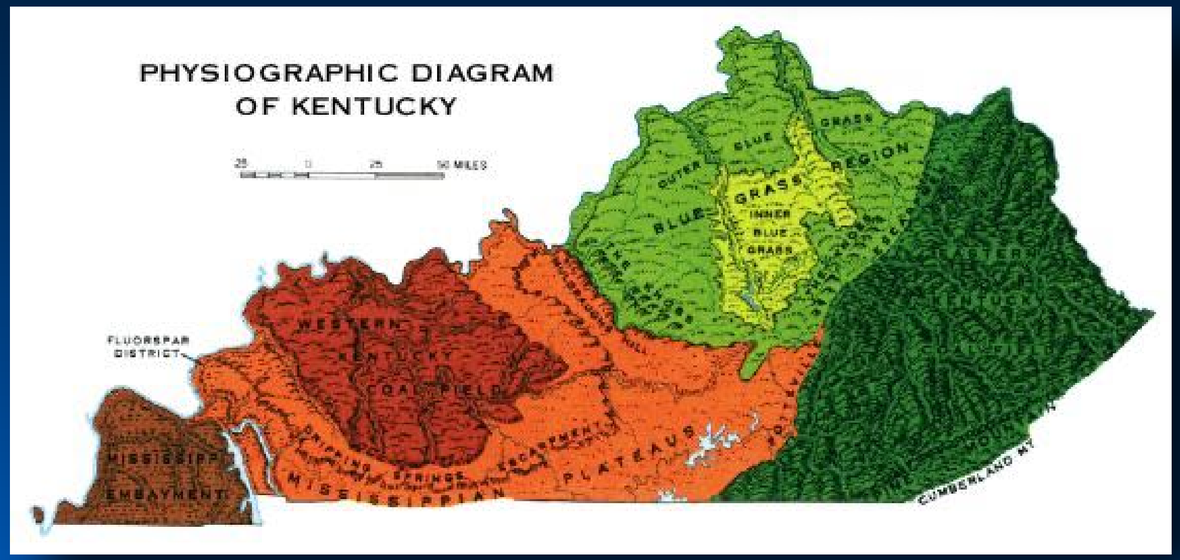
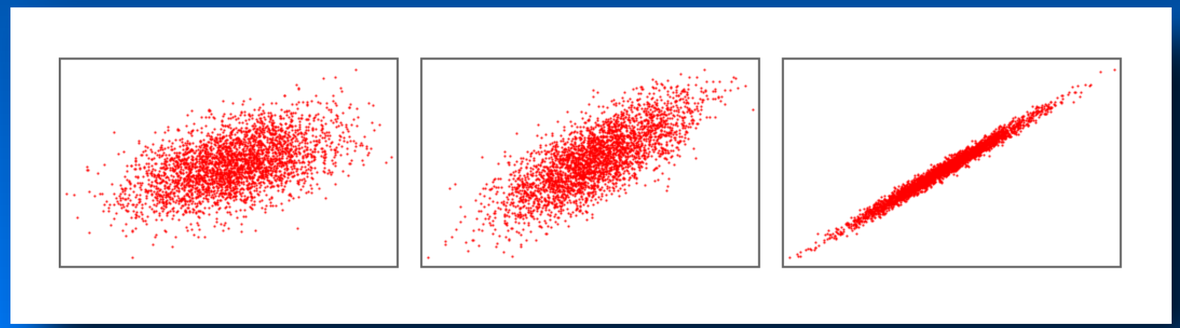
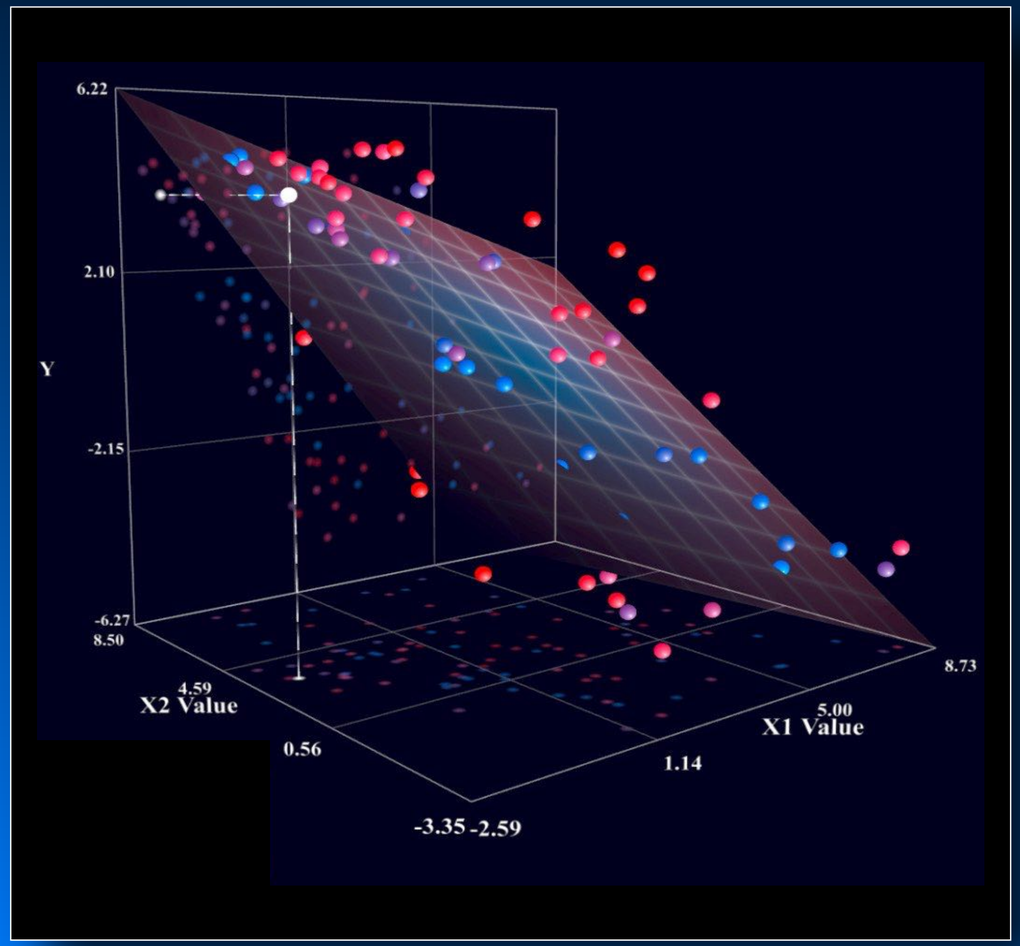
Multiple Regression

Prof. Norman MacLeod

School of Earth Sciences & Engineering, Nanjing University







Multivariate Data

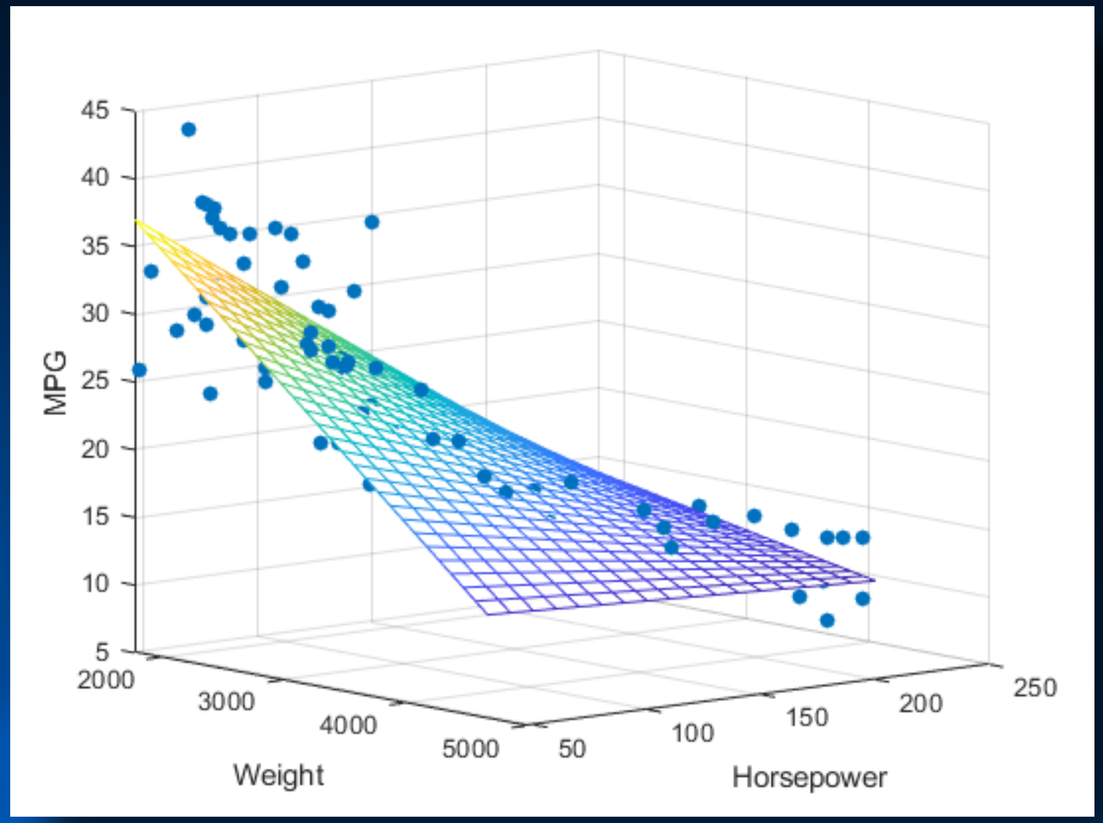
Data Configurations

Data Configuration	Description	Use
Univariate	Data consisting of a single variable.	Descriptive Statistics, Sequential Data Analysis, Time-Series Analysis
Bivariate	Data consisting of two variables	Bivariate Regression Analyses, Modeling
Multivariate	Data consisting of two or more variables	Multiple Regression Analyses, Ordination Analyses, Discriminant Analyses, Machine Learning, Artificial Intelligence

Multiple Regression

An extension of linear regression analysis in which the value of a dependent variable (y) is predicted based on the values of two or more independent variables (x).

Multiple regression problems are usually approached using matrix algebraic methods to solve a set of simultaneous equations. Multiple regression is also related to curvilinear regression where the point is to predict the value of a dependent variable (y) on the basis of a polynomial expansion of the standard linear regression relation ($y = \alpha + \beta x$).



Least-Squares Regression

A least squares regression model attempts to quantify the relation between a dependent variable (y) in terms of one or more independent variables (x) such that the square of deviations from the model (ε) are minimized.

Linear Least-Squares Regression

$$y_i = \alpha + \beta_1 x_i + \varepsilon_i$$

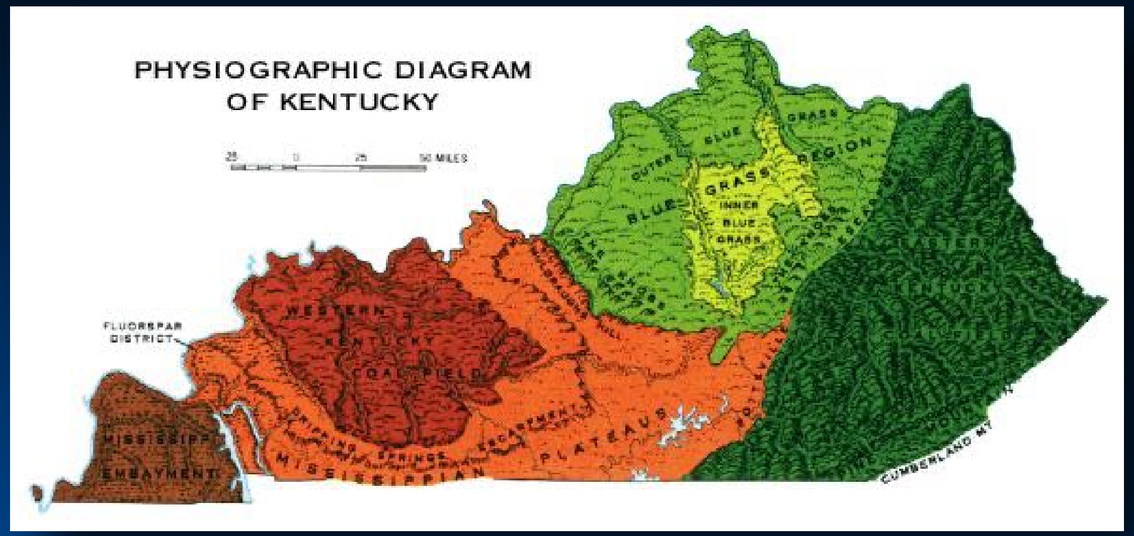
Curvilinear Least-Squares Regression

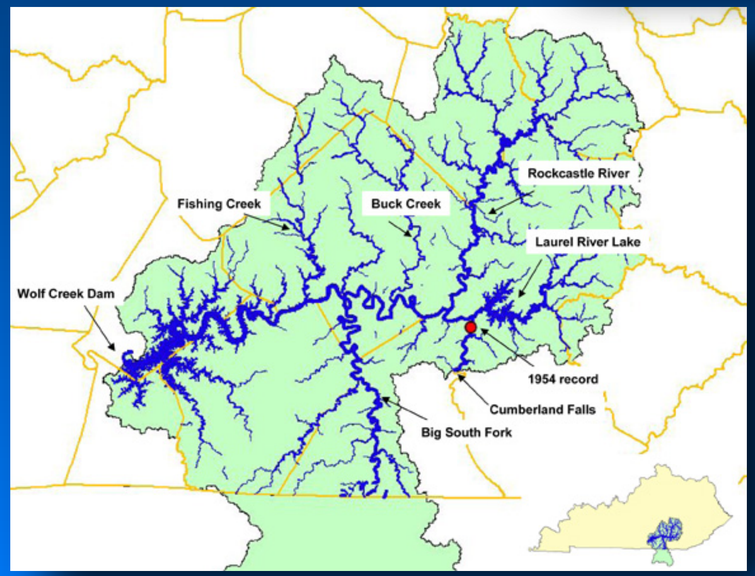
$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

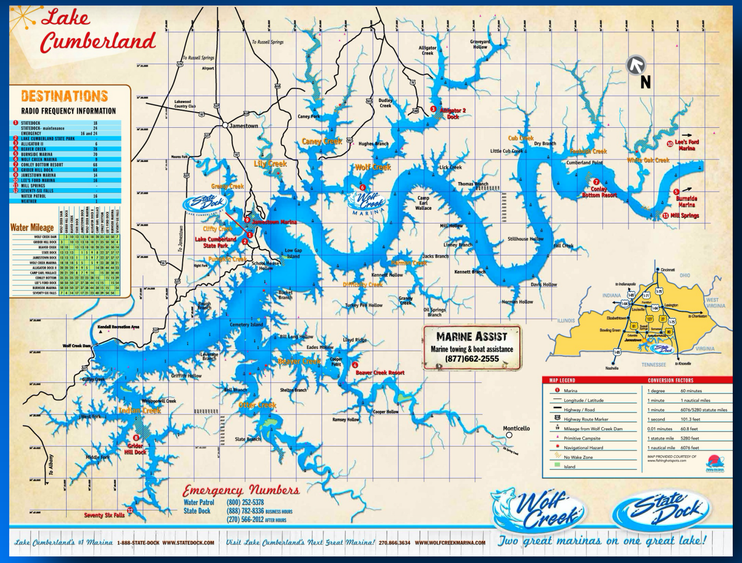
Multiple Least-Squares Linear Regression

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

Multiple Regression: Example







Multiple Regression: Example

Kentucky Stream Basin Dataset

Basin No.	Basin Magnitude	Outlet Elevation	Relief	Area	Stream Length	Drainage Density	Shape
1	14	720	570	7	154	2200	61
2	6	670	610	3	80	2667	62
3	5	860	550	11	84	763	62
4	7	870	610	11	122	1110	63
5	11	730	570	14	185	1321	52
6	14	690	590	12	200	1667	50
7	12	880	640	11	170	1545	41
8	18	760	690	28	340	1215	57
9	6	820	600	5	100	2000	41
10	5	720	480	3	80	2667	60
11	17	670	670	19	290	1526	51
12	5	660	600	5	90	1800	53

No. of Basins: 50

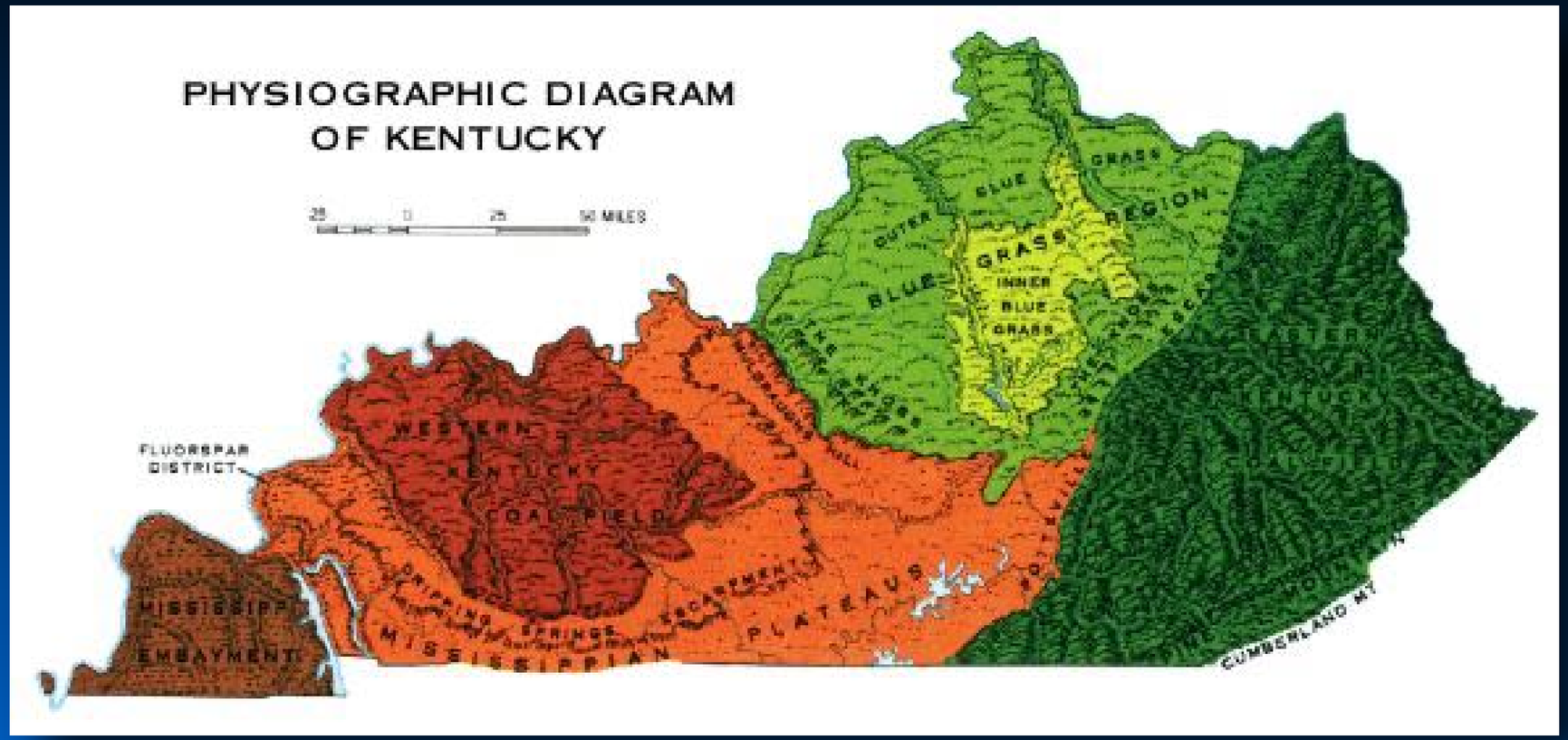
Dependent Variable:

- Basin Magnitude

Independent Variables:

- Outlet Elevation
- Relief
- Area
- Total Stream Length
- Drainage Density
- Basin Shape

Multiple Regression: Example



We would like to obtain an estimate of stream basin magnitude conditioned on the values of a series of descriptive variables, including: outlet elevation, relief, area, total stream length, drainage density and basin shape.

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Basin No.	Basin Magnitude	Area	Stream Length
1	14	7	154
2	6	3	80
3	5	11	84
4	7	11	122
5	11	14	185
6	14	12	200
7	12	11	170
8	18	28	340
9	6	5	100
10	5	3	80
11	17	19	290
12	5	5	90
13	22	18	260
14	7	17	111
15	15	15	184
16	17	21	227
17	5	4	60
18	18	20	259
19	14	9	62
20	21	6	95

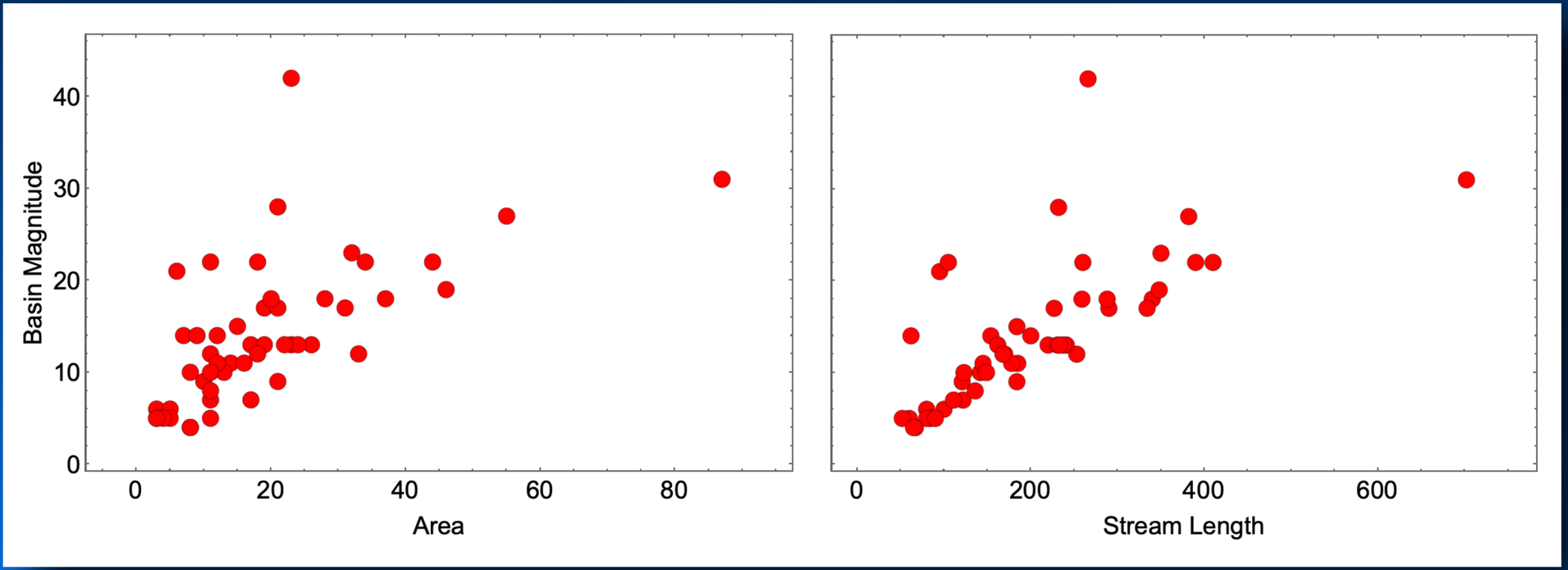
Basin No.	Basin Magnitude	Area	Stream Length
21	22	11	105
22	23	32	350
23	28	21	232
24	42	23	266
25	22	44	390
26	10	13	142
27	11	12	145
28	12	33	253
29	13	23	241
30	31	87	702
31	18	37	288
32	13	17	162
33	4	8	67
34	5	3	52
35	9	10	121
36	13	26	220
37	10	8	123
38	13	24	238
39	13	19	231
40	11	16	178

Basin No.	Basin Magnitude	Area	Stream Length
41	12	18	168
42	4	8	65
43	17	31	334
44	9	21	184
45	8	11	136
46	13	22	233
47	22	34	410
48	10	11	149
49	19	46	348
50	27	55	382

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Bivariate Plots of the Independent Variable Against the Two Dependent Variables

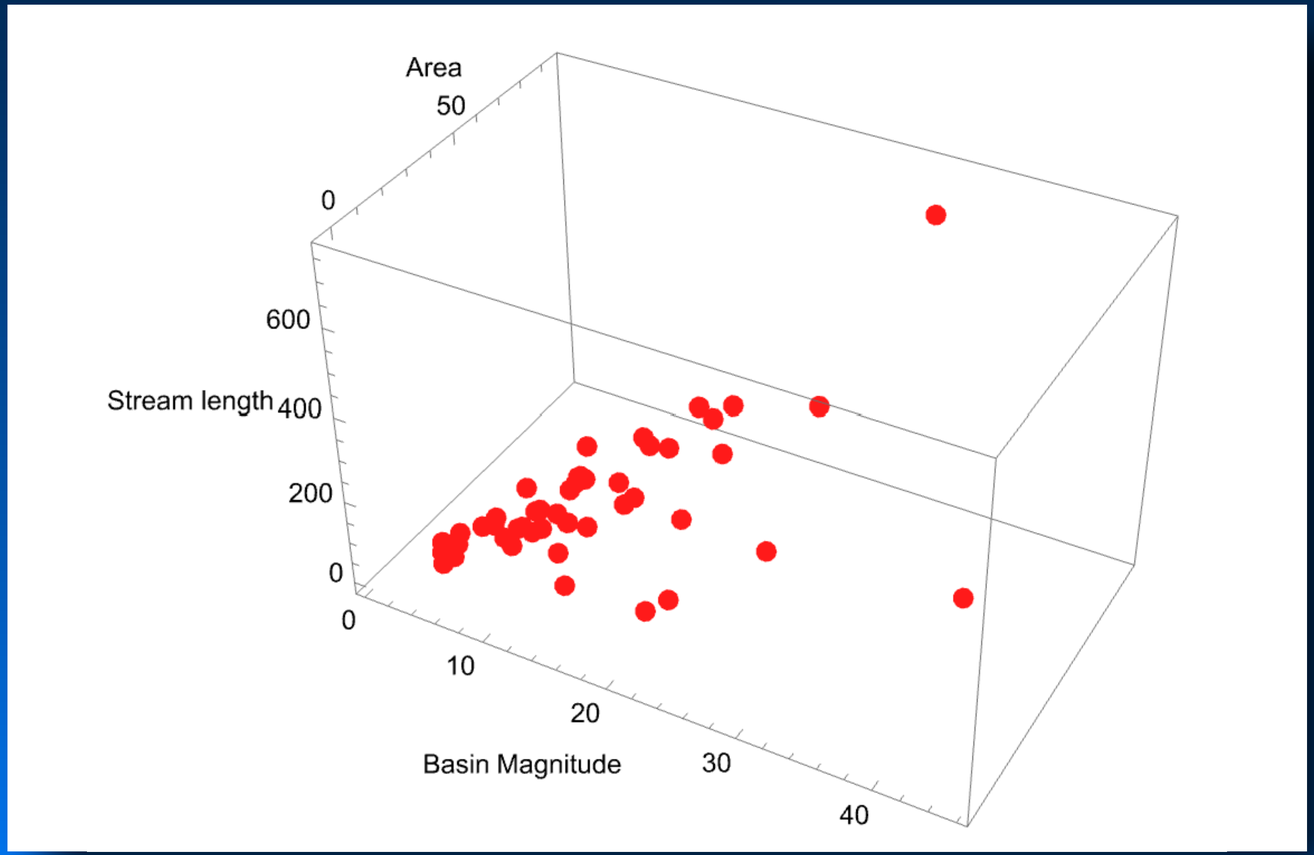


But how accurate are these representations of the data?

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

3D Plots of the Independent Variable Against the Two Dependent Variables



Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

$$y_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \epsilon_i$$

$$\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1_i} y_i \\ \sum_{i=1}^{n} x_{2_i} y_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} n & \sum_{i=1}^{n} x_{1_i} & \sum_{i=1}^{n} x_{2_i} \\ \sum_{i=1}^{n} x_{1_i} & \sum_{i=1}^{n} x_{1_i}^2 & \sum_{i=1}^{n} x_{1_i} x_{2_i} \\ \sum_{i=1}^{n} x_{2_i} & \sum_{i=1}^{n} x_{1_i} x_{2_i} & \sum_{i=1}^{n} x_{2_i}^2 \end{pmatrix}$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{1_i} & \sum_{i=1}^{n} x_{2_i} \\ \sum_{i=1}^{n} x_{1_i} & \sum_{i=1}^{n} x_{1_i}^2 & \sum_{i=1}^{n} x_{1_i} x_{2_i} \\ \sum_{i=1}^{n} x_{2_i} & \sum_{i=1}^{n} x_{1_i} x_{2_i} & \sum_{i=1}^{n} x_{2_i}^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1_i} y_i \\ \sum_{i=1}^{n} x_{2_i} y_i \end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

$$y_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \epsilon_i$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 50 & 963 & 10068 \\ 963 & 29905 & 278309 \\ 10068 & 278309 & 2738740 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 703 \\ 17214 \\ 173995 \end{pmatrix}$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0.0939 & 0.0034 & -0.0007 \\ 0.0034 & 0.0007 & -0.0001 \\ -0.0007 & -0.0001 & 0.0000 \end{pmatrix} \cdot \begin{pmatrix} 703 \\ 17214 \\ 173995 \end{pmatrix}$$

$$\begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 4.2803 \\ -0.1294 \\ 0.0609 \end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Alternative Covariance/Correlation Approach

$$y_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \epsilon_i$$

$$\begin{pmatrix} Cov(y, X_1) \\ Cov(y, X_2) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} Cov(x_1, x_1) & Cov(x_1, x_2) \\ Cov(x_2, X_1) & Cov(x_2, x_2) \end{pmatrix}$$

What is of interest in this problem are the covariation relations among the three variables. The covariance index can be used to summarize these relations much more efficiently than the sum of squares and cross products.

Covariance

Covariance = the joint variation of two variables about their common mean.

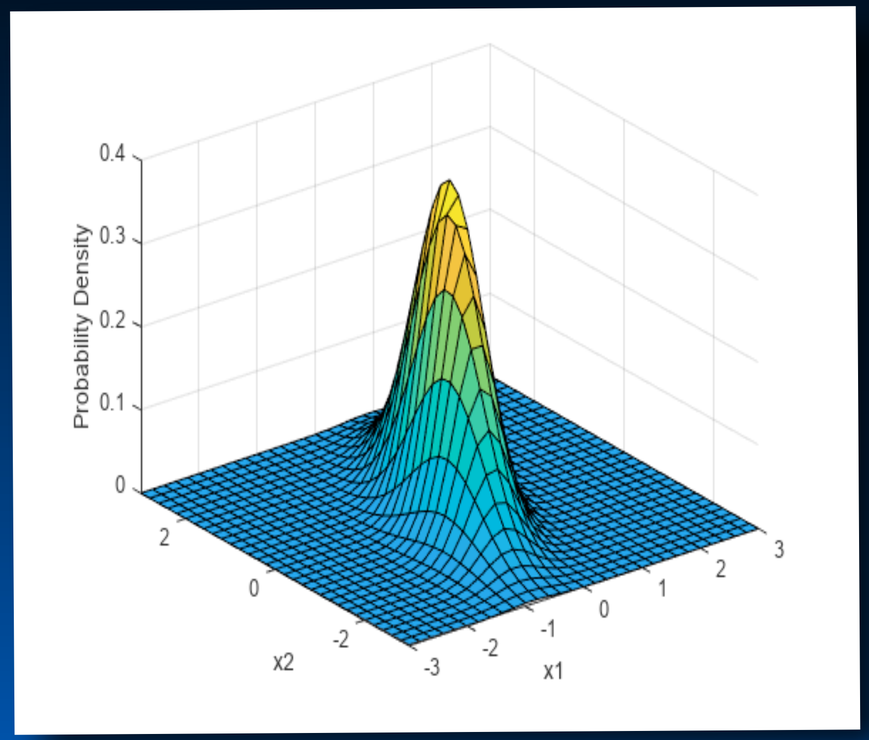
$$COV = \frac{1}{n-1} Z^T Z$$

$$Z = (I - \frac{1}{n} 11^T) X$$

Where: I = the identify matrix:

n x n matrix of 0s whose trace is filled with 1s.

1 = column vector of 1s with length n



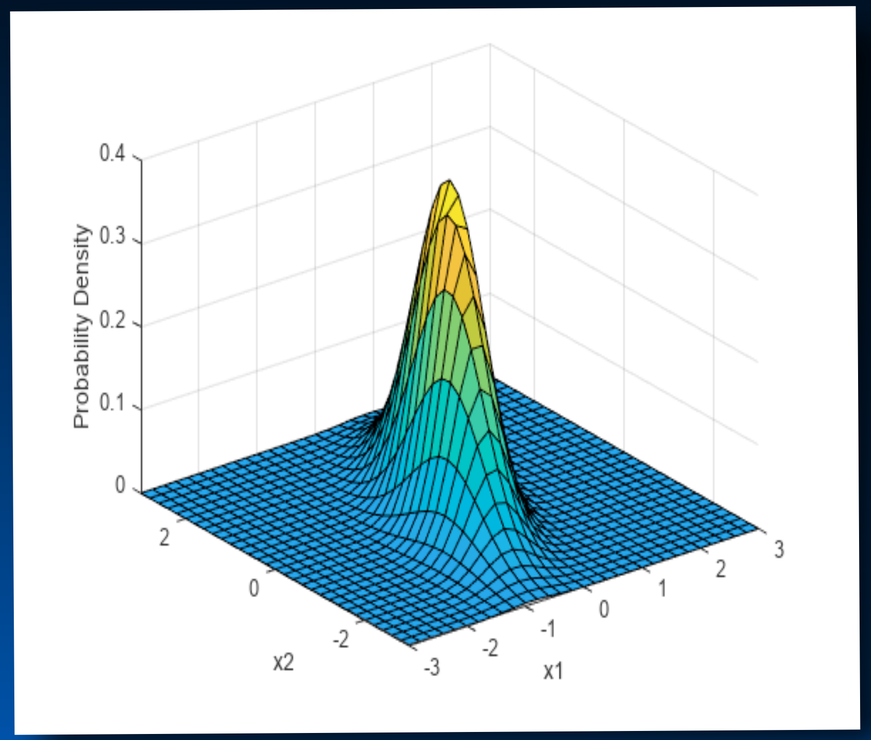
Standard Form

$$COV_{jk} = \frac{\sum_{i=1}^{n} x_{ij} x_{ik} - (\sum_{i=1}^{n} x_{ij} \cdot \sum_{i=1}^{n} x_{ik}}{n(n-1)}$$

Covariance

Because of its unbounded nature and manner in which it reflects the variances of the variables, the covariance matrix can be difficult to interpret. Therefore, it's usually employed to quantify linear relations …

- … among variables of the same type,
- … among variables of different types, but whose ranges are similar,
- … in situations where it's important to retain information about the different variables' variances.



In situations other than these it's usually better to standardize the variables prior to calculating their covariance, an operation that produces an index termed the correlation.

Correlation

Correlation = the measure of the joint variability of two standardized random variables

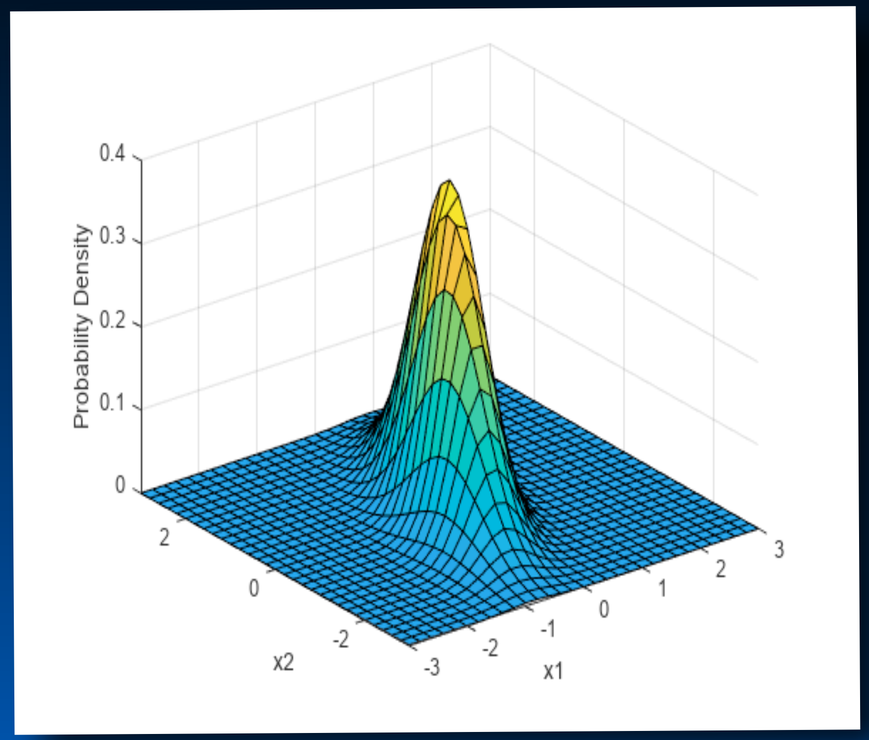
$$r = \frac{1}{n} Z^T \cdot Z$$

$$Z = (I - \frac{1}{n} 11^T) X$$

Where: I = the identify matrix:

n x n matrix of 0s whose trace is filled with 1s.

1 = column vector of 1s with length n



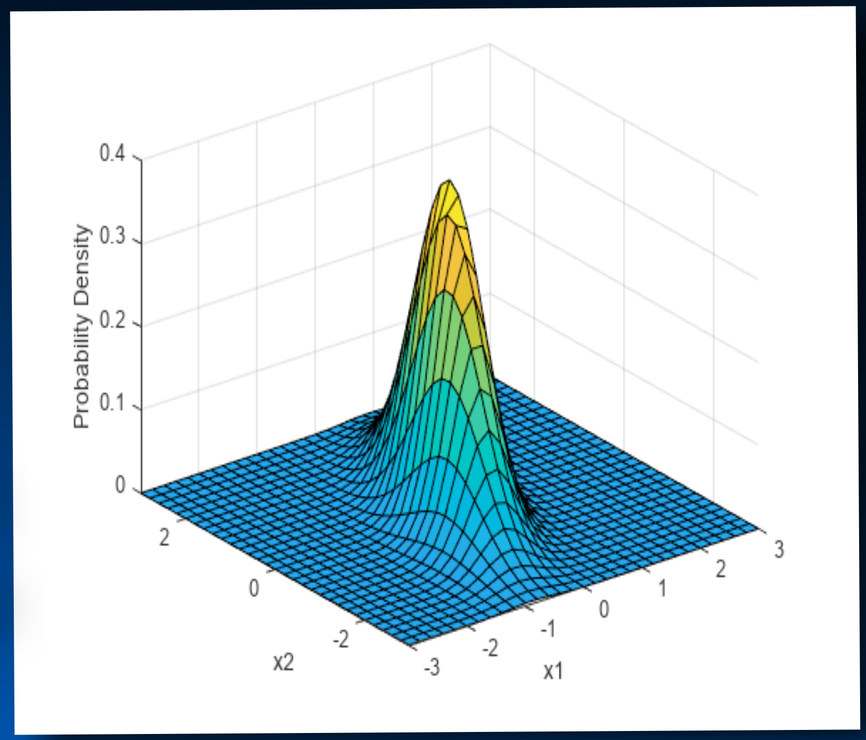
Standard Form

$$r_{jk} = \frac{\sum_{i=1}^{n} x_{ij} \cdot x_{jk} - (\frac{(\sum_{i=1}^{n} x_{ij} \cdot \sum_{i=1}^{n} x_{ik})}{n})}{\sqrt{(\sum_{i=1}^{n} x_{ij}^2 - \frac{(\sum_{i=1}^{n} x_{ij})^2}{n}) \cdot (\sum_{i=1}^{n} x_{ik}^2 - (\frac{(\sum_{i=1}^{n} x_{ik})^2}{n})}}$$

Correlation

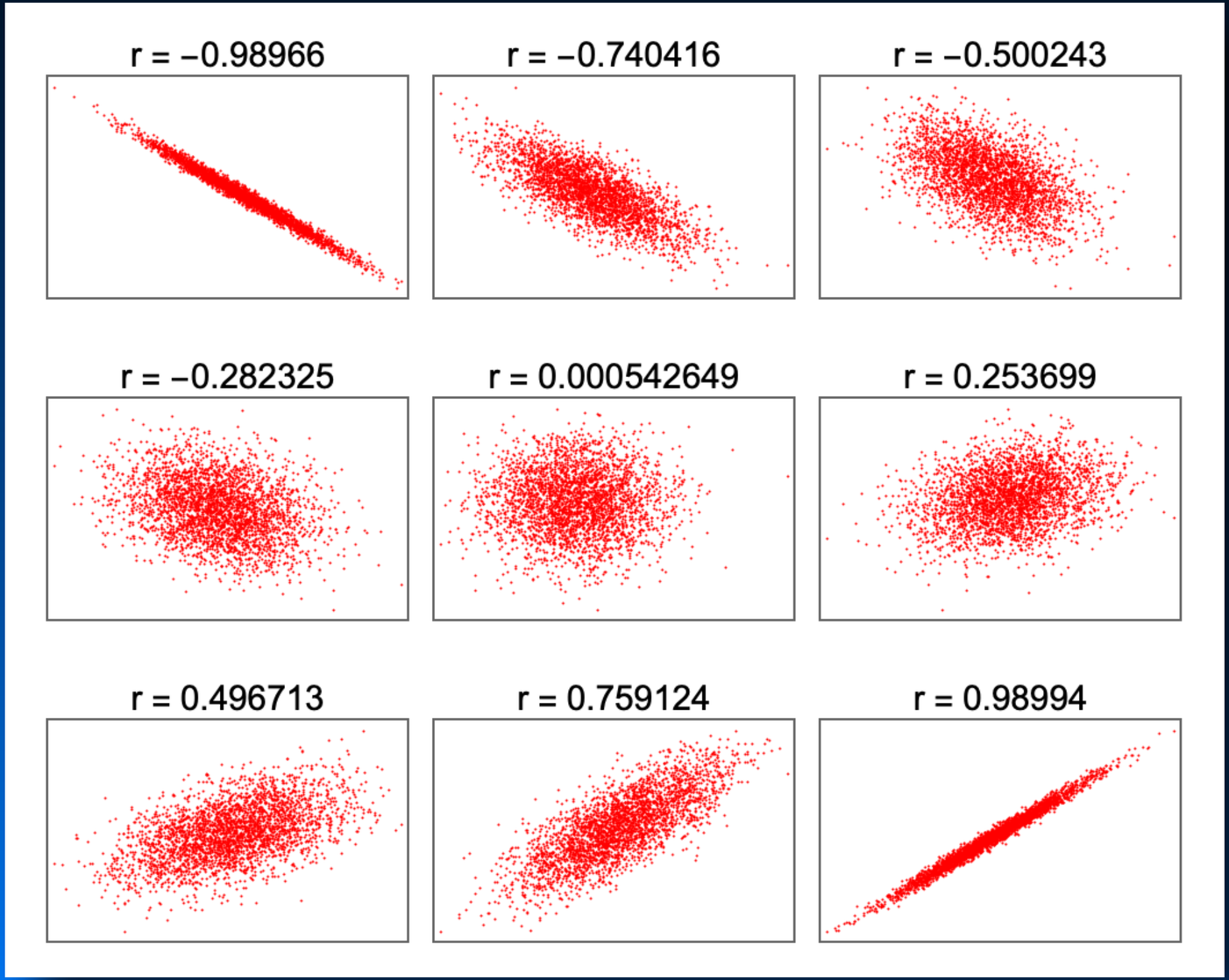
The correlation is useful because it's a bounded statistic, and to quantifies covariance in a relative sense. It's also usually employed to quantify linear relations …

- … among variables of different types type,
- … among variables whose ranges are different,
- … in situations where it's not important to retain information about the different variables' variances.



When using the correlation it must be remembered that information pertaining to differences in the magnitudes of variation among the different variables has been discarded. If all variables are of the same type and have similar observed ranges the reason(s) for standardizing the variables must be clear.

Covariance


r = −0.98966
r = −0.740416
r = −0.500243
r = −0.282325
r = 0.000542649
r = 0.253699
r = 0.496713
r = 0.759124
r = 0.98994

Covariance

Kentucky Stream Basin Dataset (Reduced)

Covariance Matrix

	Basin Magnitude	Area	Stream Length
Basin Magnitude	59.89	74.98	662.02
Area	74.98	231.79	1,722.44
Stream Length	662.02	1,722.44	14,519.30

Note, this is a square, symmetric matrix. The diagonal holds the variance values for the individual variables. The off-diagonal values express the geometric relation each variable with respect to the other variables. The numbers are large because we are working with the raw (non-standardized) values.

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

$$\begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} r_{x_1,x_1} & r_{x_1,x_2} \\ r_{x_2,x_1} & r_{x_2,x_2} \end{pmatrix}$$

	Basin Magnitude	Area	Stream Length
Basin Magnitude	1.00	0.64	0.71
Area	0.64	1.00	0.94
Stream Length	0.71	0.94	1.00

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} r_{x_1,x_1} & r_{x_1,x_2} \\ r_{x_2,x_1} & r_{x_2,x_2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

$$\begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} r_{x_1,x_1} & r_{x_1,x_2} \\ r_{x_2,x_1} & r_{x_2,x_2} \end{pmatrix}$$

	Basin Magnitude	Area	Stream Length
Basin Magnitude	1.00	0.64	0.71
Area	0.64	1.00	0.94
Stream Length	0.71	0.94	1.00

$$\begin{pmatrix} 0.64 \\ 0.71 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} 1.00 & 0.94 \\ 0.94 & 1.00 \end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

	Basin Magnitude	Area	Stream Length
Basin Magnitude	1.00	0.64	0.71
Area	0.64	1.00	0.94
Stream Length	0.71	0.94	1.00

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}1.00 & 0.94\\ 0.94 & 1.00\end{pmatrix}^{-1} \cdot \begin{pmatrix}0.64\\ 0.71\end{pmatrix}$$

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}8.5910 & -8.0756\\ -8.0756 & 8.5910\end{pmatrix} \cdot \begin{pmatrix}0.64\\ 0.71\end{pmatrix}$$

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}-0.2354\\ 0.9313\end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

	Basin Magnitude	Area	Stream Length
Basin Magnitude	1.00	0.64	0.71
Area	0.64	1.00	0.94
Stream Length	0.71	0.94	1.00

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}_{Stand.} = \begin{pmatrix} -0.2354 \\ 0.9313 \end{pmatrix}$$

These are the partial regression slopes in units of standard deviations (since they were calculated from the correlation matrix). This means they represent the rates of change in the dependent variable in each of the independent variables in isolation. They are informative, but cannot be used to reconstruct the regression plane.

Multiple Regression: Example

To complete the calculation we must convert these standardized partial regression slopes into full partial regression slopes and calculate the value of the dependent variable axis intercept.

	Basin Magnitude	Area	Stream Length
Basin Magnitude	1.00	0.64	0.71
Area	0.64	1.00	0.94
Stream Length	0.71	0.94	1.00

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{Stand.} = \begin{pmatrix}-0.2354\\ 0.9313\end{pmatrix} \qquad \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{Full} = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{Standardized} \cdot \begin{pmatrix}s_y/s_{x1}\\ s_y/s_{x2}\end{pmatrix}$$

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{Full} = \begin{pmatrix}-0.2354\\ 0.9313\end{pmatrix} \cdot \begin{pmatrix}7.74/15.22\\ 7.74/120.50\end{pmatrix} = \begin{pmatrix}-0.1197\\ 0.05981\end{pmatrix}$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

To estimate the independent variable intercept, solve for that value using the mean values of all variables.

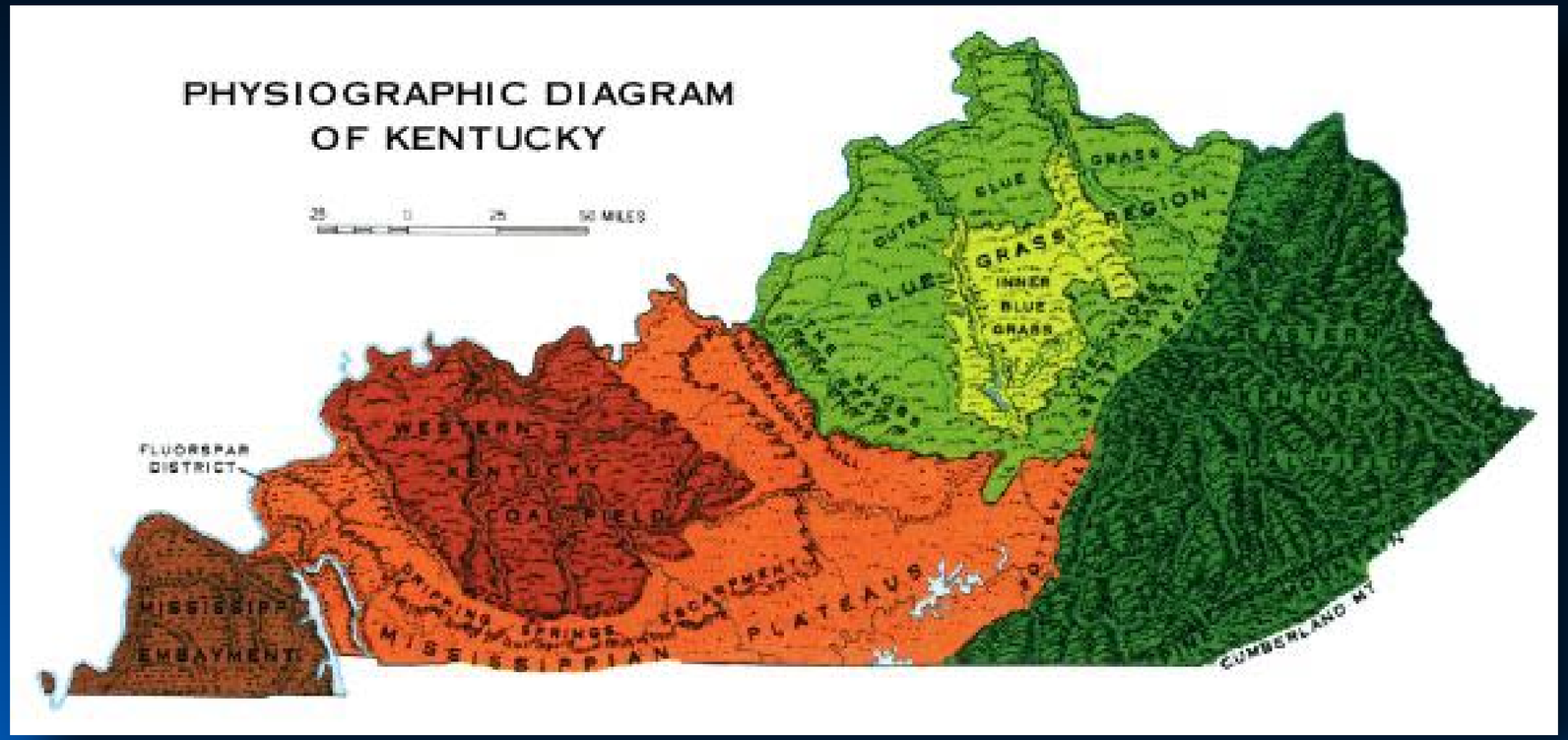
$$\alpha = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

$$\alpha = 14.06 - (-0.1197 \cdot 19.26) - (0.0598 \cdot 598.10)$$

$$\alpha = 14.06 + 2.3054 - 12.0433$$

$$\alpha = 4.32208$$

Multiple Regression: Example



$$y_i = 4.28 - 0.13x_{1_i} + 0.06_{x_{2_i}}$$

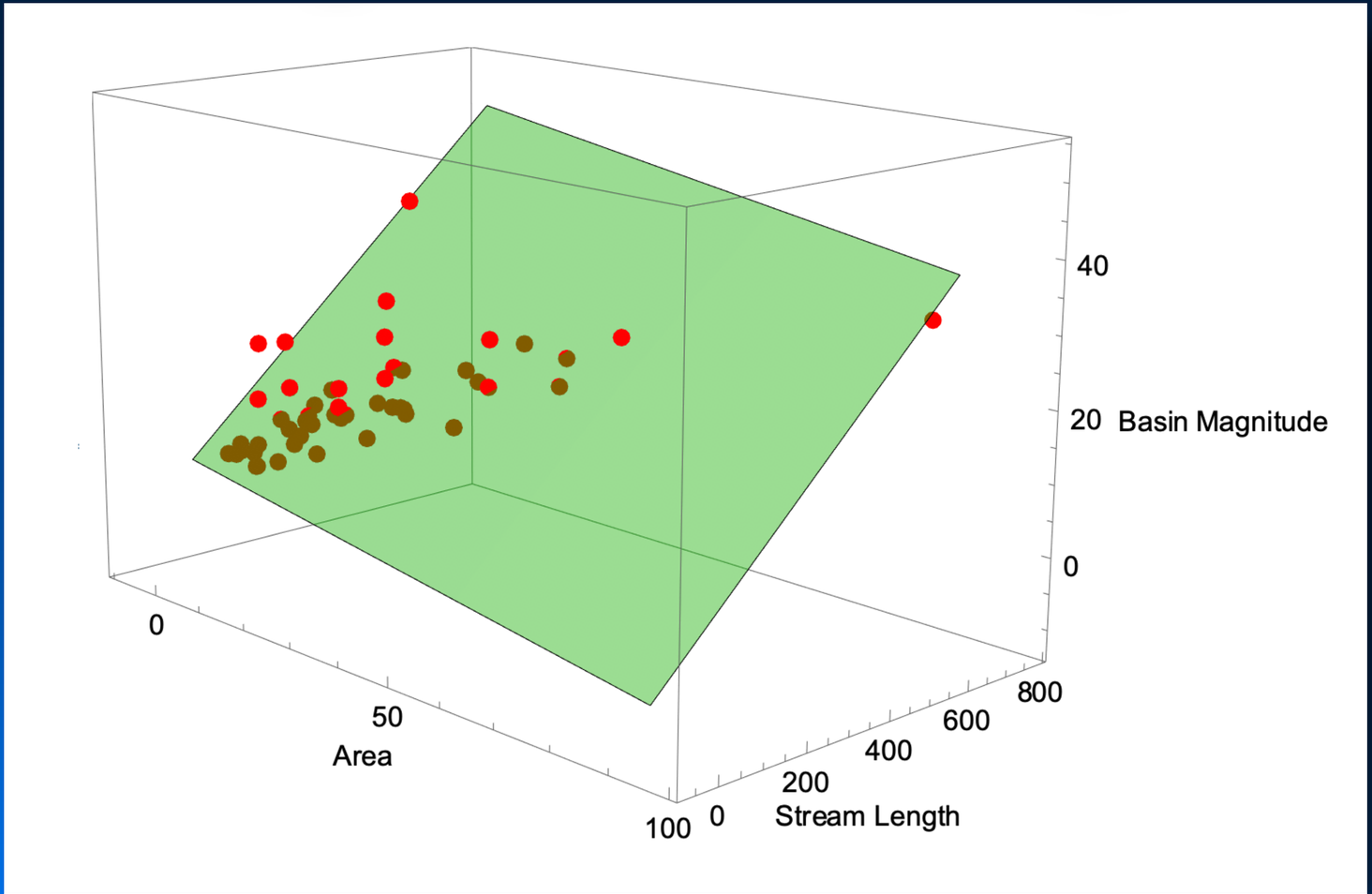
Where: y = basin magnitude;

x_1 = basin area;

x_3 = stream length.

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)



Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Coefficient of Determination (R^2)

$$R^2_{\gamma\cdot 1,\gamma\cdot 2\ldots\gamma\cdot k} = (r_{\gamma\cdot 1}\cdot\beta_1) + (r_{\gamma\cdot 2}\cdot\beta_2) + \ldots(r_{\gamma\cdot k1}\cdot\beta_k)$$

To estimate the coefficient of determination (R^2) calculate the sum of products of the correlations of individual independent variables with the dependent variable and the standardized regression coefficients.

$$R^2 = (0.64\cdot -0.25) + (0.95\cdot 0.71)$$

$$R^2 = -0.16 + 0.67 = 0.52$$

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Coefficient of Determination (R^2)

Source of Variation	Sum of Squares	Dof	Mean Squares	F-Test
Due to Model	$SS_A = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$m - 1$	$MS_A = \frac{SS_A}{m-1}$	$\frac{MS_A}{MS_E}$
Unexplained (Error)	$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$N - m$	$MS_E = \frac{SS_E}{N-m}$	
Total Variation	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$N - 1$		

Where: $\hat{y}$ = estimated value of y;
$\bar{y}$ = arithmetic mean of y;
m = number of variables;
n = number of samples.

Multiple Regression: Example

Kentucky Stream Basin Dataset (Reduced)

Coefficient of Determination (R^2)

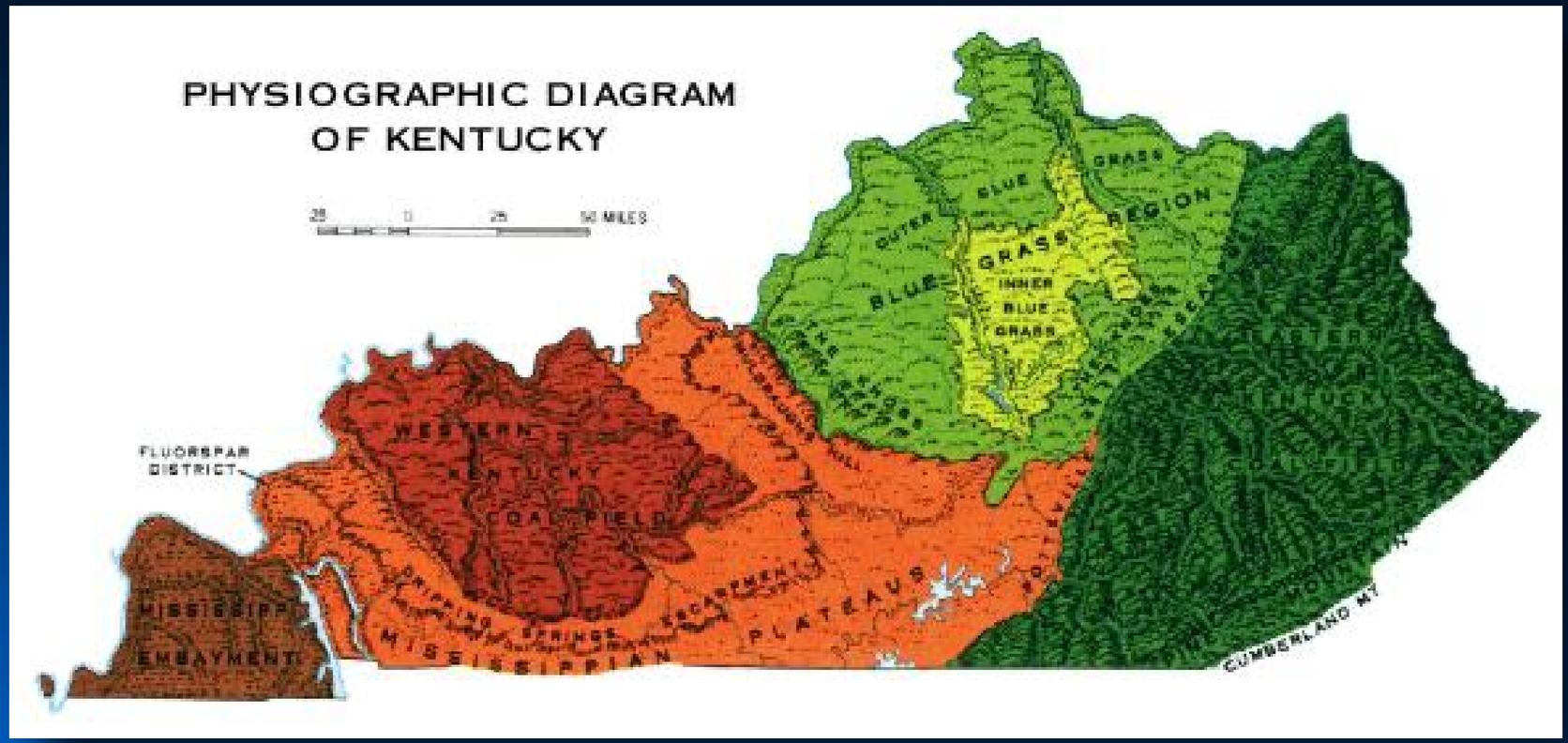
Source of Variation	Sum of Squares	Dof	Mean Squares	F-Test
Due to Model	1438.57	2	719.29	23.54
Unexplained (Error)	1436.02	47	30.55	
Total Variation	2934.82	49		

$$F_{0.05,dof:2,47} = 3.96$$

$$p_{F=23.54,dof:2,47} = 8.26x10^{-6}$$

Very, very unlikely to be the result of random sampling effects.

Multiple Regression: Example



We would like to obtain an estimate of stream basin magnitude conditioned on the values of a series of descriptive variables, including: outlet elevation, relief, area, total stream length, drainage density and basin shape.

Multiple Regression: Example

Kentucky Stream Basin Dataset

Basin No.	Basin Magnitude	Outlet Elevation	Relief	Area	Stream Length	Drainage Density	Shape
1	14	720	570	7	154	2200	61
2	6	670	610	3	80	2667	62
3	5	860	550	11	84	763	62
4	7	870	610	11	122	1110	63
5	11	730	570	14	185	1321	52
6	14	690	590	12	200	1667	50
7	12	880	640	11	170	1545	41
8	18	760	690	28	340	1215	57
9	6	820	600	5	100	2000	41
10	5	720	480	3	80	2667	60
11	17	670	670	19	290	1526	51
12	5	660	600	5	90	1800	53

No. of Basins: 50

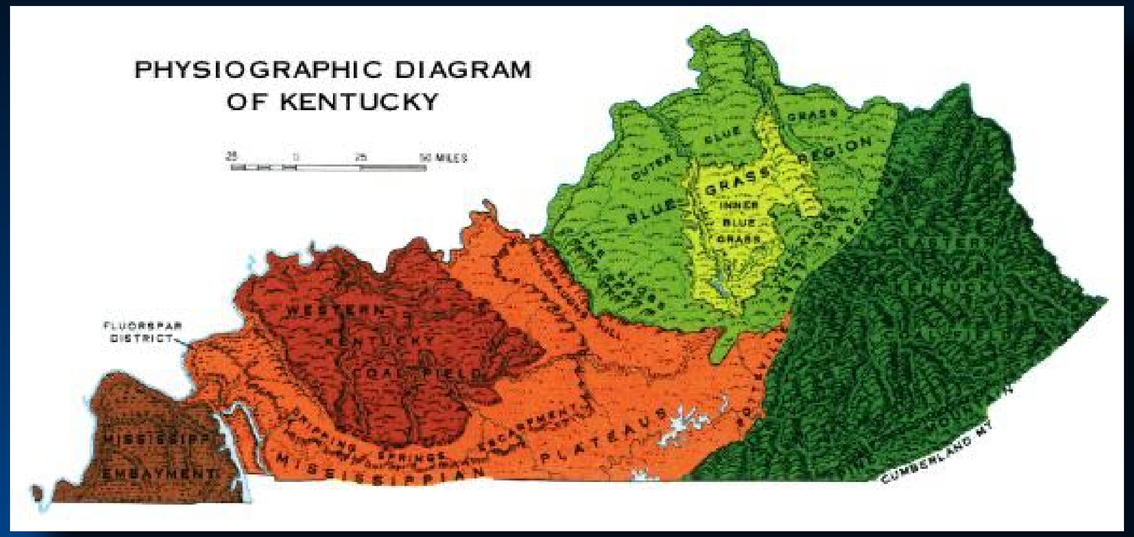
Dependent Variable:

- Basin Magnitude

Independent Variables:

- Outlet Elevation
- Relief
- Area
- Total Stream Length
- Drainage Density
- Basin Shape

Multiple Regression: Example



$$y_i = 2.244 + 0.005x_{1_i} + 0.023x_{2_i} - 2.33x_{3_i} + 0.063x_{4_i} - 0.002x_{5_i} - 0.012x_{6_i}$$

$$R^2 = 0.61$$

y = basin magnitude

Where: x_1 = outlet elevation

x_2 = relief

x_3 = area

x_4 = stream length

x_5 = drainage density

x_6 = basin shape

Multiple Regression: Example

Basin Magnitude versus Six Descriptive Variables

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Test
Due to Model	1840.25	5	368.05	14.23
Unexplained (Error)	1138.14	44	25.87	
Total Variation	2934.82	49		

$$F_{0.05,dof:5,44} = 2.83$$

$$p_{F=14.23,dof:5,44} = 2.74x10^{-6}$$

Very, very unlikely to be the result of random sampling effects.

Multiple Regression

Prof. Norman MacLeod

School of Earth Sciences & Engineering, Nanjing University