Three types of outliers:

(Dotted line is fit w/out red dot. Solid is with it.)

(Remember: Scatterplots often won’t help you find outliers when there is more than 1 covariate! That’s why we need numerical measures...)
For the red dot:
\[ e_i = -0.25 \]
\[ d_i = -0.19 \]
\[ t_i = -1.0 \]
\[ h_{ii} = -0.43 \]

test for outlier: \[ t_i \sim t(n-p-1) \] so obs i is outlier if \[ |t| > t(1-.1/(2*n),n-p-1) = 3.22 \] in this case. What do you think of the test’s conclusion for the middle frame?
In the middle frame, the fit with the red dot and fit without it are just about the same. Maybe we shouldn’t care too much about the red dot here (although effect on std errors and p-values might be larger…). The red dot is a low leverage point.

Leverage = potential of observation to influence regression fit.

$h_{ii}$ is a measure of leverage (section 10.3).

$0 \leq h_{ii} < 1, h_{11} + \ldots + h_{nn} = p$

$h_{ii}$ is a measure of distance between $x_i = (1, x_{i1}, \ldots, x_{ip-1})$ and $(1, \text{sum}(x_{i1})/n, \ldots, \text{sum}(x_{ip-1})/n)$

rules of thumb:
1. case $i$ is high leverage if $h_{ii} > 2p/n$
2. leverage > .5 is really high
To distinguish between these 3 types of outliers, we need a measure that combines the information in $t_i$ and $h_{ii}$.

For the red dot:

- $e_i = -0.25$
- $d_i = -0.19$
- $t_i = -1.0$
- $h_{ii} = -0.43$

- $e_i = -0.62$
- $d_i = -0.65$
- $t_i = -4.5$
- $h_{ii} = -0.05$

- $e_i = -0.68$
- $d_i = -1.1$
- $t_i = -6.2$
- $h_{ii} = -0.38$
Three methods to combine “outlyingness” ($t_i$) & leverage ($h_{ii}$) (10.4)

$$DFFITS_i = t_i \sqrt{h_{ii}/(1-h_{ii})}$$

($h_{ii}/(1-h_{ii}) > 1$ when $h_{ii} > 0.5$

For the red dot:

$e_i = -0.25$  $-0.62$  $-0.68$

$d_i = -0.19$  $-0.65$  $-1.1$

$t_i = -1.0$  $-4.5$  $-6.2$

$h_{ii} = -0.43$  $-0.05$  $-0.38$

$DFFITS_i = -0.88$  $-1.0$  $-4.9$

Rules of thumb:

$|DFFITS_i| > 1.0$ means outlier

$|DFFITS_i| > 2\sqrt{p/n}$ means outlier
Cook’s distance:

\[ D_i = \text{sum over } j( (Y\hat{\text{hat}}_j - Y\hat{\text{hat}}_{j(i)})^2 ) / pMSE \]

\[ D_i = \left( \frac{e_i^2}{pMSE} \right) \left( \frac{h_{ii}}{(1-h_{ii})^2} \right) \]

For the red dot:

\[ e_i = -0.25 \quad -0.62 \quad -0.68 \]
\[ d_i = -0.19 \quad -0.65 \quad -1.1 \]
\[ t_i = -1.0 \quad -4.5 \quad -6.2 \]
\[ h_{ii} = -0.43 \quad -0.05 \quad -0.38 \]
\[ \text{DFFITS}_i = -0.88 \quad -1.0 \quad -4.9 \]
\[ \text{Cook’s } D_i = 0.38 \quad 0.26 \quad 4.0 \]

Rules of thumb:

Look at \( \Pr(F>D_i) \) where \( F \sim F(p, n-p) \)
Suspect outlier if \( \Pr(F>D_i) < 0.2 \)
\( (0.69, 0.77, 0.03) \)
DFBETAs:

\[(b_k - b_{k(i)}) / \sqrt{\text{MSE}(i)c_{kk}}\]

\[c_{kk} = \text{kth element of diag}((X'X)^{-1})\]

For the red dot:

\[e_i = \phantom{-}.25 \phantom{-.62} \phantom{-.68}\]
\[d_i = -.19 \phantom{-.65} \phantom{-1.1}\]
\[t_i = -1.0 \phantom{-4.5} \phantom{-6.2}\]
\[h_{ii} = -.43 \phantom{-.05} \phantom{-3.8}\]
\[\text{DFFITS}_i = -.88 \phantom{-1.0} \phantom{-4.9}\]
\[\text{Cook’s } D_i = 0.38 \phantom{.26} \phantom{4.0}\]
\[\text{DFBETA0} = -.84 \phantom{-.62} \phantom{3.2}\]
\[\text{DFBETA1} = -.83 \phantom{.2} \phantom{4.6}\]

Rules of thumb: Outlier if |DFBETA| is

\[> 1 \text{ in small datasets}\]
\[> 2/\sqrt{n} \text{ is large datasets (.44)}\]
Which one you use depends on what you’re going to use the regression model for.

Do you care about prediction and estimated means in general? (DFFITS and Cook’s Distance are good for this)

Do you care about estimated coefficients (DFBETAS are good for this)
If you suspect an outlier, it's always good to describe its influence on an interpretable scale:

% change due to outlier:

$$100 \times \frac{\text{result without outlier} - \text{result with outlier}}{\text{result with outlier}}$$

2% change in $b_0$

-13% change in $b_1$

<2% change in $\hat{y}_i$

2% change in $b_0$

4% change in $b_1$

<2% change in $\hat{y}_i$

9% change in $b_0$

877% change in $b_1$

<9% change in $\hat{y}_i$
10.5: Multicolinearity diagnostic: The variance inflation factor (VIF)

Multicolinearity is when 2 different covariates in a multiple regression are highly correlated.

When that happens:

- Confidence intervals for parameters tend to get wider.
- Parameter estimates can do strange things.

\[ VIF_k = \frac{MSE^*}{1-R_k^2} \]

Where \( MSE^* \) is from regression on standardized response and coef
\[ y_i^* = \frac{(y_i-ybar)}{(s_y*sqrt(n-1))} \]
\[ x_{ik}^* = \frac{(x_{ik}-xbar_k)}{(s_k*sqrt(n-1))} \]

rule of thumb:
\[ \text{max over } k \ \text{VIF}_k > 10 \text{ indicates problems...} \]