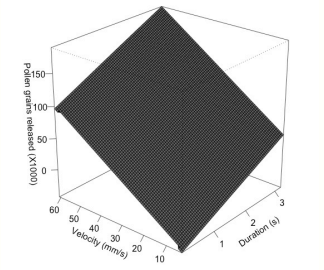
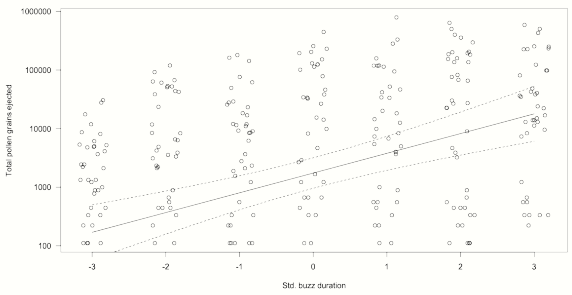
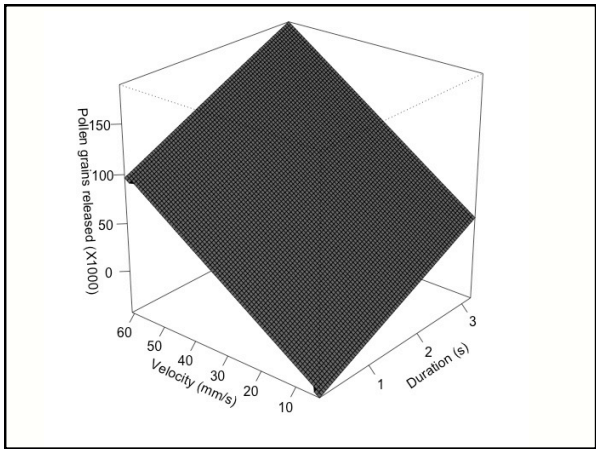
Advancing in R
Multiple regression

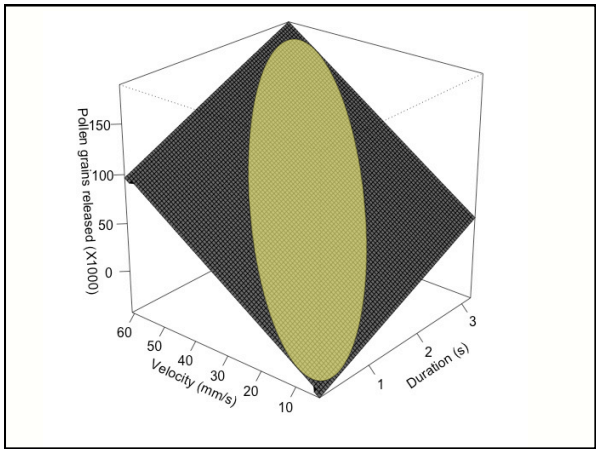


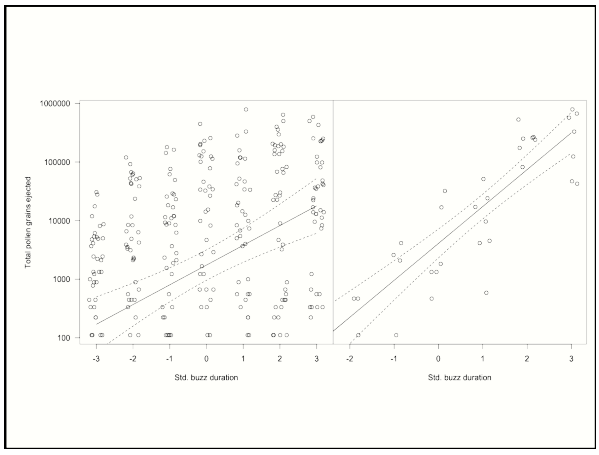
Outline

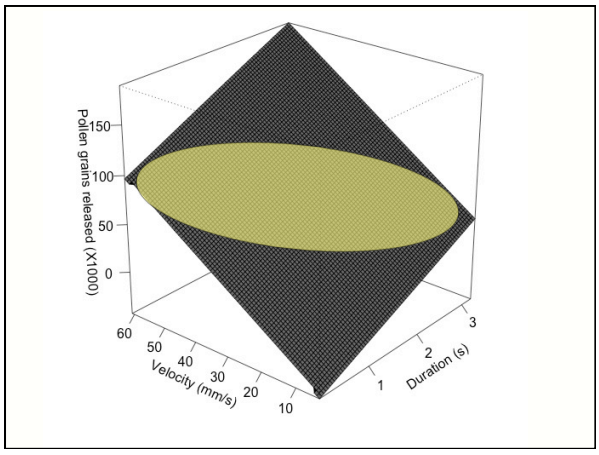
- Why multiple regression is necessary
- Multicollinearity & variance inflation
- Partial effects
- Parsimony & model simplification
- Centering and scaling (z-transformation)




Pollen grains released (X1000)
150
100
50
0
60
50
40
30
20
10
Velocity (mm/s)
1
2
3
Duration (s)


Pollen grains released (X1000)
150
100
50
0
60
50
40
30
20
10
Velocity (mm/s)
1
2
3
Duration (s)


Total pollen grains ejected
1000000
100000
10000
1000
100
-3
-2
-1
0
1
2
3
Std. buzz duration
-2
-1
0
1
2
3
Std. buzz duration


Pollen grains released (X1000)
150
100
50
0
60
50
40
30
20
10
Velocity (mm/s)
1
2
3
Duration (s)

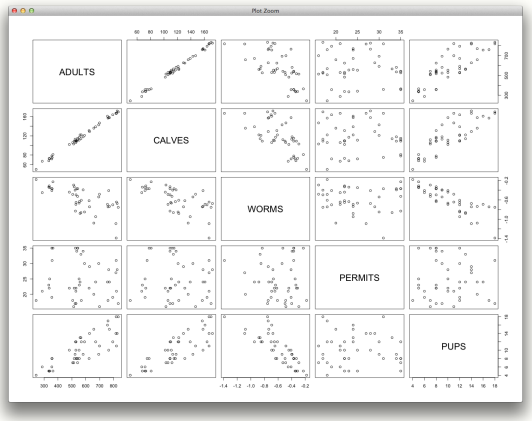
Multiple regression

- Estimates the *partial* contributions of predictors to the response *after* controlling for other variables

Multicollinearity

- Highly correlated variables can interfere with parameter estimation
- Causes increase in coefficient values called “variance inflation”
- Examine correlation plots & measure variance inflation factors (vifs)
- If vif > 10, need to adjust model
- If vif > 4-5, should investigate

```
> library(car)

> vif(MOD.1)
    ADULTS      CALVES       WORMS     PERMITS
163.604950 161.046727   1.785819   1.020219


> vif(MOD.2)
  ADULTS     WORMS   PERMITS
1.770966 1.773457 1.018879
```



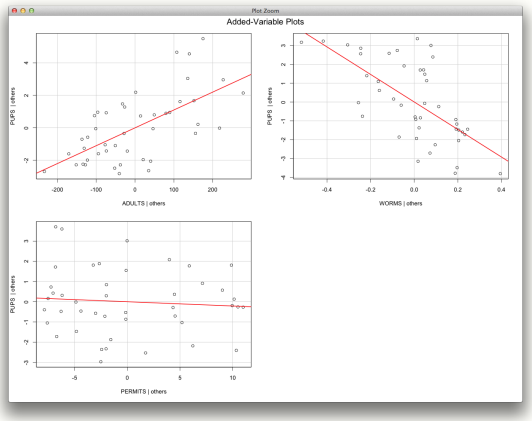

```
> summary(MOD.1)

Call:
lm(formula = PUPS ~ ADULTS + CALVES + WORMS + PERMITS, data =
WP)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1965 -0.8533 -0.0768  0.8772  3.7628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.37758    1.48397  -0.254    0.800
ADULTS      -0.01152    0.02000  -0.576    0.568
CALVES       0.11144    0.09845   1.132    0.265
WORMS       -7.43356    1.33936  -5.550 2.18e-06 ***
PERMITS     -0.01940    0.04163  -0.466    0.644
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.666 on 39 degrees of freedom
Multiple R-squared: 0.8143,   Adjusted R-squared: 0.7953
F-statistic: 42.76 on 4 and 39 DF,  p-value: 9.295e-14
```

```
> summary(MOD.2a)

Call:
lm(formula = PUPS ~ CALVES + WORMS + PERMITS, data = WP)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1134 -0.8653 -0.1479  0.9023  3.6396

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17554    1.42983  -0.123    0.903
CALVES       0.05504    0.01016   5.418 3.10e-06 ***
WORMS       -7.31703    1.31289  -5.573 1.88e-06 ***
PERMITS     -0.01996    0.04127  -0.484    0.631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.652 on 40 degrees of freedom
Multiple R-squared: 0.8127,   Adjusted R-squared: 0.7987
F-statistic: 57.87 on 3 and 40 DF,  p-value: 1.31e-14
```

Partial effects

- Estimate the partial contribution of a predictor to the response *after* controlling for other variables in the multiple regression
- Plots
  - Resids of resp var on all other predictors (omitting X)

  vs
  - Resids of X against remaining predictors



Model selection

Principal of Parsimony

- Models should have as few parameters as possible
- Linear models are preferred to nonlinear ones.
- The fewer assumptions the better.
- Models should be reduced until they are "minimally adequate".
- Simple explanations are preferred to complex ones.
- Avoids problems with overfitting your data.

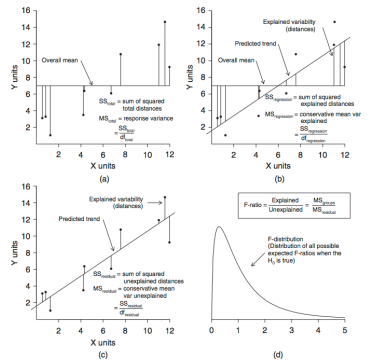


Fig 8.3 Fictitious data illustrating the partitioning of (a) total variation into components (b) explained ($MS_{regression}$) and (c) unexplained ($MS_{residual}$) by the linear trend. The probability of collecting our sample, and thus generating the sample ratio of explained to unexplained variation (or one more extreme), when the null hypothesis is true (and there is no relationship between X and Y) is the area under the F-distribution (d) beyond the sample F-ratio.

Sequential F-test

- Computes the increase in error SS (residual or unexplained variation) when a predictor is removed from the model
- The amount by which error SS increases has an F-distribution under the null hypothesis
- Use a sequential or partial F-test (these are slightly different) to determine whether a predictor significantly contributes to model fit.

In more advanced models, will use Likelihood ratio test

- Used to compare some models
- Likelihood ratio = how many times more likely are the data under one model vs the other
- Typically, use log of this ratio, the log-likelihood ratio statistic, which has a known χ^2 probability distribution

Deviance

- Measure of goodness of fit of a LM
- = -2 X the difference in log-likelihood between current model and saturated model
- Minimizing deviance = maximizing likelihood

Steps in Model Simplification see Ch. 9 of Logan or *The R Book*, or Ch. 7 of Statistics…

1) Fit the maximal model
2) Begin simplification starting with least significant highest order interactions
3) If deletion causes a nonsig incr. in deviance, leave out the term
4) If the increase in deviance is sig, keep the term
5) Repeat 2-4 until no more terms removeable (NB: cannot remove main effects if involved in higher order interactions that are retained)

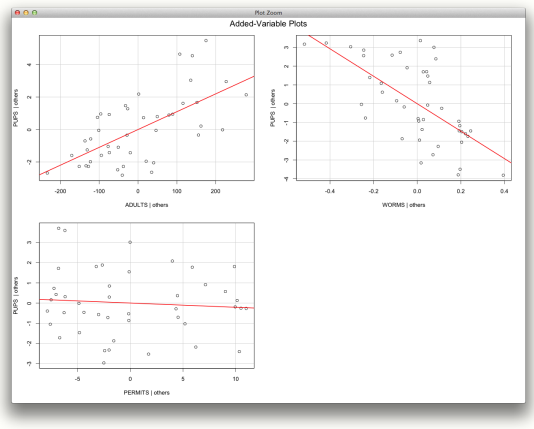
```
> MOD.3<-update(MOD.2, ~. - PERMITS)
> anova(MOD.2,MOD.3)
Analysis of Variance Table

Model 1: PUPS ~ ADULTS + WORMS + PERMITS
Model 2: PUPS ~ ADULTS + WORMS
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     40 111.84
2     41 112.55 -1  -0.71478 0.2556 0.6159
```

Centring and scaling

- Coefficients (B) provide slopes and intercept deviations in units of the predictor and response
- To compare effects across predictors having different variances and units of measurement, we can centre and scale predictors
- Produces “standardised regression coefficients”, denoted β
- Centering variables can also help reduce variance inflation when modelling interactions

```
> summary(MOD.3)

Call:
lm(formula = PUPS ~ CALVES + WORMS, data = WP)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0643 -0.8662 -0.1790  0.7327  3.7759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.693439   0.938831  -0.739    0.464
CALVES       0.055646   0.009984   5.573 1.75e-06 ***
WORMS       -7.235661   1.289849  -5.610 1.55e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.637 on 41 degrees of freedom
Multiple R-squared: 0.8116,       Adjusted R-squared: 0.8025
F-statistic: 88.34 on 2 and 41 DF,  p-value: 1.371e-15
```

```
> summary(MOD.3s)

Call:
lm(formula = PUPS ~ scale(CALVES) + scale(WORMS), data = WP)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0643 -0.8662 -0.1790  0.7327  3.7759

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.2045     0.2468  41.355  < 2e-16 ***
scale(CALVES)   1.8226     0.3270   5.573 1.75e-06 ***
scale(WORMS)   -1.8345     0.3270  -5.610 1.55e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.637 on 41 degrees of freedom
Multiple R-squared: 0.8116,       Adjusted R-squared: 0.8025
F-statistic: 88.34 on 2 and 41 DF,  p-value: 1.371e-15
```



Ecology and Evolution

Open Access

Influence of dietary specialization and resource availability on geographical variation in abundance of butterflyfish

Rebecca J. Lawton & Morgan S. Pratchett

ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville QLD, 4811, Australia

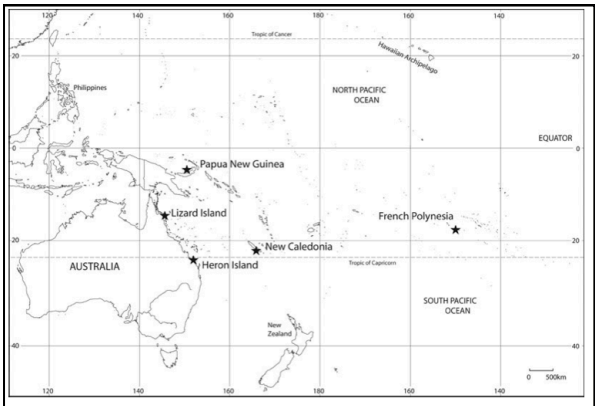


Figure 1. Map of the five locations sampled in this study. Abbreviations used throughout this paper are indicated for each location. Heron Island, Great Barrier Reef (HI); Lizard Island, Great Barrier Reef (LI); Kimbe Bay, Papua New Guinea (PNG); Noumea, New Caledonia (NC); and Moorea, French Polynesia (FP).

Table 1. Predictor variables used in multiple regression analyses for five species of butterflyfish.

Species	Dietary categories	Significant variables	Final model
C. auriga	Noncoral substrate, *Acropora*, other hard corals, other	*Acropora*	–[1]
C. vagabundus	Noncoral substrate, *Acropora*, *Pocillopora*, *Montipora*, *Porites*, other	*Pocillopora*, hard coral species diversity	*Pocillopora*, hard coral species diversity
C. citrinellus	Noncoral substrate, *Acropora*, *Pocillopora*, *Montipora*, *Porites*, Faviidae, other hard corals, other	Number of hard coral species, hard coral species diversity, total coral cover/abundance congenerics	Number of hard coral species, total coral cover/abundance congenerics
C. lunulatus	Noncoral substrate, *Acropora*, *Pocillopora*, *Montipora*, *Porites*, Faviidae, other hard corals	Noncoral substrate, total coral cover/abundance congenerics, number coral species	Noncoral substrate, total coral cover/abundance congenerics, number coral species
C. trifascialis	*Acropora*, *Pocillopora*, *Montipora*	*Acropora*, total coral cover/abundance congenerics, *Montipora*	*Acropora*, total coral cover/abundance congenerics

[1]Final model was not significant.

Table 2. Coefficients of multiple regression models for four species of butterflyfish.

Species	Final predictors	*B*	SE *B*
C. vagabundus	Pocillopora	2.354	0.437
	Hard coral diversity	−0.442	0.181
C. citrinellus	Number of coral species	−0.15	0.003
	Total coral cover/abundance congenerics	−1.201	0.292
C. lunulatus	Noncoral substrate	−0.394	0.086
	Total coral cover/abundance congenerics	−0.897	0.252
	Number coral species	−0.015	0.003
C. trifascialis	*Acropora*	0.590	0.077
	Total coral cover/abundance congenerics	−0.769	0.201

The unstandardized beta coefficients (*B*), their standard errors (SE *B*), and the standardized beta coefficients (*β*) for the predictor variables included in the final regression model for each species are presented.*$P < 0.05$,**$P < 0.01$,***$P < 0.001$.

Table 3. Final multiple regression results for abundance of five species of butterflyfish.

Species	Adjusted R^2	Sum of Squares	df	Mean square	*F*	Significance
C. auriga	−0.004	0.00	1,224	0.000	0.020	0.889
C. vagabundus	0.115	1.546	2,224	0.773	15.550	<0.001
C. citrinellus	0.163	4.194	2,224	2.097	22.751	<0.001
C. lunulatus	0.159	3.491	3,224	1.164	15.114	<0.001
C. trifascialis	0.208	3.212	2,224	1.606	30.340	<0.001

Table 2. Coefficients of multiple regression models for four species of butterflyfish.

Species	Final predictors	*B*	SE *B*	*β*
C. vagabundus	Pocillopora	2.354	0.437	0.344***
	Hard coral diversity	−0.442	0.181	−0.156*
C. citrinellus	Number of coral species	−0.15	0.003	−0.280***
	Total coral cover/abundance congenerics	−1.201	0.292	−0.256***
C. lunulatus	Noncoral substrate	−0.394	0.086	−0.306***
	Total coral cover/abundance congenerics	−0.897	0.252	−0.226***
	Number coral species	−0.015	0.003	−0.320***
C. trifascialis	*Acropora*	0.590	0.077	0.485***
	Total coral cover/abundance congenerics	−0.769	0.201	−0.242***

The unstandardized beta coefficients (*B*), their standard errors (SE *B*), and the standardized beta coefficients (*β*) for the predictor variables included in the final regression model for each species are presented.*$P < 0.05$,**$P < 0.01$,***$P < 0.001$.

Table 3. Final multiple regression results for abundance of five species of butterflyfish.

Species	Adjusted R^2	Sum of Squares	df	Mean square	*F*	Significance
C. auriga	−0.004	0.00	1,224	0.000	0.020	0.889
C. vagabundus	0.115	1.546	2,224	0.773	15.550	<0.001
C. citrinellus	0.163	4.194	2,224	2.097	22.751	<0.001
C. lunulatus	0.159	3.491	3,224	1.164	15.114	<0.001
C. trifascialis	0.208	3.212	2,224	1.606	30.340	<0.001

Suggested reading:

- Ch. 9 in Logan (Mult. regression, vifs, added variable plots)
- Or Chs. 7 (Model selection) and 11 (Mult. regression) in Crawley *Statistics*…
- Or Chs. 9 (Model selection) and 10 (Regression) in Crawley *The R Book*
- Schielzeth 2010, *Methods in Ecology and Evolution* 1:103-113 (Centring and scaling)
- Lawton & Pratchett paper on butterfly fish