10 Model Checking and Regression Diagnostics

The simple linear regression model is usually written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the ϵ_is are independent normal random variables with mean 0 and variance σ^2. The model implies (1) The average Y-value at a given $X-$value is linearly related to X. (2) The variation in responses Y at a given X value is constant. (3) The population of responses Y at a given X is normally distributed. (4) The observed data are a random sample.

A regression analysis is never complete until these assumptions have been checked. In addition, you need to evaluate whether individual observations, or groups of observations, are unduly influencing the analysis. A first step in any analysis is to plot the data. The plot provides information on the linearity and constant variance assumption.

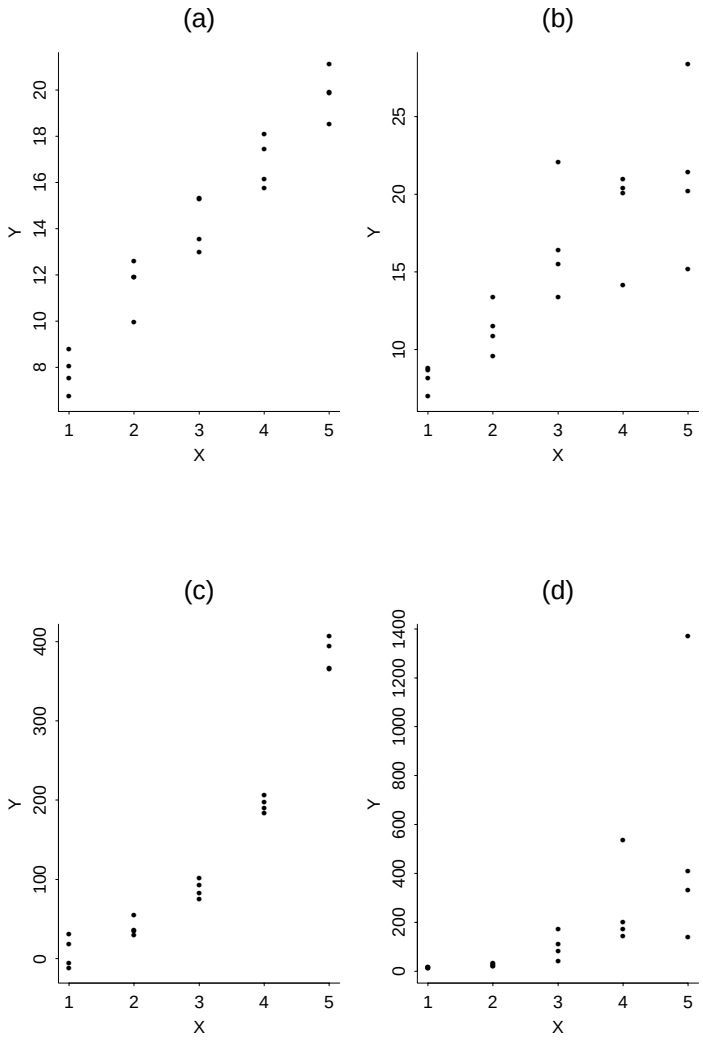


Figure (a) is the only plot that is consistent with the assumptions. The plot shows a linear relationship with constant variance. The other figures show one or more deviations. Figure (b) shows a linear relationship but the variability increases as the mean level increases. In Figure (c) we see a nonlinear relationship with constant variance, whereas (d) exhibits a nonlinear relationship with non-constant variance.

In many examples, nonlinearity or non-constant variability can be addressed by **transforming** Y or X (or both), or by fitting **polynomial models**. These issues will be addressed later.

Residual Analysis

A variety of methods for assessing model adequacy are based on the observed residuals,

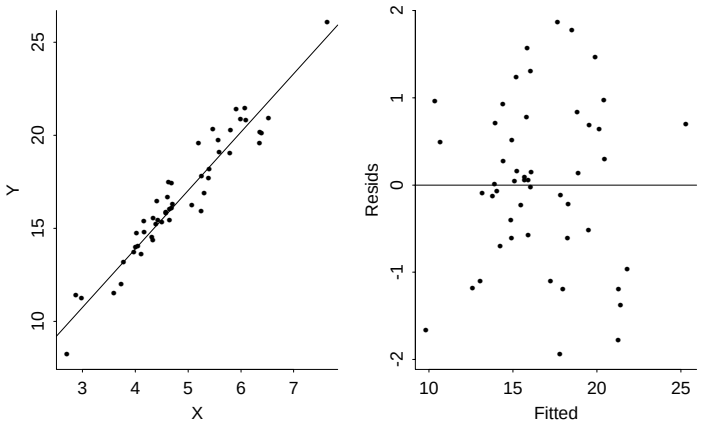
$$e_i = Y_i - \hat{Y}_i \quad \text{i.e. Observed } - \text{ Fitted values.}$$

The residuals are usually plotted in various ways to assess potential inadequacies. The observed residuals have different variances, depending on X_i, so many statisticians prefer to plot the **studentized residuals** (sometimes called the standardized residuals)

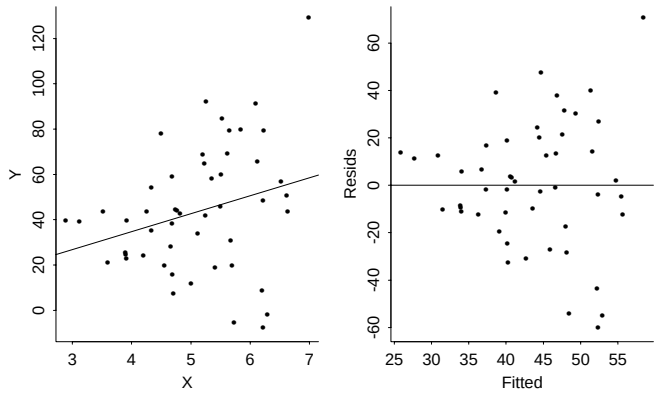
$$r_i = \frac{e_i}{SE(e_i)}.$$

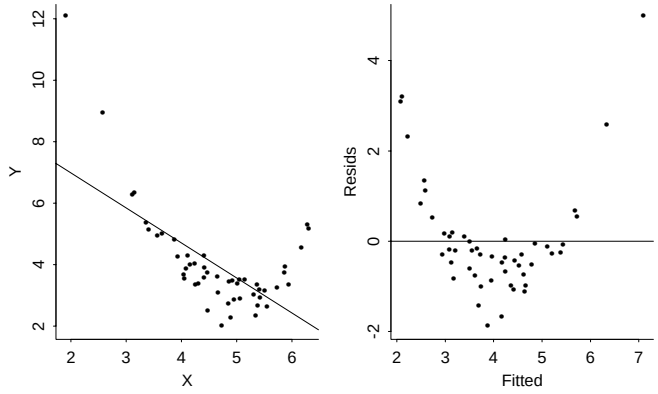
The studentized residuals have a constant variance of 1 (approximately). I will focus on diagnostic methods using the studentized residuals.

A plot of the studentized residuals r_i against the fitted values $\hat{Y}_i$ often reveals inadequacies with the model. The real power of this plot is with multiple predictor problems (multiple regression). The information contained in this plot with simple linear regression is similar to the information contained in the original data plot, except it is scaled better and eliminates the effect of the trend on your perceptions of model adequacy. The residual plot should exhibit no systematic dependence of the sign or the magnitude of the residuals on the fitted values:



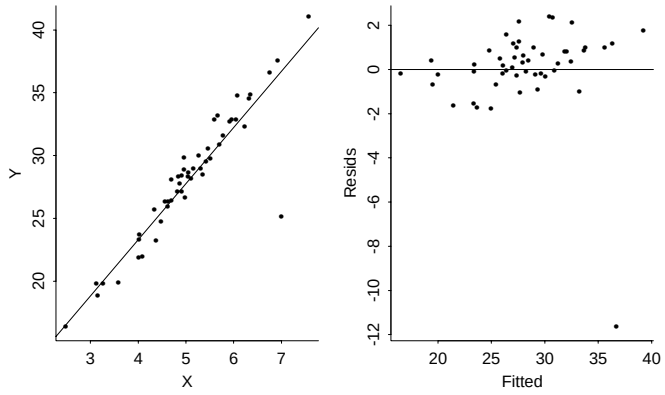
The following sequence of plots show how inadequacies in the data plot appear in a residual plot. The first plot shows a roughly linear relationship between Y and X with non-constant variance. The residual plot shows a megaphone shape rather than the ideal horizontal band. A possible remedy is a **weighted least squares** analysis to handle the non-constant variance, or to transform Y to stabilize the variance. Transforming the data may destroy the linearity.





The plot above shows a nonlinear relationship between Y and X. The residual plot shows a systematic dependence of the sign of the residual on the fitted value. Possible remedies were mentioned earlier.

The plot below shows an **outlier**. This case has a large residual and large studentized residual. A sensible approach here is to refit the model after holding out the case to see if any conclusions change.



Checking Normality

The normality assumption for the ϵ_i can be evaluated visually with a boxplot or a normal probability plot (rankit plot) of the r_i, or formally with a Shapiro-Wilk test. The rankit plot often highlights **outliers**, or poorly fitted cases. If an outlier is held out of the data and a new analysis is performed, the resulting normal scores plot may be roughly linear, but often will show a short-tailed distribution. (Why?).

You must interpret regression tests and CI with caution with non-normal data. Statisticians developed robust regression methods for non-normal data but they are not widely available in standard software packages. Minitab has a robust line fitting procedure for simple linear regression, but its features are rather limited.

Checking Independence

Diagnosing dependence among observations requires an understanding of the data collection process. There are a variety of graphical and inferential tools for checking independence for data collected over time (called a time series). The easiest check is to plot the r_i against time index and look for any suggestive patterns.

Outliers

Outliers are observations that are poorly fitted by the regression model. The response for an outlier is far from the fitted line, so outliers have large positive or negative values of the studentized residual r_i. Usually, $|r_i| > 2$ is considered large. Outliers are often highlighted in residual plots.

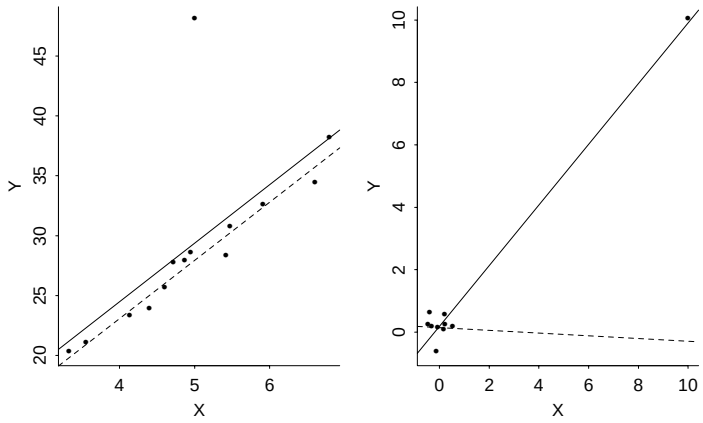
What do you do with outliers? Outliers may be due to incorrect recordings of the data or failure of the measuring device, or indications or a change in the mean or variance structure for one or more cases. Incorrect recordings should be fixed if possible, but otherwise deleted from the analysis.

Routine deletion of outliers from the analysis is not recommended. This practice can have a dramatic effect on the fit of the model and the perceived precision of parameter estimates and

predictions. Analysts who routinely omit outliers without cause tend to overstate the significance of their findings and get a false sense of precision in their estimates and predictions. To assess effects of outliers, a data analyst should repeat the analysis holding out the outliers to see whether any substantive conclusions are changed. Very often the only real effect of an outlier is to inflate MSE and hence make p-values a little larger and CIs a little wider than necessary, but without substantively changing conclusions. They can completely mask underlying patterns, however.

Influential Observations

Certain data points can play an important role in determining the position of the LS line. These data points may or may not be outliers. For example, the observation with $Y > 45$ in the first plot is an outlier relative to the LS fit. The extreme observation in the second plot has a very small r_i. Both points are highly **influential observations** - the LS line changes dramatically when these observations are held out. The influential observation in the second plot is not an outlier because its presence in the analysis determines that the LS line will essentially pass through it! In these plots the solid line is the LS line from the full data set, whereas the dashed line is the LS line after omitting the unusual point.



Dennis Cook developed a measure of the impact that individual cases have on the placement of the LS line. His measure, called **Cook's distance** or Cook's D, provides a summary of how far the LS line changes when each individual point is held out (one at a time) from the analysis. The case with the largest D has the greatest impact on the placement of the LS line. However, the actual influence of this case may be small. In the plots above, the observations I focussed on have the largest Cook's Ds.

A simple, but not unique, expression for Cook's distance for the j^{th} case is

$$D_j \propto \sum_i (\hat{Y}_i - \hat{Y}_{i[-j]})^2,$$

where $\hat{Y}_{i[-j]}$ is the fitted value for the i^{th} case when the LS line is computed from all the data except case j. Here $\propto$ means that D_j is a multiple of $\sum_i (\hat{Y}_i - \hat{Y}_{i[-j]})^2$ where the multiplier does

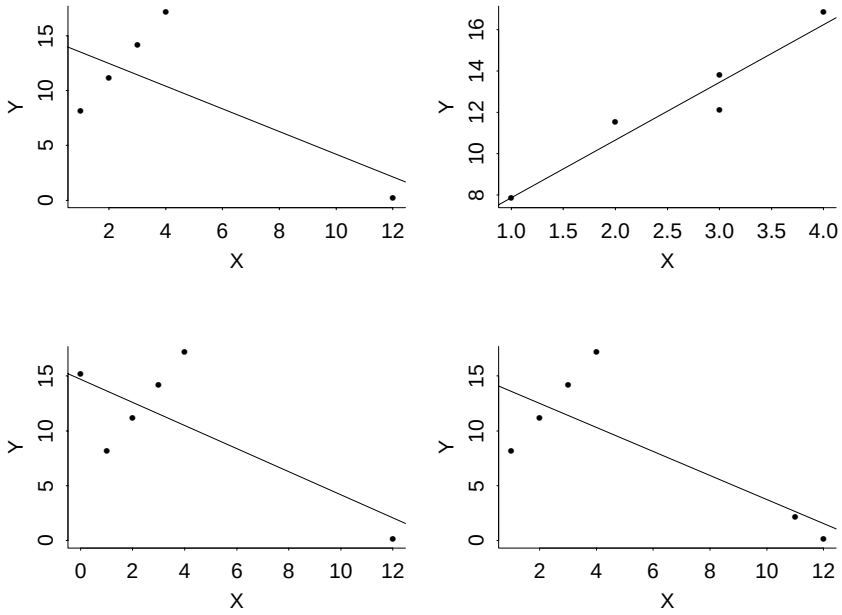
not depend on the case. This expression implies that D_j is also an overall measure of how much the fitted values change when case j is deleted.

Given a regression problem, you should locate the points with the largest D_js and see whether holding these cases out has a decisive influence on the fit of the model or the conclusions of the analysis. You can examine the relative magnitudes of the D_js across cases without paying much attention to the actual value of D_j, but there are guidelines (see below) on how large D_j needs to be before you worry about it.

It is difficult to define a good strategy for dealing with outliers and influential observations. Experience is the best guide. I will show you a few examples that highlight some standard phenomena. One difficulty you will find is that certain observations may be outliers because other observations are influential, or vice-versa. If an influential observation is held out, an outlier may remain an outlier, may become influential, or both, or neither. Observations of moderate influence may become more, or less influential, when the most influential observation is held out.

Thus, any sequential refitting of models holding out of observations should not be based on the original (full-data) summaries, but rather on the summaries that result as individual observations are omitted. I tend to focus more on influential observations than outliers.

In the plots below, which cases do you think are most influential, and which are outliers. What happens in each analysis if I delete the most influential case? Are any of the remaining cases influential or poorly fitted?



Many researchers are hesitant to delete points from an analysis. I think this view is myopic, and in certain instances, such as the Gesell example to be discussed, can not be empirically supported. Being rigid about this can lead to some silly analyses of data, but one needs a very good reason and full disclosure if any points are deleted.

Diagnostic Measures available in Minitab

`Stat > Regression > Regression > Storage` allows you to save several diagnostic measures designed to measure the effect of the i^{th} observation on the regression equation. These are

1. **Residuals**. Minitab's description is

   Residuals

   The difference between the observed values and predicted or fitted values. The residual is the part of the observation that is not explained by the fitted model. You can analyze residuals to determine the adequacy of the model.

   These are just the raw residuals we defined initially. A large residual identifies an observation poorly fit by the model.

2. **Standard residuals**. Minitab's description is

   Standardized residuals

   Also known as the Studentized residual or internally Studentized residual. The standardized residual is the residual, e_i, divided by an estimate of its standard deviation. This form of the residual takes into account that the residuals may have different variances, which can make it easier to detect outliers. Standardized residuals greater than 2 and less than -2 are usually considered large and Minitab labels these observations with an R in the table of unusual observations or fits and residuals.

   It is possible to have quite different standard errors for residuals, and this puts everything on the same scale.

3. **Deleted t residuals**. Minitab's description is

   Studentized deleted residuals

   Also called externally Studentized residual. Studentized deleted residuals are useful for identifying outliers because the i^{th} residual is calculated when the fitted regression is based on all of the cases except the i^{th} one. The residual is then divided by its estimated standard deviation. Since the Studentized deleted residual for the i^{th} observation estimates all quantities with this observation deleted from the data set, the i^{th} observation cannot influence these estimates. Therefore, unusual Y values clearly stand out. Studentized deleted residuals with large absolute values are considered large.

   If the regression model is appropriate, with no outlying observations, each Studentized deleted residual follows the t distribution with $n-1-p$ degrees of freedom.

The problem with residuals is that a highly influential value can force the residual to have a very small value. This measure tries to correct for that by looking at how well the model fits this observation without using this observation to construct the fit. It is quite possible for the deleted residual to be huge when the raw residual is tiny.

4. **Hi (leverages)**. Minitab's description is

   Leverages

   Identify observations with unusual or outlying x-values. Observations with large leverage may exert considerable influence on the fitted value and the model. Leverage values fall between 0 and 1. Experts consider a leverage value greater than 2p/n or 3p/n, where p is the number of predictors or factors plus the constant and n is the number of observations, large and suggest you examine the corresponding observation. Minitab identifies observations with leverage over 3p/n or .99, whichever is smaller, with an X in the table of unusual observations.

   High leverage values are basically outliers among the x values. Y does not enter in this calculation. These are values with the *potential* of greatly distorting the fitted model. They may or may not actually have distorted it.

5. **Cook's distance**. Minitab's description is

   Cook's distance (D)

   An overall measure of the combined impact of each observation on the fitted values. Observations with large D values may be outliers. Because D is calculated using leverage values and standardized residuals, it considers whether an observation is unusual with respect to both x- and y-values.

   To interpret D, compare it to the F-distribution with (p, n-p) degrees of freedom to determine the corresponding percentile. If the percentile value is less than 10% or 20%, the observation has little influence on the fitted values. If the percentile value is greater than 50%, the observation has a major influence on the fitted values and should be examined.

   Many statisticians make it a lot simpler than this sounds and use 1 as a cutoff value for large Cook's D. That's not much different than the Minitab recommendation. This is a very nice hybrid measure incorporating both x- and y- values. High leverage values indicate observations that have the *potential* of causing trouble, but those with high Cook's D values actually *do* disproportionately affect the overall fit. Using the cutoff of 1 can simplify an analysis, since frequently one or two values will have noticeably larger D values than other observations without actually having much effect, but it can be important to explore any observations that stand out.

6. **DFITS**. Minitab's description is

DFITS Also DFFITS.

A measure of the influence of each observation on the fitted value. Represents roughly the number of standard deviations that the fitted value changes when each case is removed from the data set. Observations with DFITS values greater than $2\sqrt{p/n}$ *are considered large and should be examined, where p is the number of predictors (including the constant) and n is the number of observations.*

These various measures often flag the same observations as unusual, but they certainly can flag different observations. At the very least I examine standardized residuals and Cooks's D values. They are invaluable diagnostic measures, but nothing is perfect. Observations can be unusual in groups – a pair of unusual high leverage values close to each other will not necessarily be flagged by Cook's D since removing just one may not affect the fit very much. Any analysis takes some careful thought.

These measures and techniques really are designed for multiple regression problems where several predictor variables are used. We are using them in simple linear regression to learn what they do and see what they have to say about data, but in truth it is fairly simple with one variable to see what may be an outlier in the x direction, to see if the data are poorly fit, etc. With more variables all that becomes quite difficult and these diagnostics are essential parts of those analyses.

Minitab Regression Analysis

There are a lot options allowed in Minitab. I will make a few suggestions here on how to start an analysis. What you find once you get started determines what more you might wish to explore.

1. **Plot the data.** With lots of variables the matrix plot is valuable as a quick screen. If you want to see better resolution on an individual scatter plot, do the individual scatter plot.

2. Do any obvious **transformations** of the data. We will discuss this in a lot more detail later. Re-plot.

3. **Fit the least squares equation**. Here is where you face all the options. At the least ask for the following.

   (a) **Graphs**: Check `Standardized` Residuals (or the `Deleted` Residuals). Standardized are more conventional and show you what actually happened, Deleted are probably the better diagnostic tool for identifying problem cases.

   Residual Plots: Asking for `Four in one` seems convenient to me. This gives you the essential plot (residuals vs. fitted values) plus a quick check on normality and possible violations of independence. If you see something that needs closer investigation, you may need to use the scatterplot menu to plot after the fit is done.

   (b) **Options**: Make sure `Fit intercept` is checked (otherwise no β_0 term is estimated and the results can be odd. This is fit by default, but check. This is where you enter **weights** – to be discussed shortly. In your first pass through the data you probably will not have any weights.

(c) **Results**: The third option is the default and seems right. With larger data sets the last option prints too much (everything can be stored in the worksheet anyway). The nice feature in the third option is automatic flagging of unusual observations.

(d) **Storage**: This is where you tell Minitab what to place in the worksheet so you can examine it in more detail. I usually ask at least for `Standardized residuals`, `Hi (leverages)`, and `Cook's distance`. These get named `SRES`, `HI`, and `COOK` in the worksheet.

4. **Examine the residual plots and results.** Check for the patterns discussed earlier.

   (a) Do you see curvature? If the sign of the residuals has a distinct pattern vs. the fitted values, the linear fit is not adequate and you need some remedy such as transformations.

   (b) Does it appear $\sigma_{Y|X}$ depends upon X (we are assuming it does not)? A megaphone pattern in residuals vs. fits is the classic (not the only) pattern to look for. Weighted least squares or transformations may be called for.

   (c) Do you see obvious outliers? Make sure you do not have a misrecorded data value. It might be worth refitting the equation without the outlier to see if it affects conclusions substantially.

   (d) Is the normality assumption reasonable? This can be very closely related to the preceding points.

   (e) Is there a striking pattern in residuals vs. order of the data? This can be an indication that the independence assumption is not valid.

5. **Check the Cook's D values.** Minitab has already flagged for you any cases with high leverage or standardized residuals. D is very useful to check separately. A nice check for high D values is a plot vs. order of the data using a `Graph > Time Series Plot`. Instead of `Connect line` under `Data View > Data Display`, I prefer `Project lines`. It can be useful to ask for a reference line at $Y = 1$ under `Time/Scale > Reference lines > Show reference lines for Y positions`. The same plot for residuals and leverage values can be useful (changing reference lines) as well.

6. If you found problem observations, omit them from the analysis and see if any conclusions change substantially. There are two good ways to do this.

   (a) Subset the worksheet using `Data > Subset Worksheet` and indicate which rows to include or exclude. This creates a new worksheet. In earlier versions of Minitab and in some other packages I found this awkward since worksheets could proliferate and specifying rows could be cumbersome. With the ability to *brush* plots and subset, this may be the best method.

   (b) Use weighted least squares with weights of 0 for the excluded observations, weights of 1 for those included. I used to prefer this method to the preceding one; I still find it as easy but it is less modern (so it works in other packages that do not support brushing). The weight variable gets entered under `Options`.

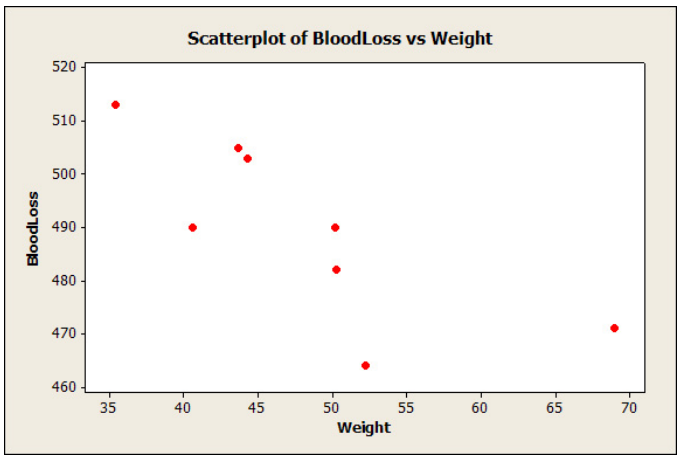
You may need to repeat all these steps many times for a complete analysis.

Residual and Diagnostic Analysis of the Blood Loss Data

We looked at much of this before, but let us go through the above steps systematically. Recall the data set (we want to predict blood loss from weight):

```
Data Display

Row  Weight  Time  BloodLoss
  1    44.3   105        503
  2    40.6    80        490
  3    69.0    86        471
  4    43.7   112        505
  5    50.3   109        482
  6    50.2   100        490
  7    35.4    96        513
  8    52.2   120        464
```

1. Plot blood loss vs. weight.



   Clearly the heaviest individual is an unusual value that warrants a closer look. Passing the mouse over that point in Minitab reveals it is the 3^{rd} observation. I might be inclined to try a transformation here to make that point a little less influential.

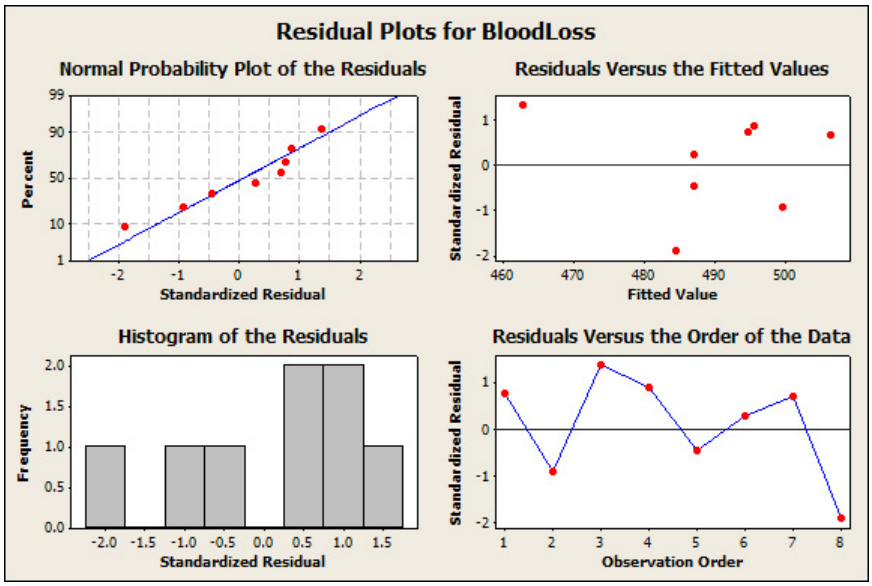
2. We will look at transformations later.

3. Results are in many pieces. The summary of the least squares fit put in the Session Window is

```
Regression Analysis: BloodLoss versus Weight

The regression equation is
BloodLoss = 552 - 1.30 Weight


Predictor      Coef  SE Coef      T      P
Constant    552.44    21.44  25.77  0.000
Weight     -1.3003   0.4364  -2.98  0.025


S = 11.6623   R-Sq = 59.7%   R-Sq(adj) = 52.9%
```

```
Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  1207.5  1207.5  8.88  0.025
Residual Error    6   816.0   136.0
Total             7  2023.5
```

The residual plots:

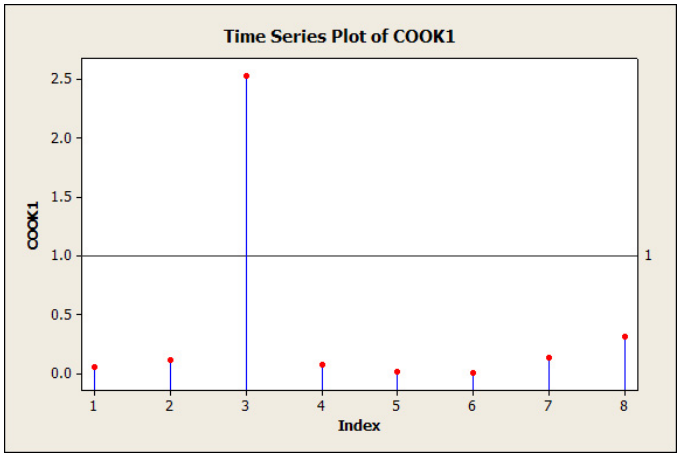


Note the new entries in the worksheet.

```
Data Display

Row  Weight  BloodLoss     SRES1       HI1     COOK1
  1    44.3        503   0.75757  0.146436  0.04923
  2    40.6        490  -0.92858  0.206150  0.11196
  3    69.0        471   1.36675  0.730118  2.52679
  4    43.7        505   0.87441  0.153515  0.06933
  5    50.3        482  -0.46321  0.131102  0.01619
  6    50.2        490   0.26065  0.130532  0.00510
  7    35.4        513   0.70348  0.354881  0.13612
  8    52.2        464  -1.90958  0.147266  0.31487
```

4. What do we see so far? Blood Loss appears significantly negatively associated with weight, and except for one very isolated residual (for obs. 3), nothing looks terribly out of whack in the residual plots.

5. Look at Cook's D. We could have anticipated this, but the 3^{rd} observation is affecting the fit by a lot more than any other values. The D-value is much larger than 1. Note that the residual is not large for this value.



6. Let us refit the equation without observation 3 to see if anything changes drastically. I will use the weighted least squares approach discussed earlier on this example. Define a variable `wt` that is 1 for all observations except obs. 3, and make it 0 for that one.

```
Data Display

wt
   1   1   0   1   1   1   1   1
```

What changes by deleting case 3? The fitted line gets steeper (slope changes from -1.30 to -2.19), adjusted R^2 gets larger (up to 58% from 53%), and S changes from 11.7 to 10.6. Because the Weight values are much less spread out, $SE(\hat{\beta}_1)$ becomes quite a bit larger (to .714, up from .436) and we lose a degree of freedom for MS Error (which will penalize us on tests and CIs). Just about any quantitative statement we would want to make using CIs would be about the same either way since CIs will overlap a great deal, and our qualitative interpretations are unchanged (Blood Loss drops with Weight). Unless something shows up in the plots, I don't see any very important changes here.

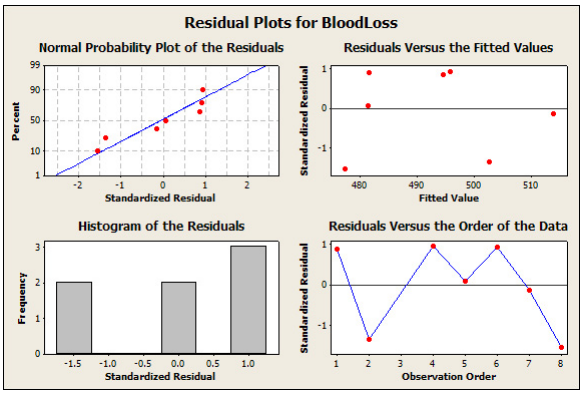
```
Regression Analysis: BloodLoss versus Weight

Weighted analysis using weights in wt


The regression equation is
BloodLoss = 592 - 2.19 Weight


7 cases used, 1 cases contain missing values
                           or had zero weight


Predictor      Coef  SE Coef      T      P
Constant     591.67    32.57  18.17  0.000
Weight     -2.1935   0.7144  -3.07  0.028


S = 10.6017   R-Sq = 65.3%   R-Sq(adj) = 58.4%


Analysis of Variance
```

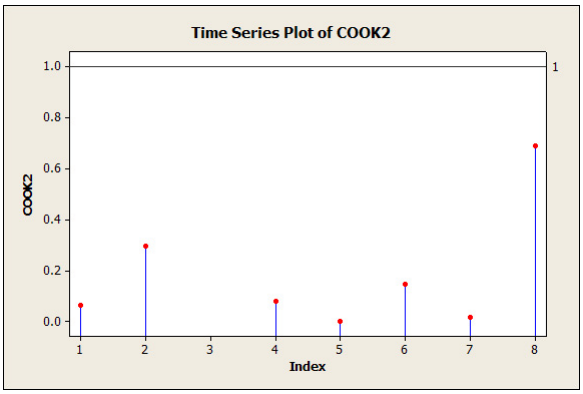
```
Source           DF      SS      MS     F      P
Regression        1  1059.7  1059.7  9.43  0.028
Residual Error    5   562.0   112.4
Total             6  1621.7
```

Note that the new diagnostic measures are added to the data set,

```
Data Display

Row  Weight  Time  BloodLoss  wt     SRES1      HI1      COOK1     SRES2      HI2      COOK2
  1    44.3   105        503   1   0.75757  0.146436  0.04923   0.86838  0.146893  0.064921
  2    40.6    80        490   1  -0.92858  0.206150  0.11196  -1.36530  0.240725  0.295492
  3    69.0    86        471   0   1.36675  0.730118  2.52679         *  0.000000         *
  4    43.7   112        505   1   0.87441  0.153515  0.06933   0.94197  0.153665  0.080551
  5    50.3   109        482   1  -0.46321  0.131102  0.01619   0.07277  0.258970  0.000925
  6    50.2   100        490   1   0.26065  0.130532  0.00510   0.92250  0.254423  0.145200
  7    35.4    96        513   1   0.70348  0.354881  0.13612  -0.14874  0.582715  0.015447
  8    52.2   120        464   1  -1.90958  0.147266  0.31487  -1.55578  0.362609  0.688491
```

Nothing very striking shows up in the residual plots, and no Cook's D values are very large among the remaining observations.





How much difference is there in a practical sense? Examine the 95% prediction interval for a new observation at Weight = 50kg. Previously we saw that interval based on all 8 observations was from 457.1 to 517.8 ml of Blood Loss. Based on just the 7 observations the prediction interval is 451.6 to 512.4 ml. There really is no practical difference here.

Gesell data

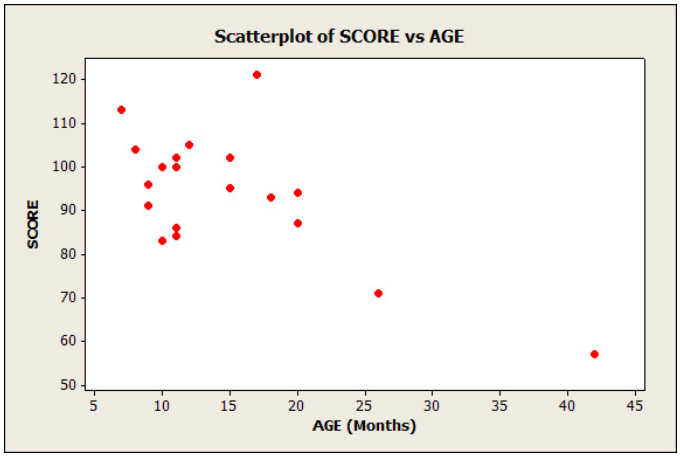
These data are from a UCLA study of cyanotic heart disease in children. The predictor is the age of the child in months at first word and the response variable is the Gesell adaptive score, for each of 21 children.

```
Data Display

Row  ID  AGE  SCORE
  1   1   15     95
  2   2   26     71
  3   3   10     83
  4   4    9     91
  5   5   15    102
  6   6   20     87
```

7	7	18	93
8	8	11	100
9	9	8	104
10	10	20	94
11	11	7	113
12	12	9	96
13	13	10	83
14	14	11	84
15	15	11	102
16	16	10	100
17	17	12	105
18	18	42	57
19	19	17	121
20	20	11	86
21	21	10	100

Let us go through the same steps as before.

1. Plot Score versus Age. Comment on the relationship between Score and Age.



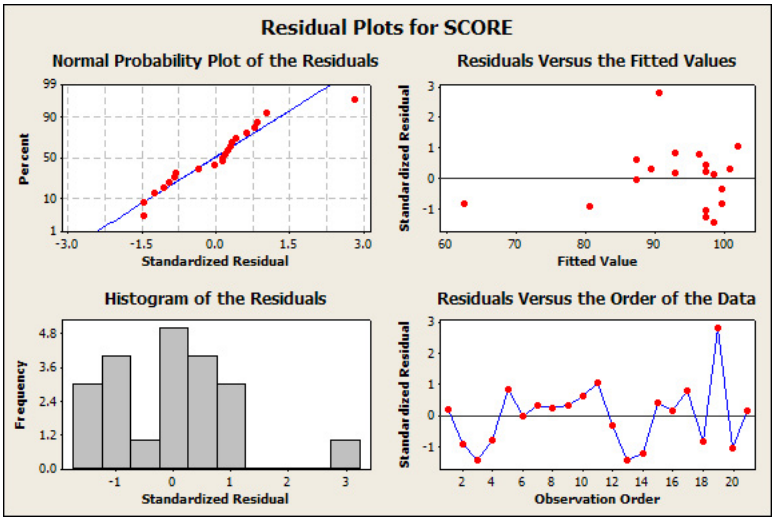
2. There are no obvious transformations to try here.

3. Fit a simple linear regression model. Provide an equation for the LS line. Does age at first word appear to be an "important predictor" of Gesell adaptive score? (i.e. is the estimated slope significantly different from zero?)

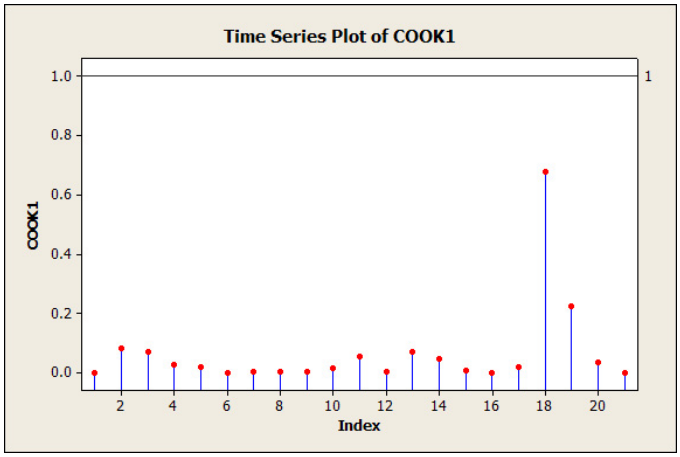
```
Regression Analysis: SCORE versus AGE

The regression equation is
SCORE = 110 - 1.13 AGE

Predictor      Coef  SE Coef      T      P
Constant    109.874    5.068  21.68  0.000
AGE         -1.1270   0.3102  -3.63  0.002

S = 11.0229   R-Sq = 41.0%   R-Sq(adj) = 37.9%

Analysis of Variance
```

```
Source           DF      SS      MS      F      P
Regression        1  1604.1  1604.1  13.20  0.002
Residual Error   19  2308.6   121.5
Total            20  3912.7

Unusual Observations

Obs    AGE    SCORE     Fit  SE Fit  Residual  St Resid
 18   42.0    57.00   62.54    8.90     -5.54     -0.85 X
 19   17.0   121.00   90.72    2.54     30.28      2.82R

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

4. Do these plots suggest any inadequacies with the model?



5. Observations 18 and 19 stand out with relatively high Cook's D. The cutoff line is only a rough guideline. Those two were flagged with high influence and standardized residual, respectively, also.



Be sure to examine the scatter plot carefully to see why 18 and 19 stand out.

6. Consider doing two additional analyses: Analyze the data after omitting case 18 only and analyze the data after omitting case 19 only. Refit the regression model for each of these two scenarios. Provide a summary table such as the following, giving the relevant summary statistics for the three analyses. Discuss the impact that observations 18 and 19 have individually on the fit of the model. I will demonstrate doing this in class by brushing the scatter plot to subset the data.

   When observation 18 is omitted, the estimated slope is not significantly different from zero (p-value = .1489), indicating that age is not an important predictor of Gesell score. This suggests that the significance of age as a predictor in the original analysis was due solely to the presence of observation 18. Note the dramatic decrease in R^2 after deleting observation 18.

   The fit of the model appears to improve when observation 19 is omitted. For example, R^2 increases noticeably and the p-value for testing the significance of the slope decreases dramatically (in a relative sense). These tendencies would be expected based on the original plot. However, this improvement is misleading. Once observation 19 is omitted, observation 18 is much more influential. Again the significance of the slope is due to the presence of observation 18.

Feature	Full data	Omit 18	Omit 19
b_0	109.87	105.63	109.30
b_1	-1.13	-0.78	-1.19
$SE(b_0)$	5.07	7.16	3.97
$SE(b_1)$	0.31	0.52	0.24
R^2	0.41	0.11	0.57
p-val for $H_0 : \beta_1 = 0$	0.002	0.149	0.000

Can you think of any reasons to justify doing the analysis without observation 18?

If you include observation 18 in the analysis, you are assuming that the mean Gesell score is linearly related to age over the entire range of observed ages. Observation 18 is far from the other observations on age (age for observation 18 is 42; the second highest age is 26; the lowest age is 7). There are no children with ages between 27 and 41, so we have no information on whether the relationship is roughly linear over a significant portion of the range of ages. I am comfortable deleting observation 18 from the analysis because it's inclusion forces me to make an assumption that I can not check using these data. I am only willing to make predictions of Gesell score for children with ages roughly between 7 and 26. However, once this point is omitted, age does not appear to be an important predictor.

A more complete analysis would delete observation 18 and 19 together. What would you expect to see if you did this?

Weighted Least Squares

Earlier I indicated that nonconstant error variance can be addressed (sometimes) with weighted least squares. Recall the LS (OLS or ordinary LS) line chooses the values of β_0 and β_1 that minimize

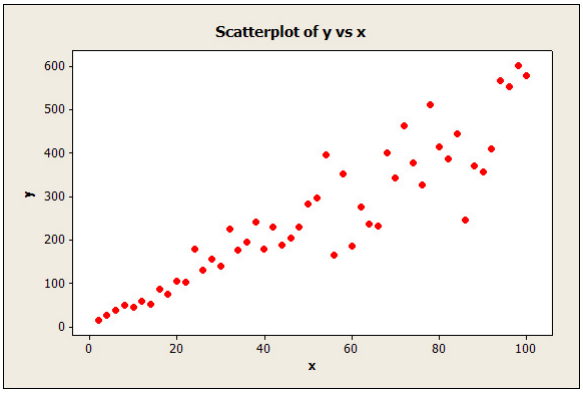
$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

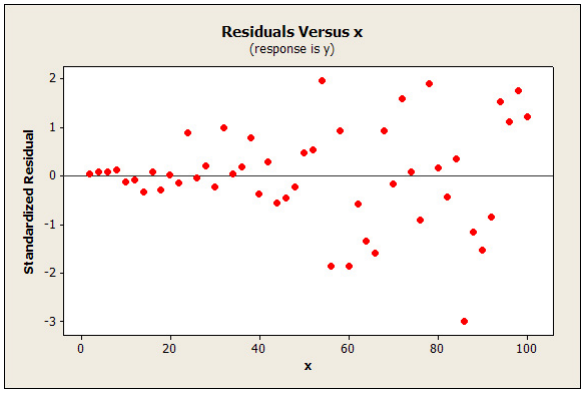
over all possible choices of β_0 and β_1. The weighted LS (WLS) line chooses the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n} w_i\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

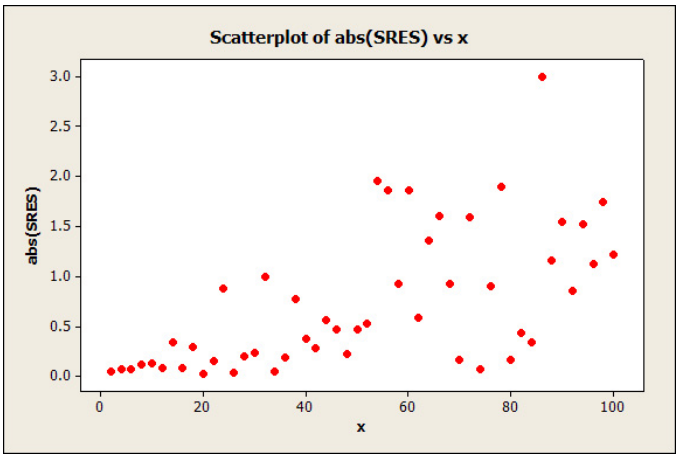
over all possible choices of β_0 and β_1. If $\sigma_{Y|X}$ depends up X, then the correct choice of weights is inversely proportional to *variance*, $w_i \propto \sigma^2_{Y|X}$.

Consider the following plot of y vs. x and standardized OLS residuals vs x. It is very clear that variability increases with x.

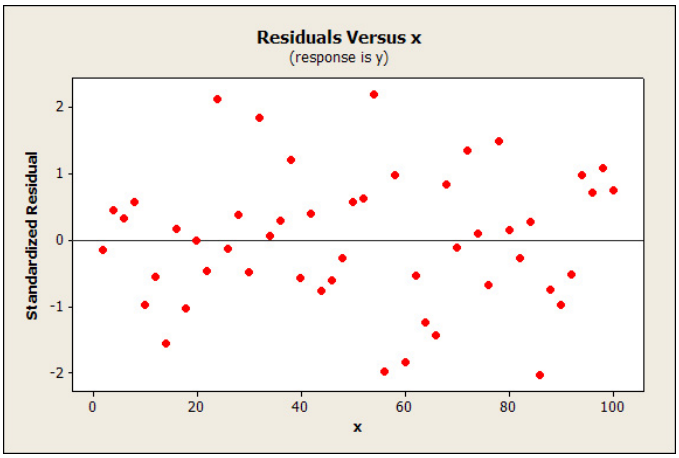




In order to use WLS to solve this problem, we need some form for $\sigma^2_{Y|X}$. Finding that form is a real problem with WLS. It can be useful to plot the absolute value of the standardized residual vs. x to see if the top boundary seems to follow a general pattern.



It is plausible the upper boundary is linear, so let us try $w_i = \frac{1}{x^2}$. Standardized residuals from this WLS fit look very good. Note that raw (nonstandardized) residuals will still have the same pattern – it is essential to use standardized residuals here.



Compare also the OLS fitted equation

```
The regression equation is y = 2.5 + 5.01 x

Predictor     Coef  SE Coef      T      P
Constant      2.55    18.37   0.14  0.890
x           5.0057   0.3134  15.97  0.000

S = 63.9662   R-Sq = 84.2%   R-Sq(adj) = 83.8%
```

to the WLS fitted equation.

```
Weighted analysis using weights in C7

The regression equation is y = 5.52 + 4.94 x

Predictor     Coef  SE Coef      T      P
Constant     5.523    1.928   2.86  0.006
x           4.9362   0.1738  28.40  0.000

S = 1.06494   R-Sq = 94.4%   R-Sq(adj) = 94.3%
```

Clearly the weighted fit looks better, although note that everything is based on the weighted SS. In practice it can be pretty difficult to determine the correct set of weights, but WLS works much better than OLS if appropriate. I actually simulated this data set using $\beta_0 = \beta_1 = 5$. Which fit actually did better?