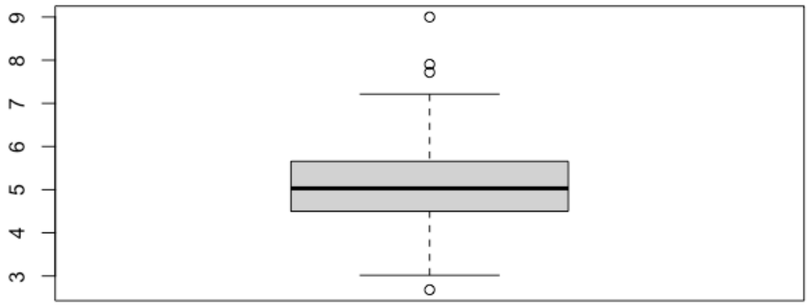
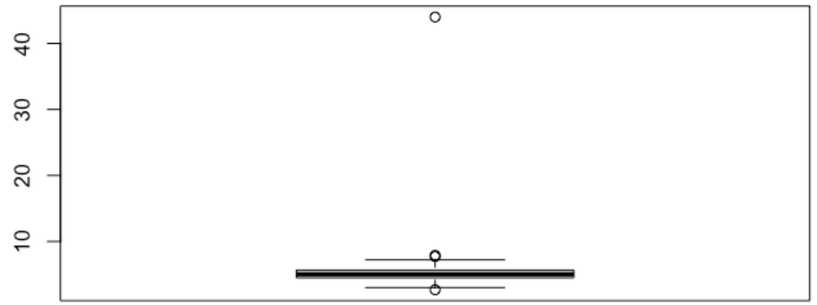
IDS 702

Influential points

Outliers

- An **outlier** is a data point whose value does not follow the general trend of the rest of the data for a given variable
- Outliers should be investigated as potential data errors/implausible values (and corrected if possible)
- If the value does not result from a data error/implausible value, do not remove it





Influential points

- Individual observations can have a large impact on the model (estimates, SE, R^2, RSE)
- Sometimes the points are obvious from EDA, but other times they are not

Note about the model in matrix notation

Leverage

- Points with extreme **predictor/covariate/feature** values are called high leverage points
- The **leverage score** measures how far away the values of the independent variables for the ith observation are from those of other observations
- The leverage score for an observation is defined as the ith diagonal element of the hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}$
- $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = p + 1$

High leverage: what to do?

- Make sure they do not result from data entry errors
- Make sure you look at the impact of those points on the estimates: just because a point is high leverage does not mean it will have a large effect on regression!
- When a point has a large effect on the regression, we say the observation is influential
    - This depends on the value of y

Cook's distance

- Quantifies the influence of the ith observation
- Measures how far, on average, predicted values will move if a given observation is removed from the dataset
- $\hat{y}_{j(i)}$ is the predicted value after excluding the ith observation

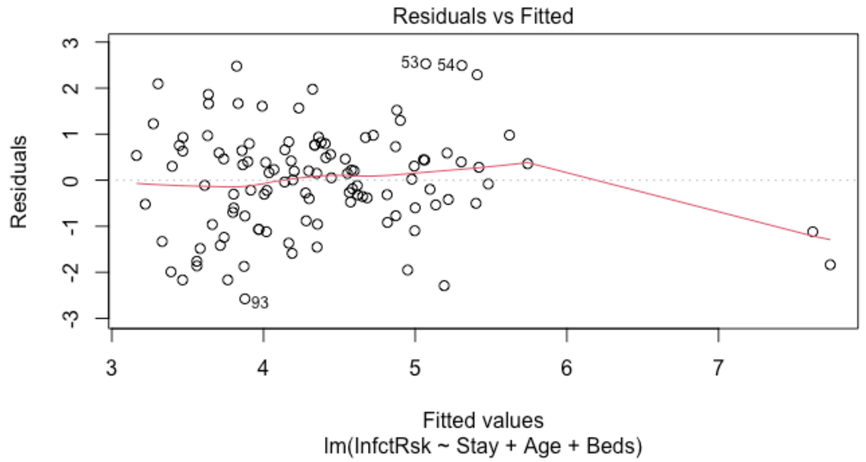
$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_j - \hat{y}_{j(i)})^2}{s_e^2(p+1)}$$

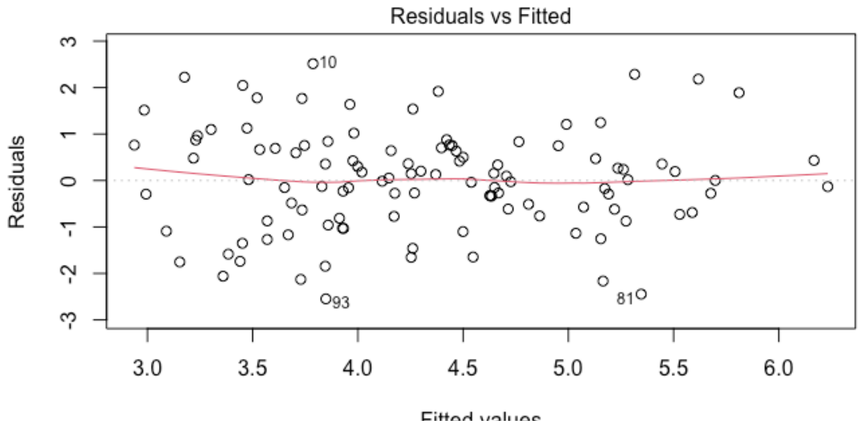
Large Cook's distance: what to do?

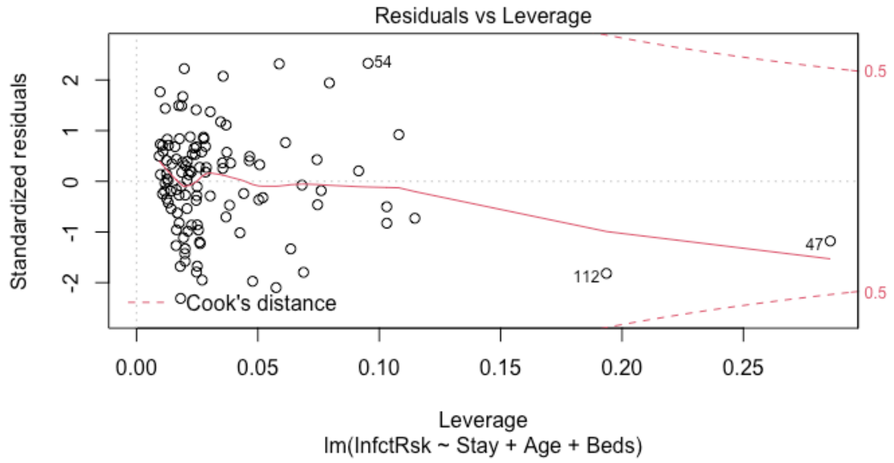
- General consensus is that $D_i > 1$ indicates an observation is an influential value, but we generally pay attention to $D_i > 0.5$
- For each observation with high Cook's distance, fit the model with and without that observation, and compare the results

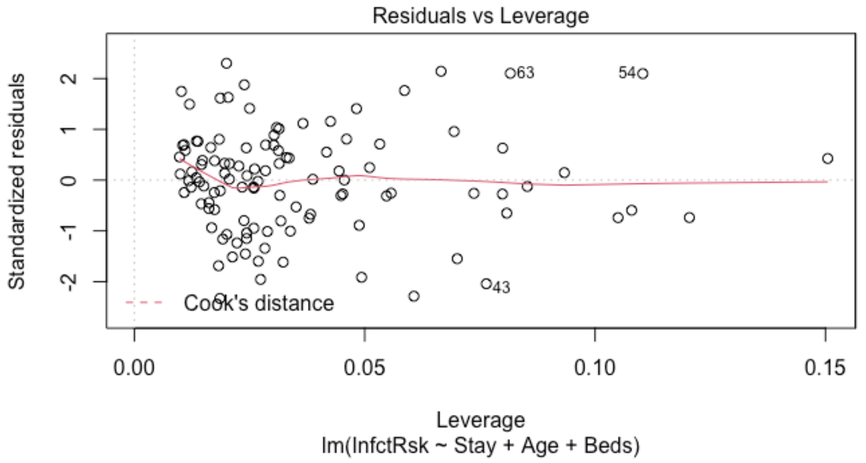
Diagnostic plots in R

plot(model, which=5) gives residuals vs leverage plot









Summary: What to do with outliers/influential observations?

- Make sure the observation does not arise from a data entry error
  - If it does, it can be changed or excluded
- May want to report results with and without influential observation(s)
  - With a large sample size, you may not see a difference
- Be wary of numeric cutoffs!