Chapter 7: Variances

In this chapter we consider a variety of extensions to the linear model that allow for more gen-

eral variance structures than the independent, identically distributed errors assumed in earlier chapters. This greatly extends the problems to which linear regression can be applied.
1 Weighted least squares

The assumption that the variance function $\text{Var}(Y|X)$ is the same for all values of the terms $X$ can be relaxed as follows.

$$\text{Var}(Y|X = x_i) = \text{Var}(e_i) = \sigma^2/w_i,$$

where $w_1, \ldots, w_n$ are known positive numbers. This leads to the use of weighted least squares, or WLS, in place of OLS, to get estimates.
In matrix terms, the model can be written as

$$Y = X\beta + e, \quad \text{Var}(e) = \sigma^2 W^{-1}. \quad (1)$$

The estimator $\hat{\beta}$ is chosen to minimize the weighted residual sum of squares function,

$$RSS(\beta) = (Y - X\beta)'W(Y - X\beta)$$

$$= \sum_i w_i(y_i - x_i\beta)^2.$$ 

The WLS estimator is given by

$$\hat{\beta} = (X'WX)^{-1}X'Wy.$$
Let $W^{1/2}$ be the $n \times n$ diagonal matrix with $i$th diagonal element $\sqrt{w_i}$, and so $W^{-1/2}$ is a diagonal matrix with $1/\sqrt{w_i}$ on the diagonal. Define $Z = W^{1/2}Y$, $M = W^{1/2}X$, and $d = \omega^{1/2}e$, and (1) is equivalent to

$$Z = M\beta + d.$$  \hfill (2)
This model can be solved using OLS,

\[ \hat{\beta} = (M' M)^{-1} M' Z = (X' W X)^{-1} X' W y, \]

which is identical to the WLS estimator.
1.1 Applications of weighted least squares

Known weight $w_i$ can occur in many ways. If the $i$th response is an average of $n_i$ equally variable observations, then $\text{Var}(y_i) = \sigma/n_i$, and $w_i = n_i$. If $y_i$ is a total of $n_i$ observations, $\text{Var}(y_i) = n_i\sigma^2$, and $w_i = 1/n_i$. If variance is proportional to some predictor $x_i$, $\text{Var}(y_i) = x_i\sigma^2$, then $w_i = 1/x_i$. 
In physics, a theoretical model of the strong interaction force predicts that

\[ E(y|s) = \beta_0 + \beta_1 s^{-1/2} + \text{relatively small terms}. \]

In an experiment, the following data are observed:

At each value of \( s \) (\( x = s^{-1/2} \)), a very large number of particles was counted, and as a result the values of \( \text{Var}(y|s = s_i) = \sigma^2/w_i \) are known almost exactly; the square roots of these values are given in the third column of the table. The fit of
Table 1: The strong interaction data.

<table>
<thead>
<tr>
<th>No.</th>
<th>x</th>
<th>y</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.345</td>
<td>367</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>0.287</td>
<td>311</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>0.251</td>
<td>295</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>0.225</td>
<td>268</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>0.207</td>
<td>253</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>0.186</td>
<td>239</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0.161</td>
<td>220</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>0.132</td>
<td>213</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>0.084</td>
<td>193</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>0.060</td>
<td>192</td>
<td>5</td>
</tr>
</tbody>
</table>
the simple regression model via WLS is summarized in Table 2. \( R^2 \) is large, and the parameter estimates are well determined.
Table 2: WLS estimates for the strong interaction data.

Coefficients:

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| (Intercept) | 148.473  | 8.079      | 18.38   | 7.91e-08 |
| x      | 530.835  | 47.550     | 11.16   | 3.71e-06 |

Residual standard error: 1.657 on 8 degrees of freedom
Multiple R-Squared: 0.9397,    Adjusted R-squared: 0.9321
F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.710e-06

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>341.99</td>
<td>341.99</td>
<td>124.63</td>
<td>3.710e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>21.95</td>
<td>2.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2 Misspecified Variances

Suppose the true regression model is

\[ E(Y|X) = X\beta, \quad \text{Var}(Y|X) = \sigma^2 W^{-1}, \]

where \( W \) has positive weights on the diagonal and zeros elsewhere. We get the weights wrong, and fit the model using OLS with the estimator

\[ \hat{\beta}_0 = (X'X)^{-1} X'Y. \]
Similar to the correct WLS estimate, this estimator is unbiased, $E(\hat{\beta}_0|X) = \beta$. However, the variance is

$$\text{Var}(\hat{\beta}_0|X) = \sigma^2(X'X)^{-1}(X'W^{-1}X)(X'X)^{-1}.$$
2.1 Accommodating misspecified variance

To estimate \( \text{Var}(\hat{\beta}_0 | X) \) requires estimating \( \sigma^2 W^{-1} \). Let \( \hat{e}_i = y_i - \hat{\beta}_0' x_i \) be the \( i \)th residual from the misspecified model. Then \( \hat{e}_i^2 \) is an estimate of \( \sigma^2 / w_i \).

It has been shown that replacing \( \sigma^2 W^{-1} \) by a diagonal matrix with the \( \hat{e}_i^2 \) on the diagonal produces a consistent estimate of \( \text{Var}(\hat{\beta}_0 | X) \).

Several variations of this estimate that are equivalent in large samples but have better small sam-
ple behavior have been proposed. For example, one estimate, often called HC3, is

$$\text{Var}(\hat{\beta}_0 | X) = (X'X)^{-1} \left[ X'd\text{ia}g \left( \frac{\hat{e}_i^2}{(1 - h_{ii})^2} \right) X \right] (X'X)^{-1},$$

where $h_{ii}$ is the $i$th leverage. An estimator of this type is often called a sandwich estimator.
Table 3: Sniffer data estimates and standard errors

<table>
<thead>
<tr>
<th></th>
<th>OLS Estimate</th>
<th>OLS SE</th>
<th>HC3 SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.15391</td>
<td>1.03489</td>
<td>1.047</td>
</tr>
<tr>
<td>TankTemp</td>
<td>-0.08269</td>
<td>0.04857</td>
<td>0.044</td>
</tr>
<tr>
<td>GasTemp</td>
<td>0.18971</td>
<td>0.04118</td>
<td>0.034</td>
</tr>
<tr>
<td>TankPres</td>
<td>-4.05962</td>
<td>1.58000</td>
<td>1.972</td>
</tr>
<tr>
<td>GasPres</td>
<td>9.85744</td>
<td>1.62515</td>
<td>2.056</td>
</tr>
</tbody>
</table>

2.2 A Test for Constant Variance

Suppose that for some parameter vector $\lambda$ and some vector of regressors $Z$

$$\text{Var}(Y|X, Z = z) = \sigma^2 \exp(\lambda'z), \quad (3)$$
where the weights are given by $w = \exp(-\lambda'z)$.

If $\lambda = 0$, then (3) corresponds to constant variance. Hence, a test of NH: $\lambda = 0$ versus $AH : \lambda \neq 0$ is a test for nonconstant variance. There is a great latitude in specifying $Z$. If $Z = Y$, then the variance depends on the response. Similarly, $Z$ may be the same as $X$ or a subset of $X$.

Assume normal errors, a score test can be used, for which the test statistic has an approximate $\chi^2(q)$
<table>
<thead>
<tr>
<th>Choice for Z</th>
<th>df</th>
<th>Test Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GasPres</td>
<td>1</td>
<td>5.50</td>
<td>0.019</td>
</tr>
<tr>
<td>TankTemp</td>
<td>1</td>
<td>9.71</td>
<td>0.002</td>
</tr>
<tr>
<td>TankTemp, GasPres</td>
<td>2</td>
<td>11.78</td>
<td>0.003</td>
</tr>
<tr>
<td>Fitted value</td>
<td>1</td>
<td>4.80</td>
<td>0.028</td>
</tr>
</tbody>
</table>

distribution and $q$ is the number of regressors in $Z$. 
3 General Correlation Structures

The generalized least squares or GLS model extends WLS one step further, and starts with

\[ E(Y \mid X) = X\beta, \quad \text{Var}(Y \mid X) = \Sigma, \]

where \( \Sigma \) is an \( n \times n \) positive definite symmetric matrix. The WLS model uses \( \Sigma = \sigma^2 W^{-1} \) for a diagonal matrix \( W \), and the OLS model uses \( \Sigma = \sigma^2 I \).
If we have \( n \) observations and \( \Sigma \) is completely unknown, then \( \Sigma \) contains \( n(n - 1)/2 \) parameters, which is much larger than the number of observations \( n \). The only hope is to introduce some structure in \( \Sigma \). Here are some examples.
Compound Symmetry

If all the observations are equally correlated, then

\[ \Sigma_{CS} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \]

which has only two parameters \( \rho \) and \( \sigma^2 \). Generalized least squares software, such as the \textit{gls} function in the \textit{nlme} package, can be used for es-
Autoregressive  This form is generally associated with time series. If data are time ordered and equally spaced, the lag-1 autoregressive covariance struc-
ture is

\[ \Sigma_{AR} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}, \]

which also contains only two parameters.

Block Diagonal A block diagonal form for \( \Sigma \) can arise if observations are sampled clusters. For example, a study of school performance might sample
$m$ children from each of $k$ classrooms. The $m$ children within a classroom may be correlated because they all have the same teacher, but children in different classrooms are independent.
The random coefficient model, as special case of mixed models, allows for appropriate inferences. Consider a population regression mean function

\[ E(y|\text{loudness} = x) = \beta_0 + \beta_1 x, \]

where subject effects are not allowed. To add them we hypothesize that each of the subjects may have his or her own slope and intercept. Let \( y_{ij}, \ i = \)
1, \ldots, 10, \; j = 1, \ldots, 5 \text{ be the log-response for subject } i \text{ measured at the } j \text{th level of loudness. For the } i \text{th subject,}

\[
E(y_{ij}|\text{loudness} = x, b_{0i}, b_{1i}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\text{loudness}_{ij}.
\]

where \(b_{0i}\) and \(b_{1i}\) are the deviations from the population intercept and slope for the \(i\)th subject, and
they are treated as random variables,

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix} \right).$$

Inferences about \((\beta_0, \beta_1)\) concern population behavior. Inferences about \((\tau_0^2, \tau_1^2)\) concern the variation of the intercepts and slopes between individuals in the population.
5 Variance Stabilizing Transformation

Suppose that the response is strictly positive, and the variance function before transformation is

$$\text{Var}(Y \mid X = x) = \sigma^2 g(E(Y \mid X = x)),$$

where $g(E(Y \mid X = x))$ is a function that is increasing with the value of its argument. For example, if the distribution of $Y \mid X$ has a Poisson distribution, then $g(E(Y \mid X = x)) = E(Y \mid X = x)$. 
For distributions in which the mean and variance are functionally related, Scheffe (1959) provides a general theory for determining transformations that can stabilize variance. Table 5 lists the common variance stabilizing transformations.
<table>
<thead>
<tr>
<th>$Y_T$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{Y}$</td>
<td>Used when $\text{Var}(Y</td>
</tr>
<tr>
<td>$\log(Y)$</td>
<td>Used if $\text{Var}(Y</td>
</tr>
<tr>
<td>$1/Y$</td>
<td>The inverse transformation stabilizes variance when $\text{Var}(Y</td>
</tr>
<tr>
<td>$\sin^{-1}(\sqrt{Y})$</td>
<td>The arcsine square-root transformation is used if $Y$ is a proportion between 0 and 1, but if can be used more generally if $y$ has a limited range by first transforming $Y$ to the range $(0,1)$, and then applying the transformation.</td>
</tr>
</tbody>
</table>

Table 5: Common variance stabilizing transformations.
6 The Delta Method

As a motivation example, we consider a polynomial regression. If the mean function with one predictor $X$ is smooth but not straight, integer powers of the predictors can be used to approximate $E(Y|X)$. The simplest example of this is quadratic regression, in which the mean function is

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$  (4)
Quadratic mean functions can be used when the mean is expected to have a minimum or maximum in the range of the predictor. The minimum or maximum will occur for the value of $X$ for which the derivative $dE(Y|X = x)/dx = 0$, which occurs at

$$x_M = -\beta_1/(2\beta_2).$$  \hspace{1cm} (5)

The delta methods provides an approximate standard error of a nonlinear combination of estimates
that is accurate in large samples.

Suppose \( g(\theta) \) is a nonlinear continuous function of \( \theta \), \( \theta^* \) is the true value of \( \theta \), and \( \hat{\theta} \) is the estimate. It follows from Taylor series expansion that

\[
g(\hat{\theta}) = g(\theta^*) + \sum_{j=1}^{k} \frac{\partial g}{\partial \theta_j} (\hat{\theta}_j - \theta^*_j) + \text{small terms} \\
\approx g(\theta^*) + \dot{g}(\theta^*)(\hat{\theta} - \theta^*),
\]
where $\dot{g}(\theta^*) = \left( \frac{\partial g}{\partial \theta_1}, \cdots, \frac{\partial g}{\partial \theta_1} \right)'$. It implies that

$$\text{Var}(g(\hat{\theta})) \approx \dot{g}(\theta^*)'\text{Var}(\hat{\theta})\dot{g}(\theta^*).$$

For quadratic regression (4), the minimum or maximum occurs at $g(\beta) = -\beta_1/(2\beta_2)$, which is estimated by $g(\hat{\beta})$. A straightforward calculation gives

$$\text{Var}(g(\beta)) = \frac{1}{4\hat{\beta}_2^2} \left( \text{Var}(\hat{\beta}_1) + \frac{\hat{\beta}_1^2}{\hat{\beta}_2^2} \text{Var}(\hat{\beta}_2) - \frac{2\hat{\beta}_1}{\hat{\beta}_2} \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \right),$$
where the variances and covariances are elements of the matrix $\sigma^2(X'X)^{-1}$, and so the estimated values are obtained from $\hat{\sigma}^2(X'X)^{-1}$.

Cakes This dataset is from a small experiment on baking packaged cake mixes. Two factors, $X_1$=baking time in minutes and $X_2$=baking temperature in degrees $F$, were varied in the experiment. The response $Y$ was the palatability score with higher
values desirable. The estimated mean function is

$$E(Y|X_1, X_2) = -2204.4850 + 25.91976X_1 + 9.9183X_2 - 0.1569X_1^2 - 0.0120X_2^2 - 0.0416X_1X_2.$$ 

When the temperature is set to 350, the estimated maximum palatability occurs when the baking time is

$$\hat{x}_M = -\frac{\hat{\beta}_1 + \hat{\beta}_{12}(350)}{2\hat{\beta}_{11}} = 36.2.$$ 

The standard error from the delta method can be computed to be 0.4 minutes. If we can believe the
normal approximation, a 95% confidence interval for $\hat{x}_M$ is $36.2 \pm 1.96 \times 0.4$ or about 35.4 to 37 minutes.
Suppose we have a sample \( y_1, \ldots, y_n \) from a particular distribution \( G \), for example a standard normal distribution. What is a confidence interval for the population median?

We can obtain an approximate answer to this question by computer simulation, set up as follows.

1. Obtain a simulated random sample \( y_1^*, \ldots, y_n^* \) from the known distribution \( G \).
2. Compute and save the median of the sample in step 1.

3. Repeat steps 1 and 2 a large number of times, say \( B \) times. The larger the value of \( B \), the more precise the ultimate answer.

4. If we take \( B = 999 \), a simple percentile-based 95% confidence interval for the median is the interval between the sample 2.5 and 97.5 percentiles, respectively.
In most interesting problems, $G$ is unknown and so this simulation is not available. Efron (1979) pointed out that the observed data can be used to estimate $G$, and then we can sample from the estimate $\hat{G}$. The algorithm becomes:

1. Obtain a random sample $y_1^*, \ldots, y_n^*$ from $\hat{G}$ by sampling with replacement from the observed values $y_1, \ldots, y_n$. In particular, the $i$-th element of the sample $y_i^*$ is equally likely to be
any of the original $y_1, \ldots, y_n$. Some of the $y_i$ will appear several times in the random sample, while others will not appear at all.

2. Continue with steps 2-4 of the first algorithm. A test at the 5\% level concerning the population median can be rejected if the hypothesized value of the median does not fall in the confidence interval computed in step 4.

Efron called this method the bootstrap.
7.1 Regression Inference without Normality

For regression problems, when the sample size is small and the normality assumption does not hold, standard inference methods can be misleading, and in these cases a bootstrap can be used for inference.

Transactions Data Each branch makes transactions of two types, and for each of the branches we have
recorded the number of transactions $T_1$ and $T_2$, as well as Time, the total number of minutes of labor used by the branch in type 1 and type 2 transactions. The mean response function is

$$E(Time|T_1, T_2) = \beta_0 + \beta_1 T_1 + \beta_2 T_2$$

possibly with $\beta_0 = 0$ because zero transactions should imply zero time spent. The data are displayed in Figure 1. The marginal response plots in the last row appear to have reasonably linear
mean functions; there appear to be a number of branches with no $T_1$ transactions but many $T_2$ transactions; and in the plot of Time versus $T_2$, variability appears to increase from left to right.

The errors in this problem probably have a skewed distribution. Occasional transactions take a very long time, but since transaction time is bounded below by zero, there cannot be any really extreme “quick” transactions. Inferences based on normal
Figure 1: Scatterplot matrix for the transactions data.
theory are therefore questionable.

A bootstrap is computed as follows.

1. Number the cases in the dataset from 1 to \( n \). Take a random sample with replacement of size \( n \) from these case numbers.

2. Create a dataset from the original data, but repeating each row in the dataset the number of times that row was selected in the random sample in step 1.
3. Repeat steps 1 and 2 a large number of times, say, $B$ times.

4. Estimate a 95% confidence interval for each of the estimates by the 2.5 and 97.5 percentiles of the sample of $B$ bootstrap samples.

Table 6 summarizes the percentile bootstrap for the transaction data.

The 95% bootstrap intervals are consistently wider than the corresponding normal intervals, indicat-
Table 6: Summary for $B = 999$ case bootstraps for the transactions data.

<table>
<thead>
<tr>
<th></th>
<th>Normal Theory</th>
<th></th>
<th>Bootstrap</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Lower</td>
<td>Upper</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>144.37</td>
<td>-191.47</td>
<td>480.21</td>
<td>136.09</td>
</tr>
<tr>
<td>$T_1$</td>
<td>5.46</td>
<td>4.61</td>
<td>6.32</td>
<td>5.48</td>
</tr>
<tr>
<td>$T_2$</td>
<td>2.03</td>
<td>1.85</td>
<td>2.22</td>
<td>2.04</td>
</tr>
</tbody>
</table>

indicating that the normal-theory confidence intervals are probably overly optimistic.
One of the important uses of the bootstrap is to get estimates of error variability in problems where standard theory is either missing, or, equally often, unknown to analyst. Suppose, for example, we wanted to get a confidence interval for the ratio $\beta_1/\beta_2$ in the transactions data. The point estimate for this ratio is just $\hat{\beta}_1/\hat{\beta}_2$, but we will not learn how to get a normal-theory confidence inter-
val for a nonlinear function of parameters like this until Chapter 6.

Using bootstrap, this computation is easy: just compute the ratio in each of the bootstrap samples and then use the percentiles of the bootstrap distribution to get the confidence interval. For these data, the point estimate is 2.68 with 95% confidence interval from 1.76 to 3.86.