MODULE 10

BIVARIATE EDA - QUANTITATIVE

Contents

BIVARIATE DATA OCCURS WHEN TWO variables are 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.

Data on the *weight* (lbs) and highway miles per gallon (*HMPG*) for 93 cars from the 1993 model year are used as an example throughout this module. Ultimately, the relationship between highway MPG and the weight of a car is described. These data are read from 93cars.csv into R and several observations of *HMPG* and *weight* are shown below.[1]

```
> 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
```

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

10.1 Response and Explanatory Variables

The **response variable** is the variable that one is interested in explaining something (i.e., variability) or in making future predictions about. The **explanatory variable** is the variable that may help explain or allow one to predict the response variable. In general, the response variable is thought to depend on the explanatory variable. Thus, the response variable is often called the **dependent variable**, whereas the explanatory variable is often called the **independent variable**.

One may identify the response variable by determining which of the two variables depends on the other. For example, in the car data, highway MPG is the response variable because gas mileage is most likely affected by the weight of the car (e.g., hypothesize that heavier cars get worse gas mileage), rather than vice versa.

In some situations it is not obvious which variable is the response. For example, does the number of mice in the field depend on the number of legumes (lots of feed=lots of mice) or the other way around (lots of mice=not much food left)? Similarly, does area depend on depth or does depth depend on area of the lake? In these situations, the context of the research question is needed to identify the response variable. For example, if the researcher hypothesized that number of mice will be greater if there is more legumes, then number of mice is the response variable. In many cases, the more difficult variable to measure will likely be the response variable. For example, researchers likely wish to predict area of a lake (hard to measure) from depth of the lake (easy to measure).

◇ **Which variable is the response may depend on the context of the research question.**

10.2 Summaries

10.2.1 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 (Figure 10.1).

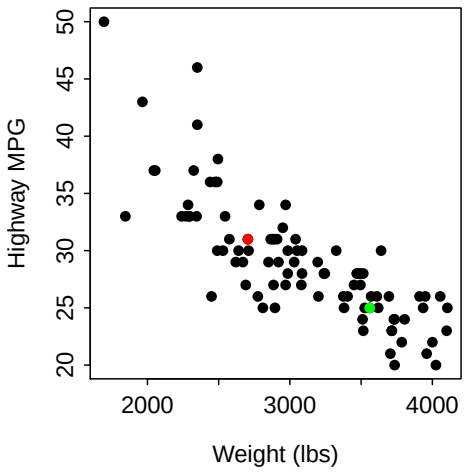

Figure 10.1. Scatterplot between the highway MPG and weight of cars manufactured in 1993. For reference to the main text, the first individual is red and the second individual is green.

Scatterplots are constructed in R with `plot()` with a formula of the form `Y~X`, where `Y` and `X` are variables to be plotted on the y- and x-axes, as the first argument, and the corresponding data.frame in `data=`. The x- and y-axis labels may be modified with `xlab=` and `ylab=`. The character plotted at each point can be changed with `pch=`,[2] which defaults to 1 or an open-circle (Figure 10.2). The scatterplot, excluding the two highlighted points, of highway MPG versus car weight (Figure 10.1) was created with the code below.

```
> plot(HMPG~Weight,data=cars93,xlab="Weight (lbs)",ylab="Highway MPG",pch=16)
```

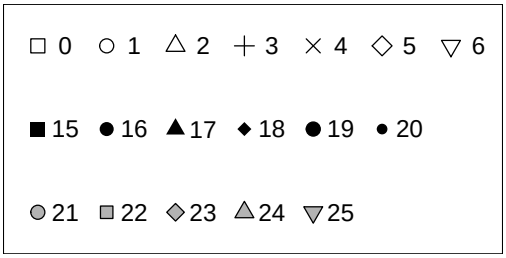

Figure 10.2. Plotting characters available in R and their numerical codes. Note that for values of 21-25 that `bg='gray70'` is used to provide the background color.

10.2.2 Correlation Coefficient

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

$$r = \frac{\sum_{i=1}^{n} \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]}{n-1} \tag{10.2.1}$$

where s_x and s_y are the sample standard deviations for the explanatory and response variables, respectively.[3] The formulae in the two sets of parentheses in the numerator are standardized values;[4] thus, the value in each parenthesis is called the standardized x or standardized y, respectively. Using this terminology, Equation (10.2.1) 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 table below illustrates these calculations for the first five individuals in the cars data.[5] 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 are the standardized x values found by dividing the values 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.

[2] This argument is short for "plotting character".
[3] See Section 6.1.4 for a review of standard deviations.
[4] See Section 8.6 for a review of standardized values.
[5] The five cars are treated as if they are the entire sample.

i	HMPG y_i	Weight 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

The meaning and interpretation of r is discussed in more detail in Section 10.3.

The correlation coefficient (r) between two quantitative variables is computed with `corr()` using a formula of the form `Y~X` or `~Y+X`, where `Y` and `X` are the names of quantitative variables, as the first argument and the corresponding data.frame in `data=`. For example, the correlation coefficient between highway MPG and weight for all cars in the car data is -0.81.

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

10.2.3 Pairs of Multiple Variables

Correlation coefficients can be computed or scatterplots can be constructed simultaneously for all pairs of many quantitative variables. A matrix of correlation coefficients is constructed with `corr()` as above using a formula of the form `~X1+X2+X3` (and so on), where the `X1`, `X2`, etc. are all quantitative variables to be used. In some instances, the data.frame may contain missing values (i.e., data that were not recorded). The individuals with missing data are efficiently removed from the correlation matrix with `use="pairwise.complete.obs"` in `corr()`.[6] The number of digits reported in the correlation matrix is controlled with `digits=`. For example, the correlation between highway MPG and size of the fuel tank is -0.786, whereas the correlation between length and weight of the car is 0.806.

```
> corr(~HMPG+FuelTank+Length+Weight,data=cars93,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
```

A matrix of scatterplots is constructed with `pairs()` using the same formula notation as in `corr()`. The plotting character can be changed, as with `plot()`, with `pch=`. Each subplot in the resulting scatterplot matrix (Figure 10.3) 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 10.3 has highway MPG on the y-axis and car weight on the x-axis.

```
> pairs(~HMPG+FuelTank+Length+Weight,data=cars93,pch=21,bg="gray70")
```

[6] Missing data are automatically removed from the scatterplots.

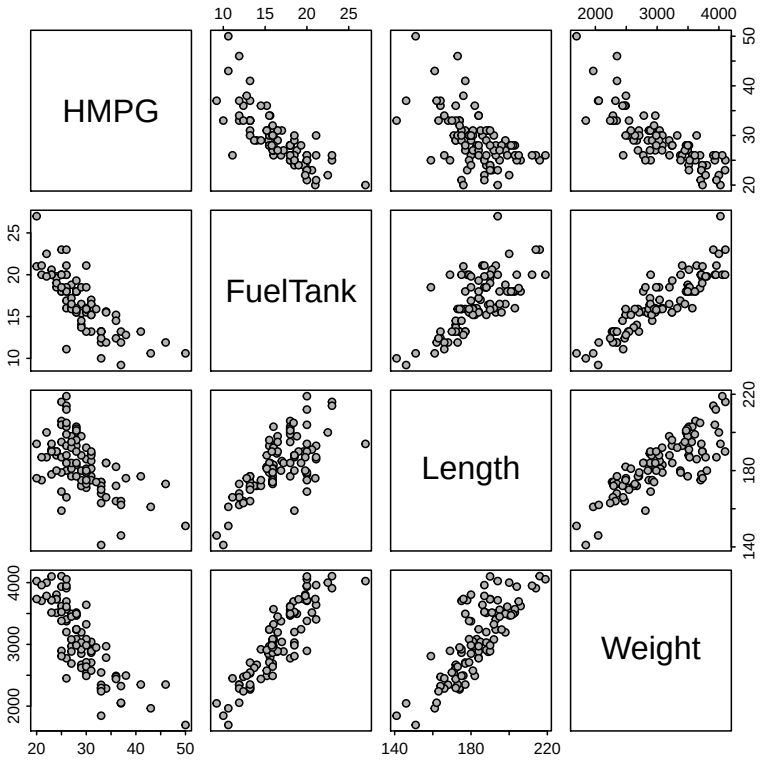

Figure 10.3. Scatterplot matrix of the highway MPG, fuel tank size, length, and weight of cars.

10.3 Items to Describe

Four characteristics should be described for a bivariate EDA with two quantitative variables:

1. **form** of the relationship,
2. presence (or absence) of **outliers**, and
3. **association** or **direction** of the relationship,
4. **strength** of the relationship.

All four of these items can be described from a scatterplot. However, for certain relationships (discussed below), strength is best described from the correlation coefficient.

10.3.1 Form and Outliers

The form of a relationship is determined by whether the "cloud" of points on a scatterplot forms a line or some sort of curve (Figure 10.5). For the purposes of this introductory course, if the "cloud" appears linear then the form will said to be linear, whereas if the "cloud" is curved then the form will be nonlinear. Scatterplots should be considered **linear** unless there is an OBVIOUS curvature in the points.

An outlier is a point that is 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.

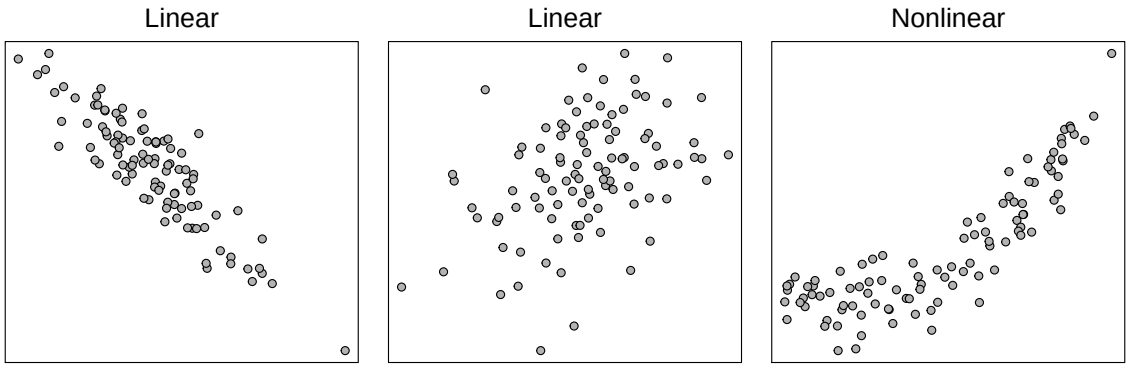

Figure 10.4. Depictions of two linear (Left and Center) and one nonlinear (Right) relationship.

10.3.2 Association or Direction

A positive association is when the scatterplot resembles an increasing function (i.e., increases from lower-left to upper-right; Figure 10.5-Left). 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 10.5-Right). 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 "shotgun blast" of points (Figure 10.5-Center). For no association, there is no tendency for individuals to be above or below average for one variable and above or below average for the other.

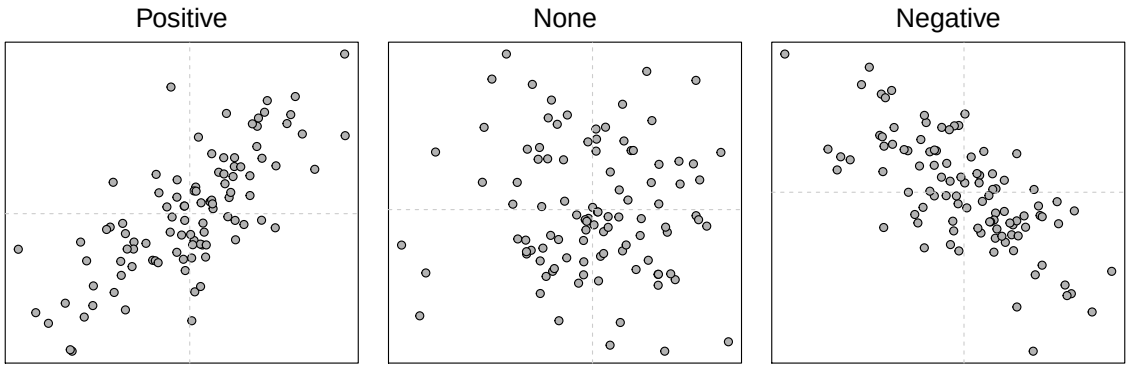

Figure 10.5. Depiction of three types of association present in scatterplots. Dashed vertical lines are at the means of each variable.

10.3.3 Strength (and Association, Again)

Strength is a summary of how closely the points cluster about the general form of the relationship. For example, if a linear form exists, then strength is how closely the points cluster around the line. Strength is difficult to define from a scatterplot because it is a relative term. However, the correlation coefficient (r; Section 10.2.2) is a measure of strength (and association) between two variables, *if the form is linear*.

The sign of r indicates the 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 strength of relationship. Absolute values nearer 1 are stronger relationships.

To better understand how r is a measure of association and strength, reconsider the steps in calculating r from Section 10.2.2. The scatterplots in Figure 10.6 represent a positive (Left) and negative (Right) association. These scatterplots have dashed lines at the mean of both the x- and y-axis variables. Because the mean is subtracted from observed values when standardizing, points that fall above the mean will have positive standardized values and points that fall below the mean will have negative standardized values. The sign for the standardized values are depicted along the axes.

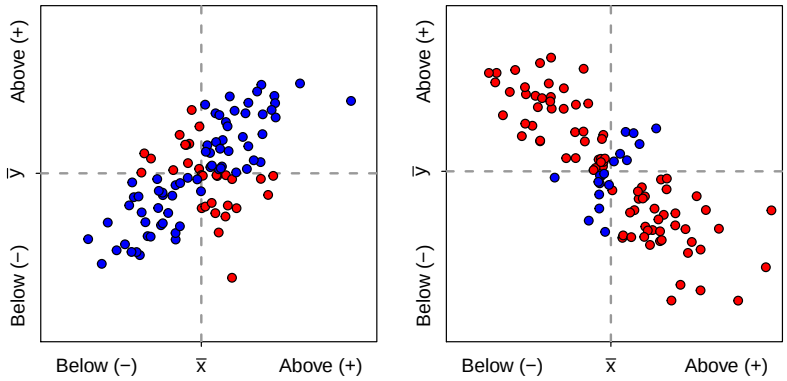

Figure 10.6. Scatterplot with mean lines superimposed and the signs of standardized values for both x and y shown for a positive (**Left**) and negative (**Right**) association. Blue points have a positive product of standardized values, whereas red points have a negative product of standardized values.

Now consider the product of standardized x's and y's in each quadrant of the scatterplots in Figure 10.6. The product of standardized values is positive (blue points) in the quadrant where both standardized values are above average (i.e., both positive signs) and both are below average. The product of standardized values is negative (red points) in the other two quadrants.

Thus, for a positive association (Figure 10.6-Left) the numerator of the correlation coefficient is positive because it is the sum of many positive (blue points) and few negative (red points) products of standardized values. The denominator (recall that it is $n-1$) is always positive. Therefore, r for a positive association is positive. Conversely, for a negative association (Figure 10.6-Right) the numerator of the correlation coefficient is negative because it is the sum of few positive (blue points) and many negative (red points) products of standardized values. Therefore, r for a negative association is negative.

Correlations range from -1 to 1. Absolute values of r equal to 1 indicate a perfect association (i.e., all points 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. How strength and association of the relationship changes along the range of r values is illustrated in Figure 10.7. Categorizations in Table 10.1 can be used as a guideline for describing the strength of relationship between two variables.

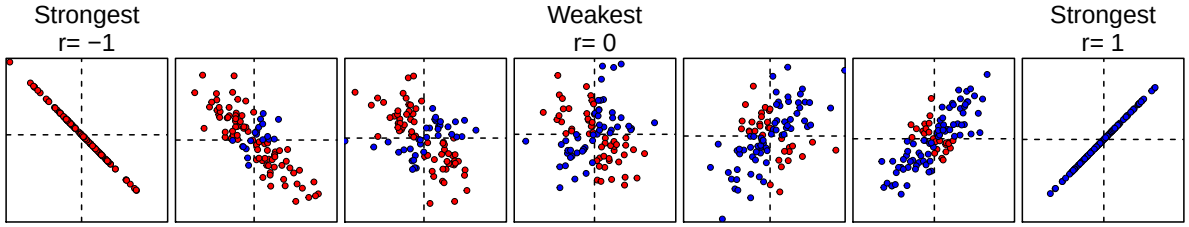

Figure 10.7. Scatterplots along the continuum of r values.

Table 10.1. 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

10.4 Example Interpretations

When performing a bivariate EDA for two quantitative variables, the form, presence (or absence) of outliers, association, and strength should be specifically addressed. In addition, you should state how you assessed strength. Specifically, you should use r to assess strength (see Section 10.3.3) **IF** the relationship is linear without any outliers. However, if the relationship is nonlinear, has outliers, or both, then strength should be subjectively assessed from the scatterplot.

Two other points to consider when performing a bivariate EDA with quantitative variables. First, if outliers are present, do not let them completely influence your conclusions about form, association, and strength. In other words, assess these items ignoring the outlier(s). If you have raw data and the form excluding the outlier is linear, then compute r with the outlier eliminated from the data. 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 is not obviously curved, 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.

Highway MPG and Weight

The following overall bivariate summary for the relationship between highway MPG and weight is made using the calculations from the previous sections.

The relationship between highway MPG and weight of cars (Figure 10.1) appears to be primarily linear (although I see a very slight concavity), negative, and moderately strong with a correlation of -0.81. 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). The correlation coefficient was used to assess strength because I deemed the relationship to be linear without any outliers.

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 linear, with two outliers at total petroleum levels greater than 3000 trillions of BTU, positive, and weak (Figure 10.8-Left). I did not use the correlation coefficient because of the outliers. If the two outliers (Texas and California) are removed then the relationship is linear, with no additional outliers, positive, and weak (r =0.53) (Figure 10.8-Right).

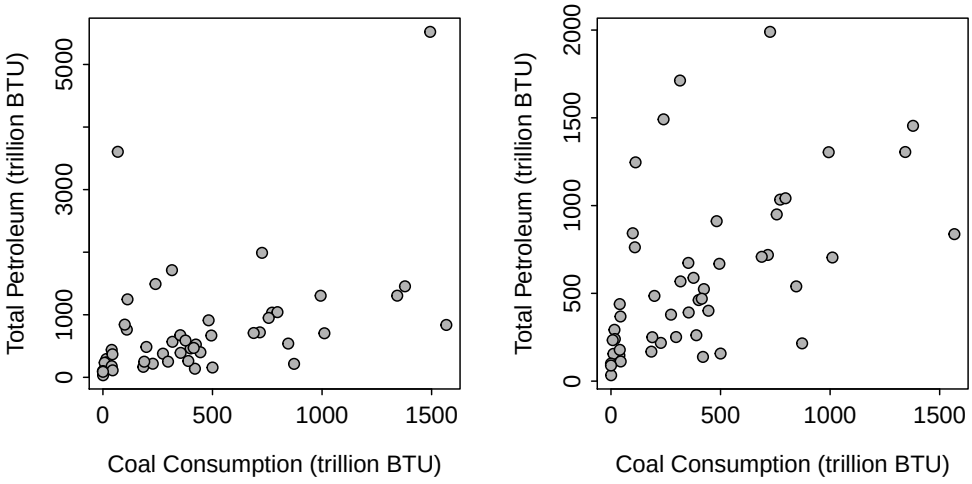

Figure 10.8. Scatterplot of the total consumption of petroleum versus the consumption of coal (in trillions of BTU) by all 50 states and the District of Columbia. The points shown in the left with total petroleum values greater than 3000 trillion BTU are deleted in the right plot.

R Appendix

```
NRG <- read.csv("data/NRG_Consump_2001.csv")
NRG1 <- NRG[-c(5,44),]
plot(TotalPet~Coal,data=NRG,pch=21,bg="gray70",xlab="Coal Consumption (trillion BTU)",
     ylab="Total Petroleum (trillion BTU)")
plot(TotalPet~Coal,data=NRG1,pch=21,bg="gray70",xlab="Coal Consumption (trillion BTU)",
     ylab="Total Petroleum (trillion BTU)")
corr(~Coal+TotalPet,data=NRG1)
```

Hatch Weight and Incubation Time of Geckos

A hobbyist hypothesized that there would be a positive association between length of incubation (days) and hatchling weight (grams) for Crested Geckos (Rhacodactylus ciliatus). To test this hypothesis she collected the incubation time and weight for 21 hatchlings (shown below). Construct a proper EDA for the relationship between incubation time and hatchling weight.

Time	53	54	56	60	60	60	60	60	63	63	77	77	78	81	82	82	83	83	84	90	90
Wt	1.5	1.7	1.4	1.0	1.4	1.5	1.7	1.8	1.4	1.5	1.1	1.6	1.5	1.9	1.4	1.5	1.3	1.7	1.6	1.4	1.8

The relationship between hatchling weight and incubation time for the Crested Geckos is linear, without obvious outliers (*though some may consider the small hatchling at 60 days to be an outlier*), without a definitive association, and weak (r=0.11) (Figure 10.9). I did compute r because no outliers were present and the relationship was linear (or, at least, it was not nonlinear).

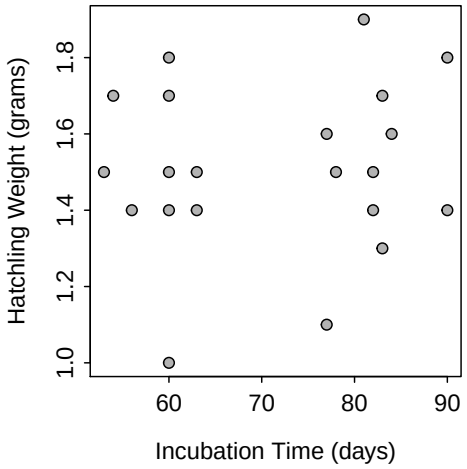

Figure 10.9. Scatterplot of hatchling weight versus incubation time for Crested Geckos.

R Appendix

```
df <- read.csv("data/Gecko.csv")
plot(hatchwt~inctime,data=df,pch=21,bg="gray70",xlab="Incubation Time (days)",
     ylab="Hatchling Weight (grams)")
corr(~inctime+hatchwt,data=df)
```

10.5 Cautions About Correlation

Examining relationships between pairs of quantitative variables is common practice. Using r can be an important part of this analysis, as described above. However, r can be abused through misapplication and misinterpretation. Thus, it is important to remember the following characteristics of correlation coefficients:

- Variables must be quantitative (i.e., if you cannot make a scatterplot, then you cannot calculate r).
- The correlation coefficient only measures strength of **LINEAR** relationships (i.e., if the form of the relationship is not linear, then 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 the variables are standardized when calculating r.
- The distinction between response and explanatory variables is not needed to compute r. 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).

Additionally, correlation is not causation! In other words, just because a strong correlation is observed it does not mean that the explanatory variable caused the response variable (an exception may be in carefully designed experiments). For example, it was found above that highway gas mileage decreased linearly as the weight of the car increased. One must be careful here to not state that increasing the weight of the car CAUSED a decrease in MPG because these data are part of an observational study and several other important variables were not considered in the analysis. For example, the scatterplot in Figure 10.10, coded for different numbers of cylinders in the car's engine, indicates that the number of cylinders may be inversely related to highway MPG and positively related to weight of the car. So, does the weight of the car, the number of cylinders, or both, explain the decrease in highway MPG?

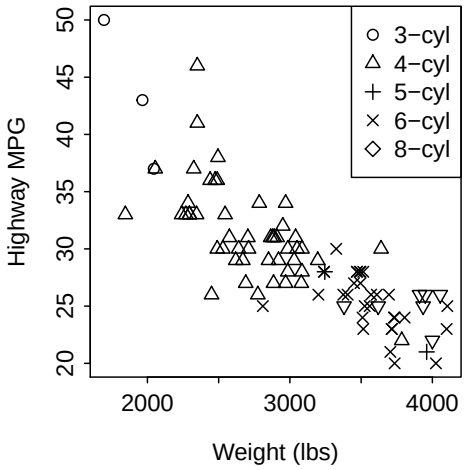

Figure 10.10. Scatterplot between the highway MPG and weight of cars manufactured in 1993 separated by number of cylinders.

More interesting examples (e.g., high correlation between number of people who drowned by falling into a pool and the annual number of films that Nicolas Cage appeared in) that further demonstrate that "correlation is not causation" can be found on the Spurious Correlations website.

Finally, the word "correlation" is often misused in everyday language. "Correlation" should only be used when discussing the actual correlation coefficient (i.e., r). When discussing the association between two variables, one should use "association" or "relationship" rather than "correlation." For example, one might ask "What is the relationship between age and rate of cancer?", but should not ask (unless specifically interested in r) "What is the correlation between age and rate of cancer?".