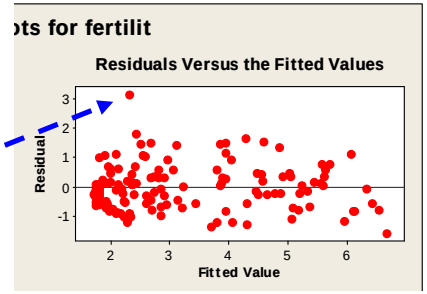
Regression Mop-up in 3 Acts

<u>Act I :</u> Outliers revisited – a way to functionally remove a suspect data-point.

Example : Poverty. *We fit a model that predicted fertility based on infant mortality. At the end, we found one unusual point – Saudi Arabia.*



How to deal with outliers :

1) Remove

- If point is not influential, probably no compelling reason to do this.
- Point may have an effect on R-squared. *In Poverty data, R-squared went from 77% to 80% when this point was removed.*
- <u>Must record removal if you do remove!!</u>

2) Ignore

3) Make an INDICATOR VARIABLE for the outlier

For the Poverty Data, this means creating a new variable that is zero for all observations except for Saudi Arabia! Include this indicator variable in the model :

$$\text{Fertility} = \beta_0 + \beta_1 \text{Infant Mort} + \beta_2 \{\text{Saudi Indicator}\} + \varepsilon$$

What this does :

- *The value of β_2 is simply the magnitude of the residual for Saudi Arabia, and the new residual for this point will be zero!*
- *Saudi Arabia no longer has any effect on the rest of the model – i.e. we get the desired R-square increase.*
- *The number of degrees of freedom is the same as if we had removed the point!*
- ***If the point is a 'true' outlier, the test of the coefficient for β_2 will be significant!***

Example : Poverty. *Here is the result from MINITAB if we use an indicator variable for Saudi Arabia :*

```
The regression equation is
fertilit = 1.55 + 0.0398 mortalit + 3.18 Indicator

Predictor      Coef   SE Coef      T      P
Constant    1.55169   0.09372  16.56  0.000
mortalit   0.039770  0.001700  23.40  0.000
Indicator    3.1845    0.7283   4.37  0.000

S = 0.724635   R-Sq = 80.3%   R-Sq(adj) = 80.0%
```

We get the higher R-squared, AND we find that Saudi Arabia is a significant outlier.

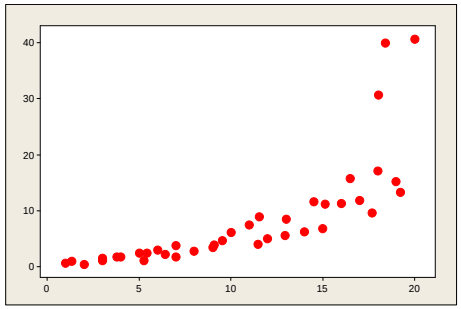
Just to show you this only works for outlier, here is the same model with an indicator variable for another country, say Honduras :

```
The regression equation is
fertilit = 1.58 + 0.0394 mortalit + 0.974 Indicator


Predictor      Coef   SE Coef      T      P
Constant    1.58280   0.09923  15.95  0.000
mortalit   0.039400  0.001802  21.87  0.000
Indicator    0.9740    0.7720   1.26  0.209


S = 0.769089   R-Sq = 77.8%   R-Sq(adj) = 77.4%
```

Act II : More on Transformations

Some visual notes on transformations :

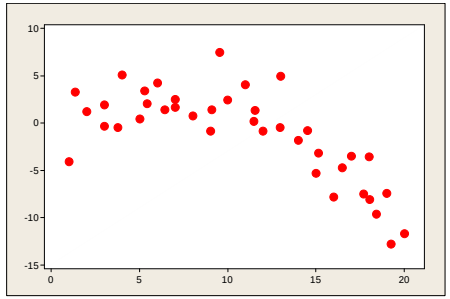
If you see heteroskedacity, consider a transformation of the Y variable – use Box-Cox or use common sense.



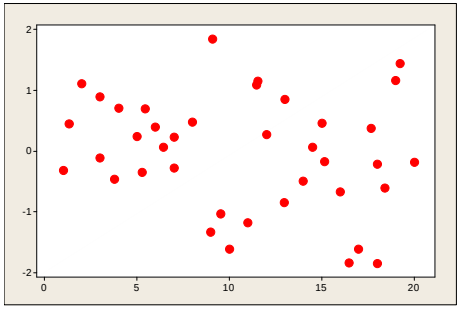
In this case, I'd try log(Y). The fact that X's are relatively uniformly spaced suggest that no transformation of X is necessary.

If you do NOT see heteroskedacity, but you see a non-linear polynomial trend, fit a model that includes additional polynomial terms in X : *(i.e. in this case, make a new variable that is the square of X and include as a predictor)*

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$



If you do NOT see heteroskedacity, and you don't see any non-linear (and non – predictable trends), you should <u>DO NOTHING!!</u> Transformations will <u>NOT HELP.</u>



Act III : Additional MINITAB output

Example : Poverty. *Let's look at the output we've been skimming past : first part we've done :*

```
The regression equation is
fertilit = 1.55 + 0.0398 mortalit + 3.18 Indicator

Predictor      Coef   SE Coef      T      P
Constant    1.55169   0.09372  16.56  0.000
mortalit   0.039770  0.001700  23.40  0.000
Indicator    3.1845    0.7283   4.37  0.000

S = 0.724635   R-Sq = 80.3%   R-Sq(adj) = 80.0%
```

*This next part we'll cover more in ANOVA – however, this is a test of whether or not there are **ANY** significant predictors in the regression model. The p-values here says 'YES'!*

```
Analysis of Variance

Source           DF      SS      MS       F      P
Regression        2  292.44  146.22  278.46  0.000
Residual Error  137   71.94    0.53
Total           139  364.38

Source     DF  Seq SS
mortalit    1  282.40
Indicator   1   10.04
```

Next, MINITAB lists 'Unusual Observations'. These are potential outliers and influential points.

- *Minitab lists any point whose standardized residual is larger than 2 in absolute value (i.e. more than 2 standard deviations from the mean).*
- *MINITAB also lists points with 'large leverge' – i.e. INFLUENTIAL POINTS. These points typically have a small residual*

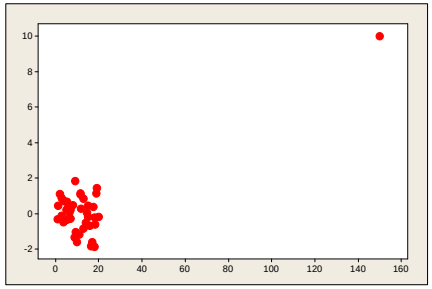
```
Unusual Observations

Obs  mortalit  fertilit     Fit  SE Fit  Residual  St Resid
 53        58    5.3600  3.8424  0.0672    1.5176      2.10R
 64        68    5.9700  4.2719  0.0765    1.6981      2.36R
 71        60    5.4500  3.9220  0.0686    1.5280      2.12R
 78        39    4.5800  3.0948  0.0616    1.4852      2.06R
 96        28    4.2000  2.6573  0.0658    1.5427      2.14R
104       129    5.1200  6.6939  0.1612   -1.5739     -2.23R
112        23    3.9700  2.4743  0.0690    1.4957      2.07R
117        22    4.2500  2.4087  0.0703    1.8413      2.55R
118        18    5.4600  5.4600  0.7246    0.0000         * X
138        76    6.1600  4.5742  0.0848    1.5858      2.20R
```



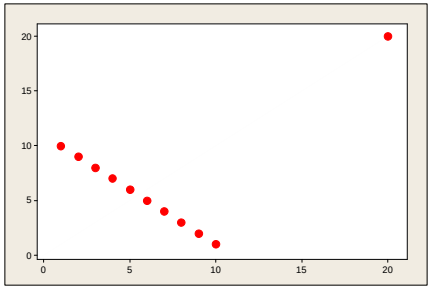
```
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

In this example, the only potentially influential point is 118, Saudi Arabia, which has a residual of zero! (since we made an indicator variable for this point).

Note that influential points can make R-squared lower OR higher.



R-squared with Influential Point : 0.66
R-squared without Influential Point : 0.04



R-squared with Influential Point : 0.16
R-squared without Influential Point : 1.00

Example : Teenage gambling behaviour (in the U.K)

A 1988 survey studied gambling habits of 47 U.K. teenagers. The variables measured were

- *gender (0=male, 1=female)*
- *status (arbitrary scale based on parents' occupation. Higher numbers → higher status)*
- *income (allowance + earnings) in pounds/wk*
- *verbal intelligence (number of words out of 12 correctly defined)*
- *estimate (from questionnaire answers) of expenditure on all forms of gambling, in pounds/yr*

Data from Ide-Smith & Lea, 1988, Journal of Gambling Behavior, 4, 110-118.