
MODULE 8

BIVARIATE EDA - QUANTITATIVE

Objectives:

1. Describe bivariate data.
2. Distinguish between response and explanatory variables.
3. Construct scatterplots of bivariate quantitative data.
4. Describe bivariate relationships with interpretations from scatterplots.
5. Describe how the correlation coefficient is calculated.
6. Use the correlation coefficient to describe the strength (and association) of the relationship between two quantitative variables.

Contents

8.1	Response and Explanatory Variables	93
8.2	Scatterplots	94
8.3	Items to Describe	96
8.4	Correlation	99
8.5	Example Interpretations	103

BIVARIATE DATA OCCURS WHEN TWO variables have been measured on the same individuals. For example, you may measure (i) the height and weight of students in class, (ii) depth and area of a lake, (iii) gender and age of welfare recipients, or (iv) number of mice and biomass of legumes in fields. This module is focused on describing the bivariate relationship between two quantitative variables. Bivariate relationships between two categorical variables is described in Module 9.

△ **Bivariate:** Data where two variables have been measured on the same individuals.

Data on the *weight* (lbs) and highway miles per gallon (stored as *HMPG*) for 93 cars from the 1993 model year will be used as an example throughout this section.¹ Ultimately, the relationship between highway MPG and the weight of a car will be examined. These are bivariate data because measurements of both variables are recorded for each individual (i.e., a car). The following commands read the data from [93cars.csv](#) into R and lists the *HMPG* and *weight* values for the first and last three cars².

```
> cars93 <- read.csv("data/93cars.csv")
> headtail(cars93, which=c("HMPG", "Weight"))
  HMPG Weight
1    31  2705
2    25  3560
3    26  3375
91    25  2810
92    28  2985
93    28  3245
```

8.1 Response and Explanatory Variables

The **response variable** is the variable that one is interested in explaining something about (i.e., variability) or in making future predictions about. Synonyms for response are dependent and predicted. The **explanatory variable** is the variable that may help explain or allow one to predict the response variable. Synonyms for explanatory are independent and predictor.

△ **Response Variable:** The variable that we are interested in explaining or predicting. Synonyms are “dependent” or “predicted” variable.

△ **Explanatory Variable:** The variable that we think may explain or allow us to predict the response variable. Synonyms are “independent” or “predictor” variable.

Deciding which variable is the response variable often depends on the context of the situation (as defined by the research question). In the first example of bivariate data given in the introduction, the response variable may be weight if interest is in predicting weight from height or it may be height if interest is in predicting height from weight.³ The response and explanatory variables for the three situations in the introduction with quantitative variables are as follows (followed by context notes):

- R = weight, E = height [want to predict weight (hard to measure) from height (easy to measure)].
- R = area, E = depth [area is hard to measure, depth is easy].
- R = number of mice in a field, E = biomass of legumes in the field [hypothesized that higher biomass leads to more mice].

In the car data, the weight of the car may help explain the highway MPG of the car (e.g., a hypothesis might be that heavier cars get worse gas mileage). Thus, highway MPG is the response variable because it is of

¹Data are from Lock (1993).

²The vector in the second argument to `headtail()` is used to show only the two variables of interest.

³The latter is usually not the case, though.

primary interest and may depend on the weight of the car. Weight is the explanatory variable as it will be used to explain the highway MPG.

◇ Which variable is the response variable depends on the context of the problem or the researcher's needs (i.e., which variable is being explained or predicted).

8.2 Scatterplots

A scatterplot is a graph where each point simultaneously represents the values of both the quantitative response and quantitative explanatory variable. The value of the explanatory variable gives the x-coordinate and the value of the response variable gives the y-coordinate of the point plotted for an individual. For example, the first individual in the cars data is plotted at x (*Weight*) = 2705 and y (*HMPG*) = 31, whereas the second individual is at x = 3560 and y = 25. The scatterplot for all individuals in the data file is shown in Figure 8.1.

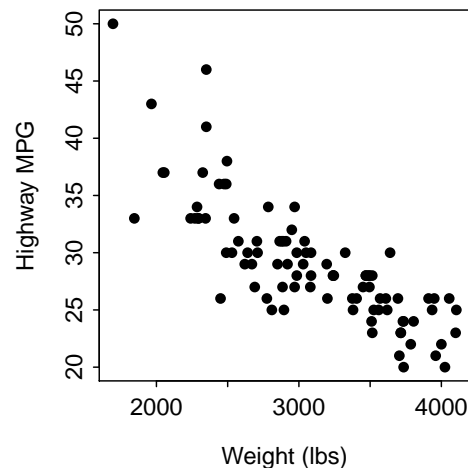


Figure 8.1. Scatterplot between the highway MPG and weight of cars manufactured in 1993.

◇ Both variables must be quantitative to construct a scatterplot.

◇ Response variables are plotted on the y-axis and explanatory variables are plotted on the x-axis.

8.2.1 Scatterplots in R

Scatterplots are constructed in R with `plot()`. This function requires a formula of the form $Y \sim X$, where Y and X are variables to be plotted on the y- and x-axes, as the first argument, and the corresponding dataframe name in `data=`.⁴ The x- and y-axis labels may be modified with `xlab=` and `ylab=`. The scatterplot of highway MPG versus car weight (Figure 8.2) was created with the code below.

⁴This function can also take the vector of x-axis data as its first argument followed by a vector of y-axis data as its second argument. The formula notation is preferred for ease of transferability to other functions.

```
> plot(HMPG~Weight,data=cars93,ylab="Highway MPG",xlab="Weight (lbs)")
```

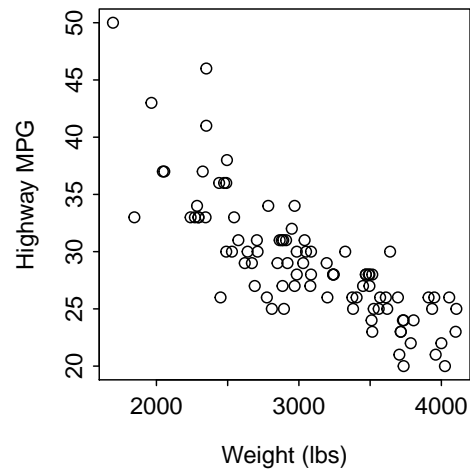


Figure 8.2. Scatterplot between the highway MPG and weight of cars manufactured in 1993 (using R default values)

The character plotted at each point can be changed with the `pch=` argument.⁵ This argument defaults to a value of 1, which is an open-circle. Numerical values used to represent other plotting characters are shown in Figure 8.3. For example, the scatterplot shown in (Figure 8.1) was created by including `pch=16` in `plot()`.

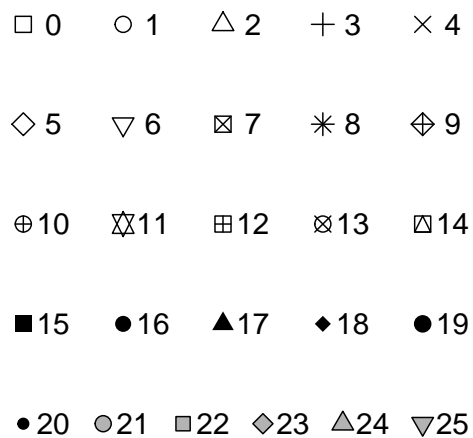


Figure 8.3. Plotting characters available in R and their numerical codes. Note that for values of 21-25 that `bg='gray70'` is used to provide the background color.

⁵This argument is short for “plotting character”.

8.3 Items to Describe

Four characteristics should be described when exploring bivariate data with a scatterplot,

1. **Association** or **Direction** of the relationship.
2. **Form** of the relationship.
3. **Strength** of the relationship.
4. Presence or absence of **outliers**.

All four of these items can be described from the scatterplot. It should be noted, though, that the strength of the relationship is best described with the correlation coefficient (see Section 8.4).

Three general statements of association are used – positive, negative, and none. A positive association is when the scatterplot resembles an increasing function (i.e., increases from lower-left to upper-right; Figure 8.4-Right). For a positive association, most of the individuals are above average or below average for both of the variables. A negative association is when the scatterplot looks like a decreasing function (i.e., decreases from upper-left to lower-right; Figure 8.4-Left). For a negative association, most of the individuals are above average for one variable and below average for the other variable. No association is when the scatterplot looks like a flat horizontal line or a “shotgun blast” of points (Figure 8.4-Middle). For no association, there are no tendencies for individuals to be above or below average for one variable and above or below average for the other.

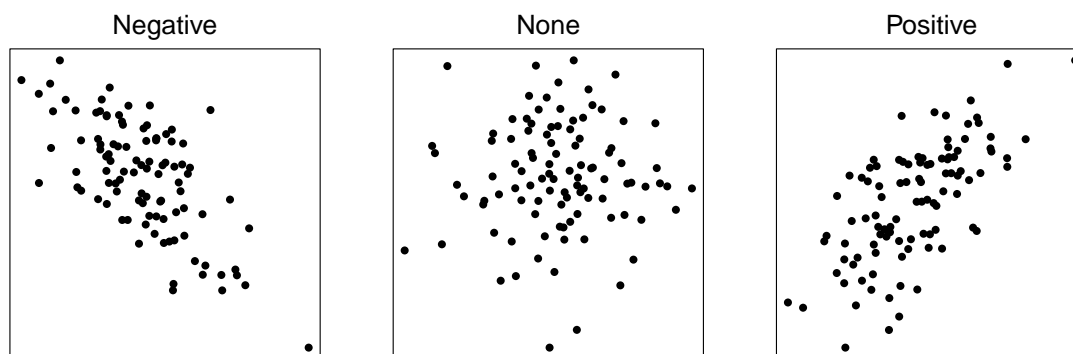


Figure 8.4. Depiction of three types of association present in scatterplots.

△ **Positive Association:** Most of the individuals are either above average or below average for both variables.

△ **Negative Association:** Most of the individuals are above average for one variable and below average for the other variable.

△ **No Association:** There are no tendencies for individuals to be above or below average for one variable and above or below average for the other variable.

For the purposes of this introductory course, form will be defined as either linear or nonlinear. By default, scatterplots will be considered linear unless there is an OBVIOUS curvature in the points. For example, all three scatterplots in Figure 8.4 are considered linear.

Strength is a summary of how closely the points cluster about the general form of the relationship. For example, for linear forms strength is how closely the points cluster around the line. Strength is difficult to define from a scatterplot because it is a relative term. The general idea of strength is depicted in Figure 8.5. However, an objective numerical measure – the correlation coefficient – is defined in Section 8.4.

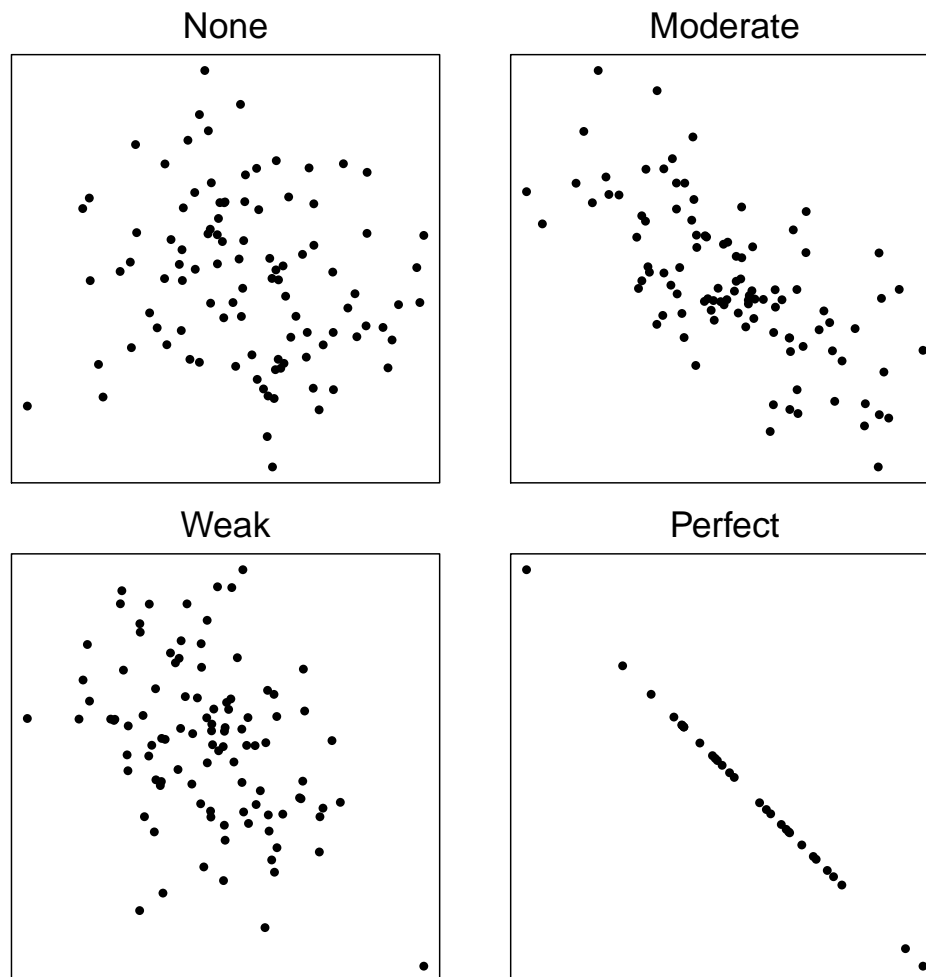


Figure 8.5. Scatterplots depicting four relative types of strength.

△ Strength: How closely the points cluster about the general form of the relationship.

◇ **Strength can only be subjectively described from a scatterplot; use the correlation coefficient to be more objective.**

Outliers are points that are far removed from the main cluster of points. Keep in mind (as always) that just because a point is an outlier doesn't mean it is wrong.

The relationship between highway MPG and the weight of cars (Figure 8.1) appears to be negative, primarily linear (although I see a very slight concavity), and moderately strong. The three points at (2400,46),

(2500,27), and (1800,33) might be considered SLIGHT outliers (these are not far enough removed for me to consider them outliers, but some people may).

A general conclusion that could be made from these results is that as the weight of the cars increases, the highway MPG attained by the car decreases in a linear fashion. While this conclusion is correct, it is also very carefully worded. We must be very careful to not state that increasing the weight of the car CAUSES a decrease in MPG. We cannot attribute cause because these data come from an observational study and because several other important variables were not considered in the analysis. For example, the scatterplot in Figure 8.6, coded for different numbers of cylinders in the car's engine, indicates that the number of cylinders may be inversely related to the highway MPG and positively related to the weight of the car. So, does the weight of the car, the number of cylinders, or both, explain the decrease in highway MPG?

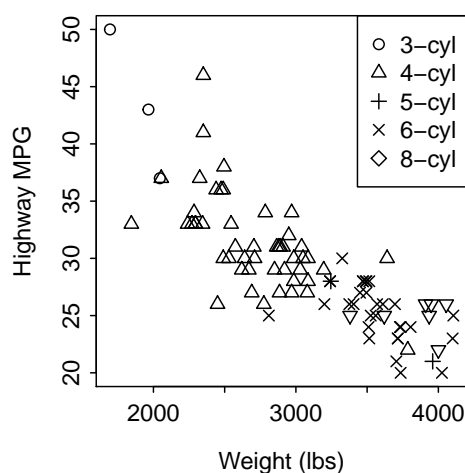


Figure 8.6. Scatterplot between the highway MPG and weight of cars manufactured in 1993 separated by number of cylinders.

Review Exercises

- 8.1** Researchers in Northern Wisconsin wanted to explain the role of the whitetail deer as a keystone herbivore (Waller and Alverson 1997). As a part of their analysis, they examined the relationship between the mean number of hemlock saplings on 14 x 21 m sections of a woodlot and a browsing index (a complicated measurement that gives the amount of food a deer has been eating in a given area). Use the data in the table below to make a scatterplot of the mean number of hemlock saplings versus the browsing index and describe the bivariate relationship from it. [Answer](#)

mean no. hemlock saplings	0.95	2.89	2.97	3.94	4.74	5.10	6.64	7.13
browse index	0.31	0.35	0.49	0.50	0.61	0.63	0.86	0.90

8.4 Correlation

The sample correlation coefficient, abbreviated as r , is calculated with

$$r = \frac{\sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n - 1} \quad (8.4.1)$$

where s_x and s_y are the sample standard deviations for the explanatory and response variable, respectively.⁶ The formulas in the two sets of parentheses in the numerator are standardized values; thus, the value in each parenthesis is called the standardized x or standardized y, respectively.⁷ Using this terminology, the formula for the correlation coefficient reduces to these steps:

1. For each individual, standardize x and standardize y.
2. For each individual, find the product of the standardized x and standardized y.
3. Sum all of the products from step 2.
4. Divide the sum from step 3 by $n-1$.

◊ The sample correlation coefficient is abbreviated with r . The population correlation coefficient is abbreviated with ρ .

The table below illustrates these calculations for the first five individuals in the cars data.⁸ In the table note that the “i” column is an index for each individual, the x_i and y_i columns are the observed values of the two variables for individual i , \bar{x} was computed by dividing the sum of the x_i column by n , s_x was computed by dividing the sum of the $(x_i - \bar{x})^2$ column by $n - 1$ and taking the square root, and the “std x” column is the standardized x values found by dividing the value in the $x_i - \bar{x}$ column by s_x . Similar calculations were made for the y variable. The final correlation coefficient is the sum of the last column divided by $n - 1$. Thus, the correlation between car weight and highway mpg for these five cars is -0.54.

	HMPG	Weight							
i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	std. y	std. x	(std. y)(std. x)
1	31	2705	3.4	-632	11.56	399424	1.26	-1.71	-2.15
2	25	3560	-2.6	223	6.76	49729	-0.96	0.6	-0.58
3	26	3375	-1.6	38	2.56	1444	-0.59	0.1	-0.06
4	26	3405	-1.6	68	2.56	4624	-0.59	0.18	-0.11
5	30	3640	2.4	303	5.76	91809	0.89	0.82	0.73
sum	138	16685	0	0	29.2	547030	0	0	-2.17

There are easier formulae for calculating r than that illustrated above. However, the formula and method above illustrates some intuitive concepts about r . The correlation coefficient is a measure of both association and strength. The sign of r indicates the direction or association between the two variables. A positive r means a positive association and a negative r means a negative association. The absolute value of r (i.e., the value ignoring the sign) is an indicator of the strength of relationship. Absolute values nearer 1 are stronger relationships. Each of these concepts is discussed further next.

⁶See Section 5.6.3 for a review of standard deviations.

⁷See Section 7.5 for a review of standardized values.

⁸The five cars are treated as if they are the entire sample.

A positive association occurs when both variables measured on an individual tend to be above or below average together. To illustrate this concept, examine the scatterplot in Figure 8.7-Left that has lines superimposed at the means of both the x and y variables. Standardized values for measurements larger than the mean are positive, because the difference between the larger observed value and the mean is positive. With similar reasoning, standardized values for measurements smaller than the mean are negative.

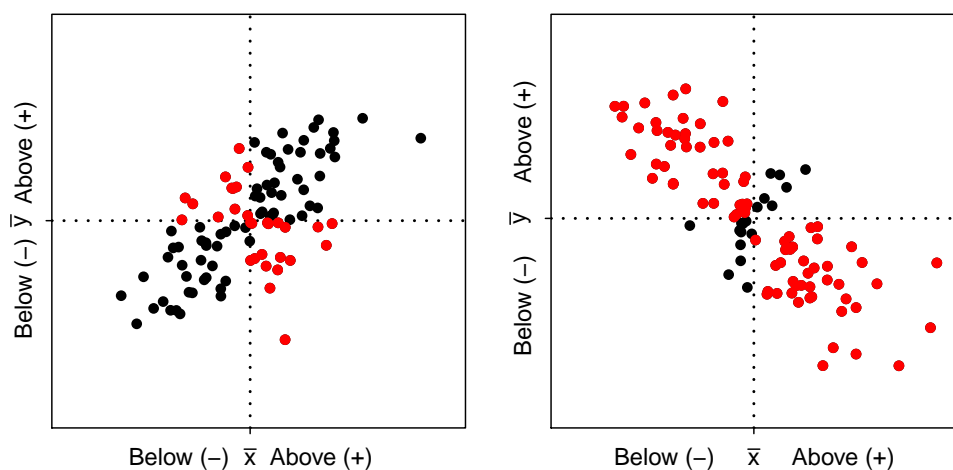


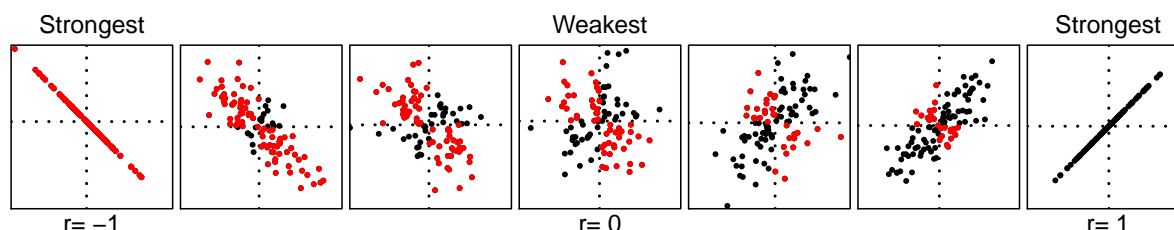
Figure 8.7. Scatterplot with mean lines superimposed and the signs of standardized values for both x and y shown for a positive (**Left**) and negative (**Right**) association.

Now consider the product of standardized x's and y's in each quadrant of Figure 8.7-Left. In the quadrant that corresponds to above average for both standardized values (i.e., both positive signs) the product is positive (denoted by black dots). In the quadrant that corresponds to below average for both standardized values the product is also positive. In the other two quadrants the product is negative (denoted by red dots). From Figure 8.7-Left it is seen that, for a positive association, the numerator of the correlation coefficient is the sum of many positive products of standardized x's and y's (black dots) and few negative products (red dots). Thus, the numerator is positive. The denominator (recall it is $n-1$) is always positive. Thus, the correlation for a positive association is positive.

A negative association is examined in the same manner with Figure 8.7-Right. The signs of the products in the quadrants are the same as described above. With the negative association, the numerator is the sum of many negative products (red dots) and a few positive products (black dots). Thus, the numerator is negative. Therefore, the correlation for a negative association is negative.

◇ **The correlation coefficient is positive for positive associations and negative for negative associations.**

Correlations range from -1 to 1. Absolute values of r equal to 1 indicate a perfect correlation; i.e., all points fall exactly on a line. A correlation of 0 indicates no association. Thus, absolute values of r near 1 indicate strong relationships and those near 0 are weak. The range of correlation values and a few scatterplots illustrating how the strength and direction of the relationship changes along this scale is illustrated in Figure 8.8. The categorizations in Table 8.1 can be used as a rough guideline for categorizing the strength of a relationship between two variables.

Figure 8.8. Scatterplots along the continuum of r values.Table 8.1. Classifications of strength of relationship for absolute values of r by type of study.

Strength of Relationship	Uncontrolled/ Observational	Controlled/ Experimental
Strong	> 0.8	> 0.95
Moderate	> 0.6	> 0.9
Weak	> 0.4	> 0.8

◇ Absolute values of correlation coefficients nearer one are stronger.

The following are important characteristics of correlation coefficients:

- The variables must be quantitative (i.e., if you should not make a scatterplot, then don't calculate r).
- The correlation coefficient only measures strength of LINEAR relationships (i.e., if the form of the relationship is not linear, then the r is meaningless and should not be calculated).
- The units that the variables are measured in do not matter (i.e., r is the same between heights and weights measured in inches and lbs, inches and kg, m and kg, cm and kg, and cm and inches). This is because of the standardization of the two variables in the calculation of r .
- The distinction between response and explanatory variables is not needed. That is, the correlation of GPA and ACT scores is the same as the correlation of ACT scores and GPA.
- Correlation coefficients are between -1 and 1.
- Correlation coefficients are strongly affected by outliers (simply, because both the mean and standard deviation, used in the calculation of r , are strongly affected by outliers).
- Correlation is not causation – just because a strong correlation is observed it doesn't mean that the explanatory variable caused the response variable (an exception may be in carefully designed experimental studies).

◇ The word “correlation” is often mis-used in everyday language. This word is used only when discussing the actual correlation coefficient (i.e., r). When discussing the association between two variables, one should use the word “relationship” rather than “correlation” (e.g., “What is the relationship between age and rate of cancer?”).

8.4.1 Correlations in R

The correlation coefficient (r) between two quantitative variables is computed with `corr()`. With two quantitative variables Y and X , `corr()` can take a formula of the form $Y \sim X$ as the first argument and the

corresponding data.frame name in `data=`.⁹

```
> corr(HMPG~Weight,data=cars93)
[1] -0.8106581
```

The correlation coefficient can be simultaneously computed for all pairs of variables in a data.frame that contains ONLY quantitative variables. For example, to find the correlations between each pair of highway MPG, size of the fuel tank, length, and weight of cars in `cars93`, then these variables must first be isolated and assigned to a new data.frame.

```
> cars93a <- cars93[,c("HMPG","FuelTank","Length","Weight")]
> str(cars93a)
'data.frame': 93 obs. of 4 variables:
 $ HMPG      : int  31 25 26 26 30 31 28 25 27 25 ...
 $ FuelTank  : num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
 $ Length    : int  177 195 180 193 186 189 200 216 198 206 ...
 $ Weight    : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
```

In some instances, the data.frame may contain some missing values (i.e., data that was not recorded). The individuals with missing data are efficiently removed when computing r by including `use="pairwise.complete.obs"` in `corr()`. Thus, the correlations between all pairs of these four variables is computed below (note use of `digits=` to control the number of decimal points returned).

```
> corr(cars93a,use="pairwise.complete.obs",digits=3)
      HMPG FuelTank Length Weight
HMPG    1.000   -0.786  -0.543  -0.811
FuelTank -0.786    1.000   0.690   0.894
Length   -0.543   0.690    1.000   0.806
Weight   -0.811   0.894   0.806    1.000
```

These results are called a correlation matrix where each cell in the matrix represents the r between variables that label the corresponding row and column. Thus, the correlation between highway MPG and size of the fuel tank is -0.786. The correlation matrix has all 1s on the main diagonal because the correlation between a variable and itself is always 1 (i.e., a perfect relationship). In addition, the matrix is symmetric about the main diagonal because the correlation between X and Y is the same as the correlation between Y and X .

◊ If the vector submitted to `corr()` has missing data, then the individuals with missing data should be excluded by including the `use="pairwise.complete.obs"` argument in `corr()`.

A scatterplot matrix is a visual that corresponds to the correlation matrix (Figure 8.9). Each subplot in the scatterplot matrix is a scatterplot with the variable listed in the same column on the x-axis and the variable listed in the same row on the y-axis. For example, the scatterplot in the upper-right corner of Figure 8.9 has highway MPG on the y-axis and car weight on the x-axis. A scatterplot matrix is constructed in R by submitting the “reduced” data frame to `pairs()`.

```
> pairs(cars93a)
```

⁹`corr()` can also use `~Y+X`.

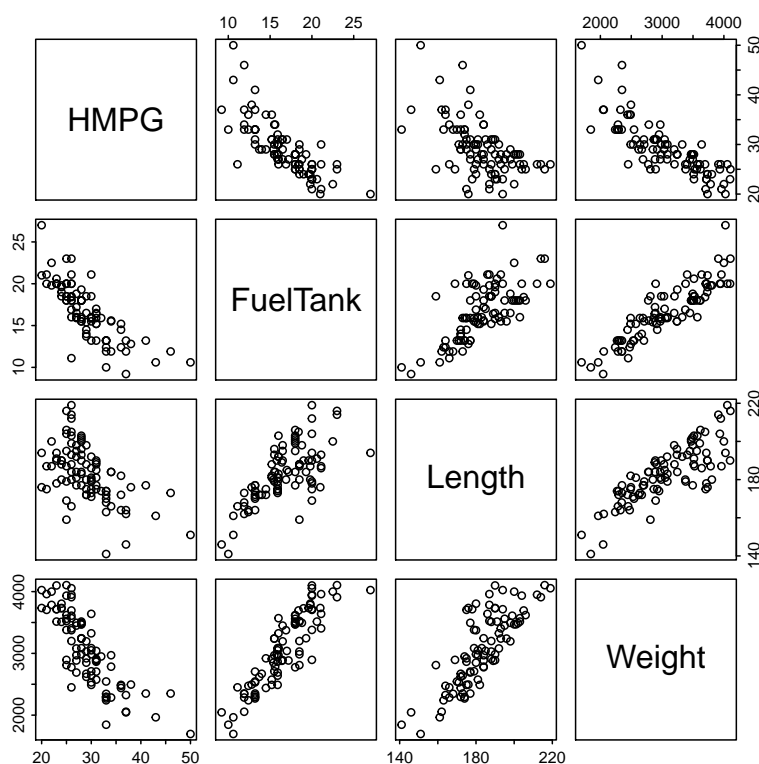


Figure 8.9. Scatterplot matrix of the highway MPG, fuel tank size, length, and weight of cars.

8.5 Example Interpretations

8.5.1 Highway MPG and Weight

The following overall bivariate summary for the relationship between highway MPG and weight is made from the analyses in the previous sections. The relationship between highway MPG and the weight of cars (Figure 8.1) appears to be negative, primarily linear (although I see a very slight concavity), and moderately strong with a correlation of -0.79. The three points at (2400,46), (2500,27), and (1800,33) might be considered SLIGHT outliers (these are not far enough removed for me to consider them outliers, but some people may).

8.5.2 State Energy Usage

A 2001 report from the [Energy Information Administration](#) of the Department of Energy details the total consumption of a variety of energy sources by state in 2001. Construct a proper EDA for the relationship between total petroleum and coal consumption (in trillions of BTU).

The relationship between total petroleum and coal consumption is generally positive, linear, weak, with two outliers at total petroleum levels greater than 3000 trillions of BTU (Figure 8.10-Left). I did not compute a correlation coefficient because of the outliers. The two outliers were Texas and California. After removing them from the data set the relationship is clearly positive, linear, weak ($r = 0.53$), with no additional outliers (Figure 8.10-Right).

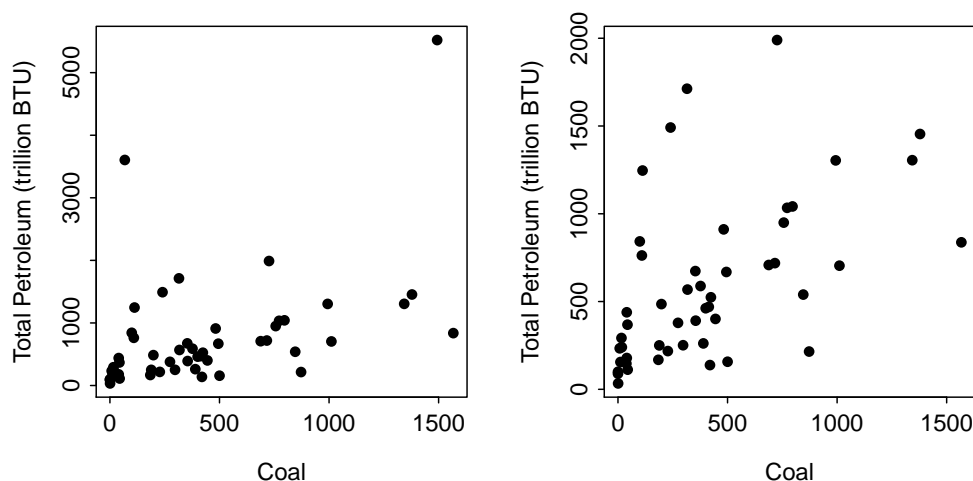


Figure 8.10. Scatterplot of the total consumption of petroleum versus the consumption of coal (in trillions of BTU) by all 50 states and the District of Columbia. The points shown in the left with total petroleum values greater than 3000 trillion BTU are deleted in the right plot.

This example illustrates several key points in the description of a bivariate EDA. First, the descriptions of association, strength, and form should not be influenced by the presence of outliers. In other words, describe association, strength, and form ignoring any outliers present in the data. If you don't have the ability to compute r without the outliers (e.g., you are just given r for the entire data set), then **DO NOT** report r because it is too strongly influenced by the outliers. Second, the form of weak relationships is difficult to describe because, by definition, there is very little clustering to a form. As a rule-of-thumb, if the scatterplot does not have an obvious curvature to it, then it is described as linear by default.

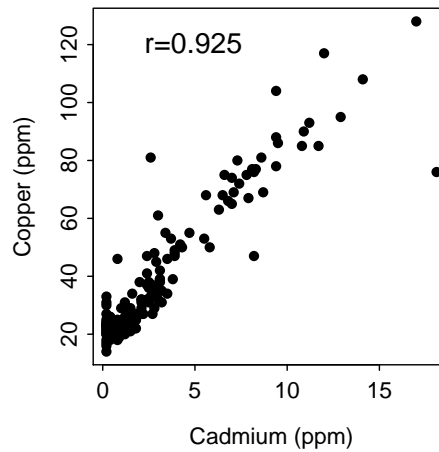
◇ Outliers should not influence the descriptions of association, strength, and form.

◇ The form is linear unless there is an OBVIOUS curvature.

Review Exercises

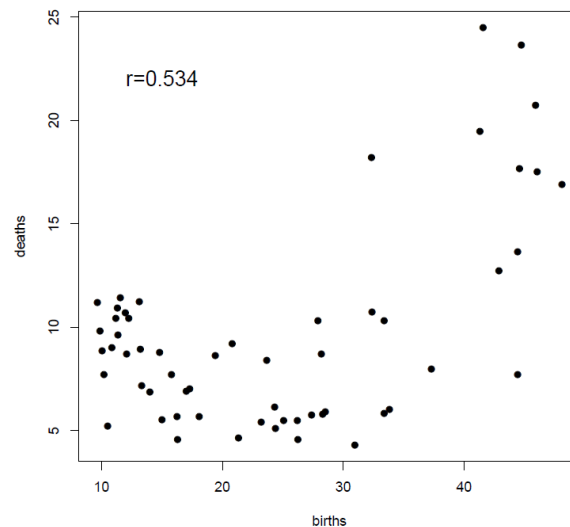
- 8.2 Calculate the correlation coefficient between the mean number of hemlock saplings and deer browse index given in Review Exercise 8.1. [Answer](#)
- 8.3 The concentration of cadmium and copper in the topsoil of 115 15mX15m plots along the river Meuse in the village Stein in New Zealand was recorded by van Rijn and Rikken¹⁰. Use the scatterplot below to describe the bivariate relationship between these two variables. [Answer](#)


¹⁰These data are available in `data(meuse)` of the `sp` package.



- 8.4** Ten variables were measured on 57 countries and reported in the International Vital Statistics (1996). A scatterplot of the birth and death rates is shown below. Write a brief description of this bivariate relationship.

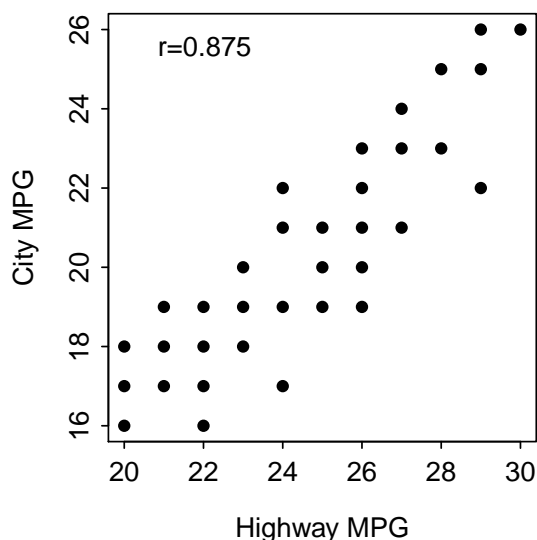
[Answer](#)



- 8.5**  Allen *et al.* (1997) investigated the impact of the density of red-imported fire ants (RIFA) on the recruitment of white-tailed deer fawns (an index of does to fawns). A modified version of their data is recorded in [RIFA.csv](#). Use this information to write a brief description of this bivariate relationship.

[Answer](#)

- 8.6** Researchers at Chevrolet attempted to determine the relationship between gas mileage (MPG) of Lumina in the city (CITY) and on the highway (HIGHWAY). Their results are shown below. Use this information to write a brief description of this bivariate relationship. [Answer](#)

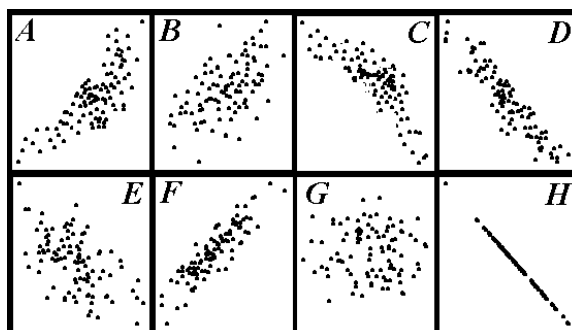


8.7 Mladenoff *et al.* (1997) estimated the territory size (km^2) of wolf (*Canis lupus*) packs and the density of whitetail deer (number/ km^2 ; *Odocoileus virginianus*) in the same areas in northern Wisconsin. Their data is recorded in [Wolves2.csv](#). Load these data into R and generate results to write a brief description of this bivariate relationship. [Answer](#)

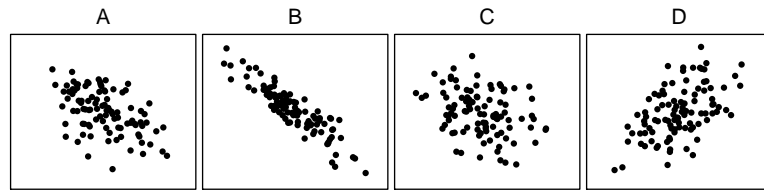
8.8 The Park Management team of Kejimikujik National Park, Nova Scotia examined the relationship between the length and weight of yellow perch (*Perca flavescens*) captured from Grafton Lake in the park in 2000 following the removal of a dam (Brylinsky 2001). Their data is stored in [PerchGL.csv](#). Load these data into R, isolate just the results from 2000 (i.e., use `filterD()`), and generate results to describe this bivariate relationship. [Answer](#)

8.9 It has been said that you can roughly estimate the temperature from the number of cricket chirps heard. To determine if this relationship existed, an entomologist recorded the number of chirps in a 15-second interval by crickets held at different temperatures. The researcher's data is recorded in [Chirps.csv](#). Load these data into R and generate results to write a brief description of this bivariate relationship. [Answer](#)

8.10 Five of the scatterplots below correspond to the following correlation coefficients — 0.89, -0.48, -0.92, 0.56, 0.00. Identify the scatterplot that each correlation corresponds to. Some scatterplots will not be used. [Answer](#)



8.11 Order the following graphs from (i) lowest to highest value of r and (ii) weakest to strongest. [Answer](#)



8.12 Order the following graphs from (i) lowest to highest value of r and (ii) weakest to strongest. [Answer](#)

