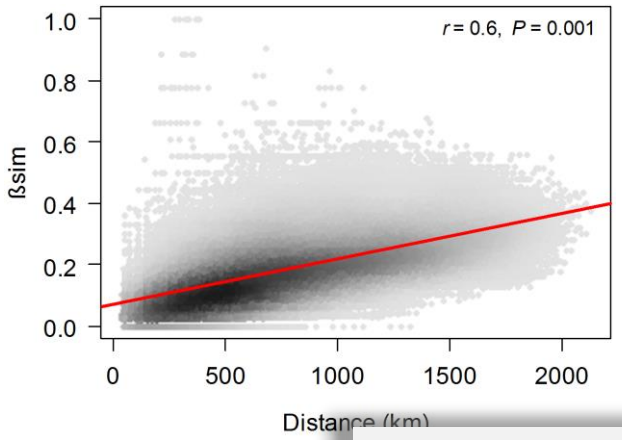
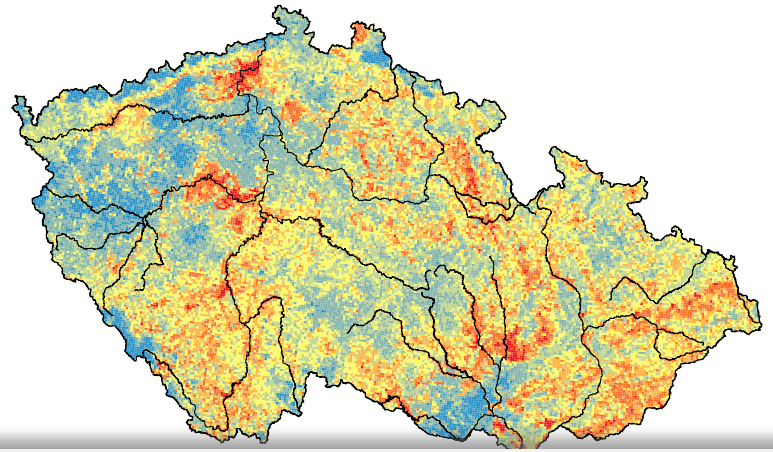
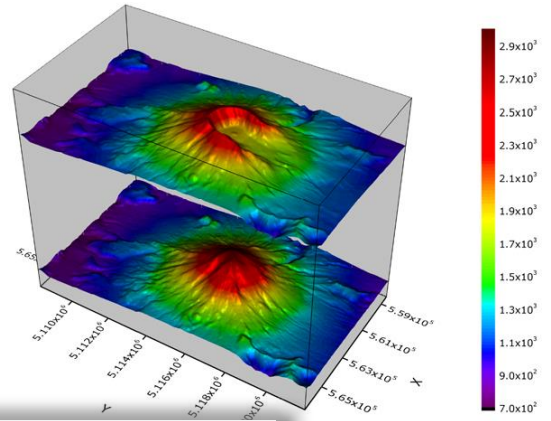
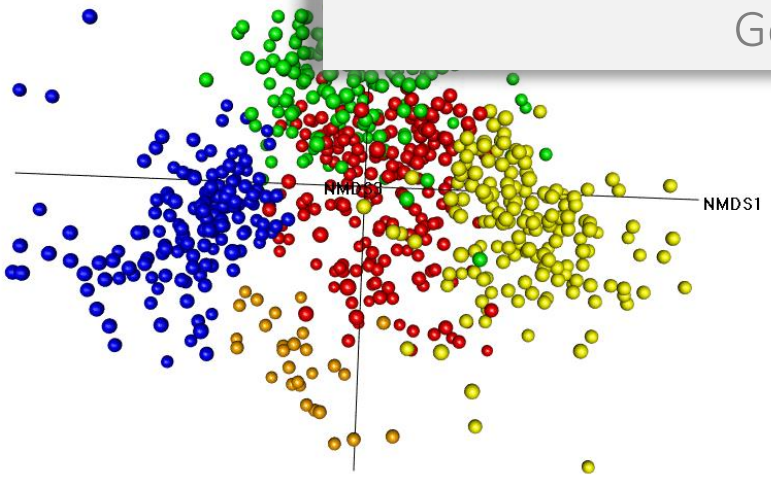


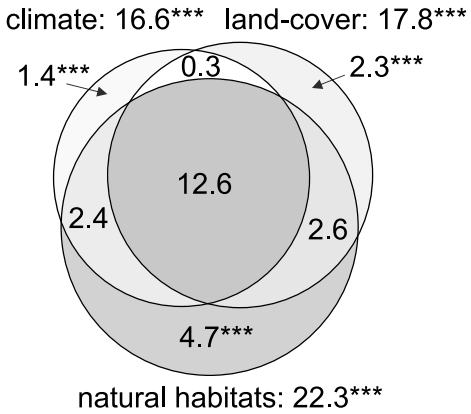




Metody fyzické geografie 3:
Biogeografie & ekologie

Jan Divíšek
Geografický ústav & Ústav botaniky a zoologie



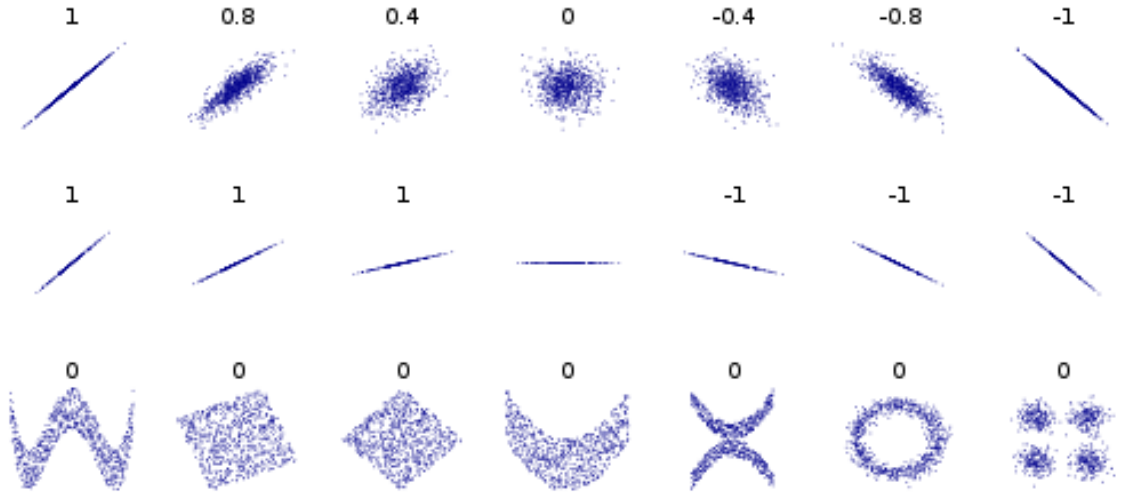


Korelační a regresní analýza

Korelační analýza

```
cor()
cor.test()
```

- Korelace = vzájemný vztah mezi dvěma procesy nebo veličinami
- Ve statistice popisuje vzájemný lineární vztah mezi veličinami *x* a *y*
- Míru korelace vyjadřuje **korelační koeficient**, který může nabývat hodnot od −1 až po +1.
  - Pearsonův korelační koeficient (*r*)
  - Spearmanův koeficient pořadové korelace (ρ nebo r_s)



https://en.wikipedia.org/wiki/Correlation_and_dependence

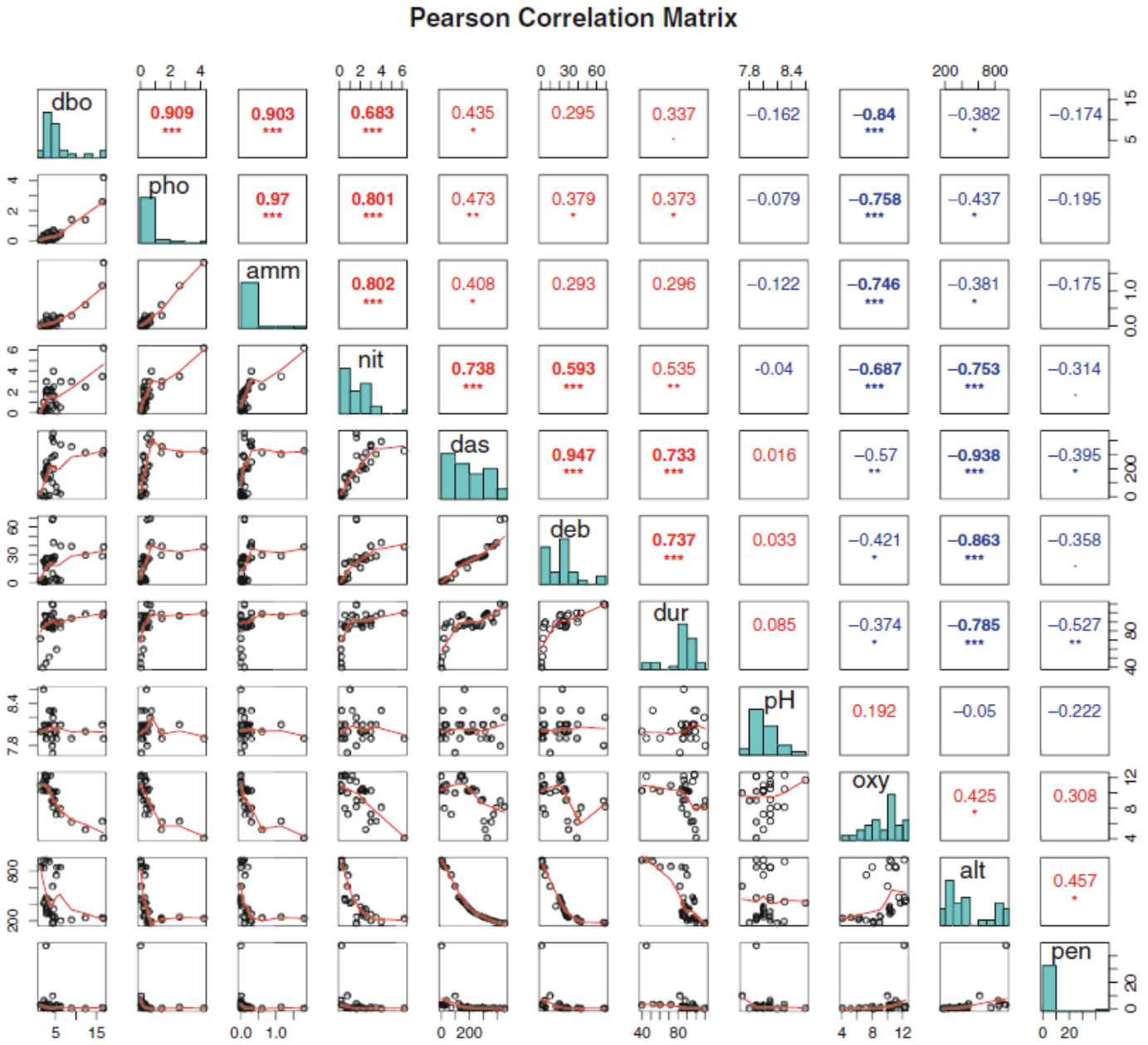


Fig. 3.3 Multipanel display of pairwise relationships between environmental variables with Pearson's *r* correlations

Borcard et al. (2011)

Lineární regrese

`lm()`

- Lineární regresní model popisuje vztah mezi jednou závislou proměnnou a jednou nebo více vysvětlujícími proměnnými

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- y_i: hodnota vysvětlované proměnné pro dané pozorování
- α: absolutní člen
- β: směrnice přímky (určuje sklon)
- x_i: hodnota vysvětlující proměnné pro dané pozorování
- ε_i: náhodná složka (chyba)

$$\varepsilon_i \sim N(0, \sigma^2), \mathrm{cor}(\varepsilon_i, \varepsilon_{i'}) = 0 \text{ pro } i \neq i'$$

- $\varepsilon_i \sim N(0, \sigma^2)$: chyby mají normální rozdělení (N) s nulovou střední hodnotou (0) a rozptylem (σ^2) který je stejný pro všechna pozorování
- $\mathrm{cor}(\varepsilon_i, \varepsilon_{i'}) = 0$: vzájemná korelace chyb je nulová

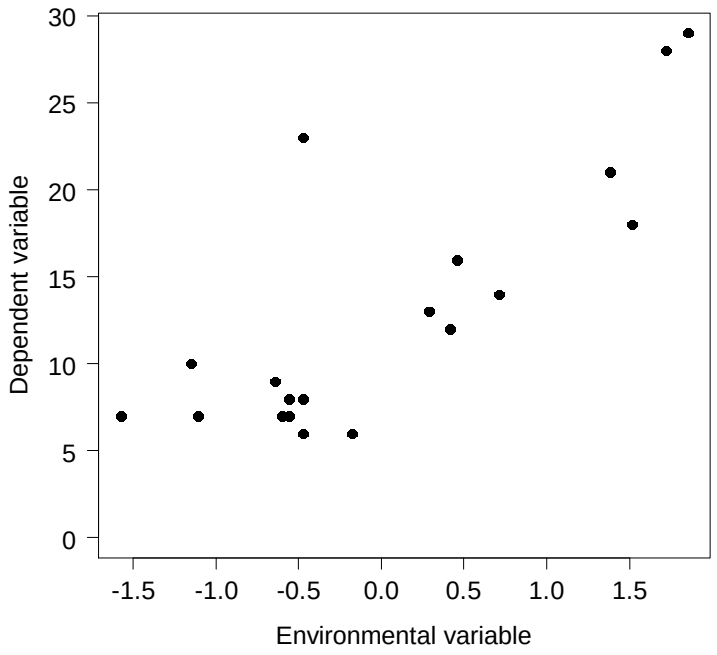
Lineární regrese

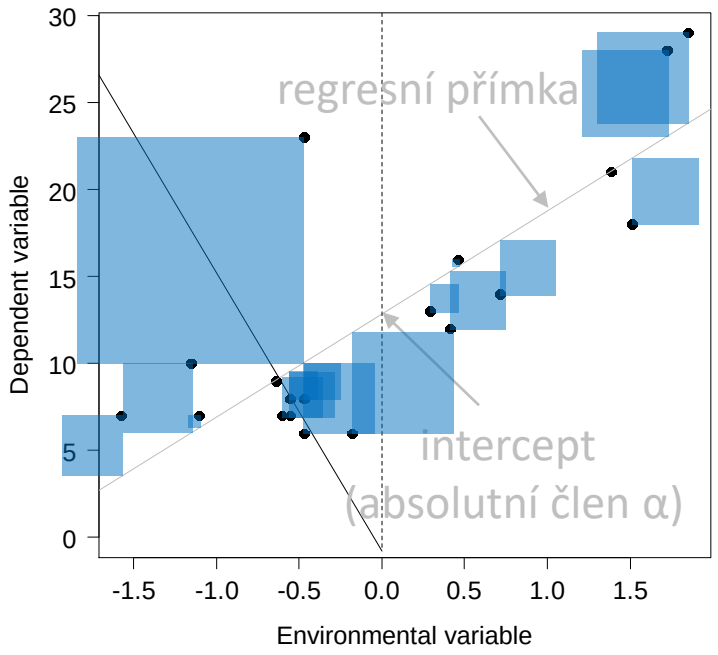
- Lineární regresní model popisuje vztah mezi jednou závislou proměnnou a jednou nebo více vysvětlujícími proměnnými

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

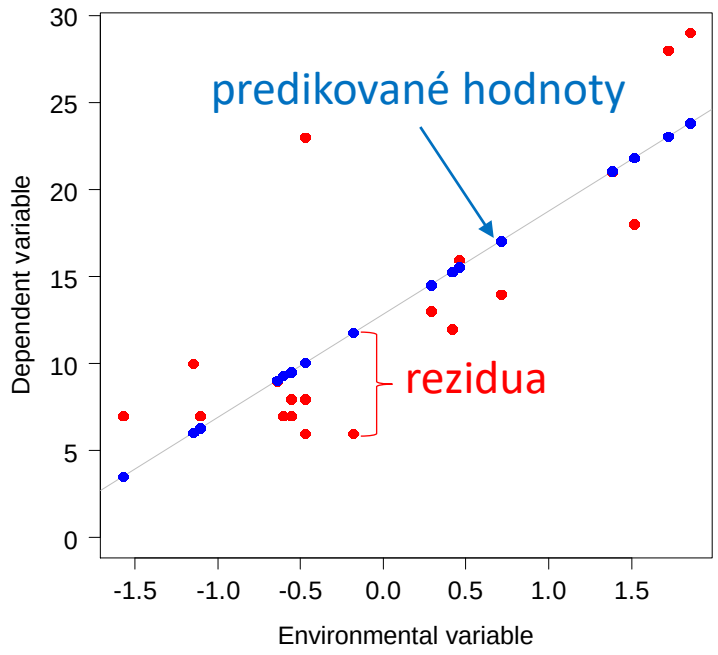
kde $\varepsilon \sim N(0, \sigma^2)$, nezávisle pro různá měření

Lineární regrese





`lm()`

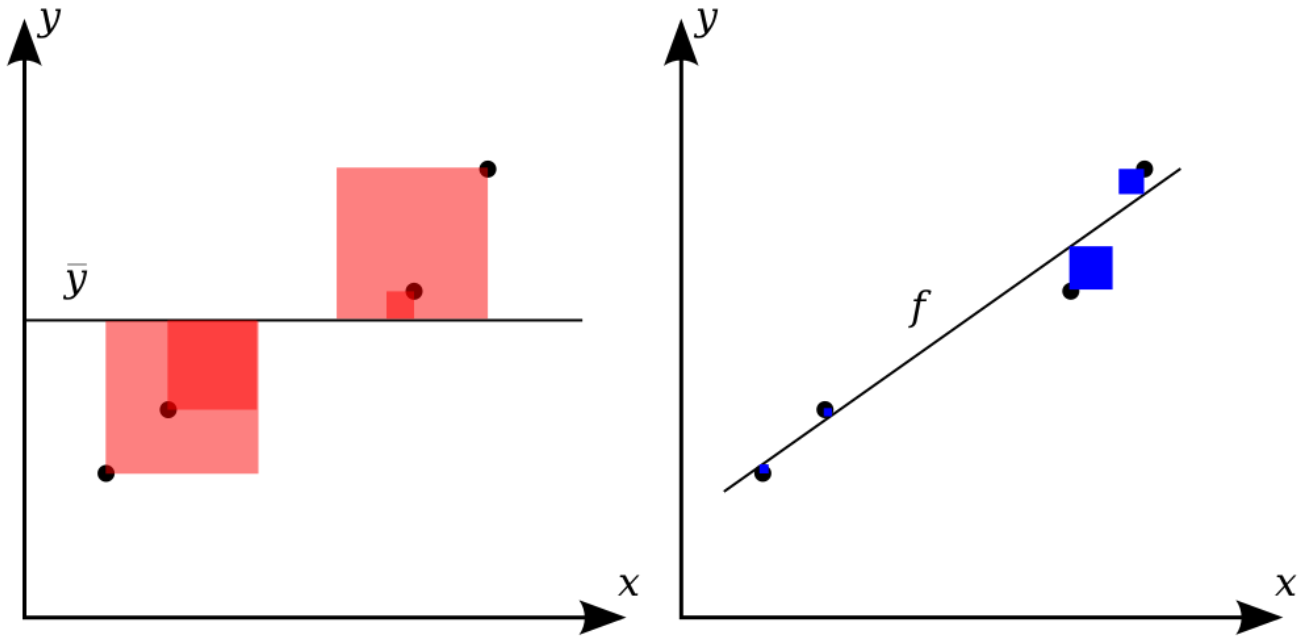


`predict()`
`resid()`

Vegetační snímky z Terénního cvičení z krajinné ekologie 2013 na Kralickém Sněžníku.

Vysvětlená variabilita v regresi

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

http://en.wikipedia.org/wiki/Coefficient_of_determination

Lineární regresní model v R

```
m <- lm(Y ~ X)
anova(m)
```

```
Analysis of Variance Table

Response: Species
          Df Sum Sq Mean Sq F value    Pr(>F)
Altitude   1 503.89  503.89  17.488 0.0005604 ***
Residuals 18 518.66   28.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

průměrné sumy čtverců (Sum Sq/Df)

hodnoty F-testového kritéria (podíl Mean Sq pro danou proměnnou a Mean Sq pro rezidua)

stupně volnosti (Degrees of freedom)

statistická významnost

sumy čtverců (Sum of Squares)

pro vysvětlující proměnné (model sum of squares): $MSS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

pro reziduály (residual sum of squares): $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$TSS = MSS + RSS$

Lineární regresní model v R

```
m <- lm(Y ~ X1 + X2 + … + X5)
anova(m)
```

```
Analysis of Variance Table

Response: Species
           Df Sum Sq Mean Sq F value    Pr(>F)
Altitude    1 503.89  503.89 31.3657 6.541e-05 ***
Slope       1   2.61    2.61  0.1622  0.693243
pH          1 182.77  182.77 11.3768  0.004551 **
Moisture    1  76.63   76.63  4.7702  0.046465 *
E3_cover    1  31.73   31.73  1.9753  0.181690
Residuals 14 224.91   16.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lineární regresní model v R

```
m <- lm(Y ~ X1 + X2 + … + X5)
anova(m)
```

	Slope	pH	Moisture	E3_cover
Altitude	0.296	-0.759	0.268	-0.331
Slope		-0.221	0.085	-0.408
pH			-0.229	0.461
Moisture				0.149

```
Analysis of Variance Table

Response: Species
           Df Sum Sq Mean Sq F value    Pr(>F)
pH          1 667.23  667.23 41.5325 1.536e-05 ***
Slope       1   6.50    6.50  0.4044   0.53511
Altitude    1  15.55   15.55  0.9678   0.34192
Moisture    1  76.63   76.63  4.7702   0.04647 *
E3_cover    1  31.73   31.73  1.9753   0.18169
Residuals  14 224.91   16.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lineární regresní model v R

```
summary(m)
```

```
Call:
lm(formula = Species ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4215 -1.9238  0.6839  2.3229  6.6599

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8500     0.8962  14.338 9.24e-10 ***
Altitude     -2.0206     1.4600  -1.384   0.1880
Slope         0.1462     1.0439   0.140   0.8906
pH            4.0380     1.5358   2.629   0.0198 *
Moisture      1.6358     1.0076   1.623   0.1268
E3_cover      1.6526     1.1758   1.405   0.1817
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.008 on 14 degrees of freedom
Multiple R-squared:   0.78,     Adjusted R-squared:  0.7015
F-statistic:  9.93 on 5 and 14 DF,  p-value: 0.0003219
```

hodnoty t-testu nulové hypotézy (H0) o tom, že skutečná hodnota daného koeficientu je nulová

střední chyba

statistická významnost

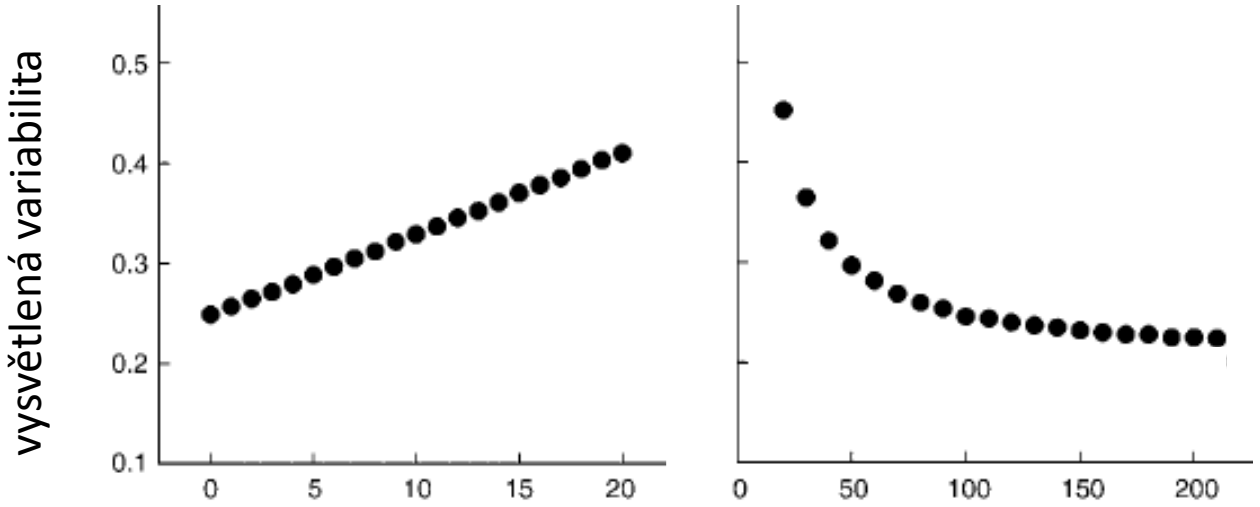
odhady koeficientů

variabilita v `Y` vysvětlená modelem (R^2)

adjustovaný R^2

Vysvětlená variabilita (R^2)

- vysvětlená variabilita stoupá s počtem vysvětlujících proměnných (i když jsou náhodné) a klesá s počtem vzorků v datovém souboru
- platí pro mnohonásobnou regresi i pro přímou (kanonickou) ordinační analýzu



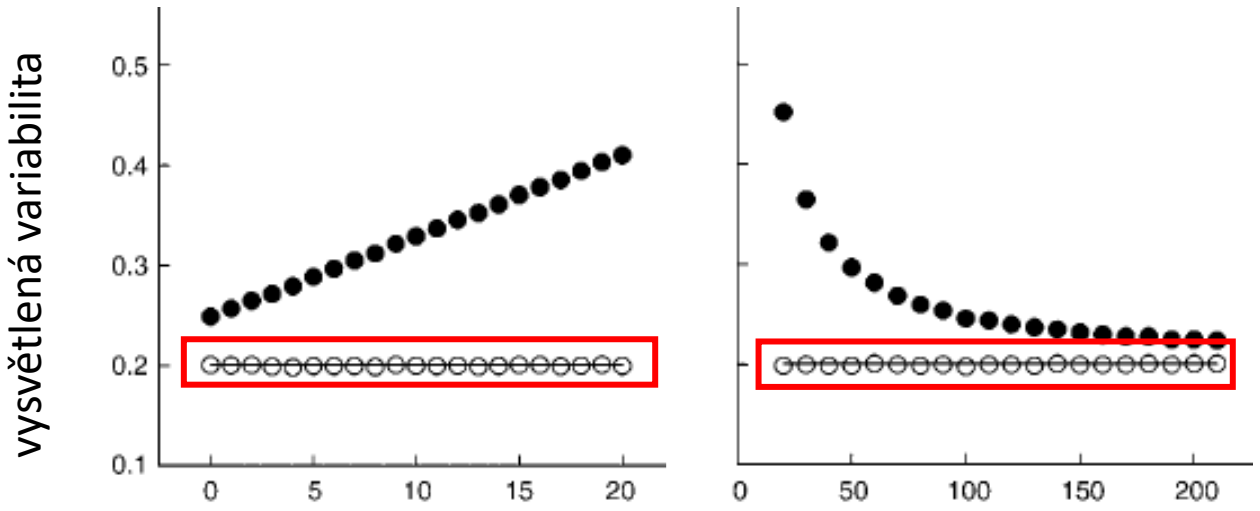
počet vysvětlujících proměnných

počet vzorků v datovém souboru

Peres-Neto et al. (2006) *Ecology*

Vysvětlená variabilita (R^2) a adjustovaný R^2

- adjustovaný R^2 se nemění s počtem vysvětlujících proměnných a počtem vzorků v souboru



počet vysvětlujících proměnných

počet vzorků v datovém souboru

Peres-Neto et al. (2006) *Ecology*

Výpočet adjustovaného R^2

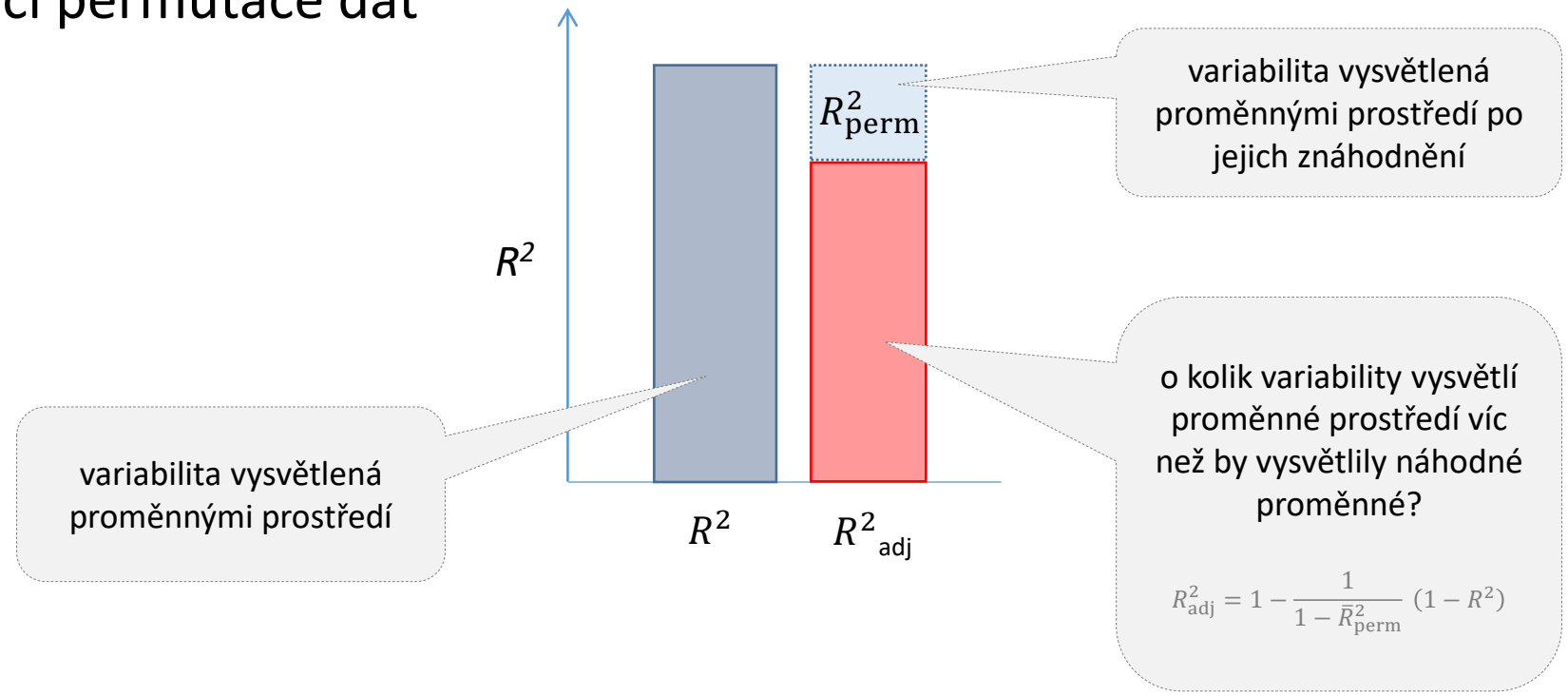
`RsquareAdj {vegan}`

- pomocí Ezekielovy formule

$$R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}} = 1 - \frac{n-1}{n-p-1}(1 - R^2_{\mathbf{Y}|\mathbf{X}})$$

n ... počet vzorků
p ... počet vysvětlujících proměnných
$R^2_{Y|X}$ ... vysvětlená variabilita bez adjustace

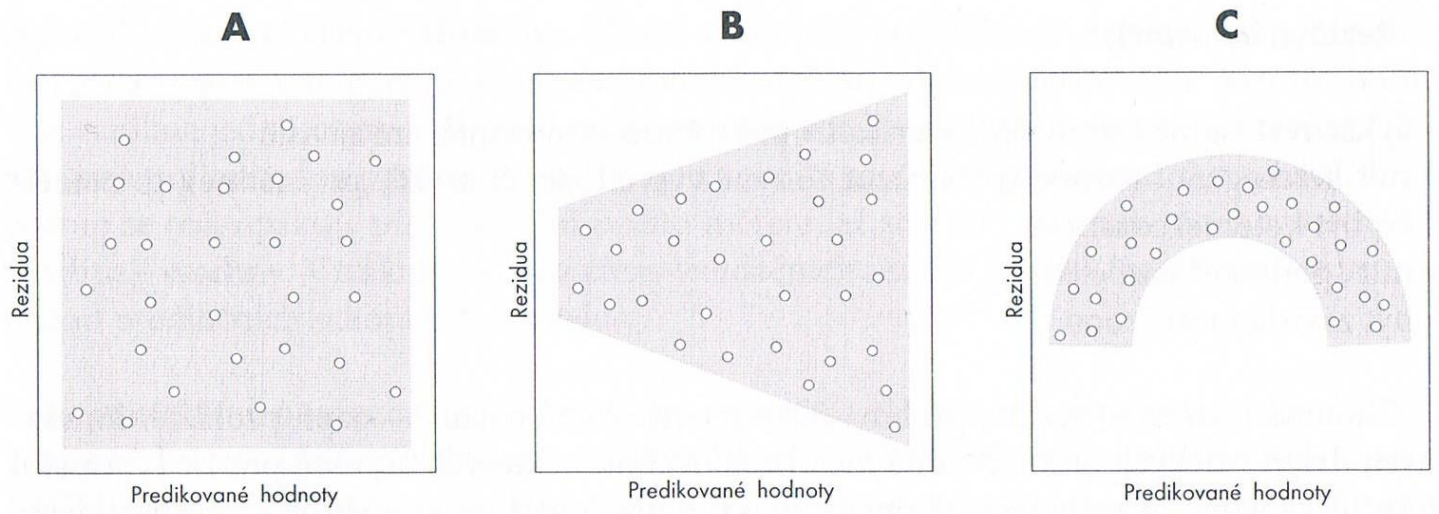
- pomocí permutace dat



Rezidua lineárního regresního modelu

`resid()`

- Rezidua by neměla:
  - vykazovat trendy vůči kterékoliv proměnné, vysvětlující ani závislé
  - mít heterogenní rozptyl (přes různé úrovně vysvětlující či závislé proměnné), tj. neměla by bý heteroskedastická
  - mít „podivné“ rozdělení (předpokládá se normální)
  - být závislá mezi sebou (autokorelovaná)



Obr. 4-1 Závislost reziduí na predikovaných hodnotách. **A**. Homogenní rozptyl. **B**. Rozptyl rostoucí se střední hodnotou. **C**. Zakřivený trend v reziduích. Pro názornost je trend zvýrazněn šedou plochou.

Pekár & Brabec (2009)

Zobecněné lineární modely (GLM)

`glm()`

- Umožňují modelovat proměnné, které nesplňují předpoklady lineárního modelu
  - „nenormální“ rozložení dat (Lognormální, Poissonovo, Binomické atp.)
  - rozptyl se mění s průměrem
  - …

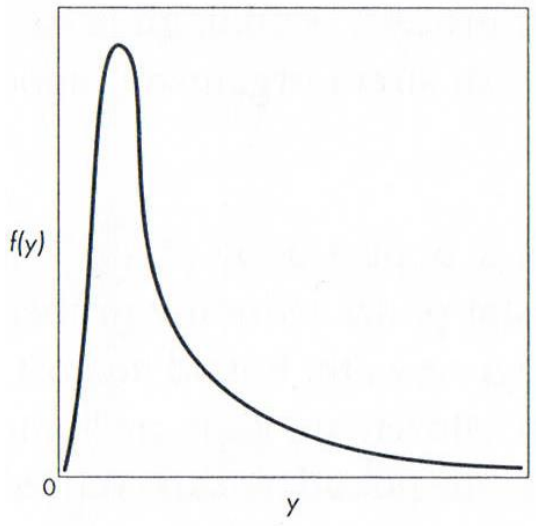
$$\eta_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ji} \qquad y_i = \hat{y}_i + \varepsilon_i, \text{kde } g(\hat{y}_i) = \eta_i$$

kanonická link funkce

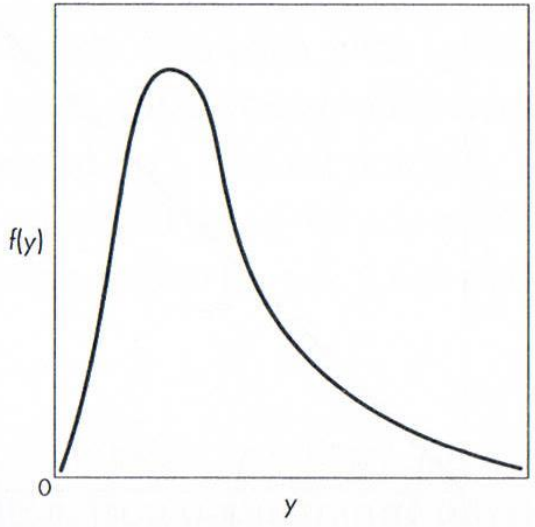
- Parametry GLM
  - Transformační funkce (link) – volí se podle typu rozložení dat
  - Lineární prediktor
  - Náhodná složka

Typy rozdělení

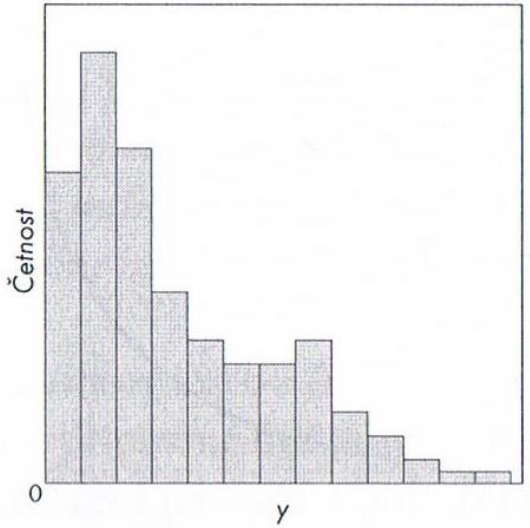
Lognormální rozdělení



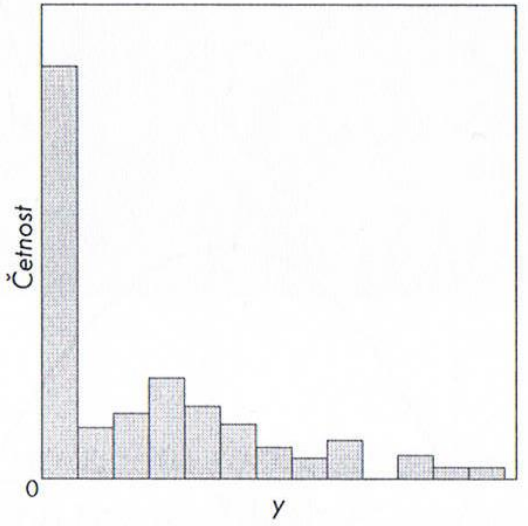
Gamma rozdělení



Poissonovo rozdělení

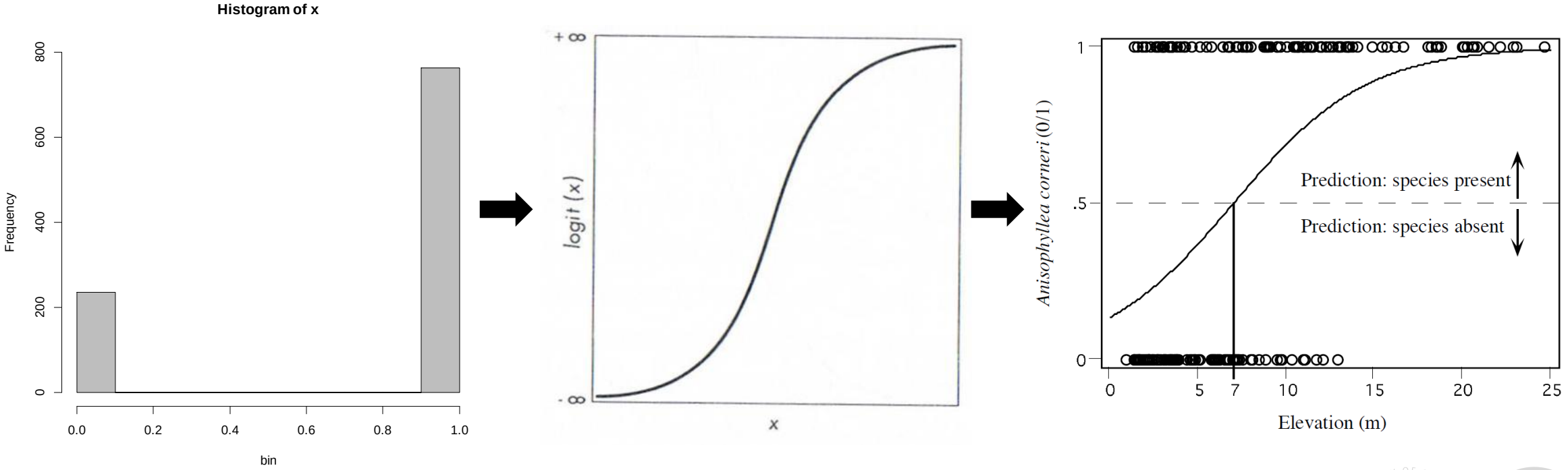


Negativně binomické rozdělení



Binární proměnné

- Např. přítomnost (1) × nepřítomnost (0) druhu
- Binární proměnná + GLM + logit link → **logistická regrese (*logistic regression*)**



Legendre & Legendre (2012)

Kanonické link funkce

Rozdělení	Jméno linku	Link funkce	Rozptyl	Vysvětlovaná proměnná	
				Hodnoty	Typy údajů
Gaussovo (normální)	identity	$\hat{y}$	1	jakékoliv reálné	s ohledem na ostatní možnosti skoro žádné
Gamma	inverse	$\frac{1}{\hat{y}}$	$\hat{y}^2$	kladné reálné	velikosti, hmotnosti, jejich podíly
Poissonovo	log	$\log(\hat{y})$	$\hat{y}$	celé nezáporné	počty případů
Binomické	logit	$\log\frac{\hat{y}}{1-\hat{y}}$	$\frac{\hat{y}(1-\hat{y})}{n}$	podíly z počtů	pravděpodobnosti jevů či výsledků

Pekár & Brabec (2009); Šmilauer (2007)

GLM v prostředí R

```
m <- glm(Y ~ X1 + X2 + … + X5, family = poisson)
anova(m, test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: Species

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         19     71.975
Altitude    1   41.217        18     30.758 1.363e-10 ***
Slope       1    0.454        17     30.305  0.500606
pH          1   10.081        16     20.224  0.001498 **
Moisture    1    5.241        15     14.983  0.022061 *
E3_cover    1    0.720        14     14.263  0.396138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

analogie součtu čtverců (MSS) v LM; pro GLM s Gaussovým linkem totožné se součtem čtverců; jinak se výpočet liší

analogie reziduálnímu součtu čtverců (RSS) v LM; pro GLM s Gaussovým linkem totožné se součtem čtverců; jinak se výpočet liší

stupně volnosti (Degrees of freedom)

Výsledek χ^2 testu (pokud Poissnovo rozdělení); pokud byl vhodnější F-test nutné specifikovat v argumentu 'test'

GLM v prostředí R

```
m <- glm(Y ~ X1 + X2 + … + X5, family = poisson)
summary(m)
```

…

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.44461    0.06888  35.490   <2e-16 ***
Altitude    -0.21654    0.11605  -1.866   0.0620 .
Slope       -0.03097    0.06996  -0.443   0.6581
pH           0.24419    0.11070   2.206   0.0274 *
Moisture     0.13165    0.07494   1.757   0.0789 .
E3_cover     0.07156    0.08460   0.846   0.3976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 71.975  on 19  degrees of freedom
Residual deviance: 14.263  on 14  degrees of freedom
AIC: 111.69

Number of Fisher Scoring iterations: 4
```

tzv. Waldovy statistiky – jejich předpoklady často nejsou splněny

odhady koeficientů a jejich chyby; pokud je použita logaritmická link funkce lze převést na jednotky `Y` pomocí exponenciální funkce `exp()`

statistická významnost

analogie TSS

analogie RSS

Akaikeho informační kritérium

Rezidua v GLM

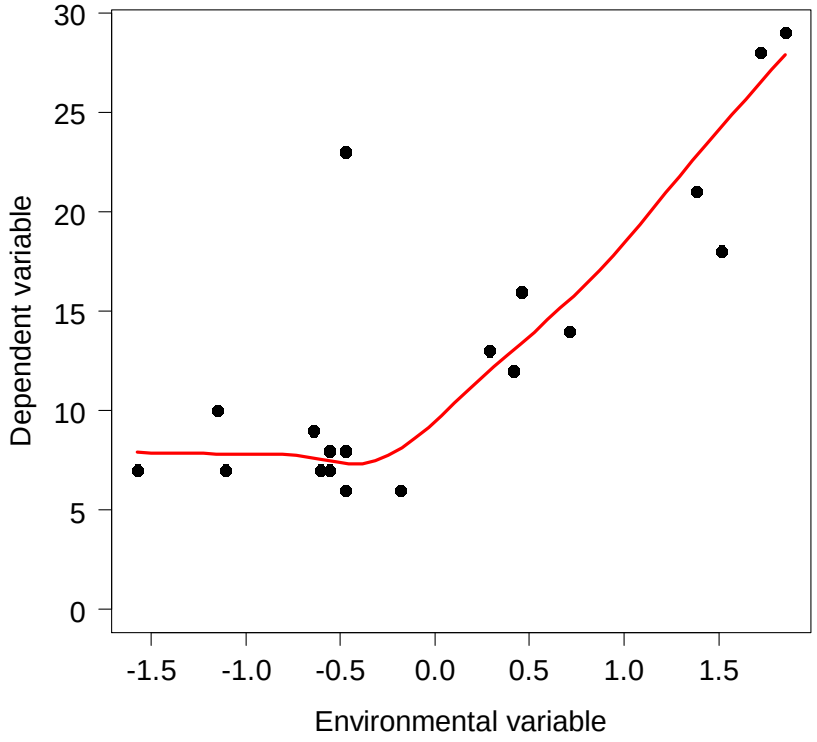
`resid()`

- U modelů s jiným než identickým linkem (Gaussovo rozdělení) je několik typů reziduí
  - Rezidua na trasformované škále (např. log při Poissonově rozdělení): `type = "working"`
  - Pearsonova rezidua (obdoba standardizovaných reziduí v LM): `type = "pearson"`
  - Prostá reziuda na původní škále: `type = "response"`

Další vychytávky (nelineární trendy)

- **Lowess and loess smoothing methods**
  - Neparametrický odhad trendu pořízený na základě velmi flexibilní lokální regrese
  - Fituje křivku na data → dobré pro ukázání vztahu proměnných
  - Vhodné jen pro n < 1000, pokud více → GAM

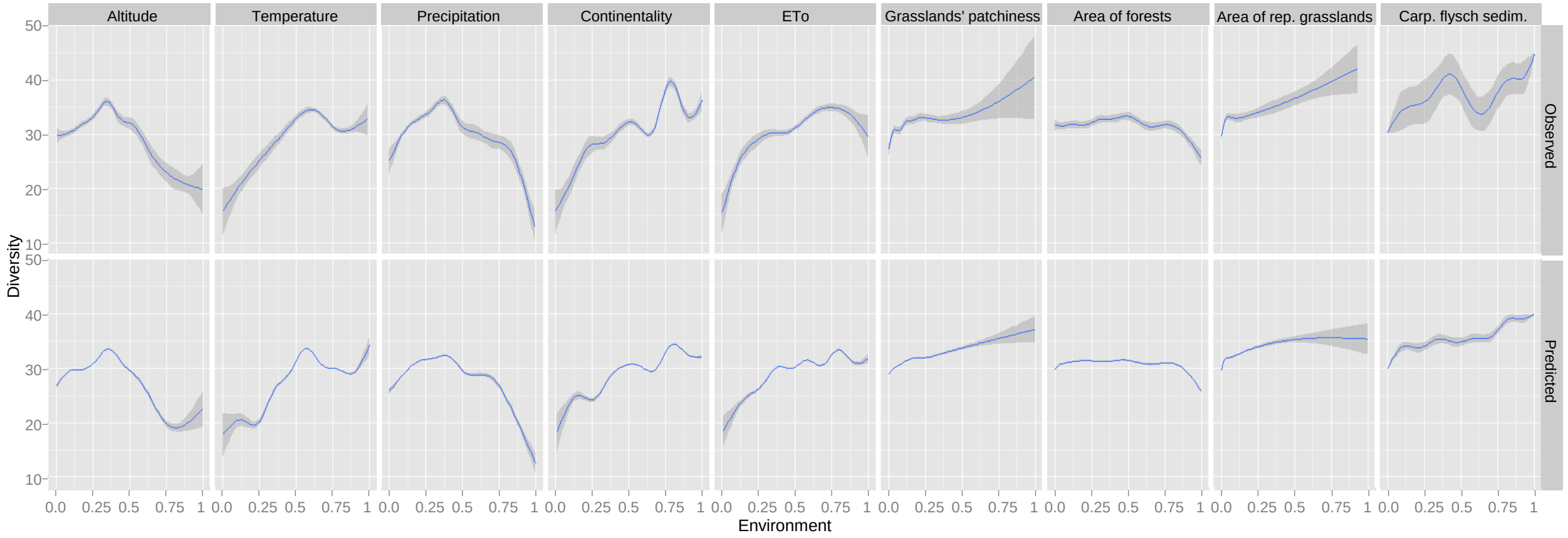
```
loess.smooth()
lowess()
```



Další vychytávky (nelineární trendy)

- Generalized Aditive Models (GAM)

`gam()`



Forward selection a variation partitioning v lineární regresi

Lineární regresní model v R

```
m <- lm(Y ~ X1 + X2 + … + X5)
anova(m)
```

```
Analysis of Variance Table

Response: Species
           Df Sum Sq Mean Sq F value    Pr(>F)
Altitude    1 503.89  503.89 31.3657 6.541e-05 ***
Slope       1   2.61    2.61  0.1622  0.693243
pH          1 182.77  182.77 11.3768  0.004551 **
Moisture    1  76.63   76.63  4.7702  0.046465 *
E3_cover    1  31.73   31.73  1.9753  0.181690
Residuals 14 224.91   16.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lineární regresní model v R

```
m <- lm(Y ~ X1 + X2 + … + X5)
anova(m)
```

	Slope	pH	Moisture	E3_cover
Altitude	0.296	-0.759	0.268	-0.331
Slope		-0.221	0.085	-0.408
pH			-0.229	0.461
Moisture				0.149

```
Analysis of Variance Table

Response: Species
           Df Sum Sq Mean Sq F value    Pr(>F)
pH          1 667.23  667.23 41.5325 1.536e-05 ***
Slope       1   6.50    6.50  0.4044   0.53511
Altitude    1  15.55   15.55  0.9678   0.34192
Moisture    1  76.63   76.63  4.7702   0.04647 *
E3_cover    1  31.73   31.73  1.9753   0.18169
Residuals  14 224.91   16.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Forward selection

- Metoda pro výběr souboru „nejlepších“ vysvětlujících proměnných z celého setu proměnných, které mám k dispozici
- Cílem je redukovat počet proměnných, ale zachovat maximální vysvětlenou variabilitu
- Dobře použitelné pro ekologické studie s korelovanými proměnnými (nikoliv pro laboratorní experimenty s propracovaným designem)
- Použitelné v lineární regresi a vícerozměrných metodách (RDA, CCA)
- V R několik funkcí
  - `ordistep {vegan}`
  - `ordiR2step {vegan}`
  - `forward.sel {packfor}`

Jak pracuje forward selection?

- Předem je nutné otestovat signifikanci celého modelu, tj. se všemi vysvětlujícími proměnnými → pokud není signifikantní, nemá smysl dělat FS
- Kroky forward selection:
  1. Každá vysvětlující proměnná se použije v samostatném modelu → zaznamená se vysvětlená variabilita
  2. Seřadí proměnné podle vysvětlené variability od „nejlepší“ po „nejhorší“
  3. Zjistí zda variabilita vysvětlená nejlepší proměnnou je statisticky signifikantní (v regresi použije F-test), pokud není → zastaví výběr
  4. Zjistí kolik variability vysvětlí každá ze zbylých proměnných zatímco první vybraná proměnná je zahrnuta v modelu jako kovariáta
  5. Seřadí proměnné podle vysvětlené variability a pro nejlepší proměnnou otestuje statistickou významnost jejího příspěvku do modelu, pokud nevýznamný → zastaví výběr
  6. Opakuje body 4 a 5 dokud další proměnné významně přispívají do modelu

Jak pracuje forward selection?

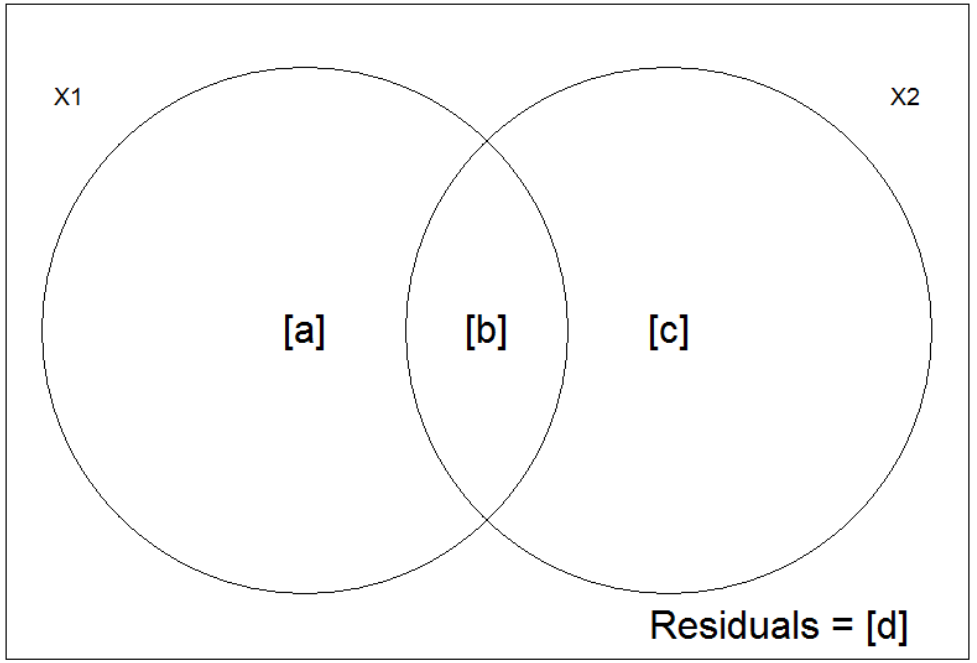
- Kritéria pro zastavení výběru
  1. Statistická signifikance
  2. Adjustovaný R^2 globálního modelu (tj. modelu se všemi proměnnými)

- Lze použít v lineární regresi a přímé ordinaci (RDA, CCA)
- Alternativy k forward selection
  - backward selection
  - forward-backward selection

Rozklad variance (*variation partitioning*)

`varpart {vegan}`

- Umožňuje rozložit variabilitu vysvětlenou danými proměnnými na následující části:
  - [a] Variabilitu vysvětlenou čistým vlivem první proměnné (nebo sadou proměnných)
  - [b] Variabilitu vysvětlenou sdíleným vlivem první a druhé proměnné (případně první a druhou sadou proměnných)
  - [c] Variabilitu vysvětlenou čistým vlivem druhé proměnné (nebo sadou proměnných)
- Je možné použít i více proměnných (jejich sad), ale většinou se končí u 3 až 4
- Lze testovat statistickou signifikanci „čistých vlivů“
- Pokud se skupiny liší počtem proměnných → adjustovaný R^2
- Čím více jsou proměnné korelované tím větší bude sdílená variabilita

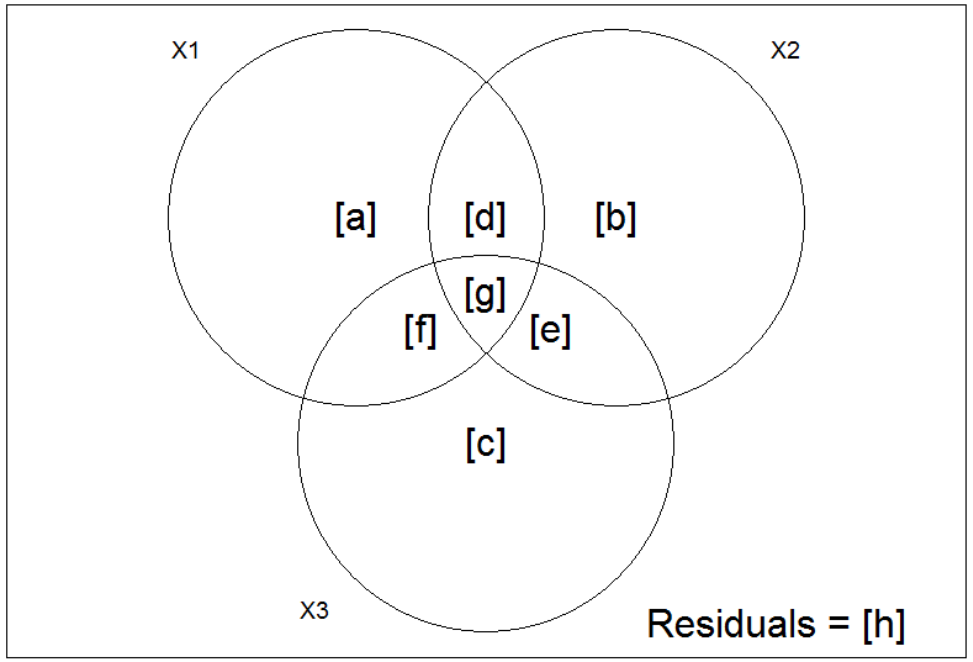


Borcard, el al. (1992)

Rozklad variance (*variation partitioning*)

`varpart {vegan}`

- Umožňuje rozložit variabilitu vysvětlenou danými proměnnými na následující části:
  - [a] Variabilitu vysvětlenou čistým vlivem první proměnné (nebo sadou proměnných)
  - [b] Variabilitu vysvětlenou sdíleným vlivem první a druhé proměnné (případně první a druhou sadou proměnných)
  - [c] Variabilitu vysvětlenou čistým vlivem druhé proměnné (nebo sadou proměnných)
- Je možné použít i více proměnných (jejich sad), ale většinou se končí u 3 až 4
- Lze testovat statistickou signifikanci „čistých vlivů“
- Pokud se skupiny liší počtem proměnných → adjustovaný R^2
- Čím více jsou proměnné korelované tím větší bude sdílená variabilita

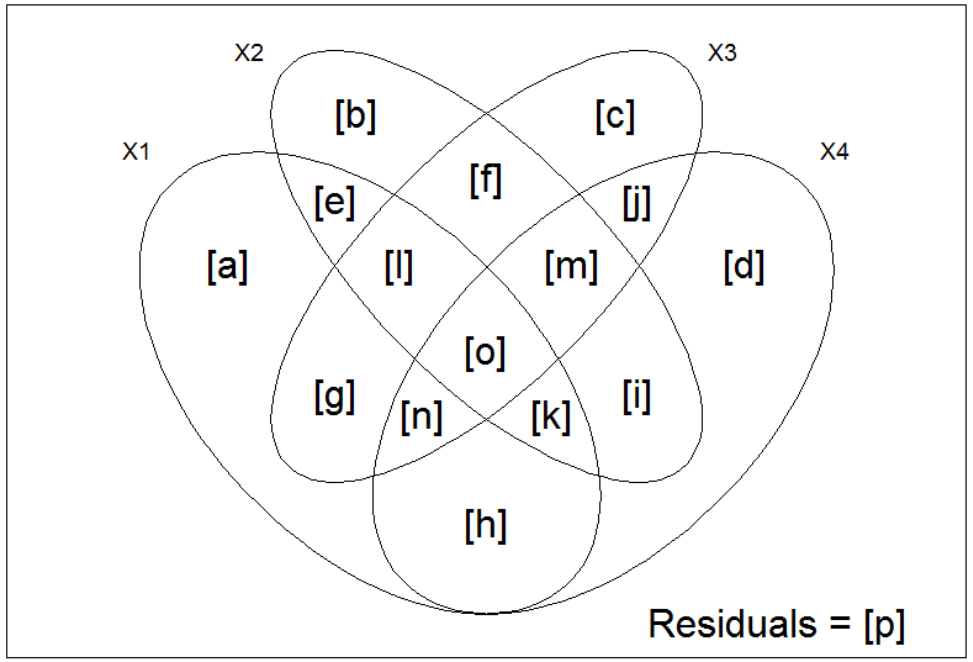


Borcard, el al. (1992)

Rozklad variance (*variation partitioning*)

`varpart {vegan}`

- Umožňuje rozložit variabilitu vysvětlenou danými proměnnými na následující části:
  - [a] Variabilitu vysvětlenou čistým vlivem první proměnné (nebo sadou proměnných)
  - [b] Variabilitu vysvětlenou sdíleným vlivem první a druhé proměnné (případně první a druhou sadou proměnných)
  - [c] Variabilitu vysvětlenou čistým vlivem druhé proměnné (nebo sadou proměnných)
- Je možné použít i více proměnných (jejich sad), ale většinou se končí u 3 až 4
- Lze testovat statistickou signifikanci „čistých vlivů“
- Pokud se skupiny liší počtem proměnných → adjustovaný R^2
- Čím více jsou proměnné korelované tím větší bude sdílená variabilita



Borcard, el al. (1992)

Diversity and Distributions, (Diversity Distrib.) (2013) **19**, 423–432

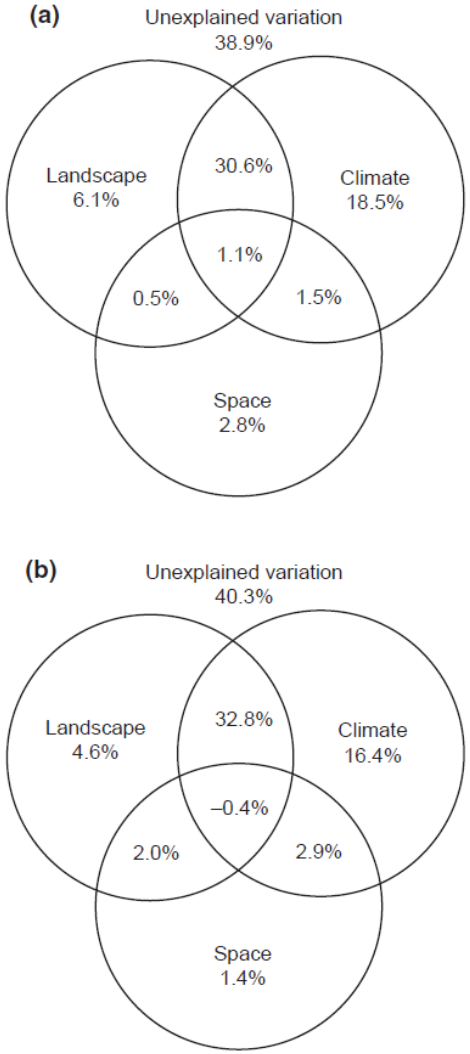
BIODIVERSITY RESEARCH

Does local habitat fragmentation affect large-scale distributions? The case of a specialist grassland bird

Luís Reino[1,2,3]*, Pedro Beja[3], Miguel B. Araújo[1,4,5], Stéphane Dray[6] and Pedro Segurado[2]

Table 1 Summary statistics (mean ± standard deviation) of predictor variables considered in the analyses of factors affecting the presence or absence of the calandra lark in 10 × 10 km squares across the Iberian Peninsula

Variables (Abbreviation; Units)	Presence	Absence
Total annual precipitation (PREC; mm)	504.0 ± 102.4	730.6 ± 276.3
Mean annual air temperature (TANN; °C)	12.9 ± 23.4	12.6 ± 30.3
Annual temperature range (TRAN; °C)	28.6 ± 26.1	24.9 ± 40.6
Mean slope (SLOP;%)	3.6 ± 2.4	8.5 ± 5.4
Habitat area (AREA; km^2)		
Total	29.9 ± 25.2	6.9 ± 11.3
Effective	11.9 ± 21.6	1.5 ± 5.4
Number of patches (NUMP; N)		
Total	7.5 ± 5.8	7.5 ± 7.1
Effective	6.2 ± 4.3	5.2 ± 5.03
Edge density (EDEN; km ha^{-1})		
Total	0.59 ± 0.31	0.25 ± 0.28
Effective	0.72 ± 0.34	0.40 ± 0.37



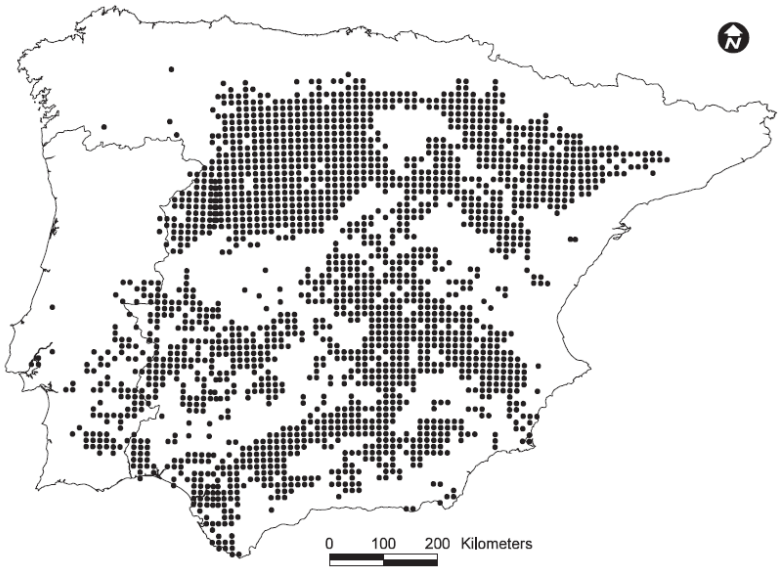


Figure 3 Variation partitioning Venn diagrams representing the pure and shared contributions of climatic/topographic, landscape and spatial sets of variables to the explained variation in the distribution of calandra lark in the Iberian Peninsula. Landscape variables were computed considering either the total (a) or the effective habitat (b).

kalandra zpěvná (*Melanocorypha calandra*)

Literatura

- Legendre, P. & Legendre, L. (2012): Numerical ecology. Third Edition. Elsevier, Amsterdam.
- Borcard, D., Gillet, F. & Legendre, P. (2011): Numerical ecology with R. Springer, New York.
- Borcard, D., Legendre, P. & Drapeau, P. (1992): Partialling out the spatial component of ecological variation. Ecology, 73: 1045–1055
- Pekár, S. & Brabec, M. (2009): Moderní analýza biologických dat. Scientia, Praha.