CHAPTER 16

Line Up, Please

Sheepdog demonstration, Lone Pine Sanctuary, Brisbane, QLD. Photo credit: Jeff Stanton

Data users are often interested in questions about relationships and prediction. For example, those interested in athletics might want to know how the size of the fan base of a team is connected with attendance on game day. In this chapter, our Australian colleague, Robert de Graaf, introduces the techniques of linear regression, a very important data science tool.

Using R to Find Relationships between Sets of Data via Multiple Regression, by Robert W. de Graaf

Finding relationships between sets of data is one of the key aims of data science. The question of 'does x influence y' is of prime concern for data analysts – are house prices influenced by incomes, is the growth rate of crops improved by fertilizer, do taller sprinters run faster?

The work horse method used by statisticians to interpret data is linear modeling, which is a term covering a wide variety of methods, from the relatively simple to very sophisticated. You can get an idea of how many different methods there are by looking at the Regression Analysis page in Wikipedia and checking out the number of entries listed under 'Models' on the right hand sidebar (and, by the way, the list is not exhaustive).

The basis of all these methods is the idea that is possible to fit a line to a set of data points which represents the effect an "independent" variable is having on a "dependent" variable. It is easy to visualize how this works with one variable changing in step with another variable. Figure one shows a line fitted to a series of points, using the so called "least squares" method (a relatively simple mathematical method of finding a best fitting line). Note that although the line fits the points fairly well, with an even split (as even as it can be for five points!) of points on either side of the line, none of the points are precisely on the line – the data do not fit the line precisely. As we discuss the concepts in regression analysis further, we will see that understanding these discrepancies is just as important as understanding the line itself.

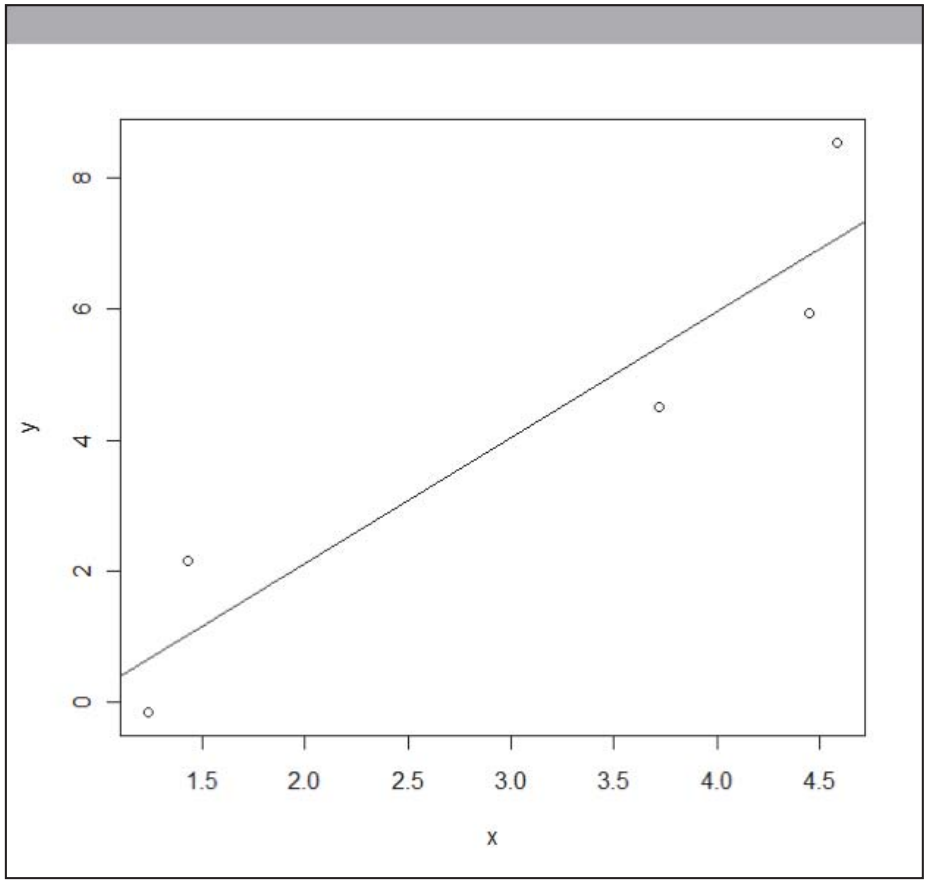


Figure 1: A line fitted to some points

The graph in figure 1 above shows how the relationship between an input variable – on the horizontal x axis – relates to the output values on the y axis.

The original ideas behind linear regression were developed by some of the usual suspects behind many of the ideas we've seen already, such as Laplace, Gauss, Galton, and Pearson. The biggest in-

dividual contribution was probably by Gauss, who used the procedure to predict movements of the other planets in the solar system when they were hidden from view, and hence correctly predict when and where they would appear in view again.

The mathematical idea that allows us to fit lines of best fit to a set of data points like this is that we can find a position for the line that will minimize the distance the line is from all the points. While the mathematics behind these techniques can be handled by someone with college freshman mathematics the reality is that with even only a few data points, the process of fitting with manual calculations becomes very tedious, very quickly. For this reason, we will not discuss the specifics of how these calculations are done, but move quickly to how it can be done for us, using R.

Football or Rugby?

We can use an example to show how to use linear regression on real data. The example concerns attendances at Australian Rules Football matches. For all of you in the U.S., Australian Rules Football is closer to what you think of as rugby and not so much like American football. The data in this example concerns matches/games at the largest stadium for the sport, the Melbourne Cricket Ground (note: Australian Rules football is a winter sport, cricket is a summer sport). The Melbourne Cricket Ground or MCG is also considered the most prestigious ground to play a football match, due to its long association with Australian Rules Football.

Australian Rules Football is the most popular sport to have been developed in Australia. The object is to kick the ball through the larger goal posts, scoring six points. If the ball is touched before it crosses the line under the large posts, or instead passes through the smaller posts on either side of the large goal posts, a single point is scored. The rules ensure that possession of the ball is always changing. There is full body tackling and possession is turned over frequently. This leads to continuous and exciting on-field action.

The main stronghold of Australian Rules Football is in the state of Victoria (south eastern Australia), where the original league, the VFL, became the AFL after its league of mostly Melbourne suburban teams added teams to represent areas outside Victoria, such as West Coast (Perth, the capital of the state of Western Australia) and Adelaide (capital of the state of South Australia). Note that Melbourne is the capital city of Victoria. Much of the popularity of the VFL was based on the rivalries between neighboring suburbs, and teams with long histories, like the Collingwood Football Club - based in one of Melbourne's most working class suburbs – have large and loyal fan bases.

While it isn't necessary to know anything about how the sport is played to understand the example, it is useful to know that the Australian Football League, the premiere organization playing Australian Rules Football, was formerly the Victorian Football League, and although teams from outside the Australian state of Victoria have joined, more than half the teams in the league are Victorian, even though Victoria is only one of six Australian states.

Getting the Data

The data are available from OzDasl, a website which provides public domain data sets for analysis, and the MCG attendance data has its own page at http://www.statsci.org/data/oz/afl.html.

The variable of interest is MCG attendance, and named 'MCG' in the dataset. Most statisticians would refer to this variable as the dependent variable, because it is the variable that we are most interested in predicting: It is the "outcome" of the situation we are trying to understand. Potential explanatory, or independent, variables include club membership, weather on match day, date of match, etc. There is a detailed description of each of the variables available on the website. You can use the data set to test your own theories of what makes football fans decide whether or not to go to a game, but to learn some of the skills we will test a couple of those factors together.

Before we can start, we need R to be able to find the data. Make sure that you download the data to the spot on your computer that R considers the "working" directory. Use this command to find out what the current working directory is:

```
> getwd()
```

After downloading the data set from OzDasl into your R working directory, read the data set into R as follows:

```
> attend<-read.delim("afl.txt", header=TRUE)
```

```
> attach(attend)
```

We include the optional 'header = TRUE' to designate the first row as the column names, and the 'attach' commands turns each of the named columns into a single column vector.

Once we've read the data into R, we can examine some plots of the data. With many techniques in data science, it can be quite valuable to visualize the data before undertaking a more detailed analysis. One of the variables we might consider is the combined membership of the two teams playing, a proxy for the popularity of the teams playing.

```
> plot(MCG ~ Members, xlab = "Combined membership
of teams playing", ylab = "MCG Match Day Atten-
dance" )
```

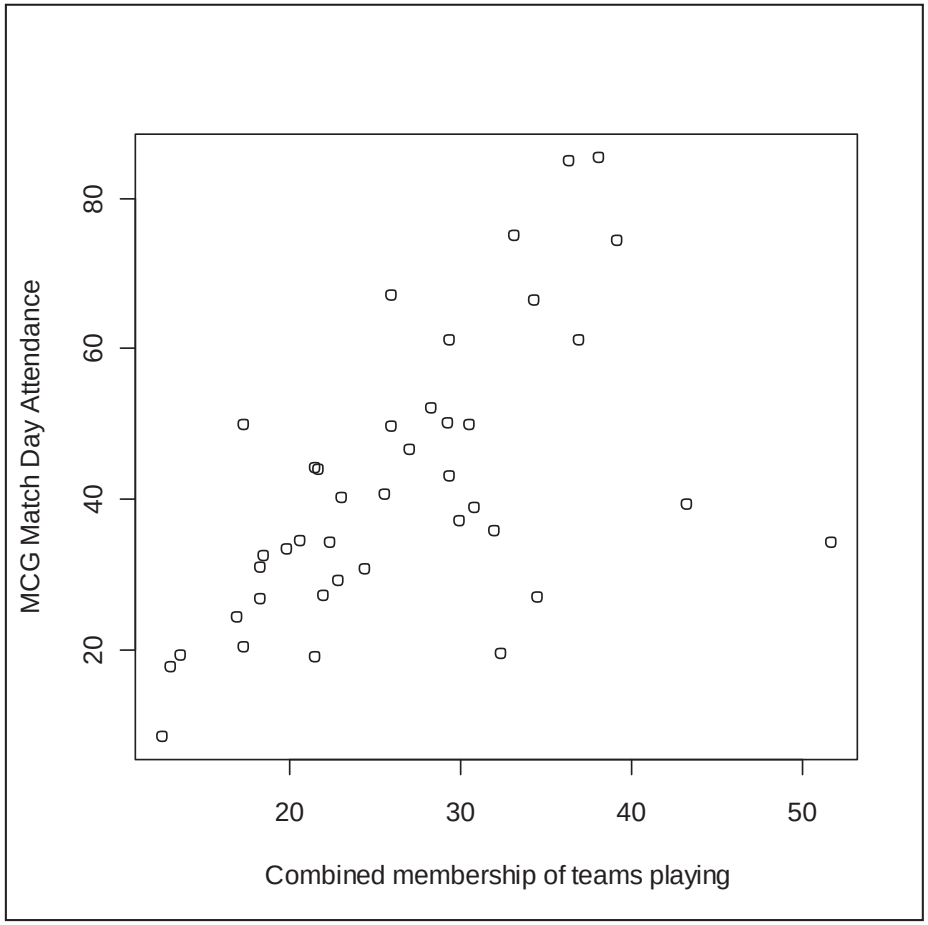


Figure 2: Scatterplot of membership versus attendance

We see evidence of a trend in the points on the left hand side of the graph, and a small group of points representing games with very high combined membership but that don't seem to fit the trend applying to the rest of data. If it wasn't for the four "outliers" on the right hand side of the plot, we would be left with a plot showing a very strong relationship.

As a next step we can use R to create a linear model for MCG attendance using the combined membership as the single explanatory variable using the following R code:

```
> model1 <- lm(MCG ~ Members-1)

> summary(model1)
```

There are two steps here because in the first step the lm() command creates a nice big data structure full of output, and we want to hang onto that in the variable called "model1." In the second command we request an overview of the contents of model1.

It is important to note that in this model, and in the others that follow, we have added a '-1' term to the specification, which forces the line of best fit to pass through zero on the y axis at zero on the x axis (more technically speaking, the y-intercept is forced to be at the origin). In the present model that is essentially saying that if the two teams are so unpopular they don't have any members, no one will go to see their matches, and vice versa. This technique is appropriate in this particular example, because both of the measures have sensible zero points, and we can logically reason that zero on X implies zero on Y. In most other models, particularly for survey data that may not have a sensible zero point (think about rating scales ranging from 1 to 5), it would not be appropriate to force the best fitting line through the origin.

The second command above, summary(model1), provides the following output:

```
Call:

lm(formula = MCG ~ Members - 1)


Residuals:

    Min      1Q  Median      3Q     Max

-44.961  -6.550   2.954   9.931  29.252


Coefficients:

        Estimate Std. Error t value Pr(>|t|)

Members  1.53610    0.08768   17.52   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1


Residual standard error: 15.65 on 40 degrees of
freedom

Multiple R-squared: 0.8847,

Adjusted R-squared: 0.8818

F-statistic: 306.9 on 1 and 40 DF,  p-value: <
2.2e-16
```

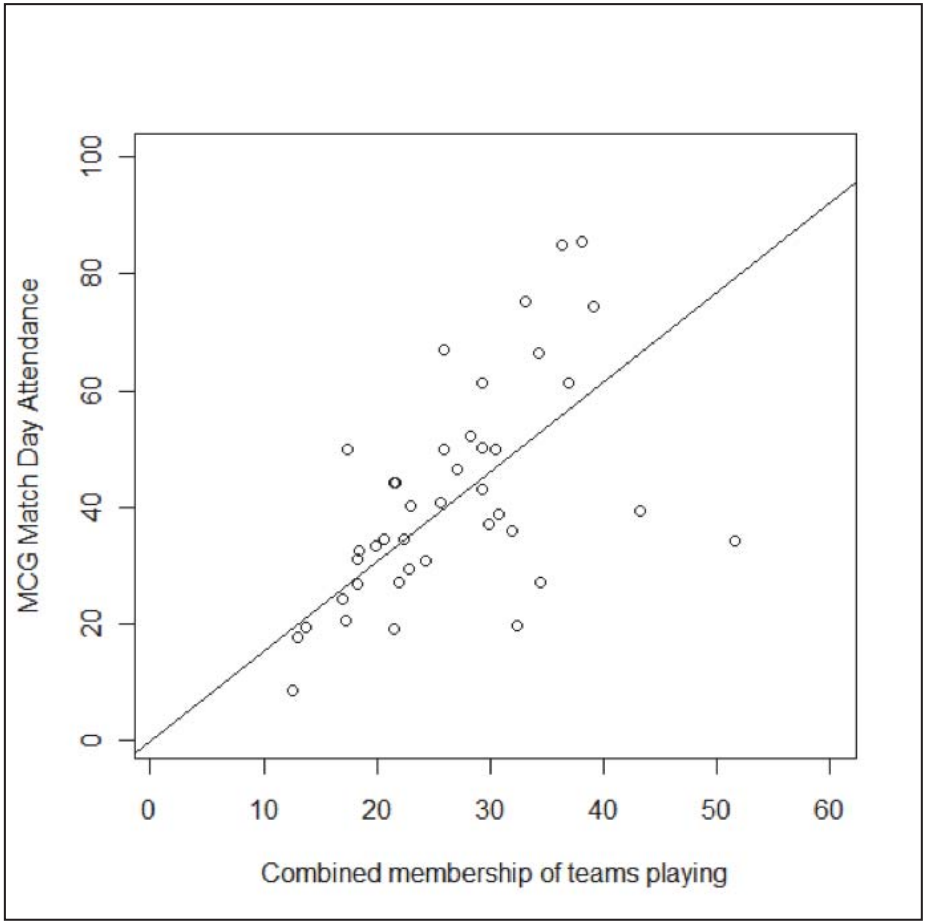
Wow! That's a lot information that R has provided for us, and we need to use this information to decide whether we are happy with this model. Being 'happy' with the model can involve many factors, and there is no simple way of deciding. To start with, we will look at the r-squared value, also known as the coefficient of determination.

The r squared value – the coefficient of determination – represents the proportion of the variation which is accounted for in the dependent variable by the whole set of independent variables (in this case just one independent variable). An r-squared value of 1.0 would mean that the X variable(s), the independent variable(s), perfectly predicted the y, or dependent variable. An r-squared value of zero would indicated that the x variable(s) did not predict the y variable at all. R-squared cannot be negative. The r-squared of .8847 in this example means that the Combined Members variable account for 88.47% of the MCG attendance variable, an excellent result. Note that there is no absolute rule for what makes an r-squared good. Much depends on the purpose of the analysis. In the analysis of human behavior, which is notoriously unpredictable, an r-squared of .20 or .30 may be very good.

In figure 3, below, we have added a line of best fit based on the model to the x-y plot of MCG attendance against total team membership with this command:

```
> abline(model1)
```

While the line of best fit seems to fit the points in the middle, the points on the lower right hand side and also some points towards the top of the graph, appear to be a long way from the line of best fit.



Adding Another Independent Variable

We discussed at the beginning of this chapter the origin of Australian Rules Football in Victoria, where the MCG is located. While most of the teams in the AFL are also Victoria teams, and therefore have a supporter base which can easily access the MCG, a number of the teams are from other states, and their supporters would

need to make a significant interstate journey to see their team play at the MCG. For example, the journey from Sydney to Melbourne is around eight hours by car or two by plane, whereas from Perth, where most West Coast's supporter base is located, is close to five hours by air – and two time zones away. Australia is a really huge country.

The dataset doesn't have a variable for interstate teams but fortunately there are only four teams that are interstate: Brisbane, Sydney, Adelaide, and West Coast, abbreviated respectively as "Bris", "Syd", "Adel", and "WC". We can make a binary coded variable to indicate these interstate teams with a simple command:

```
> away.inter <-
ifelse(Away=="WC" |
       Away=="Adel"|
       Away=="Syd"|
       Away=="Bris",1,0)
```

The code above checks the values in the column labeled 'Away', and if it finds an exact match with one of the names of an interstate team, it stores a value of 1. Otherwise it stores a value of 0. Note that we use a double equals sign for the exact comparison in R, and the vertical bar is used to represent the logical 'OR' operator. These symbols are similar, although not precisely the same, as symbols used to represent logical operators in programming languages such as C and Java. Having created the new 'Away team is interstate' variable, we can use this variable to create a new linear regression model that includes two independent variables.

```
> model2<-lm(MCG~Members+away.inter-1)
```

```
> summary(model2)


Call:

lm(formula = MCG ~ Members + away.inter - 1)


Residuals:

     Min       1Q   Median       3Q      Max

-30.2003  -8.5061   0.0862   8.5411  23.5687

Coefficients:

          Estimate Std. Error t value Pr(>|t|)

Members   1.69255    0.07962  21.257  < 2e-16 ***

away.inter -22.84122 5.02583  -4.545  5.2e-05 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1


Residual standard error: 12.82 on 39 degrees of
freedom

Multiple R-squared: 0.9246,
Adjusted R-squared: 0.9208

F-statistic: 239.2 on 2 and 39 DF,  p-value: <
2.2e-16
```

Note that the r-squared value is now 0.9246, which is quite a bit higher than the 0.8847 that we observed in the previous model. In this new model, the two independent variables working together account for 92.46% of the dependent variable. So together, the total fan base and the status as an away team are doing a really great job of predicting attendance. This result is also intuitive – we would expect that football fans, regardless of how devoted they are to their team, are more likely to come to games if they're a moderate car ride away, compared to a plane journey.

Because we have two independent variables now, we have to look beyond the r-squared value to understand the situation better. In particular, about one third of the way into the output for the lm() command there is a heading that says "Estimate." Right below that are slope values for Members and for away.inter. Notice that the slope (sometimes called a "B-weight") on Members is positive: This makes sense because the more fans the team has the higher the attendance. The slope on away.inter is negative because when this variable is 1 (in the case of interstate teams) the attendance is lower) whereas when this variable is 0 (for local teams), attendance is higher.

How can you tell if these slopes or B-weights are actually important contributors to the prediction? You can divide the unstandardized B-weight by its standard error to create a "t value". The lm() command has done this for you and it is reported in the output above. This "t" is the Student's t-test, described in a previous chapter. As a rule of thumb, if this t value has an absolute value (i.e., ignoring the minus sign if there is one) that is larger than about 2, you can be assured that the independent/predictor variable we are talking about is contributing to the prediction of the dependent variable. In this example we can see that Members has a humongous t value of 21.257, showing that it is very important in the prediction. The away.inter variable has a somewhat more modest, but still important value of -4.545 (again, don't worry about the minus sign when judging the magnitude of the t value).

We can keep on adding variables that we think make a difference. How many variables we end up using depends, apart from our ability to think of new variables to measure, somewhat on what we want to use the model for.

The model we have developed now has two explanatory variables – one which can be any positive number, and one which is two levels. We now have what could be considered a very respectable r-squared value, so we could easily leave well enough alone. That is not to say our model is perfect, however – the graphs we have prepared suggest that the 'Members' effect is actually different if the away team is from interstate rather than from Victoria – the crowd does not increase with additional combined membership as quickly with an away team, which is in line with what we might expect intuitively.

One thing we didn't mention was the actual prediction equation that one might construct from the output of lm(). It is actually very simple and just uses the estimates/B-weights from the output:

MCG = (21.257 * Members) - (4.545 * away.inter)

This equation would let us predict the attendance of any game with a good degree of accuracy, assuming that we knew the combined fan base and whether the team was interstate. Interestingly, statisticians are rarely interested in using prediction equations like

the one above: They are generally more interested in just knowing that a predictor is important or unimportant. Also, one must be careful with using the slopes/B-weights obtained from a linear regression of a single sample, because they are likely to change if another sample is analyzed - just because of the forces of randomness.

Conclusion

The material we have covered is really only a taste of multiple regression and linear modeling. On the one hand, there are a number of additional factors that may be considered before deciding on a final model. On the other hand, there are a great number of techniques that may be used in specialized circumstances. For example, in trying to model attendance at the MCG, we have seen that the standard model fits the data some of the time but not others, depending on the selection of the explanatory variables.

In general, a simple model is a good model, and will keep us from thinking that we are better than we really are. However, there are times when we will want to find as many dependent variables as possible. Contrast the needs of a manager trying to forecast sales to set inventory with an engineer or scientist trying to select parameters for further experimentation. In the first case, the manager needs to avoid a falsely precise estimate which could lead her to be overconfident in the forecast, and either order too much stock or too little. The manager wants to be conservative about deciding that particular variables make a difference to prediction variable. On the other hand the experimenter wants to find as many variables as possible for future research, so is prepared to be optimistic about whether different parameters affect the variables of interest.

Chapter Challenge

We intentionally ignored some of the output of these regression models, for the sake of simplicity. It would be quite valuable for you to understand those missing parts, however. In particular, we ignored the "p-values" associated with the t-tests on the slope/B-weight estimates and we also ignored the overall F-statistic reported at the very bottom of the output. There are tons of great recourses on the web for explaining what these are.

For a super bonus, you could also investigate the meaning of the "Adjusted" r-squared that appears in the output.

Sources

http://en.wikipedia.org/wiki/Australian_rules_football

http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html

http://www.ddiez.com/teac/r/linear_models.php

R Functions Used in This Chapter

abline - plots a best fitting line on top of a scatterplot

attach - makes a data structure the "focus" of attention

getwd - show the current working directory for R

ifelse - a conditional test that provides one of two possible outputs

lm - "linear models" and for this chapter, multiple regression

plot - general purpose graphing function, many uses in R

summary - produces an overview of an output structure