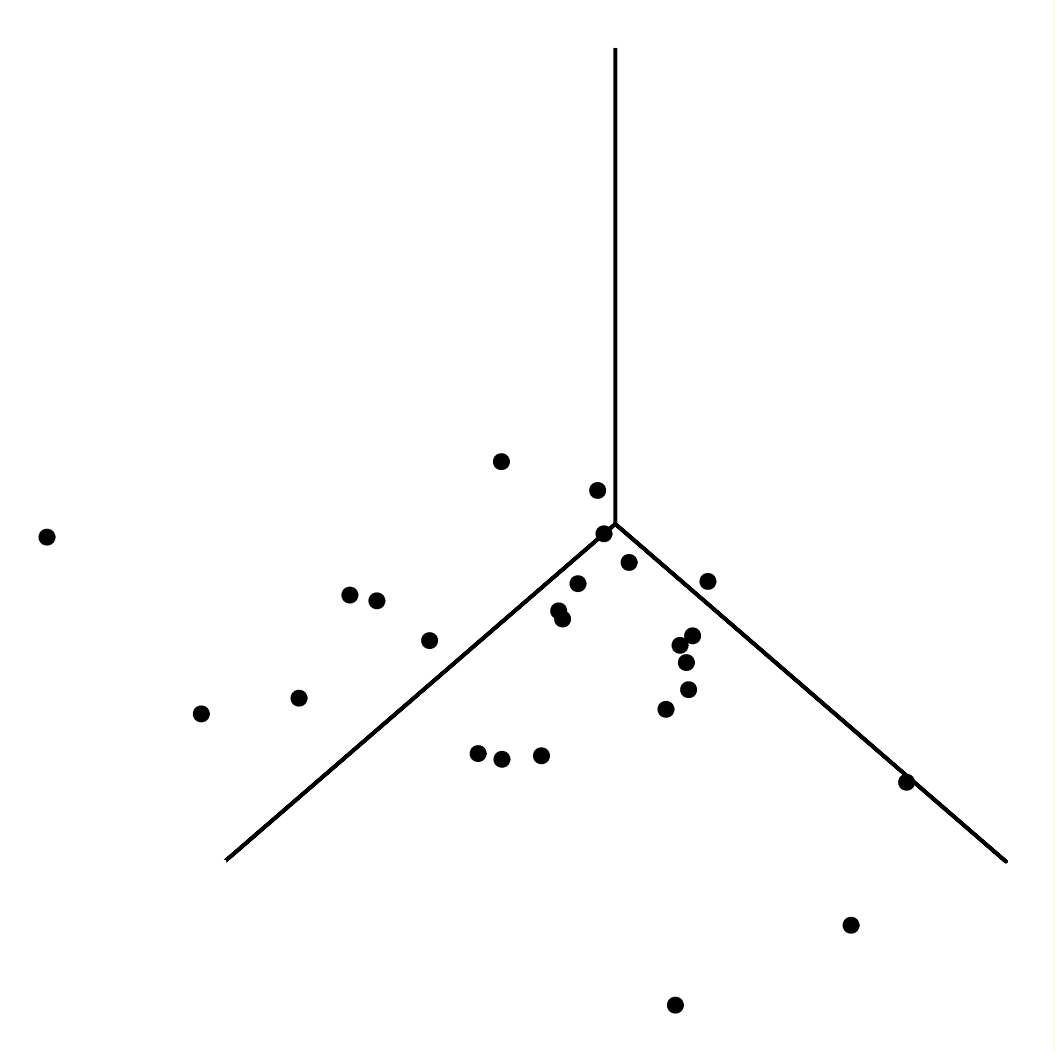
MLR Diagnostics: outliers

Stat 230

April 27 2022

Overview



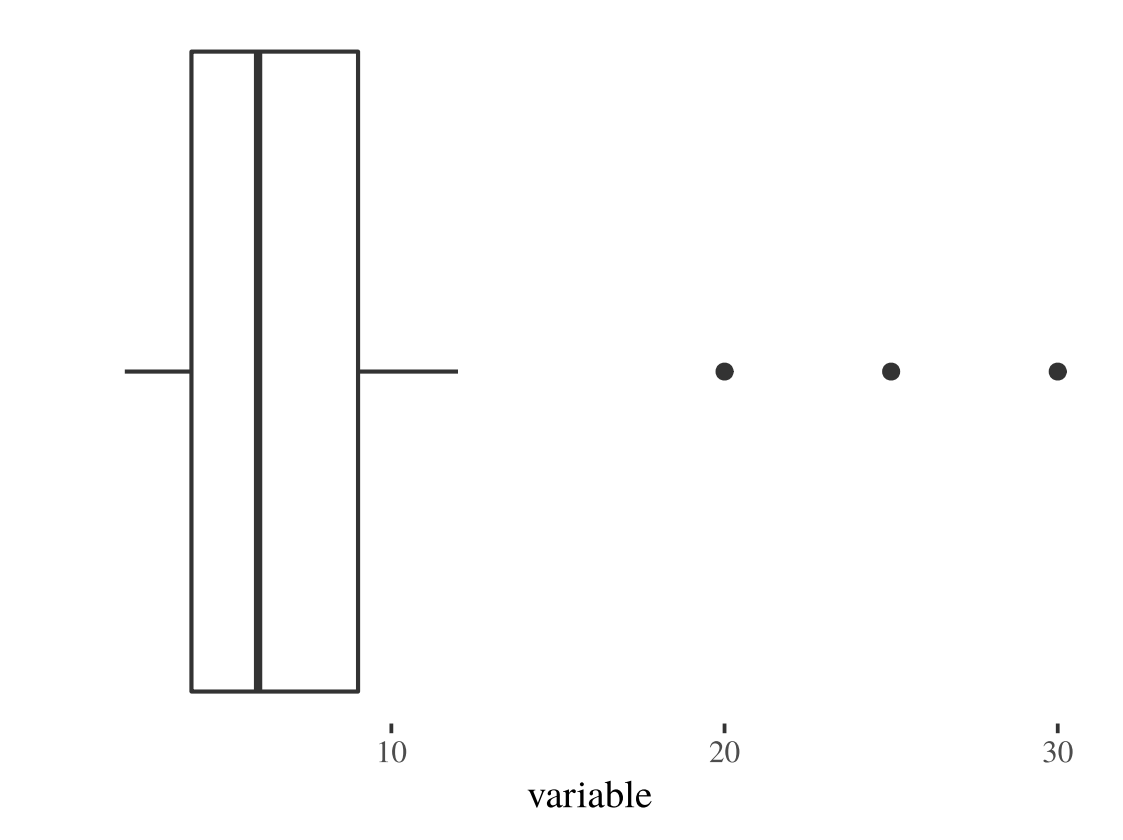
Today:

Assessing "outliers" in regression

- leverage
- standardized residuals
- Cook's distance

Outliers

For one variable, outliers are often found by measuring how different value is from a mean or median value.



Outliers in Regression

For two or more variables, outliers can be cases that have

- an unusual y value
- an unusual x_1 value
- an unusual combination of $\left(x_1, y\right)$ values
- an unusual combination of $\left(x_1, x_2\right)$ values
- an unusual combination of $\left(x_1, x_2, x_3\right)$ values
- and so on ...

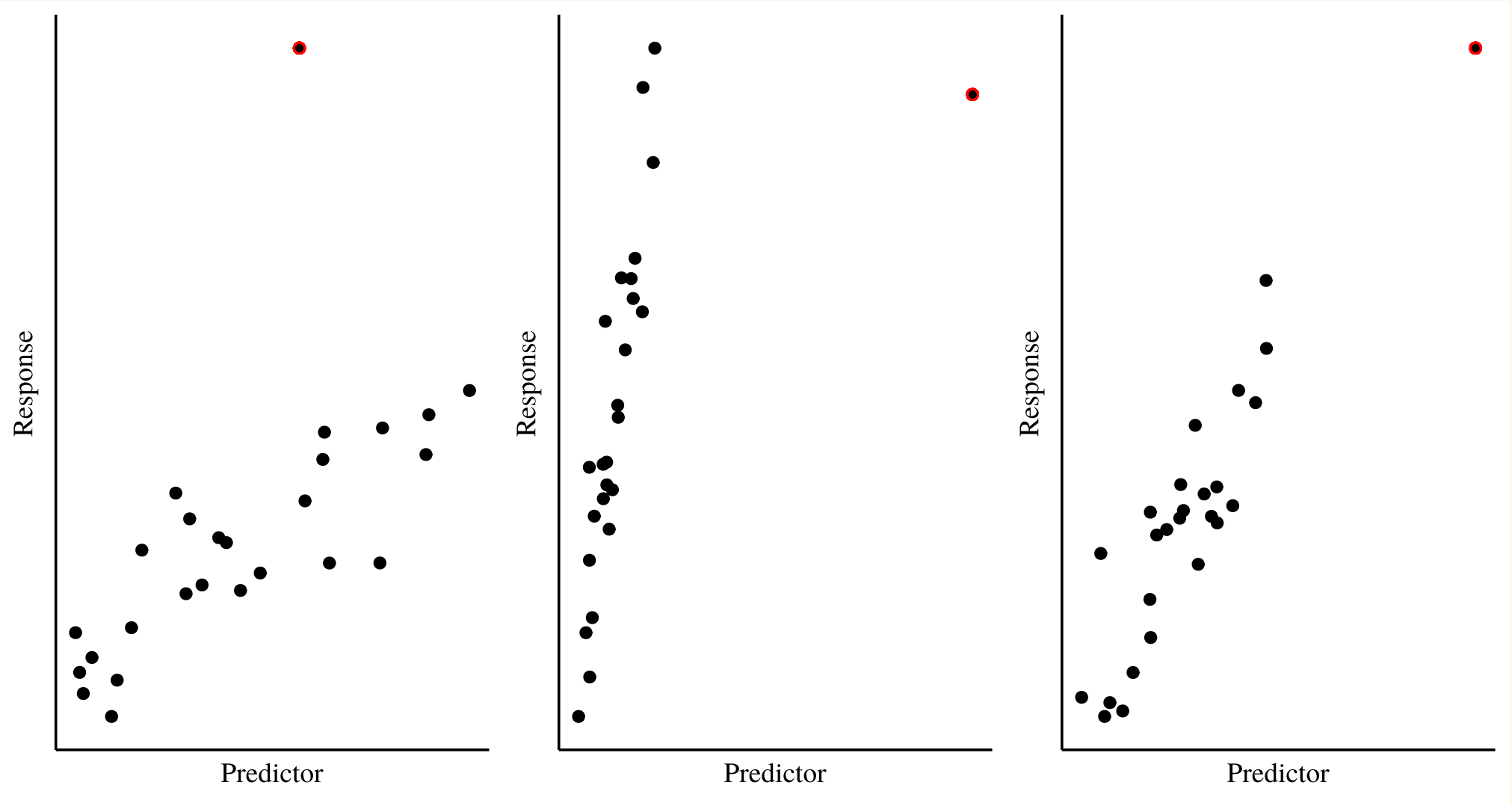
Leverage

- Measures how unusual a case's predictor values are compared to average predictor values for all cases.
  - does not depend on the response
- **SLR**: the leverage of case i equals

$$h_i = \frac{1}{n-1}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

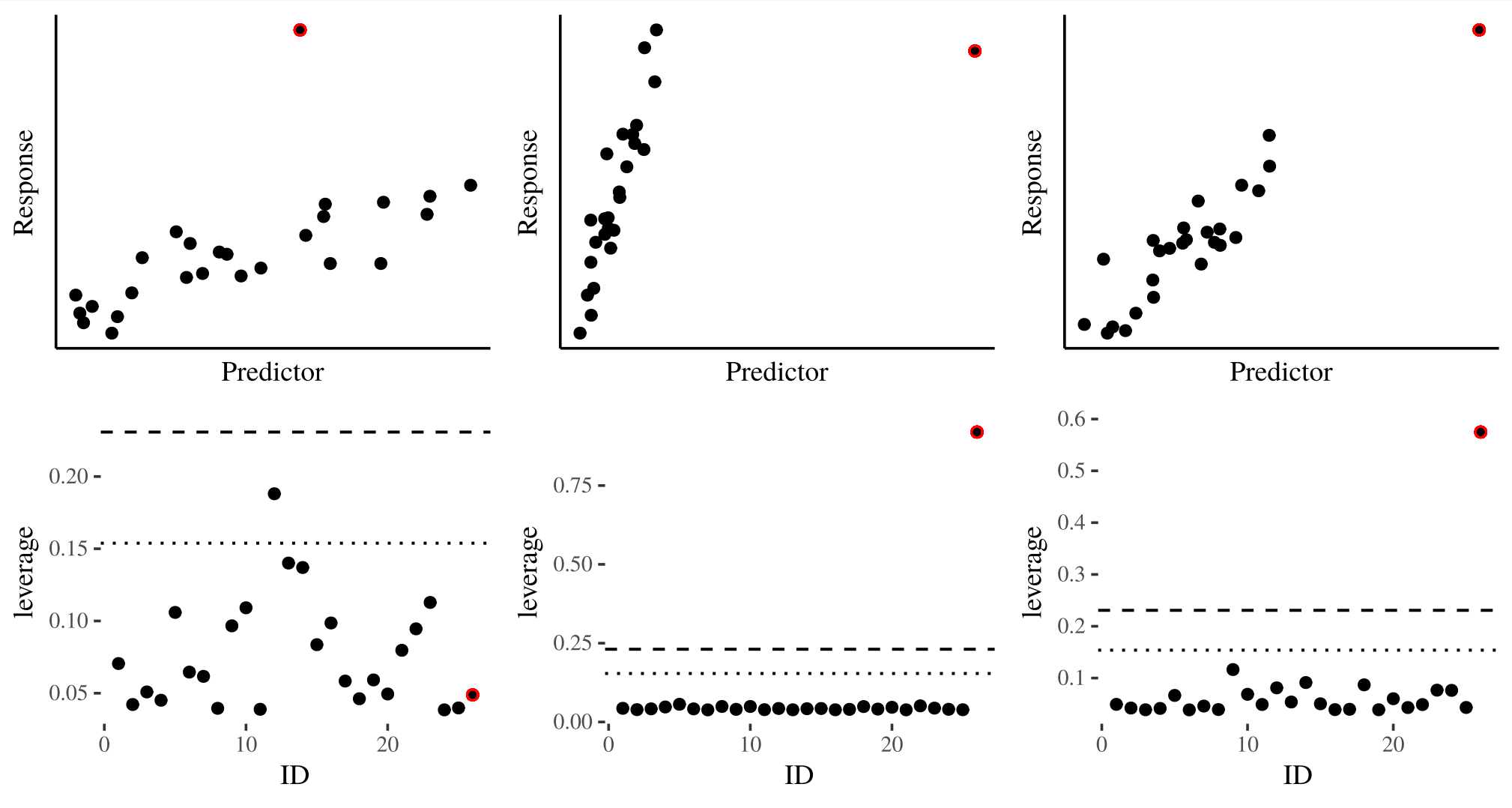
- **MLR**: h_i measures the distance of case i's predictors from the center of the predictor "point cloud"

Leverage



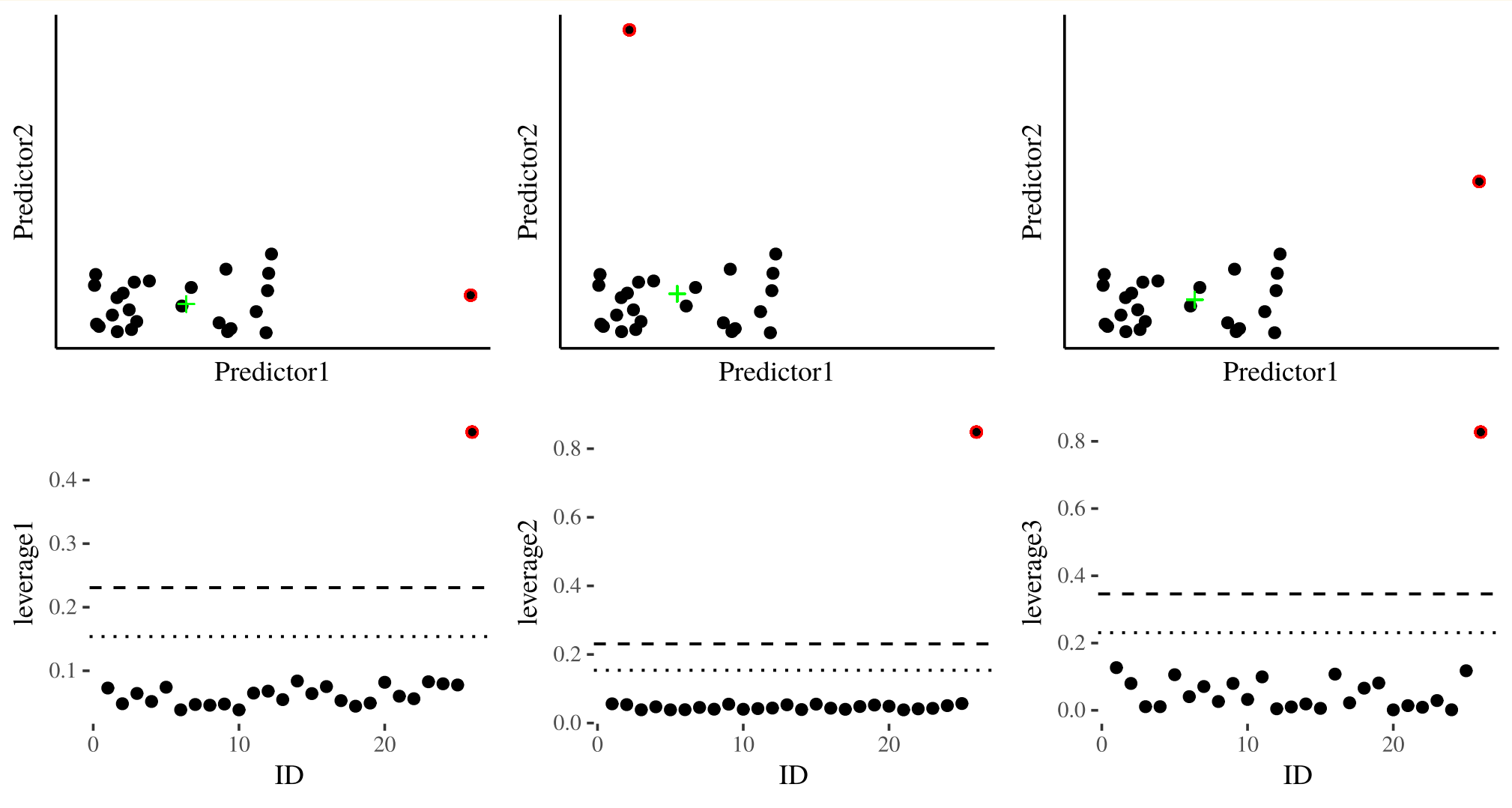
Which red case below will have high leverage?

Leverage



Leverage is high for the red case in the middle and right data sets

Leverage



In MLR, we look at the distance between a case's predictor values (red dot) and the point cloud center (green pluses)

Leverage

Mathematically, $\circ\ \frac{1}{n} \leq h_i < 1$

- for all cases, mean leverage is always $\bar{h} = (p+1)/n$

Guidelines for "large" leverage

- $h_i > 2(p+1)/n$: potential for some influence
- $h_i > 3(p+1)/n$: potential for large influence

Leverage

- A case with high leverage has the potential to be influential
- **influential**: regression line/surface is pulled towards the case

Why?

- The SE of each observed residual $r_i = y_i - \hat{y}_i$ is inversely related to leverage:

$$SE\left(r_i\right) = \hat{\sigma}\sqrt{1 - h_i}$$

Leverage

$$SE\left(r_i\right) = \hat{\sigma}\sqrt{1 - h_i}$$

Large $h_i \approx 1$: residual has little variability and $\hat{y}_i$ will be very close to Y_i

- If Y_i for case i is not following the "trend" of the other $n - 1$ cases, then this case will pull the regression line/surface towards it.

Small $h_i \approx 1/n$: residual has variability of about σ and $\hat{y}_i$ that can vary widely from Y_i.

Standardized residual

- The ratio of a residual to its SE :

$$studr_i = \frac{r_i}{SE\left(r_i\right)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

- also known as internally *studentized* residuals

Standardized residual

$$studr_i = \frac{r_i}{SE\left(r_i\right)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

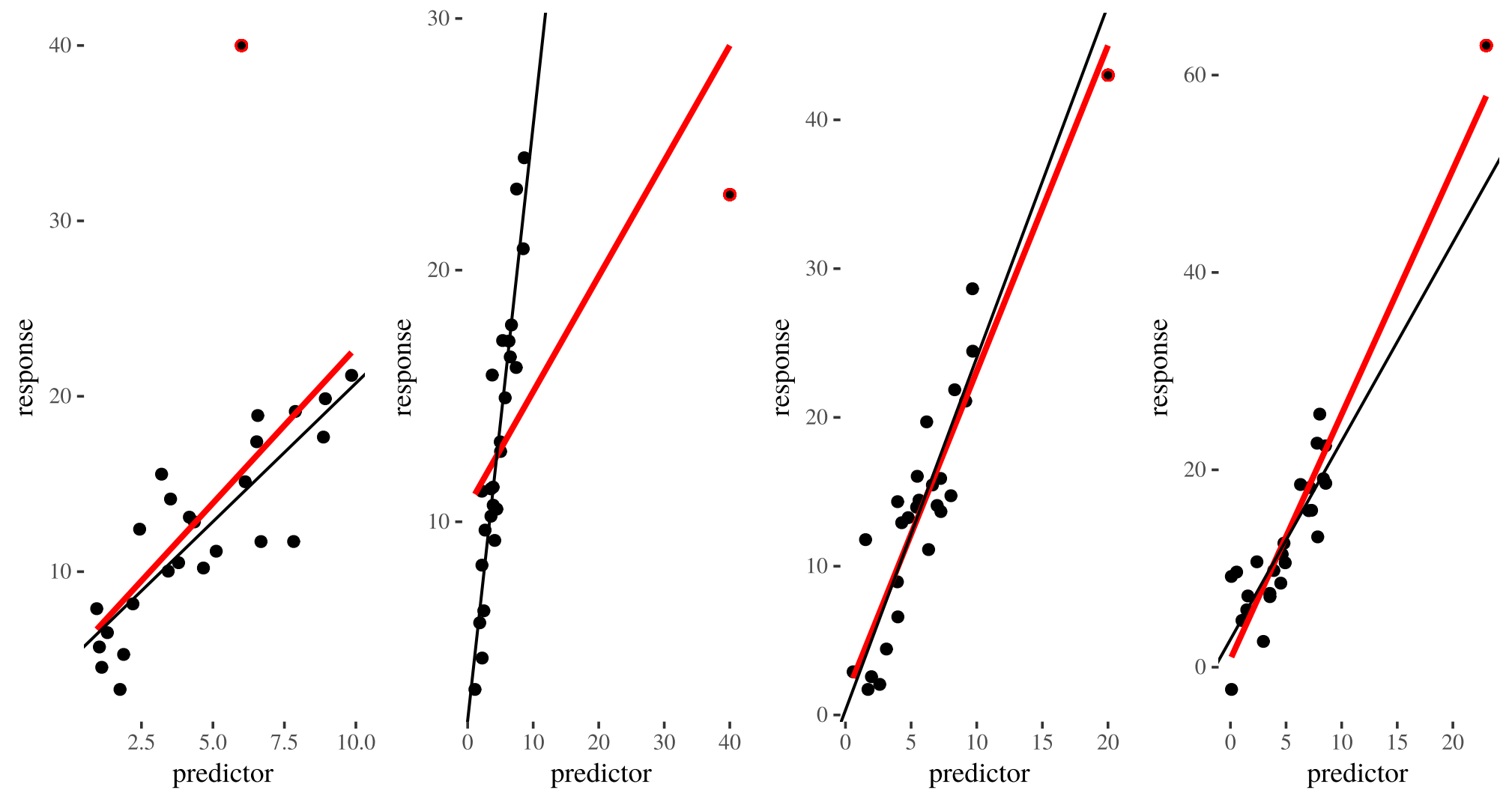
- The reasons for a case with a large standardized residual are
- a large r_i value (just poorly predicted) but "regular" leverage value. These cases will have typical predictor values but an unusual response.
- a "regular" r_i value but a small SE due to a large leverage value.

Guidelines to flag unusual cases are

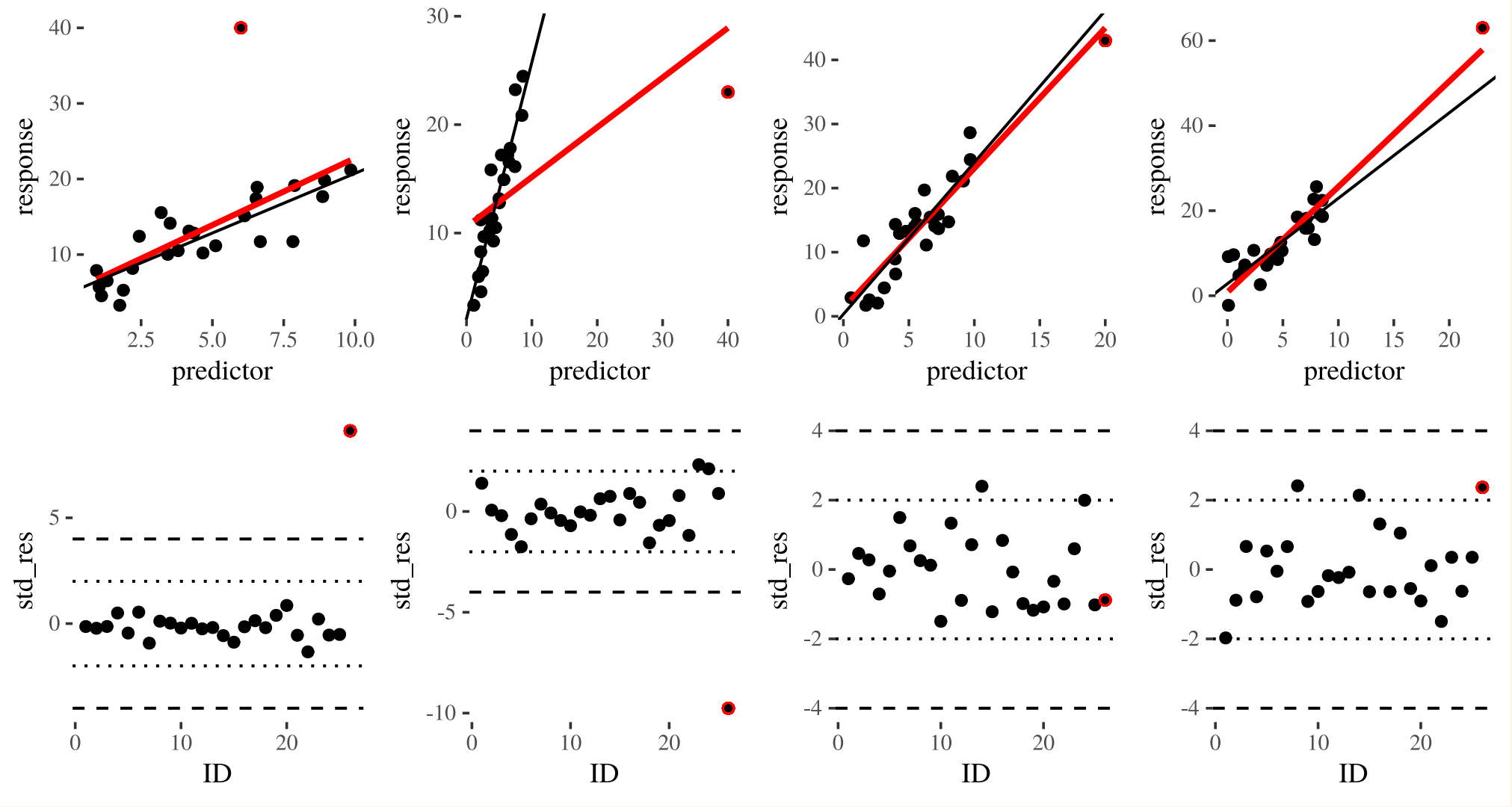
- $|studr_i| > 2$ for smaller data sets
- $|studr_i| > 4$ for larger data sets

Standardized residual

Which red case below will have large standardized residual?



Standardized residual



Cook's distance

Cook's distance: a measure of a case's influence on the fitted regression line/surface.

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

- $\hat{Y}_{j(-i)}$ = predicted response for case j using a model fit that excludes case i.
- $\hat{Y}_j$ = predicted response for case j using a model fit that uses all cases.
- If these two predictions deviate a lot, overall all cases, then we would say that case i is influential in the regression fit.
- $\frac{h_i}{1-h_i}$ can be shown to be the distance from x_i to the centroid of remaining data

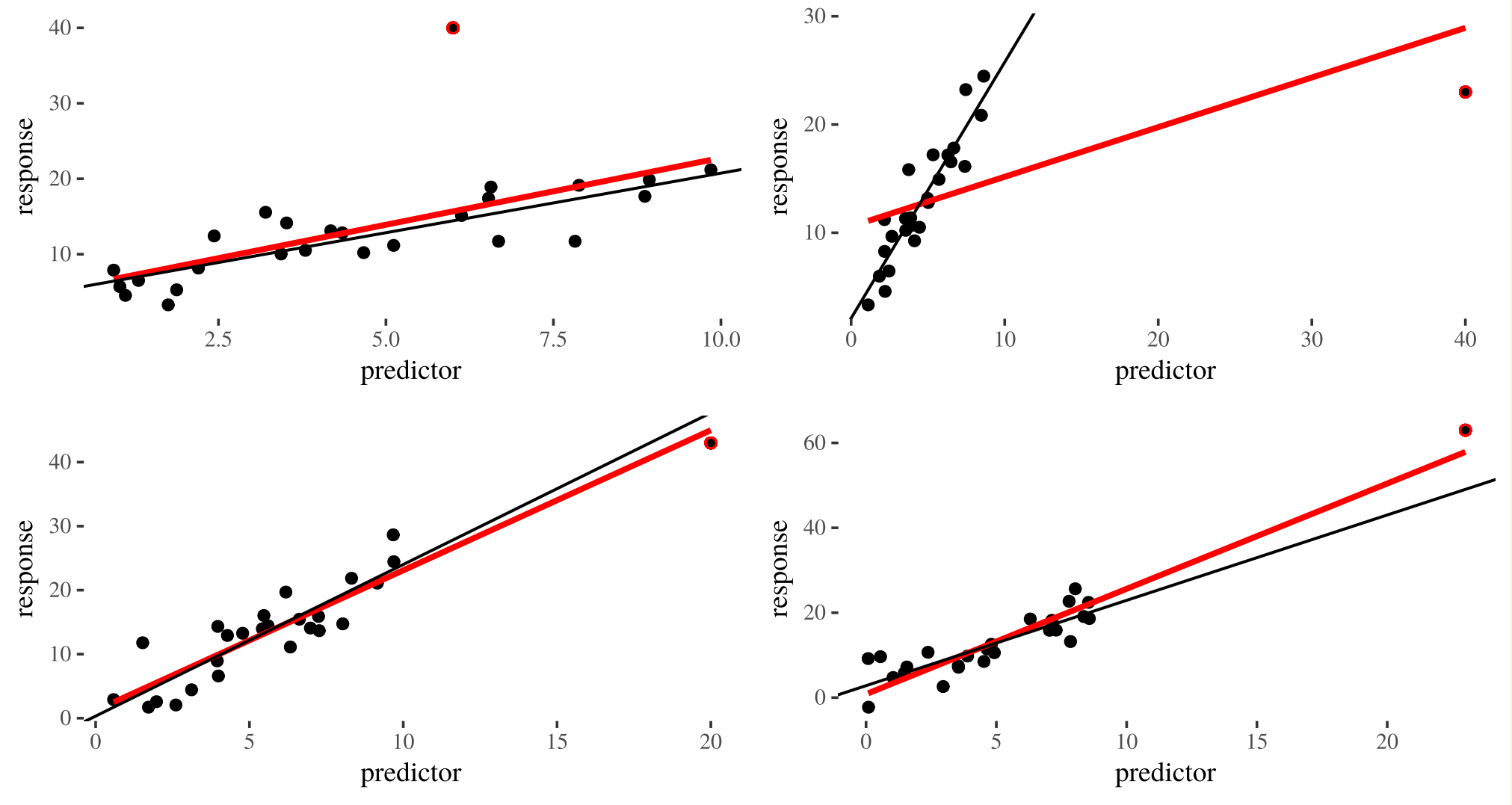
Cook's distance

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

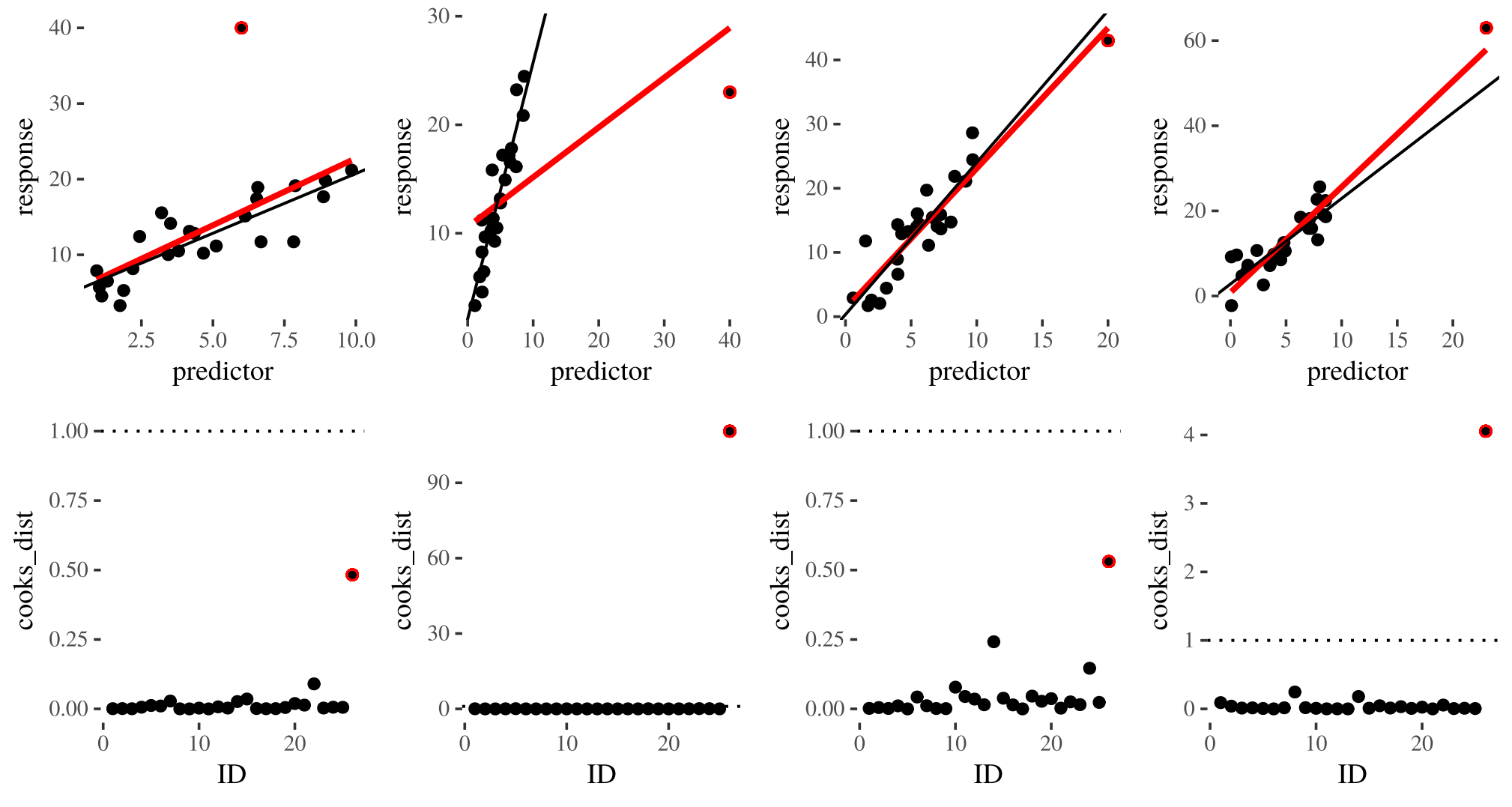
- Cases have large Cook's distance if it has
- high standardized residual,
- high leverage,
- or a combination of both.

Cook's distance

Largest Cook's distance case?



Cook's distance



Cook's distance

Y outlier (1st from left): This red case has a very large residual value r_i but it doesn't have high leverage. It is not very influential.

X outlier (2nd): This red case has a large negative standardized residual and large leverage, so it is very influential $\left(D_i \approx 100\right)$.

X and Y outlier (3rd and 4th):

- The left case doesn't have an unusual standardized residual value though it has large leverage, so it isn't very influential because it follows the trend of the other cases.
- The right case has the same leverage as the left case, but this time it has a larger standardized residual than the left case and it is deemed to be influential $\left(D_i > 3\right)\big)$.

Cook's distance

Guidelines for possible influence:

- Flag cases whose D_i is really unusual (sticks out from the rest)
- Flag cases whose $D_i > 1$ for large data sets or $D_i > qf(0.5, p + 1, n - p - 1)$ for smaller data sets
- for really big data sets: will be hard for one case to be overly influential
- `ggResidpanel` uses another common cutoff: 4/n

Outlier strategy

(1) After EDA, fit potential model

(2) Check residual plots, make transformations if needed, repeat 1-2

(3) Check case-influence statistics. If an "outlier" is found, fit model with and without the case

- Keep the case if conclusions don't change

If conclusions change:

- Omit case if part of a different population
- if not in a different population, omit case if it has high leverage and report reduced predictor range
- else, may need to learn more about the data collection to understand what makes the case different

Case influence disgnostics in R

Base-R options

- Cook's distance vs. row number: `plot(my_lm, which = 4)`
- Standardized residuals (y-axis) against leverage (x-axis) with contours given by Cook's distance: `plot (my_lm, which=5)`
- add `id.n` to get row numbers for the n most "extreme" cases

Case influence disgnostics in R

ggResidpanel package `resid_panel(my_lm, plots =c("cookd", "lev"))`:

- `cookd`: Cook's distance vs. row number
- `lev`: Standardized residuals (y-axis) against leverage (x-axis) with contours given by Cook's distance

`resid_interact(my_lm, plots = c("cookd", "lev"))`:

- clickable plots

Case influence disgnostics in R

`ggplot2` + `broom` option

```
# is not possible using moderndive yet!
data_aug <- augment(my_lm)  # variables used in the lm
# all variables in the lm data set
regression_points <- augment(data = mydata, my_lm)
```

What is added?

- `hat`: leverage
- `cooksd`: Cook's distance
- `.std.resid`: standardized residuals

Case influence disgnostics in R

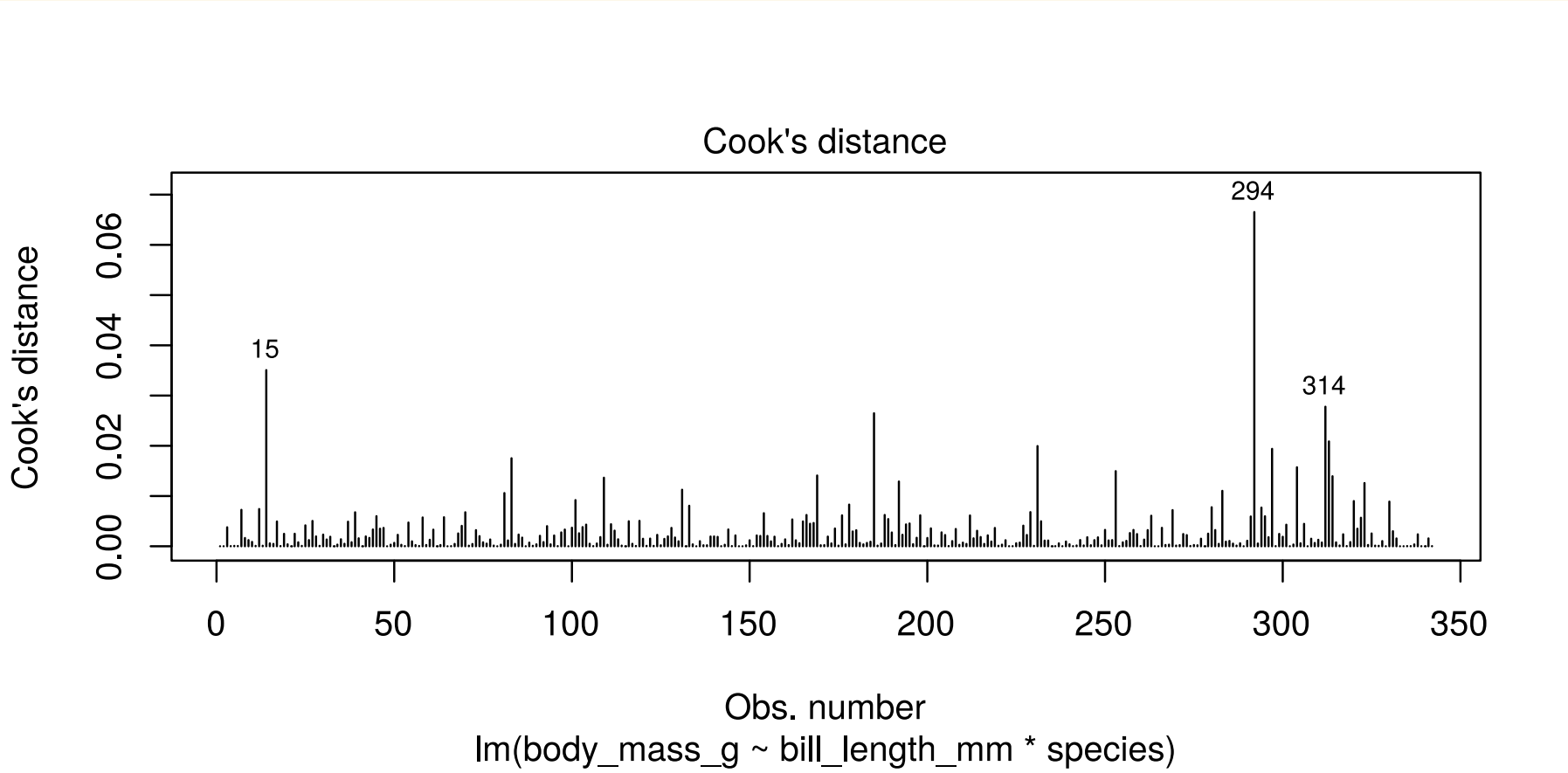
GGally Case influence stats vs. predictors:

```
ggnostic (my_lm, columnsY =c(".std.resid", ".hat", ".cooksd"))
```

- `.std.resid`: standardized residual
- `.hat`: leverage
- `.cooksd`: Cook's distance

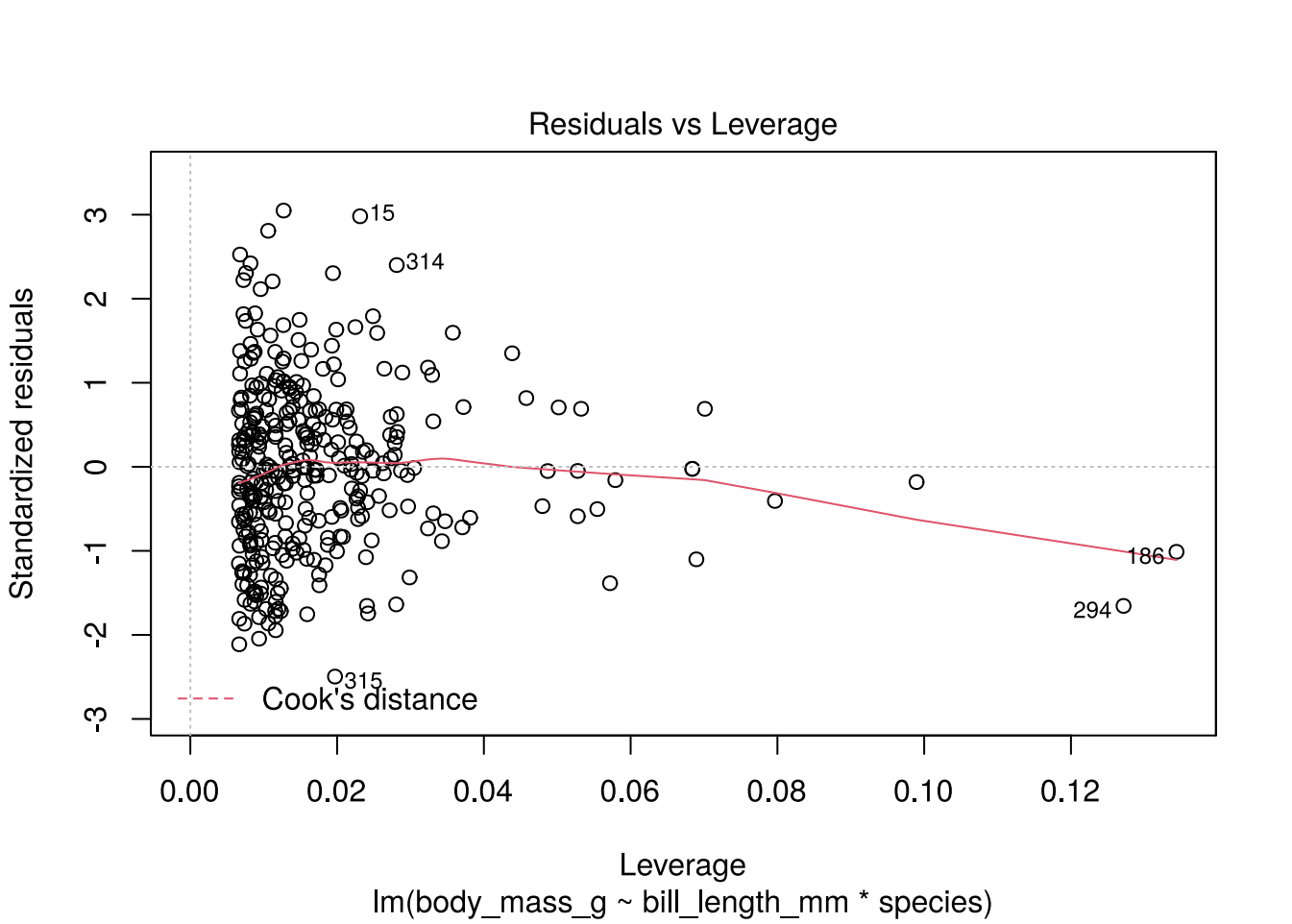
Penguins example

```
peng_interaction_lm <- lm(body_mass_g ~ bill_length_mm * species, data= penguins)
plot(peng_interaction_lm, which = 4)
```



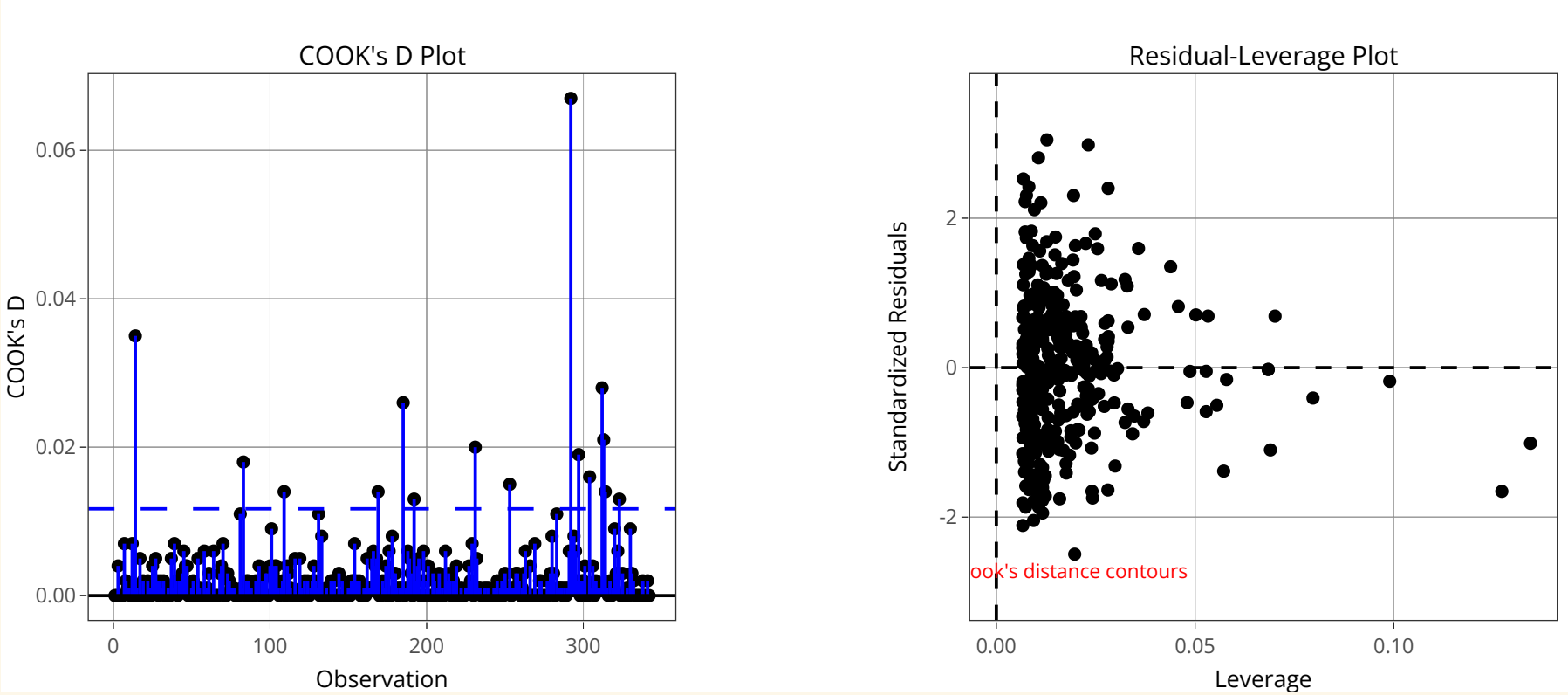
Penguins example

```
plot(peng_interaction_lm, which = 5, id.n = 5)
```



Interactive plots

```
library(ggResidpanel)
resid_interact(peng_interaction_lm, plots = c("cookd", "lev"))
```



Penguins example

```
regression_points <- augment(peng_interaction_lm)
regression_points <- regression_points %>% mutate(ID = row_number())
```

ID 292 has the largest Cook's distance and 2nd largest leverage

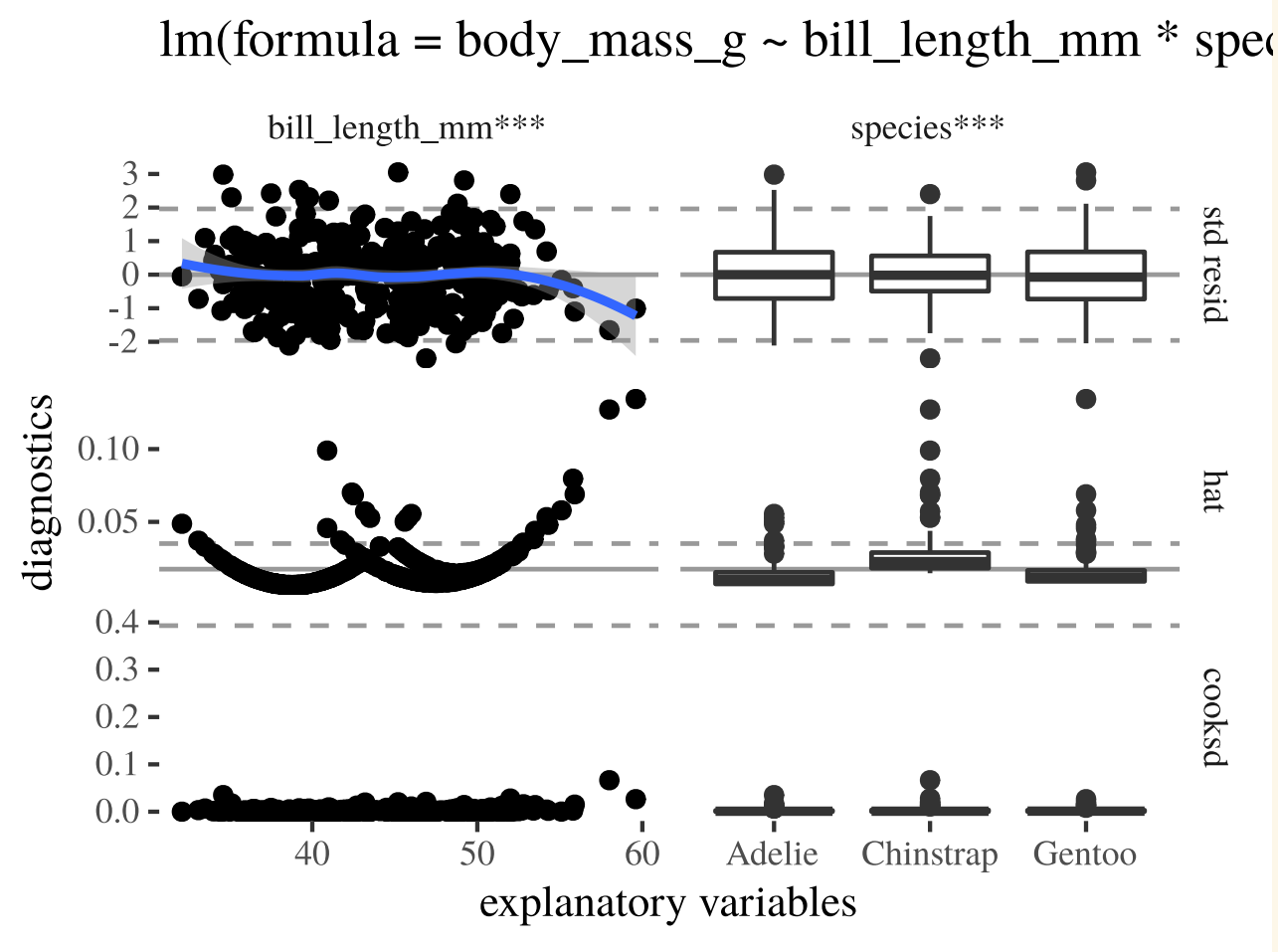
```
regression_points %>%
  slice_max(.cooksd, n = 2) %>% select(-1, -5, -8)
# A tibble: 2 × 8
  body_mass_g bill_length_mm species   .resid   .hat .cooksd .std.resid    ID
        <int>          <dbl> <fct>      <dbl>  <dbl>   <dbl>      <dbl> <int>
1        3700           58   Chinstrap  -575. 0.127   0.0665      -1.66   292
2        4400           34.6 Adelie     1095. 0.0231  0.0351       2.98    14
```

```
regression_points %>%
  slice_max(.hat, n = 2) %>% select(-1, -5, -8)
# A tibble: 2 × 8
  body_mass_g bill_length_mm species   .resid  .hat .cooksd .std.resid    ID
        <int>          <dbl> <fct>      <dbl> <dbl>   <dbl>      <dbl> <int>
1        6050           59.6 Gentoo     -350. 0.134  0.0265      -1.01   185
2        3700           58   Chinstrap  -575. 0.127  0.0665      -1.66   292
```

Penguins example

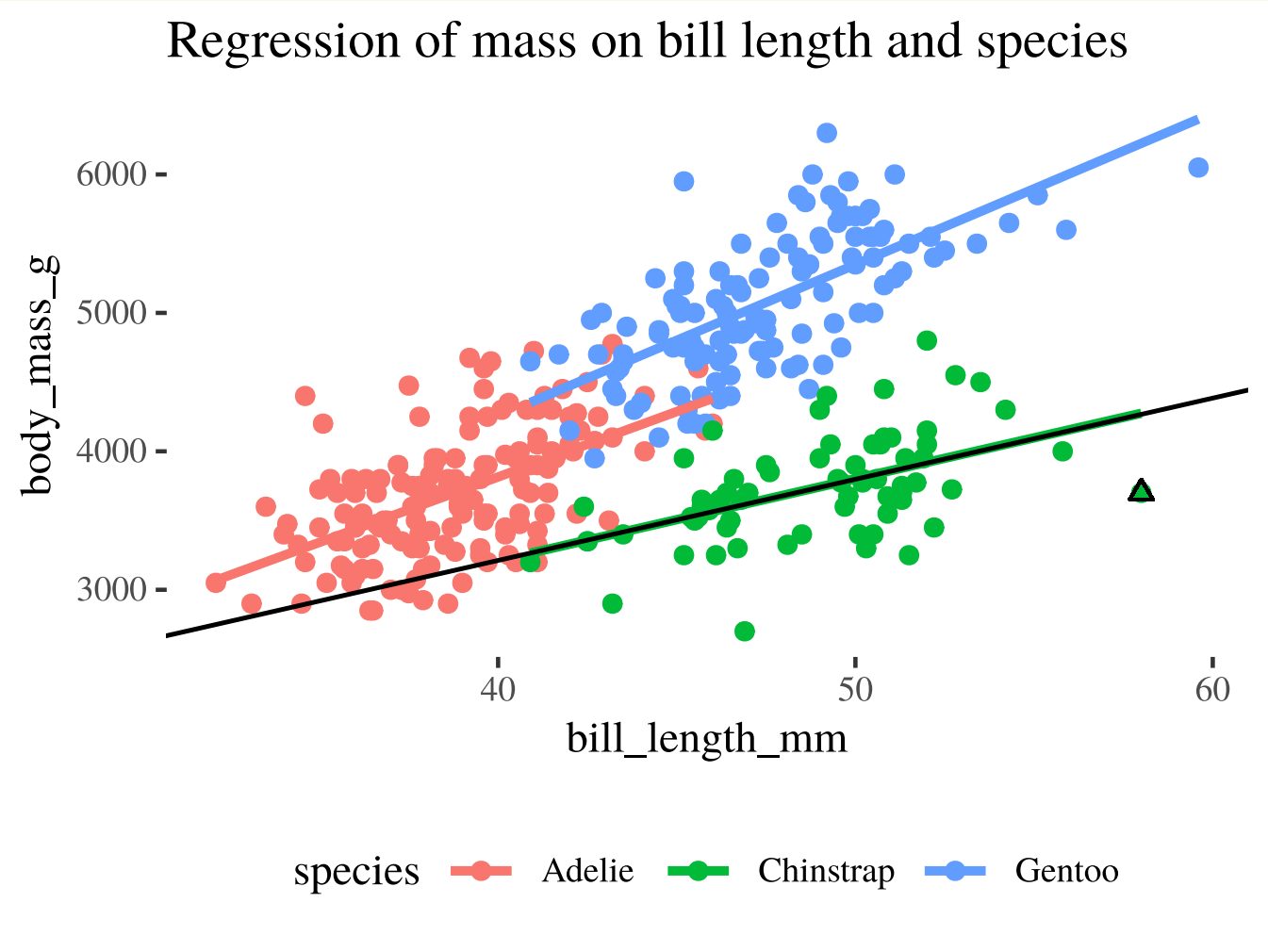
```
library(GGally)
ggnostic(peng_interaction_lm,
         columnsY = c(".std.resid",
                      ".hat",
                      ".cooksd"))
```

- case 292 (largest Cook's D) has the second largest bill length
- case 185 has larger leverage (largest bill length) but smaller std. residual



`Penguins` example

- Case 292 does not have a concerning Cook's distance, but what if we want to check by fitting a model that excludes this case?



Penguins example

```
# remove 292 and refit the model
peng_interaction_lm_no292 <- update(peng_interaction_lm, subset = -292)
```

```
get_regression_table(peng_interaction_lm_no292)
```

```
# A tibble: 6 × 7
  term                    estimate std_error statistic p_value lower_ci upper_ci
  <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                   34.9     444.      0.079   0.937   -838.     907.
2 bill_length_mm              94.5      11.4     8.28    0         72.1    117.
3 species: Chinstrap         832.      801.      1.04    0.3     -744.    2409.
4 species: Gentoo           -159.      684.     -0.232   0.817  -1504.    1186.
5 bill_length_mm:species…    -35.9      17.8    -2.02    0.044    -70.9    -0.897
6 bill_length_mm:species…     15.0      15.8     0.947   0.344    -16.1    46.0
```

Penguins example

- Removing case 292 makes the effect of bill length on mass larger for Chinstrap
- closer to the slope of Adelie!

```
get_regression_table(peng_interaction_lm)
```

```
# A tibble: 6 × 7
  term                  estimate std_error statistic p_value lower_ci upper_ci
  <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                34.9      443.      0.079   0.937   -837.     907.
2 bill_length_mm           94.5       11.4     8.29    0         72.1    117.
3 species: Chinstrap      811.       800.      1.01    0.311   -762.    2385.
4 species: Gentoo        -159.       683.     -0.232   0.816  -1503.    1185.
5 bill_length_mm:species…  -35.4      17.7     -1.99    0.047    -70.3    -0.474
6 bill_length_mm:species…   15.0      15.8      0.948   0.344    -16.1    46.0
```

Penguins example

- Removing this one case makes the interaction between species and bill length significant ($F = 4.28$, $df = 335$, $p = 0.01449$) !
- This case does seem to have some influence on the model.

```
anova(peng_interaction_lm_no292)
```

```
Analysis of Variance Table

Response: body_mass_g
                     Df   Sum Sq  Mean Sq F value  Pr(>F)
bill_length_mm        1 78178002 78178002 564.451 < 2e-16 ***
species               2 93520403 46760201 337.612 < 2e-16 ***
bill_length_mm:species   2  1187789   593894   4.288 0.01449 *
Residuals           335 46398407   138503
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```