Stat 412/512

Case Influence Statistics

Feb 2 2015

Charlotte Wickham

stat512.cwick.co.nz

Regression in your field

See website.

You may complete this assignment in pairs.

Find a journal article in your field of study (or a field of interest) that uses multiple linear regression to answer their question of interest.

Write a one page report that includes:

- a citation to the article
- a brief summary of the background of the problem and their question of interest
- a full specification of their regression model (i.e. full β form) defining all the variables included
- a statement about how their question of interest translates into questions about parameters in their regression model
- a summary of the results they report

Due Feb 27th

Case 11.01 Alcohol Metabolism

Women get drunk quicker than men. Women also develop alcohol related liver disease more readily.

Theory: a particular enzyme responsible for alcohol metabolism in the stomach is more active in men.

"first pass metabolism" = alcohol metabolized in the stomach so it doesn't reach the bloodstream

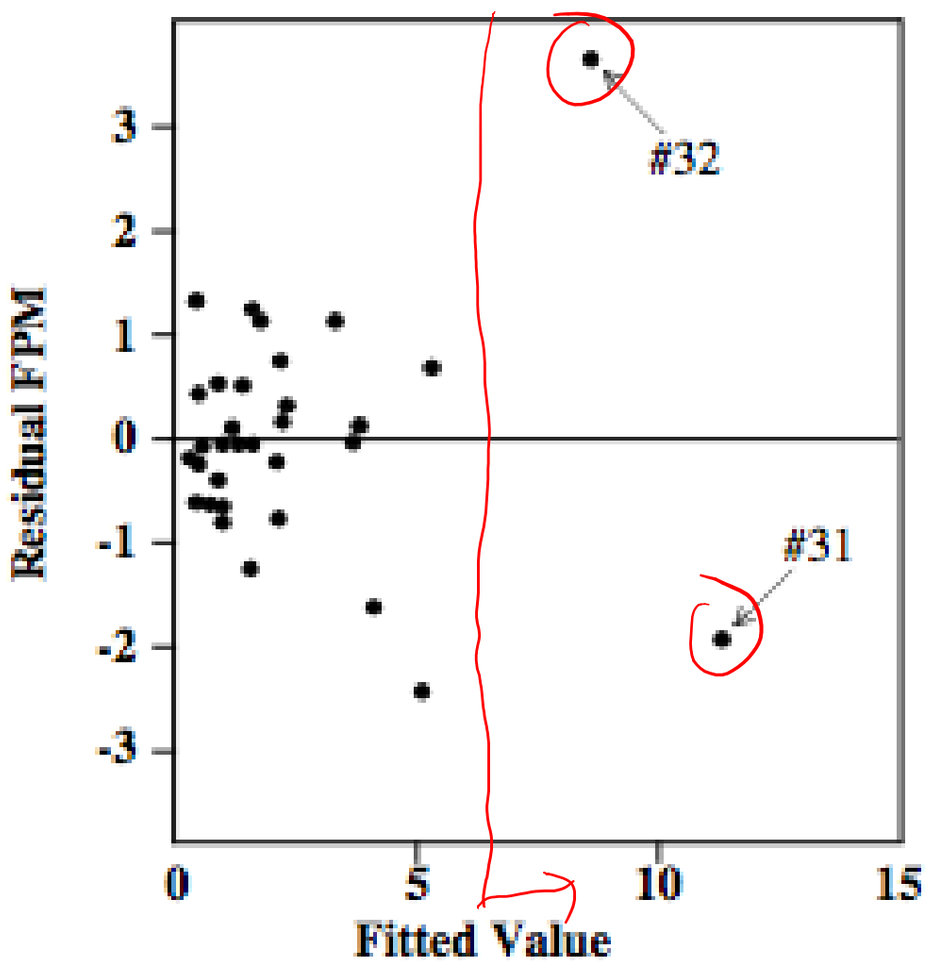
To determine first pass metabolism, compare blood alcohol levels after drinking to after intravenous alcohol.

Also measure enzyme activity.

μ{ *First pass metabolism* | *gast*, female, *alcoholic*} =
gast + female + alcoholic + female x alcoholic +
gast x female + gast x alcoholic +
gast x female x alcoholic

Display 11.7 **p. 313**

Residual plot from the regression of first-pass metabolism on gastric activity, sex indicator, alcoholism indicator, and all 2nd and 3rd-order interactions



subset(,)

Outliers

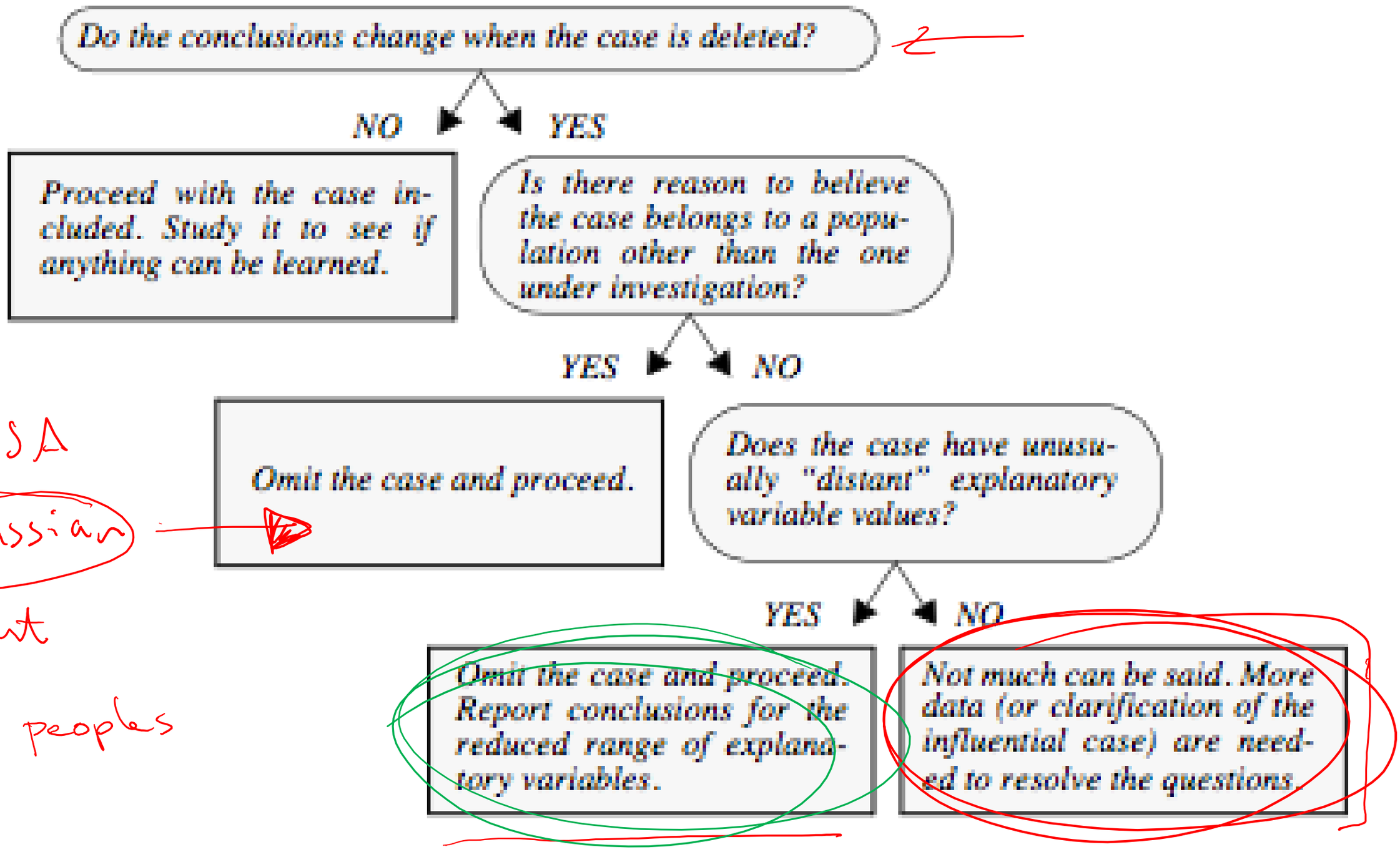
Least squares estimates are not **resistant** to outliers.

Identify outliers early on, so you don't end up tailoring the model to to fit a few unusual observations.

An observation is said to be **influential** if the fitted model depends unduly on its value.

For example, removing it: changes the estimate of parameters greatly, changes conclusions, or changes which terms are included in the model.

A strategy for dealing with suspected influential cases



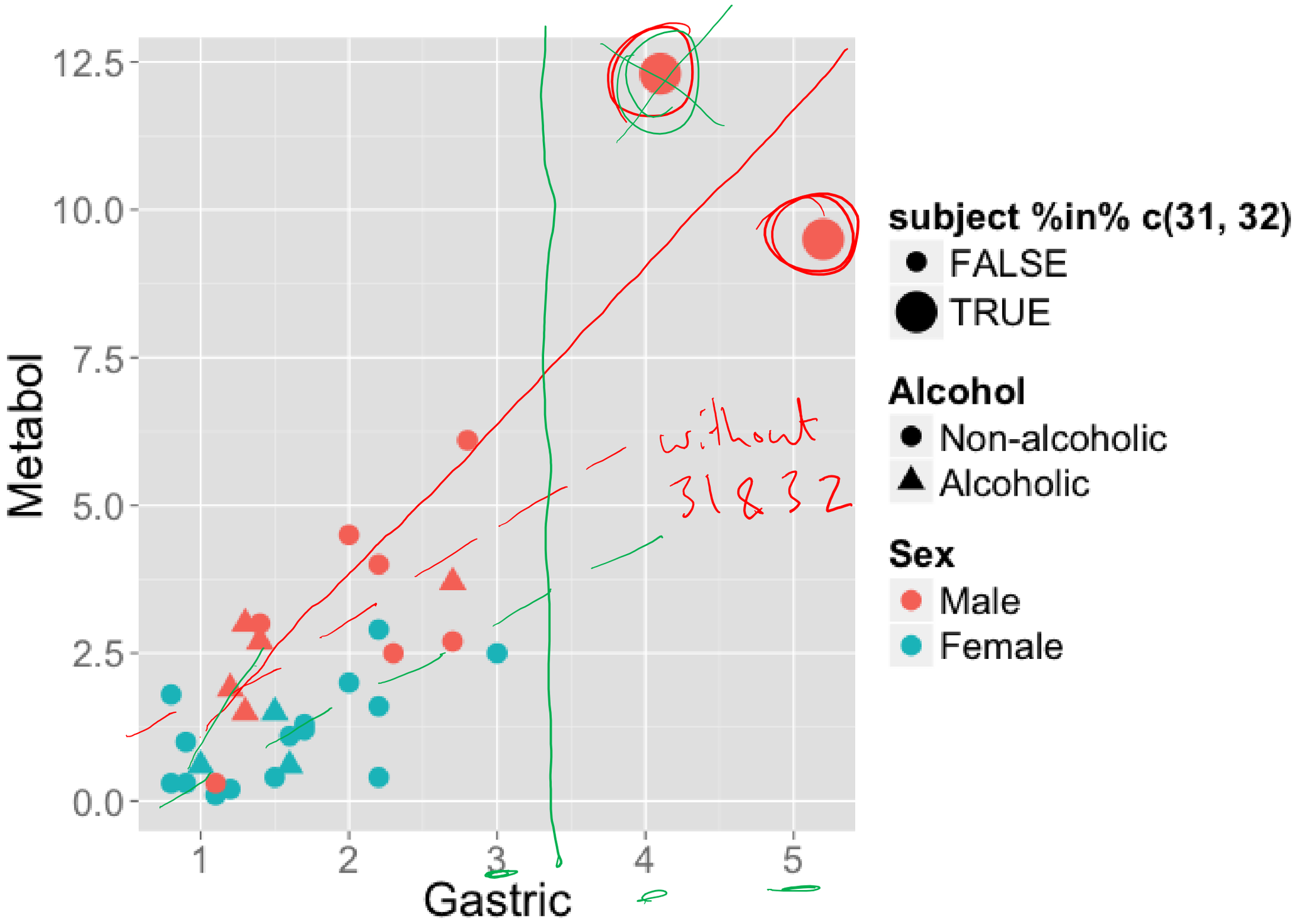
with 31 & 32

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -1.6597     0.9996  -1.660   0.1099
Gastric                             2.5142     0.3434   7.322 1.46e-07 ***
SexFemale                           1.4657     1.3326   1.100   0.2823
AlcoholAlcoholic                    2.5521     1.9460   1.311   0.2021
Gastric:SexFemale                  -1.6734     0.6202  -2.698   0.0126 *
Gastric:AlcoholAlcoholic           -1.4587     1.0529  -1.386   0.1786
SexFemale:AlcoholAlcoholic         -2.2517     4.3937  -0.512   0.6130
Gastric:SexFemale:AlcoholAlcoholic  1.1987     2.9978   0.400   0.6928
```

without 31 & 32

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -0.6797     1.3091  -0.519  0.60878
Gastric                             1.9212     0.6082   3.159  0.00455 **
SexFemale                           0.4858     1.4665   0.331  0.74360
AlcoholAlcoholic                    1.5722     1.8119   0.868  0.39493
Gastric:SexFemale                  -1.0805     0.7211  -1.498  0.14826
Gastric:AlcoholAlcoholic           -0.8658     0.9631  -0.899  0.37839
SexFemale:AlcoholAlcoholic         -1.2718     3.4669  -0.367  0.71725
Gastric:SexFemale:AlcoholAlcoholic  0.6058     2.3158   0.262  0.79608
```

Do our conclusions change?



Whether we have evidence males and females have different slopes depends heavily on if observations 31 and 32 are included.

Safe option, drop 31 & 32 and only make inferences for people with a Gastric AD activity level < 3

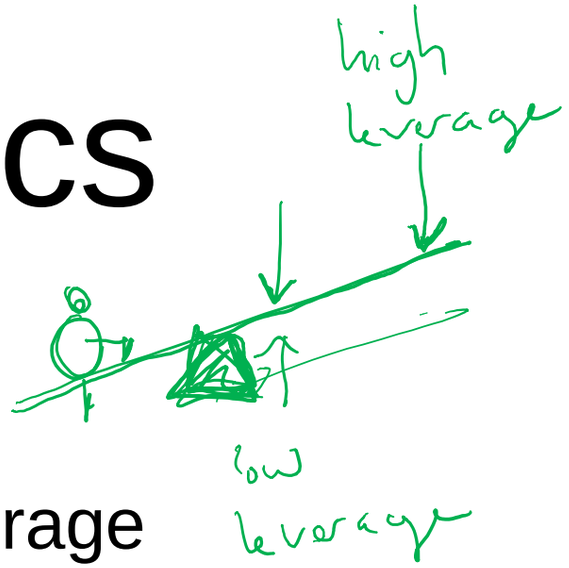
Case influence statistics

Case influence statistics help identify **observations** that may be influential.

Sometimes influential observations won't show up in our usual plots.

Case influence statistics

a number for every observation



Leverage

Measures the distance of the observation from the average explanatory values (taking correlation into account). High leverage = unusual combination of explanatory values = possibility to be influential.

Studentized residuals

The residual divided by its expected variation. High residual = observation far from fitted line.

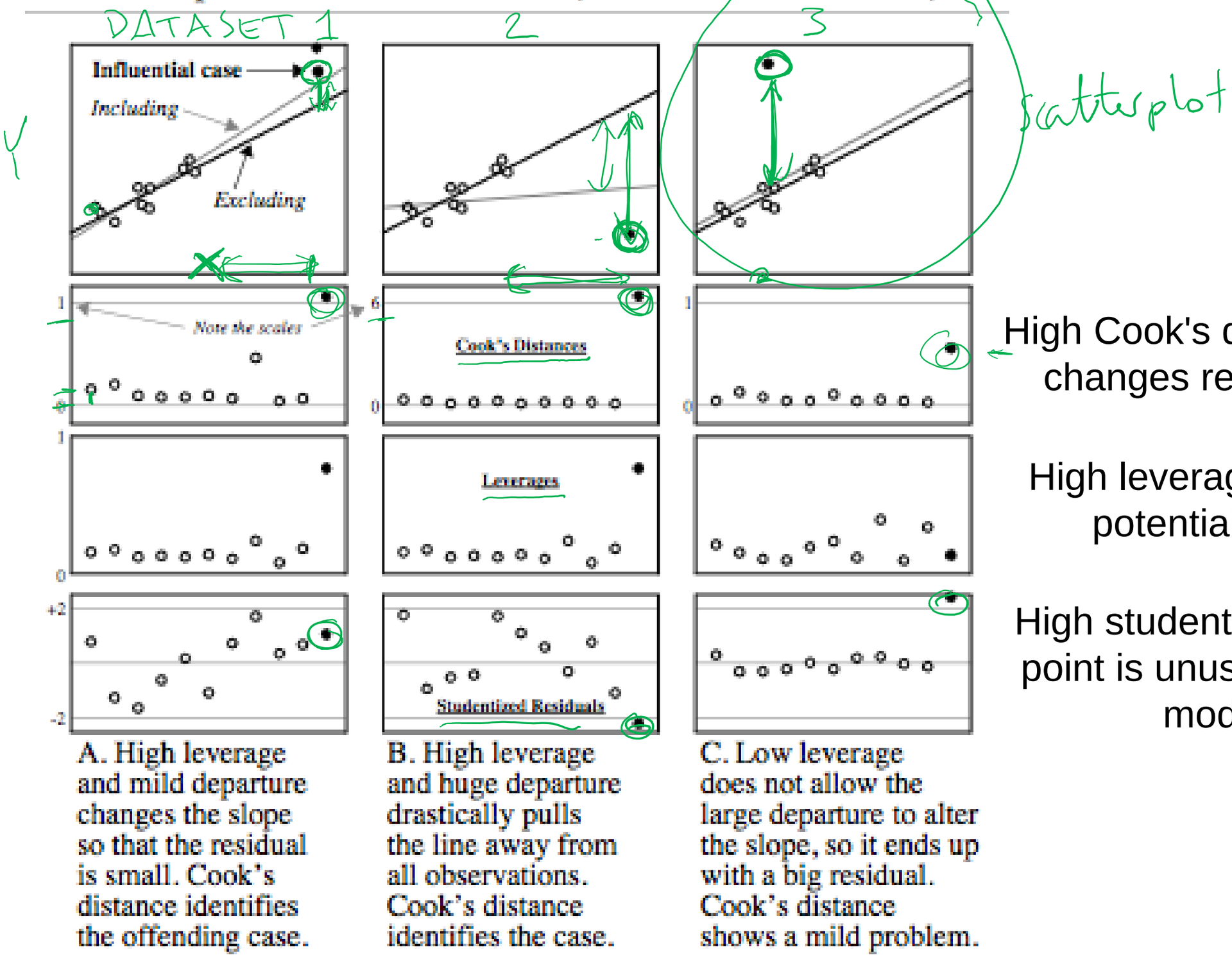
Cook's distance

The effect on estimated parameters when the observation is dropped out. High Cook’s distance = influential on parameter estimates.

you don't need to know the formulas

use them together to understand influential points

Three examples of influential cases in simple linear regression. The top row shows regression lines with and without the influential case included. The next three rows show the resulting case influence statistic plots: Cook's distances, leverages, and Studentized residuals. The horizontal axes for the case statistic plots show the case numbers (=11 for the influential case).



A. High leverage and mild departure changes the slope so that the residual is small. Cook's distance identifies the offending case.

B. High leverage and huge departure drastically pulls the line away from all observations. Cook's distance identifies the case.

C. Low leverage does not allow the large departure to alter the slope, so it ends up with a big residual. Cook's distance shows a mild problem.

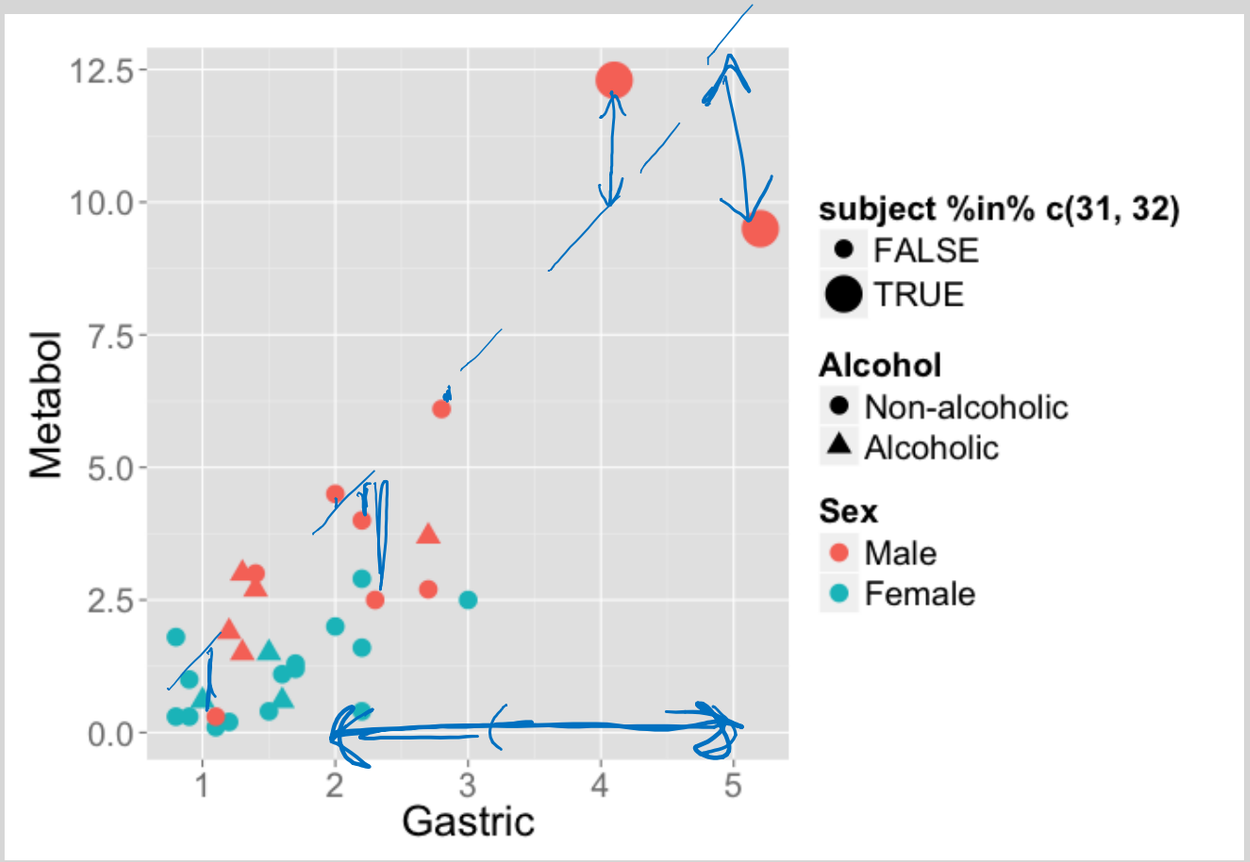
High Cook's distance, means point changes regression estimates

High leverage, means point has potential to be influential

High studentized residual, means point is unusual compared to the modelled mean.

Your turn

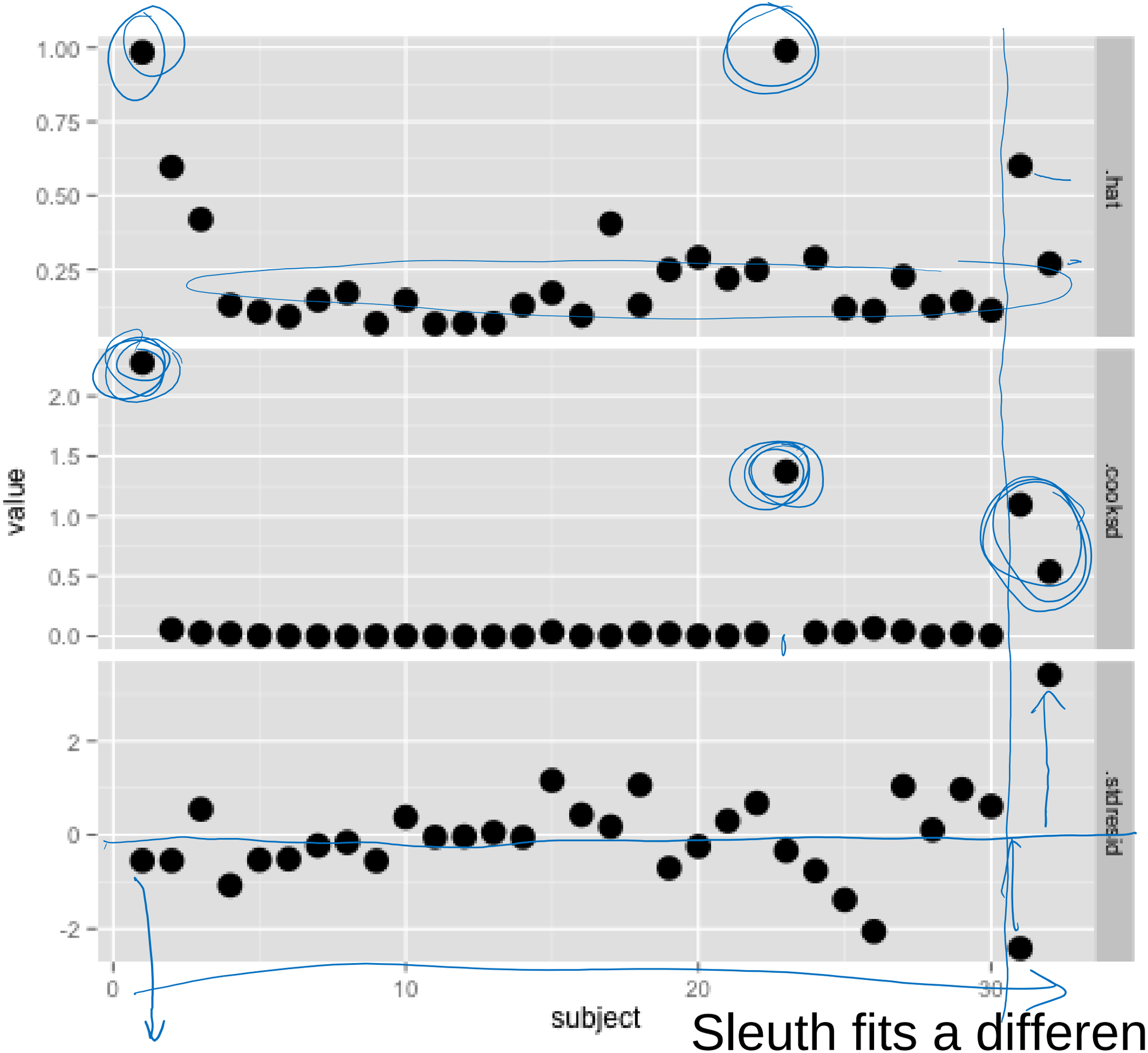
Which would you expect to be large for observations 31 & 32?



leverage large

studentized residual moderate

Cook's distance pretty high



leverage

Cook's distance

studentized residuals

Sleuth fits a different model


12.5
10.0
7.5
5.0
2.5
0.0
Metabol
1
2
3
4
5
Gastric
subject %in% c(1, 23, 31, 32)
FALSE
TRUE
Alcohol
Non-alcoholic
Alcoholic
Sex
Male
Female

Next step: Find a simple, good fitting model.

You can test whether terms are necessary with F-tests.

Generally, if you think something is important, leave it in whether it is significant or not.

Some people say (and I agree): if it was important enough for you to think about including it, leave it in.

Be careful about interpreting individual t-tests, especially if it involves a term that is elsewhere in the model.

Case 11.01 Alcohol Metabolism

Not many alcoholics, and an extra sum of square F-test with reduced model:

μ{ *First pass metabolism* | *gast*, female, *alcoholic*} =

gast + female + gast x female

has a p-value of 0.93.

So, use this reduced model, i.e. no effect of alcoholism.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.06952    0.80195   0.087 0.931580
Gastric            1.56543    0.40739   3.843 0.000704 ***
SexFemale         -0.26679    0.99324  -0.269 0.790352
Gastric:SexFemale -0.72849    0.53937  -1.351 0.188455
```

#1

#2

No evidence of different intercepts, if different slopes are in the model.

#1

No evidence of different slopes, if different intercepts are in the model.

#2

Here:

a multiplicative rather than additive difference makes more sense,

both having the same zero intercept makes sense,

μ{ *First pass metabolism* | *gast*, female, *alcoholic*} =

$\beta_1$gast + $\beta_2$gast x female

```
                   Estimate Std. Error t value Pr(>|t|)
Gastric              1.5989     0.1249  12.800 3.20e-13 ***
Gastric:SexFemale   -0.8732     0.1740  -5.019 2.63e-05 ***
```

"Males had a higher first-pass metabolism than females even after accounting for differences in gastric AD activity (two-sided p-value = 0.0003 for a t-test for equality of male and female slopes when both intercepts are zero.)

For a given level of gastric AD activity the mean first-pass metabolism for men is estimated to be 2.20 times as large as the mean first-pass alcohol metabolism for women."

a different way to interpret a difference in slopes, but only if there are **no intercepts**.