Influential and high-leverage observations, outliers

Influential observations

- Idea: how much does my fit change after taking out this observation?
- There are different ways to measure this
- For example: Cook's distance, DFFITS, etc.

Cook's distance

- Cook's distance of observation *i* is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

$\hat{y}_j$ predicted value for observation j with all the observations

$\hat{y}_{j(i)}$ predicted value for observation j after taking out the i-th observation

s^2 our usual estimator of the residual variance σ^2

- *How big is big?* Different recommendations… Some people say $D_i > 1$
- I recommend looking closely at any observation that seems to "stick out"

Leverage, outliers, and influence

- Leverage: measures how far away x_i is from the other x values [goes from 0 to 1, from "average x" to "very unusual x"]
- High leverage: unusual value of x_i, which may or may not be well predicted by our line
- Big residual $|e_i|$: point that is badly predicted by our line (outliers)
- Observations with high leverage and big residuals are highly influential, because Cook's distance can be written as

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right]$$

h_i: leverage of observation i
e_i: residual of observation i

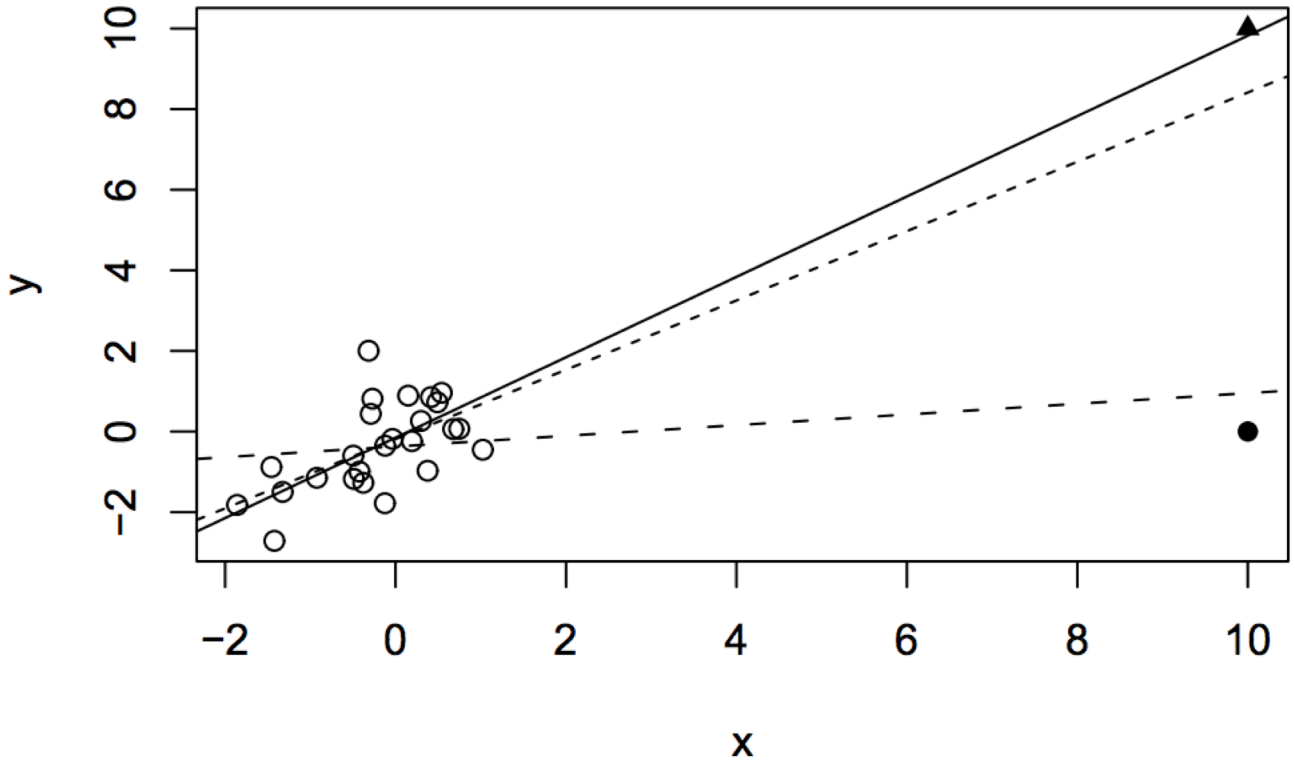

Figure 7.2: Outliers can conceal themselves. The solid line is the fit including the ▲ point but not the • point. The dotted line is the fit without either additional point and the dashed line is the fit with the • point but not the ▲ point.

Source: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Multiple linear regression

Multiple linear regression

- The same, but with more variables
- Find the coefficients that minimize in-sample predictive error
- We can find CIs and hypothesis tests if we make assumptions
- We assume

$$y_i \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Independence of outcomes y_i for i in 1:n (given the x_{ij}).
- Normality
- Homoscedasticity (equal variance across observations, which doesn't depend on x_{ij})
- Linearity (i.e. E[Y | X] is a linear comb. of the Xs)

Model building

General problem: Variable selection

- You have an outcome y and predictors $x_1, x_2, \ldots, x_p$
- Do put all p predictors in the model?
- Some reasons we might not want to include all of them
  - In the application, the client might be interested in knowing which variables seem to be "active" ("predictive")
  - If you don't need some of them, you might be able to get rid of them and get more precise estimates and predictions [*there are some caveats here*]

Two classes of approaches

- All subsets
  - Fit *all possible models* (with all the possible subsets of predictors in and out of the model)
  - *Rank/score* the model according to some criterion
    - Almost infinitely many possibilities, no single criterion is uniformly better than the rest
- Search strategies
  - Look for *good models*, without exploring all the subsets
  - Sometimes you just have to do this because the model space is too big, and you can't go through all subsets…

How to *score* models?

- You go through all subsets, find a "score"… A score like what?
- We have seen R^2

variability in y

Residual sum of squares

variability in predictions

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- It's easy to use: it goes from 0 to 1
- Tempting to use it as a "goodness-of-fit" statistic
- However, it can be highly deceptive when the relationship between y and x isn't linear

$$R^2_{\text{adjusted}} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

- Unfortunately, R^2 can't get worse as you add in more variables [the residual sum of squares can't get worse after adding a variable… Worst case scenario, the coefficient of that variable is set to 0, and we're done]
- Fortunately, somebody found out a way to penalize the so that there isn't a *bias* towards bigger models
- If all predictors are garbage: $E[R^2] = p/(n-1)$
  - BAD! It increases as we put in bogus predictors
  - Adjusted R^2 is modified so that $E[R^2_{adj}] = 0$ if all predictors are bad

BIC and C_p

- BIC: smaller is better
  - Again, it looks at the tradeoff between smaller residual sum of squares (RSS) and the fact that bigger models (tend to) have smaller residual sum of squares
  - So, it has a term that increases in RSS and some penalty on model "complexity" (p * log n)
- C_p: Pick smallest model whose C_p is roughly p
  - Idea: Same tradeoff between small RSS and penalizing big models
  - Can be derived by thinking how E(RSS) should behave if the model is "correct"

Searching for *good* models

- Sometimes you can't go through all models
- Some strategies for finding *good models*
  - Forward selection: start with no variables, and keep on adding variables one at a time until it doesn't pay off (according to some criterion)
  - Backward selection: start with all of the variables, and keep on dropping variables until it doesn't pay off (according to some criterion)
  - Stepwise selection: start with no variables, and keep on adding variables one at a time until it doesn't pay off. If a variable that seemed useful at some previous step isn't useful anymore, you drop it
- You can use p-values as the criterion to include/exclude variables
- You can use other criteria, such as BIC, etc.

Don't compare model scores if you transformed *y*!

Two fitted models, obtained by different transformations of the response, are plotted on the original scale in Figures 1 and 2. Figure 1 is obtained by fitting a model of the form

$$Y_1^* = \alpha + \beta x + \gamma x^2 + e, \qquad (1)$$

where $Y_1^* = Y/x^{3/2}$, by ordinary least squares and then expressing the prediction equation and the prediction interval limits back in the original scale. Figure 2 is obtained in the same way by fitting

$$Y_2^* = \alpha + \beta x + \gamma x^2 + e, \qquad (2)$$

with $Y_2^* = \log_e(Y)$. Note that both linear models contain a constant term.

Source:

Transformations and R^2

Alastair Scott &Chris Wild

Don't compare model scores if you transformed *y*!

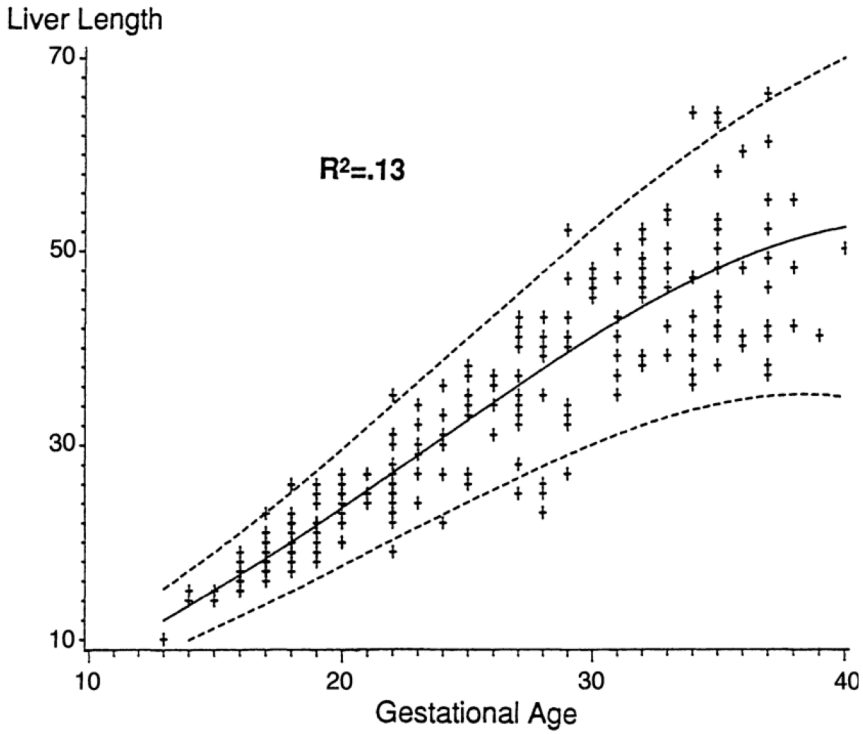

Figure 1. Fitted Model Based on $Y_1^* = Y/x^{3/2}$.

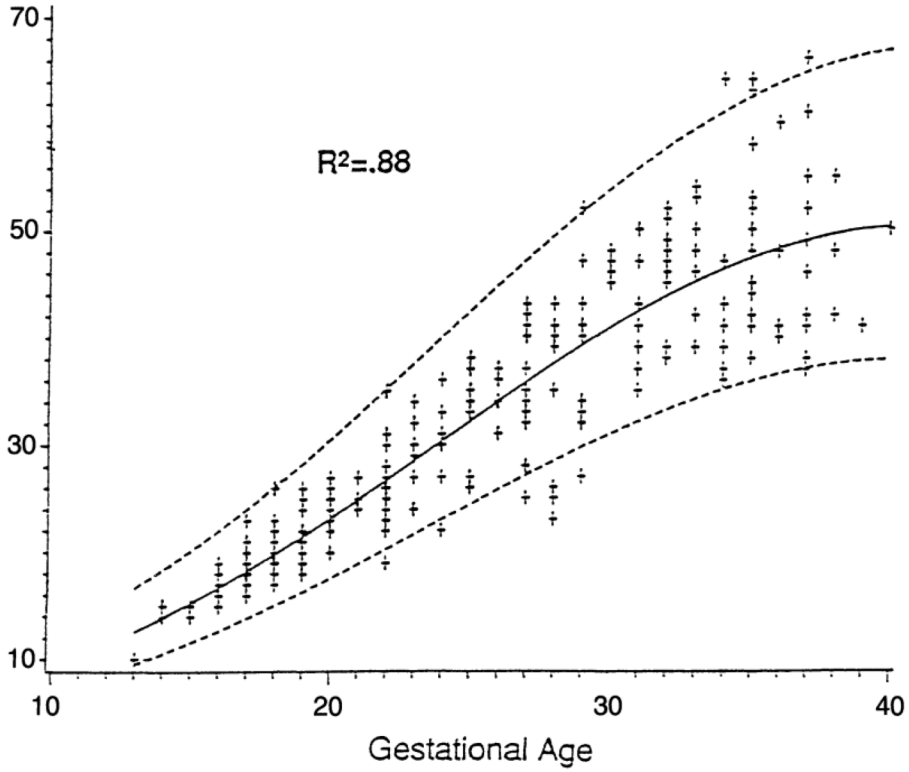

Figure 2. Fitted Model Based on $Y_2^* = \log Y$.

Source:

Transformations and R^2

Alastair Scott &Chris Wild