Overview

Welcome to ADA 2. The course will be similar to ADA1. We will use R to analyze data sets. The techniques used will sometimes be extensions from techniques in ADA1, and sometimes seem completely new.

Grades will be based on homeworks, two in-class tests, and a final project. The final project will be a data analysis that is presented to the class using slides, or if you wish to not present, you can turn in a poster as it might be presented at a conference. You can either work in small groups or as individuals for your project, and you can also use data from your research area in other departments. However, you must use techniques from the class. More details will be given about the final project as we progress through the semester.

Overview

The assessment will be broken down as follows:

- Homeworks, 50% (each homework will be equally weighted, and the lowest hw grade will be dropped)
- In-class tests, 15% each
- Final project, 20%

Topics in the course will include multiple regression (i.e., multiple predictor variables), model selection, experimental design, ANCOVA, logistic regression, principal component analysis, cluster analysis, and discriminant analysis. The last few examples are part of **multivariate analysis**.

Overview

We'll begin with a review of what occurred in ADA1 and a review of R, the software used for the course. Most students in the course will have had ADA1, but not all will have, and might need a review of R especially.

R review

To use R, we will usually read data from a file. For this class, these files will usually be either `.txt`, `.dat`, or `.csv`. The last version is a comma delimited file, where each field is separated by a comma instead of space or tabs. If you have data in an Excel spreadsheet, you can save it as a `.csv` file so that it can be read into R.

To read in data from a file, you can either type the website of the data (if you are online), or you can read the data from your computer. If you read data from your computer, you need to specify the path to the file, or have the file in the same directory as your R session.

R review

Here we'll give an example of reading in a file from my website

```
x <- read.table("http://www.math.unm.edu/~james/chile.txt",header=T)
```

If copying and pasting this code generates errors in R, then try typing the text directly inside R without copying and pasting (common problems are the quoation marks or the tilde sign copying and pasting differently). Also, try going to the website and see if you can download the data.
If the data exists in your computer in your current directory, you can type

```
x <- read.table("chile.txt",header=T)
```

R review

Here is an example of not reading in a file correctly. In this case, I didn't type `http://` at the start of the URL, and R was unable to find the file. This happens a lot when you either get the URL wrong or try to read in data that is mistakenly in a different directory from your R session.

```
> x <- read.table("www.math.unm.edu/~james/chile.txt")
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file ’www.math.unm.edu/~james/chile.txt’:
  No such file or directory
```

R review

Here we extract a variable and print a subset of the observations

```
> x$Length[x$group=="Chimayo"]
 [1] 14.0 15.5 12.5 16.0  8.5 12.5
  15.0 13.0 11.0 10.0 12.8
> x[x$Length>13,]
      group Length Width Thickness
14 Casados   13.5   3.5      1.55
15 Casados   14.0   3.0      1.77
16 Casados   15.0   3.0      1.59
18 Casados   13.5   3.0      1.58
21 Casados   14.0   4.0      1.99
22 Casados   13.3   3.3      1.71
23 Chimayo   14.0   3.5      1.80
24 Chimayo   15.5   3.5      1.81
26 Chimayo   16.0   3.5      1.82
29 Chimayo   15.0   3.5      1.95
```

R review

We'll continue the review using this data set. The data set consists of measurements of length, width, and thickness of green chiles cultivated at four locations in New Mexico in a particular year. There are a total of 44 observations (rows), and four variables (columns). Here are some things you can do with a data set that will be used to review some R functions

```
> head(x) # show the first 6 observations to check that
# data was read in correctly
> head(x)
    group Length Width Thickness
1 Alcalde   10.5   3.0      1.53
2 Alcalde    7.0   3.5      1.76
3 Alcalde   10.5   3.5      1.82
4 Alcalde   11.5   4.0      1.58
5 Alcalde   11.5   3.5      1.84
6 Alcalde    9.5   3.0      1.86
```

```
> table(x$group)
Alcalde Casados Chimayo Cochiti
     11      11      11      11
> mean(x$Length)
[1] 11.075
> sd(x$Width)
[1] 0.5032597
> by(x$Length,x$group,mean)
x$group: Alcalde
[1] 9.25
------------------------------------------------------------
x$group: Casados
[1] 13.3
------------------------------------------------------------
x$group: Chimayo
[1] 12.8
------------------------------------------------------------
x$group: Cochiti
[1] 8.95
```

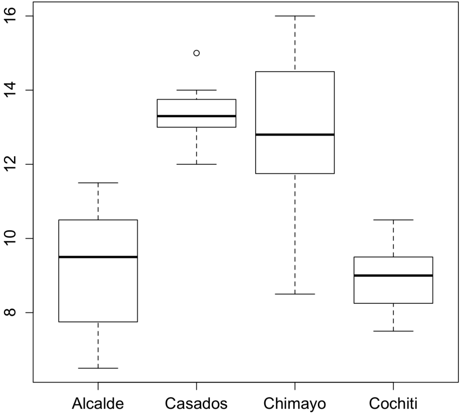
To use by() with combinations of variables, select a list of the columns used in the data set as follows: (here this is for columns 1 and 3):

```
> by(x$Length,x[,c(1,3)],mean)
group: Alcalde
Width: 2
[1] NA
--------------------------------------------------------------
group: Casados
Width: 2
[1] NA
--------------------------------------------------------------
group: Chimayo
Width: 2
[1] 12.5
--------------------------------------------------------------
group: Cochiti
Width: 2
[1] 9
```

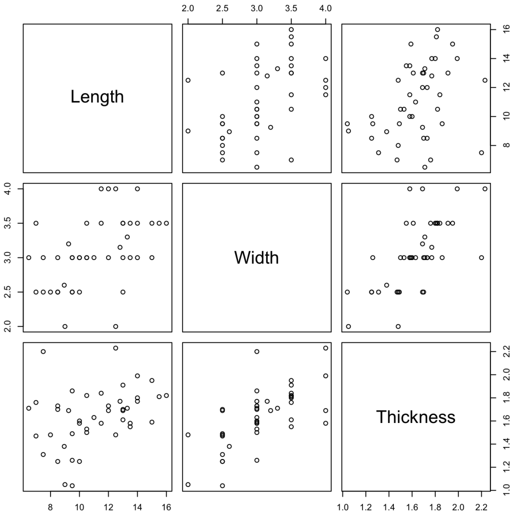
Here are some graphical ways to look at the data. Here cex.axis=1.3 increases the label sizes by 30% to make them more readable. The code pairs(x[,2:4]) means to make pairwise scatterplots for all rows, columns 2 through 4 of the data (so column 1 isn't included).

```
> boxplot(x$Length ~ x$group, cex.axis=1.3)
> pairs(x[,2:4])
> pairs(x[,-1]) # is equivalent
```

Boxplots



Boxplots


Length
Width
Thickness
2.0
2.5
3.0
3.5
4.0
8
10
12
14
16
1.0
1.2
1.4
1.6
1.8
2.0
2.2

R can also be used as a calculator, or to enter data directly by typing it in.

```
> a <- 2
> b <- c(4,5,7)
> a
[1] 2
> b
[1] 4 5 7
> a*b
[1] 8 10 14
> a+b
[1] 6 7 9
> b^2
[1] 16 25 49
> sqrt(b)
[1] 2.000000 2.236068 2.645751
> log(b)
[1] 1.386294 1.609438 1.945910
```

As a review of some of the statistical procedures we used last semester, here is a table

one sample t test	one group/population, quantitative data
matched pairs t test	one sample t-test on differences, quantitative data
two sample t-test	1 quantitative variable, 1 group variable (2 groups)
anova	quantitative response, 1 group variable with 2 or more
simple linear regression	1 predictor, 1 response, both quantiative
correlation test	2 quantitative variables that are paired, but not (e.g.,GPA and SAT score)
proportion test	binary variable (0 or 1), test whether frequency of 1s equal some proportion
2-sample proportion test	binary variable for two groups, test for equality of proportions
chi-square test (one way)	several categories, testing for equality of proportions or proportions matching theoretical values
chi-square test (two-way)	two table test of association, e.g., college major versus political affiliation

As a review of some of the statistical procedures we used last semester, here is a table

one sample t test	`t.test(x)`
matched pairs t test	`t.test(x-y)` or `diff <- x-y; t.test(diff)`
two sample t-test	`t.test(x,y)` or `t.test(x ∼ group)`
anova	`a <- aov(x ∼ group)` or `a <- lm(x ∼ group)`
simple linear regression	`model <- lm(y ∼ x)`
correlation test	`cor.test(x,y)`
proportion test	`prop.test(x,p=.5)`
2-sample proportion test	e.g., `prop.test(c(45,55),c(100,100))`
chi-square test (one way)	`chisq.test()`
chi-square test (two-way)	`chisq.test(M)` where `M` is a matrix

Many of the above procedures could be employed with the `chile.txt` data set. For example, we could

- use a t-test to see whether the length of the chile pods differs for Chimayo versus Cochiti
- use a t-test to see whether the width of the child pods differes for Chimayo versus Casados
- use ANOVA to test whether length is the same for all four types of chiles
- use regression to determine the relationship between length and width of child pods (either ignoring chile type or separately for each type)
- test the correlation of length and thickness
- Classify chiles as long verus short (say, greater than 11cm) and test whether the proportion of child pods that are long is greater than 50%. (proprtion test).
- test whehter the proportion of chile pods classified as long differs by location (this would be a one-way chi-squared)

In this exampe, a t-test is done for the Length of the Cochiti chiles. Note that the hypothesis test is testing whether the length is 0. The confidence interval is probably more useful.

```
> t.test(x$Length[x$group=="Cochiti"])

One Sample t-test

data:  x$Length[x$group == "Cochiti"]
t = 30.101, df = 10, p-value = 3.833e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 8.287493 9.612507
sample estimates:
mean of x
     8.95
```

```
> t.test(x$Length[x$group=="Cochiti"],
x$Length[x$group=="Alcalde"])

Welch Two Sample t-test

data:  x$Length[x$group == "Cochiti"] and x$Length[x$group == "Al
t = -0.48637, df = 15.546, p-value = 0.6335
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
 -1.610688  1.010688
sample estimates:
mean of x mean of y
     8.95      9.25
```

```
> a <- aov(x$Length ~ x$group)
> summary(a)
            Df Sum Sq Mean Sq F value  Pr(>F)
x$group      3  173.5   57.83   22.45 1.1e-08 ***
Residuals   40  103.0    2.58
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

logic and for-loops in R

```
> sum(Long)
[1] 25
> length(Long)
[1] 44
> prop.test(25,44)

1-sample proportions test with continuity correction

data:  25 out of 44, null probability 0.5
X-squared = 0.56818, df = 1, p-value = 0.451
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4114174 0.7131820
sample estimates:
        p
0.5681818
```

logic and for-loops in R

Although you can do a lot of statistics in R without logic and for-loop programming in R, it would be a shame to take this course without learning a little about logic control and for-loops. These are extremely general concepts used in programming that is ubiquitous in our lives (Google, Facebook, cell phones, etc.).

R can evaluate statements as true or false. These true and false values can then be converted into 1s (TRUEs) and 0s (FALSEs).

```
> x$Length>10
 [1]  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
[13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[25]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
[37] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
> as.numeric(x$Length>10)
 [1] 1 0 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1
[39] 0 0 0 0 0 0
```

for loops

As another example, suppose you have Likert scale data, such as 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=stronngly agree. Now suppose you want to transform the data so that you collapse categories, and just want 1=disagree, 2=neutral, 3=agree.

One way of processing the data is with a for loop. For example

```
> data <- c(5,3,4,3,1,2,1,5,4,3)
> newdata <- 1:length(data)
> for(i in 1:length(newdata)) {
+ if(data[i] <= 2) newdata[i] <- 1
+ if(data[i] == 3) newdata[i] <- 2
+ if(data[i] >= 4) newdata[i] <- 3
+ }
> newdata
 [1] 3 2 3 2 1 1 1 3 3 2
```

for loops

Another way of achieving the same result is

```
> data <- c(5,3,4,3,1,2,1,5,4,3)
> newdata <- 1*(data <= 2) + 2*(data==3) + 3*(data >= 4)
> newdata
 [1] 3 2 3 2 1 1 1 3 3 2
```

logic and for-loops in R

For loops allow you to process a data set one row at a time or one column at a time, or to do a repetetive thing one at a time.
A simple example of a for loop is

```
> for(i in 1:10) {
+ print(i^2)
+ }
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
[1] 49
[1] 64
[1] 81
[1] 100
```

logic and for-loops in R

Often in R you can avoid for loops by processing a vector all at once, such as testing whether each element is above or below a certain threshold. But sometimes it is useful to do things one at a time. For loops can also be useful for simulation. Here we simulate the probability of getting four aces in a poker hand. Cards are numbered 1 through 52, and I assume cards 1 through 4 are aces.

```
> cards <- 1:52
> fourAces <- 1:100000
> for(i in 1:100000) {
+ hand <- sample(cards,5,replace=TRUE)
+ num_aces = sum(hand <= 4)
+ fourAces[i] <- (num_aces == 4)
+ }
> sum(fourAces)/100000
[1] 0.00016 # estimated probability is 0.016%
```

Multiple Regression

This ends the review. For more review, please see slides from ADA 1.

We'll now begin the topic of Regression. The idea for multiple regression is similar to simple regression, but now there are multiple predictors. There is still a single response variable, which is assumed to be normally distributed for combination of predictors, and there can be any number of predictor variables. The predictor variables can be numeric, binary (0/1), or even categorical. We'll start with examples where all predictors are quantitative. Note that the predictors are not assumed to be normally distributed.

Multiple regression

We'll use the `chile.txt` data set as an example. Initially, we'll analyze all 44 chile peppers and not use their location, and we'll analyze the length of the chile pod as a function of the width and thickness. In this case there are two predictors and one response variable.

Note that this is an arbitrary choice on my part—I could have chosen to predict width based on length and thickness, thickness based on length and width, and so on. Which variable you want to consider a response and which the predictors can depend on your research question. In an experiment, you normally use as predictors those variables under control. For agricultural experiments, you might use as predictors the type of fertilizer (or nitrogen content in the fertilizer), the amount of water used, etc. as predictors, and crop yield would be the response.

Often predictors are not completely under experimental control, for example the amount of rainfall in plots in an agricultural experiment, the abundance of parasites, etc., might not be controlled but could still be relevant.

Multiple regression

We can think of the model as

$$\text{length} = \beta_0 + \beta_1 \times \text{width} + \beta_2 \times \text{thickness} + \text{error}$$

More informally, we might just write

$$\text{length} = \text{width} + \text{thickness} + \text{error}$$

where the addition sign isn't literal addition since there are regression coefficients involved.
The regression analysis will find the optimal values of β_0, β_1 and β_2 to minimize the sum of squared residuals.

Multiple regression

A more formal way to write the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where

- y_i is the length of the ith chile
- x_{1i} is the width of the ith chile
- x_{2i} is the thickness of the ith chile
- ε_i is the residual, or the deviation from the actual response to the mean response
- β_0 is the intercept (the predictred length when length and thickness are 0)
- β_1 is the coefficient for the length
- β_2 is the coefficient for the width

```
> m1 <- lm(x$Length ~ x$Width + x$Thickness)
> summary(m1)
Call:
lm(formula = x$Length ~ x$Width + x$Thickness)
Residuals:
    Min      1Q  Median      3Q     Max
-5.0991 -1.4546 -0.2972  1.5957  4.0967

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2619     2.4315   1.342   0.1871
x$Width       2.1134     0.9221   2.292   0.0271 *
x$Thickness   0.8183     1.8380   0.445   0.6585
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.282 on 41 degrees of freedom
Multiple R-squared:  0.2277,Adjusted R-squared:  0.1901
F-statistic: 6.045 on 2 and 41 DF,  p-value: 0.005003
```

Multiple regression

The output can be interpreted very similarly to simple linear regression. Here the column of estimated coefficients gives you the estimated values for β_0, β_1, and β_2. The fitted model is

$$\text{Expected length } = 3.2619 + 2.1134 * \text{width} + 0.8183 * \text{thickness}$$

This gives a formula for predicting the expected length of a new chile pod with a given width and thickness.

These estimated coefficients can be notated as b_0, b_1,and b_2, or as $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$. Putting a "hat" symbol over a parameter indicates that it is an estimate of the parameter. Thus for this example,

$$b_0 = \widehat{\beta}_0 = 3.2619$$
$$b_1 = \widehat{\beta}_1 = 2.1134$$
$$b_2 = \widehat{\beta}_2 = 0.8183$$

Multiple regression

As an example of using the regression equation, we might first note that the range of values for width and thickness is 2.0 to 3.0 for width and 1.04 to 2.20 for thickness. To predict the length of a new chile pod, we might consider only examples within those ranges. Otherwise, we risk extrapolation, which is dangerous to do in linear regression since the same relationships might not hold outside of the range of the data.

Suppose we want to predict the average elength of a chile pod with width of 2.5 and thicknss of 1.3. (We don't really need to worry about the units here, but I think they are in cm.)

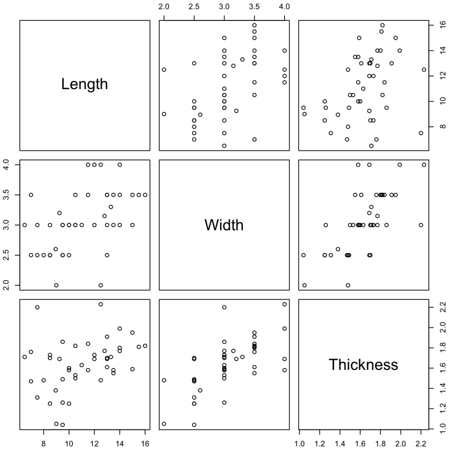
We can plug these values into the regression equation to get

$$\text{Expected length} = 3.2619 + 2.1134 * (2.5) + 0.8183 * (1.3) = 9.6$$

in cm.

Multiple regression

Note the positive correlation between length and width, and also length and thickness.



Multiple regression

Note the signs of the estimated coefficients. The estimate for β_1 is 1.837, which is positive. This means that for every 1 cm increase in width, when thickness is held constant, you expect an increase in length by 1.837 cm.

The estimate of β_2 is 0.8183, which means that for every unit increase in thickness (assuming width is held constant), you expect the length to increase by 0.8183 units.

Multiple regression

The difficulty with interpreting the regression coefficients in multiple regression is that they determine the effect of the variable when other variables are in the model. Thickness and width are not independent of one another (they are fairly strongly correlated), so if width is in the model, some of the effect of thickness is already taken into account. This makes it a little unpredictable what will happen when both thickness and width are in the model.

Multiple regression

We can imagine instead doing simple linear regressions of length against width or thickness.

```
> m1 <- lm(x$Length ~ x$Width + x$Thickness)
> m2 <- lm(x$Length ~ x$Width)
> m3 <- lm(x$Length ~ x$Thickness)
> m1$coefficients
(Intercept)     x$Width x$Thickness
  3.2619337   2.1134285   0.8182887
> m2$coefficients
(Intercept)     x$Width
   3.771047    2.384964
> m3$coefficients
(Intercept) x$Thickness
   5.169273    3.604549
```

Multiple regression

Each of these models gives different predictions for the length of a chile pod. The first model uses both width and thickness to make a prediction. The second model uses width but ignores thickness. And the third model uses thickness but ignores width. Because width and thickness are highly correlated, we might expect all three models to make similar predictions.

Generally, we prefer models with as few predictors as possible. There are several reasons for this. One is that when making predictions, it is cheaper (for example, it takes less time) to measure fewer variables. For the chile example, you might wonder if you can just use width to predict the length instead of width and thickness, for example. The topic of **model selection** deals with trying to determine which model is preferable when there is a choice of several.

Multiple regression

To continue with the example of a chile pod with width 2.5 and thickness 1.3, the three models predict the following lengths

```
> m1$coefficients
(Intercept)     x$Width x$Thickness
  3.2619337   2.1134285   0.8182887
> m2$coefficients
(Intercept)     x$Width
   3.771047    2.384964
> m3$coefficients
(Intercept) x$Thickness
   5.169273    3.604549
```

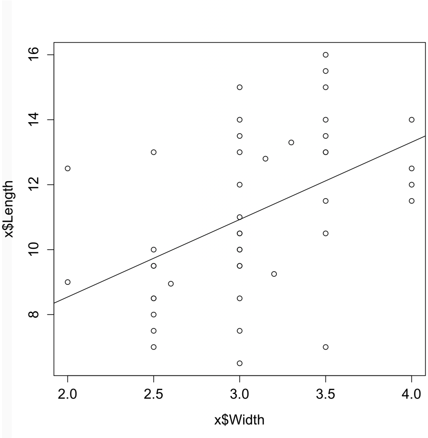
$$\begin{aligned}
m1 :& 3.2619 + 2.1134 * (2.5) + 0.8183 * (1.3) = 9.61\\
m2 :& 3.7710 + 2.3850 * (2.5) = 9.73\\
m3 :& 5.1693 + 3.6045 * (1.3) = 9.86
\end{aligned}$$

The three models give slightly different predictions for these values of width and thickness. You could also compare the fitted values, the predicted lengths on all the observed combinations of length and width:

```
> m1$fitted.values
      34       35       36       37       38       39       40
8.938373 9.975457 8.512369 8.536036 9.407451 9.080375 8.634964 9.5
      42       43       44
7.750766 9.080375 8.956455
> m2$fitted.values
      34       35       36       37       38       39       40
8.916667 9.083333 8.916667 8.916667 9.083333 8.916667 8.750000 8.9
      42       43       44
9.083333 8.916667 8.950000
> m1$fitted.values - m2$fitted.values
          34           35           36           37           38
 0.021706638  0.892123716 -0.404298094 -0.380631165  0.324117405
          40           41           42           43           44
-0.115035621  0.660713737 -1.332567666  0.163708216  0.006454617
```

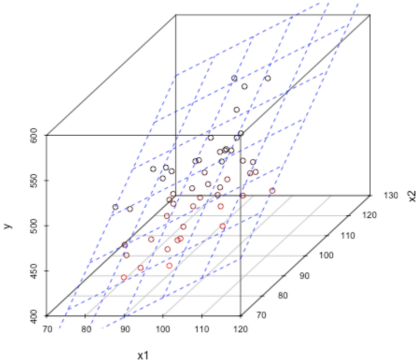
Visualizing multiple regression

Before trying to visualize multiple regression, recall simple linear regression, here from model m2.

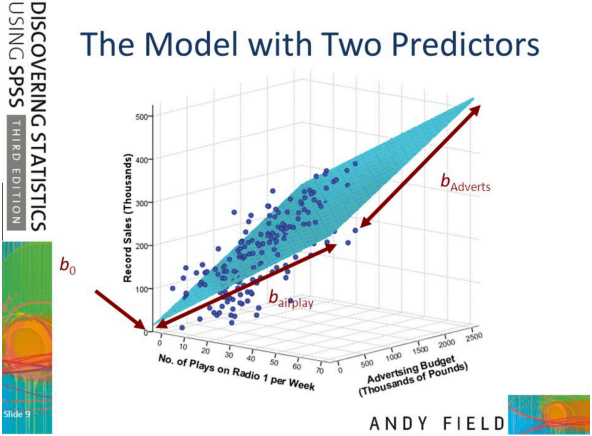


Visualizing multiple regression

source:
https://my.vertica.com/blog/machine-learning-series-linear-regression/



Visualizing multiple regression



The Model with Two Predictors

Multiple regression

Usually, we don't bother to make a 3D scatterplot with one response and two predictors. It is difficult to visualize, and the plot depends on the angle you use. Instead it is more common to just use a scatterplot matrix.

Most regression problems involve more than two predictors, and this gets even more difficult to visualize. If there are three predictors, the plot would have to be three dimensional. For p predictors, there would be $p + 1$ dimensions (the extra dimension is for the response).

With two predictors, the regression problem is to find the plane that minimizes the sum of squared distances from points to the plane. For more than three predictors, things get more abstract, and mathematically you are minimizing distances from points to "hyperplanes".

Multiple regression

One way of thinking about what occurs with multiple regression is that for each value of thickness, we get a different regression line for length versus width, and similarly, for each value of width, we get a different regression line for length versus width.

From the model

```
> m1$coefficients
(Intercept)     x$Width x$Thickness
  3.2619337   2.1134285   0.8182887
```

If we fix the thickness at 1.5, the model predicts

$$\begin{aligned}\text{Length} &= 3.2619 + (0.8183) * (1.5) + 2.1134 * \mathit{Width}\\ &= 4.4893 + 2.1134 * \mathit{Width}\end{aligned}$$

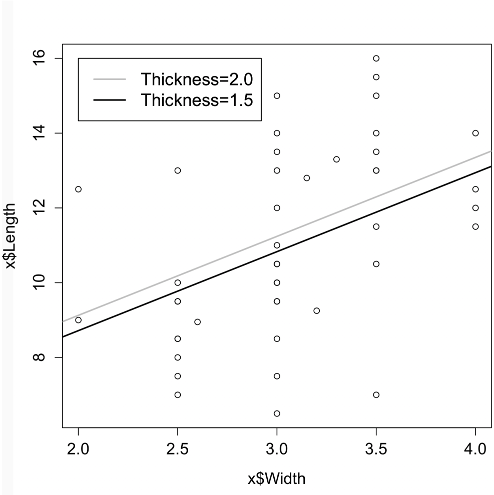
Multiple regression

At thickness equals 2.0, the model predicts

$$\begin{aligned}\text{Length} &= 3.2619 + (0.8183) * (2.0) + 2.1134 * \mathit{Width} \\ &= 4.8985 + 2.1134 * \mathit{Width}\end{aligned}$$

This is the same slope but a slightly different intercept.

Visualizing multiple regression



Multiple regression

Something to notice about the regression output is that R gives tests of significance for the individual predictors. The column of p-values is giving tests for whether or not the coefficients β_0, β_1, and β_2 are equal to 0 or not.

Usually whether or not $\beta_0 = 0$ is not as interesting–this is whether the intercept is 0. What is more interesting is whether β_1 or β_2 is equal to 0. If one of these is equal to 0, that means that the independent variable does not significantly improve the ability to predict the response variable. For the chile example, width appears to significantly predict length, but thickness does not, at least when both variables are in the model. We usually use $p \leq .05$ to indicate statistical significance.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2619     2.4315   1.342   0.1871
x$Width       2.1134     0.9221   2.292   0.0271 *
x$Thickness   0.8183     1.8380   0.445   0.6585
Multiple R-squared:  0.2277,Adjusted R-squared:  0.1901
```

Multiple regression

Since thickness does not seem to be significant when width is in the model, we might think about dropping thickness from the model. Note that doing so will change the estimate of the effect of width and it's p-value.

```
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.771      2.125   1.774  0.08324 .
x$Width        2.385      0.685   3.482  0.00118 **
---
Residual standard error: 2.26 on 42 degrees of freedom
Multiple R-squared:  0.224,Adjusted R-squared:  0.2055
```

Note that the p-value has decreased from 0.027 to 0.001.

Multiple regression

Since thickness had a p-value that was greater than .05 (actually about 0.66) when width and thickness were both in the model, can we conclude that thickness is not significantly associated with length? We can fit the model with just thickness as a predictor to check.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.169      2.397   2.156   0.0368 *
x$Thickness    3.605      1.447   2.492   0.0167 *
---
Residual standard error: 2.395 on 42 degrees of freedom
Multiple R-squared:  0.1288,Adjusted R-squared:  0.108
```

Note that when width is not in the model, thickness actually is significantly associated with length (p=.0167). However, when width is in the model, thickness is not adding much extra information. Essentially, you can think of thickness as being redundant when width is in the model. This is partly because thickness and width are correlated. They are not giving independent information about length. However, width seems to be giving somewhat more information than thickness.

Multiple regression

The topic of model selection—which we'll get into more—deals with choosing which of these multiple models to prefer. Usually we prefer to have fewer predictors if the extra predictors do not provide much extra information. In this case, it would be usual to use width but not thickness as a predictor.

Later on, we'll look at more formal methods for model selection.

Multiple regression

Something that is frequently used in the output is the R-squared (R^2) value. In simple linear regression (i.e., one predictor), R^2 is the square of the correlation between the response and the predictor. The mutliple R^2 is the square of the correlation between the response and the fitted values, i.e., between y and $\widehat{y}$. For example,

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2619     2.4315   1.342   0.1871
x$Width       2.1134     0.9221   2.292   0.0271 *
x$Thickness   0.8183     1.8380   0.445   0.6585

Residual standard error: 2.282 on 41 degrees of freedom
Multiple R-squared:  0.2277,Adjusted R-squared:  0.1901

> cor(m1$fitted.values,x$Length)^2
[1] 0.2277329
```

Multiple regression

R^2 is also interpreted as the amount of variability in the response that is "accounted for", or predicted by the model. In other words, there is variability in the length of the chile peppers. Some of the variability is associated with the width and thicknesses, but even taking these into account, there is still additional variability.

Generally, the more predictor variables you have, the more variability in the response you can predict, so the higher R^2 becomes. However, adding more variables sometimes barely increases R^2, indicating that adding a new variable doesn't necessarily contribute much useful information.

Multiple regression

Because R^2 always increases with more variables, using (multple) R^2 by itself doesn't necessarily indicate the best model. Another approach is to use **adusted** R^2, which penalizes for the number of predictor variables. This is also in the R output. You don't need the formula for this, but it is

$$R^2_{adj} = 1 - (1 - R^2)\left[\frac{n-1}{n-k-1}\right]$$

The adjusted R^2 is based on R^2, but is modified depending on the overall sample size, n, and the number of predictors, k. The adjusted R^2 value does not necessarily increase when you add more predictors. Note that if n is much larger than k, then the adjusted R^2 is very close to the multiple R^2.

Multiple regression

For the chile pepper example, we get

model	R^2	adjusted R^2
width+thickness	0.2277	0.1901
width	0.2240	0.2055
thickness	0.1288	0.1080

Note that the model with width but not thickness has the highest adjusted R^2, and that its multiple R^2 value is not much lower than the model with both predictors.