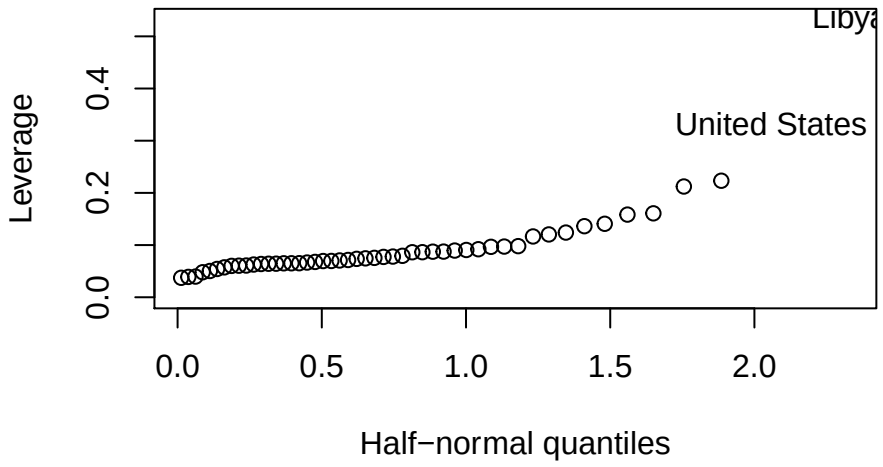
```
##          Libya  United States          Japan        Ireland South Rhodesia
##     0.53145676     0.33368800     0.22330989     0.21223634     0.16080923
##         Canada        Jamaica         France         Sweden        Austria
##     0.15840239     0.14076016     0.13620478     0.12398898     0.12038393
## United Kingdom        Uruguay       Portugal         Greece        Finland
##     0.11651375     0.09794717     0.09714946     0.09662073     0.09204246
##    Netherlands        Bolivia        Belgium        Germany     Luxembourg
##     0.09061400     0.08947114     0.08748248     0.08735739     0.08634787
##      Venezuela          Malta          China          Spain     Costa Rica
##     0.08628365     0.07940290     0.07795899     0.07732854     0.07546780
##         Tunisia    Switzerland          India        Iceland         Brazil
##     0.07456729     0.07359423     0.07145213     0.07049590     0.06955944
##       Paraguay      Australia          Italy       Malaysia   South Africa
##     0.06937188     0.06771343     0.06651170     0.06523300     0.06510405
##           Peru         Zambia    Philippines        Ecuador        Denmark
##     0.06504891     0.06433163     0.06425415     0.06372651     0.06271782
##          Korea      Guatamala       Honduras       Colombia    New Zealand
```

```
##      0.06079915      0.06049212      0.06008079      0.05730171      0.05421789
##       Nicaragua          Norway          Turkey          Panama           Chile
##      0.05035056      0.04793213      0.03964224      0.03897459      0.03729796
```

So, Libya, the US, Japan, and Ireland have leverages higher than 0.2 and should be looked at closer as they have the potential to be influential. Typically a point which has high leverage paired with a large residual it is likely to be influential. None of these countries had large residuals, so we'll have to dig a little more.

The `halfnorm()` function of the `faraway` package provides a useful way to visualize these leverages.

```
halfnorm(mdl_inf$hat, labs = names(mdl_inf$hat), ylab = 'Leverage')
```



Influence

Influential points in the data are points that if removed would cause noteworthy changes in the estimates of the model coefficients. They typically have large leverages paired with unusual response values, therefore giving it more say in the fitted model values compared to other points.

Cook's Distance

Cook's distance is a measure of the distance between predicted values between models fit with and without point i. Your textbook says points with $D_i > 1$ are considered to be influential.

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

Now, I'll calculate the cook's distance for our model using the `cooks.distance()` function with the model object as it's only argument.

```
mdl_cook <- cooks.distance(mdl)
sort(mdl_cook, decreasing = TRUE)
```

```
##          Libya          Japan         Zambia        Ireland    Philippines
##   2.680704e-01   1.428162e-01   9.663275e-02   5.439637e-02   4.522120e-02
##           Peru        Iceland       Paraguay         Sweden          Chile
##   4.401457e-02   4.352902e-02   4.157229e-02   4.055963e-02   3.781324e-02
##          Korea     Costa Rica        Denmark        Jamaica      Venezuela
##   3.555386e-02   3.207537e-02   2.879580e-02   2.402677e-02   1.886141e-02
##         Greece         France United Kingdom         Brazil  United States
##   1.590102e-02   1.547176e-02   1.496628e-02   1.402735e-02   1.284481e-02
##          Malta      Guatamala        Tunisia       Malaysia        Uruguay
##   1.146827e-02   1.067111e-02   9.562447e-03   9.113404e-03   8.532329e-03
##          China    Switzerland        Belgium         Panama        Ecuador
##   8.156984e-03   7.334746e-03   7.154674e-03   6.333674e-03   5.818699e-03
## South Rhodesia    New Zealand        Finland         Turkey     Luxembourg
##   5.267290e-03   4.379219e-03   4.364051e-03   4.224370e-03   3.993882e-03
##          Italy       Colombia       Portugal        Austria      Australia
##   3.919100e-03   1.879460e-03   9.733900e-04   8.175997e-04   8.035888e-04
##        Bolivia          Spain         Norway       Honduras      Nicaragua
##   7.278744e-04   5.659085e-04   5.558570e-04   4.741920e-04   3.226479e-04
##         Canada          India    Netherlands   South Africa        Germany
##   3.106199e-04   2.965778e-04   2.744377e-04   2.405063e-04   4.736572e-05
```

So, we see that for our data and model we don't get any D_i larger than 1, with Libya having the greatest value at 0.268.

How does the model change if we delete Libya from the data set?

```r
mdl_lib <- lm(sr ~ ., savings, subset = (row.names(savings) != 'Libya'))
summary(mdl_lib)
```

```
##
## Call:
## lm(formula = sr ~ ., data = savings, subset = (row.names(savings) !=
##     "Libya"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0699 -2.5408 -0.1584  2.0934  9.3732
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.5240460  8.2240263   2.982  0.00465 **
## pop15       -0.3914401  0.1579095  -2.479  0.01708 *
## pop75       -1.2808669  1.1451821  -1.118  0.26943
## dpi         -0.0003189  0.0009293  -0.343  0.73312
## ddpi         0.6102790  0.2687784   2.271  0.02812 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.795 on 44 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.2968
## F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

Here's the original model summary:

```
##
## Call:
## lm(formula = sr ~ ., data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

So, the biggest difference I see is in the coefficient for `ddpi`, changing from 0.41 to 0.61 – quite the change considering the magnitude of the coefficient.

So, what do we do with these points?

The answer to this question is heavily dependent on context and similar to how we treat outliers. That is, if we know the point is influential due to experimental error it is safe to delete the point. Otherwise we should leave it in. There are methods of regression that are less susceptible to changes caused by the presence of influential points.