Descriptive Statistics – Associations

Kate Davis

February 13, 2015

Introduction to Associations

We now consider the statistical associations between data. A data set can contain multiple data points per observation, and understaning how those data are associated within the observation is a key goal of descriptive and inferential statistical analysis. Examples of data sets are observations as patients, with data points as vital signs, or observations as states or counties and data points as crime statistics.

Some associations that are often examined are statistics including covariance, correlations, and contingencies, and visualizations such as scatterplots and mosaic plots.

US Arrest Statistics by Crime and State

This data set contains statistics, in arrests per 100,000 residents for assault (X) and murder (Y) in each of the 50 states in 1973, along with the percent of population living in urban areas (Z).

	Assault	Murder	UrbanPop
Min.	45.0	0.80	32.0
1st Qu.	109.0	4.08	54.5
Median	159.0	7.25	66.0
Mean	171.0	7.79	65.5
3rd Qu.	249.0	11.30	77.8
Max.	337.0	17.40	91.0
Sum Sq Deviation	340316.0	929.55	10266.5
Variance	6806.3	18.59	205.3
Standard Deviation	82.5	4.31	14.3

Table 1: Summary Statistics for US Arrests in 1973

Scatterplots

A **Scatterplot** of any two of the three variables shows the relationship on the same observation, in this case, State.

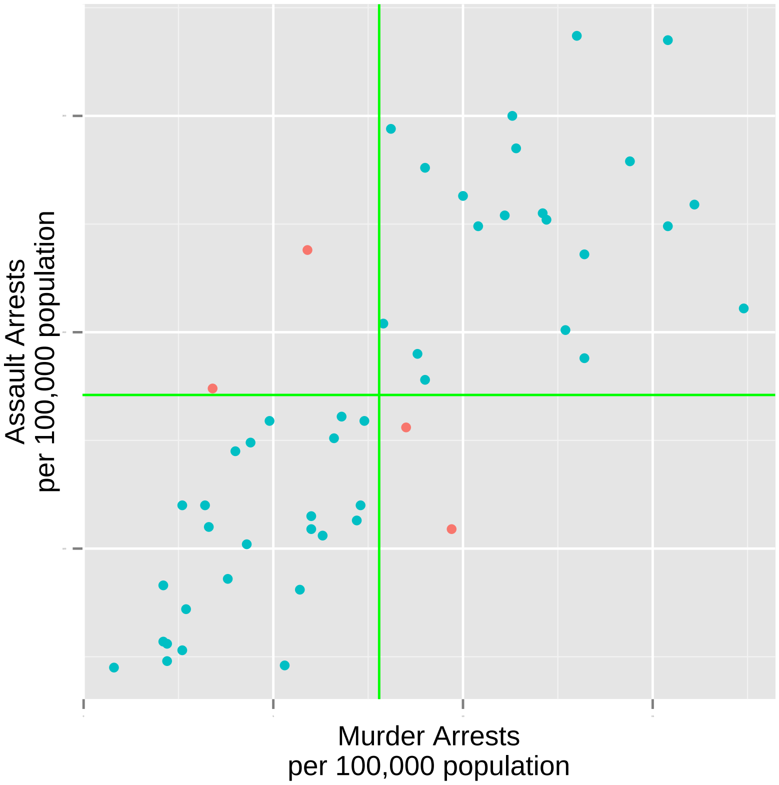

Figure 3: A scatterplot of Assault vs. Murder Arrests shows a possible linear relationship

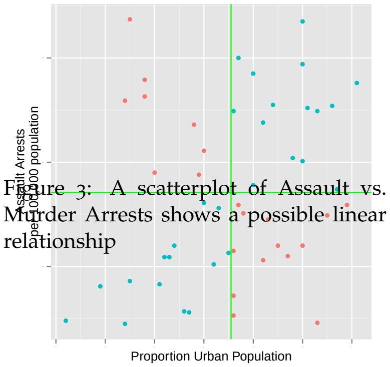

Figure 1: A scatterplot of Assault Arrests vs. Proportion Urban Population does not show a relationship

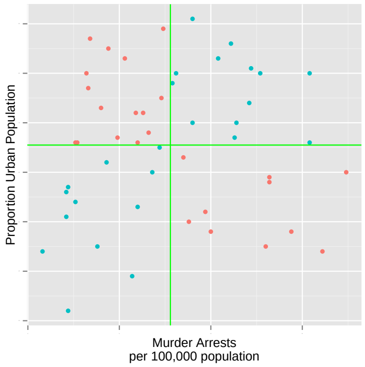

Figure 2: A scatterplot of Murder Arrests vs. Proportion Urban Population does not show a relationship

Covariance

We would like to quantify the linear relationship between each of the variables. The **Covariance**, a measure of strength of the association between any two variables X and Y, denoted $Cov(X,Y)$ is calculated by first multiplying the deviations from their means, $Dev_{\bar{x}}$ and $Dev_{\bar{y}}$, then summing over all observations and dividing by N, the number of observations. This is very similar to the population variance calculation, and the variance can be thought of as the covariance of a variable with itself ie. $Var(X) = Cov(X,X)$.

$$Cov(X,Y) = \frac{\Sigma_{i=1}^{N} Dev_{\bar{x}} Dev_{\bar{y}}}{N}$$

The Covariance of Assault Arrests with Murder Arrests is 285.2416 "Assault Arrests x Murder Arrests". These units make very little sense.

We cannot compare covariances among variables in a data set if the units are different.

Table 2: Covariances, Part 1

Observation	Assault	Murder	UrbanPop	$Dev_{x\bar{y}}$	$Dev_{x\bar{z}}$	$Dev_{x\bar{z}}$
Alabama	236	13.20	58	351.65	-487.50	-40.57
Alaska	263	10.00	48	203.32	-1610.00	-38.67
Arizona	294	8.10	80	38.13	1783.50	4.49
Arkansas	190	8.80	50	19.19	-294.50	-15.66
California	276	9.00	91	127.05	2677.50	30.86
Colorado	204	7.90	78	3.63	412.50	1.38
Connecticut	110	3.30	77	273.89	-701.50	-51.64
Delaware	238	5.90	72	-126.63	435.50	-12.28
Florida	335	15.40	80	1248.04	2378.00	110.34
Georgia	211	17.40	60	384.40	-220.00	-52.85
Hawaii	46	5.30	83	311.25	-2187.50	-43.58
Idaho	120	2.60	54	264.69	586.50	59.68
Illinois	249	10.40	83	203.58	1365.00	45.68
Indiana	113	7.20	65	34.22	29.00	0.29
Iowa	56	2.20	57	642.85	977.50	47.52
Kansas	115	6.00	66	100.24	-28.00	-0.90
Kentucky	109	9.70	52	-118.42	837.00	-25.78
Louisiana	249	15.40	66	593.58	39.00	3.81
Maine	83	2.10	51	500.72	1276.00	82.50
Maryland	300	11.30	67	452.79	193.50	5.27
Massachusetts	149	4.40	85	74.58	-429.00	-66.10
Michigan	255	12.10	74	362.04	714.00	36.63
Minnesota	72	2.70	66	503.91	-49.50	-2.54
Mississippi	259	16.10	44	731.28	-1892.00	-178.67
Missouri	178	9.00	70	8.47	31.50	5.45
Montana	109	6.00	53	110.98	775.00	22.38
Nebraska	102	4.30	62	240.81	241.50	12.21
Nevada	252	12.20	81	357.21	1255.50	68.35
New Hampshire	57	2.10	56	648.66	1083.00	54.05
New Jersey	159	7.40	89	4.68	-282.00	-9.16

Table 3: Covariances, Part 2

Observation	Assault	Murder	UrbanPop	$Dev_{\bar{x}\bar{y}}$	$Dev_{\bar{x}\bar{z}}$	$Dev_{\bar{x}\bar{z}}.1$
New Mexico	285	11.40	70	411.54	513.00	16.25
New York	254	11.10	86	274.73	1701.50	67.85
North Carolina	337	13.00	45	864.86	-3403.00	-106.80
North Dakota	45	0.80	44	880.74	2709.00	150.28
Ohio	120	7.30	75	24.99	-484.50	-4.66
Oklahoma	151	6.60	68	23.80	-50.00	-2.98
Oregon	159	4.90	67	34.68	-18.00	-4.33
Pennsylvania	106	6.30	72	96.85	-422.50	-9.69
Rhode Island	174	3.40	87	-13.17	64.50	-94.39
South Carolina	279	14.40	48	713.88	-1890.00	-115.68
South Dakota	86	3.80	45	339.15	1742.50	81.80
Tennessee	188	13.20	59	91.97	-110.50	-35.16
Texas	201	12.70	80	147.30	435.00	71.19
Utah	120	3.20	80	234.09	-739.50	-66.55
Vermont	48	2.20	32	687.57	4120.50	187.26
Virginia	156	8.50	63	-10.65	37.50	-1.77
Washington	145	4.00	73	98.54	-195.00	-28.43
West Virginia	81	5.70	39	188.10	2385.00	55.38
Wisconsin	53	2.60	66	612.42	-59.00	-2.59
Wyoming	161	6.80	60	9.90	55.00	5.45
Total	8538.0	389.4	3277.0	14262.1	15301.0	214.9
Total/N	171.0	7.8	65.5	285.2	306.0	4.3
	$\bar{x}$	$\bar{y}$	$\bar{z}$	$Cov(X,Y)$	$Cov(X,Z)$	$Cov(Y,Z)$

Linear Correlation

A standardized Covariance is the **Linear Correlation**, calculated by dividing each Covariance by the Standard Deviations of each of the variables:

$$Corr(X,Y) = \frac{Cov(Y,X)}{(StdDev(X)StdDev(Y))}$$

The Correlation of Assault Arrests with Murder Arrests is 0.80187 with no units, so the correlations of multiple pairs of variables can be compared.

Correlations are always between -1 and 1, and are a quantification of the linear relationship between two variables. A correlation of zero means that there is linear relationship between two variables, although there may be a non-linear relationship. A correlation of 1 or -1 is indicates a perfect positive or negative linear relationship. $Corr(X,X) = 1$ always.

Correlation does not imply Causation! Even if two variables have a high or perfect correlation, there is not necessarily causation. Causation means X depends on Y or Y depends on X.

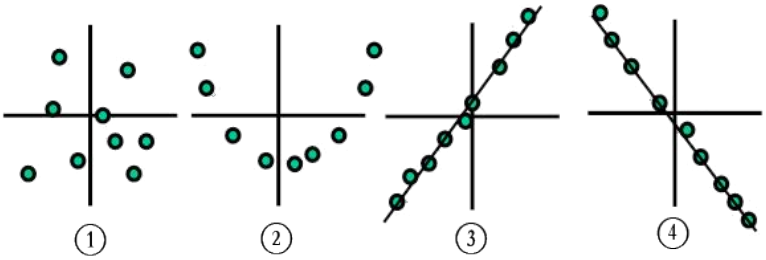

Figure 4: Examples of Correlation

① $Corr(X,Y) = 0$ No Correlation
② $Corr(X,Y) = 0$ No Linear Correlation
③ $Corr(X,Y) approaching + 1$ Positive Linear Correlation
④ $Corr(X,Y) approaching - 1$ Negative Linear Correlation

The Squared value of the correlation, 64.3% is a measure of the "shared variance" of two variable, and the complement 35.7% is the proportion of variance not explained by the association.

Covariance and Correlation Extras

The covariance relationship between multiple variables can be expressed in a variance-covariance matrix:

$$\begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{array}{c} X \\ Y \\ Z \end{array} & \begin{pmatrix} Var(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Var(Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Var(Z) \end{pmatrix} \end{array}$$

$$= \begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{array}{c} X \\ Y \\ Z \end{array} & \begin{pmatrix} 6806.32 & 285.2416 & 306.02 \\ 285.2416 & 18.59106 & 4.2984 \\ 306.02 & 4.2984 & 205.33 \end{pmatrix} \end{array}$$

The Variance-Covariance Matrix (also referred to as the VCV or simply the Covariance matrix) is a key part of multivariate statistics and methods, including:

- Principal Components Analysis (PCA)
- Factor Analysis
- Hierarchical Clustering

Similarly, the correlation matrix expresses all the correlations among variables.

$$\begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{array}{c} X \\ Y \\ Z \end{array} & \begin{pmatrix} Corr(X,X) & Corr(X,Y) & Corr(X,Z) \\ Corr(Y,X) & Corr(Y,Y) & Corr(Y,Z) \\ Corr(Z,X) & Corr(Z,Y) & Corr(Y,Y) \end{pmatrix} \end{array}$$

$$= \begin{array}{cc} & \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{array}{c} X \\ Y \\ Z \end{array} & \begin{pmatrix} 1 & 0.80187 & 0.25886 \\ 0.80187 & 1 & 0.06957 \\ 0.25886 & 0.06957 & 1 \end{pmatrix} \end{array}$$

The Correlation Matrix is a key part of multivariate statistics and methods, including:

- Canonical Correlation Analysis
- Portfolio Analysis and Optimization

When two variables have very different distributions, two non-parametric methods can assess the association on the ranks of the variables: ρ, the Spearman Rank Correlation, and τ, the Kendall Rank Correlation.

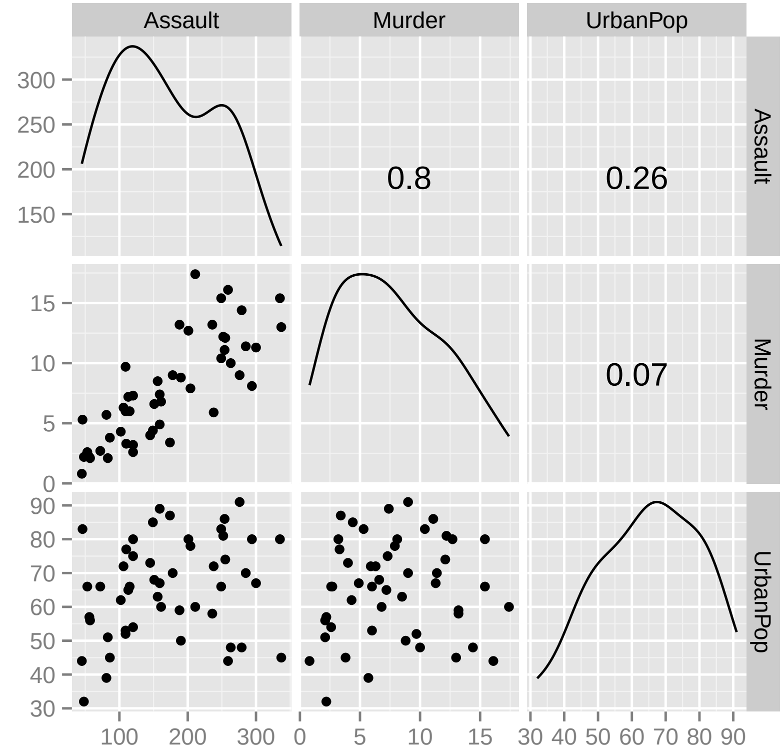

Figure 5: Scatterplot Matrix: Multiple scatterplots and correlations can be combined in one visualization

Simple Linear Regression

When a linear correlation exists between two variables, we can explore causation using a **Simple Linear Regression**, also called Ordinary Least Squares (OLS), regressing a dependent variable, denoted Y, on an independent variable, denoted X as a line with the form:

$$Y = \alpha + \hat{\beta} X$$

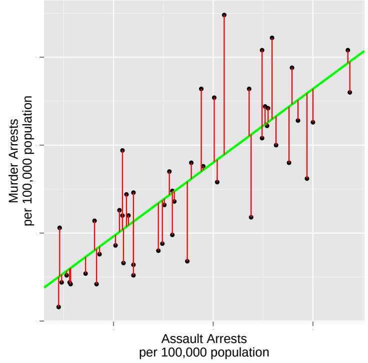

Figure 6: Green regression line with prediction error, as noted in red on the chart

This is very similar to the traditional algebra formula $y = mx + b$ with slope m and y-intercept b. In this case, the slope is $\hat{\beta}$.

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)} = Corr(X,Y)\frac{StdDev(Y)}{StdDev(X)}$$

Note that $\hat{\beta}$ is very close to the correlation value of $\frac{Cov(X,Y)}{StDev(X)StDev(Y)}$, but with $StDev(X)$ replacing $StDev(Y)$, to indicate the dependency of Y upon X.

For the US Arrests data set,

$$\hat{\beta} = \frac{285.2416}{6806.32} = 0.04191$$

The linear regression always goes through the point $(\bar{x}, \bar{y})$, so returning to algebra, any point plus the slope determines the line:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

The predicted value for any y_i is $\hat{y}_i$, and the prediction error is $\hat{\epsilon} = y_i - \hat{y}_i$.

Some properties of the Simple Linear Regression:

- $\Sigma_{i=1}^{N} \hat{\epsilon} = 0$
- $\Sigma_{i=1}^{N} x_i \hat{\epsilon} = 0$
- The predicted values $\hat{y}_i$ minimize the sum of the squared prediction errors, $\Sigma_{i=1}^{N} \hat{\epsilon}^2$, often referred to as Sum Squared Errors, or SSE.
- The regression equation is valid to predict $\hat{y}$ values in the range of X, that is, on the interval (min(X),max(X)), and any prediction will be in the range of (min(Y),max(Y))