IDS 702: Module 1.6

Outliers and influential points

Dr. Olanrewaju Michael Akande

Leverage, influence, and standardized residuals

- Individual observations can have large impact on the estimates of coefficients and SEs.
- Sometimes these points are obvious from scatter plots, and sometimes they are not, especially in multivariate data.
- Concepts and metrics of leverage, influence, and standardized residuals can help identify impactful and unusual points.
- An outlier is a data point whose value does not follow the general trend of the rest of the data.
- When does a data point have high leverage? When is a data point influential? How can we identify them?
- Those are the questions we seek to answer in this module.

Leverage

- Points with **extreme predictor/covariate/feature values** are called high leverage points.
- That is, the predictor values for these points are far outside the range of values for most of the other points.
- Thus, leverage has nothing to do with values of the response variable $\boldsymbol{y}$.
- Leverage points POTENTIALLY have large impact on the estimates of coefficients and SEs.

How?

- First, note that the leverage score h_{ii}, for observation i, is defined as the i^{th} diagonal element of the projection or hat matrix.

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

QUICK EXERCISE

- Just to see what the hat matrix (and leverage scores) looks like, you will compute it for a very simple example.
- Open R/RStudio on your computer. Suppose the design matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.0 \\ 1 & 2.0 \\ 1 & 2.5 \\ 1 & 3.5 \\ 1 & 50.0 \end{bmatrix}$$

  that is, we have one predictor and an intercept. You can set this up in R using the matrix function.
- Compute the corresponding hat matrix for this design matrix

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

- Compare that leverage score to the original rows of $\boldsymbol{X}$.
- Which diagonal element is the largest? What do you think about that observation?

Leverage

- Recall that

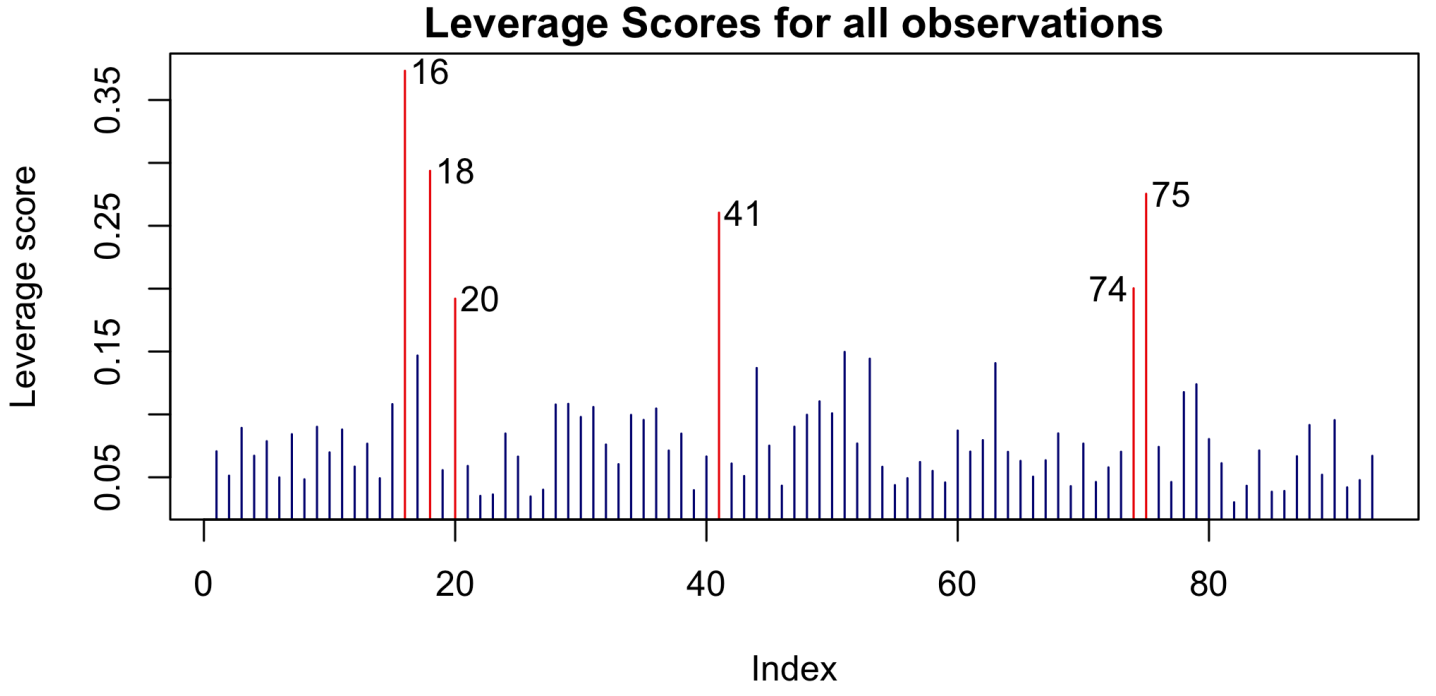
$$\hat{\boldsymbol{y}} = \left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}.$$

- The leverage score h_{ii} for observation i measures how far away the values of the independent variables for the i^{th} observation are from those of other observations.
- That leverage score then clearly impacts predictions since, again, $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. Think about the exercise you just completed.
- Some properties of h_{ii}:
  - $0 \leq h_{ii} \leq 1$.
  - $\mathbb{V}\text{ar}[e_i] = (1 - h_{ii})\sigma^2$.
  - High leverage points are often determined by paying attention to any observation for which $h_{ii} > 2(p+1)/n$.
  - Points with h_{ii} close to 1 will have more of an impact on model fit.

Back to our example

Let's identify any high leverage points. Here, $2(p+1)/n = 16/93 \approx 0.17$.

```
n <- nrow(model.matrix(regwagecsquares)); p <- ncol(model.matrix(regwagecsquares))
lev_scores <- hatvalues(regwagecsquares) #can also use influence(regwagecsquares)$hat
plot(lev_scores,col=ifelse(lev_scores > (2*p/n), 'red2', 'navy'),type="h",
     ylab="Leverage score",xlab="Index",main="Leverage Scores for all observations")
text(x=c(1:n)[lev_scores > (2*p/n)]+c(rep(2,4),-2,2),y=lev_scores[lev_scores > (2*p/n)],
     labels=c(1:n)[lev_scores > (2*p/n)])
```



High leverage: what to do?

- Points with high leverage deserve special attention:
  - Make sure that they do not result from data entry errors.
  - Make sure that they are in scope for the types of individuals for which you want to make predictions.
  - Make sure that you look at the impact of those points on estimates, especially when you have interactions in the model.
- Just because a point is a high leverage point does not mean it will have a large effect on regression.
- When a point has a large effect on the regression, we say that the observation is influential.
- Whether or not a high leverage point actually affects the regression line depends on the value of the response variable y.

Cook's distance

- What if a point has a large impact on the estimates of the regression coefficients?
  - Dropping that point should change the coefficients significantly.
  - Consequently, a significant change in the coefficients should also change that point's predicted y_i value by a lot.
- For every point, we could delete it, re-run the regression, and see which points lead to big changes in the predicted y_i's; very time consuming!
- However, Cook's distance gives a formula for quantifying the influence of the i^{th} observation, if it is removed from the sample. We have

$$D_i = \sum_{j=1}^{n} \frac{\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{s_e^2(p+1)}$$

where $\hat{y}_{j(i)}$ is the predicted value after excluding the i^{th} observation.

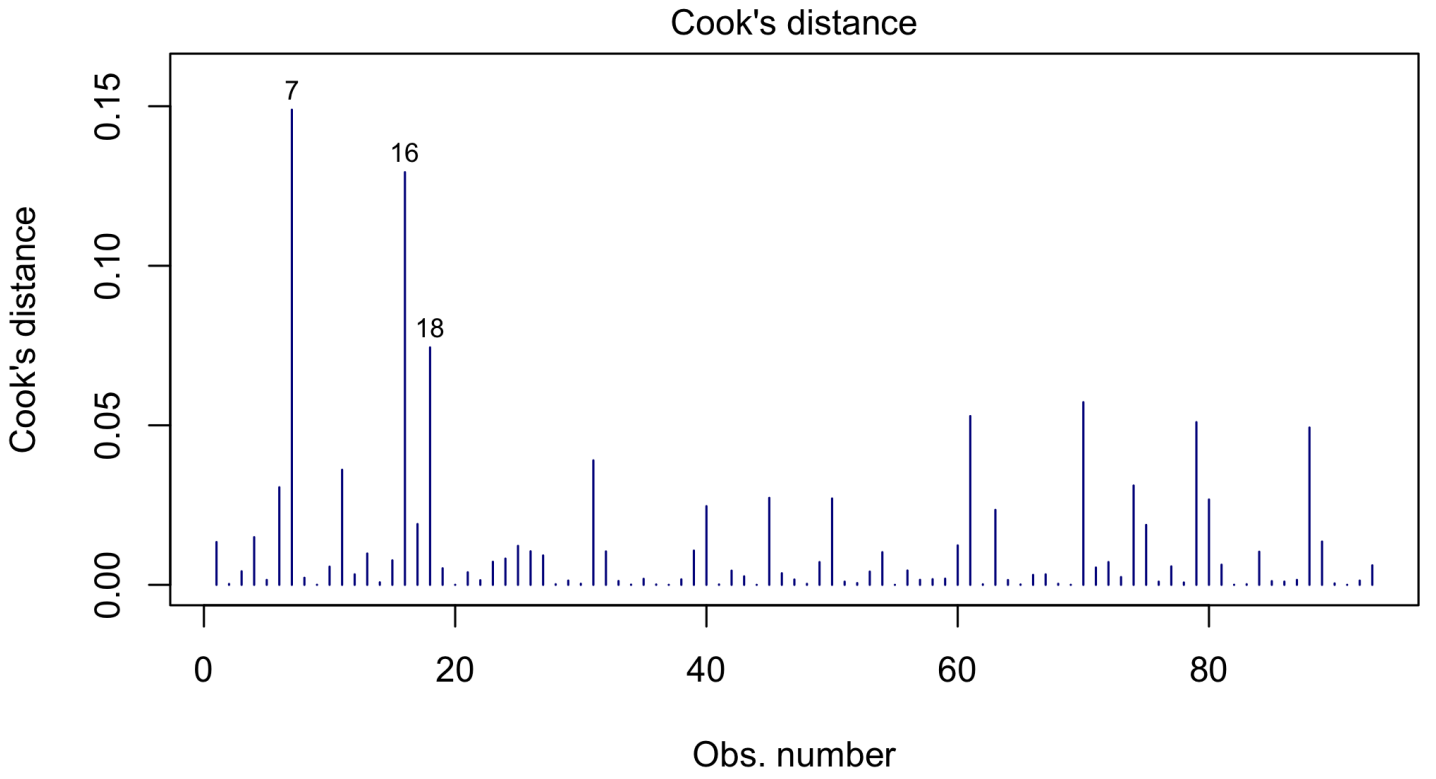
Big Cook's distances: what to do?

- Examine Cook's distances to look for large values.
  - Make sure there are no data entry errors in those points.
  - For each point with high Cook's distance, fit the model with and without that point, and compare the results.
- The consensus seems to be that $D_i > 1$ indicates an observation is an influential value, but we generally pay attention to observations with $D_i > 0.5$.
- If the results (predictions or scientific interpretations) do not change much, just report the final model based on all data points and you don't really need to report anything about the Cook's distances.
- If results change a lot, you have several options...

Back to our example

Can we try to identify any influential points?

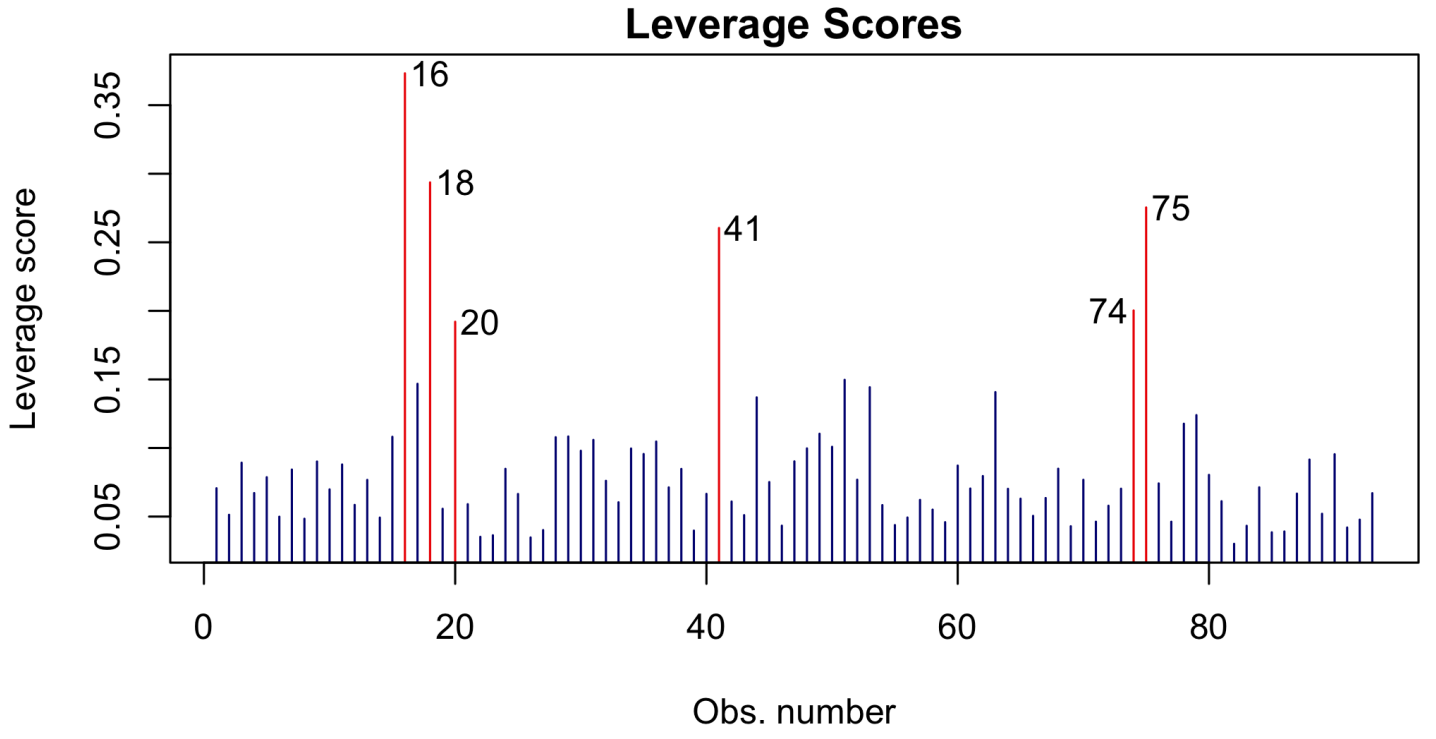
```
plot(regwagecsquares,which=4,col=c("blue4"))
```



Let's compare to the leverage score

Which of the potentially influential points actually have high leverage?

```
plot(lev_scores,col=ifelse(lev_scores > (2*p/n), 'red2', 'navy'),type="h",
     ylab="Leverage score",xlab="Obs. number",main="Leverage Scores")
text(x=c(1:n)[lev_scores > (2*p/n)]+c(rep(2,4),-2,2),y=lev_scores[lev_scores > (2*p/n)],
     labels=c(1:n)[lev_scores > (2*p/n)])
```



Cook's distance: What to do if large changes in results?

- It is generally OK to drop observations based on PREDICTOR values if

  1. It is scientifically meaningful to do so; and

  2. You intended to fit a model over the smaller X range to begin with (and just forgot). When this is the case, you should mention this in your analysis write-up and be careful when making predictions to avoid extrapolation.

- It is generally NOT OK to drop an observation based on its RESPONSE value (assuming no data errors in that value). These are legitimate observations and dropping them is essentially cheating by changing the data to fit the model.

- You should try transformations or collect more data.

Standardized residuals (also called internally studentized residuals)

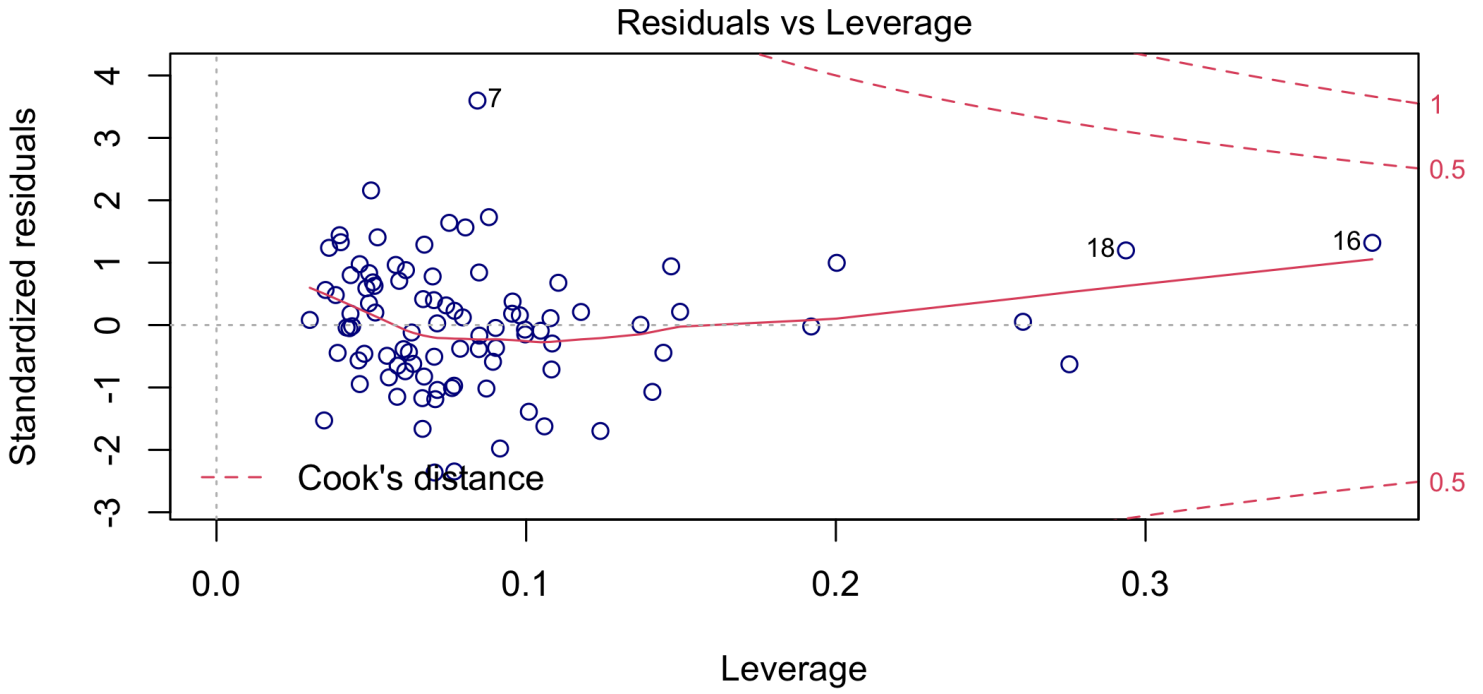
- How do we best identify outliers, i.e., points that don't fit the pattern implied by the line? We look for points with relatively large residuals.
- It would be nice to have a common scale to interpret what a "big" residual is, across all problems.
- As with most metrics in statistics, we look at each residual divided by its standard error (hence the term standardized residual).
- The SE of any residual (that is, e_i and not ϵ_i) depends on the values of the predictors.
- As such, it turns out that the residuals for high leverage predictors have smaller variance than residuals for low leverage predictors.
- Intuition: the regression line tries to fit high leverage points as closely as possible, which results in smaller residuals for those points.

Standardized residuals (also called internally studentized residuals)

- Standardized residuals have a Normal(0,1) distribution.
- Values with large standardized residuals are outliers.
- How large is large? Well, remember that 95% of any normal distribution should lie within 2 standard deviations of the mean...
- Values with large standardized residuals are not necessarily influential on the regression line. A point can be an outlier without impacting the line. We need to examine their Cook's distance to determine influence.
- It turns out that the Cook's distance D_i can also be expressed using the leverage score h_{ii} and square of the internally Studentized residuals.
- Bottomline: make a plot of the standardized residuals to check for outliers, but also find a way to add leverage scores and investigate observations with high Cook's distance in the same plot.
- Very easy to do in R.

Standardized residuals: What to do if large outliers?

```
plot(regwagecsquares,which=5,col=c("blue4"))
```



Are there any outliers or influential points?

Standardized residuals: What to do if large outliers?

- As before, it is generally OK to drop observations based on PREDICTOR values if

  1. It is scientifically meaningful to do so; and

  2. You intended to fit a model over the smaller X range to begin with (and just forgot). When this is the case, you should mention this in your analysis write-up and be careful when making predictions to avoid extrapolation.

- It is generally NOT OK to drop an observation based on its RESPONSE value (assuming no data errors in that value). These are legitimate observations and dropping them is essentially cheating by changing the data to fit the model.

- You should try transformations or collect more data.

- Or just do nothing! It can be okay to have some outliers. Examine their influence on your results and report them.

What's next?

Move on to the readings for the next module!