Table 8.3. Standard errors are based on the estimate $\tilde{\sigma} = 1.1$. The estimate for the intercept corresponding to all factors at their lowest level is 3410×10^{-6}. Bearing in mind that the analysis is performed here on the reciprocal scale and that a large positive parameter corresponds to a small claim, we may deduce the following. The largest average claims are made by policyholders in the youngest four age groups, i.e. up to age 34, the smallest average claims by those aged 35–39, and intermediate claims by those aged 40 and over. These effects are in addition to effects due to type of vehicle and vehicle age. The value of claims decreases with car age, although not linearly. There are also marked differences between the four car groups, group D being the most expensive and group C intermediate. No significant difference is discernible between car groups A and B.

It should be pointed out that the parameter estimates given here are contrasts with level 1. In a balanced design the three sets of estimates corresponding to the three factors would be uncorrelated while the correlations within a factor would be 0.5. Even where, as here, there is considerable lack of balance, the correlations do not deviate markedly from these values.

It is possible to test and quantify the assertions made above by fusing levels 1–4, levels 6–8 of PA and levels 1 and 2 of CG. The deviance then increases to 129.8 on 116 d.f., which is a statistically insignificant increase.

The preceding analysis is not the only one possible for these data. In fact a multiplicative model corresponding to a logarithmic link function would lead to similar qualitative conclusions. As is shown in Chapter 10, the data themselves support the reciprocal model better but only marginally so, and it might be argued that quantitative conclusions for these data would be more readily stated and understood for a multiplicative model.

8.4.2 *Clotting times of blood*

Hurn *et al.* (1945) published data on the clotting time of blood, giving clotting times in seconds (y) for normal plasma diluted to nine different percentage concentrations with prothrombin-free plasma (u); clotting was induced by two lots of thromboplastin. The data are shown in Table 8.4. A hyperbolic model for lot 1 was fitted by Bliss (1970), using an inverse transformation of the data,

and for both lots 1 and 2 using untransformed data. We analyse both lots using the inverse link and gamma errors.

Initial plots suggest that a log scale for u is needed to produce inverse linearity, and that both intercepts and slopes are different for the two lots. This claim is confirmed by fitting the following model sequence:

Model	*Deviance*	*d.f.*
1	7.709	17
X	1.018	16
$L + X$	0.300	15
$L + L.X$	0.0294	14

Here $x = \log u$ and L is the factor defining the lots. Clearly all the terms are necessary and the final model produces a mean deviance whose square root is 0.0458, implying approximately a 4.6% standard error on the y-scale. The two fitted lines, with standard errors for the parameters shown in parentheses, are

lot 1: $\hat{\mu}^{-1} = -0.01655(\pm 0.00086) + 0.01534(\pm 0.00143)x$

lot 2: $\hat{\mu}^{-1} = -0.02391(\pm 0.00038) + 0.02360(\pm 0.00062)x$

The plot of the Pearson residuals $(y - \hat{\mu})/\hat{\mu}$ against the linear predictor $\hat{\eta}$ is satisfactory, and certainly better than either (i) the use of constant variance for Y where the residual range decreases with $\hat{\eta}$ or (ii) the use of constant variance for $1/Y$ where the analogous plot against $\hat{\mu}$ shows the range increasing with $\hat{\mu}$. Note that constant variance for $1/Y$ implies (to the first order) $\text{var}(Y) \propto \mu^4$. Thus the assumption of gamma errors (with $\text{var}(Y) \propto \mu^2$) is 'half-way' between assuming $\text{var}(Y)$ constant and $\text{var}(1/Y)$ constant.

The estimates suggest that the parameters for lot 2 are a constant multiple (about 1.6) of those for lot 1. If true this would mean that $\boldsymbol{\mu}_2 = k\boldsymbol{\mu}_1$, where the suffix denotes the lot. This model, though not a generalized linear model, has simple maximum-likelihood equations for estimating α, β and k where

$$\boldsymbol{\mu}_1 = 1/\boldsymbol{\eta}_1, \qquad \boldsymbol{\eta}_1 = \alpha + \beta \mathbf{x},$$
$$\boldsymbol{\mu}_2 = k\boldsymbol{\mu}_1.$$

Table 8.4 *Mean clotting times in seconds (y) of blood for nine percentage concentrations of normal plasma (u) and two lots of clotting agent*

	Clotting time	
u	*Lot* 1	*Lot* 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

These are equivalent to fitting α and β to data $\mathbf{y}_1$ and $\mathbf{y}_2/k$, combined with the equation $\sum(y_2/\mu_1 - k) = 0$. The resulting fit gives $\hat{k} = 0.625$ with deviance = 0.0332 and having 15 d.f. Comparing this with the fit of separate lines gives a difference of deviance of 0.0038 on one degree of freedom against a mean deviance of 0.0021 for the more complex model. The simpler model of proportionality is not discounted, with lot 2 giving times about five-eighths those of lot 1.

8.4.3 *Modelling rainfall data using two generalized linear models*

Histograms of daily rainfall data are usually skewed to the right with a 'spike' at the origin. This form of distribution suggests that such data might be modelled in two stages, one stage being concerned with the pattern of occurrence of wet and dry days, and the other with the amount of rain falling on wet days. The first stage involves discrete data and can often be modelled by a stochastic process in which the probability of rain on day t depends on the history of the process up to day $t - 1$. Often, first-order dependence corresponding to a Markov chain provides a satisfactory model. In the second stage we require a family of densities on the positive line for the quantity of rainfall. To be realistic, this family of densities should be positively skewed and should have variance increasing with μ. The gamma distribution