Identify points with high leverage

Recall the matrix format of MLR:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

$\boldsymbol{H}$ is called "hat matrix" because it's putting "hat" on $\boldsymbol{y}$ (yeah...). If we actually perform the matrix multiplication on the right side of this equation, we can see that

$$+\cdots+\boxed{h_{ii}y_i}+\cdots$$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n$$

Identify points with high leverage

$\hat{y}_i = 0(y_1) + 0(y_2) + \cdots + (1)y_i + \cdots (0)(y_n) = y_i$

Since

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n$$

- h_{ii} quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$.
- if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$. On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$.
- h_{ii} is defined as the **leverage** of the i_{th} data point.

Q: What happens if $h_{ii} = 1$ & $h_{ij} = 0 \ \forall j \neq i$?

$$H = X(X^T X)^{-1} X^T$$

$\Rightarrow$ projection

$$H = U \Lambda U^T$$

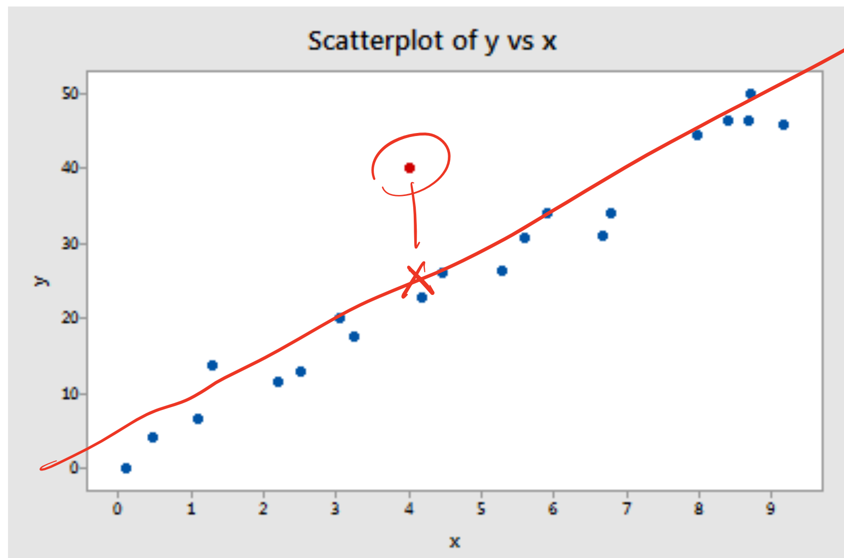
$$H^2 = U \Lambda \cancel{U^T U} \Lambda U^T = U \Lambda^2 U^T$$

$$\Lambda = \Lambda^2 = \lambda_i = \lambda_i^2$$

Properties of leverages

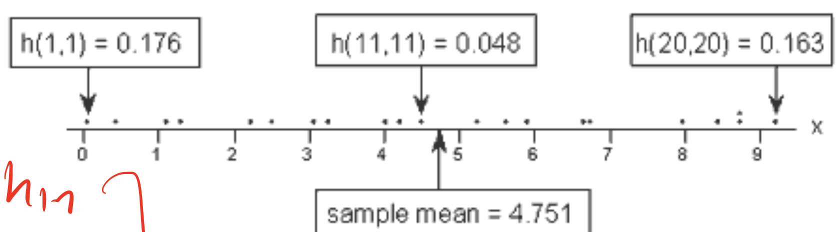
- h_{ii}= measure of the distance between the x value for the i_{th} data point and the mean of the x values for all n data points.

"hat_matrix_diag" in get_influence()



$h_{ii} \approx 0$

(x is not extreme in the range of predictor data)



X_i on the edge of pt. cloud

$\Rightarrow$ higher leverage

$$\underset{n\times n}{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & & \\ & & & & \\ & & & & \end{bmatrix}$$

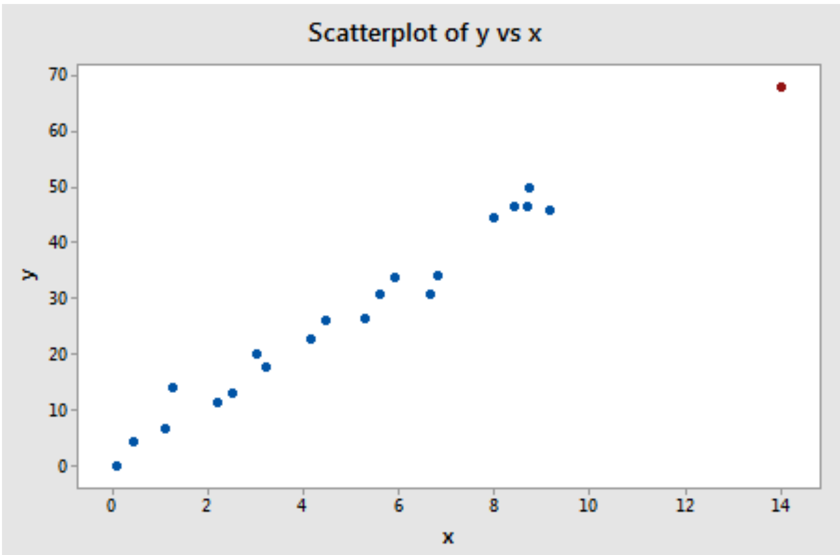
$$h_{ii} = x_i (X^T X)^{-1} x_i^T$$

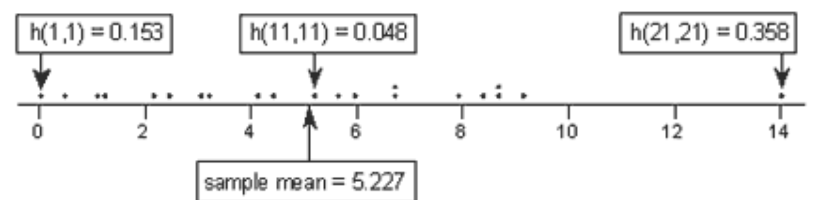
$$\hat{y}_i = \sum_{j=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSx}\right) y_j$$

Mahalanobis distance

Properties of leverages

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ & & & \\ & & & \end{bmatrix}$$





Hint: $H^2 = H \Leftarrow i^{th}$ diag of $H = \underline{h_{ii}} = \sum_{j=1}^{n} h_{ij}^2 \geq \underline{h_{ii}^2}$

Other notes:

- h_{ii} is between 0 and 1.
- $\boxed{\sum_{i=1}^{n} h_{ii} = p}$

Can you show these?

$tr(H) = p \Rightarrow \bar{h}_{\cdot} \approx \frac{p}{n}$

$= tr(X(X^TX)^{-1}X^T)$

Identify data points with high leverage

The great thing about leverages is that they can help us identify x values that are extreme and therefore potentially influential on our regression analysis.

- A common rule is to flag any observation whose leverage value, h_{ii}, is more than 3 times larger than the mean leverage value
- $\bar{h} = \frac{\sum h_{ii}}{n} = \frac{p}{n}$
- Flag any observation whose leverage value $h_{ii} > \frac{3p}{n}$

high leverage point

Use externally studentized residuals to detect outliers

There are several ways to check for outliers in regression. For our class, we will focus on using the "externally studentized residual."

$e_i = y_i - \hat{y}_i$

Def:

$$\text{stud}(e_i) = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \sim t_{(n-1)-p}$$

where $MSE_{(i)} = \hat{\sigma}^2$ from the model fit on all observations except the i^{th} data point.

$$MSE_{(i)} = \hat{\sigma}^2_{(i)} = \frac{SSE_{(i)}}{(n-1)-p} = \frac{\sum_{j=1, j\neq i}^{n} (e_{j(i)})^2}{n-1-p}$$

Idea:

$Var(\underline{e}) = \sigma^2 (I - H)$

$\Rightarrow \hat{SE}(e_i) = \sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}$

but if i^{th} data point is an outlier, it could skew our estimate of σ^2... so we leave it out to be safe! ("External")

Since $\text{stud}(e_i) \sim t_{(n-1)-p}$ dist, if

$$|\text{stud}(e_i)| > t^*_{(n-1)-p, (1-\alpha/2)} \Rightarrow \text{"high discrepancy point"}$$

then we say the i^{th} data point is an outlier.

Influential Points

in Y

in X

- Influential points are a combination of outliers and high leverage points.
- To identify influential points, the basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining n–1 observations. Then, we compare the results using all n observations to the results with the i^{th} observation deleted to see how much influence the observation has on the analysis.

Cook's distance

$MSE = \hat{\sigma}^2 = \frac{SSE}{n-p}$

$MSE_{(i)} = \hat{\sigma}^2_{(i)} = \frac{SSE_{(i)}}{n-p-1}$

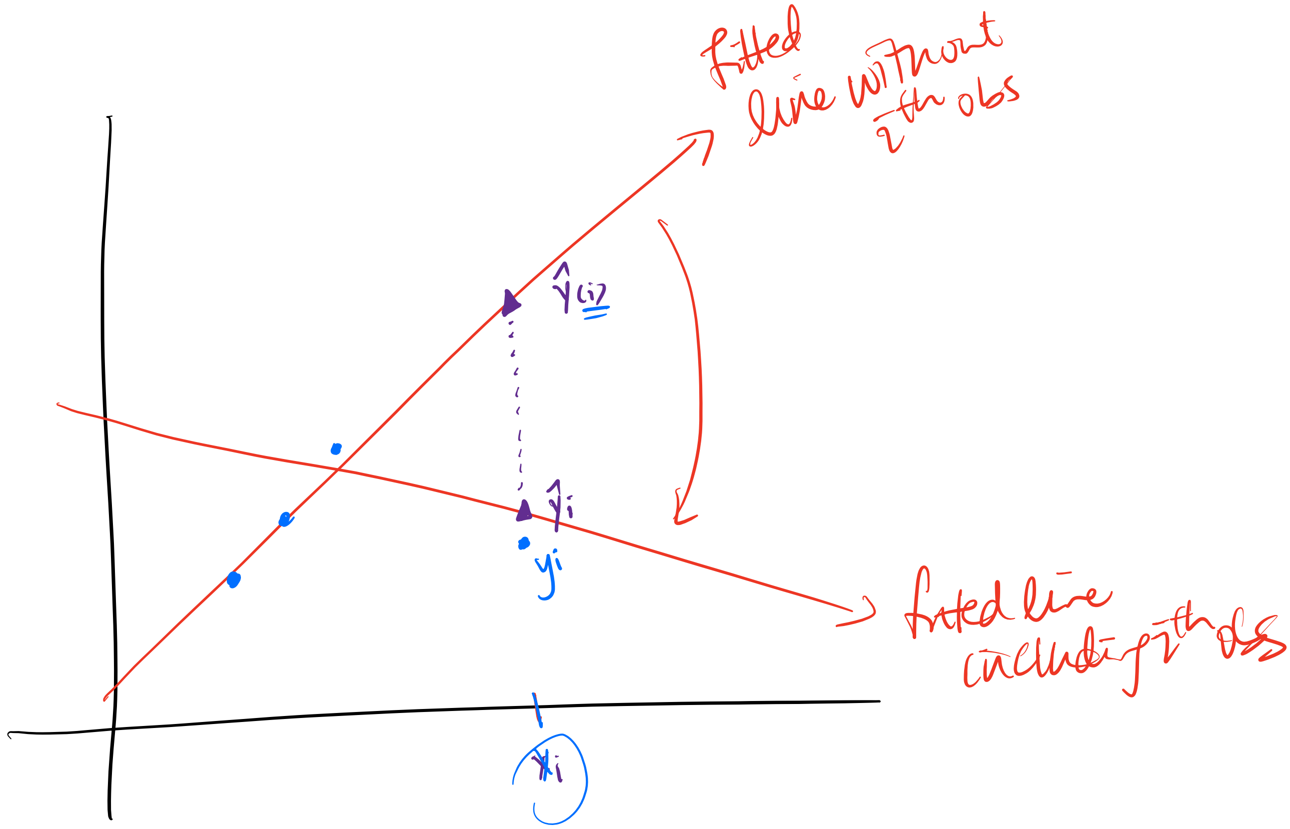
Cook's distance measure, denoted D_i, is defined as:

combination of high leverage & high discrep. pts ⇒ high D_i

$$D_i = \boxed{\frac{\sum(\hat{y}_i - \hat{y}_{(i)})^2}{p * MSE}} \overset{(\dots)}{=\!=} \frac{(y_i - \hat{y}_i)^2}{p * MSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

It looks a little messy, but the main thing to recognize is that Cook's D_i depends on both the residual, e_i, and the leverage, h_{ii}. That is, both the x value and the y value of the data point play a role in the calculation of Cook's distance.

- D_i directly summarizes how much all of the fitted values change when the i^{th} observation is deleted.
- A data point having a large $\boxed{D_i\ (> \frac{4}{n})}$ indicates that the data point strongly influences the fitted values. "Influential point"


fitted line without i th obs
$\hat{y}_{(i)}$
$\hat{y}_i$
y_i
fitted line including i th obs
x_i

A Strategy for Dealing with Problematic Data Points

THERE IS NO SYSTEMATIC WAY! That is, the various measures that we have learned in this lesson can lead to different conclusions about the extremity of a particular data point.Some recommended strategy for dealing with problematic data points:

- Check for obvious data errors:
  - If the error is just a data entry or data collection error, correct it.
  - If the data point is not representative of the intended study population, delete it.
- Consider the possibility that you might have just misformulated your regression model:
  - Did you leave out any important predictors?
  - Should you consider adding some interaction terms?

coming up in Model Selection

A Strategy for Dealing with Problematic Data Points

Decide whether or not deleting data points is warranted:

- Do not delete data points just because they do not fit your preconceived regression model.
- You must have a good, objective reason for deleting data points.
- If you delete any data after you've collected it, justify and describe it in your reports.
- If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.