Constructing and Using Scatterplots

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Constructing and Using Scatterplots

Introduction to Regression Analysis

We described regression earlier as the study of relationships, with the goal of modeling the relationship between two or more variables, with an eye toward assessing causality.

Regression can also be described as the study of dependence. It is used to answer such questions as:

1. Do changes in diet result in changes in cholesterol level?
2. Does an increase in the size of classes result in a reduction in learning?
3. Can a runner's marathon time be predicted from her 5km time?
4. What factors in an insurance company's database can be used to successfully predict whether a claim is fradulent?

Goals of Regression Analysis

1. The goal of regression is to summarize observed data in a simple, elegant, and useful way.
2. Our simplest examples will involve two variables, one of which is predicted from the other.
3. We'll now look at a few examples from Chapter 1 of ALR4, using a tool that it absolutely essential for the analysis of regression data – the *scatterplot*.

Inheritance of Height

- One of the first uses of regression was to study inheritance of traits from generation to generation.
- During the period 1893–1898, E. S. Pearson organized the collection of $n = 1375$ heights of mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18.
- Pearson and Lee (1903) published the data, which are in the data file `Heights`.

Inheritance of Height

The alr4 library must be loaded before we begin. If the alr4 library has been loaded, the Heights data set is automatically available. However, because R allows you to have many data files loaded and available at the same time, and because different data files may have variables with the same name, the system has to be able to avoid "clashes."

Any of the alr4 data files may be referenced and inspected by name. For example, we can take a quick look at the first few lines of the Heights data set as follows:

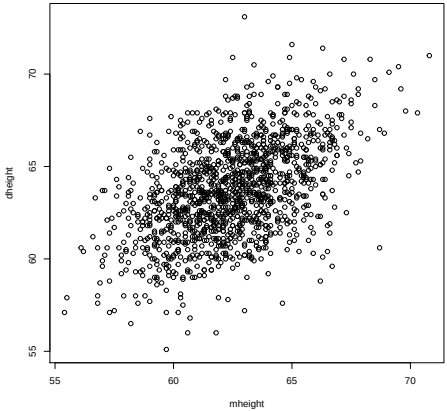
```
> head(Heights)

  mheight dheight
1    59.7    55.1
2    58.2    56.5
3    60.6    56.0
4    60.7    56.8
5    61.8    56.0
6    55.5    57.9
```

Inheritance of Height

- Note that the `Heights` data file contains two variables, `mheight` and `dheight`, representing the mother's and daughter's height for each mother-daughter pair.
- Next, we produce a scatterplot showing the height of the daughter (`dheight`) and the height of the mother (`mheight`).

Inheritance of Height

```
> plot(dheight ~ mheight,data=Heights)
```



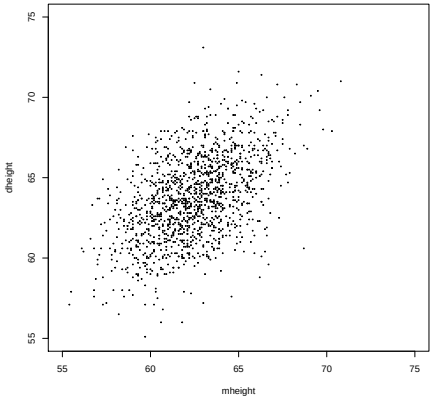
Producing the Scatterplot

Some comments are in order.

- The range of heights appears to be about the same for mothers and for daughters.
- Because of this, we might be better off drawing the plot so that the lengths of the horizontal and vertical axes are the same, and the scales are the same. We can force this by use of the `xlim` and `ylim` options.
- Some computer programs automate the sizing of the X and Y axes, but others may require you to do this for yourself. Fortunately, in R it is very easy to experiment.
- Notice how I've changed the plotting character and the size of the points as well, using the `pch` and `cex` specifications.

Producing the Scatterplot

```
> plot(dheight ~ mheight,data=Heights,
+      xlim=c(55,75),ylim=c(55,75),pch=20,cex=.3)
```



Jittering the Scatterplot

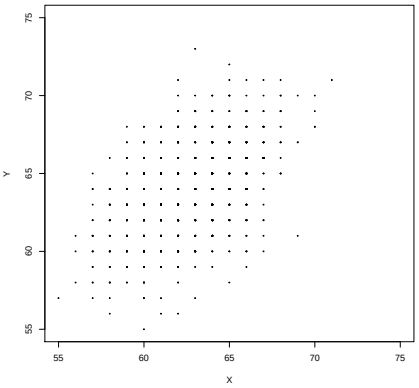
- Weisberg tells us in the text that the original data as published were rounded to the nearest inch.
- In order to avoid an unfortunate problem with such rounded data, Weisberg displaced the data randomly in the X and Y directions by using a uniform random number generator on the range from -0.5 to $+0.5$, then rounding to a single decimal place.
- This type of operation is called *jittering* the scatterplot.
- What problem was he fixing?

Jittering and Un-jittering

- We can round the data back to the nearest inch by using the `round` function in R. This will give us an idea of what we would see if we did not jitter the plot.
- Let's do that, then plot the rounded variables, and see what the new scatterplot looks like. Code is shown below.

Jittering and Un-jittering

```
> X<-round(Heights$mheight)
> Y<-round(Heights$dheight)
> plot(X,Y,xlim=c(55,75),ylim=c(55,75),pch=20,cex=.3)
```

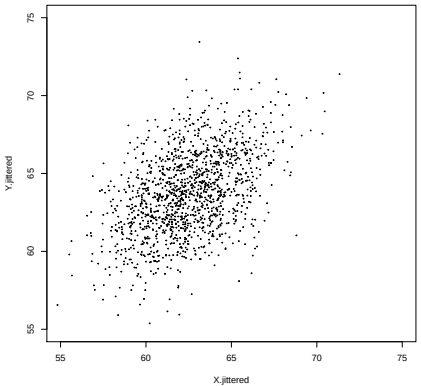


Jittering and Un-jittering

- R has a built-in `jitter` function
- Let's try it with our rounded data.

Jittering and Un-jittering

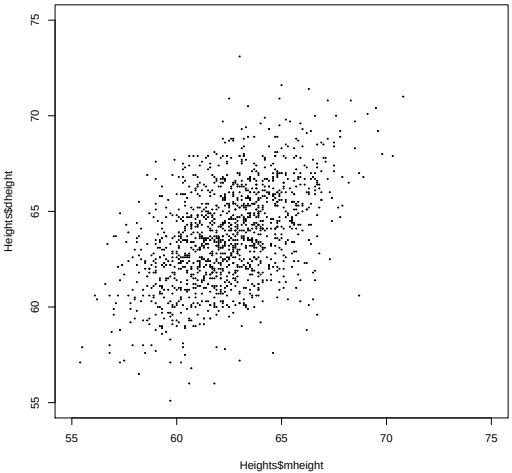
```
> X.jittered <- jitter(X,amount=.5)
> Y.jittered <- jitter(Y,amount=.5)
> plot(X.jittered,Y.jittered,xlim=c(55,75),ylim=c(55,75),pch=20,cex=.3)
```



Examining the Scatterplot

- We examine the scatterplot to see if there is an identifiable dependency.
- If X and Y were independent, then the conditional distribution of Y for a given value of X would not change.
- This is clearly not the case here since as we move across the scatterplot from left to right, the scatter of points is different for each value of the predictor.

Examining the Scatterplot



Examining the Scatterplot

We can see this even more clearly in Weisberg's figure 1.2, in which we show only points corresponding to mother-daughter pairs with *mheight* rounding to either 58,64, or 68 inches.

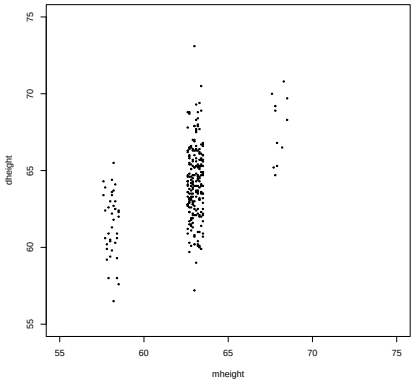
We establish a selection condition with the code below.

```
>  sel <- (57.5 < Heights$mheight) & (Heights$mheight <= 58.5) |
+         (62.5 < Heights$mheight) & (Heights$mheight <= 63.5) |
+         (67.5 < Heights$mheight) & (Heights$mheight <= 68.5)
```

Examining the Scatterplot

Then we plot the figure.

```
> plot(Heights$mheight[sel],Heights$dheight[sel],xlim=c(55,75),ylim=c(55,75),pch=20,cex=.5,
+          xlab="mheight",ylab="dheight")
```



Examining the Scatterplot

We see that within each of these three strips or slices

- The mean of *dheight* is increasing from left to right, and
- The vertical variability in *dheight* seems to be more or less the same for each of the fixed values of *mheight* in the strip.

Examining the Scatterplot

The scatter of points in the graph appears to be more or less elliptically shaped, with the axis of the ellipse tilted upward. We will see in the textbook Section 4.3 that summary graphs that look like this one suggest use of the simple linear regression model.

This model is discussed in detail in Chapter 2 of ALR.

Finding Unusual Cases

- Scatterplots are also important for finding separated points, which are either points with values on the horizontal axis that are well separated from the other points or points with values on the vertical axis that, given the value on the horizontal axis, are either much too large or too small.
- In terms of this example, this would mean looking for very tall or short mothers or, alternatively, for daughters who are very tall or short, given the height of their mother.

Finding Unusual Cases

- These two types of separated points have different names and roles in a regression problem. Extreme values on the left and right of the horizontal axis are points that are likely to be important in fitting regression models and are called leverage points by Weisberg.
- The separated points on the vertical axis, here unusually tall or short daughters give their mother's height, are potentially outliers in Weisberg's terminology. These are cases that are somehow different from the others in the data.

Forbes' Data

- In an 1857 article, a Scottish physicist named James D. Forbes discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water.
- Forbes knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes.
- In the middle of the nineteenth century, barometers were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure.

Forbes' Data

- Forbes collected data from $n = 17$ locations in the Alps and in Scotland.
- He measured at each location pressure in inches of mercury with a barometer and boiling point in degrees Fahrenheit.
- Let's take a look at the scatterplot.

Examining the Scatterplot

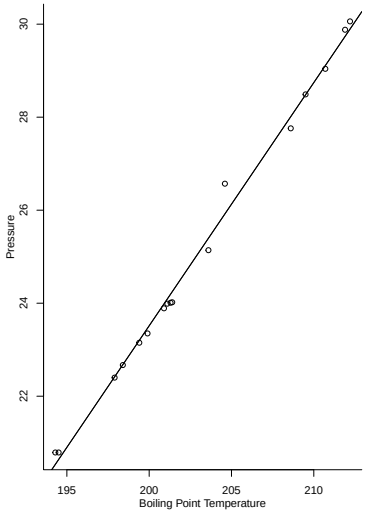
Here is the scatterplot. Of course we have to load the data first. After plotting the data, we add the *best-fitting OLS line* to the plot. This is the straight line that best fits the data according to the Ordinary Least Squares criterion, which we shall discuss in detail later.

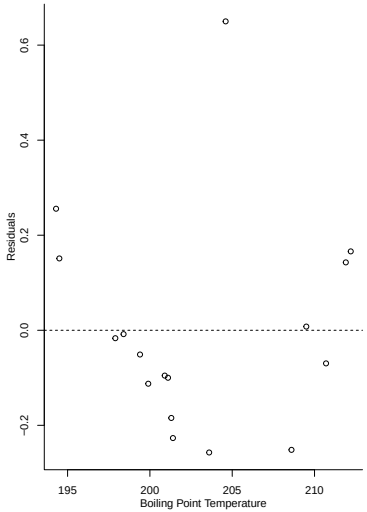
Examining the Scatterplot

Figure 1.3 in the text shows the plot, along side a plot of the model *residuals*.

```
> attach(Forbes)
> oldpar <-par(mfrow=c(1,2),mar=c(4,3,1,.5)+.1,mgp=c(2,1,0))
> plot(bp,pres,xlab="Boiling Point Temperature",
+   ylab="Pressure",bty="l")
> m0 <- lm(pres~bp)
> abline(m0)
> abline(m0)
> plot(bp,residuals(m0), xlab="Boiling Point Temperature",
+    ylab="Residuals",bty="l")
> abline(h=0,lty=2)
> par(oldpar)
```

Examining the Scatterplot





Evaluating Residuals

Look closely at the graph on the left, and you will see that there is a small systematic error with the straight line: apart from the one point that does not fit at all, the points in the middle of the graph fall below the line, and those at the highest and lowest temperatures fall above the line. This is much easier to see in the residual plot on the right.

In examining the residual plot, we look for residuals that are small and that are dispersed around the zero line with approximately equal variability as we move from left to right along the horizontal axis.

In this case, we can see that the residuals do not have the pattern that we want.

Transforming the Dependent Variable

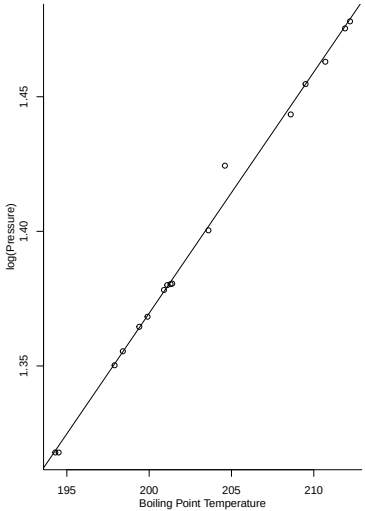
The variable on the vertical axis is the *dependent variable* in the analysis. The variable on the horizontal axis is the *independent variable*. Often, transforming the dependent variable non-linearly can improve the linearity of the scatterplot.

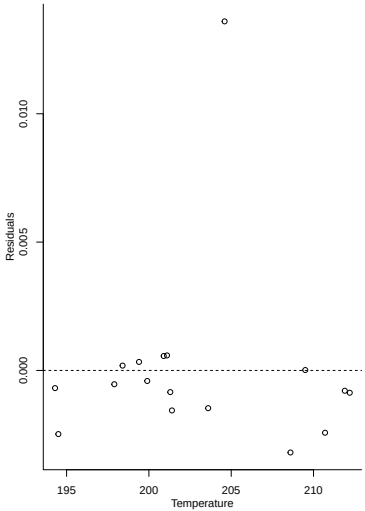
Forbes had a physical theory that suggested that log(*pressure*) is linearly related to *temp*. Forbes (1857) contains what may be the first published summary graph corresponding to his physical model.

Plotting the Transformed Variables

```
> oldpar <- par(mfrow=c(1,2),mar=c(4,3,1,.5)+.1,
+   mgp=c(2,1,0),bty="l")
> plot(bp,logb(pres,10),
+   xlab="Boiling Point Temperature",ylab="log(Pressure)")
> m0 <- lm(logb(pres,10)~bp)
> abline(m0)
> plot(bp,residuals(m0),
+   xlab="Temperature", ylab="Residuals")
> abline(h=0,lty=2)
> par(oldpar)
> detach("Forbes")
```

Residuals of the Transformed Model





Introduction

In Psychology 310, we discussed the basic algebra of regression and correlation, and how it relates to *conditional distributions* in the case where the data are well-approximated by a *bivariate normal* distribution.

These ideas are presented in a slightly different way by Weisberg in ALR. Let's review the key ideas. For more detail, go to the Psychology 310 website and read the relevant handouts.

Variance

The *variance* of a variable is its average squared deviation score, or the expected value of the squared deviation. We have the formula.

$$\sigma_x^2 = \text{Var}(x) = E(X - E(X))^2 \tag{1}$$

Recall that, in the sample, an unbiased estimator is obtained by dividing by $n - 1$ rather than dividing by n. So the *sample variance* S_x^2 is

$$S_x^2 = 1/(n-1)\sum_{i=1}^{n}(X_i - \overline{X}_\bullet)^2 \tag{2}$$

Covariance

The *covariance* of a two variables is the average cross-product of their deviation scores, or the expected value of the product of their deviations. We have the formula.

$$\sigma_{xy} = \text{Cov}(x, y) = E(X - E(X))(Y - E(Y)) \tag{3}$$

The *sample covariance* S_{xy} is

$$S_{xy} = 1/(n-1) \sum_{i=1}^{n} (X_i - \overline{X}_\bullet)(Y_i - \overline{Y}_\bullet) \tag{4}$$

Correlation

The correlation ρ_{xy} between two variables is the average cross-product of their standard scores, or

$$\rho_{xy} = E(Z_x Z_y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{5}$$

The sample correlation is calculated correspondingly as

$$r_{xy} = 1/(n-1) \sum_{i=1}^{n} Z_{x_i} Z_{y_i} = \frac{S_{xy}}{S_x S_y} \tag{6}$$

The OLS Best-Fitting Straight Line

The Ordinary Least Squares line of best fit to a data set is the line that minimizes the sum of squared residuals in the up-down (Y) direction. This line has a slope of $\beta_1 = \rho_{yx}\sigma_y/\sigma_x$ and an intercept of $\beta_0 = \mu_y - \beta_1\mu_x$, with corresponding (non-Greek) formulas in the sample.

With modern software like R, of course we will never have to compute any of these quantities, unless it is for fun.

However, our predicted scores are of the form

$$\hat{Y} = \beta_1 X + \beta_0 \tag{7}$$

Conditional Distributions — The Mean

When Y and X have a bivariate normal distribution, the conditional distribution of Y given X is normal, with a conditional mean that follows the OLS linear regression rule, that is

$$E(Y|X = x) = \beta_0 + \beta_1 x \tag{8}$$

where β_1 and β_0 are the slope and intercept of the OLS regression line.

Conditional Distributions – The Variance

The conditional distribution of Y given X has a variance that is constant, specifically,

$$\text{Var}(Y|X = a) = \sigma^2_\epsilon = (1 - \rho^2_{xy})\sigma^2_y \tag{9}$$

Conditional Distribution of Heights

Weisberg discusses the conditional distribution ideas we reviewed above in Section 1.2–1.3 of ALR. The "Mean Function" that gives the conditional mean of Y given X is simply the OLS regression line.

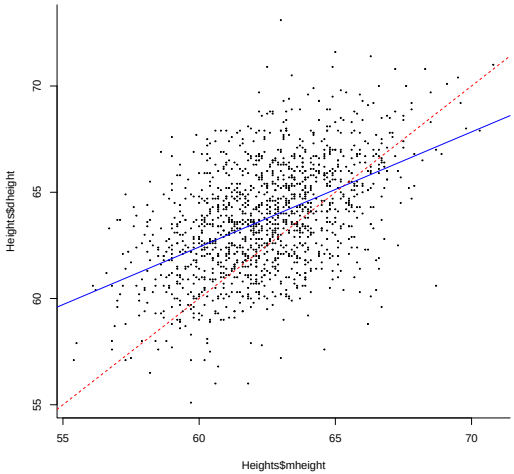
Mean Function

- In Figure 1.8, Weisberg presents the conditional mean line estimated by the OLS regression line, and contrasts it with an "identity" line that represents daughters having, on average, the same height as their mothers.
- By contrasting the two lines, you can see that daughters of tall mothers tend to be taller than average, but somewhat shorter than their mothers.
- Likewise, daughters of short mothers tend to be shorter than the average woman, but taller than their mothers.
- This is the well-known phenomenon of "regression toward the mean" discussed in detail in Psychology 310.

Regression Toward the Mean

```
> ## Scatterplot mheight on horizontal,
> ##   dheight on vertical
> ## In an L-shaped box
> ## Smaller than normal points
> ## Point character is a bullet
> plot(Heights$mheight,Heights$dheight,bty="l",cex=.3,pch=20)
> ## Next draw line with b0=0
> ##   b1=1 dotted red
> abline(0,1,lty=2,col="red")
> ## Next draw regression line
> ##    dheight~mheight solid blue
> abline(lm(Heights$dheight~Heights$mheight),lty=1,col="blue")
```

Regression Toward the Mean



Nonlinear Mean Functions

Depending on the type of data and the nature of the relationship between X and Y, the conditional mean function need not be linear. We'll have a lot more to say about that.

Variance Functions

A frequent assumption in fitting *linear* regression models is that the variance function is the same for every value of X. This is usually written as

$$\text{Var}(Y|X = x) = \sigma^2 \tag{10}$$

where σ^2 is a generally unknown positive constant.

However, we'll also deal with a variety of situations where the variance function is non-constant.

Anscombe's Quartet

It is essential to always examine the scatterplot for bivariate data. The same summary statistics and regression coefficients (means, variances, covariances, correlation, b_0, b_1) can yield very different scatterplots.

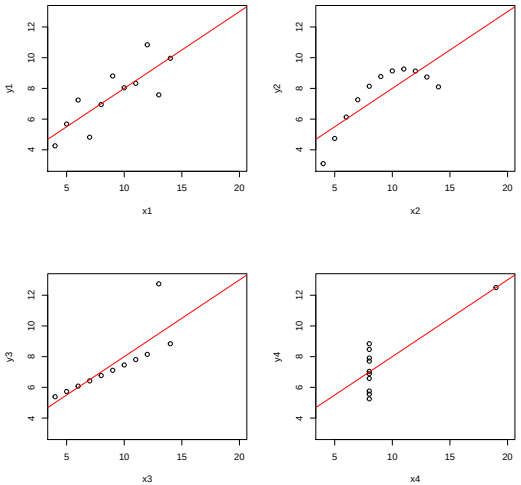
Anscombe (1973) dramatized this phenomenon with 4 small data sets that have come to be known as "Anscombe's Quartet."

They are plotted on the next slide. What do we see (C.P.)?

Anscombe's Quartet

```
> par(mfrow=c(2,2))
> attach(anscombe)
> plot(x1,y1,xlim=c(4,20),ylim=c(3,13))
> abline(lm(y1~x1),col="red")
> plot(x2,y2,xlim=c(4,20),ylim=c(3,13))
> abline(lm(y2~x2),col="red")
> plot(x3,y3,xlim=c(4,20),ylim=c(3,13))
> abline(lm(y3~x3),col="red")
> plot(x4,y4,xlim=c(4,20),ylim=c(3,13))
> abline(lm(y4~x4,col="red"))
```

Anscombe's Quartet



The Loess Smoother

We can estimate the mean function with a model, as we have done with linear regression assuming bivariate normality.

However, we can also "let the data speak for themselves" with various nonparametric techniques. One approach is *smoothing*. The *loess* smoother:

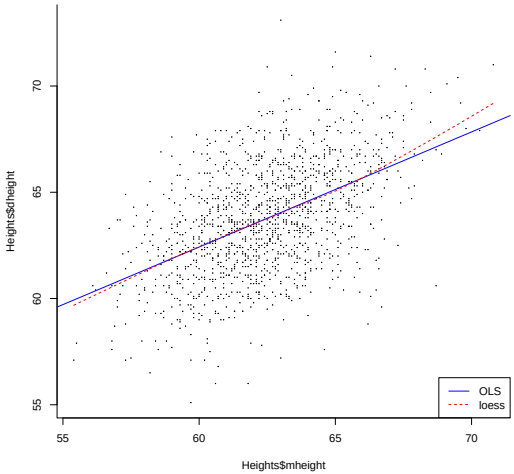
1. Steps across the plot and, for each value x of X, gathers all the points within a certain *span* of x.
2. A regression line is fit to these data (not, in general, by simple linear regression!), and then
3. The conditional mean at x is calculated from that regression line.
4. The points resulting from this process are then graphed as a continuous line.

The Loess Smoother

R automates the calculation and plotting of the loess smoother. Here is some code to generate a smoothed line for the `heights` data.

```
> plot(Heights$dheight~Heights$mheight,cex=.1,pch=20,bty="l")
> abline(lm(Heights$dheight~Heights$mheight),lty=1)
> lines(lowess(Heights$dheight~Heights$mheight,f=6/10,iter=1),lty=2)
> legend("bottomright", c("OLS", "loess"),
+   lty = c(1, 2),col=c("blue","red"))
```

The Loess Smoother



The Scatterplot Matrix

- When we have several potential predictors, a *scatterplot matrix* can help us immediately spot which predictors have an exploitable relationship with the criterion, and
- Also make it relatively easy to spot categorical variables and outliers

Section 1.6 of ALR discusses construction of a scatterplot matrix for data from an analysis of fuel consumption in the U.S.. Let's load the data

```
> attach(fuel2001)
```

and take a quick look.

```
> head(fuel2001)
```

Variable Definitions

TABLE 1.2 Variables in the Fuel Consumption Data[a]

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
State	State name
Fuel	$1000 \times Fuelc/Pop$
Dlic	$1000 \times Drivers/Pop$
log(*Miles*)	Base-two logarithm of *Miles*

Source: "Highway Statistics 2001," http://www.fhwa.dot.gov/ohim/hs01/index.htm.

[a] All data are for 2001, unless otherwise noted. The last three variables do not appear in the data file but are computed from the previous variables, as described in the text.

Additional Calculated Variables

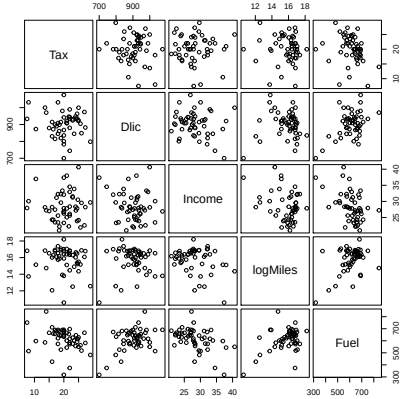
Both *Drivers* and *FuelC* are state totals, so these will be larger in states with more people and smaller in less populous states. Income is computed per person. To make all these comparable and to attempt to eliminate the effect of size of the state, we:

- Compute rates $Dlic = Drivers/Pop$ and $Fuel = FuelC/Pop$, and rescale *Income* to be in thousands.
- Also replace *Miles* by its (base-two) logarithm before doing any further analysis. (Justification for replacing *Miles* with log(*Miles*) is deferred to ALR Problem 7.7.)

```
> fuel2001$Dlic <- 1000*fuel2001$Drivers/fuel2001$Pop
> fuel2001$Fuel <- 1000*fuel2001$FuelC/fuel2001$Pop
> fuel2001$Income <- fuel2001$Income/1000
> fuel2001$logMiles <- logb(fuel2001$Miles,2)
> names(fuel2001)

 [1] "Drivers"  "FuelC"    "Income"   "Miles"    "MPC"      "Pop"
 [7] "Tax"      "Dlic"     "Fuel"     "logMiles"
```

The Scatterplot Matrix

```
> pairs(Tax~Dlic+Income+logMiles+
+   Fuel,data=fuel2001,gap=0.4,
+   cex.labels=1.5)
```



The Scatterplot Matrix

The row of the scatterplot matrix determines the variable that is on the vertical axis, and the column of the scatterplot matrix determines the variable that is on the horizontal axis of any scatterplot.

For example, the upper-right plot is in row 1 and column 5. It shows a plot of *Fuel* on the horizontal axis and *Tax* on the vertical axis.