Solution to Exercise 14.2 (Version 1, 1/1/16)

from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8

Exercise 14.2

Samples of foliage from plots of red pine were analysed to establish whether foliar nutrients could predict growth (Bliss, 1970, Exercise 18.8). File FOLIAR.DAT contains the plot number (*Plot*) with the quantity (mg) of potassium (variate *K*) and calcium (variate *Ca*) found in foliar samples of given weight (variate *SampleWt*, g) together with the increase in height (variate *IncHt*, ft) and basal area (variate *IncBA*, sq ft/acre) over a five-year period. Construct biologically-meaningful explanatory variates and fit MLR models for height and basal area. Comment on the fit of your models, examine them for any evidence of misspecification, and write down and interpret the best predictive model in each case. Give a 95% CI for the increase in height and basal area for a plot with 100 mg K and 20 mg Ca in a sample of weight 20 g.

Data 14.2 (FOLIAR.DAT)

Plot	K	Ca	SampleWt	IncHt	IncBA
1	104.3	26.5	22.00	9.95	170
3	70.8	22.2	17.34	8.30	153
9	105.0	30.0	18.28	9.10	164
15	95.5	42.8	21.58	10.75	157
16	123.2	60.8	25.75	9.78	152
17	78.3	36.3	17.63	11.20	152
20	128.5	49.7	26.61	10.40	148
21	58.7	22.5	12.67	8.43	163
23	136.0	56.7	25.65	10.82	165
25	72.0	19.2	16.63	6.03	154
30	144.5	62.5	23.99	12.32	183
31	79.3	38.8	21.10	11.47	144
36	76.2	38.5	20.40	10.30	144
39	102.0	37.7	19.81	8.88	166
40	120.2	18.7	22.30	5.70	152
41	57.3	14.7	13.98	5.58	139
42	77.3	25.3	18.27	7.23	156
43	79.3	26.5	20.47	8.80	150

Solution 14.2

We have two response variates here, increase in height (*IncHt*) and increase in basal area (*IncBA*) and three potential explanatory variates: sample weight (*SampleWt*) and potassium (*K*) or calcium (*Ca*) content of each sample. In order to relate growth to nutrients, we need to standardize to a common sample size; we will use 20g as this is close to the average sample weight. We thus calculate the nutrient concentrations in a 20g sample as

$$Kconc = K / (SampleWt/20)$$
$$Caconc = Ca / (SampleWt/20)$$

We will look at a multiple regression using both nutrient variates for each response in turn. With only two explanatory variables, we can fit the model in both orders and look at the sequential ANOVA tables to identify the best predictive model.

Response variate: *IncHt*

We start with exploratory analysis. A scatterplot matrix for the two explanatory variates and *IncHt* is in Figure S14.2.1. The corresponding correlation matrix takes the form

```
  Kconc          -
 Caconc      0.3319     -
  IncHt      0.1629   0.8263       -
             Kaconc   Caconc     IncHt
```

The two explanatory variates have a small positive correlation. Calcium concentration (*Caconc*) has a large positive correlation with *IncHt*, which looks approximately linear, but the potassium concentration (*Kaconc*) appears to have no direct relationship with *IncHt*.

Our first multiple regression model can be written in symbolic form as

Response:	*IncHt*
Explanatory component:	*[1]* + *Caconc* + *Kconc*

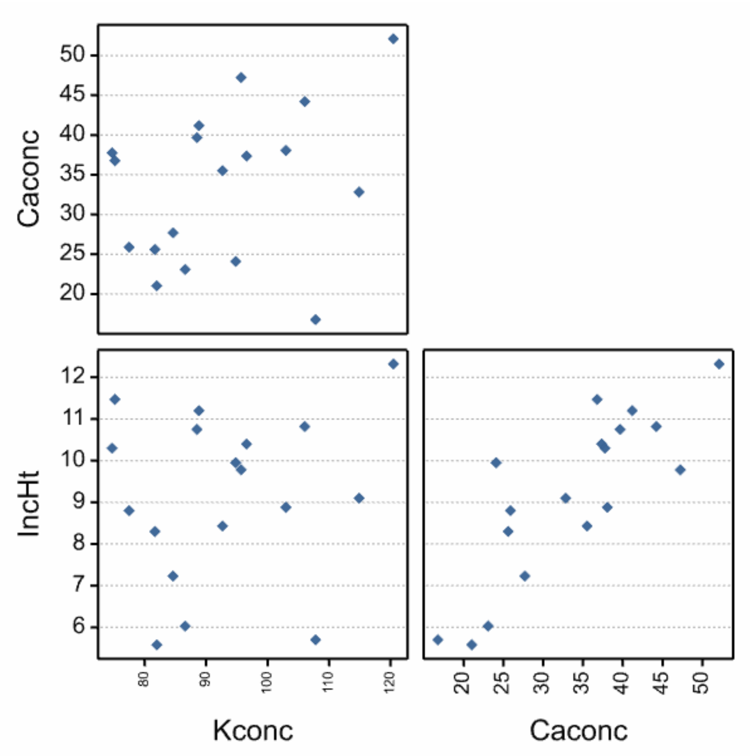

Figure S14.2.1. Scatterplot matrix of explanatory variates with response variate *IncHt*.

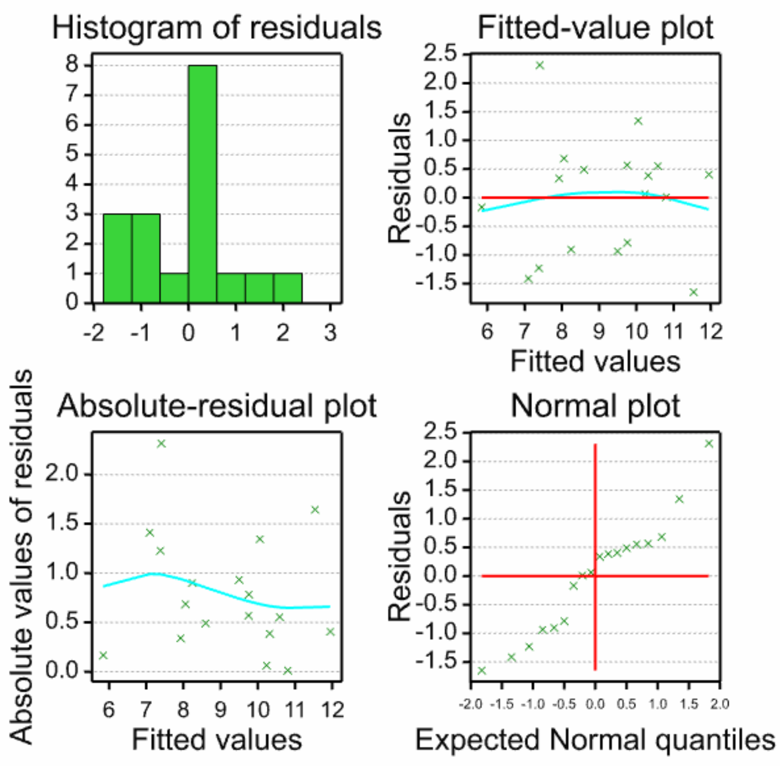

Figure S14.2.2. Composite set of residual plots based on standardized residuals from MLR model for *IncHT* with both explanatory variates.

A composite set of residual plots from this model is shown in Figure S14.2.2. The residual plots are not wholly consistent with the assumptions of Normal deviations but this is a small data set, with only 18 observations, and there are no systematic patterns so we judge them acceptable. The sequential ANOVA table is in Table S14.2.1. After accounting for calcium concentration, there is no evidence ($F_{1,15}$ = 0.689, P = 0.420) that potassium concentration (*Kconc*) explains any variation in increase in height. If we fit the variables in the other order (not shown) then, as we might expect from the exploratory plots, there is still no evidence of a relationship with potassium but calcium concentration remains significant. We can therefore simplify the model to a SLR with calcium concentration as the explanatory variate. In symbolic form:

Response: *IncHt*
Explanatory component: *[1] + Caconc*

Table S14.2.1 Sequential ANOVA table for multiple regression with two explanatory variates for response *IncHt*.

Term added	Incremental df	Incremental SS	Mean square	Variance ratio	*P*
+ *Caconc*	1	46.704	46.704	33.763	< 0.001
+ *Kconc*	1	0.953	0.953	0.689	0.420
Residual	15	20.749	1.383		
Total	17	68.405			

Residual plots from the reduced model (not shown) remain acceptable, as does a plot of the fitted model. The summary ANOVA table is shown in Table S14.2.2 and the parameter estimates in Table S14.2.3. The calcium concentration accounts for 66.3% of the variation in height increase (adjusted R^2 = 0.663). The fitted model can be written in terms of generic variables as

$$IncHt(Ca) \;=\; 3.410 + 0.171Ca\,,$$

and this formula can be used to make predictions of increase in height. The question asks us to predict for 100 mg of K and 20mg Ca in a sample of 20g. We have standardized to a sample of 20g, so we do not need to adjust the quantities, and we have found no relationship with K, so we do not use this in the prediction. We therefore predict as

$$IncHt(Ca = 20) \;=\; 3.410 + 0.171 \times 20 \;=\; 6.826\,.$$

We can use statistical software to get a SE for this prediction, then calculate a 95% CI for the prediction as

$$\left(\text{Prediction} - \text{SE} \times t_{16}^{0.975},\, \text{Prediction} + \text{SE} \times t_{16}^{0.975}\right),$$

which evaluates as

(5.799, 7.853).

Table S14.2.2 Summary ANOVA table for SLR with response *IncHt* and *Caconc* as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	P
Model	1	46.704	46.704	34.433	< 0.001
Residual	16	21.702	1.356		
Total	17	68.405			

Table S14.2.3 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (P) for a SLR model for increase in height (*IncHt*) with explanatory variate *Caconc*.

Term	Parameter	Estimate	SE	t	P
[1]	α	3.410	1.0191	3.346	0.004
Caconc	β	0.171	0.0291	5.868	< 0.001

Response variate: *IncBA*

We follow a similar path with increase in basal area. The summary correlation matrix takes the form

```
  Kconc         -
Caconc      0.3319        -
 IncBA      0.7513   0.3893        -
            Kaconc   Caconc    IncBA
```

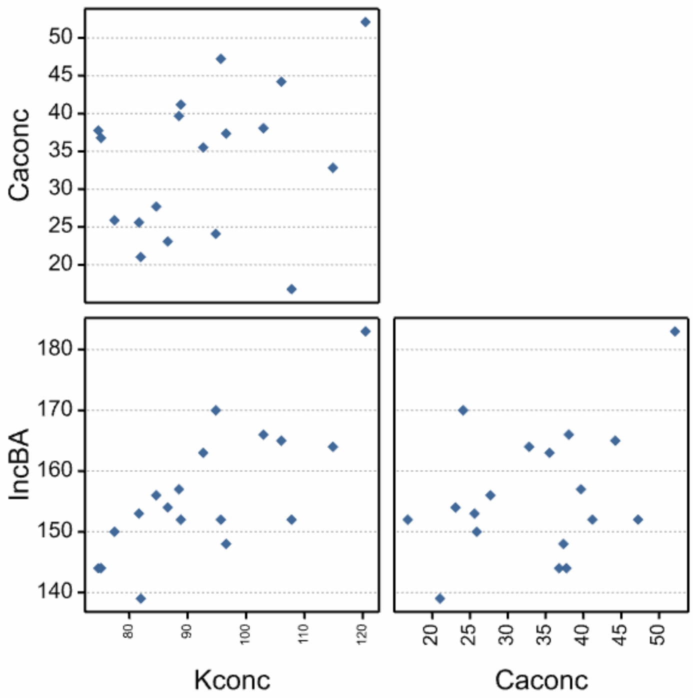

Figure S14.2.3. Scatterplot matrix of explanatory variates with response variate *IncBA*.

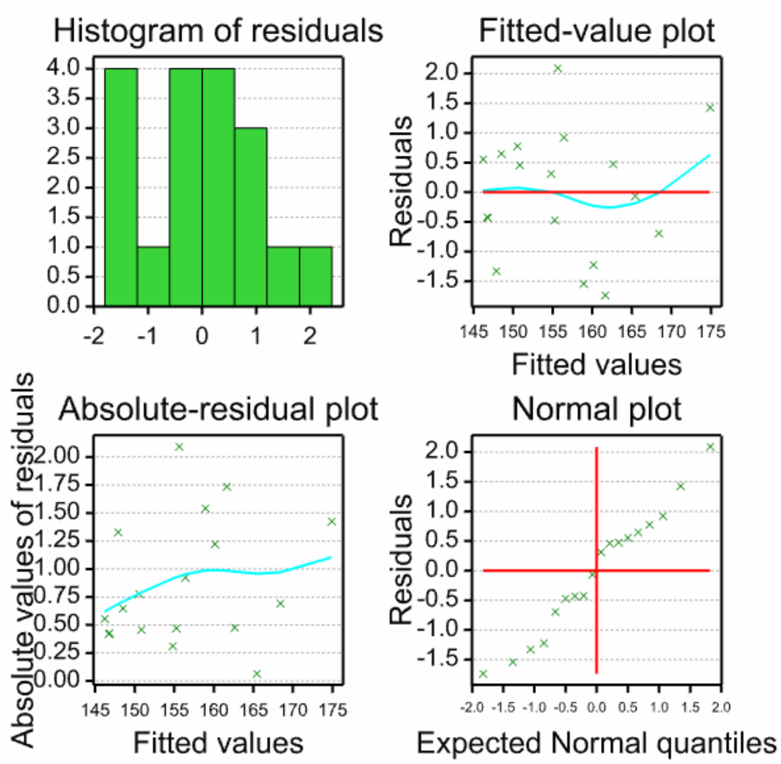

Figure S14.2.4. Composite set of residual plots based on standardized residuals from MLR model for *IncBA* with both explanatory variates.

Both explanatory variables have a positive correlation with increase in basal area, but the relationship is stronger for potassium (*Kconc*) here. Our first multiple regression model can be written in symbolic form as

Response: *IncBA*
Explanatory component: *[1] + Kconc + Caconc*

A composite set of residual plots from this model is shown in Figure S14.2.4, and again are broadly acceptable although there is a slight suggestion of variance heterogeneity. The sequential ANOVA table is in Table S14.2.4. After accounting for potassium concentration, there is no evidence ($F_{1,15}$ = 0.798, P = 0.386) that calcium concentration explains any variation in increase in basal area. If we fit terms in the other order then both terms show evidence of explaining variation in the response. However, since we do not need calcium once we have included potassium then, in the interests of parsimony, we simplify the model to a SLR with potassium concentration as the only explanatory variate. In symbolic form:

Response: *IncBA*
Explanatory component: *[1] + Kconc*

Table S14.2.4 Sequential ANOVA table for multiple regression with two explanatory variates for response *IncBA*.

Term added	Incremental df	Incremental SS	Mean square	Variance ratio	P
+ *Kconc*	1	1104.82	1104.82	20.479	< 0.001
+ *Caconc*	1	43.06	43.06	0.798	0.386
Residual	15	809.23	53.95		
Total	17	1957.11			

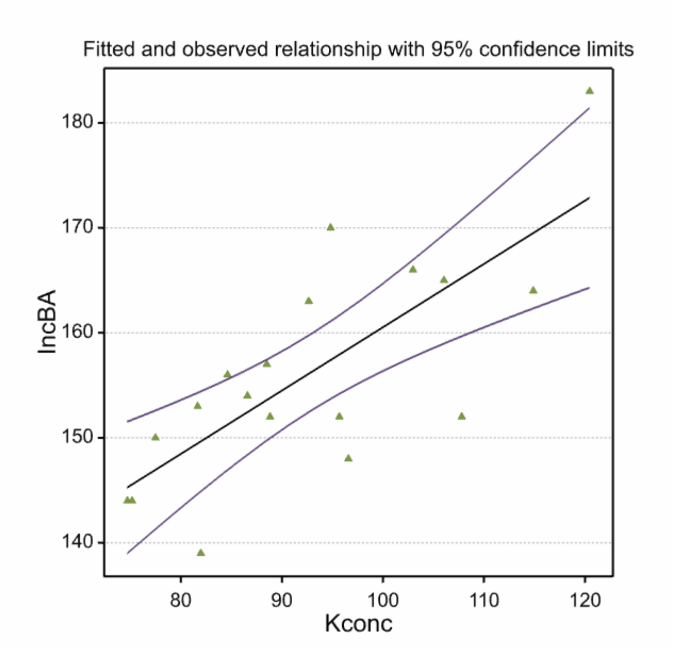

Figure S14.2.5. Fitted SLR for response *IncBA* with explanatory variable potassium concentration (*Kconc*).

Residual plots from the reduced model (not shown) remain acceptable, although there is a strong suggestion of variance heterogeneity. A plot of the fitted model (Figure S14.2.5) shows that there is more scatter about the fitted line at higher values of potassium concentration, but more data would be required to diagnose variance heterogeneity with any confidence.

The summary ANOVA table is shown in Table S14.2.5 and the parameter estimates in Table S14.2.6. The potassium concentration accounts for 53.7% of the variation in increase of basal area (adjusted R^2 = 0.537). The fitted model can be written in terms of generic variables as

$$IncBA(K) \ = \ 100.22 + 0.603K\,,$$

and this formula can be used to make predictions of increase in basal area. The question asks us to predict for 100 mg of K and 20mg Ca in a sample of 20g. Here we have found no relationship with Ca, so we do not use this in the prediction. We therefore predict as

$$IncBA(K = 100) \ = \ 100.22 + 0.603 \times 100 \ = \ 160.53\,.$$

We can use statistical software to get a SE for this prediction, then calculate a 95% CI for the prediction as previously to get

(156.37, 164.69).

Table S14.2.5 Summary ANOVA table for SLR with response *IncBA* and *Kconc* as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
Model	1	1104.82	1104.82	20.741	< 0.001
Residual	16	852.29	53.27		
Total	17	1957.11			

Table S14.2.6 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for increase in basal area (*IncBA*) with explanatory variate *Kconc*.

Term	Parameter	Estimate	SE	t	*P*
[1]	α	100.22	12.416	8.072	< 0.001
Kconc	β	0.603	0.1324	4.554	< 0.001