

Data Set 5: Effects of Nutrients on Shrimp Growth

Statistical setting

This Handout is an example of extreme collinearity of the independent variables, and of the methods used for diagnosing this problem.

Background and Data

These data are from a study by Shaun Moss (ex-Zoology grad student, now Director of the Shrimp Department at the Oceanic Institute). He was interested in the effects of various water filtration treatments on the growth of Pacific white shrimp, *Litopenaeus vannamei*, a species used in the aquaculture industry. He found very strong treatment effects on shrimp growth, with shrimp grown in less finely filtered water growing faster than those in finely filtered water. Shaun then wanted to determine which of several water nutrient parameters were responsible for this effect. Unfortunately, the parameters were very strongly correlated with each other – they all were affected similarly by the filtering treatments – making this determinations essentially impossible.



The response variable was the mean growth (mg/day) of a group of shrimp reared together in a tank. The water parameters — the independent variables — were particulate organic carbon ('POC'; mg/l), ATP (ug/l), algal numbers ('AN'; $10^7/l$) and bacterial numbers ('BN'; $10^9/l$). There were five filtration treatments with three replicates (tanks) each, for a total sample size of $n = 15$. All four water variables were higher in the less-filtered treatments, in which shrimp grew faster.

The Data

POC	ATP	AN	BN	GROWTH
3.645	5.060	9.195	6.431	1.373
3.967	6.425	13.222	6.188	1.480
3.695	5.682	17.388	5.921	1.597
2.258	1.923	4.972	3.574	1.151
2.048	2.388	9.003	3.724	1.181
2.366	1.623	8.473	3.838	1.263
1.762	2.127	7.500	5.270	1.256
2.064	2.550	10.105	4.003	1.258
1.845	1.748	9.028	4.244	1.070
7.188	13.496	32.237	7.272	1.706
6.440	11.850	25.195	7.522	1.918
7.894	11.994	26.256	6.632	1.860
0.416	0.182	0.000	0.356	1.035
0.436	0.149	0.000	0.390	1.047
0.386	0.131	0.000	0.389	0.833

Data Exploration

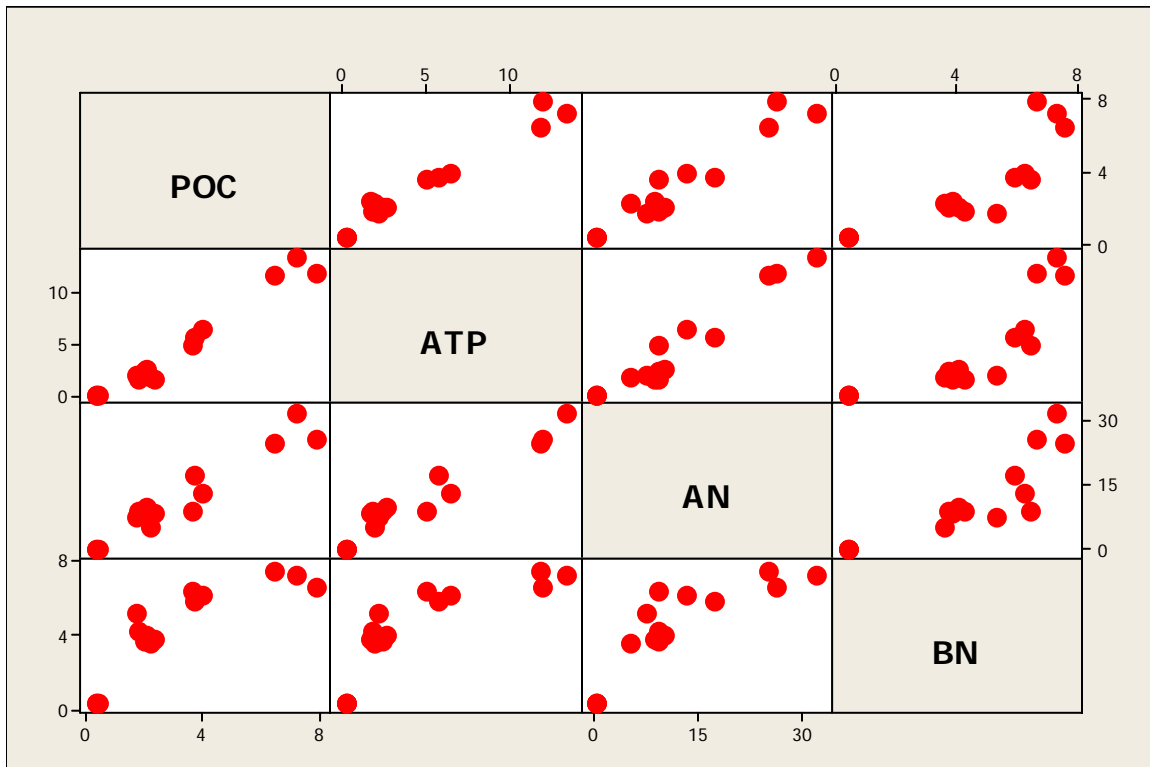
Collinearity

Correlations

	ATP	AN	BN
POC	0.9819	0.9607	0.8539
ATP		0.9675	0.8271
AN			0.8542
BN			

All four variables are strongly correlated, with the smallest correlation coefficient being a very high 0.8271 (ATP and BN); ATP and POC are almost perfectly correlated ($r = 0.9819$).

Scatterplots



These plots show visually the very high correlations among the variables. In addition they show that the slightly lower correlations between BN and the other variables are at least partly because the associations, while strong, are somewhat nonlinear.

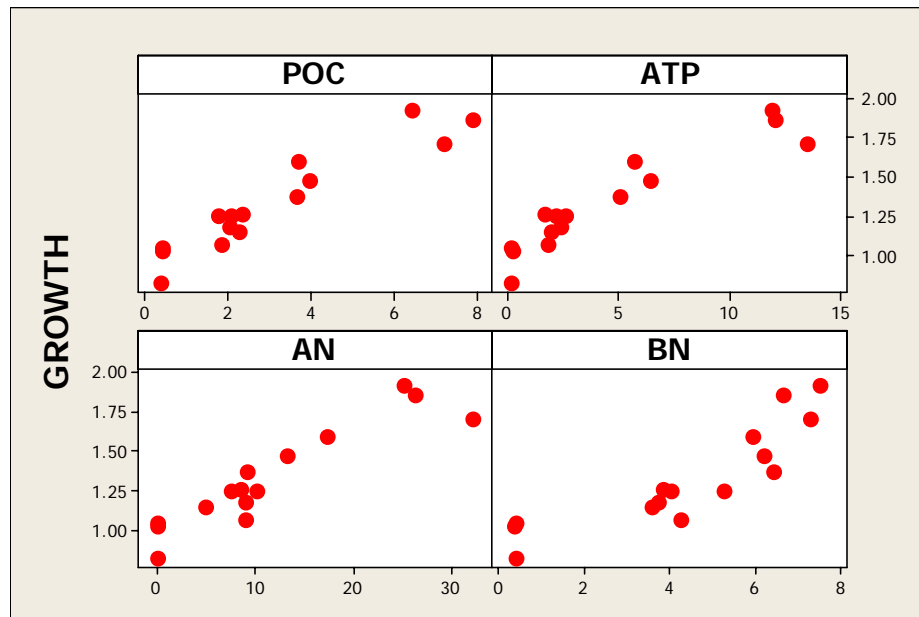
Relationships with Response Variable

Correlations

	POC	ATP	AN	BN
GROWTH	0.946	0.938	0.925	0.872

Shrimp growth was very strongly positively associated with all four nutrient variables; once again the smallest (though still very large) correlation was with BN.

Scatterplots



These plots illustrate the very strong relationships, and again suggest that the slightly lower correlation with BN reflects slight curvature in that relationship.

Collinearity Diagnostics

Variance Inflation Factors

Variable	DF	Variance Inflation
INTERCEPT	1	0.00000000
POC	1	33.28180698
ATP	1	38.01152362
AN	1	18.43683379
BN	1	4.18920542

There VIF values – three of them well above 10 – clearly show the problem in these data. It is noteworthy that the VIF for BN is considerably less: although BN is strongly correlated with the other variables, the difference between correlations around 0.85 (between BN and the others) and correlations above 0.96 (among the other three variables) turns out to be quite substantial in terms of collinearity. But even that VIF of 4.19 indicates a problem.

Condition Indices

Intercept Adjusted:

	Eigenvalue	Condition Index	Variance Proportions			
			POC	ATP	AN	BN
1	3.72607	1.00000	0.0021	0.0018	0.0038	0.0143
2	0.21675	4.14620	0.0080	0.0162	0.0109	0.8430
3	0.04183	9.43772	0.2002	0.0356	0.8587	0.0141
4	0.01535	15.57846	0.7898	0.9464	0.1266	0.1285

Including Intercept:

	Eigenvalue	Condition Index	Variance Proportions				
			Intercept	POC	ATP	AN	BN
1	4.57073	1.00000	0.0055	0.0005	0.0006	0.0010	0.0024
2	0.35828	3.57175	0.2859	0.0012	0.0065	0.0046	0.0043
3	0.04798	9.75984	0.4790	0.0055	0.0225	0.0022	0.8224
4	0.01662	16.58107	0.0100	0.1434	0.0643	0.9076	0.0329
5	0.00638	26.77196	0.2196	0.8494	0.9061	0.0845	0.1379

These measures give a somewhat different conclusion than do the variance inflation factors: with the possible exception of the last eigenvalue in the analysis including the intercept, none of these eigenvalues indicate severe problems of ill-conditioning. That last eigenvalue suggests that the greatest collinearity is between POC and ATP; the intercept-adjusted analysis, as well as the simple bivariate correlations given early, also indicate this as the greatest problem.

It should be noted that while the condition indices (with the one possible exception) do not exceed the threshold for possible computational problems from ill-conditioning, they do confirm that there is substantial confounding among the variables: over 90% of the variation in the independent variables can be attributed to the first principal component which simply measures how much of all the nutrients there was.

Regression Analyses

The nature of the problem posed by these correlations among explanatory variables can be seen in regression analyses for the four-variable model, contrasted with the four different one-variable models, as well as in the model-selection results. (For this study formal hypothesis testing is less appropriate than exploratory model selection, since there were no *a priori* hypotheses. Both approaches are shown here simply to illustrate the problems caused by collinearity.)

Model with All Four Independent Variables

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	1.29235	0.32309	26.479	0.0001
Error	10	0.12202	0.01220		
C Total	14	1.41437			
	Root MSE	0.11046	R-square	0.9137	
	Dep Mean	1.33520	Adj R-sq	0.8792	
	C.V.	8.27307			

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	0.922350	0.07316453	12.607	0.0001
POC	1	0.053055	0.07093303	0.748	0.4717
ATP	1	0.023377	0.04002118	0.584	0.5721
AN	1	0.000110	0.01287517	0.009	0.9933
BN	1	0.032508	0.02473841	1.314	0.2182

The effect of the multicollinearity is striking: while the overall model is highly significant and explains a very high proportion of the variability in shrimp growth, none of the individual variables is close to significant in its marginal contribution, *i.e.* in a model containing the other variables.

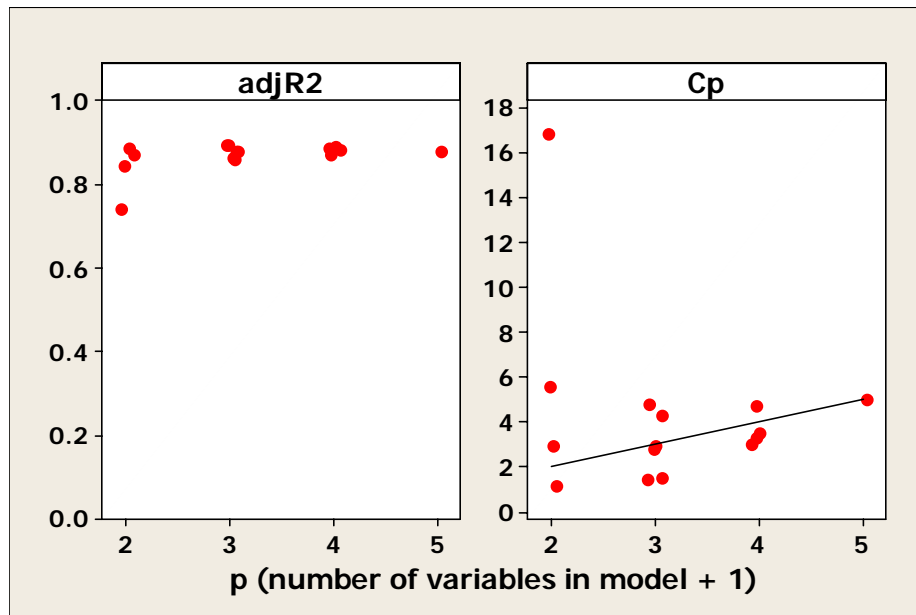
Single-Variable Models

Variable	Estimate	Std. Error	<i>t</i>	<i>P</i>	<i>R</i> ²
POC	0.1252	0.01191	10.505	0.0001	0.895
ATP	0.0656	0.00672	9.753	0.0001	0.880
AN	0.0299	0.00339	8.809	0.0001	0.857
BN	0.1134	0.01769	6.412	0.0001	0.760

These results obviously are very different from those for each parameter in the full model above: taken one at a time, not only are the estimates quite different (as is usually the case with even moderate confounding), but the standard errors are much smaller, so that now each variable is very highly significant when considered alone.

Model Selection: All Possible (First-Order) Models

	R-square	Adjusted R-square	.C(p)	Vars in Model			
1	0.89462113	0.88651506	1.21490	POC			
	0.87977354	0.87052535	2.93594		ATP		
	0.85651770	0.84548060	5.63162			AN	
	0.75974693	0.74126592	16.84872				BN
2	0.90971837	0.89467143	1.46491	POC			BN
	0.90885646	0.89366587	1.56482		ATP		BN
	0.89828535	0.88133291	2.79016	POC		AN	
	0.89698691	0.87981806	2.94067	POC	ATP		
	0.88482695	0.86563145	4.35018		ATP	AN	
	0.88083554	0.86097480	4.81284			AN	BN
3	0.91372858	0.89020001	3.00007	POC	ATP		BN
	0.91078573	0.88645456	3.34119	POC		AN	BN
	0.90890281	0.88405812	3.55945		ATP	AN	BN
	0.89883242	0.87124126	4.72675	POC	ATP	AN	
4	0.91372921	0.87922090	5.00000	POC	ATP	AN	BN



With the exception of the model with only BN, all these models are essentially indistinguishable by the R^2_{adj} criterion. The C_p criterion distinguishes among them somewhat better, but even using it the three best models have almost identical values despite being fairly different in the variables they include (POC only, POC + BN, and ATP + BN). Similarly three of the four 3-variable ($p=4$) models have quite similar values of C_p .

Conclusions and Solutions

The only solid conclusion that can be drawn is that while shrimp growth increased with the increase in all the nutrients, it is impossible to determine which nutrients were responsible since they are so tightly correlated. The results suggest that perhaps BN has slightly less of an effect on shrimp growth than do the other parameters, but otherwise the slight differences in R^2 for the single-variable models are meaningless. The differences in coefficients in the four-variable model (even if they had been from a standardized regression model) also are made meaningless by the instability in estimates caused by the collinearity.

There is not likely to be a statistical remedy to these problems. There is no statistical basis for removing some variables rather than others. Principal components regression (not shown) finds that the first principal component of the X 's, which simply measures the amount of all nutrients together, had a strong effect on growth, but no other principal components (which reflect relative differences in amounts of different nutrients) have noticeable effects.

The proper way to address this situation, instead, is to obtain new data in which the nutrients are not so highly correlated: to experimentally manipulate the levels of the nutrients separately. [Since the question of which nutrient(s) were important was at best peripheral to Shaun's study, he chose not to conduct such an experiment.]