Statistics for the Sciences

Multiple Linear Regression Models

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Introduction
- MLR Models
- Model formulation
- Scatter plot matrix
- Estimation
- ANOVA
- Statistical inferences
- Lab

Introduction

- `loyn.csv`: Loyn (1987) selected 56 forest patches in southeastern Victoria, Australia, and related the abundance of forest birds in each patch to six predictor variables: patch area (ha), distance to nearest patch (km), distance to nearest larger patch (km), grazing stock (1 to 5 indicating light to heavy), altitude (m) and years since isolation (years).

```
##    abund area yearisol dist distl graze alt
## 1    5.3  0.1     1968   39    39     2 160
## 2    2.0  0.5     1920  234   234     5  60
## 3    1.5  0.5     1900  104   311     5 140
## 4   17.1  1.0     1966   66    66     3 160
## 5   13.8  1.0     1918  246   246     5 140
## 6   14.1  1.0     1965  234   285     3 130
## 7    3.8  1.0     1955  467   467     5  90
## 8    2.2  1.0     1920  284  1829     5  60
## 9    3.3  1.0     1965  156   156     4 130
## 10   3.0  1.0     1900  311   571     5 130
```

Introduction

- Response variable `abund`, the target that we wish to predict
- with the following five **predictors** or independent variables as input
  - area
  - yearisol
  - dist
  - distl
  - graze
  - alt
- The aim was to develop a **best** predictive model relating bird abundance to these predictors. Perhaps we can use a model

$$\text{abund} \approx f(\text{area, dist, distl, graze, alt, yearisol})$$

Introduction

- We can refer to the input vector collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}$$

- Now we can write our model as

$$Y = f(X) + \varepsilon$$

where ε captures measurement errors and other discrepancies.
  - One such model is `Multiple Linear Regression Models`

MLR Models

- Data:

Y	X_1	$\cdots$	X_k
y_1	x_{11}	$\cdots$	x_{1k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	$\cdots$	x_{nk}

- Y: Response variable
- $X_1, X_2, \ldots, X_k$: Predictors or independent variables

MLR Models

Definition. A linear statistical model relating a random response Y to a set of independent variables $X_1, X_2, \ldots, X_k$ is of the form

$$Y|_{X_1=x_1,\ldots,X_k=x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ε is a random variable, and the variables $X_1, X_2, \ldots, X_k$ assume known values.

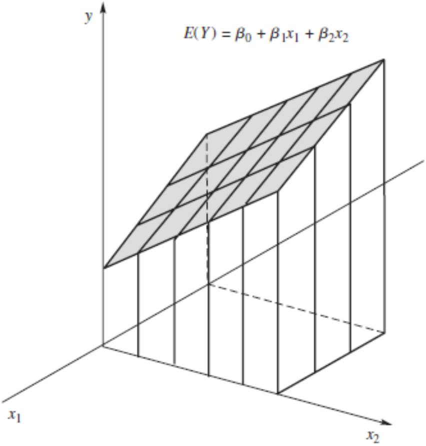
- We will assume that $E(\varepsilon) = 0$, and hence that

$$E(Y|_{X_1=x_1,\ldots,X_k=x_k}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

- When $k = 1$, the model is the simple linear regression model.

MLR Models

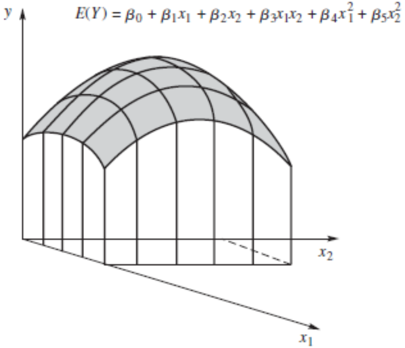
Plot of $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



MLR Models

- **New predictors** can be created by transforming available predictors

Plot of $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$



MLR Models

- Matrix notation: We define the following matrices

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then the MLR model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

where ε has a multivariate distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 I_n$, and I_n is a n-dimensional identity matrix.

MLR Models

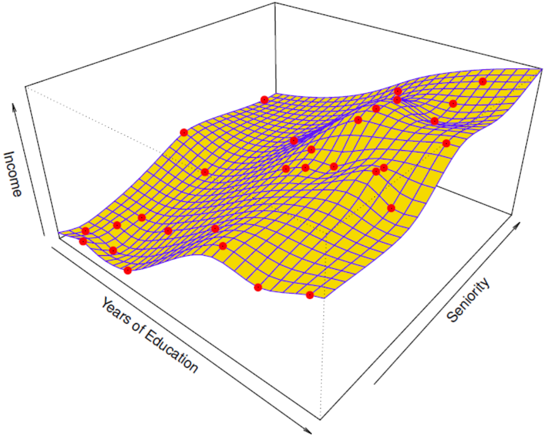
- $\mathbf{Y}$ is generally assumed to have a multivariate normal distribution $\mathbf{Y} \sim MVN(\mathbf{X}\beta, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the Variance-Covariance Matrix of $\mathbf{Y}$:

$$\mathbf{\Sigma} = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & \cdots & Cov(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \cdots & Var(Y_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

- **Question** What does `linear` in multiple linear regression models mean?

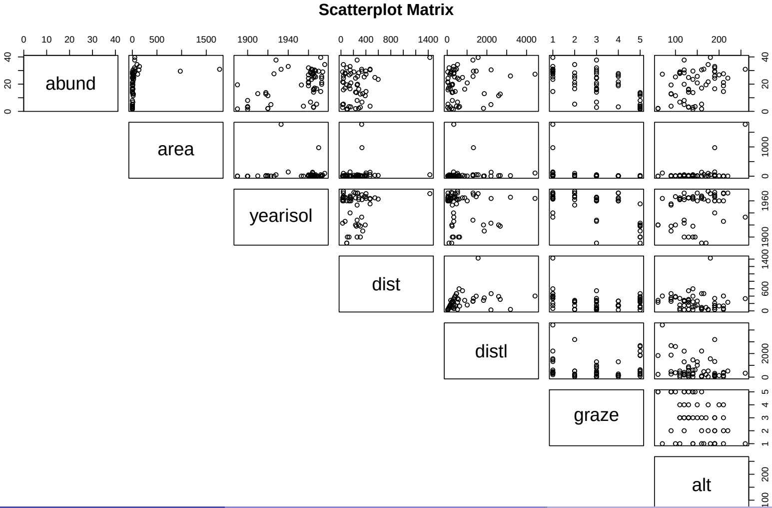
MLR Models

- We can build a very complex model without fitting errors (all fitted values are equal to the corresponding observed values). It is called **overfitting**.
  - Overfitting performs very bad in predictions.
  - It is often possible to get more accurate predictions with a simpler, instead of a complicated model.



Scatter plot matrix

- How do we build a reasonable **linear** model for given Y's and predictors $X_1, X_2, \ldots, X_k$?
  - We need to check the relationship between Y and each X_i



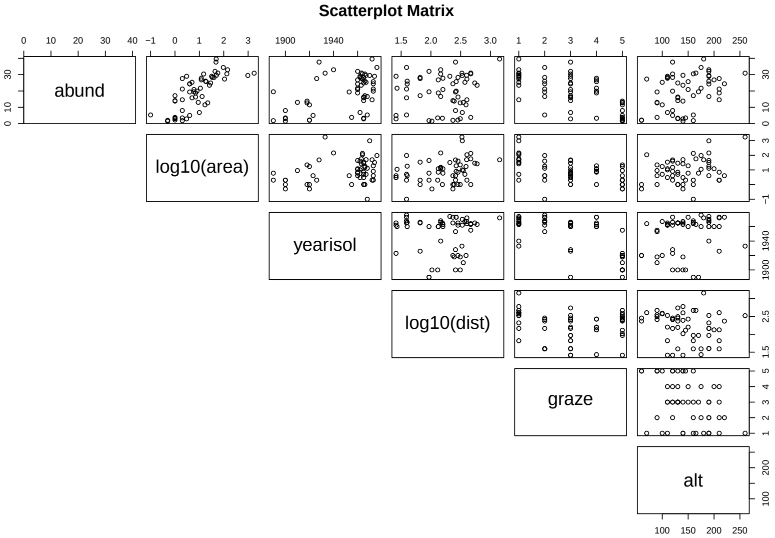
Scatter plot matrix

- We may check the correlation matrix as well

```
##           abund      area     yearisol      dist       distl       graze
## abund         1 0.2559702  0.503357741 0.2361125  0.08715258 -0.68251138
## area            1.0000000 -0.001494192 0.1083429  0.03458035 -0.31040242
## yearisol                   1.000000000 0.1132175 -0.08331686 -0.63556710
## dist                                   1.0000000  0.31717234 -0.25584182
## distl                                              1.00000000 -0.02800944
## graze                                                          1.00000000
## alt
##                  alt
## abund      0.3858362
## area       0.3877539
## yearisol   0.2327154
## dist      -0.1101125
## distl     -0.3060222
## graze     -0.4071671
## alt        1.0000000
```

Scatter plot matrix

- Now let's remove predictor `distl` and consider `log` transformation of `area` and `dist`

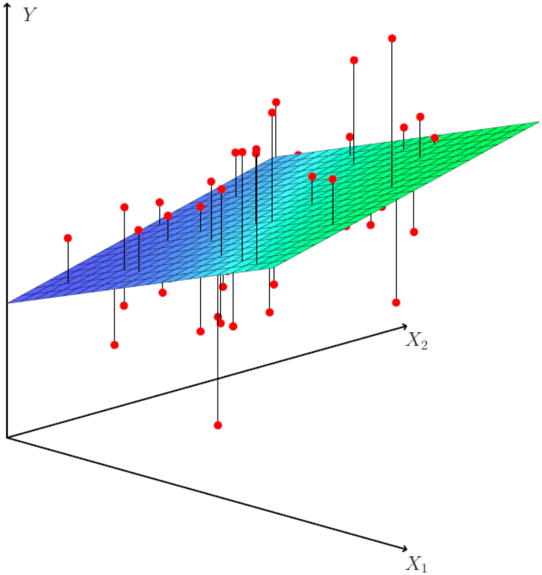


Scatter plot matrix

- The new correlation matrix

```
##          abund  logarea  yearisol     logdist      graze        alt
## abund        1 0.7400358 0.5033577  0.12672333 -0.6825114  0.3858362
## logarea        1.0000000 0.2784145  0.30216662 -0.5590886  0.2751428
## yearisol                 1.0000000 -0.01957223 -0.6355671  0.2327154
## logdist                             1.00000000 -0.1426392 -0.2190070
## graze                                           1.0000000 -0.4071671
## alt                                                        1.0000000
```

Estimation by Least Squares



Analysis of Variance

- The Analysis of Variance for MLR models can be summarized in the following table.

Source	df	SS	MS	F
Regression	k	SSR	$MSR = SSR/k$	MSR/MSE
Error	n-1-k	SSE	$MSE = SSE/(n-1-k)$	
Total	n-1	SS_{total}		

where $SSR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$, $SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ and $SS_{total} = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

- **Note.** The F-test is for $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ versus $H_a : \beta_i \neq 0$ for some $i = 1, 2, \ldots, k$. And the F-test statistic (Exercise 11.84(a)) has an F distribution under H_0 with $df_1 = k, df_2 = n - 1 - k$.
- H_0 is rejected only if the calculated test statistic F^* is large: given significance level α, H_0 is rejected only if $F^* \geq F_{df_1, df_2, 1-\alpha}$.

Analysis of Variance

- **The Coefficient of Multiple Determination**, R^2, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{SS_{total}}.$$

- R^2 is
  - The proportion of variation in the response explained by the regression.
  - The proportion by which the unexplained variation in the response is reduced by the regression.

- One problem with using R^2 to measure the quality of model fit, is that it can always be increased by adding another regressor.
- The **Adjusted Coefficient of Multiple Determination**, R_a^2, is a measure that adjusts R^2 for the number of regressors in the model. It is defined as

$$R_a^2 = 1 - \frac{\text{SSE}/(n-1-k)}{SS_{total}/(n-1)}.$$

Statistical inference problems

- Suppose that the MLR model is $Y_i|_{X_1=x_{1i},\ldots,X_k=x_{ki}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, i = 1, \ldots, k$
- Inferences about individual parameters: $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$, $i = 1, 2, \ldots, k$
- Inferences about a set of parameters: testing $H_0 : \beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$ versus H_a : At least one of the $\beta_i, i = r+1, \ldots, k$ differs from 0 which is checking if a reduced model is sufficient:
- Model R (Reduced model):

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

- Model C (Complete model):

$$\begin{aligned} E(Y) = \beta_0 &+ \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r X_r \\ &+ \beta_{r+1} X_{r+1} + \beta_{r+2} X_{r+2} + \cdots + \beta_k X_k \end{aligned}$$

Statistical inference problems

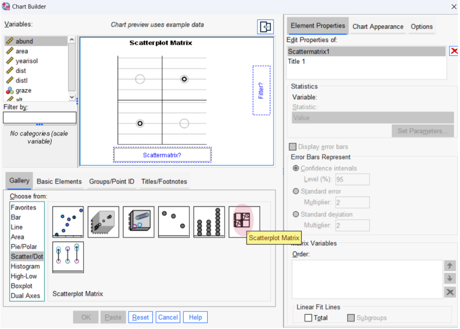
Let $\mathbf{x} = \mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_k^*)$ be a vector of new observation of the predictors.

- Predicting the average Value of Y: $E(Y) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*$
- Predicting a Particular Value of $Y = Y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^* + \varepsilon$

Remark. Again, prediction intervals for the actual value of Y are longer than confidence intervals for $E(Y)$ if both confidence levels are the same and both are determined for the same value of $\mathbf{x} = \mathbf{x}^*$.

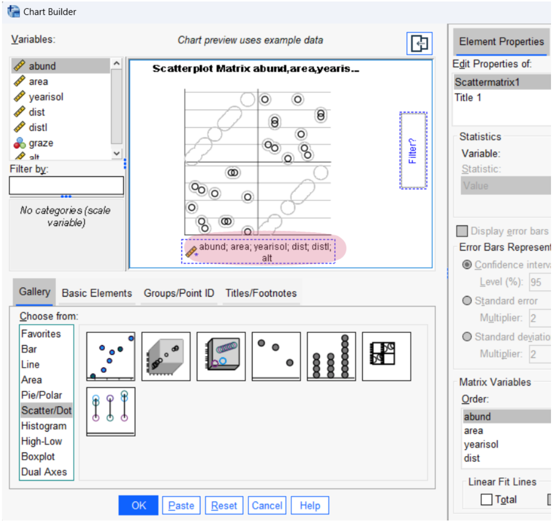
Lab

- `loyn.csv`: Loyn (1987) selected 56 forest patches in southeastern Victoria, Australia, and related the abundance of forest birds in each patch to six predictor variables: patch area (ha), distance to nearest patch (km), distance to nearest larger patch (km), grazing stock (1 to 5 indicating light to heavy), altitude (m) and years since isolation (years).
- After importing data, Click on `Graphs` in the top menu → Select `Chart Builder...` to plot a scatter plot matrix
  - In the Chart Builder dialog box, drag the `Scatterplot Matrix` under icon `Scatter/Dot` from the Gallery tab into the Chart Preview area.



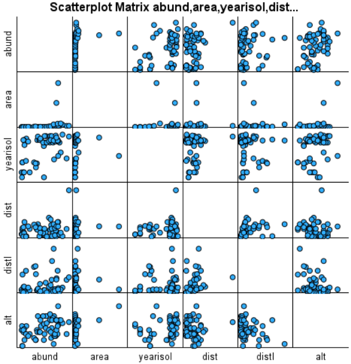
Lab

- Drag all numerical variables to the `Scattermatrix?` box.
  - `graze` is a nominal variable.



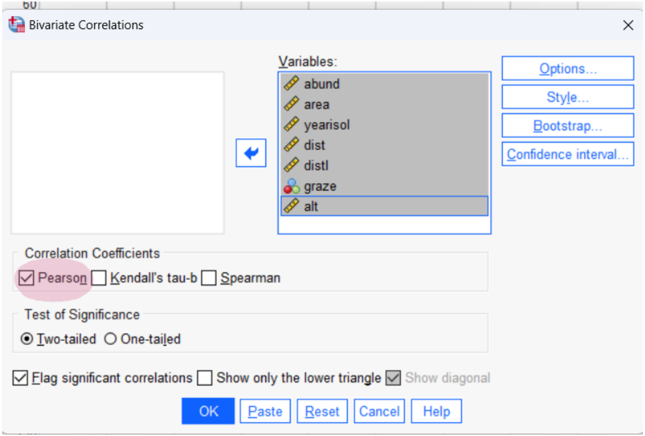
Lab

- Click OK to generate the scatter plot matrix.
- Or We can get a scatter plot matrix by clicking on `Graphs` → `Scatter/Dot ...` → `Matrix Scatter`



Lab

- To get the correlation matrix, click on `Analyze` → `Correlate` → `Bivariate..`, and then Add all numerical variables to the `Variables` box.



Lab

Correlations

Correlations

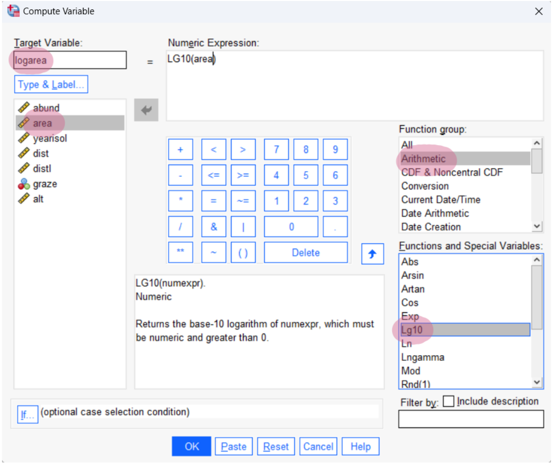
		abund	area	yearisol	dist	distl	graze	alt
abund	Pearson Correlation	1	.256	.503**	.236	.087	-.683**	.386**
	Sig. (2-tailed)		.057	<.001	.080	.523	<.001	.003
	N	56	56	56	56	56	56	56
area	Pearson Correlation	.256	1	-.001	.108	.035	-.310*	.388**
	Sig. (2-tailed)	.057		.991	.427	.800	.020	.003
	N	56	56	56	56	56	56	56
yearisol	Pearson Correlation	.503**	-.001	1	.113	-.083	-.636**	.233
	Sig. (2-tailed)	<.001	.991		.406	.542	<.001	.084
	N	56	56	56	56	56	56	56
dist	Pearson Correlation	.236	.108	.113	1	.317*	-.256	-.110
	Sig. (2-tailed)	.080	.427	.406		.017	.057	.419
	N	56	56	56	56	56	56	56
distl	Pearson Correlation	.087	.035	-.083	.317*	1	-.028	-.306*
	Sig. (2-tailed)	.523	.800	.542	.017		.838	.022
	N	56	56	56	56	56	56	56
graze	Pearson Correlation	-.683**	-.310*	-.636**	-.256	-.028	1	-.407**
	Sig. (2-tailed)	<.001	.020	<.001	.057	.838		.002
	N	56	56	56	56	56	56	56
alt	Pearson Correlation	.386**	.388**	.233	-.110	-.306*	-.407**	1
	Sig. (2-tailed)	.003	.003	.084	.419	.022	.002	
	N	56	56	56	56	56	56	56

**. Correlation is significant at the 0.01 level (2-tailed).

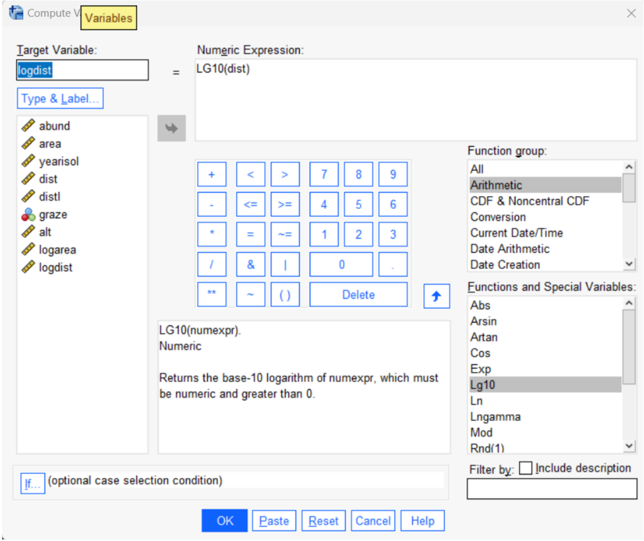
*. Correlation is significant at the 0.05 level (2-tailed).

Lab

- Conduct log transformation for two variables `area` and `dist`
- Click on `Transform` → `Compute Variable ...`

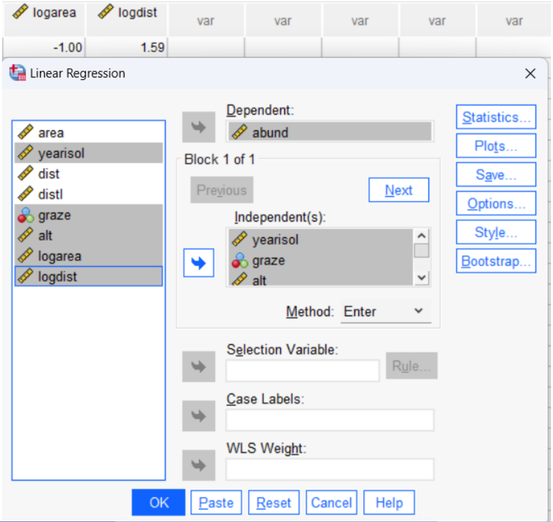


Lab


Compute Variables
Target Variable:
logdist
Type & Label...
abund
area
yearisol
dist
distl
graze
alt
logarea
logdist
=
Numeric Expression:
LG10(dist)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Abs
Arsin
Artan
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd(1)
Delete
LG10(numexpr).
Numeric
Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
If...
(optional case selection condition)
Filter by:
Include description
OK
Paste
Reset
Cancel
Help

Lab

- Follow the procedure fitting a simple linear regression model, to fit an MLR, Click on `Analyze` → `Regression` → `Linear ...`
  - Model: `abund ~ logarea + logdist + graze + alt + yearisol`



Lab

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.827[a]	.684	.653	6.3256

a. Predictors: (Constant), logdist, yearisol, alt, logarea, graze

b. Dependent Variable: abund

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4337.272	5	867.454	21.679	<.001[b]
	Residual	2000.656	50	40.013		
	Total	6337.929	55			

a. Dependent Variable: abund

b. Predictors: (Constant), logdist, yearisol, alt, logarea, graze

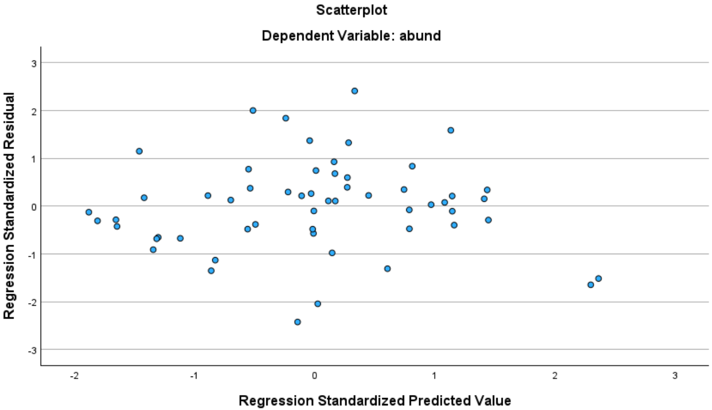
Lab

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-131.847	88.640		-1.487	.143	-309.886	46.191
	yearisol	.077	.044	.182	1.744	.087	-.012	.165
	graze	-1.676	.921	-.230	-1.819	.075	-3.526	.174
	alt	.021	.023	.087	.937	.353	-.025	.067
	logarea	7.295	1.336	.552	5.460	<.001	4.612	9.979
	logdist	-1.303	2.319	-.050	-.562	.577	-5.961	3.354

a. Dependent Variable: abund

Lab



Scatterplot
Dependent Variable: abund
Regression Standardized Residual
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Regression Standardized Predicted Value

License