Scatterplots and Correlation

Reading

Section 7.1.4

Practice Problems

7.5.1 (Page 356) 7.3, 7.4, 7.6, 7.7, 7.8, 7.9, 7.10, 7.15, 7.16, 7.17, 7.18

Notes

Scatterplots

Scatterplots visually display the relationship between two quantitative/scalar variables.

One of the first things we want to do is *describe* the relationship seen in the graph:

Clusters Are there multiple clusters? A "cluster" is a group of points that exhibit a common behavior. They do not have to be near each other, but they do have to share a pattern.

If there are multiple clusters, we usually describe each separately, as far as the following items are concerned.

Form Does the pattern resemble a straight line? Some other curve?

Direction Do the y values increase as the x values increase? Do they decrease instead? These would make the direction *positive*/*negative*.

Strength How close to the points resemble the form?

Outliers Are there points that deviate from the pattern in some way?

Let us look at an example:

This is an image of the relation between city mileage and highway mileage for various car types.

We notice a single cluster, with an overall positive direction, pretty close to linear form. This makes sense to some extent: Some cars are more efficient than others, so they will have better performance both in city and highway.

And in general we don't expect a car to be much more efficient on the highway but not as much more efficient in the city, hence the linear form.

The strength of the relation is that the relation appears to be fairly strong, as the points stay relatively close to the linear direction. There is one very marked *outlier*: A

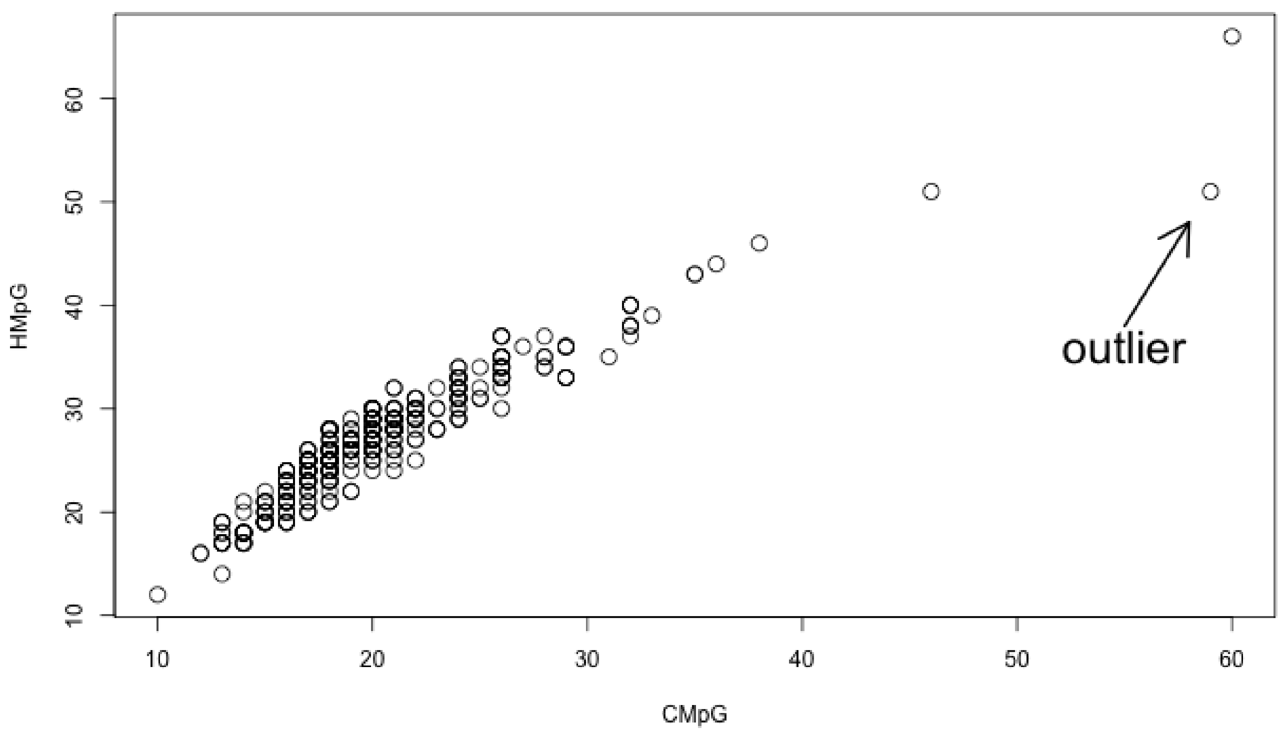

Figure 1: Example of a Scatterplot

car that is doing extremely well in the city, with a very high CMpG, but its highway mileage, while very good, is not just as good.

There are two more cars that are “far from the pack”, but I would not classify them as outliers as they are still part of the linear pattern, just a bit further ahead from the rest.

Let us look at another example:

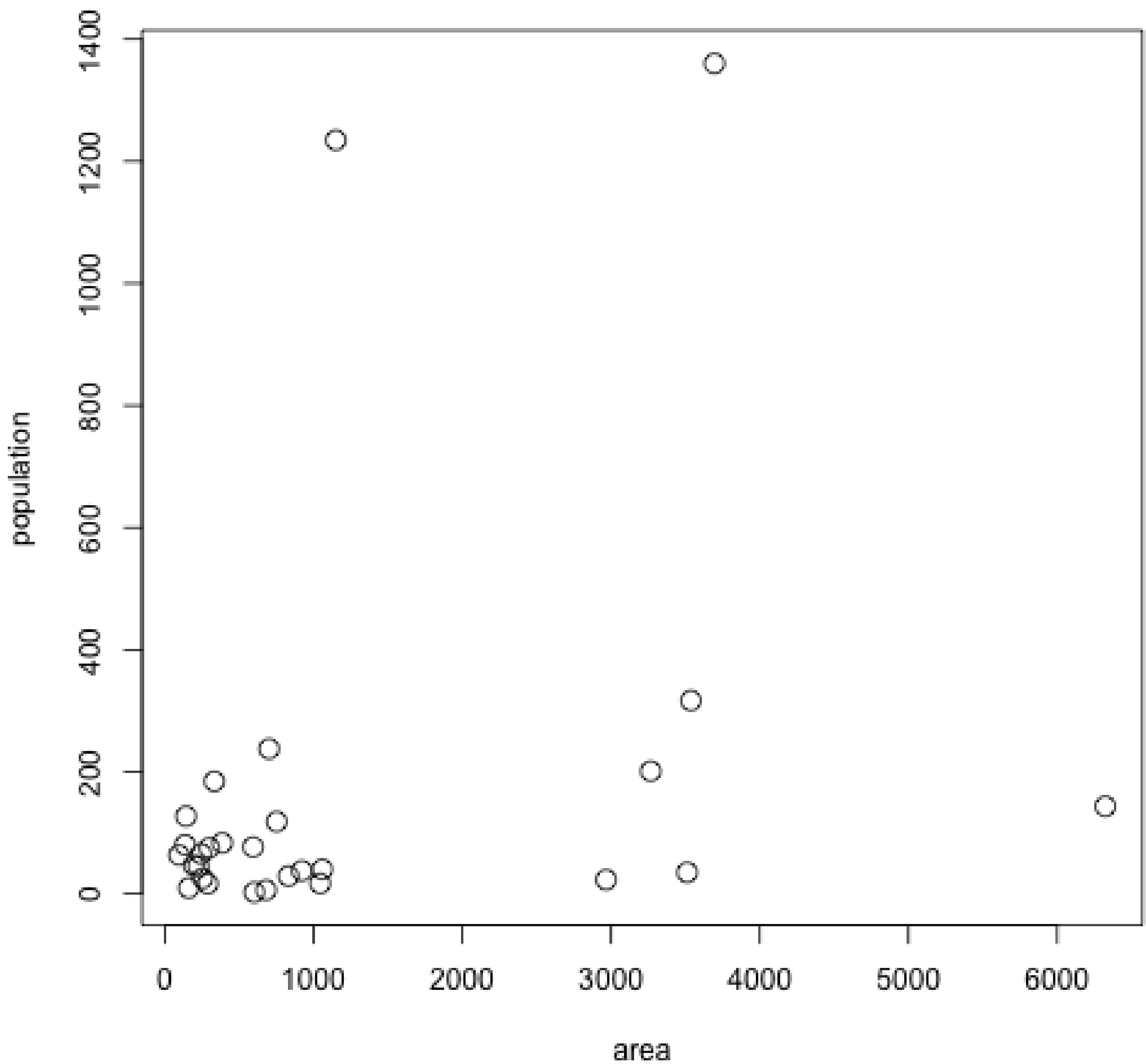

Figure 2: Another example of a Scatterplot

This example shows the relationship between a country’s population and its land mass. See if you can identify the various outliers.

Overall here we can't really talk about a pattern. There are some interesting outliers:

- a country with very little area but a large population;
- a country with moderate area and very large population;
- a country with very large area but relatively small population;
- a cluster of countries with moderate area but relatively small population;

Overall not much of a pattern. We could look for a pattern by "zooming in" on the bottom left of the graph:

Even in this zoomed-in version we do not see much of a pattern going on. Overall we can conclude that there is not much of a relationship between a country's area and its population.

Correlation

The **correlation coefficient** is a specific number, denoted by r, that we compute from our data.

It is only appropriate to use for linear relationships without too many outliers.

In that case it can measure the *strength* and *direction* of the linear relationship.

Some key facts about the correlation coefficient, before we look at its computation:

- It is a unitless number. No matter what measuring units the variables have, r does not have any.
- It does not change if the variables undergo linear transformations: It only depends on their z-scores.
- It is always a value between -1 and 1.
- $r = 1$ means a perfect positive linear relation (all points exactly on a line).
- $r = -1$ means a perfect negative linear relation (all points exactly on a line).
- Values close to those indicate strong association.
- Values close to 0 indicate a very weak association.

For example, our mileage example earlier has an $r = 0.94$, while the population example has an $r = 0.306$, which is typical of a very weak positive correlation.

If we "zoom in" on the lower left part of that graph, we find a correlation of $r = -0.0856$, practically 0.

Now we describe the computation. First the formula:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Basically:

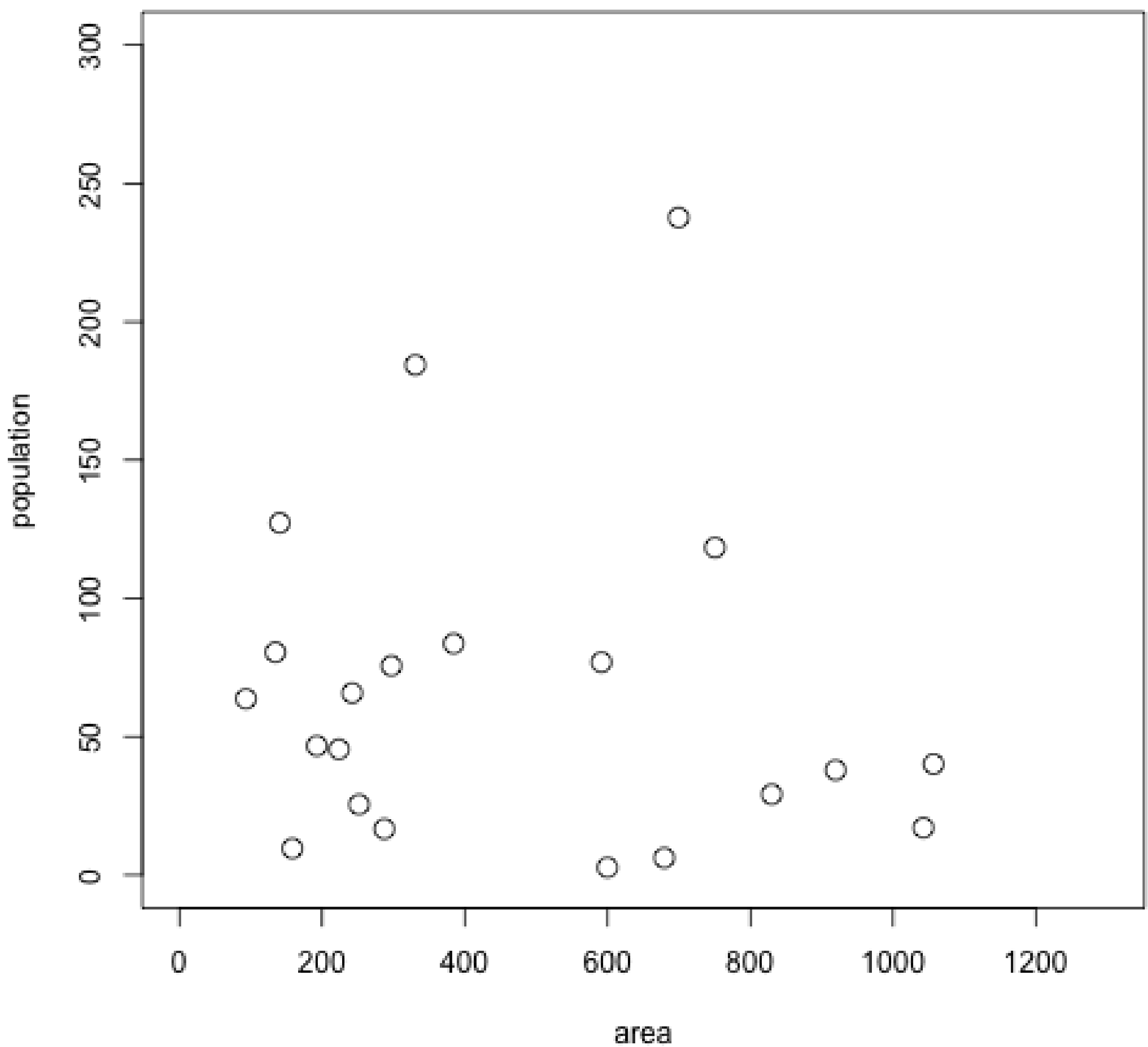

Figure 3: A zoomed-in version

- Compute the standardized values of the x's and the y's.
- Pairwise multiply the corresponding standardized values.
- Average those products, dividing by $n-1$.

The idea behind this is as follows:

- The standardized values are negative when below the mean, and positive when above the mean.
- If we have a positive association, then high x values (above the mean) will tend to be paired with high y values (above the mean), and conversely for low x values. Therefore positive standardized values from x are paired with and multiplied with positive standardized values from y, while negatives are paired with negatives. The result is that almost all of the pairwise products are positive, resulting in a large positive r.
- Conversely, if we have a negative association, then high x values are paired with low y values, and low x values are paired with high y values. In terms of standardized values, it means that positives are paired with negatives, resulting in negative pairwise products. When we average them all up, we end up with negative r.
- Finally, if there is no clear association, then we end up with a mixture of products, which cancel each other out in the summation. This results in r close to 0.