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

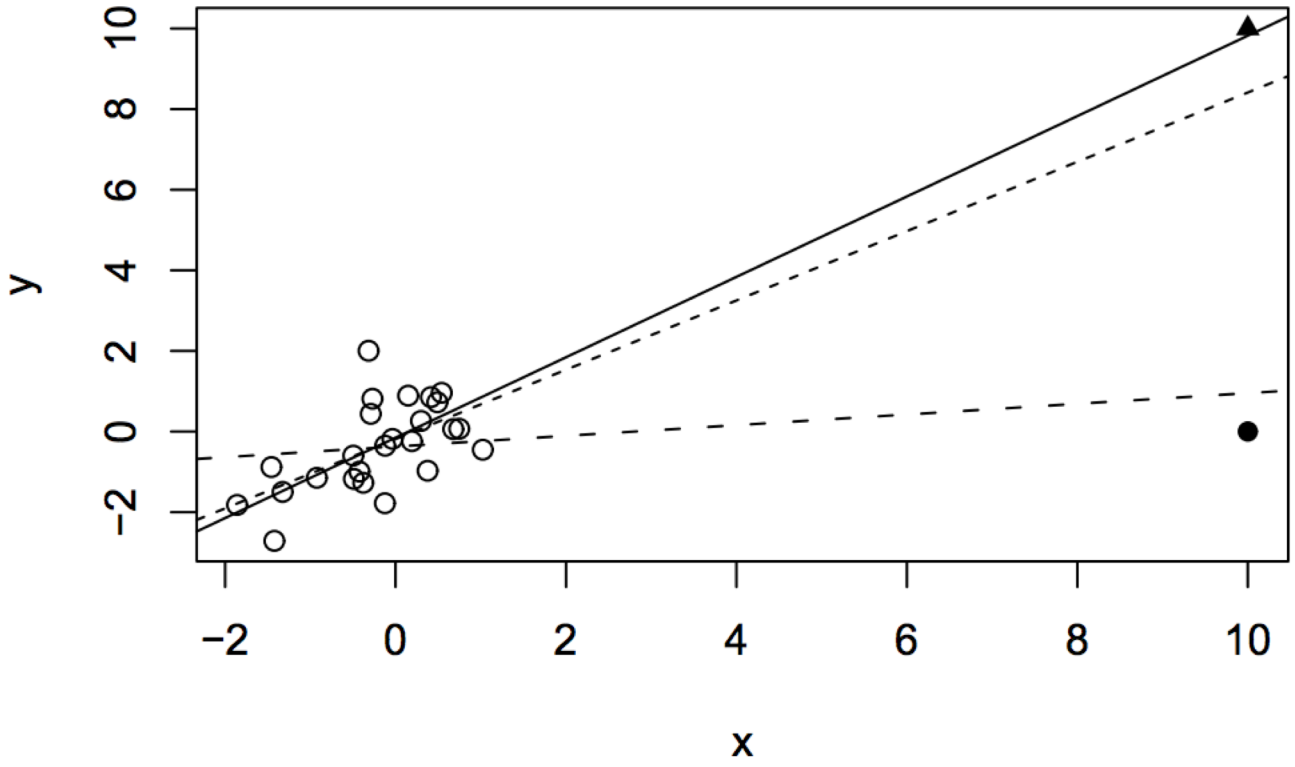


Figure 7.2: Outliers can conceal themselves. The solid line is the fit including the ▲ point but not the • point. The dotted line is the fit without either additional point and the dashed line is the fit with the • point but not the ▲ point.

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)

Regression & diagnostics with SAS

PROC REG

- You can fit linear regression models with PROC REG
- For example:

```
PROC REG data=iq;
    MODEL PIQ = brain height weight;
RUN;
```

This fits a model where "PIQ" is the outcome and the predictors are "brain", "height", and "weight"

- You can find info about PROC REG in class code and the handouts

Sums of squares (defined in previous lecture and later in this slideshow)

Number of Observations Read	38
Number of Observations Used	38

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5572.74444	1857.58148	4.74	0.0072
Error	34	13322	391.81789		
Corrected Total	37	18895			

Root MSE	19.79439	R-Square	0.2949
Dependent Mean	111.34211	Adj R-Sq	0.2327
Coeff Var	17.77799		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	111.35361	62.97110	1.77	0.0860
Brain	1	2.06037	0.56345	3.66	0.0009
Height	1	-2.73193	1.22943	-2.22	0.0330
Weight	1	0.00055994	0.19707	0.00	0.9977

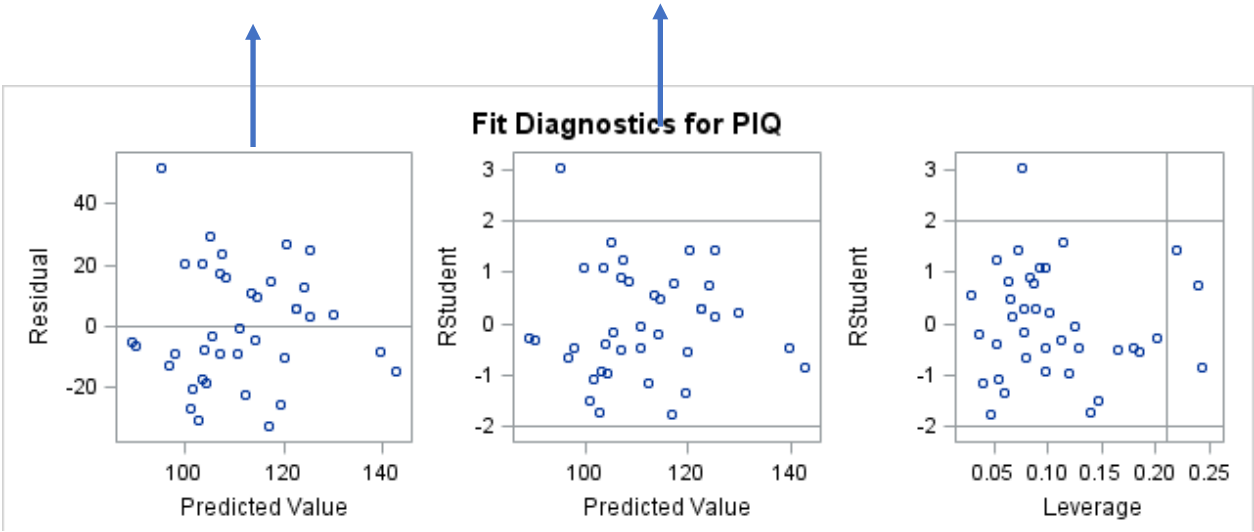
P-value for test with null: none of the predictors are useful

P-values for individual variables

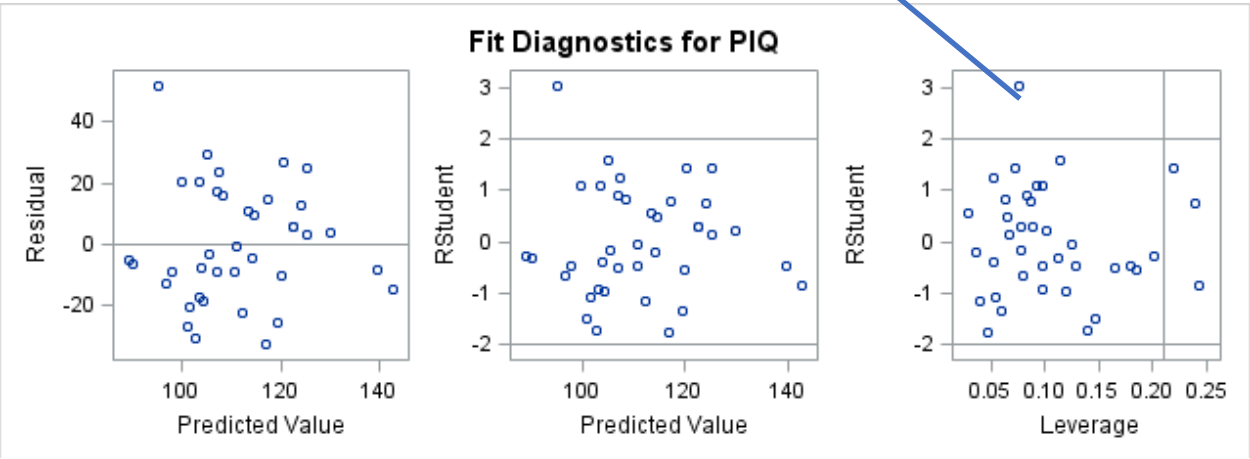
coefficients

If the model fits the data well, we should observe no "nonrandom" patterns (e.g. a parabola) and the spread of y-axis should not depend strongly on the values on the x-axis

Same idea as the previous plot, but the y-axis has been standardized so that, if the model fits the data well, roughly 95% of the points lie within the [-2,2] band

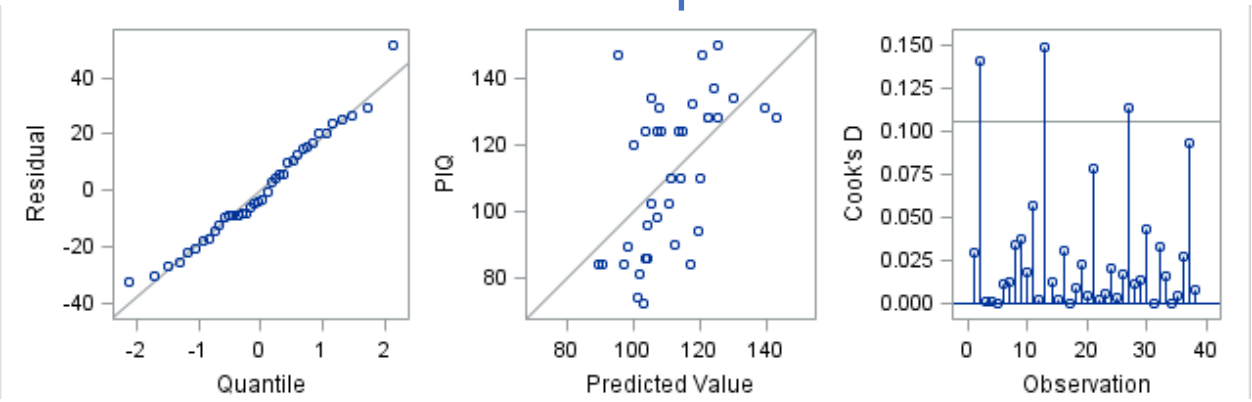


High values on the x-axis indicate high leverage points. SAS has a rule of thumb to flight "high leverage" points, but in general I look at observations that "stick out"

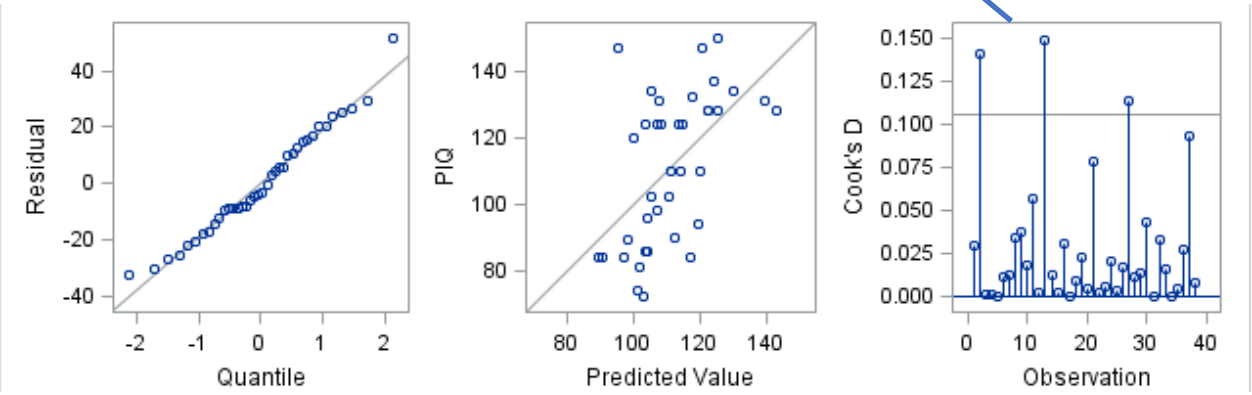


If the model fits the data well, the points should be on the line. Helpful for assessing the assumption of normality.
More here (or ask me)

x-axis: predicted values
y-axis: actual values
If the model does a good job at predicting, the points should align nicely around y = x

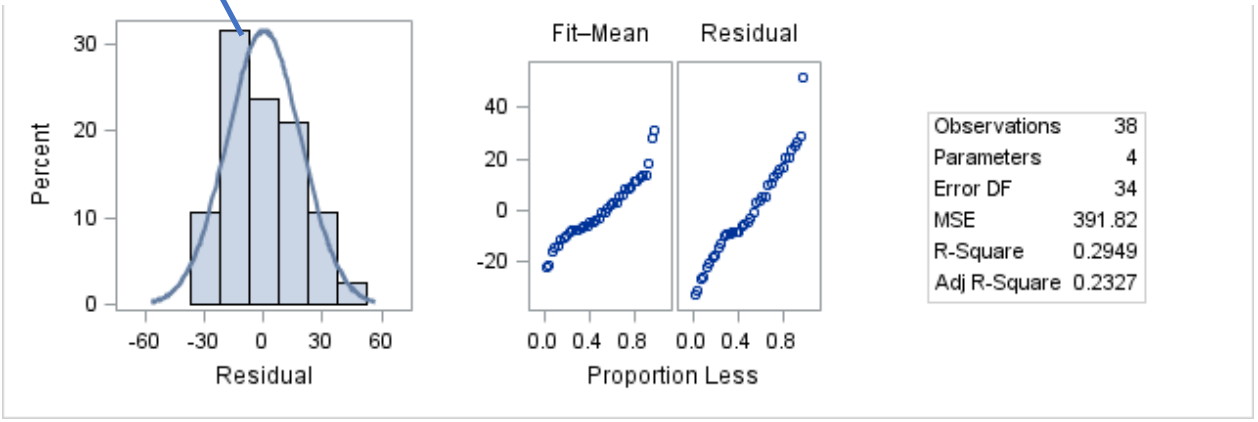


x-axis: observation number
y-axis: Cook's distances
SAS has a rule for flagging "big" Cook's distances. In general, I look for observations that stick out



Histogram of residuals vs best normal fit. If the residuals are roughly normal, the histogram and the overlaid curve should look similar. Useful for assessing normality of residuals

I don't like this plot! If you want to learn more, click here



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

Sums of squares

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
- It can be 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*!

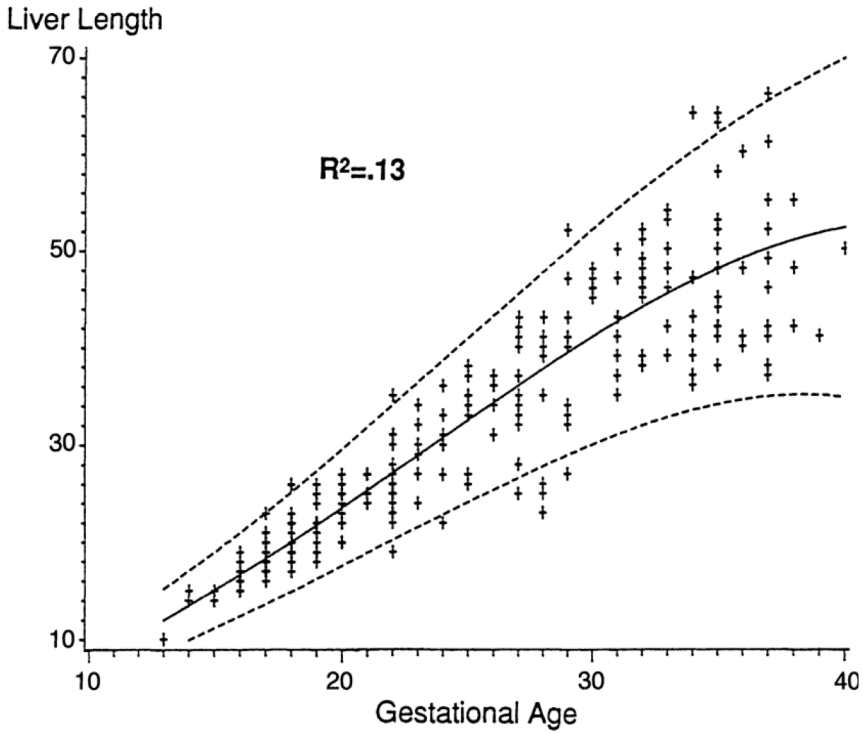


Figure 1. Fitted Model Based on $Y_1^* = Y/x^{3/2}$.

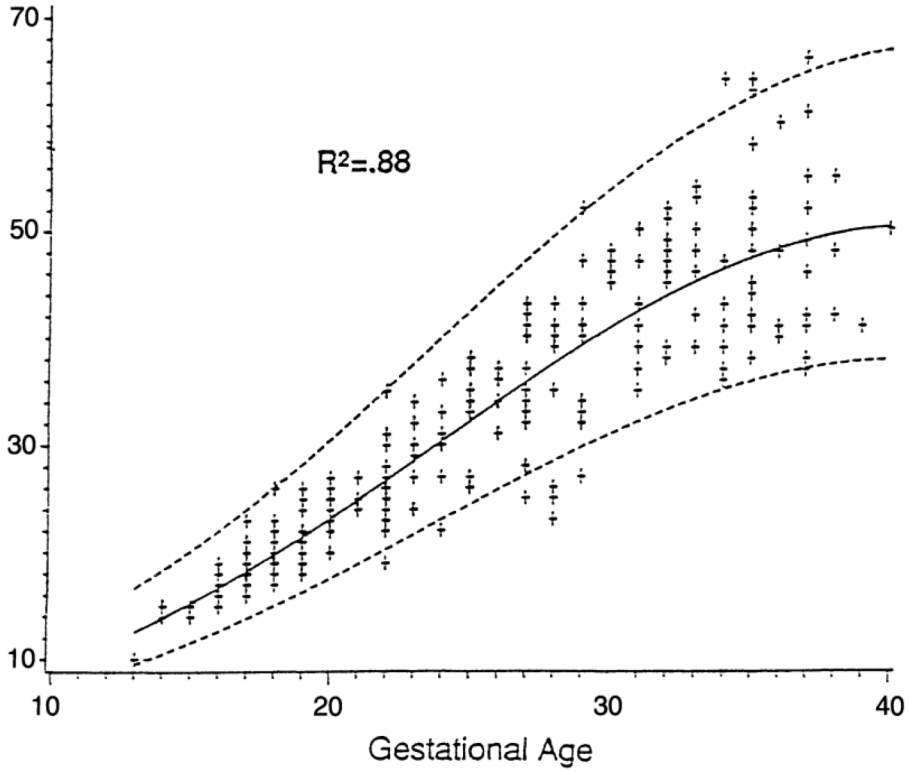


Figure 2. Fitted Model Based on $Y_2^* = \log Y$.

Source:

Transformations and R^2

Alastair Scott &Chris Wild