Solution to Additional Exercise 2.11

The dataset consists of measurements of concentrations of rubidium and bromide in potato slices at different times after their immersion in a solution containing these ions. Our interest is modelling the absorption of the ions over time.

y_{ij} = concentration in mg per 1000 g of water,
z_{ij} = time in hours (actually, the times are the same for the two ions: $z_{rj} = z_{bj}$),

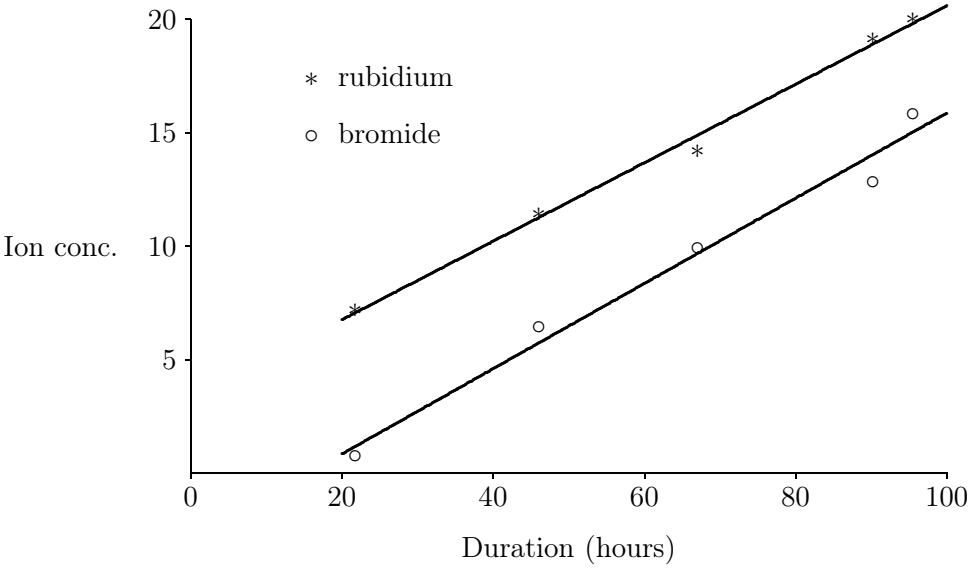
for the jth ($j = 1, \ldots, 5$) measurement of ion i ($i = r, b \sim$ rubidium, bromide).

2 separate regression models

Initially, we analyse the two ions completely separately, that is,

$$\begin{aligned} y_{rj} &= \mu_r + \gamma_r z_{rj} + \varepsilon_{rj}, \quad j = 1, \ldots, 5 \\ y_{bj} &= \mu_b + \gamma_b z_{bj} + \varepsilon_{bj}, \quad j = 1, \ldots, 5 \end{aligned} \tag{1}$$

where the rubidium errors $\varepsilon_{r1}, \ldots, \varepsilon_{r5}$ are assumed i.i.d. and $\mathrm{N}(0, \sigma_r^2)$, and the bromide errors $\varepsilon_{b1}, \ldots, \varepsilon_{b5}$ are assumed i.i.d. and $\mathrm{N}(0, \sigma_b^2)$. The plot shows the data points and the estimated regression lines from separate regressions. Both regression lines seem to fit the 5 data points very well. Parameter estimates (for this and the subsequent models) are collected in the table below.



The table and the plot shows the two regression lines to be close to parallel, whereas their intercepts clearly differ. In fact, the relation for bromide ions cannot be extended much further to the left in time, as concentrations are non-negative.

Separate regressions in the same model

Combining the two regression lines into one model makes the further assumption that the standard deviations about the lines are the same. (It also assumes the errors in the two equations to be

independent, see the Additional notes on the last page for further discussion.) The two estimated standard deviations from model (1) differ by a ratio of 0.971/0.443 = 2.2, which considering the small data sets (and corresponding small degrees of freedom for the estimates) is no serious deviation from equal standard deviations (and is far from statistically significant). The combined model can be written,

$$y_{ij} = \mu_i + \gamma_i z_{ij} + \varepsilon_{ij}, \quad i = r, b;\ j = 1, \ldots, 5 \tag{2}$$

where μ_i and γ_i are the intercept and slope, respectively, for the regression of ion i, $i = r, b$. The estimated regression parameters are the same as in model (1) because the assumptions about the mean part of the model is unchanged. The standard deviation is obtained by averaging the variances in the separate regression models, see the table below.

The ANOVA table, as listed in the expanded Minitab output, is as follows:

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	3	317.145	98.93%	317.145	105.715	185.47	0.000
hour	1	247.552	77.22%	134.237	134.237	235.51	0.000
ion	1	69.169	21.58%	15.097	15.097	26.49	0.002
hour*ion	1	0.424	0.13%	0.424	0.424	0.74	0.421
Error	6	3.420	1.07%	3.420	0.570		
Total	9	320.565	100.00%				

The interaction term `ion*time` corresponds to the two regressions being non-parallel. The F-test of the table shows that there is no evidence at all against taking the slopes (absorption rates) to be equal for rubidium and bromide.

Parallel regressions on time for rubidium and bromide

The model with parallel regression lines corresponds to the classical analysis of covariance, where the "covariate" (time) enters the model only with a single regression term. The model can be written,

$$y_{ij} = \mu_i + \gamma z_{ij} + \varepsilon_{ij}, \quad i = r, b;\ j = 1, \ldots, 5, \tag{3}$$

where the parameter γ represents the common slope of the two regressions. The parameters μ_r and μ_b still represent the intercepts, and their difference represents the difference between bromide and rubidium concentrations, now not only at (the extrapolated) hour 0 but at all hours of the experiment. The estimated common slope is $\hat{\gamma} = 0.180$ with a standard error of 0.0085 and a t-statistic of 21.2 for the hypothesis H_0: $\gamma = 0$ (which of course is strongly significant in $t(7)$).

The last question, whether the two regressions are identical, can either be answered from a t-test based on the estimated difference in the intercepts ($\hat{\mu}_r - \hat{\mu}_b = 5.26.$ with SE = 0.4687) or by an F-statistic derived from the sequential SS in the above ANOVA table: F=[69.169/1] / 0.57 = 121, which is strongly significant in $F(1, 6)$ as well.

Table of parameter estimates (with standard errors) from models (1)−(3):

	rubidium			bromide		
model	interc.	slope	st.dev.	interc.	slope	st.dev.
(1): separate models	3.295 (.501)	0.173 (.0072)	0.443	-2.922 (1.099)	0.188 (.016)	0.971
(2): separate regressions	3.295 (.854)	0.173 (.012)	0.755	-2.922 (.854)	0.188 (.012)	0.755
(3): parallel regressions	2.816 (.638)	0.180 (.0085)	0.741	-2.444 (.638)	0.180 (.0085)	0.741

(Technical note: In order to get standard errors in Minitab for all parameter estimates of the regression

line, models (2) and (3) must be formulated using dummy variables without an overall intercept term.)

In conclusion, the parallel regressions model is the best model considered for these data, and it does seem to give a quite accurate description of the data. The relevant parameter estimates are given in the table. Clearly, one would want more data points to validate the regressions and obtain a more firmly based estimate of the variation about the line.

Additional notes

Above, little has been said above model checking; the graph shows a reasonable fit of the two lines to the data points. However, in the rubidium sample a closer look (involving residuals and diagnostics) reveals that the third is seriously below the line formed by the other points. Its deletion residual is huge (and very significant) in the regression for rubidium only, because the standard deviation about the line of the four other points is very, very small. With the small dataset it seems most reasonable to attribute this to a coincidence, but it would seem appropriate for the experimentator to take a closer look at that observation.

Another question is whether the data can be analysed without taking into account that the rubidium and bromide values are obtained from the *same* samples (potatoes). This might introduce a correlation between the two values, and one might be tempted to analyse the two variable separately instead of putting them into a common model. First, it *is* possible to analyse them as two outcomes from the same experiment using a so-called MANOVA (multivariate ANOVA) analysis (Section 12.3 in RC), and it shows only weak (and negative) correlation between the errors of the two values. This analysis is beyond the course, but it does justify the analysis without taking the correlation into account. Second, by analysing them separately one does not at all take into account a possible correlation between the variables, and it becomes more difficult to compare the results.