To formalize the notion of influence consider fitting the model $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$ with all of the data, and fit it again on the remaining $n-1$ observations when the i^{th} observation is removed.

Let $\hat{\vec{\beta}}_{(i)}$ denote the estimate of $\vec{\beta}$ without the i^{th} observation, and let $\hat{\vec{\beta}}$ be the usual LSE of $\vec{\beta}$ based on all of the data. Thus interest lies in comparing $\hat{\vec{\beta}}_{(i)}$ with $\hat{\vec{\beta}}$

- if they're very different then we say observation i has a large influence.
- if they're similar, then we say observation i is not very influential.

We formalize this decision using <u>Cook's D-statistic</u> which is a standardized distance metric between $\hat{\vec{\beta}}_{(i)}$ and $\hat{\vec{\beta}}$. The standardization accounts sampling variability, so that we ignore differences that may have just happened by chance.

$$D_i = \frac{(\hat{\vec{\beta}} - \hat{\vec{\beta}}_{(i)})^T (X^TX)(\hat{\vec{\beta}} - \hat{\vec{\beta}}_{(i)})}{\hat{\sigma}^2 (p+1)}$$

This formula suggests that we would need to fit $n+1$ regressions inorder to calculate D_i, $i=1,2,\dots,n$, **BUT** we don't have to. It turns that $\hat{\vec{\beta}}_{(i)}$ can be computed using only information from the regression on all of the data.

It can be shown (by you, on A4) that

$$\hat{\vec{\beta}}_{(i)} = \hat{\vec{\beta}} - \left(\frac{e_i}{1-h_{ii}}\right)(X^TX)^{-1}\vec{x}_i$$

where $\vec{x}_i$ is the i^{th} row of X corresponding to the "deleted" observation.

Rearranging this equation yields:

$$\hat{\vec{\beta}} - \hat{\vec{\beta}}_{(i)} = \left(\frac{e_i}{1-h_{ii}}\right)(X^TX)^{-1}\vec{x}_i$$

Substituting this into the Cook's-D formula:

$$D_i = \frac{\left(\frac{e_i}{1-h_{ii}}\right)\vec{x}_i^T (X^TX)^{-1}(X^TX)\left(\frac{e_i}{1-h_{ii}}\right)(X^TX)^{-1}\vec{x}_i}{\hat{\sigma}^2(p+1)}$$

$$= \frac{e_i^2\, \vec{x}_i^T (X^TX)^{-1}\vec{x}_i}{(1-h_{ii})^2\, \hat{\sigma}^2 (p+1)}$$

h_{ii}

Recall $d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

d_i^2 is in here

$$= \frac{d_i^2 h_{ii}}{(1-h_{ii})(p+1)}$$

This version of the formula illustrates what makes an influential observation influential: an observation i is influential when it is an outlier in both the y <u>and</u> x dimensions. (d_i and h_{ii} <u>both</u> have to be large).

To identify highly influential observations, find the one(s) with Cook's-D values much larger than the others. Also $D_i > 0.5$ is definitely cause for concern.

<u>Example</u>: Air Quality

Modeling the relationship between air quality (ozone) and three explanatory variables: solar radiation, windspeed, temperature. We found that all three variables significantly influence the response and that together they explain about 61% of the variation in ozone values.

When evaluating the residuals it became evident that an outlier (in the y dimension) existed and that the constant variance and normality assumptions are suspect.

Using leverages and Cook-D values we found that observation 30 had high leverage but not high influence, so it's not a concern. However, observation 77 (the outlier in the y dimension) had high influence and so we deleted it.

The analysis without it, although more accurate, still had issues with the residuals that need to be taken care of....

<u>Variance Stabilizing Transformations</u>

When the plot of residuals vs. fitted values indicates non-constant variance, we want to alter our regression model. What we do, is pick a <u>transformation</u> of the response variable that <u>stabilizes the variance</u>.

$g(\cdot)$

i.e. $Var[y_i]$ is non-constant

Choose $g(\cdot)$ such that $Var[g(y_i)]$ is constant

We then perform the regression using $g(y_i)$ as the response, rather than y_i:

$$g(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

<u>So how do we choose $g(\cdot)$?</u>

In general $y_i = \mu_i + \varepsilon_i$, where $\mu_i = E[y_i]$.

When the variance of y_i is non-constant it is typically because $Var[y_i]$ is some function of μ_i:

$$Var[y_i] = Var[\varepsilon_i] = [h(\mu_i)]^2\sigma^2$$

where $h(\cdot)$ is some unknown function. The idea is to find $g(y_i)$ such that $Var[g(y_i)]$ is constant. We'll do this by considering a first-order Taylor series approximation of $g(y_i)$ around μ_i:

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

Thus, the variance of the transformed response can be approximated by:

$$Var[g(y_i)] \approx Var[g(\mu_i) + (y_i - \mu_i)g'(\mu_i)]$$

$$= [g'(\mu_i)]^2 Var[y_i]$$

$$= [g'(\mu_i)]^2[h(\mu_i)]^2\sigma^2$$

So to stabilize the variance we need to choose a transformation $g(\cdot)$ such that

$$g'(\mu_i) = \frac{1}{h(\mu_i)}$$