CHAPTER 6

Problems with the Error

The standard assumption about the error term ε is that it is independent and identically distributed (i.i.d.) from case to case. That is, var $\varepsilon = \sigma^2 I$. Furthermore, we also assume that the errors are normally distributed in order to carry out the usual statistical inference. We have seen that these assumptions can often be violated and we must then consider alternatives. When the errors are not i.i.d., we consider the use of *generalized least squares* (GLS). When the errors are independent, but not identically distributed, we can use *weighted least squares* (WLS), which is a special case of GLS. Sometimes, we have a good idea how large the error should be, but the residuals may be much larger than we expect. This is evidence of a *lack of fit*. When the errors are not normally distributed, we can use *robust regression*.

6.1 Generalized Least Squares

Until now we have assumed that var $\varepsilon = \sigma^2 I$, but sometimes the errors have nonconstant variance or are correlated. Suppose instead that var $\varepsilon = \sigma^2 \Sigma$ where σ^2 is unknown but Σ is known — in other words, we know the correlation and relative variance between the errors, but we do not know the absolute scale. Right now, it might seem redundant to distinguish between σ and Σ, but we will see how this will be useful later.

We can write $\Sigma = SS^T$, where S is a triangular matrix using the Choleski decomposition. Now we can transform the regression model as follows:

$$\begin{aligned} y &= X\beta + \varepsilon \\ S^{-1}y &= S^{-1}X\beta + S^{-1}\varepsilon \\ y' &= X'\beta + \varepsilon' \end{aligned}$$

Now we find that:

$$\text{var } \varepsilon' = \text{var } (S^{-1}\varepsilon) = S^{-1}(\text{var } \varepsilon)S^{-T} = S^{-1}\sigma^2 SS^T S^{-T} = \sigma^2 I$$

So we can reduce GLS to ordinary least squares (OLS) by a regression of $y' = S^{-1}y$ on $S^{-1}X$ which has error ε' that is i.i.d. So we simply reduce the problem to one that we have already solved. In this transformed model, the sum of squares is:

$$(S^{-1}y - S^{-1}X\beta)^T(S^{-1}y - S^{-1}X\beta) = (y - X\beta)^T S^{-T}S^{-1}(y - X\beta) = (y - X\beta)^T\Sigma^{-1}(y - X\beta)$$

which is minimized by:

$$\hat\beta = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$$

We find that:

$$\text{var } \hat\beta = (X^T\Sigma^{-1}X)^{-1}\sigma^2$$

Since $\varepsilon' = S^{-1}\varepsilon$, diagnostics should be applied to the residuals, $S^{-1}\hat{\varepsilon}$. If we have the right Σ, then these should be approximately i.i.d.

The main problem in applying GLS in practice is that Σ may not be known and we may have to estimate it. To illustrate this we will use a built-in R dataset known as Longley's regression data. Our response is the number of people employed, yearly from 1947 to 1962 and the predictors are gross national product (GNP) and population 14 years of age and over. The data originally appeared in Longley (1967).

Fit a linear model:

```
> data(longley)
> g <- lm(Employed ~ GNP + Population, longley)
> summary(g,cor=T)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.9388    13.7850    6.45  2.2e-05
GNP           0.0632     0.0106    5.93  5.0e-05
Population   -0.4097     0.1521   -2.69    0.018

Residual standard error: 0.546 on 13 degrees of freedom
Multiple R-Squared: 0.979,      Adjusted R-squared: 0.976
F-statistic:  304 on 2 and 13 DF,  p-value: 1.22e-11

Correlation of Coefficients:
           (Intercept) GNP
GNP         0.98
Population -1.00        -0.99
```

The correlation between the coefficients for GNP and Population is strongly negative while the correlation between the corresponding variables:

```
> cor(longley$GNP,longley$Pop)
[1] 0.99109
```

is strongly positive.

In data collected over time such as this, successive errors could be correlated. The simplest way to model this is the autoregressive form:

$$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i$$

where $\delta_i \sim N(0,\tau^2)$. We can estimate this correlation ρ by:

```
> cor(residuals(g)[-1],residuals(g)[-16])
[1] 0.31041
```

Under this assumption $\Sigma_{ij} = \rho^{|i-j|}$. For simplicity, let's assume we know that $\rho = 0.31041$. We now construct the Σ matrix and compute the GLS estimate of β along with its standard errors. The calculation is for demonstration purposes only:

```
> x <- model.matrix(g)
> Sigma <- diag(16)
> Sigma <- 0.31041^abs(row(Sigma)-col(Sigma))
> Sigi <- solve(Sigma)
> xtxi <- solve(t(x) %*% Sigi %*% x)
> (beta <- solve(t(x) %*% Sigi %*% x,t(x) %*%
```

```
  Sigi %*% longley$Empl))
                   [,1]
  (Intercept) 94.89889
  GNP          0.06739
  Population  -0.47427
> res <- longley$Empl - x %*% beta
> (sig <- sqrt((t(res) %*% Sigi %*% res)/g$df))
             [,1]
  [1,] 0.5424432
> sqrt(diag(xtxi))*sig
[1] 13.94477260  0.01070339  0.15338547
```

Compare with the model output above where the errors are assumed to be uncorrelated. Another way to get the same result is to regress $S^{-1}y$ on $S^{-1}x$ as we demonstrate here:

```
> sm <- chol(Sigma)
> smi <- solve(t(sm))
> sx <- smi %*% x
> sy <- smi %*% longley$Empl
> summary(lm(sy ~ sx-1))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
sx(Intercept)  94.8989    13.9448    6.81  1.3e-05
sxGNP           0.0674     0.0107    6.30  2.8e-05
sxPopulation   -0.4743     0.1534   -3.09   0.0086

Residual standard error: 0.542 on 13 degrees of freedom
```

In practice, we would not know that the $\rho = 0.31$ and we will need to estimate it from the data. Our initial estimate is 0.31, but once we fit our GLS model we would need to reestimate it as:

```
> cor(res[-1],res[-16])
[1] 0.35642
```

and then recompute the model again with $\rho = 0.35642$. This process would be iterated until convergence. This is cumbersome. A more convenient approach may be found in the nlme package of Pinheiro and Bates (2000), which contains a GLS fitting function. We can use it to fit this model:

```
> library(nlme)
> g <- gls(Employed ~ GNP + Population,
  correlation=corAR1(form= ~Year), data=longley)
> summary(g)
Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
    Phi
0.64417

Coefficients:
              Value Std.Error t-value p-value
```

```
(Intercept) 101.858   14.1989  7.1736  <.0001
GNP           0.072    0.0106  6.7955  <.0001
Population   -0.549    0.1541 -3.5588  0.0035

Residual standard error: 0.68921
Degrees of freedom: 16 total; 13 residual
```

We see that the estimated value of ρ is 0.64. However, if we check the confidence intervals for this:

```
> intervals(g)
Approximate 95% confidence intervals

 Coefficients:
                 lower       est.       upper
(Intercept) 71.183204 101.858133 132.533061
GNP          0.049159   0.072071   0.094983
Population  -0.881491  -0.548513  -0.215536

 Correlation structure:
        lower    est.   upper
Phi -0.44335 0.64417 0.96451

 Residual standard error:
  lower    est.   upper
0.24772 0.68921 1.91748
```

we see from the interval, $(-0.44, 0.96)$, that it is not significantly different from zero. So there is no evidence of serial correlation.

6.2 Weighted Least Squares

Sometimes the errors are uncorrelated, but have unequal variance where the form of the inequality is known. When Σ is diagonal, the errors are uncorrelated but do not necessarily have equal variance. WLS can be used in this situation. We can write $\Sigma = \text{diag}(1/w_1, \ldots, 1/w_n)$, where the w_i are the *weights* so $S = \text{diag}(\sqrt{1/w_1}, \ldots, \sqrt{1/w_n})$. So we can regress $\sqrt{w_i} y_i$ on $\sqrt{w_i} x_i$ (although the column of ones in the X-matrix needs to be replaced with $\sqrt{w_i}$). Cases with low variability should get a high weight, high variability a low weight. Some examples:

1. Errors proportional to a predictor: $\text{var}(\varepsilon_i) \propto x_i$ suggests $w_i = x_i^{-1}$.
2. When the Y_i are the averages of n_i observations, then $\text{var}\, y_i = \text{var}\, \varepsilon_i = \sigma^2/n_i$, which suggests $w_i = n_i$. Responses that are averages arise quite commonly, but take care that the variance in the response really is proportional to the group size. For example, consider the life expectancy for different countries. At first glance, one might consider setting the weights equal to the populations of the countries, but notice that there are many other sources of variation in life expectancy that would dwarf the population size effect. Setting $w_i = n_i$ is only likely to be sensible for small n_i.

When weights are used, the residuals must be modified. Use $\sqrt{w_i}\hat{\varepsilon}_i$ for diagnostics.

Elections for the French presidency proceed in two rounds. In 1981, there were 10 candidates in the first round. The top two candidates then went on to the second round, which was won by François Mitterand over Valéry Giscard-d'Estaing. The losers in the first round can gain political favors by urging their supporters to vote for one of the two finalists. Since voting is private, we cannot know how these votes were transferred; we might hope to infer from the published vote totals how this might have happened. Anderson and Loynes (1987) published data on these vote totals in every fourth department of France:

```
> data(fpe)
> fpe
                   EI  A  B  C  D E F G H J K  A2  B2
Ain               260 51 64 36 23 9 5 4 4 3 3 105 114
Alpes              75 14 17  9  9 3 1 2 1 1 1  32  31
...
```

`A` and `B` stand for Mitterand's and Giscard's votes in the first round, respectively, while `A2` and `B2` represent their votes in the second round. `C-K` are the first round votes of the other candidates while `EI` is *electeur inscrits* or registered voters. All numbers are in thousands. The total number of voters in the second round was greater than the first — we can compute the difference as `N`.

We will treat this group effectively as another first round candidate (we could reasonably handle this differently). Now we can represent the transfer of votes as:

$$A2 = \beta_A A + \beta_B B + \beta_C C + \beta_D D + \beta_E E + \beta_F F + \beta_G G + \beta_H H + \beta_J J + \beta_K K + \beta_N N$$

where β_i represents the proportion of votes transferred from candidate i to Mitterand in the second round. Now we would expect these transfer proportions to vary somewhat between departments, so if we treat the above as a regression equation, there will be some error from department to department. The error will have a variance in proportion to the number of voters because it will be like a variance of a sum rather than a mean. Since the weights should be inversely proportional to the variance, this suggests that the weights should be set to `1/EI`. Notice also that the equation has no intercept, hence the `-1` in the model formula. We fit the appropriate model:

```
> g <- lm(A2 ~ A+B+C+D+E+F+G+H+J+K+N-1, fpe, weights=1/EI)
> coef(g)
       A        B        C        D        E        F        G
 1.06713 -0.10505  0.24596  0.92619  0.24940  0.75511  1.97221
       H        J        K        N
-0.56622  0.61164  1.21066  0.52935
```

Note that the weights do matter — see what happens when we leave them out:

```
> lm(A2 ~ A+B+C+D+E+F+G+H+J+K+N-1, fpe)$coef
       A        B        C        D        E        F        G
 1.07515 -0.12456  0.25745  0.90454  0.67068  0.78253  2.16566
       H        J        K        N
-0.85429  0.14442  0.51813  0.55827
```

which causes substantial changes for some of the lesser candidates. Furthermore, only the relative proportions of the weights matter — for example, suppose we multiply the weights by 53:

```
> lm(A2 ~ A+B+C+D+E+F+G+H+J+K+N-1, fpe, weights=53/EI)$coef
       A        B        C        D        E        F        G
 1.06713 -0.10505  0.24596  0.92619  0.24940  0.75511  1.97221
       H        J        K        N
-0.56622  0.61164  1.21066  0.52935
```

This makes no difference.

Now there is one remaining difficulty, unrelated to the weighting, in that proportions are supposed to be between zero and one. We can impose an *ad hoc* fix by truncating the coefficients that violate this restriction either to zero or one as appropriate. This gives:

```
> lm(A2 ~ offset(A+G+K)+C+D+E+F+N-1, fpe, weights=1/EI)$coef
      C       D       E       F       N
0.22577 0.96998 0.39020 0.74424 0.60854
```

We see that voters for the Communist candidate D apparently almost all voted for the Socialist Mitterand in the second round. However, we see that around 20% of the voters for the Gaullist candidate C voted for Mitterand. This is surprising since these voters would normally favor the more right wing candidate, Giscard. This appears to be the decisive factor. We see that of the larger blocks of smaller candidates, the Ecology party voters, E, roughly split their votes as did the first round nonvoters. The other candidates had very few voters and so their behavior is less interesting.

This analysis is somewhat crude and more sophisticated approaches are discussed in Anderson and Loynes (1987).

In cases where the form of the variance of ε is not completely known, we may model Σ using a small number of parameters. For example:

$$\text{var } \varepsilon_i = \gamma_0 + \gamma_1 x_1$$

might seem reasonable in a given situation. The iteratively reweighted least squares (IRWLS) fitting algorithm is:

1. Start with $w_i = 1$.
2. Use least squares to estimate β.
3. Use the residuals to estimate γ, perhaps by regressing $\hat{\varepsilon}^2$ on x.
4. Recompute the weights and go to 2.

Continue until convergence. There are some concerns about this because the estimation of the γ has some uncertainty and consumes some degrees of freedom. This affects the subsequent inference about β. An extensive investigation of this may be found in Carroll and Ruppert (1988).

Another approach is to model the variance and jointly estimate the regression and weighting parameters using a likelihood-based method. This can be implemented in R using the `gls()` function in the `nlme` library.

6.3 Testing for Lack of Fit

How can we tell whether a model fits the data? If the model is correct, then $\hat{\sigma}^2$ should be an unbiased estimate of σ^2. If we have a model that is not complex enough to fit the data or simply takes the wrong form, then $\hat{\sigma}^2$ will overestimate σ^2. The situation is illustrated in Figure 6.1. Alternatively, if our model is too complex and overfits the data, then $\hat{\sigma}^2$ will be an underestimate.

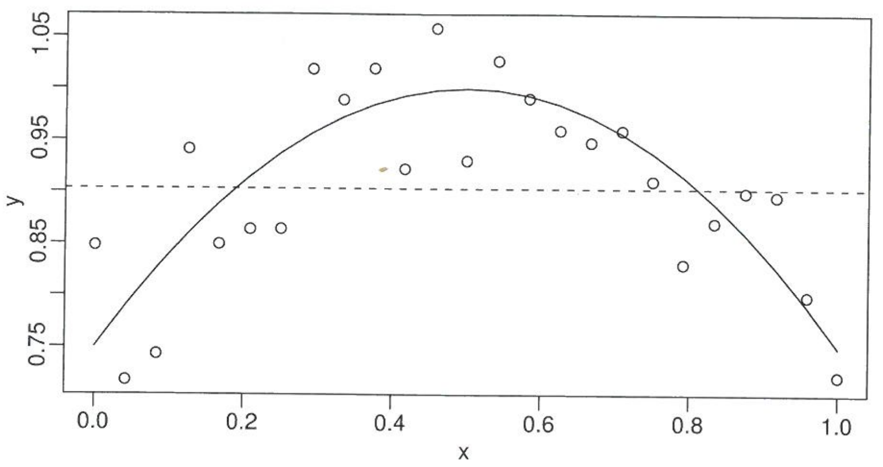

Figure 6.1 *True quadratic fit shown with the solid line and incorrect linear fit shown with the dotted line. Estimate of σ^2 will be unbiased for the quadratic model, but far too large for the linear model.*

This suggests a possible testing procedure — we should compare $\hat{\sigma}^2$ to σ^2. The usual problem, of course, is that we do not know the true value of σ^2 and so the comparison cannot be made. In a few cases, we might actually know σ^2 – for example, when measurement error is the only source of variation and we know its variance because we are very familiar with the measurement device. This is rather uncommon and we do not discuss it here — see Weisberg (1985) for an example. A more realistic possibility is that we have *replication* in our data that allows an estimate of σ^2 that does not depend on any particular model.

The $\hat{\sigma}^2$ that is based in the chosen regression model needs to be compared to some model-free estimate of σ^2. We can do this if we have repeated y for one or more fixed x. These replicates do need to be truly independent. They cannot just be repeated measurements on the same subject or unit. Such repeated measures would only reveal the within subject variability or the measurement error. We need to know the between subject variability, as this reflects the σ^2 described in the model. Let y_{ij} be the i^{th} observation in the group of replicates j.

The "pure error" estimate of σ^2 is given by SS_{pe}/df_{pe} where:

$$SS_{pe} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

and degrees of freedom $df_{pe} = \sum_j (\#replicates - 1) = n - \#groups$.

If you fit a model that assigns one parameter to each group of observations with