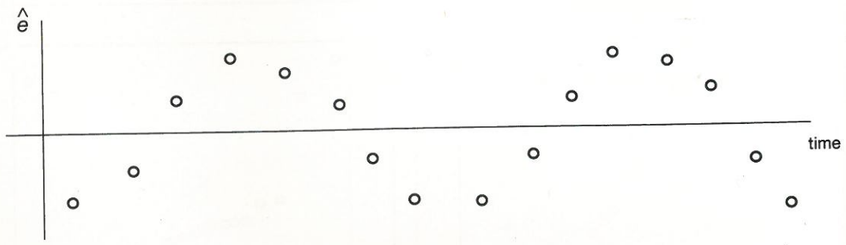
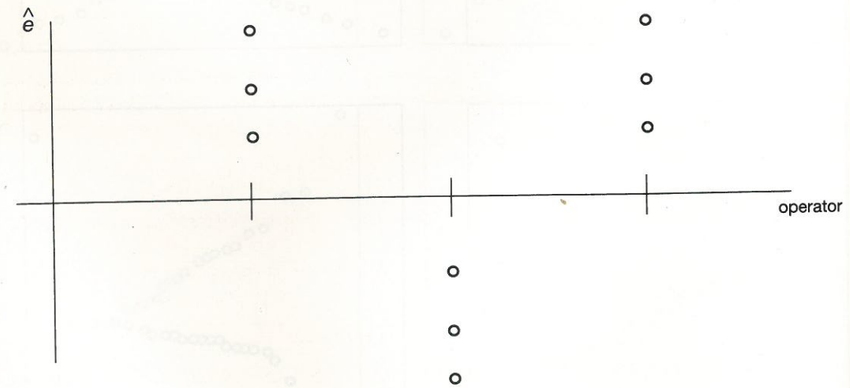
probability plot, that is, the sample correlation between the s_i and $\hat{e}_i$. Atkinson (1981) proposed a simulation method for constructing approximate confidence "envelopes" for the normal probability plot. Using these, the normality assumption is rejected if the normal probability plot fails to lie between its upper and lower envelope. Because the required test statistics, plots, or significance tables are not readily available, none of these tests are extensively used.

Daniel and Wood (1971) give numerous examples of normal probability plots for samples of sizes 8, 16, 32, 64, and 384 from a normal distribution. These provide additional training sets for learning to interpret normal probability plots. One should be alerted to the fact that many statistical programs, including those in BMDP, reverse things and plot the



(a) Residuals showing seasonal trend



(b) Residuals showing an operator effect

Figure 3.10: Plots of residuals against external variables.

normal scores s on the fitted values $\hat{e}$. Since both plotting methods are used, one must read the labels on a normal probability plot carefully to avoid misinterpretations.

Figure 1.8 shows a residual normal probability plot for the Grade Data in Chapter 1. It is reasonably linear and shows no significant departure from normality. The SAS specifications for normal probability plotting are given in Figure 1.3. There the RANK and PLOT procedures are used. The UNIVARIATE procedure can also be used to give normal probability plots.

3.3.3 Plots on External Variables

It is often informative to plot $\hat{e}$ and y against external variables that are not part of the fitted regression model. Anscombe and Tukey (1963) mention time and geographic location (position in a field, height above floor, latitude, etc.) as prime examples of such external variables. When plotting against external variables, one is looking for hidden effects. Joiner (1981) has given a striking example in which a simple linear regression of y on the order i in which the data was listed gave a significantly better fit than that in two previously published analyses using what seemed to be natural and appropriate independent variables (Problem 13, Chapter 4).

Figure 3.10(a) illustrates a "seasonal trend" that one might discover by plotting the fitted residual $\hat{e}$ against time. Figure 3.10(b) shows how a residual plot for the output of a machine against the operator used to produce the output might look. There is a clear operator effect. Once discovered one can often deal with external variables by expanding the model and making the variables internal. Methods for doing this will be discussed in Chapters 4 and 5.

3.4 WEIGHTED REGRESSION

Most regression programs have an option to compute **weighted least squares** fits. In the context of simple linear regression this means finding α and β to minimize the weighted least squares criterion

$$Q_w(\alpha,\beta)=\sum_{i=1}^{n} w_i(y_i-\alpha-\beta x_i)^2 \tag{3.7}$$

where the w_i are **weights** specified by the user. The main purpose for these is to deal with responses y_i that have unequal variances.

We will summarize here a number of properties of weighted least squares estimates, the proofs for which will be given in Chapter 6. Assume the data model

$$y_i=\alpha+\beta x_i+e_i$$

is unbiased. Then both the ordinary and weighted least squares estimates of α and β are unbiased. If, moreover, the y_i are uncorrelated and the weights are chosen so the weighted responses

$$\tilde{y}_i=\sqrt{w_i}\,y_i \tag{3.8}$$

have equal variance σ^2, then the weighted least squares estimates are optimal in the sense that among all linear unbiased estimates they have minimum variance. As a consequence, their variance cannot exceed that of the ordinary least squares estimates and can be considerably smaller (Problem 27, Chapter 6). The optimality here follows from the Gauss-Markov Theorem (Chapter 6) and because of this the resulting estimates are sometimes called **Gauss-Markov estimates**. The required weights, that is weights that make the transformed responses (3.8) have equal variance, are called **optimal** or **Gauss-Markov weights**.

Optimal weighting has other nice consequences. If the y_i are independent and normally distributed, the confidence intervals and tests of Section 2.4 apply with the understanding that the residual sum of squares RSS is replaced throughout by the **weighted residual sum of squares**

$$RSS_w = \sum w_i\,(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{3.9}$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote the weighted least squares estimates. This replacement is done automatically by regression programs that allow user-specified weights. The weighted residual mean square $RMS_w = RSS_w/(n-2)$ estimates the common variance σ^2 of the weighted responses $\tilde{y}_i$. Note that the value of σ^2 depends on the choice of weights. Multiplying a set of weights by a positive constant c gives another set of weights. The estimates of α and β are unchanged, but σ^2 and its estimate are multiplied by c.

From (3.8)

$$w_i = \frac{\sigma^2}{var\, y_i}$$

so the optimal weights are inversely proportional to the variances of the y_i. If the y_i have equal variances, $w_i = 1$ gives optimal weights and weighted least squares reduces to ordinary least squares.

As a simple example, if each y_i is a mean of a sample of size n_i and has variance σ^2/n_i, the choice $w_i = n_i$ is natural and optimal.

In general it may be more difficult to find appropriate weights. Often the variances of the y_i are, or are approximately functions of their expectations $f_i = \alpha + \beta x_i$. For example, when the y_i represent counts from a Geiger counter and this is the only source of error, their distributions are usually Poisson. Then one expects the variance of each y_i to equal its expected value because the variance of a Poisson distribution equals its mean. If the primary errors are those arising from aliquoting (measuring) a blood sample to be placed into a Geiger counter, one expects the standard deviations of the y_i to be proportional to their expectations, and hence their variances to be proportional to the squares of their expectations. This suggests the use of weights $w_i = 1/f_i$ in the first case and $w_i = 1/f_i^2$ in the second, but, of course, the f_i are in general unknown because α and β are unknown. Often it is sufficient to approximate f_i by the fitted value $\hat{y}_i$ from an unweighted least squares fit and use weights $w_i = 1/\hat{y}_i$ and $w_i = 1/\hat{y}_i^2$ in the two cases mentioned. A possibly better method of approximation, called iterative reweighting, is discussed in Chapter 9. If the y_i are quite close to the f_i it may be sufficient to approximate f_i by y_i and use weights $w_i = 1/y_i$ or $w_i = 1/y_i^2$, but this, while convenient,

is not recommended for general use because it can lead to serious biases when the errors are not small.

One may not know ahead of time how the variances of the y_i are related to their expectations. The fan-shaped residual plot in Figure 3.2(b) suggests that the spread of the fitted residuals is roughly proportional to the fitted values and hence that the standard deviations of the observations y_i are roughly proportional to their expectations. This suggests the use of weights $w_i = 1/\hat{y}_i^2$.

If one is successful in choosing weights so the weighted observations (3.8) have equal variances, a plot of $\sqrt{w_i}\ \hat{e}_i$ on $\hat{y}$ should look like that in Figure 3.11. If this plot is still fan-shaped to the right, one might use weights inversely proportional to a higher power of $\hat{y}_i$ or, if it is fan-shaped to the left, a lower power.

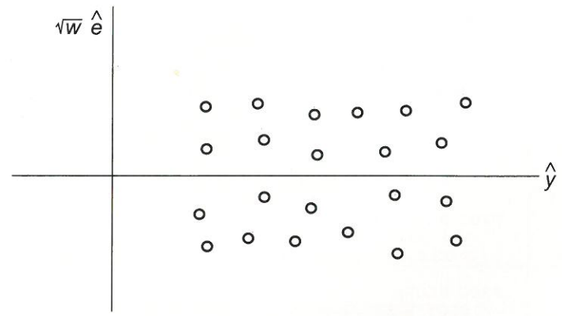


Figure 3.11: A weighted residual plot showing a successful choice of weights.

The appropriate and natural normal probability plot when using weights is that for the weighted residuals $\sqrt{w_i}\ \hat{e}_i$ on normal scores s_i. Another appropriate plot is that of Studentized residuals (Chapter 4) on normal scores. Not all regression programs provide appropriately weighted residual plots when weights are used, so again it is necessary to read plot labels carefully to avoid being misled.

Example 3.3

The residual on fit plot for the air pressure data, Figure 3.5, is somewhat fan-shaped opening to the right. This motivates weighting the large responses less than the small responses by, for example, using weights $w_i = 1/\hat{P}_i$ where the $\hat{P}_i$ are the fitted values from an unweighted least squares fit. A SAS problem specification for this is given in Figure 3.12. It will be discussed in a moment. The weighted least squares fit is

$$\hat{P} = 0.110 + 1405/V$$

```
DATA;
   INPUT V P;
   VI=1/V;
   CARDS;
   48  29.1

   ( DATA FROM TABLE 3.1 HERE )

   12 117.6
   ;

PROC REG;
   MODEL P=VI;
   OUTPUT P=LSFIT;

DATA;
   SET;
   W=1/LSFIT;

PROC REG;
   MODEL P=VI;
   WEIGHT W;
   OUTPUT P=FIT R=RES;

DATA;
   SET;
   WRES=SQRT(W)*RES;

PROC RANK NORMAL=VW;
   VAR WRES;
   RANKS NSCORE;

PROC PLOT;
   PLOT WRES*FIT='*' / VREF=0 VPOS=30;
   PLOT WRES*NSCORE='*' / VPOS=30;
   LABEL WRES='WEIGHTED RESIDUAL' NSCORE='NORMAL SCORE';

RUN;
```

Figure 3.12: SAS problem specification for the weighted regression in Example 3.3.

This is very close to the unweighted fit. Figure 3.13 shows the weighted residual plot. Most of the fan shape has been removed. Although it is not shown, the normal probability plot for the weighted residuals from the weighted fit appears more linear than that for the unweighted residuals from the unweighted fit. While these are improvements, they are only minor. Here and in perhaps most linear regression problems, weights do not play a major role. As noted, this will change when we consider nonlinear problems in Chapters 8 and 9.

In Figure 3.12 the first PROC REG step computes the unweighted least squares fit. This is used in the following DATA step to compute the weights W which are used in the second PROC REG step to produce the weighted fit. The DATA step that follows this computes the weighted residuals WRES. The SET command in these two DATA steps tells SAS to use the most current data set. ♦

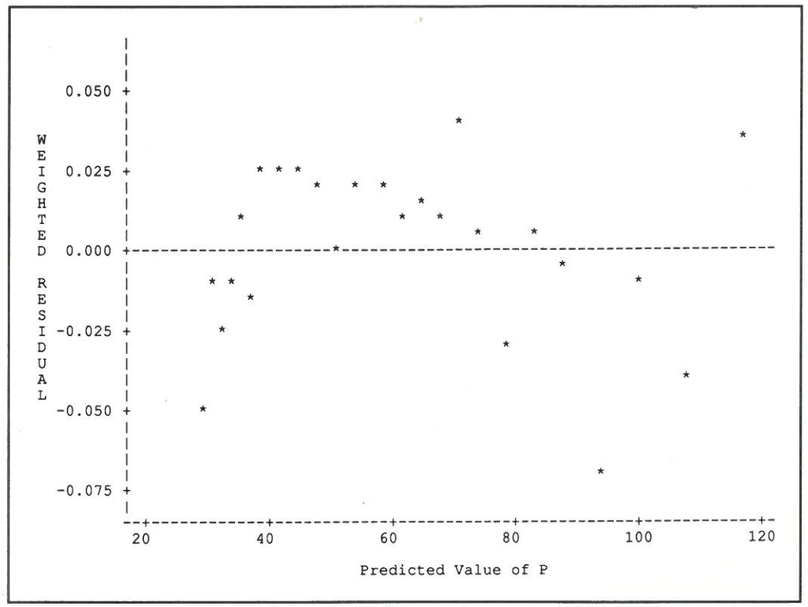


Figure 3.13: Weighted residual plot of $\sqrt{w}\,\hat{e}$ on $\hat{y}$ for a weighted fit of the model (3.6) to the air pressure data in Table 3.1.

3.5 STRUCTURE OF THE POWER TRANSFORMS

In Section 3.1 we used transformations to reduce some nonlinear models to simple linear regression models. Transformations can also be used to deal with problems of bias, inequality of variance, and non-normality, but not necessarily simultaneously. It is important to understand what effect transformations have on each of these problems, and for this it is useful to know something about the structure of a very important class of transformations called the **power transforms:**

$$-1/y, -1/\sqrt{y}, \log y, \sqrt{y}, y \tag{3.10}$$

The minus signs that appear with the first two power transforms are used to make them all increasing functions of y. This simplifies the moving rules given later. It is assumed for now that y ranges over positive values. We consider here the structure of the power transforms and explain in particular why $\log y$ appears in the middle of the list. One reason, in fact, for choosing the list (3.10) over some other list is that it is centered on $\log y$ which has well-documented effectiveness both for bias removal and variance stabilization.

We consider first the concept of equivalent transformations. Assume $T(y)$ is any