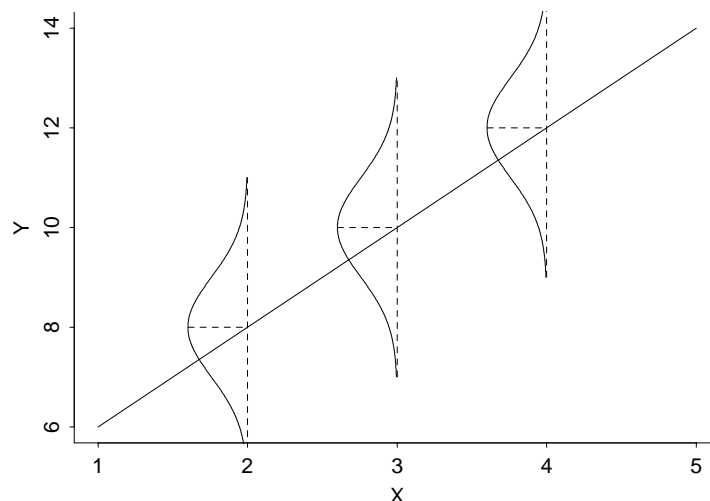


2 The Linear Regression Model

The following statistical model is assumed as a means to provide error estimates for the LS line, regression coefficients, and predictions. Assume that the data (X_i, Y_i) , $i = 1, \dots, n$ are a sample of (X, Y) values from the population of interest, and

Visual representation of regression model with population regression line



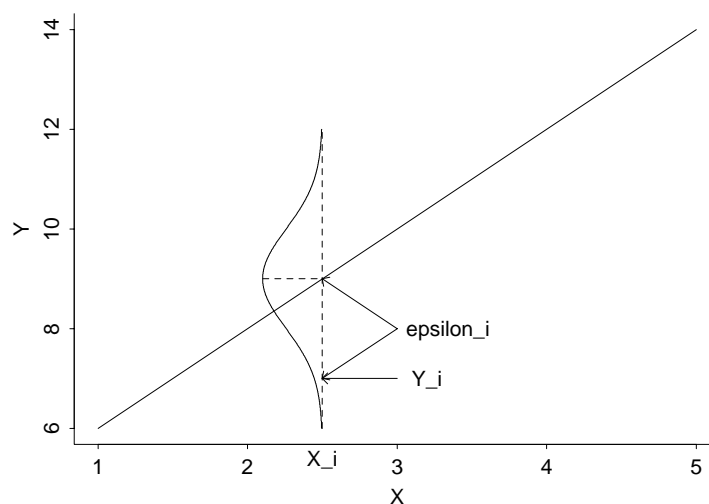
1. The mean in the population of all responses Y at a given X value (called $\mu_{Y|X}$ by SW) falls on a straight line, $\beta_0 + \beta_1 X$, called the population regression line.
2. The variation among responses Y at a given X value is the same for each X , and is denoted by $\sigma_{Y|X}^2$.
3. The population of responses Y at a given X is normally distributed.
4. The pairs (X_i, Y_i) are a random sample from the population. Alternatively, we can think that the X_i s were fixed by the experimenter, and that the Y_i are random responses at the selected predictor values.

The model is usually written in the form

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

(i.e. Response = Mean Response + Residual), where the ϵ_i s are, by virtue of assumptions 2, 3 and 4, independent normal random variables with mean 0 and variance $\sigma_{Y|X}^2$. The picture below might help you visualize this. Note that the population regression line is unknown, and is estimated from the data using the LS line.

Visual representation of population regression model notation



Back to the Data

There are three unknown population parameters in the model: β_0 , β_1 and $\sigma_{Y|X}^2$. Given the data, the LS line

$$\hat{Y} = b_0 + b_1 X$$

estimates the population regression line $\beta_0 + \beta_1 X$. The LS line is our best guess about the unknown population regression line. Here b_0 estimates the intercept β_0 of the population regression line and b_1 estimates the slope β_1 of the population regression line.

The i^{th} **observed residual** $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 X_i$ is the i^{th} **fitted value**, estimates the **unobservable residual** ϵ_i . (ϵ_i is unobservable because β_0 and β_1 are unknown.) The Residual MS from the ANOVA table is used to estimate $\sigma_{Y|X}^2$:

$$s_{Y|X}^2 = \text{Res MS} = \frac{\text{Res SS}}{\text{Res df}} = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n - 2}.$$

CI and Tests for β_1

A CI for β_1 is given $b_1 \pm t_{crit} SE_{b_1}$, where the standard error of b_1 under the model is

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum_i (X_i - \bar{X})^2}},$$

and where t_{crit} is the appropriate critical value for the desired CI level from a t -distribution with $df = \text{Res } df$.

To test $H_0 : \beta_1 = \beta_{1,0}$ (a given value) against $H_A : \beta_1 \neq \beta_{1,0}$, reject H_0 if $|t_s| \geq t_{crit}$, where

$$t_s = \frac{b_1 - \beta_{1,0}}{SE_{b_1}},$$

and t_{crit} is the t -critical value for a two-sided test, with the desired size and $df = \text{Res } df$. Alternatively, you can evaluate a p-value in the usual manner to make a decision about H_0 .

The parameter estimates table in **Stata** gives the standard error, t -statistic, p-value for testing $H_0 : \beta_1 = 0$, and a 95% CI for β_1 . Analogous summaries are given for the intercept, but these are typically of less interest.

Testing $\beta_1 = 0$

Assuming the mean relationship is linear, consider testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. This test can be conducted using a t -statistic, as outlined above, or with an ANOVA F -test, as outlined below.

For the analysis of variance (ANOVA) F -test, compute

$$F_s = \frac{\text{Reg MS}}{\text{Res MS}}$$

and reject H_0 when F_s exceeds the critical value (for the desired size test) from an F -table with numerator $df = 1$ and denominator $df = n - 2$; see SW, page 654. The hypothesis of zero slope (or no relationship) is rejected when F_s is large, which happens when a significant portion of the variation in Y is explained by the linear relationship with X . **Stata** gives the F -statistic and p-value with the ANOVA table output.

The p-values from the t -test and the F -test are always equal. Furthermore this p-value is equal to the p-value for testing no correlation between Y and X , using the t -test described earlier. Is this important, obvious, or disconcerting?

A CI for the Population Regression Line

I can not overemphasize the **power** of the regression model. The model allows you to estimate the mean response at any X value in the range for which the model is reasonable, even if little or no data is observed at that location.

We estimate the mean population response among individuals with $X = X_p$

$$\mu_p = \beta_0 + \beta_1 X_p,$$

with the fitted value, or the value of the least squares line at X_p :

$$\hat{Y}_p = b_0 + b_1 X_p.$$

X_p is not necessarily one of the observed X_i s in the data. To get a CI for μ_p , use $\hat{Y}_p \pm t_{crit} SE(\hat{Y}_p)$, where the standard error of \hat{Y}_p is

$$SE(\hat{Y}_p) = s_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}}.$$

The t -critical value is identical to that used in the subsection on CI for β_1 .

CI for Predictions

Suppose a future individual (i.e. someone not used to compute the LS line) has $X = X_p$. The best prediction for the response Y of this individual is the value of the least squares line at X_p :

$$\hat{Y}_p = b_0 + b_1 X_p.$$

To get a CI (prediction interval) for an individual response, use $\hat{Y}_p \pm t_{crit} SE_{pred}(\hat{Y}_p)$, where

$$SE_{pred}(\hat{Y}_p) = s_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}},$$

and t_{crit} is identical to the critical value used for a CI on β_1 .

For example, in the blood loss problem you may want to estimate the blood loss for an 50kg individual, and to get a CI for this prediction. This problem is different from computing a CI for the mean blood loss of all 50kg individuals!

Comments

1. The prediction interval is wider than the CI for the mean response. This is reasonable because you are less confident in predicting an individual response than the mean response for all individuals.
2. The CI for the mean response and the prediction interval for an individual response become wider as X_p moves away from \bar{X} . That is, you get a more sensitive CI and prediction interval for X_p s near the center of the data.

A Further Look at the Blood Loss Data using Stata

We obtain a prediction interval for an individual and confidence intervals for mean blood loss in **Stata** as follows (but note that there are a lot of ways to do this). In a separate **Stata** data set we create a variable that contains the weight values at which we would like to predict blood loss. This is done either with the **input** command or (preferably) using the data editor, an Excel-like spreadsheet utility. We illustrate the use of **input**. We desire predictions at weights of 30, 35, 40, 45, 50, 55, 60, 65, 70, and 75 kg. Examine the following **Stata** code.

```
clear
input weight
30
35
40
45
50
55
60
65
70
75
end
save weight.dta
use bloodloss
append using weight
regress loss weight
predict loss_hat,xb
predict se_line, stdp
predict se_pred, stdf
```

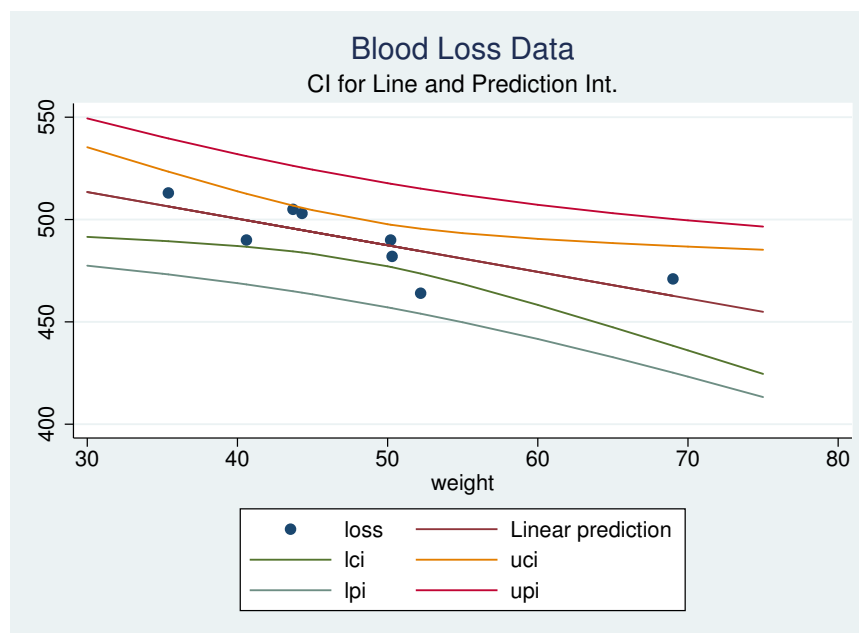
```

generate lci=loss_hat+invttail(6,0.025)*se_line
generate uci=loss_hat+invttail(6,0.025)*se_line
generate lpi=loss_hat+invttail(6,0.025)*se_pred
generate upi=loss_hat+invttail(6,0.025)*se_pred
graph twoway (scatter loss weight) (line loss_hat weight) ///
              (line lci weight,sort)(line uci weight,sort) ///
              (line lpi weight,sort)(line upi weight, sort) ///
              , title(Blood Loss Data) subtitle(CI for Line and Prediction Int.)

```

The above commands create a new data set called **weight**, append those weight values to the bloodloss data set (leaving values of weight and time missing) perform regression using only the original data set (cases with missing values of X or Y are discarded), and then save the predicted values \hat{Y}_p (fitted values on the regression line) for each value of the variable **weight** as well as the standard errors for the fitted line, $SE(\hat{Y}_p)$, and standard errors for prediction, $SE_{pred}(\hat{Y}_p)$. The confidence interval for the line and prediction interval is computed and plotted. After this program is run (from a do-file) the data set looks as follows.

. list, clean	weight	time	loss	loss_hat	se_line	se_pred	lci	uci	lpi	upi
1.	44.3	105	503	494.8375	4.46279	12.48698	483.9175	505.7576	464.283	525.392
2.	40.6	80	490	499.6487	5.295104	12.80805	486.6921	512.6054	468.3086	530.9889
3.	69	86	471	482.7195	9.965039	15.33982	438.3359	487.103	425.1843	500.2546
4.	43.7	112	505	495.6177	4.569383	12.52547	484.4369	506.7986	464.969	526.2665
5.	50.3	109	482	487.0356	4.222673	12.40319	476.703	497.3681	456.686	517.3851
6.	50.2	100	490	487.1656	4.213473	12.40006	476.8556	497.4756	456.8237	517.5074
7.	35.4	96	513	506.4104	6.947425	13.57479	489.4107	523.4102	473.1941	539.6267
8.	52.2	120	464	484.5649	4.475415	12.4915	473.614	495.5159	453.9994	515.1306
9.	30	.	.	513.4322	8.954061	14.70317	491.5224	535.342	477.4548	549.4095
10.	35	.	.	506.9306	7.088681	13.64762	489.5852	524.2759	473.536	540.3251
11.	40	.	.	500.4289	5.463197	12.87846	487.061	513.7969	468.9165	531.9413
12.	45	.	.	493.9273	4.355063	12.44888	483.2708	504.5838	463.466	524.3886
13.	50	.	.	487.4257	4.196375	12.39426	477.1575	497.6938	457.098	517.7533
14.	55	.	.	480.924	5.076955	12.71942	468.5012	493.3469	449.8008	512.0474
15.	60	.	.	474.4224	6.592747	13.39673	458.2905	490.5543	441.6418	507.203
16.	65	.	.	467.9207	8.406906	14.37652	447.3498	488.4917	432.7427	503.0988
17.	70	.	.	461.4191	10.36392	15.60189	436.0595	486.7787	423.2427	499.5956
18.	75	.	.	454.9175	12.39631	17.01989	424.5848	485.2502	413.2713	496.5636



Given the model $\text{Blood Loss} = \beta_0 + \beta_1 \text{Weight} + \epsilon$:

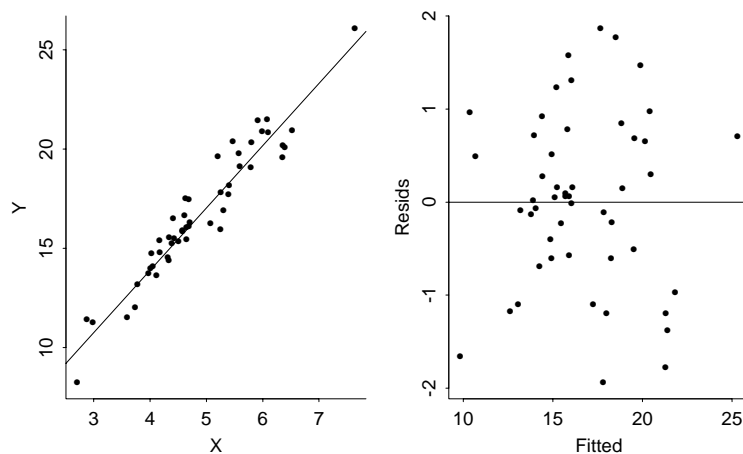
- The LS line is: Predicted Blood Loss = 552.442 - 1.30 Weight.
- The R^2 is .597 (i.e. 59.7%); see Lecture 1.
- The F -statistic for testing $H_0 : \beta_1 = 0$ is $F_{obs} = 8.88$ with a p -value = .0247. The Error MS is $s_{Y|X}^2 = 136.008$; see ANOVA table.
- The Parameter Estimates table gives b_0 and b_1 , their standard errors, t -statistics and p -values for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The t -test and F -test p -values for testing that the slope is zero are identical.
- Prediction and CI: The estimated average blood loss for all 50kg patients is $552.442 - 1.30033 * 50 = 487.43$. We are 95% confident that the mean blood loss of all 50kg patients is between (approximately) 477 and 498 ml. A 95% prediction interval for the blood loss of a single 50 kg person is less precise (about 457 to 518 ml).

As a summary we might say that weight is important for explaining the variation in blood loss. In particular, the estimated slope of the least squares line (Predicted Blood loss = 552.442 - 1.30 Weight) is significantly different from zero (p -value = .0247), with weight explaining approximately 60% (59.7%) of the variation in blood loss for this sample of 8 thyroid operation patients.

Checking the regression model

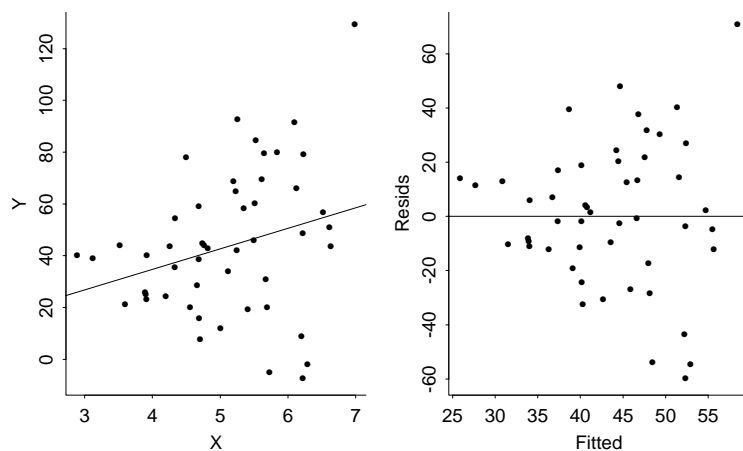
A regression analysis is never complete until the assumptions of the model have been checked. In addition, you need to evaluate whether individual observations, or groups of observations, are unduly influencing the analysis. A first step in any analysis is to plot the data. The plot provides information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

In addition to plotting the data, a variety of methods for checking models are based on plots of the residuals, $e_i = Y_i - \hat{Y}_i$ (i.e. Observed - Fitted). The command `rvpplot` in **Stata** plots the e_i against the predictor values X_i . Alternatively (and equivalently for simple linear regression), the command `rvfplot` plots e_i against the fitted values \hat{Y}_i , as illustrated in the plots below. Regardless of which you use, the residual plot should exhibit no systematic dependence of the sign or the magnitude of the residuals on the fitted values.

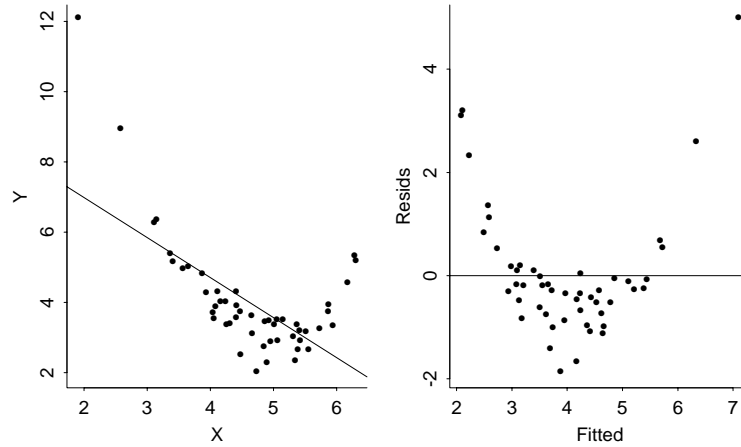


The real power of this plot (e_i against \hat{Y}_i) is with multiple predictor problems (multiple regression). For simple linear regression, the information in this plot is similar to the information in the original data plot, except that the residual plot eliminates the effect of the trend on your perceptions of model adequacy.

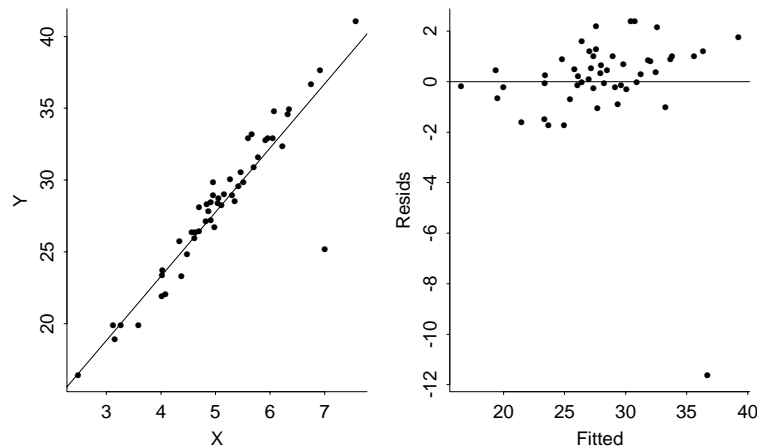
The following plots show how inadequacies in the data plot appear in a residual plot.



The first plot (above) 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 next 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. A possible remedy is to transform the data.



The last plot (above) shows an outlier. This point has a large residual. A sensible approach is to refit the model after deleting the case and see if any conclusions change.

Checking Normality

The normality assumption can be evaluated with a boxplot or a normal quantile plot of the residuals (**Stata** command `graph [residuals], box and qnorm`). A formal test of normality using the residuals

can be based on the Wilk-Shapiro test (discussed in last semester's lab) using the **Stata** command **swilk**.

Checking Independence

Diagnosing dependence among observations usually requires some understanding of the mechanism that generated the data. There are a variety of graphical and inferential tools for checking independence for data collected over time (called a time series). The easiest thing to do is plot the r_i against time index and look for any suggestive patterns.

Outliers

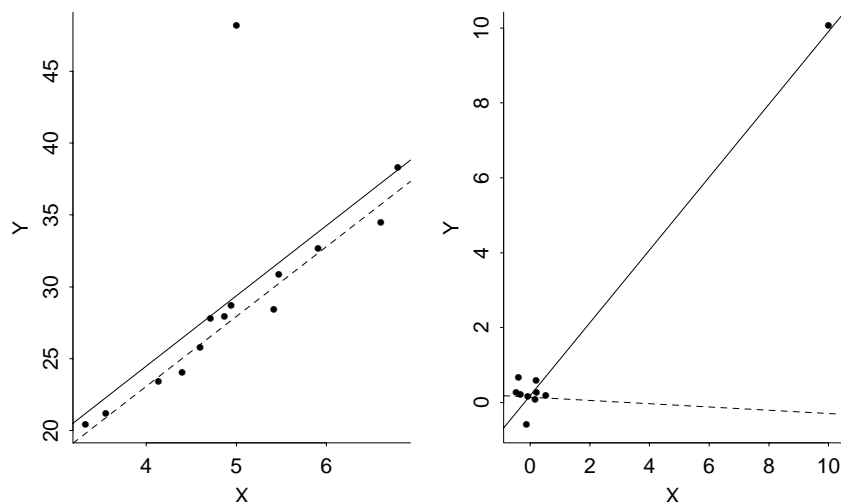
Outliers are observations that are poorly fitted by the regression model. The response for an outlier is far from the fitted line, so outliers have large positive or negative values of the residual e_i .

What do you do with outliers? Outliers may be due to incorrect recordings of the data or failure of the measuring device, or indications of 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. At the very least, a data analyst should repeat the analysis with and without the outliers to see whether any substantive conclusions are changed.

Influential observations

Certain data points can play a very 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 below is an outlier relative to the LS fit. The extreme observation in the second plot has a very small e_i . Both points are highly **influential observations** - the LS line changes dramatically when these observations are deleted. 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.



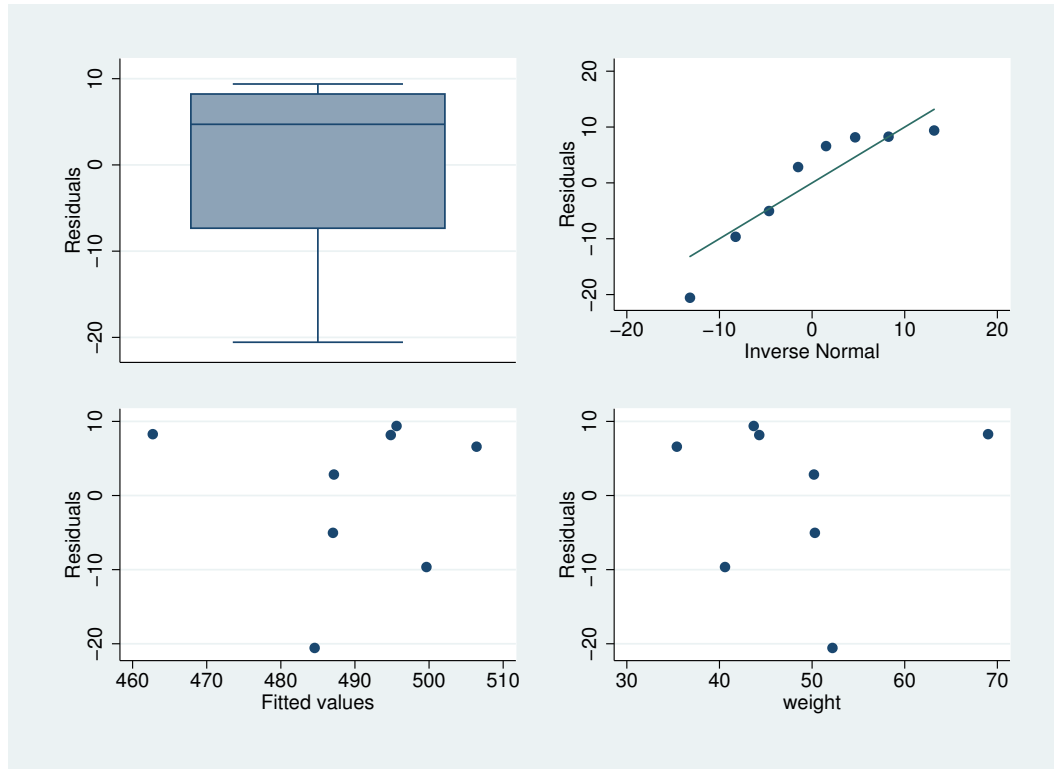
A standard measure of the influence that individual cases have on the LS line is called **Cook's Distance**, which is available as `predict cooksd`, `cooksd` for example. For simple linear regression most influential cases can be easily spotted by carefully looking at the data plot. If you identify cases that you suspect might be influential, you should hold them out (individually) and see if any important conclusions change. If so, you need to think hard about whether the cases should be included or excluded from the analysis. We will obtain and interpret Cook's distances later.

A Final Look at the Blood Loss Data

We create various diagnostic plots and perform the Shapiro-Wilk test of normality on the residuals using the **Stata** commands

```
use bloodloss
regress loss weight
predict res, r
swilk r
graph box r, saving(boxplot)
qnorm r, saving(probplot)
rvfplot, saving(respredplot)
rvpplot weight, saving(resweightplot)
graph combine boxplot.gph probplot.gph respredplot.gph resweightplot.gph, saving(all)
```

Residual plots for the blood loss problem follow. Do we see any marked problems with influential cases, outliers, or non-normality? Also, go back in the notes and look at the data plot.



The results of the Shapiro-Wilk normality test on the residuals:

Shapiro-Wilk W test for normal data					
Variable	Obs	W	V	z	Prob>z
res	8	0.84852	2.110	1.328	0.09204