Multiple Regression: A Model for the Mean

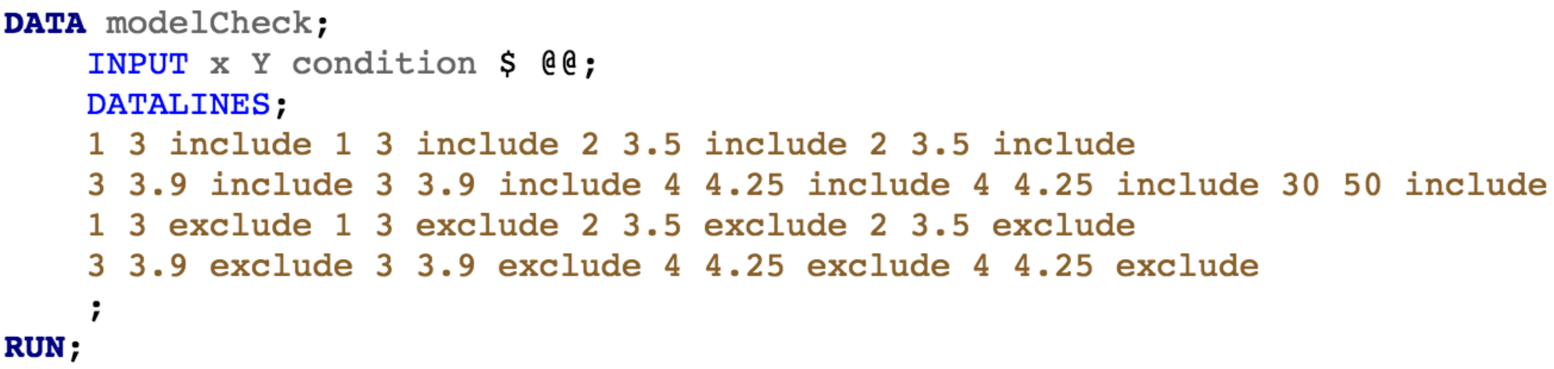
LEVERAGE

STUDENTIZED RESIDUALS

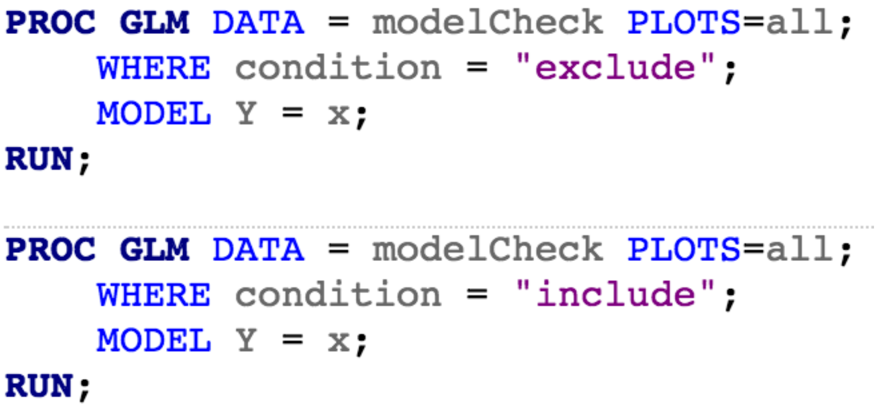
COOKS-D

Influential Observations

An Example Dataset



```
DATA modelCheck;
    INPUT x Y condition $ @@;
    DATALINES;
    1 3 include 1 3 include 2 3.5 include 2 3.5 include
    3 3.9 include 3 3.9 include 4 4.25 include 4 4.25 include 30 50 include
    1 3 exclude 1 3 exclude 2 3.5 exclude 2 3.5 exclude
    3 3.9 exclude 3 3.9 exclude 4 4.25 exclude 4 4.25 exclude
    ;
RUN;
```



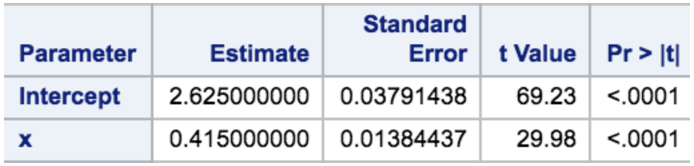
```
PROC GLM DATA = modelCheck PLOTS=all;
    WHERE condition = "exclude";
    MODEL Y = x;
RUN;
```

```
PROC GLM DATA = modelCheck PLOTS=all;
    WHERE condition = "include";
    MODEL Y = x;
RUN;
```

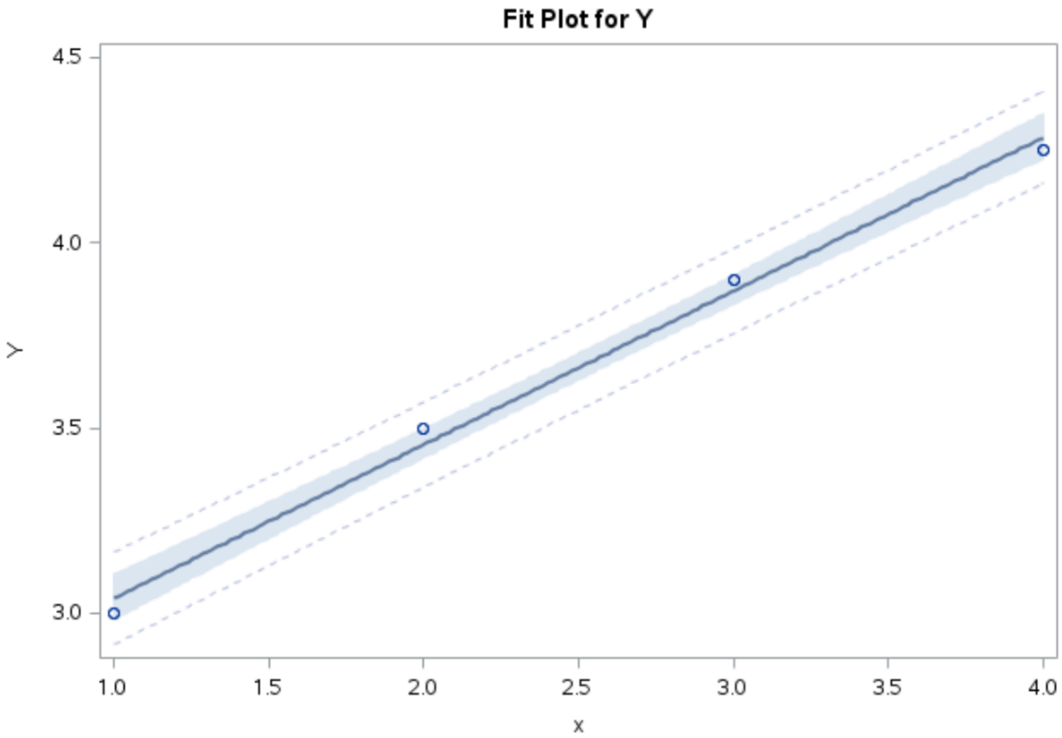
Influential Observations

- An influential observation is an observation that has a disproportionate affect on the estimated regression model
- Addressing influential observations is related to, but distinct from, assumption checking
- As the estimated regression model is based on average squared deviations, it is very sensitive to observations that are extreme in some direction
- The general idea is similar to the “outlier strategy” in Chapter 3.3.1
- However, the details are a little different due to there being outliers in *X* or in *Y*

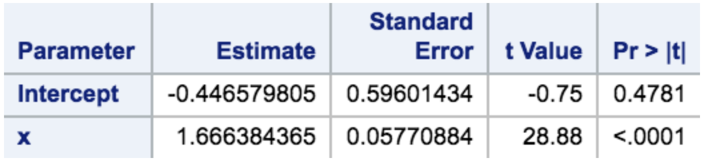
Influential Observations: Included



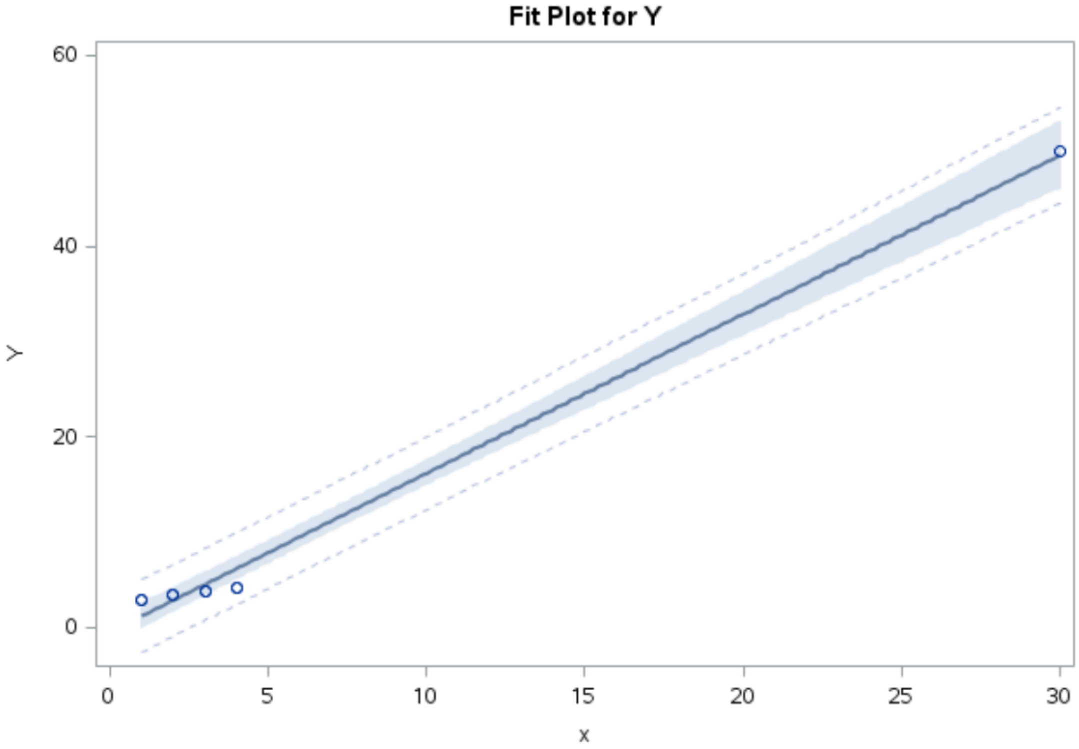
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.625000000	0.03791438	69.23	<.0001
x	0.415000000	0.01384437	29.98	<.0001



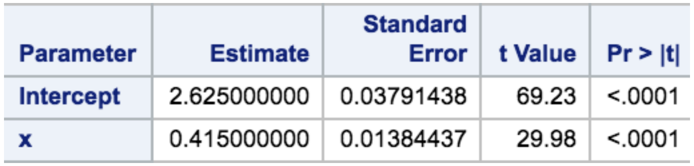
Influential Observations: Excluded



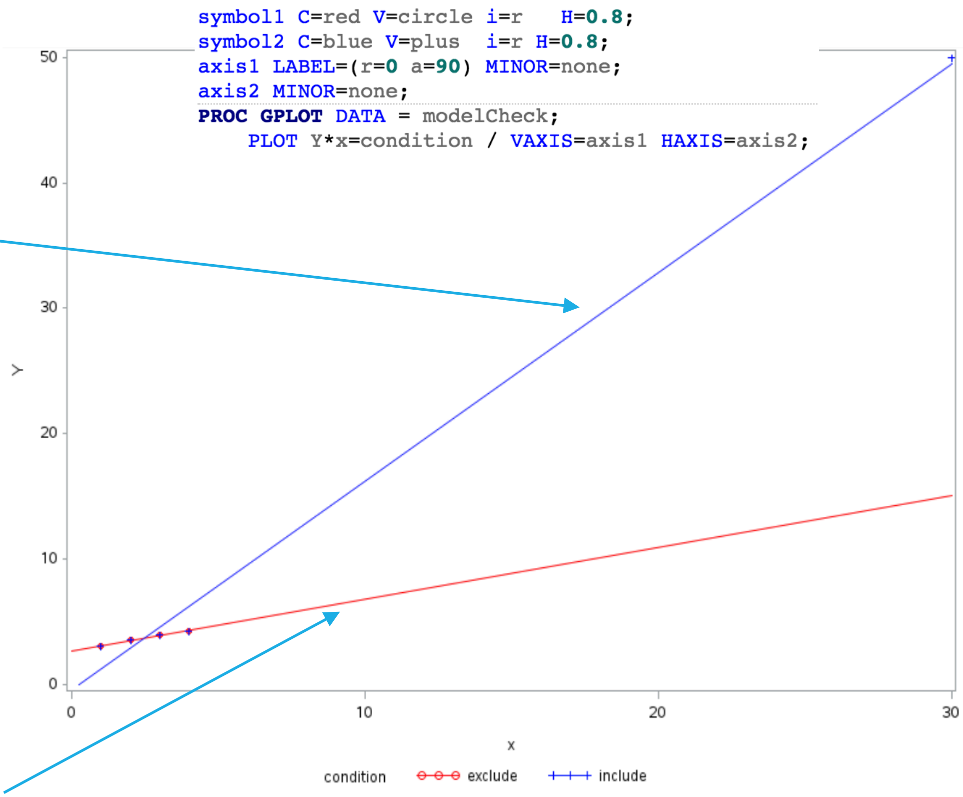
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.446579805	0.59601434	-0.75	0.4781
x	1.666384365	0.05770884	28.88	<.0001



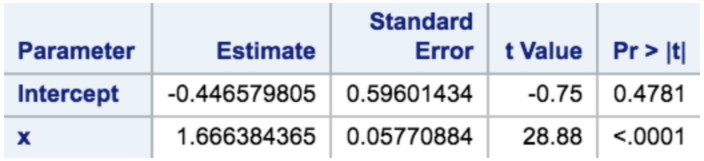
Influential Observations: Included and Excluded



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.625000000	0.03791438	69.23	<.0001
x	0.415000000	0.01384437	29.98	<.0001



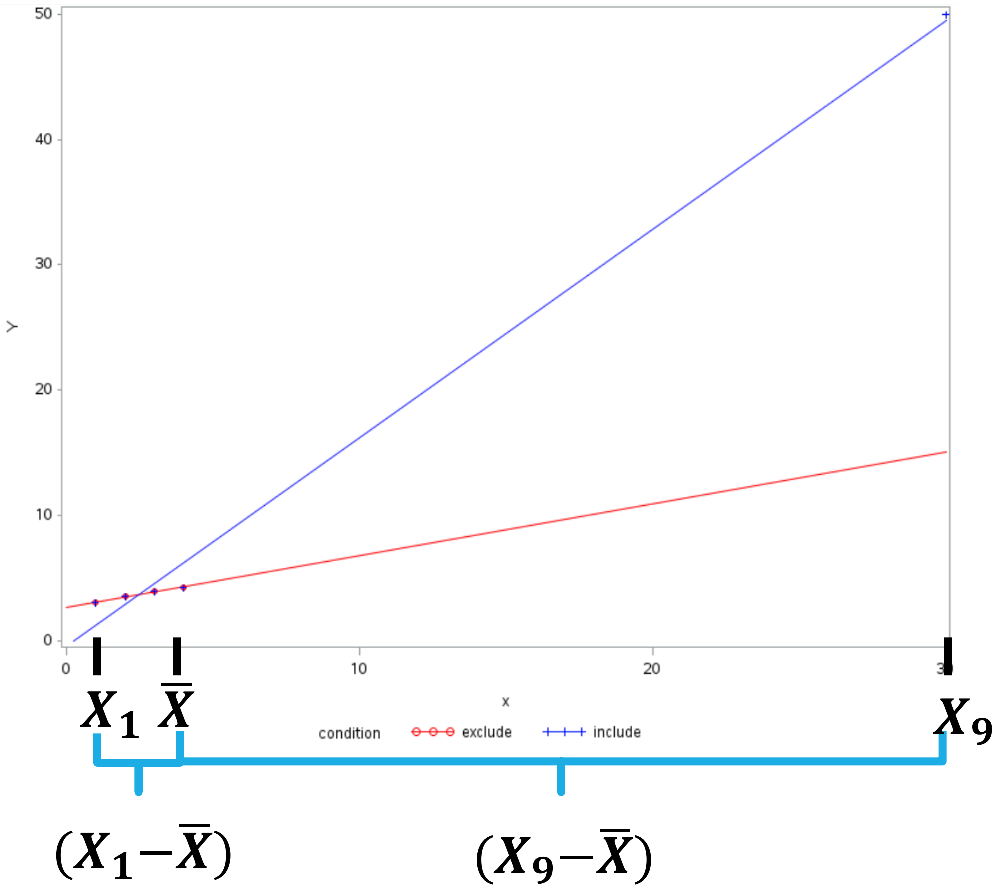
```
symbol1 C=red V=circle i=r   H=0.8;
symbol2 C=blue V=plus  i=r H=0.8;
axis1 LABEL=(r=0 a=90) MINOR=none;
axis2 MINOR=none;
PROC GPLOT DATA = modelCheck;
    PLOT Y*x=condition / VAXIS=axis1 HAXIS=axis2;
```



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.446579805	0.59601434	-0.75	0.4781
x	1.666384365	0.05770884	28.88	<.0001

The Idea of Leverage

Note: Here, there are $n = 9$ observations, with each of the four “circles” on the left being duplicated



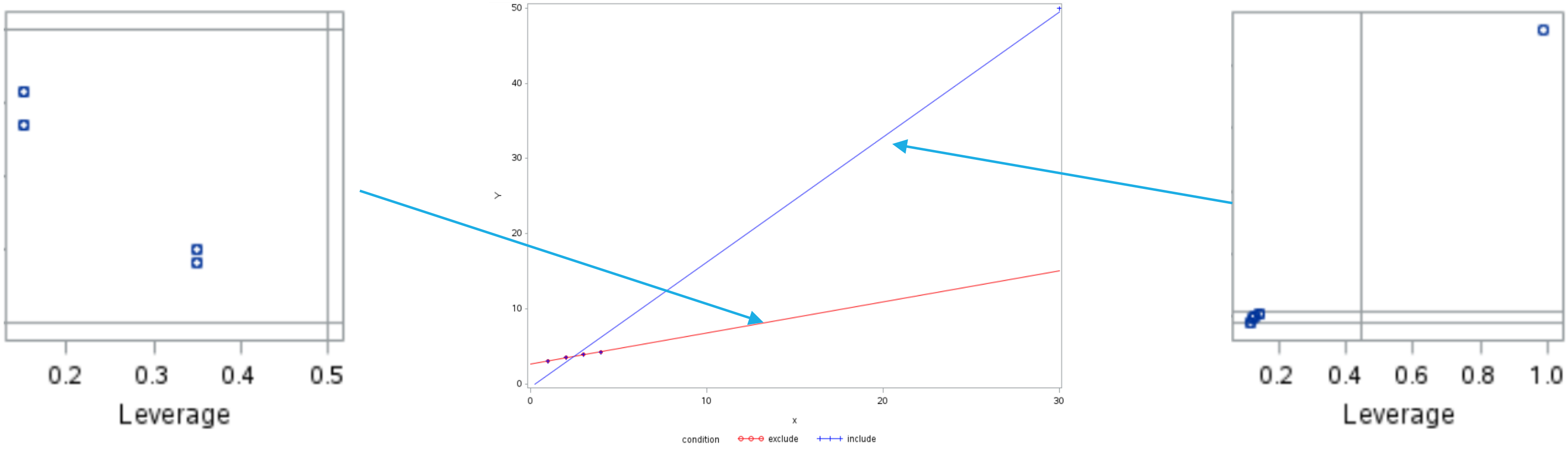
Definition of Leverage for Simple Linear Regression

The LEVERAGE of the i^{th} observation is defined to be h_i where:

$$h_i = \frac{1}{n-1}\left(\frac{X_i - \bar{X}}{s_X}\right)^2 + \frac{1}{n} = \left(\frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}\right)^2 + \frac{1}{n}$$

In words, this can be expressed equivalently as:

- The squared distance of X_i from the mean in units of standard deviation of X
- Or the proportion of total sum of squares of X contributed by X_i



Notation for the Mean

- Y is the response variable
- $x_1, \dots, x_p$ are the explanatory variables
- $\mu\{Y|X\}$ is the "mean of Y as a function of X $= (x_1, \dots, x_p)$"

This gets estimated from data:

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

via solving for the "least squares" solution:

$\hat{\mu}\{Y|X\} = \text{minimizer}(\sum_{i=1}^{n}(Y_i - \mu\{Y_i|X_i\})^2)$

over functions $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$

The quality of this solution depends on whether the assumptions are reasonably met

Definition of Leverage for Multiple Linear Regression

$$\mathbb{X} = [1, x_1, x_2, \dots, x_p] = \begin{bmatrix} 1, X_1{}^T \\ \vdots \\ 1, X_n{}^T \end{bmatrix} = \begin{bmatrix} 1, & X_{11} & \cdots & X_{1p} \\ \vdots & & \ddots & \vdots \\ 1, & X_{n1} & \cdots & X_{np} \end{bmatrix}, \qquad \mathbb{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

Data: (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n)

Estimate model via:

$\hat{\mu}\{Y|X\} =$ minimizer$(\sum_{i=1}^{n}(Y_i - \mu\{Y_i|X_i\})^2)$
over functions $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$

This corresponds to $\hat{\mu}\{Y|X\} = X^T\hat{\beta}$, where

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

Note that we can write the vector of fitted values as:

$$\widehat{\mathbb{Y}} = \mathbb{X}\hat{\beta} = \mathbb{H}\mathbb{Y}$$

This “$\mathbb{H}$” is an important object..

Definition of Leverage for Multiple Linear Regression

The leverage of the i^{th} observation is defined to be h_i where:

$h_i = \mathbb{H}_{ii}$,

(that is, the i^{th} diagonal entry of the *n* by *n* matrix $\mathbb{H} = \mathbb{X}\,(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$)

Note that ${}^{1}\!/_{n} \leq h_i \leq 1$ and $\sum_{i=1}^{n} h_i = p$

How large is too large? Somewhat arbitrary cut-off: ${}^{2p}\!/_{n}$

Write $\mathbb{X} = \mathbb{U}\mathbb{D}\mathbb{V}^T$

(as in the notes _linearAlgebraProbabilityOnly.pdf)

The orthogonal matrix $\mathbb{U}$ has n rows and p columns

$\mathbb{H}_{ii} = ||U_i||_2^2$

(It is also the partial derivative of $\hat{\mu}\{Y_i|X_i\}$ w/ respect to Y_i)

Normalized Residuals

The residuals $resid_i = Y_i - \hat{\mu}\{Y_i|X_i\}$ are like an estimate of ε_i in the model:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$

It makes sense to report a "normalized" version of the residuals

Looking at the above definition of $resid_i$

$$\text{Var}(resid_i) = \sigma^2(1 - h_i)$$

So, we can normalize the residuals if we have an estimate of σ^2 by reporting

$$\frac{resid_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}}$$

Normalized Residuals

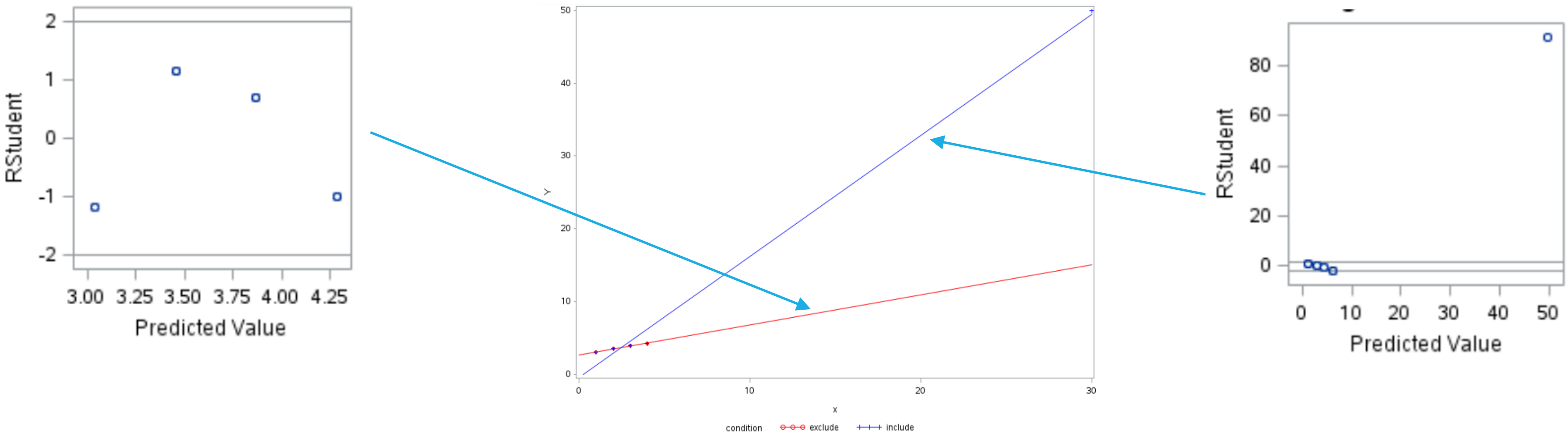
A common convention: use the word “studentized” instead of normalized

This leads to the definition of STUDENTIZED RESIDUAL:

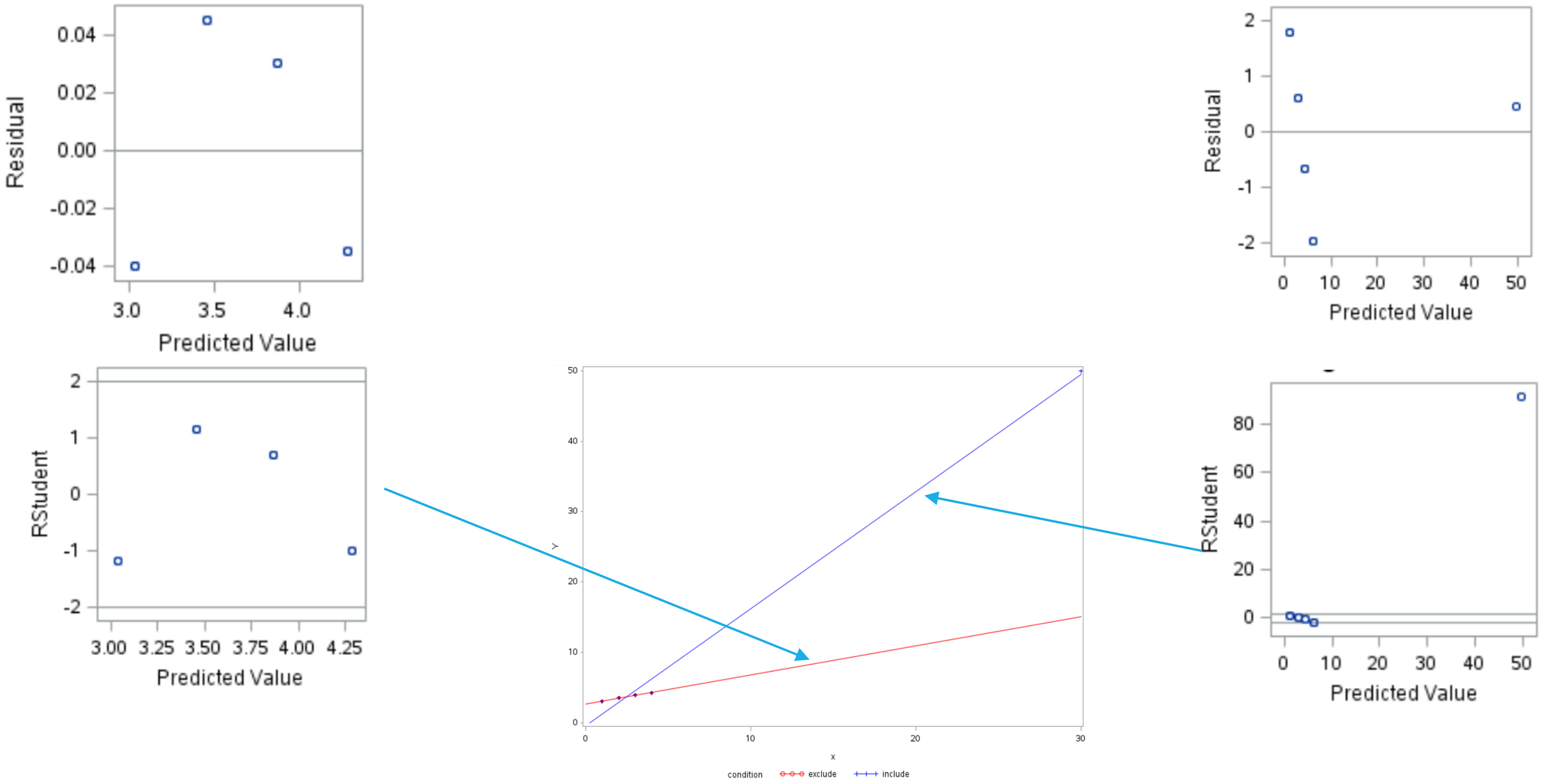
$$studres_i = \frac{resid_i}{\sqrt{\hat{\sigma}^2(1-h_i)}}$$

Observation i is considered extreme if $|studres_i| > 2$

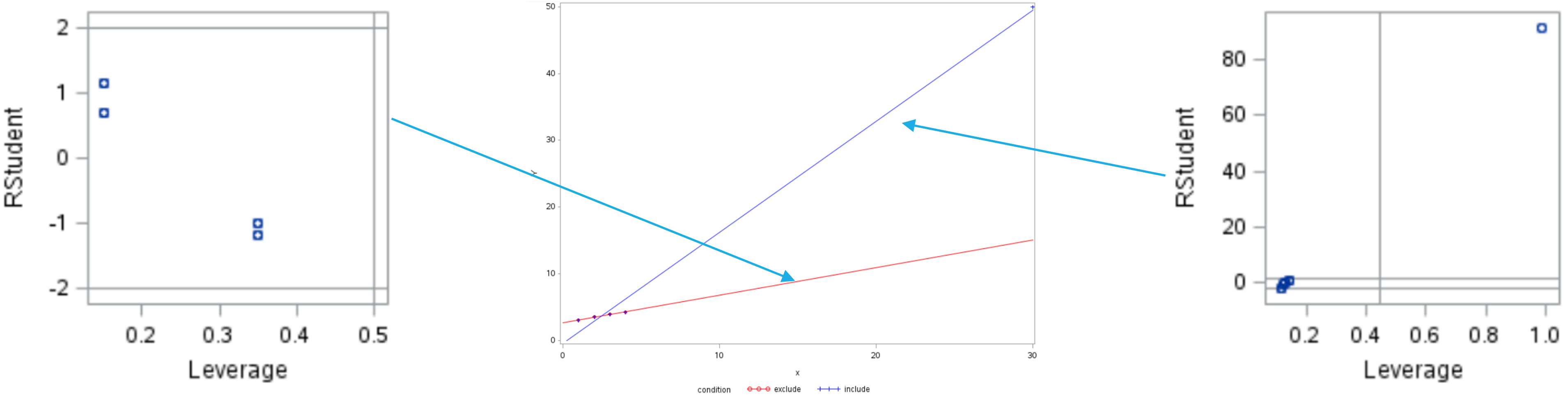
(Important: We would expect 5% of observations to exceed this threshold)



Raw Residuals vs. Studentized Residuals



Putting Leverage and Studentized Residuals Together



Cook’s D(istance): The Key Idea

It can be a difficult task, especially when $p > 3$, to disentangle:

- Is the residual small due to a good fitting model or does that point have enough leverage to ”pull” the model fit towards it?
- A direct way to measure the affect of an observation is via a sensitivity analysis
- This means, fitting the model with an observation and without an observation and measuring the change
- A large change indicates that observation is having an outsized effect on the model fit

Cook’s D(istance): The Key Idea

Let’s define the least squares solution with the k^{th} observation removed:

$\hat{\mu}_{(k)}\{Y|X\} =$ minimizer$(\sum_{i \neq k}(Y_i - \mu\{Y_i|X_i\})^2)$

over functions $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$

This corresponds to $\hat{\mu}\{Y|X\} = X^T \hat{\beta}_{(k)}$, where

$\hat{\beta}_{(k)} = (\mathbb{X}_{(k)}{}^T \mathbb{X}_{(k)})^{-1} \mathbb{X}_{(k)}{}^T \mathbb{Y}_{(k)}$

Now, we can compare the model with and without the k^{th} observation by comparing $\mathbb{X}\hat{\beta}$ to $\mathbb{X}\hat{\beta}_{(k)}$

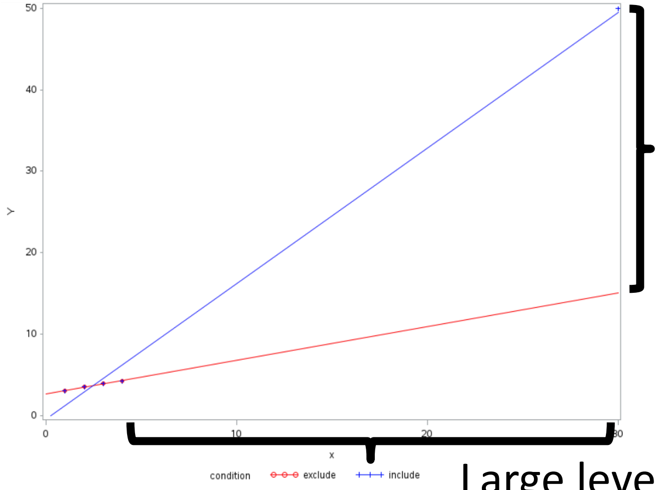
Cook’s D(istance)

$\hat{\sigma}^2$ = MSE for the least squares fit to all n observations

$$D_k = \frac{(\hat{\beta}-\hat{\beta}_{(k)})\mathbb{X}^T\mathbb{X}(\hat{\beta}-\hat{\beta}_{(k)})}{p\hat{\sigma}^2} = \frac{{studres_k}^2 h_k}{p(1-h_k)}$$

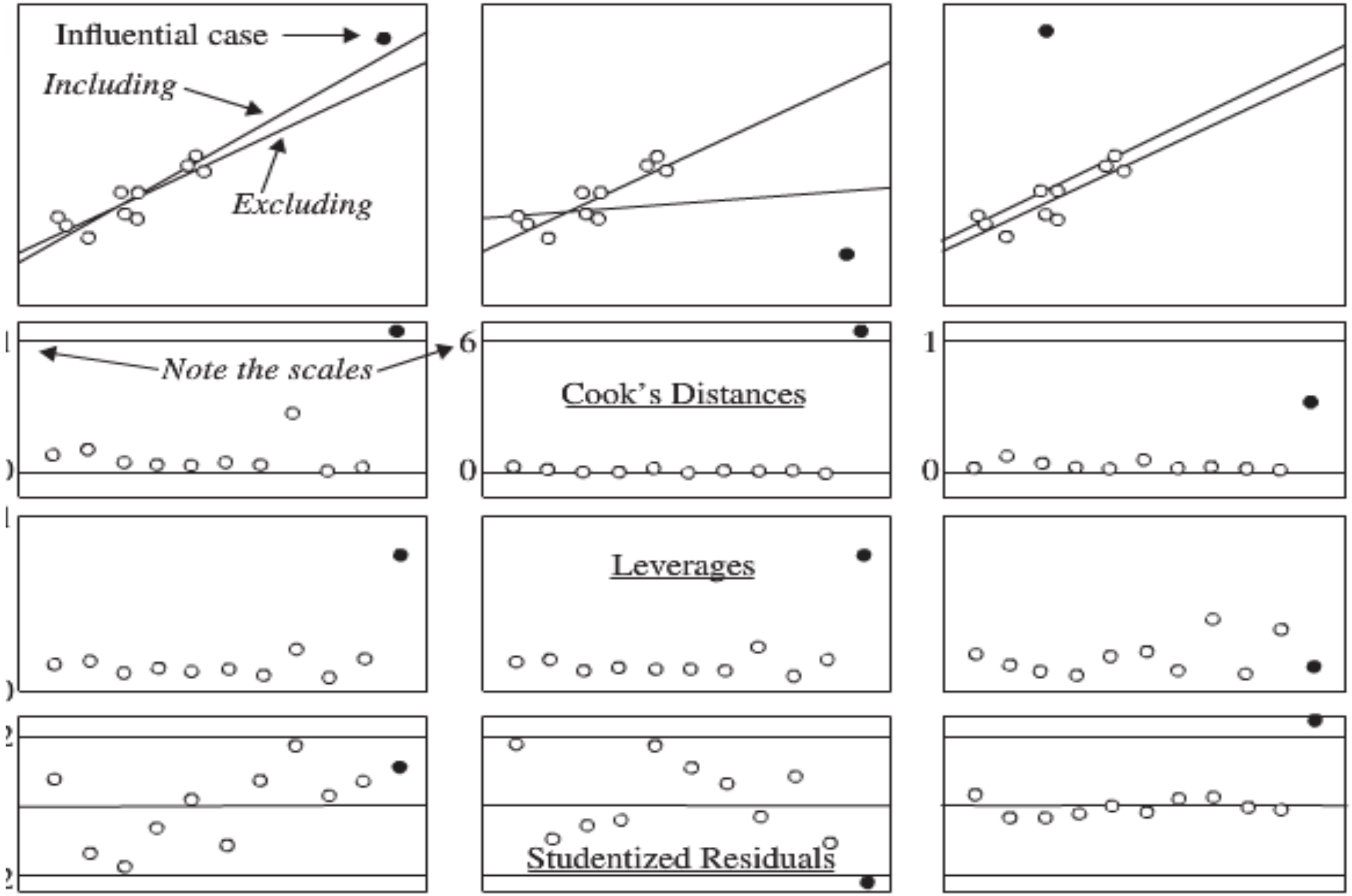
Hence, Cook’s D is large if both:

- h_k is close to 1 (that is, if the observation has high leverage)
- $|studres_k|$ is large



$(\hat{\beta}-\hat{\beta}_{(k)})\mathbb{X}^T\mathbb{X}(\hat{\beta}-\hat{\beta}_{(k)})$

Large leverage



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.

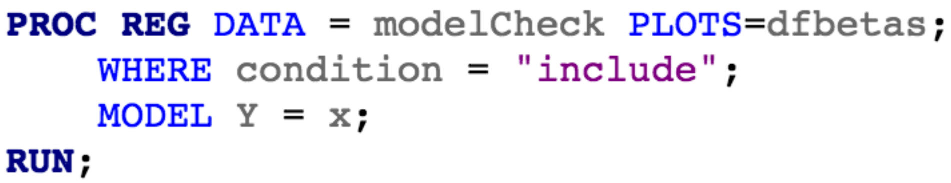
Applying Sensitivity Analysis to the Coefficients

Cook’s D is in terms of the fitted values

We can extend this idea to the coefficients as well

This can be helpful because a high leverage observation might not affect the coefficient estimate of an explanatory variable of interest

We need to use PROC REG to get the necessary output:



```
PROC REG DATA = modelCheck PLOTS=dfbetas;
    WHERE condition = "include";
    MODEL Y = x;
RUN;
```

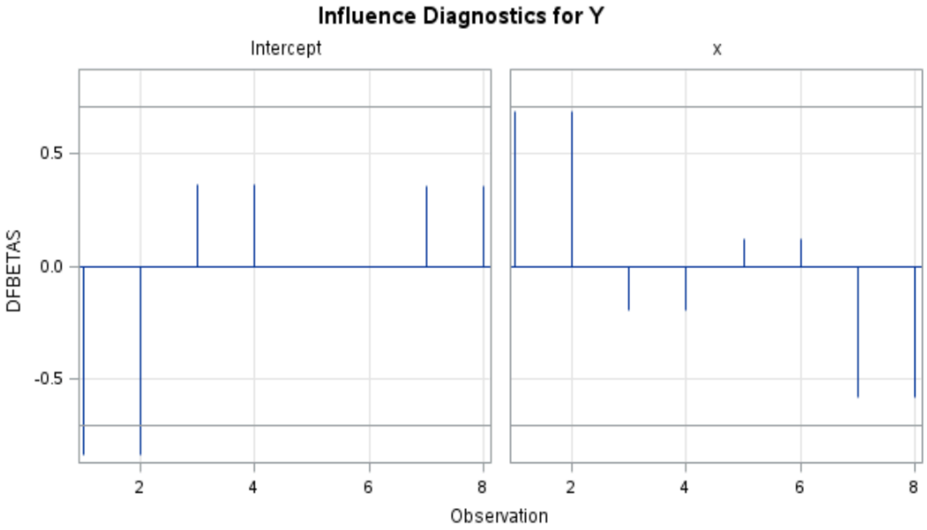
See the following website for details on computation:

(https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect040.htm)

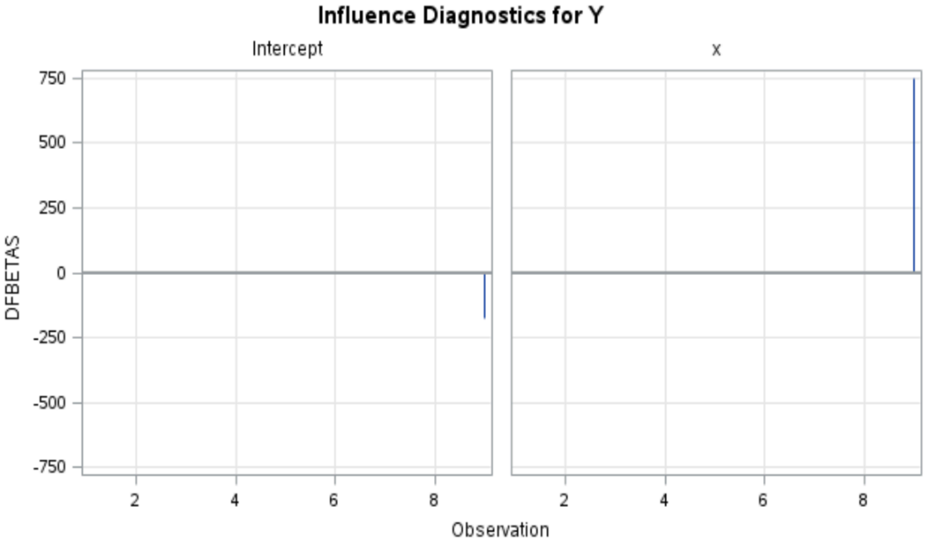
Applying Sensitivity Analysis to the Coefficients: dfbetas

There are $p + 1$ explanatory variables, so there will be $p + 1$ plots

There are n observations, so there will be n values for each plot



“Exclude”



“Include”