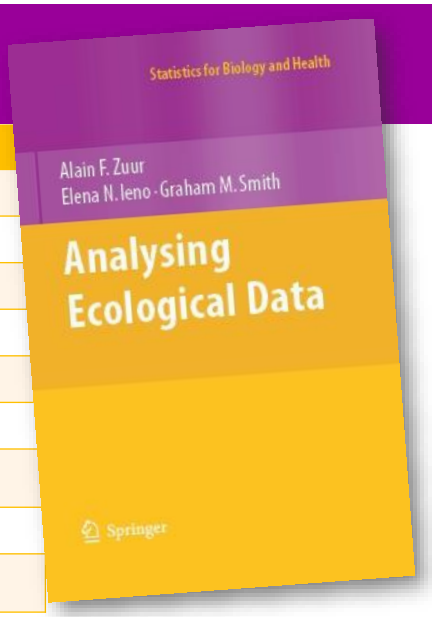
Chapter 5. Linear Regression

MAR 536 Biological Statistics II

January 28 2020

Schedule

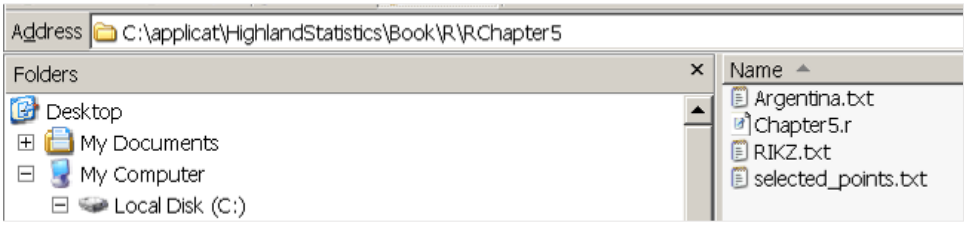
Type	Day	Date	Reading	Topic
Lecture	Tue.	Jan 21	Zuur et al. Chap. 1-4	Introduction, statistical rethinking
Lab	Wed.	Jan 22		R Lab 1
Lecture	Thu.	Jan 23	Bolker 2008 Chap. 4	Probability review
Lecture	Tue.	Jan 28	Zuur et al. Chapter 5	Linear regression review
Lab	Wed.	Jan 29		R Lab 2
Lecture	Thu.	Jan 30		data exploration, checking
Lecture	Tue.	Feb 04	Hilborn & Mangel Chapter 7	Likelihood
Lab	Wed.	Feb 05		R Lab 3
Lecture	Thu.	Feb 06	Zuur et al. Section 6.1	Extending the linear model (GLM) (Poisson)



- Linear Regression Review (based on Zuur et al. 2007 Chapter 5)

Lecture Outline

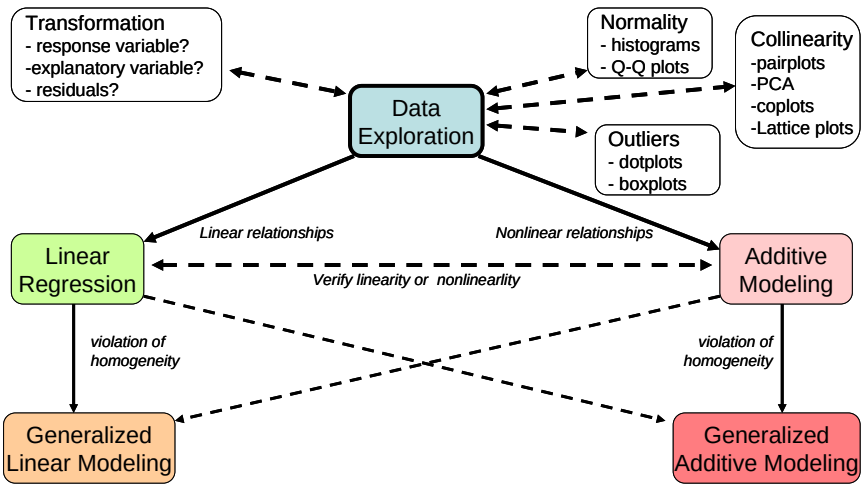
- Bivariate linear regression
  - Back to basics
  - Significance tests
  - Model validation
  - Assessing assumptions
  - Influential points
- Multiple linear regression
  - Model selection
- Partial linear regression
  - Example: variance partitioning



3

Exploratory Approach

1. Inspect scatterplot
2. Fit a regression line
3. Residual analysis
   - GLM for non-normality
   - GAM for nonlinear patterns
- Principles of linear regression underpin GLM and GAM



4

Example Data: Dutch Sandy Beach Community

- Chapter 27
  - Abundance of 75 invertebrate species sampled from 45 sites
  - Species diversity (richness)
  - Vertical position in beach relative to sea level (i.e., 'NAP'<0 for subtidal zone)

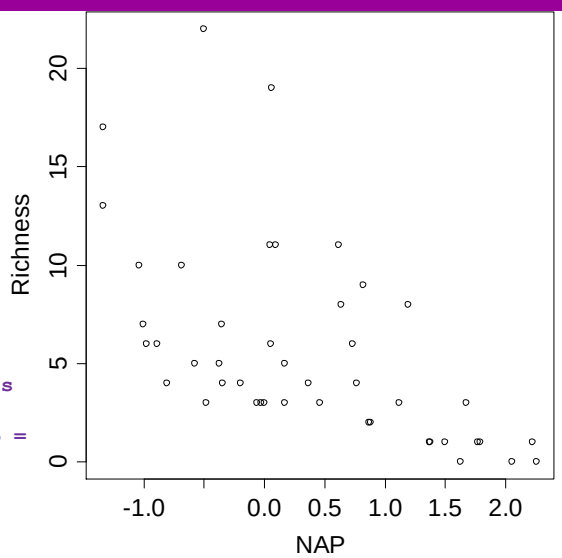
```
RIKZ <- read.table(file = "RIKZ.txt",header = TRUE)
RIKZ$Richness <- rowSums(RIKZ[,2:76] > 0)
```

5

Species Richness by Tidal Zone ('NAP')

- Scatterplot suggests that a linear relationship might be appropriate.

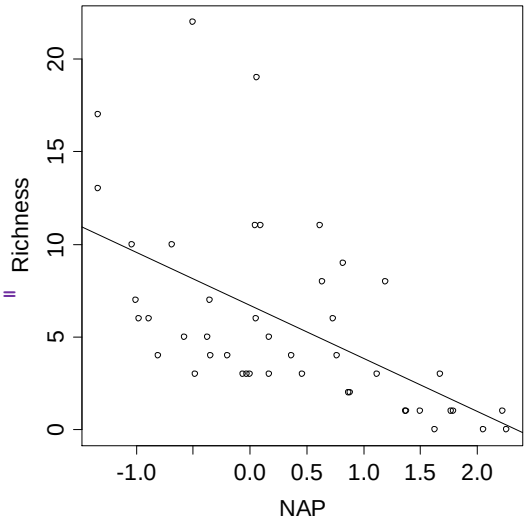
```
par(mar = c(4.5,4.5,0.5,0.5), cex.lab = 1.5, cex.axis
= 1.5)
plot(RIKZ$NAP,RIKZ$Richness, ylab = "Richness", xlab =
"NAP")
```



6

Species Richness by Tidal Zone ('NAP')

- Significant linear trend fit to data.
- Observations are distributed around the predicted line.
- Species richness decreases with elevation in the tidal zone.




```
plot(RIKZ$NAP,RIKZ$Richness, ylab = "Richness", xlab =
"NAP")
RIKZ_model.1<-lm(Richness ~ NAP, data = RIKZ)
abline(RIKZ_model.1)

Analysis of Variance Table
Response: Richness
           Df Sum Sq Mean Sq F value    Pr(>F)
NAP         1 357.53  357.53  20.660 4.418e-05 ***
Residuals 43 744.12   17.31
```


Statistical Model Underlying Linear Regression

- Bivariate linear regression model of response variable (*Y*) and explanatory variable (*X*):
- Model is based on the entire population, but we need to use data for sample estimates, making 4 assumptions:
  1. Fixed X_i
  2. Normality of ε_i
  3. Homogeneity of variance
  4. Independence

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

$$Y_i = a + X_i b + e_i$$

$$\varepsilon_i \approx N\left(0, \sigma_i^2\right)$$

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$$

Significance of Regression Parameters (ANOVA Table)

- The Analysis of Variance (ANOVA) table partitions total variability:

Source	*SS*	*df*	*MS*	*Expected_MS*
Regression	$\sum_n (\hat{Y} - \bar{Y})^2$	1	$\frac{\sum_n (\hat{Y} - \bar{Y})^2}{1}$	$\sigma_\varepsilon^2 + \beta^2 \sum_n (X_i - \bar{X})^2$
Residual	$\sum_n (Y_i - \hat{Y})^2$	$n-2$	$\frac{\sum_n (Y_i - \hat{Y})^2}{n-2}$	σ_ε^2
Total	$\sum_n (Y_i - \bar{Y})^2$	$n-1$	$\frac{\sum_n (Y_i - \bar{Y})^2}{n-1}$	

```
Analysis of Variance Table
Response: Richness
          Df Sum Sq Mean Sq F value     Pr(>F)
NAP        1 357.53  357.53  20.660 4.418e-05 ***
Residuals 43 744.12   17.31
```


Significance of Regression Parameters

- The ANOVA table can be used to test the null hypothesis that the slope is zero (β=0):

```
Analysis of Variance Table
Response: Richness
          Df Sum Sq Mean Sq F value     Pr(>F)
NAP        1 357.53  357.53  20.660 4.418e-05 ***
Residuals 43 744.12   17.31
```

- Or significance of the slope and intercept can be tested using a *t*-ratio:

```
summary(RIKZ_model.1)$coefficients
```

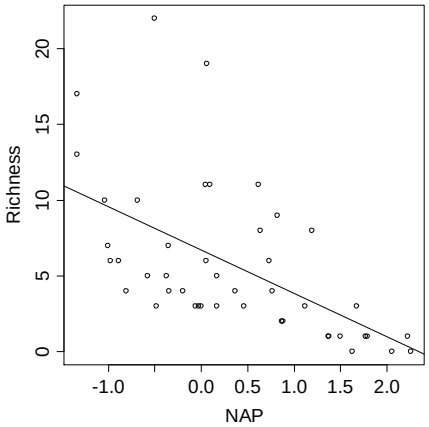
$$t = \frac{b}{s_b}$$

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  6.685662  0.6577579 10.164320 5.251419e-13
NAP         -2.866853  0.6307186 -4.545376 4.417521e-05
```

Coefficient of Determination

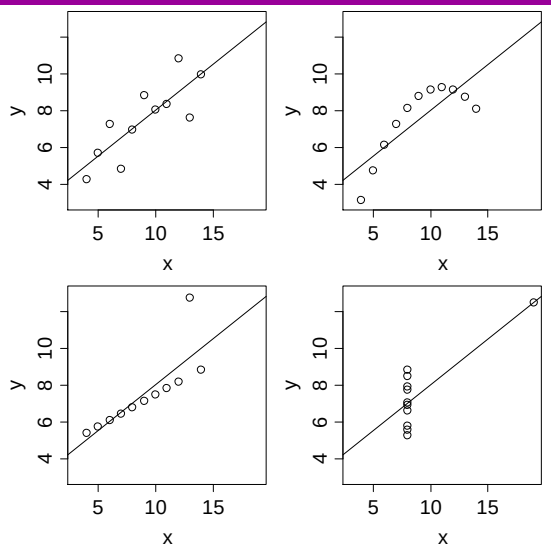
- R^2: Proportion of total variance in *Y* explained by *X*.
- NAP explained 32% of variance in richness (R^2=0.32).

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$



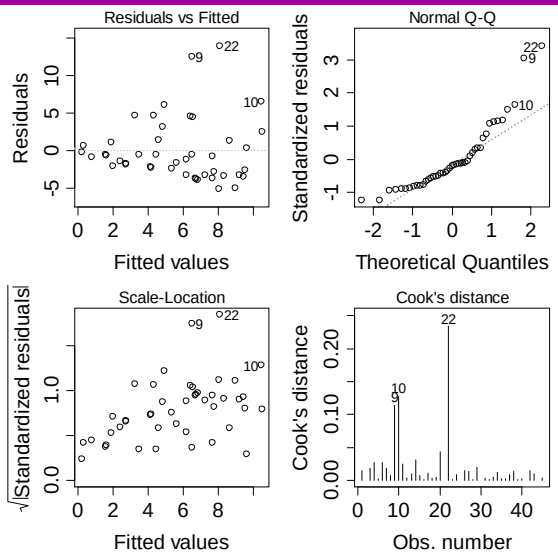
Model Validation

- Anscombe (1973) – all four data sets have the same intercept, slope, significance and R^2 (0.67).
- R^2 alone cannot be used for model choice.



Testing the Assumptions of Regression

- Richness-NAP regression:
  - Scatterplot of residuals on fitted values and scale-location plot show increasing variance.
  - Q-Q plot shows non-normality
  - Cook's distance indicates 3 influential observations.

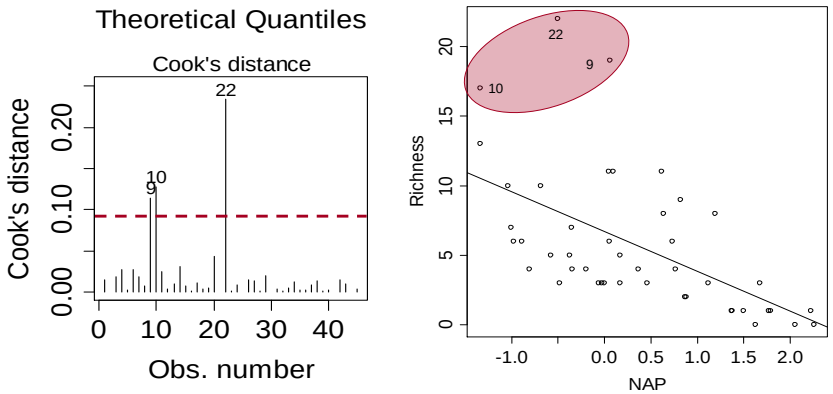



```
RIKZ_model.1<-lm(Richness ~ NAP, data = RIKZ)
plot(RIKZ_model.1, which = c(1:4), add.smooth = F,
cex.id = 1)
```


13

Influential Points

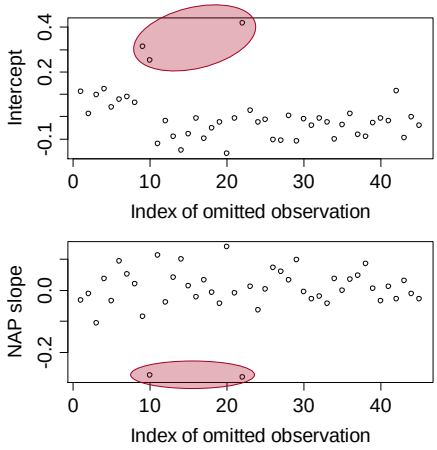
- Cook's *D* measures change in all regression statistics when an observation is removed.
- Expected $D<4/(n-k-1)$; for richness-NAP regression, $D<4/(45-1-1)=0.09$

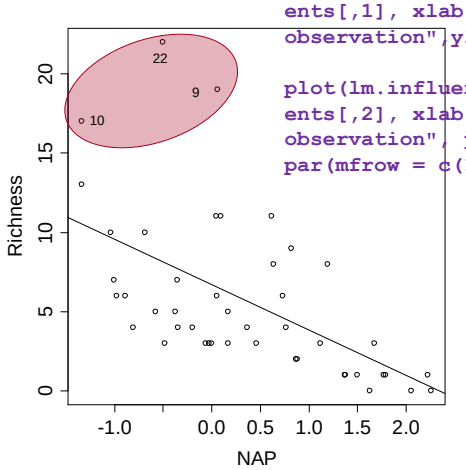


14

Influential Points

- Jackknife method removes each observation and measures variability in regression statistics.





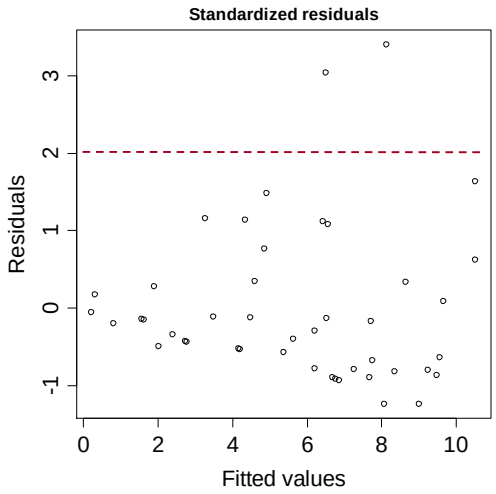

```
plot(lm.influence(RIKZ_model.1)$coefficients[,1], xlab = "Index of omitted observation", ylab = "Intercept")

plot(lm.influence(RIKZ_model.1)$coefficients[,2], xlab = "Index of omitted observation", ylab = "NAP slope")
par(mfrow = c(1,1))
```


Influential Points

- Standardized residuals (z) greater than <-2 or >+2 indicate outliers.

$$z_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\left(\sum_n \left(Y_i - \hat{Y}_i\right)^2 \Big/ (n-2)\right)}}$$



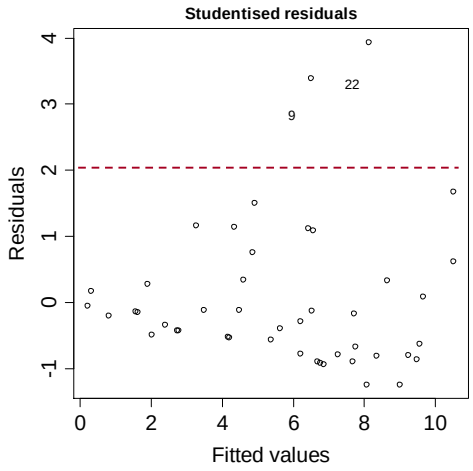

```
plot(predict(RIKZ_model.1),stdres(RIKZ_model.1),
     xlab = "Fitted values", ylab = "Residuals",
     main = "Standardized residuals")
```

Influential Points

- Studentized residuals are *t*-distributed for detection of outliers, accounting for sample sizes.
- For richness-NAP regression, $t_{0.05}$=2.02.

$$z_i' = \frac{Y_i - \hat{Y}_i}{\sqrt{\left(\sum_n \left(Y_i - \hat{Y}_i\right)^2 \Big/ (n-2)\right)\sqrt{1-h_i}}}$$

A studentized deleted residual removes each Y_i from the denominator.

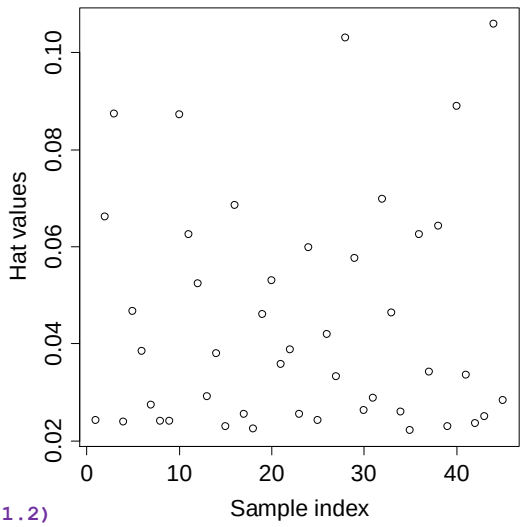



```
plot(predict(RIKZ_model.1),studres(RIKZ_model.1),
     xlab = "Fitted values", ylab = "Residuals",
     main = "Studentised residuals")
```


Influential Points

- Leverage (*h*, 'hat value'):

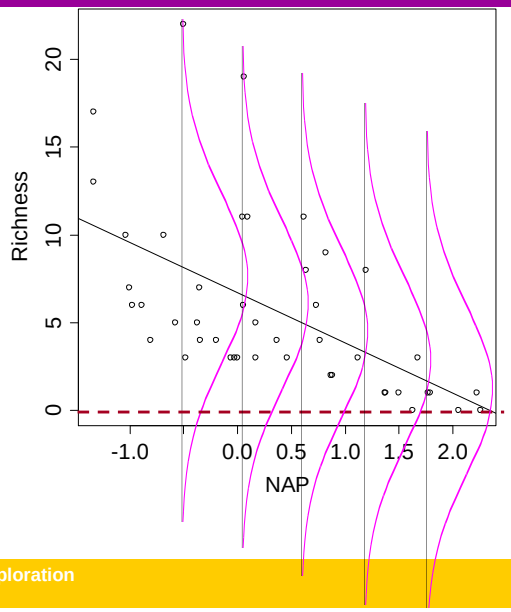
$$h_i = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum_n \left(X_i - \bar{X}\right)^2}$$




```
plot(lm.influence(RIKZ_model.1)$hat,
     xlab = "Sample index", ylab = "Hat values", cex = 1.2)
```

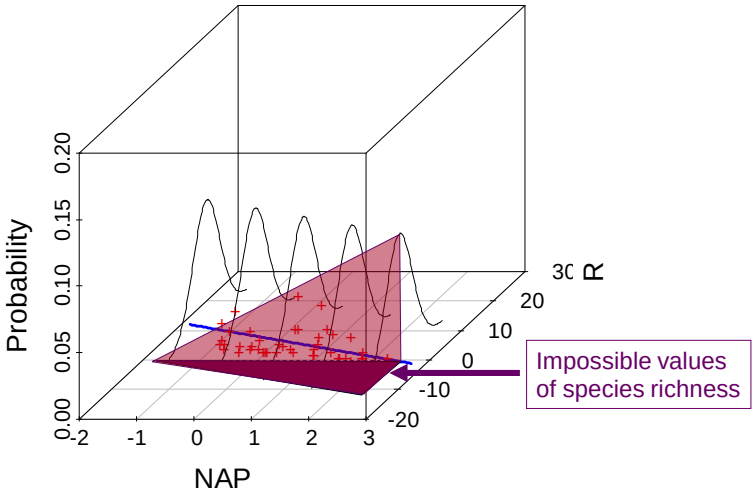
Implications of Normality Assumption

- Species richness is expected to be zero at NAP ~2.4.
- Large probability of negative richness (?) at even intermediate NAP levels!?!?!?
- Unequal variances (use GLM with a more appropriate distribution assumption)



Implications of Normality Assumption

- Scatterplot3d
- Species richness should not be expected to have a normal distribution, and alternative distributions should be explored.



Testing the Assumptions of Regression

- Residual analyses can detect heterogeneity, non-normality and lack of independence.
- Mild non-normality and error in *X* are not fatal flaws.
- Homogeneous variance is more important.
- If residuals are obviously non-random:
  - Transform
  - Add other explanatory variables
  - Add interactions
  - Add quadratic terms
  - Use smoothers (additive models)
  - Use Generalized Least Squares for unequal spread
  - Use Generalized Linear Models with alternative variance structures
  - Use Generalized Additive Models for nonlinear relationships
  - Apply mixed modeling

5.2. Multiple Linear Regression

- NAP was only one of several explanatory variables (grain size, humus, beach angle, exposure, week) for species richness.
- Multiple regression formula of response *Y*, with multiple explanatory variables X_1 to X_p:

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

$$R_i = \alpha + \beta_1 NAP_i + \beta_2 Grain_i + \beta_3 Humus_i + Week_i + \beta_4 Angle_i + noise_i$$

- Each partial regression coefficient (β) measures the effect of one unit of the explanatory variable on species richness.
- Week is treated as a nominal value.

ANOVA Table

- Null hypothesis: all partial regression coefficients (β) equal 0.

Source	*SS*	*df*	*MS*
Re*gression*	$\sum_n \left(\hat{Y} - \bar{Y}\right)^2$	p	$\frac{\sum_n \left(\hat{Y} - \bar{Y}\right)^2}{p}$
Re*sidual*	$\sum_n \left(Y_i - \hat{Y}\right)^2$	$n - p - 1$	$\frac{\sum_n \left(Y_i - \hat{Y}\right)^2}{n - p - 1}$
Total	$\sum_n \left(Y_i - \bar{Y}\right)^2$	$n - 1$	$\frac{\sum_n \left(Y_i - \bar{Y}\right)^2}{n - 1}$

```
RIKZ_model.2<-lm(Richness ~ angle2+NAP+grainsize+humus+factor(week), data = RIKZ)
summary(RIKZ_model.2)$coefficients
summary(RIKZ_model.2)
```


ANOVA Table

- If regression is significant, *t*-ratios can be used to test significance of each explanatory variable:

```
Call:
lm(formula = Richness ~ angle2 + NAP + grainsize + humus +
factor(week), data = RIKZ)

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.298448   7.967002   1.167 0.250629
angle2          0.016760   0.042934   0.390 0.698496
NAP            -2.274093   0.529411  -4.296 0.000121 ***
grainsize       0.002249   0.021066   0.107 0.915570
humus           0.519686   8.703910   0.060 0.952710
factor(week)2  -7.065098   1.761492  -4.011 0.000282 ***
factor(week)3  -5.719055   1.827616  -3.129 0.003411 **
factor(week)4  -1.481816   2.720089  -0.545 0.589182

Residual standard error: 3.092 on 37 degrees of freedom
Multiple R-Squared: 0.679,      Adjusted R-squared: 0.6182
F-statistic: 11.18 on 7 and 37 DF,  p-value: 1.664e-07
```

Comparing Nested Models

- If the explanatory variables in one model (the 'nested model,' model 1) are a subset of those in another (the 'full model,' model 2),

$$Model1: \quad Y_i = \alpha + \varepsilon_i$$
$$Model2: \quad Y_i = \alpha + \beta Angle_i + \varepsilon_i$$

- Models can be compared with an *F* statistic based on residual sums-of-squares (*RSS*) and parameters in model 1 (*p*+1) and model 2 (*q*+1):

$$F = \frac{(RSS_1 - RSS_2)/(p-q)}{RSS_2/(n-p)}$$

Full ANOVA Table

- Each *F* statistic tests the sequential significance of the additional explanatory variable:

```
RIKZ_model.2<-lm(Richness ~ angle2+NAP+grainsize+humus+factor(week),
data = RIKZ)
anova(RIKZ_model.2)

Analysis of Variance Table

Response: Richness
             Df Sum Sq Mean Sq F value    Pr(>F)
angle2        1 124.86  124.86 13.0631 0.0008911 ***
NAP           1 319.32  319.32 33.4071 1.247e-06 ***
grainsize     1 106.76  106.76 11.1692 0.0019116 **
humus         1  19.53   19.53  2.0433 0.1612721
factor(week)  3 177.51   59.17  6.1902 0.0016200 **
Residuals    37 353.66    9.56
```

Model Selection

- Finding the best set of explanatory variables is subjective unless a criterion is specified.
- All criteria are functions of goodness-of-fit (e.g., residual sum-of-squares), number of parameters (e.g., explanatory variables, *p*) and sample size (*n*)
- Adjusted R^2 (portion of explained variance; larger values are better):

$$R^2_{adjusted} = 1 - \frac{SS_{residual}/[n-(p+1)]}{SS_{total}/[n-1]}$$

- Akaike Information Criterion (model dispersion; smaller values are better):

$$AIC = 2p + nLn(SS_{residual}) \ldots + c$$

Model Selection

- Starting with the 'full model,' AIC can judge the removal of an explanatory variable:

```
Start:  AIC=108.78
Richness ~ angle2 + NAP + grainsize + humus + factor(week)

                Df Sum of Sq    RSS    AIC
- humus          1      0.03 353.70 106.78
- grainsize      1      0.11 353.77 106.79
- angle2         1      1.46 355.12 106.96
<none>                      353.66 108.78
- factor(week)   3    177.51 531.17 121.08
- NAP            1    176.37 530.03 124.98
```

- Model is better without humus (AIC decreases from 108.78 with all variables to 106.78 without humus).

Model Selection

- Regression without humus

```
Step:  AIC=106.78
Richness ~ angle2 + NAP + grainsize + factor(week)

               Df Sum of Sq    RSS    AIC
- grainsize     1      0.12 353.82 104.80
- angle2        1      1.55 355.24 104.98
<none>                     353.70 106.78
- factor(week)  3    197.00 550.70 120.70
- NAP           1    180.31 534.01 123.32
```

- Grain size should be removed.

Model Selection

- Regression without humus or grainsize

```
Step:  AIC=104.8
Richness ~ angle2 + NAP + factor(week)

               Df Sum of Sq    RSS    AIC
- angle2        1      3.19 357.00 103.20
<none>                     353.82 104.80
- NAP           1    213.45 567.26 124.04
- factor(week)  3    303.64 657.46 126.68
```

- Remove angle.

Model Selection

- Regression without humus, grainsize or angle

```
Step:  AIC=103.2
Richness ~ NAP + factor(week)

               Df Sum of Sq    RSS    AIC
<none>                      357.00 103.20
- NAP           1    210.33 567.33 122.04
- factor(week)  3    387.11 744.12 130.25
```

- Retain NAP and week (AIC increases when they are removed)

Model Selection

- Backward elimination results.

```
step(RIKZ_model.2, direction = "backward")
#     One variable is dropped in turn
drop1(RIKZ_model.2, test = "F")

Single term deletions

Model:
Richness ~ angle2 + NAP + grainsize + humus + factor(week)
             Df Sum of Sq    RSS    AIC F value     Pr(F)
<none>                    353.66 108.78
angle2        1      1.46 355.12 106.96  0.1524 0.6984955
NAP           1    176.37 530.03 124.98 18.4514 0.0001211 ***
grainsize     1      0.11 353.77 106.79  0.0114 0.9155704
humus         1      0.03 353.70 106.78  0.0036 0.9527102
factor(week)  3    177.51 531.17 121.08  6.1902 0.0016200 **
```

Model Selection

- Just like R^2, AIC should only be used as a general guide (e.g., can 'accept' non-significant effects) for valid models.
- *F*-tests, model fit, and residual patterns should also be considered.
- Selection criteria ignore multiple comparisons (i.e., type-I error) and collinearity (assume independent explanatory variables)
- Generalized models with alternative distribution assumptions will have similar measures of model fit (e.g., deviance explained) model selection criteria (e.g., DIC: Deviance Information Criteria)

5.3. Partial linear regression

- Three reasons to introduce partial regression:
  1. Determine if an explanatory variable should be included in a model.
  2. The concept is used in some multivariate methods
  3. Variance partitioning is easier to explain in a bivariate relationship
- Example – Argentinean benthic data (Chapter 28 – need to install 'vegan' package in R)

```
Argentina<-read.table(file = "Argentina.txt",header = TRUE)
library(vegan)
```

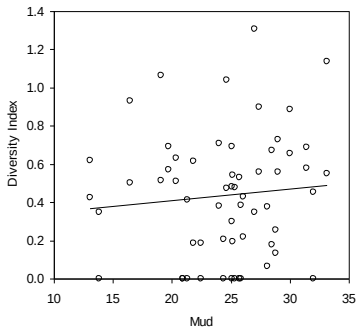
Partial linear regression

- Mud and transect may be important in explaining biodiversity (*H*), but they're effects look insignificant:

ANOVA

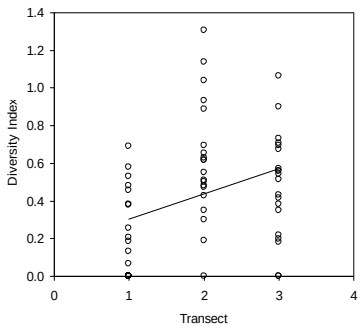
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.92819	0.46409	5.12146	0.00901
Residual	57	5.16520	0.09062		
Total	59	6.09339			

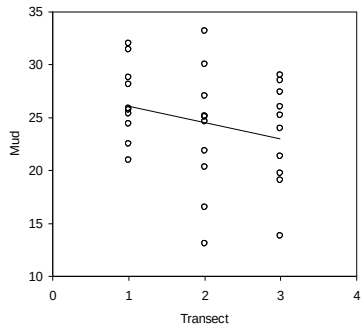
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.18171	0.25052	-0.72534	0.47121
Mud	0.01277	0.00827	1.54502	0.12788
Transect	0.15322	0.04929	3.10831	0.00293



Partial Regression

```
par(mar = c(4.5,4.5,0.5,0.5), cex.lab = 1.5, cex.axis = 1.5)
H<-diversity(Argentina[,2:5], index = "shannon")
H_model<-lm(H ~ factor(Transect), data = Argentina)
Mud_model<-lm(Mud ~ factor(Transect), data = Argentina)
partial_lm<-lm(residuals(H_model) ~ residuals(Mud_model))
summary(partial_lm)$coefficients
plot(residuals(Mud_model), residuals(H_model), xlab = "Mud",
ylab = "Diversity Index")
abline(partial_lm)
```





- How big of a contribution does mud make in explaining biodiversity?
- It may only appear important because it's collinear with transect.
- Partial regression tests the relationship between diversity and mud while filtering the effects of transect.

Partial Linear Regression

- Two approaches:
  - Filter out effects of each variable on the response variable (Quinn & Keough 2002)
  - Decomposition of variation (Legendre & Legendre 1998).

Partial Regression – Method 1

1. Filter out effects of explanatory variables on response variable.
   - Given one response variable *Y* and 3 explanatory variables *X*, *W* and *Z*:
   - ε_i is the variation in *Y* that cannot be explained by *X*, *W* and *Z*.

$$Y_i = const. + \beta_1 X_i + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i$$

   - ε_{1i} is the variation in *Y* that can't be explained by *W* and *Z*.

$$Y_i = const. + \beta_2 W_i + \beta_3 Z_i + \varepsilon_{1i}$$

Partial Regression – Method 1

2. Regress the 'removed' explanatory variable on the others:

$$X_i = const. + \beta_6 W_i + \beta_7 Z_i + \varepsilon_{2i}$$

- ε_{2i} is the variation in *X* that can't be explained by *W* and *Z*.

Partial Regression – Method 1

1. ε_{1i} is the variation in *Y* after filtering out the effects of *W* and *Z*.
2. ε_{2i} is the variation in *X* after filtering out the effects of *W* and *Z*.
3. Regress ε_{1i} on ε_{2i}:

$$\varepsilon_{1i} = \beta\varepsilon_{2i} + noise$$

- This 3rd model shows the relationship of *Y* and *X* after partitioning out the effects of *W* and *Z*.

Partial Regression – Method 1

1. For the benthic diversity example, regress diversity on transect:

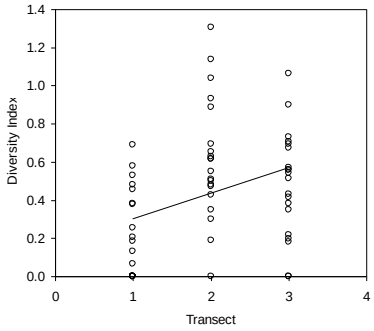
$$H_i = const. + factor(transect) + \varepsilon_{1i}$$

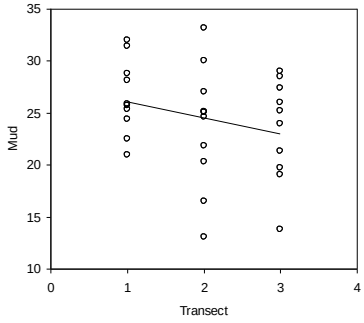
2. Regress mud on transect

$$Mud_i = const. + factor(transect) + \varepsilon_{2i}$$

3. Regress residuals of regression 1 on the residuals of regression 2.

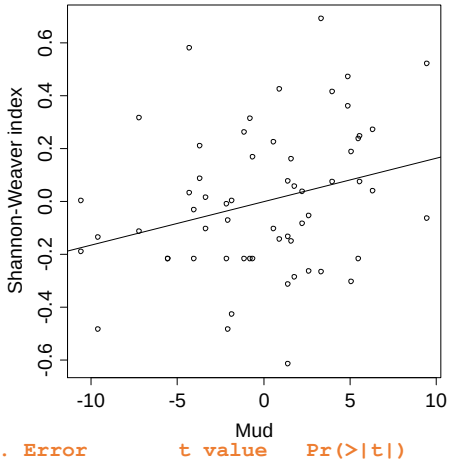
$$\varepsilon_{1i} = \beta\varepsilon_{2i} + noise$$





Partial Regression – Method 1

- Slope of the regression of ε_{1i} on ε_{2i} indicates that mud explains some variation in diversity, even after the effect of transect is filtered out.




```
                          Estimate  Std. Error       t value   Pr(>|t|)
(Intercept)           -7.050783e-18 0.034323828 -2.054195e-16 1.00000000
residuals(Mud_model)   1.643364e-02 0.007362991  2.231924e+00 0.02949761
```

Partial Linear Regression

- Using a more general multivariate notation, matrices **X** and **W** contain *p* variables, respectively.
- Parameters are β and *v*.

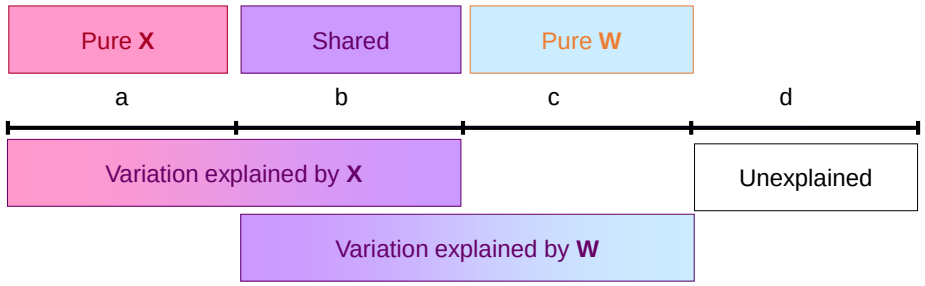
$$y_i = \mathbf{X_i}\beta + \mathbf{W_i}v + \varepsilon_i$$

- Regress *y* and each variable x_j against **W**.

$$y_i = \mathbf{W_i}v_0 + \varepsilon_{0i}$$
$$x_{1i} = \mathbf{W_i}v_1 + \varepsilon_{1i}$$
$$\ldots$$
$$x_{pi} = \mathbf{W_i}v_p + \varepsilon_{pi}$$

Partial Regression – Method 2

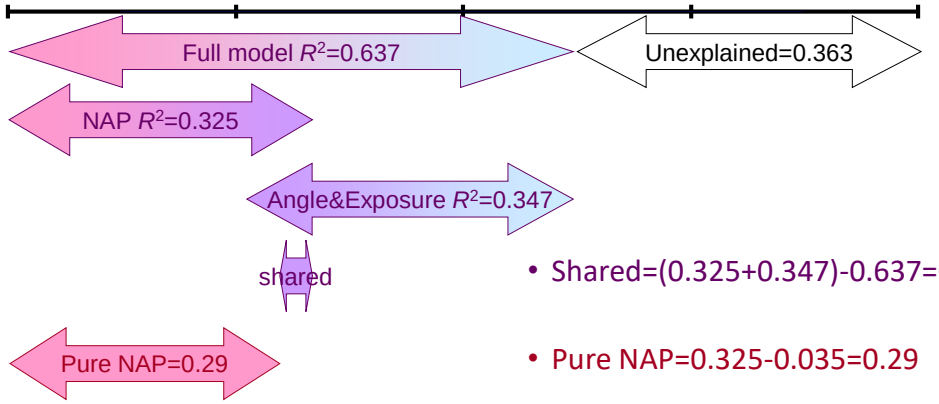
- Compare the R^2 from the full model with both explanatory variables (a+b+c) to R^2 from regressions of *Y* on *X* (a+b) and *Y* on *W* (b+c) to partition variance into variance explained purely by X and purely by W:



Partial Regression – Method 2

$$R_i = \alpha + \beta_1 NAP_i + \beta_2 Angle_i + Exposure_i + noise_i$$

- Dutch beach community



- Shared=(0.325+0.347)-0.637=0.035
- Pure NAP=0.325-0.035=0.29

Summary

- Model validation:
  - goodness of fit (e.g., R^2)
  - evaluation of assumptions (equal variance, normality, linearity)
  - residual analyses
- Model selection:
  - model fit (e.g., diagnostics and R^2)
  - parsimony (e.g., adjusted R^2, AIC)
- Effect of variables may be masked by effects of other variables:
  - Other effects can be 'filtered.'
  - Variance can be decomposed into 'pure' effects.

Schedule

Type	Day	Date	Reading	Topic
Lecture	Tue.	Jan 21	Zuur et al. Chap. 1-4	Introduction, statistical rethinking
Lab	Wed.	Jan 22		R Lab 1
Lecture	Thu.	Jan 23	Bolker 2008 Chap. 4	Probability review
Lecture	Tue.	Jan 28	Zuur et al. Chapter 5	Linear regression review
Lab	Wed.	Jan 29		R Lab 2
Lecture	Thu.	Jan 30		data exploration, checking
Lecture	Tue.	Feb 04	Hilborn & Mangel Chapter 7	Likelihood
Lab	Wed.	Feb 05		R Lab 3
Lecture	Thu.	Feb 06	Zuur et al. Section 6.1	Extending the linear model (GLM) (Poisson)

- Questions on regression?

47