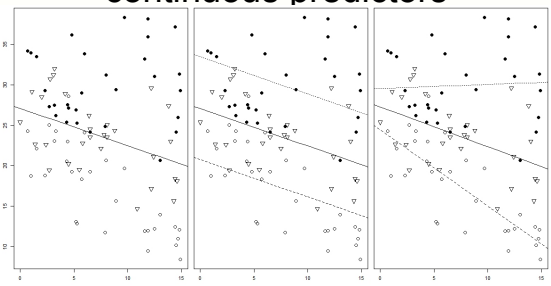
Advancing in R
ANCOVA: factorial and continuous predictors



Outline

- Brief summary of (multiple) regression
- Factorial predictors in linear models
- Interpreting coefficients
- Interactions between factors and covariates
- Which model? Likelihood and AIC

Linear models up until now

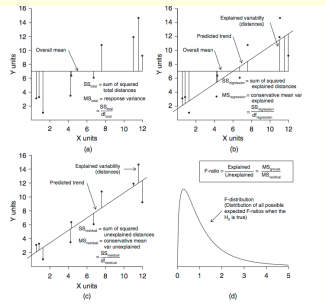


Fig 8.3 Fictitious data illustrating the partitioning of (a) total variation into components (b) explained ($MS_{regression}$) and (c) unexplained ($MS_{residual}$) by the linear trend. The probability of collecting our sample, and thus generating the sample ratio of explained to unexplained variation (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.

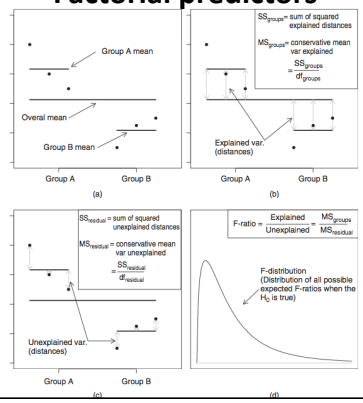
Regression

$y_i = a + bX_i + \varepsilon_i$

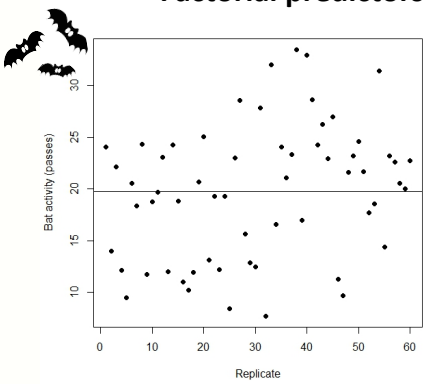
e.g.
Temperature ~ Altitude
Pollen ~ Buzz duration
Pups ~ Permits

Factorial predictors

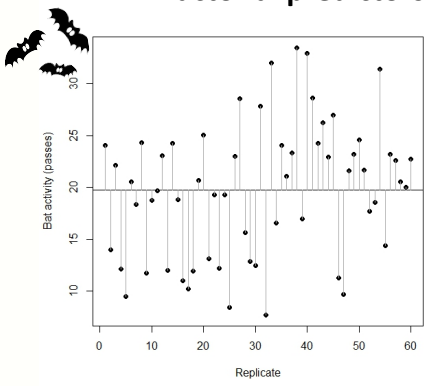
Factorial predictors



Factorial predictors



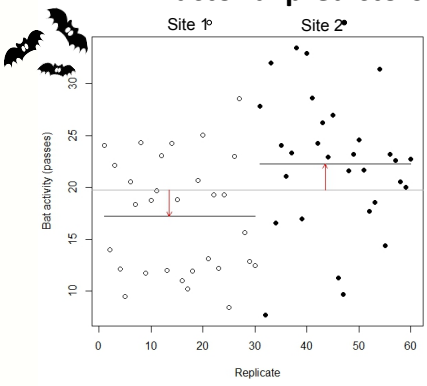
Factorial predictors



SS_{total} = sum of squared total distances

MS_{total} = response variance = SS_{total} / df_{total}

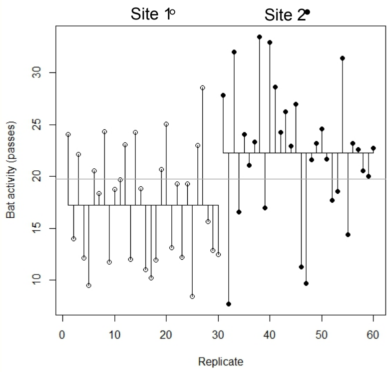
Factorial predictors



SS_{site} = sum of squared explained (by group) distances

MS_{site} = mean variance explained = SS_{site} / df_{site}

Factorial predictors



$SS_{residual}$ = sum of squared unexplained distances

$MS_{residual}$ = mean unexplained variance = $SS_{residual}$ / $df_{residual}$

Factorial predictors: Analysis of Variance

Factors consist of a number of levels:

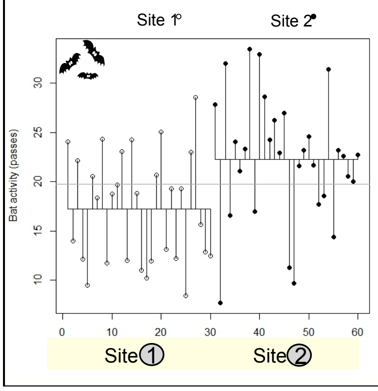
```
[1]  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 2 2 [56] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2

[1] "1" "2"
```

Analysis of Variance (ANOVA):

```
> mod1 <- aov(PASSES ~ factor(SITE), data = BATS))
> summary(mod1)
```

	Df	Sum Sq	Mean Sq	F	Pr(>F)
f(SITE)	1	379.2	379.2	10.44	0.00204
Residuals	58	2107.5	36.3		

Factorial predictors in a linear model



$y_i = a + bX_i + \varepsilon_i$

"Site effect"

What are *a* and *b* ?

A) Means of site 1 and 2.
B) The mean of site 1, and the difference between mean 1 and 2
C) The overall mean and the slope of the site effect
D) None of the above

Factorial predictors in a linear model

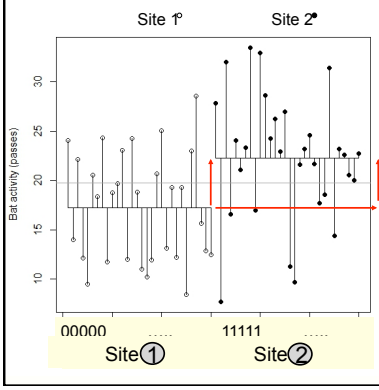
$y_i = a + bX_i + \varepsilon_i$

```
> mod1 <- lm(PASSES ~ factor(SITE), data=BATS)
> summary(mod1)
```

Coefficients:

	Estimate	Std. Err.	t-value	Pr(>|t|)
(Intercept)	17.230	1.101	15.656	<2.00E-16
f(SITE)2	5.028	1.556	3.231	0.00204

Factorial predictors in a linear model



$y_i = a + bX_i + \varepsilon_i$

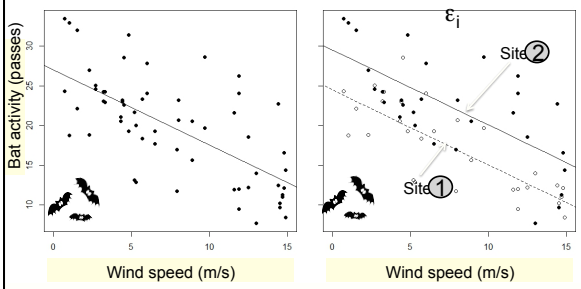
factor(SITE)2= b = 5.0

(Intercept) = a = 17.2

Analysis of Covariance (ANCOVA)

Passes = Wind speed … = Wind speed + Site

$y_i = a + bX_i + \varepsilon_i$ $y_i = a + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$



Analysis of Covariance (ANCOVA)

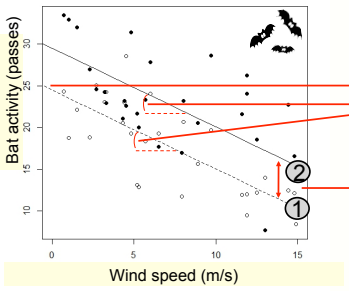
Passes = Wind speed + Site = $a + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$

```
> mod <- lm(PASSES ~ WIND + factor(SITE), BATS)
> summary(mod)

Coefficients:
                Estimate  Std. Err.  t-value Pr(>|t|)
(Intercept)      24.4808     1.1625  21.060  <2.00E-16
WIND             -0.9453     0.1155  -8.185   3.39e-11
factor(SITE)2     5.0280     1.0645  -8.185   1.56e-05
```

Analysis of Covariance (ANCOVA)

Passes = Wind speed + Site

$y_i = a + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$



(Intercept) = a = 24.5

Wind = b_1 = -0.9

factor(SITE)2 = b_2 = 5.0

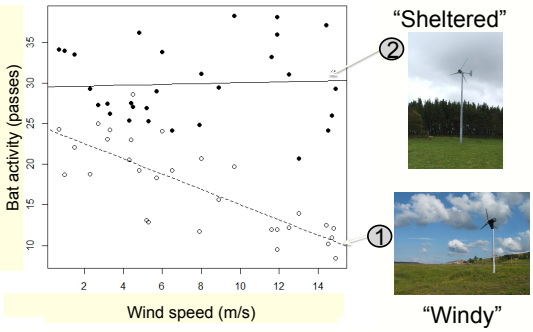
Analysis of Covariance: Interactions

Site ①
"Windy"

Site ②
"Sheltered"

Analysis of Covariance: Interactions

Passes = Wind speed + Site + (Wind speed * Site)



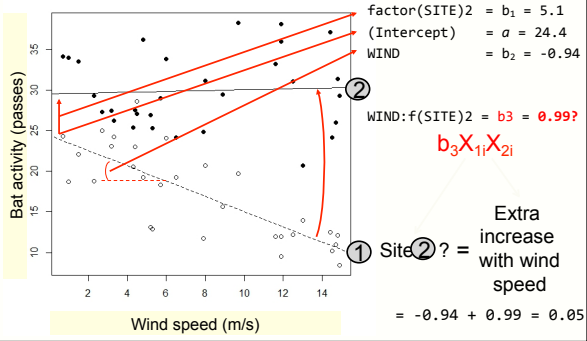
Analysis of Covariance: Interactions

$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + \varepsilon_i$

```
> mod2 <- lm(PASSES ~ WIND + factor(SITE) + WIND*factor(SITE), BATS2)
> summary(mod2)

Coefficients:
                    Estimate Std. err. t-value  Pr(>|t|)
(Intercept)          24.4434   1.4745  16.578 <2.00E-16
WIND                 -0.9404   0.1648  -5.707  4.53e-07
factor(SITE)2         5.1028   2.0852   2.447    0.0176
WIND:factor(SITE)2    0.9902   0.2330   4.249  8.19e-05
```

Analysis of Covariance: Interactions

$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + \varepsilon_i$



Analysis of Covariance: varying slopes

More than two factor levels?

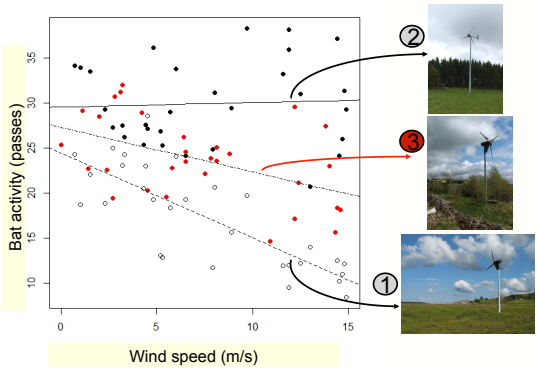
Analysis of Covariance: varying slopes

More than two factor levels?

```
[1] "1" "2" "3"
```

```
Coefficients:
                    Estimate  Std. err. t-value  Pr(>|t|)
(Intercept)          24.4434    1.4696  16.633    < 2e-16
WIND                 -0.9404    0.1642  -5.726   1.55e-07
factor(SITE)2         5.1028    2.0783   2.455     0.0161
factor(SITE)3         2.8307    2.0516   1.380     0.1713
WIND:factor(SITE)2    0.9902    0.2323   4.263   5.24e-05
WIND:factor(SITE)3    0.4486    0.2348   1.911     0.0595
```

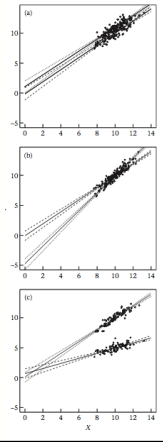
Analysis of Covariance: varying slopes



ANCOVA assumes homogeneity of slopes.

If slopes are not equal, cannot easily interpret effects (unless predictor is mean-centred).

See Engqvist 2005 (in Dropbox folder) for more details.

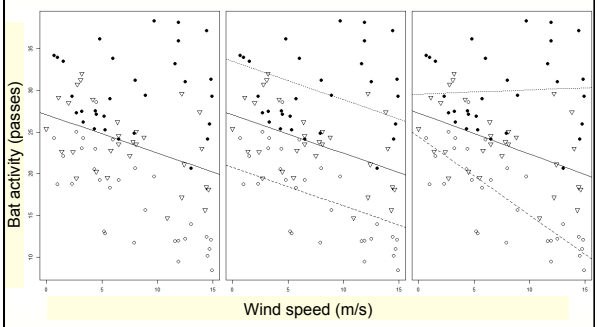


Analysis of Covariance

Summary 1

- Factorial predictors can be combined with covariates in a Linear Model
- Effect of covariate may vary by factor levels (interactions)
- Coefficients interpreted relative to baseline (intercept, first factor level)

Which model? Likelihood and AIC



Which model? Likelihood and AIC

$y_i = a + bX_i + \varepsilon_i$
(*Passes* = *wind*)

$y_i = a + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$
(*Passes* = *wind* + *site*)

$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + \varepsilon_i$
(*Passes* = *wind* + *site* + *wind * site*)

$y_i = a + b_1X_{1i} + b_2X_{2i} + b_4X_{4i} + \varepsilon_i$
(*Passes* = *wind* + *site* + … *temperature?*)

Which model? Likelihood revisited

Likelihood = *Pr* (D | H_x) = *Pr* (Data | Model)

Model 1: Passes = Wind + Site
Model 2: Passes = Wind + Site + Wind * Site

```
> mod1 <- lm(PASSES ~ WIND + factor(SITE), BATS3)
> mod2 <- lm(PASSES ~ WIND + factor(SITE) + WIND*factor(SITE), BATS3)
> logLik(mod1)
'log Lik.' -261.4188 (df=5)
> logLik(mod2)
'log Lik.' -252.5812 (df=7)
> anova(mod2, mod1)
```

```
Model 1: PASSES ~ WIND + factor(SITE) + WIND * factor(SITE)
Model 2: PASSES ~ WIND + factor(SITE)
  Res. Df RSS     Df  Sum of Sq. F      Pr(>F)
1      84  1443.5
2      86  1756.8  -2    -313.25 9.1141        0.0002617
```

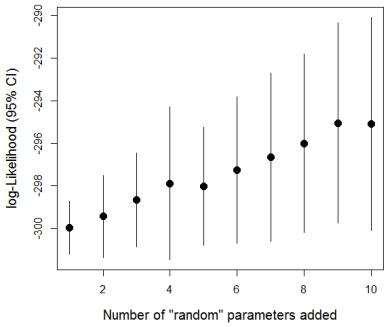
Which model? Likelihood revisited

Random “noise” terms usually increase likelihood!

```
> BATS3$RANDOM1 <- rnorm(nrow(BATS3), mean=0, sd=1)
> BATS3$RANDOM2 <- runif(nrow(BATS3), 0, 10)
> BATS3$RANDOM3 <- rpois(nrow(BATS3), 10)
> formula(mod2)
PASSES ~ WIND + factor(SITE) + WIND * factor(SITE)
> logLik(mod2)
'log Lik.' -252.5812 (df=7)
> mod3 <- update(mod2, .~. +RANDOM1)
> logLik(mod3)
'log Lik.' -252.5245 (df=8)
> mod4 <- update(mod3, .~. +RANDOM2)
> logLik(mod4)
'log Lik.' -251.2074 (df=9)
> mod5 <- update(mod4, .~. +RANDOM3)
> logLik(mod5)
'log Lik.' -249.9625 (df=10)
```

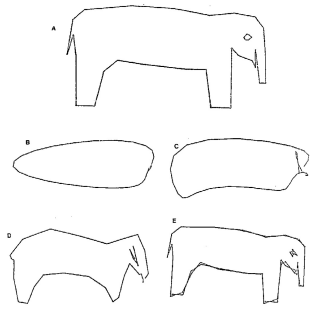
Which model? Likelihood vs. Parsimony

Random “noise” terms usually increase likelihood!



Which model? Likelihood vs. Parsimony

How many parameters to draw an elephant?



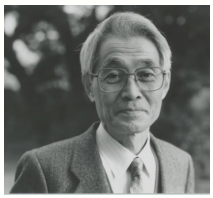
from Burnham & Anderson (1998), Fig. 1.2.

Which model? Likelihood and AIC

Akaike Information Criterion

AIC = -2 Log(Likelihood, model) + 2*K*

= Model Deviance

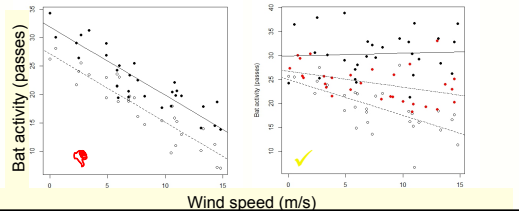


- Balances likelihood (fit) with number of parameters (***K***)
- Lower AIC = relatively better “fit”
- Only meaningful relative to other models

Which model? Comparing by AIC

Five different models (some not "nested"!)

Model	Passes …	Log-Lik	Par's	AIC
1	~ WIND + f (SITE)	-261.4	5	532.8
2	~ WIND + f (SITE) + WIND * f (SITE)	-252.6	7	519.2 ✓
3	~ TEMPERATURE	-304.5	3	615.0
4	~ TEMPERATURE + RAIN + DAY	-303.5	5	617.1
5	~ WIND + TEMPERATURE	-300.4	4	608.7



Which model? Likelihood and AIC

Summary 2

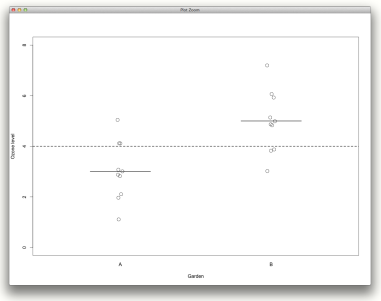
- Likelihood = *Pr* (Data | Model)
- Measures model "fit"
- More parameters is not always better: likelihood can increase with "noise"
- AIC balances likelihood with numbers of parameters used
- AIC can be used to compare across many different models, not just nested ones

Suggested reading:

- Chs. 10, 12, 15 in Logan
- Chs. 7, 9, 10 in Crawley *Statistics*…
- Chs. 11, 12 in Crawley *The R Book*

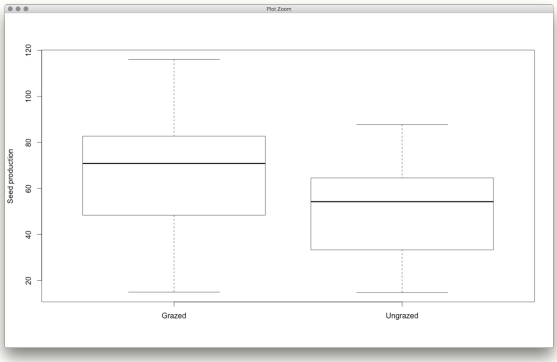
Practical exercise: 2 analyses

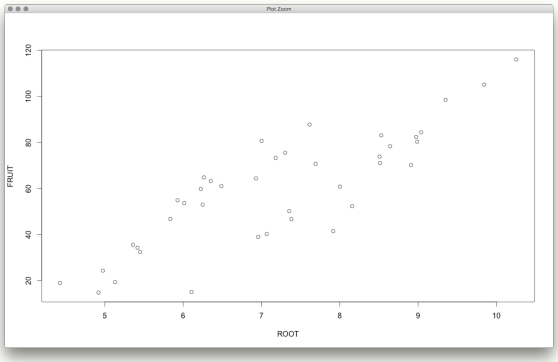
- "oneway" dataset
- Measure levels of ozone pollution in 2 gardens
- N = 10 ozone measures per garden
- Do gardens differ in ozone levels?
- Predictor variable?
- Response variable?
- Plot?

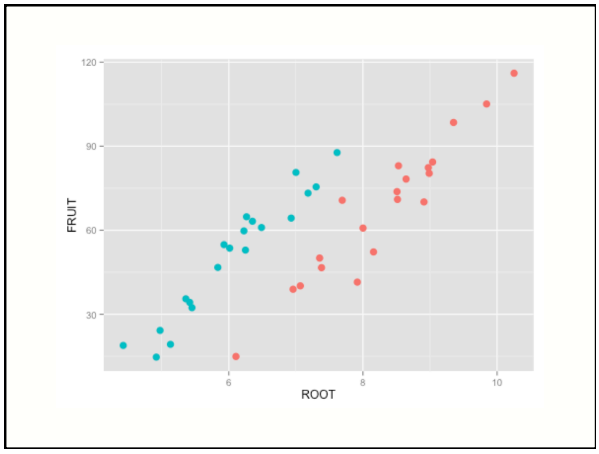


Practical exercise: 2 analyses

- “compensationo” dataset
- Study of seed production in two pastures differing in grazing treatment
- “Nuisance” covariate: root diameter
- N = 20 plots harvested per grazing treatment
- How do grazing and root size affect seed yield?
- Predictor variables?
- Response variable?
- Plot?






120
90
60
30
FRUIT
6
8
10
ROOT