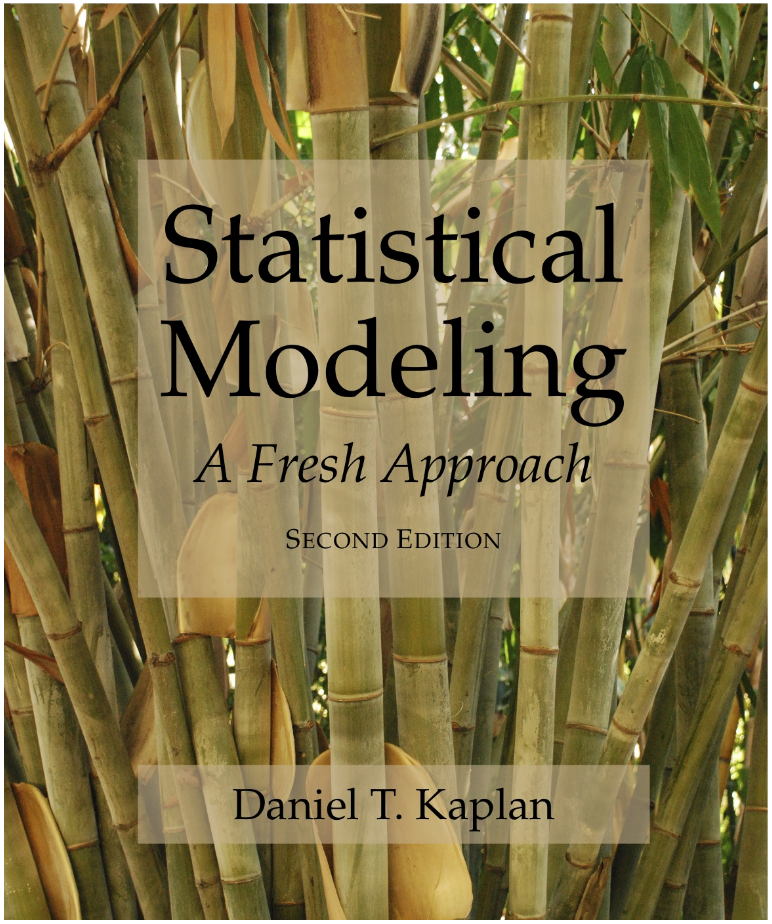


4 Group-wise Models

Seek simplicity and distrust it. — Alfred North Whitehead (1861-1947), mathematician and philosopher

This chapter introduces a simple form of statistical model based on separating the cases in your data into different groups. Such models are very widely used and will seem familiar to you, even obvious. They can be useful in very simple situations, but can be utterly misleading in others. Part of the objective of this chapter is to guide you to understand the serious limitations of such models and to critique the inappropriate, though all too common, use of group-wise models.

4.1 "Grand" and Group-wise Models

Consider these two statements:

> *Adults are about 67 inches tall.*[1]

and

> *Worldwide per-capita income in 2010 was about $12,000 per year.*

Such statements lump everybody together into one group.

You're also familiar with statements that divide things up into separate groups. For example:

[1] 1.70 meters in metric units.

Women are about 64 inches tall; men are about 69 inches tall.

and this sort of information about per-capita income:

Per-capita Income in 2010

Country	Amount	Country	Amount
Qatar	$88,559	Tajikistan	$1935
Luxembourg	$81,383	Bangladesh	$1572
Singapore	$56,522	Haiti	$1165
United States	$47,284	Afghanistan	$907
Switzerland	$41,663	Zimbabwe	$434

Source: International Monetary Fund
World Economic Outlook Database - April 2011

These sorts of statements, either in the **grand mean** form that puts everyone into the same group, or **group-wise mean** form where people are divided up into several groups, are very common.

People are so used to this sort of division by sex or citizenship that it's easy to miss that the point that these are only some of many possible variables that might be informative. For instance, other variables that might also contribute to account for variation in people's height are nutrition, parents' height, ethnicity, etc.

It's understandable to interpret such statements as giving "facts" or "data," not as models. But they are representations of the situation which are useful for some purposes and not for others: in other words, models.

The group-wise income model, for instance, accounts for some of the person-to-person variation in income by dividing things up country by country. In contrast, the corresponding "grand" model is simply that per capita income worldwide is $12,000 — everybody in one group! The group-wise model is much more informative because there is so much spread: some people are greatly above the mean, some much, much lower, and a person's country accounts for a lot of the variation.

The idea of averaging income by countries is just one way to display how income varies. There are many other ways that one might choose to account for variation in income. For example, income is related to skill level, to age, to education, to the political system in force, to the natural resources available, to health status, and to the population level and density, among many other things. Accounting for income with these other variables might provide different insights into the sources of income and to the association of income with other outcomes of interest, e.g., health. The table of country-by-country incomes is a statistical model in the sense that it attempts to explain or account for variation in income.

These examples have all involved situation where the cases are people, but in general you can divide up the cases in your data, whatever they be, into groups

as you think best. The simplest division is really no division at all, putting every case into the same group. This might descriptively be called **all-cases-the-same** models, but you will usually hear them referred to by the statistic on which they are often based: the **grand mean** or the grand median. Here, "grand" is just a way of distinguishing them from group-wise quantities: grand versus group.

Given an interest in using models to account for variation, an all-cases-the-same model seems like a non-contender from the start — if everybody is the same, there is no variation. Even so, grand models are important in statistical modeling; they providing a starting point for measuring variation.

4.2 Accounting for Variation

Models explain or account for some (and sometimes all) of the case-to-case variation. If cases don't vary one from the other, there is nothing to model!

It's helpful to have a way to measure the "amount" of variation in a quantity so that you can describe how much of the overall variation a model accounts for. There are several such standard measures, described in Chapter 3:

- the standard deviation
- the variance (which is just the standard deviation squared)
- the inter-quartile interval
- the range (from minimum to maximum)
- coverage intervals, such as the 95% coverage intervals

Each of these ways of measuring variation has advantages and disadvantages. For instance, the inter-quartile interval is not much influenced by extreme values, whereas the range is completely set by them. So the inter-quartile interval has advantages if you are interested in describing a "typical" amount of variation, but disadvantages if you do not want to leave out even a single case, no matter how extreme or non-typical.

It turns out that the variance in particular has a property that is extremely advantageous for describing how much variation a model accounts for: the variance **partitions the variation** between that explained or accounted for by a model, and the remaining variation that remains unexplained or unaccounted for. This latter is called the **residual variation**.

To illustrate, consider the simple group-wise mean model, "Women are 64.1 inches tall, while men are 69.2 inches tall." How does this model account for variation?

Measuring the person-to-person variation in height by the variance, gives a variation of 12.8 square-inches. That's the total variation to be accounted for.

Now imagine creating a new data set that replaces each person's actual height with what the model says. So all men would be listed at a height of 69.2 inches, and all women at a height of 64.1 inches. Those model values also have a variation, which can be measured by their variance: 6.5 square inches.

Now consider the residual, the difference between the actual height and the heights according to the model. A women who is 67 inches tall would have a residual of 2.9 inches — she's taller by 2.9 inches than the model says. Each person has his or her own residual in a model. Since these vary from person to person, they also have a variance, which turns out to be 6.3 inches for this group-wise model of heights.

Notice the simple relationship among the three variances:

$$\begin{array}{ccccc} \text{Overall} & & \text{Model} & & \text{Residual} \\ 12.8 & = & 6.5 & + & 6.3 \end{array}$$

This is the partitioning property of the variance: the overall case-by-case variation in a quantity is split between the variation in the model values and the variation in the residuals.

You might wonder why it's the variance — the square of the standard deviation, with it's funny units (square inches for height!) — that works for partitioning. What about the standard deviation itself or the IQR or other ways of describing variation? As it happens, the variance, uniquely, has the partitioning property. It's possible to calculate any of the other measures of variation, but they won't generally be such that the variation in the model values plus the variation in the residuals gives exactly the variation in the quantity being modeled. It's this property that leads to the variance being an important measure, even though the standard deviation contains the same information and has more natural units.

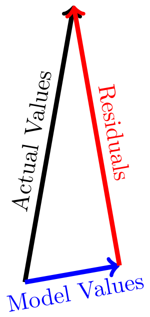
The reason the variance is special can be explained in different ways, but for now it suffices to point out an analogous situation that you have seen before. Recall the Pythagorean theorem and the way it describes the relationship between the sides of a right triangle: if a and b are the lengths of the sides adjoining the right angle, and c is the length of the hypotenuse, then $a^2 + b^2 = c^2$. One way to interpret this is that sides a and b partition the hypotenuse, but only when you measure things in terms of square lengths rather than the length itself.

To be precise about the variance and partitioning ... the variance has this property of partitioning for a certain kind of model — groupwise means and the generalization of that called **linear, least-squares models** that are the subject of later chapters — but those models are by far the most important. For other kinds of models, such as **logistic models** described in Chapter 16 there are other measures of variation that have the partitioning property.

Aside. 4.1 The Geometry of Partitioning

The idea of **partitioning** is to divide the overall variation into parts: that accounted for by the model and that which remains: the residual.

This is done by assigning to each case a model value. The difference between the actual value and the model value is the residual. Naturally, this means that the model value *plus* the residual add up to the actual value. One way to think of this is in terms of a triangle: the model value is one side of the triangle, the residual is another side, and the actual values are the third side.



For reasons to be described in Chapter 8 it turns out that this triangle involve a right angle between the residuals and the model values. As such, the Pythagorean theorem applies and the square of the triangle side lengths add in the familiar way:

$$\text{Model values}^2 + \text{Residuals}^2 = \text{Actual values}^2.$$

4.3 Group-wise Proportions

It's often useful to consider proportions broken down, group by group. For example, In examining employment patterns for workers, it makes sense to consider mean or median wages in different groups, mean or median ages, and so on. But when the question has to do with employment termination — whether or not a person was fired — the appropriate quantity is the proportion of workers in each group who were terminated. For instance, in the job termination data introduced on page 60, About 10% of employees were terminated. This differs from job level to job level, as seen in the table below. For instance, fewer than 2% of Principals (the people who run the company) were terminated. Staff were the most likely to be terminated.

	Administrative	Manager	Principal	Senior	Staff
Retained	91.73	91.84	98.35	90.04	88.43
Terminated	8.27	8.16	1.65	9.96	11.57

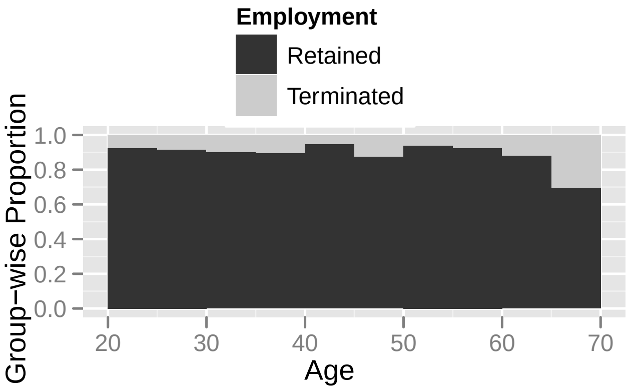


Figure 4.1: The proportion of workers who were terminated broken into groups according to age.

Figure 4.1 shows another way of looking at the termination data: breaking down the groups according to age and, within each group, showing the fraction of workers who were terminated. The graph suggests that workers in their late 60's were substantially more likely to be fired. This might be evidence for age discrimination, but there might be other reasons for the pattern. For instance, it could be that those employees in their 60s were people who failed to be promoted in the past, or who were making relatively high salaries, or who were planning to retire soon. A more sophisticated model would be needed to take such factors into account.

4.4 What's the Precision?

The main point of constructing group-wise models is to be able to support a claim that the groups are different, or perhaps to refute such a claim. In thinking about differences, it's helpful to distinguish between two sorts of criteria:

- Whether the difference is substantial or important in terms of the phenomenon that you are studying. For example, Administrative workers in the previous example were terminated at a rate of 8.27%, whereas Managers were terminated at a rate of 8.16%. This hardly makes a difference.

- How much evidence there is for any difference at all. This is a more subtle point.

In most settings, you will be working with *samples* from a population rather than the population itself. In considering a group difference, you need to take into account the random nature of the sampling process, which leads to some randomness is the group-wise statistics. If different cases had been included in the sample, the results would be different.

Quantifying and interpreting this **sampling variability** is an important component of statistical reasoning. The techniques to do so are introduced in Chapter 5 and then expanded in later chapters. But before moving on to the techniques for quantifying sampling variability, consider the common format for presenting it: the **confidence interval**.

A confidence interval is a way of expressing the precision or repeatability of a statistic, how much variation would likely be present across the possible different random samples from the population. In format, it is a form of coverage interval, typically taken at the 95% level, and looks like this:

$$68.6 \pm 3.3$$

The first number, 68.6 here, is called the **point estimate**, and is the statistic itself. The second number, the **margin of error**, reflects how precise the point estimate is. There is a third number, sometimes stated explicitly, sometimes left implicit, the **confidence level**, which is the level of the coverage interval, typically 95%.

Like other coverage intervals, a confidence interval leaves out the extremes. You can think of this as arranging the confidence interval to reflect what *plausibly* might happen, rather than the full extent of the possibilities, no matter how extreme or unlikely.

4.5 Misleading Group-wise Models

Group-wise models appear very widely, and are generally simple to explain to others and to calculate, but this does not mean they serve the purposes for which they are intended. To illustrate, consider a study done in the early 1970s in Whickham, UK, that examined the health consequences of smoking. [11] The method of the study was simple: interview women to find out who smokes and who doesn't. Then, 20 years later, follow-up to find out who is still living.

Data File
whickham.csv

Examining the data file whickham.csv shows that, after 20 years, 945 women in the study were still alive out of 1315 total: a proportion of 72%. Breaking this proportion into group-wise into smokers and non-smokers gives

Non-smokers : 68.6% were still alive.

Smokers : 76.1% were still alive.

Before drawing any conclusions, you should know what is the precision of those estimates. Using techniques to be introduced in the next chapter, you can calculate a 95% confidence interval on each proportion:

Non-smokers : 68.6 ± 3.3% were still alive.

Smokers : 76.1 ± 3.7% were still alive.

The 95% confidence interval on the difference in proportions is 7.5±5.0 percentage points. That is, the data say that smokers were more likely than non-smokers to have stayed alive through the 20-year follow up period.

Perhaps you are surprised by this. You should be. Smoking is convincingly established to increase the risk of dying (as well as causing other health problems such as emphysema).

The problem isn't with the data. The problem is with the group-wise approach to modeling. Comparing the smokers and non-smokers in terms of mortality doesn't take into account the other differences between those groups. For instance, at the time the study was done, many of the older women involved had grown up at a time when smoking was uncommon among women. In contrast, the younger women were more likely to smoke. You can see this in the different ages of the two groups:

Among non-smokers the average age is 48.7 ± 1.3.

Among smokers the average age is 44.7 ± 1.4.

You might think that the difference of 4 years in average ages is too small to matter. But it does, and you can see the difference when you use modeling techniques to incorporate both age and smoking status as explaining mortality.

Since age is related to smoking, the question the group-wise model asks is, effectively, "Are younger smokers different in survival than older non-smokers?" This is probably not the question you want to ask. Instead, a meaningful question would be, "**Holding other factors constant**, are smokers different in survival than non-smokers?"

You will often see news reports or political claims that attempt to account for or dismiss differences by appealing to "other factors." This is a valuable form of argument, but it ought to be supported by quantitative evidence, not just an intuitive sense of "small" or "big." The modeling techniques introduced in the following chapters enable you to do consider multiple factors in a quantitative way.

A relatively simple modeling method called **stratification** can illustrate how this is possible.

Rather than simply dividing the Whickham data into groups of smokers and non-smokers, divide it as well by age. Here's a table of survival percentages done this way:

	Age Group			
	18-30	31-40	41-53	54-64
Non-Smokers	98.2	95.5	87.6	66.9
Smokers	97.6	94.1	80.2	58.1

Within each age group, smokers are less likely than non-smokers to have been alive at the 20-year follow-up, especially in the older groups. By comparing people of similar ages — stratifying or **disaggregating** the data by age — the model is effectively "holding age constant."

You may rightly wonder whether the specific choice of age groups plays a role in the results. You also might wonder whether it's possible to extend the approach to more than one stratifying variable, for instance, not just smoking status but overall health status. The following chapters will introduce modeling techniques that let you avoid having to divide variables like age into discrete groups and that allow you to include multiple stratifying variables.

4.6 Computational Technique

Calculating means and other simple statistics is a matter of using the right function in R. The mosaic package — which you should load in to R as shown in Section 2.5.1 — makes it straightforward to calculate either a "grand" statistic or a "group-wise" statistic. To illustrate:

Load the mosaic package, needed only once per session:

```
> require(mosaic)
```

Read in data you are interested in analyzing, for instance the Cherry-Blossom 2008 data described earlier:

DATA FILE
Cherry-
Blosssom-
2008.csv

```
> runners = fetchData("Cherry-Blossom-2008.csv")
> names( runners )
```

```
[1] "position" "division" "total"    "name"     "age"      "place"    "net"
[8] "gun"      "sex"
```

Calculate a grand mean on the "gun" time — the time from the start of the race, signalled by a gun, and when each runner crossed the finish line:

```
> mean( gun, data=runners )
```

```
[1] 93.7
```

Other "grand" statistics include:

```
> median( gun, data=runners )
[1] 93.7
> sd( gun, data=runners )
[1] 15
```

To tell R that you want to break the statistics down by groups, use the ˜ notation, pronounced "**tilde**." You will be using this notation frequently in building models. It means, "model by" or "broken down by" or "versus."

```
> mean( gun ˜ sex, data=runners )
   F    M
98.8 88.3
```

Other statistics work the same way, for instance,

```
> sd( gun ˜ sex, data=runners )
   F    M
13.3 14.7
```

Another example ... wage broken down by sector of the economy, using data :

```
> cps = fetchData("cps.csv")
> mean( wage ˜ sector, data=cps )
clerical    const    manag    manuf    other     prof    sales  service
    7.42     9.50    12.70     8.04     8.50    11.95     7.59     6.54
```

In the Whickham smoking data example, the outcome for each person was not a number but a category: Alive or Dead at the time of the follow-up.

```
> w = fetchData("whickham.csv")
> names(w)
[1] "outcome" "smoker"  "age"
> with( w, levels(outcome) )
[1] "Alive" "Dead"
```

To find the proportion of people who were alive at the end of the 20-year follow-up period, you can use a computational trick. Convert the `outcome` variable to `TRUE` or `FALSE` to indicate whether an individual is alive, then take the mean of the true/false values. R, like many computer languages, treats `TRUE` as 1 and `FALSE` as 0 for the purposes of doing arithmetic.

```
> mean( outcome=="Alive", data=w )
[1] 0.719
```

Here's the breakdown according to smoking status:

```
> mean( outcome=="Alive" ˜ smoker, data=w )
   No   Yes
0.686 0.761
```

A more meaningful question is whether smokers are different from non-smokers when holding other variables constant, such as age. To address this question, you need to add age into the model.

It might be natural to consider each age — 35, 36, 37, and so on — as a separate group, but you won't get very many members of each group. And, likely, the data for 35 year-olds has quite a lot to say about 36 year-olds, so it doesn't make sense to treat them as completely separate groups.

You can use the `cut()` function to divide up a quantitative variable into groups. You get to specify the breaks between groups. Using `transform()`, you can add the new variable to an existing data frame.

```
> w = transform(w, ageGroups=cut(age,breaks=c(0,30,40,53,64,75,100)))
> mean( outcome=="Alive" ˜ ageGroups, data=w )
  (0,30]  (30,40]  (40,53]  (53,64]  (64,75] (75,100]
   0.979    0.948    0.832    0.625    0.201    0.000
> mean( outcome=="Alive" ˜ smoker + ageGroups, data=w )
   No.(0,30]   Yes.(0,30]   No.(30,40]   Yes.(30,40]   No.(40,53]   Yes.(40,53]
       0.982        0.976        0.955         0.941        0.876         0.802
  No.(53,64]  Yes.(53,64]   No.(64,75]   Yes.(64,75]  No.(75,100] Yes.(75,100]
       0.669        0.581        0.214         0.158        0.000         0.000
```

The mean has been calculated group-by-group. This is a very widely used technique, but there is a better approach that will be introduced in later chapters: use quantitative variables directly without dividing them into groups.

4.6.1 Model Values and Residuals

A group-wise model tells you the model value for each group. There is additional information that you will want to generate about models. Two fundamental aspects of a model are the **fitted model values** for each case and the **residual** for each case. To make it easy to do these calculations, R has a set of modeling functions that keep track of the data used in creating the model. Later chapters will introduce the `lm()` function — `l` for "linear", `m` for "model" — that is central to statistical modeling. To create a model based on groupwise means, use the `mm()` function — the first `m` for means, the second `m` for "model." `mm()` does the same sorts of calculations as `mean()`, but packages up its results in a different way:

```
> kids = fetchData("kidsfeet.csv")
> mod = mm( width ˜ sex, data=kids )
> mod
Groupwise Model
Coefficients:
   B    G
9.19 8.78
```

The fitted() function takes as an argument a model such as generated by mm(). fitted() computes the fitted model value for each case in the data used to fit the model. To do this, fitted() takes each case in turn, figures out which group it belongs to and the corresponding model value, and then returns the set of model value for each of the cases. For example:

```
> fitted( mod )
```

```
 [1] 9.19 9.19 9.19 9.19 9.19 9.19 9.19 8.78 8.78 9.19 9.19 9.19 9.19 9.19 8.78
[16] 8.78 8.78 8.78 8.78 8.78 9.19 9.19 8.78 8.78 8.78 9.19 8.78 9.19 9.19 9.19
[31] 8.78 8.78 8.78 9.19 9.19 8.78 8.78 8.78 8.78
```

The residuals are found by subtracting the case-by-case model value from the actual values for each case. This is accomplished by the resid() function, which works in much the same way as fitted():

```
> resid(mod)
```

```
 [1] -0.790 -0.390  0.510  0.610 -0.290  0.510  0.410  0.016  0.516 -0.390
[11]  0.610 -0.290 -0.090  0.610  0.516 -0.884 -0.084  0.016  0.216  0.716
[21]  0.010 -0.590 -0.484  0.216 -0.684  0.210  0.716  0.310 -0.290  0.110
[31]  0.516 -0.184 -0.184 -0.190 -0.590 -0.284  0.216 -0.884  0.016
```

For each case, adding the residual to the fitted value will produce the value of the response variable.

The var() function calculates variance. There is a simple additive relationship among the variances of the response variable, the model's fitted values, and the corresponding residual values:

```
> var( width, data=kids )  # overall variation
```

```
[1] 0.2597
```

```
> var( fitted(mod) ) # variation in model values
```

```
[1] 0.04222
```

```
> var( resid(mod) ) # residual variation
```

```
[1] 0.2175
```

Reading Questions

1. Which is larger: variance of residuals, variance of the model values, or the variance of the actual values?
2. How can a difference in group means clearly shown by your data nonetheless be misleading?
3. What does it mean to partition variation? What's special about the variance — the square of the standard deviation — as a way to measure variation?

For exercises, see www.mosaic-web.org/StatisticalModeling/Exercises.