1 Bivariate analysis

What can we glean from the crackers data set when we look at two numeric variables simultaneously. Can we see what determines the calories per serving? Is there a relationship between `Sodium()` and `Fiber()`? etc. To answer these it helps to look at bivariate relationships.

1.1 Scatterplots

Looking at two numeric variables simultaneously is often done using a *scatterplot*. These are produced in `R` with the `plot()` function. This function can be used several ways. To make a scatterplot or the variables `x` and `y`, we'll use it with an argument like:

```
plot(y ~ x).
```

If only a subset of the data is desired, you can specify this by the indices, or using a logical expression using syntax like

```
plot(y ~ x, subset= ...).
```

Simply replace `...` by a logical expression, or a data vector of indices.

For example, a scatterplot of grams per serving on the x axis and calories per serving on the y axis is done, as follows (Figure 1):

```
> plot(Calories ~ Grams)
```

The scatterplot shows two distinct clusters of data. Within each cluster there appears to be very little trend.

Question 1: Verify that the command

```
> plot(Calories ~ Grams, subset = Calories >= 100)
```

plots just the upper cluster. What subset command using `Grams` instead of `Calories` will produce this same graphic?

Question 2: Make a scatterplot with `Crackers` on the x axis and `Calories` on the y axis. Does there appear to be any trends?

Question 3: From your graph of `Crackers` versus `Calories` there is an "outlier" around (45,90). To identify that point in the data set can be done using the mouse. Type the command

```
> identify(Crackers, Calories, labels = Product)
```

Now click on a point, and the point will be labeled with the corresponding product name. Right click to stop the process. [1]

What product is the outlier?

Question 4: As an exploratory device, multiple scatterplots can be made at once in a graphic called a scatterplot matrix. The `pairs()` function will do so. For example, to make all pairs of scatterplots of the variables 5 through 11 we have the command

[1] In Mac OS X you may need to hit the ESC key.

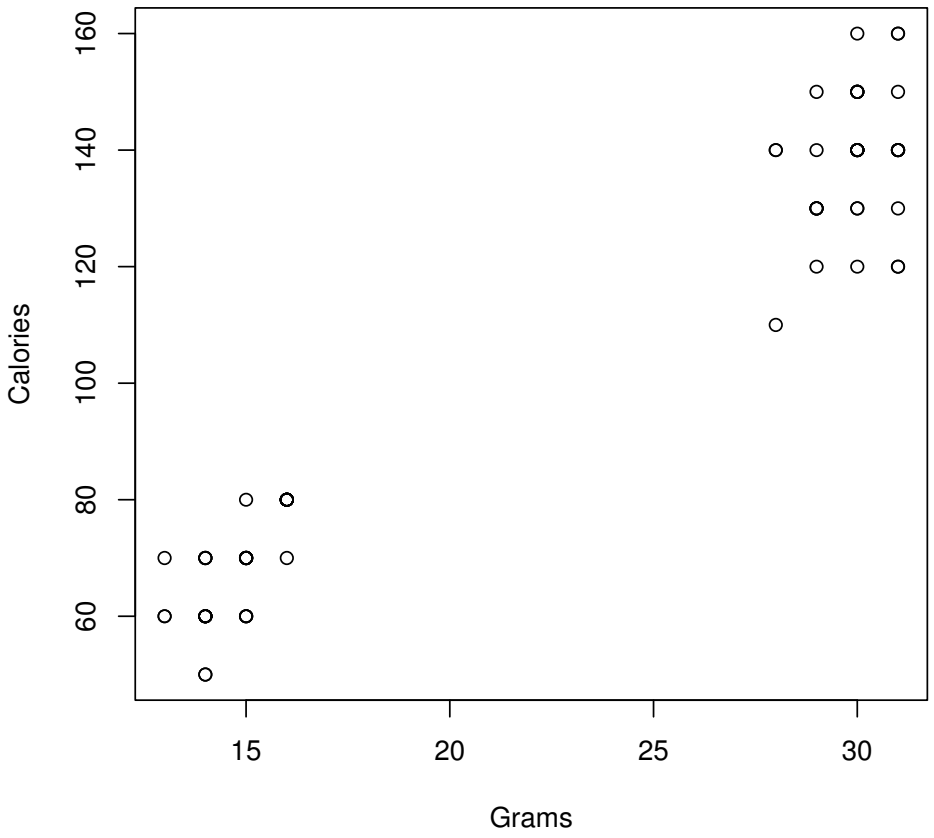

Figure 1: Scatterplot of grams per serving predicting calories per serving

```
> pairs(crackers[5:11])
```

Which relationship is closest to a straight line?

Question 5: Make a scatterplot with `Carbohydrates` on the x axis and `Calories` on the y axis. Does the amount of carbohydrates per serving seem to affect the number of calories per serving?

Question 6: Make an indicator variable, `low.carb`, which is `TRUE` if the number of carbohydrates per serving is less than 15. Make parallel boxplots of the number of calories per serving broken up by the values of `low.carb`. Explain the differences in the boxplot.

1.2 Correlation

The Pearson correlation coefficient is a numeric summary of the strength of a linear relationship between two variables. The Spearman correlation coefficient is a numeric sum-

mary of the *monotonic* relationship between two variables. They are both computed by the `cor()` function. The default is to return the Pearson coefficient. When the extra argument `method="spearman"` is used the Spearman coefficient is returned.

For instance the correlation between the calories per serving and the carbohydrates per serving is computed with

```
> cor(Calories, Carbohydrates)

[1] 0.8670158
```

Question 7: What is the correlation between `Fat.Grams` and `Fat.Calories`? Make a scatterplot and guess the correlation first.

Question 8: What is the correlation between the crackers per serving (`Crackers`) and calories per serving (`Calories`)?

Question 9: Make a scatterplot of crackers per serving (`Crackers`) and calories per serving (`Calories`). Does the relationship appear to be linear? Monotonic? If you said yes to monotonic, compute the Spearman correlation coefficient and compare to the Pearson correlation coefficient found in the previous question.

Question 10: The correlation between `Calories` and `Carbohydrates` is positive. However, a scatterplot shows two distinct clusters. These are inidicated by the indicator variable defined as

```
> low.carb = Carbohydrates < 15
```

Find the correlation for just the low-carbohydrate data and compare to that for the non-low-carbohydrate data. Then compare to the value of 0.867 to the two just found. This is an example of data with two clusters throwing off the interpretation of correlation.

(This can be done with syntax like `Calories[low.carb]` or `Calories[!low.carb]`.)