Regressions tutorial

Grusha Prasad

July 15, 2018

What is linear regression?

Linear regression allows us to describe the relationship between a dependent variable (y) and one or more independent variables (x). Unlike with correlation where we are interested in the extent to which x and y are related to each other, with regression we are specifically interested in predicting y from x. Given this goal, people often talk about regression in terms of causation. For example consider the following equation:

$$y = 3x + 4$$

One way to describe this model: The model predicts that one unit increase in x causes/ is associated with 3 units increase in y. Therefore the coefficient of x is the degree to which x can impact y. This description has a causal flavour because it describes it as a difference within an individual or a group (i.e. how would the response of an indvidual change if they were given the 'treatment'). However Gelman and Hill argue that we should be thinking of regressions as a difference between individuals and groups.

> "Linear regression is a method that summarizes how the average values of a numerical outcome variable vary over subpopulations defined by linear functions of predictors." (pg 31)

Another way to describe the model: The model predicts that two groups that have a one unit difference in x on average tend to have two units difference in y. The coefficient of x is predicted *average difference* in y for groups that vary in x.

Brief description of the dataset

```r
data(iris)
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
##                 
```

There are 4 continuous variables and one categorical variable. In the remainder of this document we will predict the Petal.Length from the other variables.

Categorical predictors

One predictor

Let us start by looking at just two species.

```
two_species <- subset(iris, Species != 'virginica')
two_species$Species <- factor(two_species$Species)  #removes ghost levels
lm(two_species$Petal.Length ~ two_species$Species)
```

```
##
## Call:
## lm(formula = two_species$Petal.Length ~ two_species$Species)
##
## Coefficients:
##                   (Intercept)  two_species$Speciesversicolor
##                         1.462                          2.798
```

In order to understand what the intercept and the coefficients mean, it is important to understand how the contrasts are coded for the variable

```
contrasts(two_species$Species)
```

```
##            versicolor
## setosa              0
## versicolor          1
```

As a default, R uses dummy coding - which means that it treats one of the levels as a "baseline" and compares all the other levels with this baseline. (It picked 'setosa' as the basline because the levels were organized alphabetically)

With dummy coding with one predictor, the intercept is the mean petal length of the baseline category (i.e. setosa)

```
mean(subset(two_species, Species == 'setosa')$Petal.Length)
```

```
## [1] 1.462
```

The coefficient is the mean difference in petal length between baseline cateogry and the category it is being compared to.

```
mean(subset(two_species, Species == 'versicolor')$Petal.Length - subset(two_species, Species == 'setosa
```

```
## [1] 2.798
```

The same idea applies when there are more than two levels to a predictor. There are if a variable has k levels, there are k-1 coefficients to estimate

```
lm(iris$Petal.Length ~ iris$Species)
```

```
##
## Call:
## lm(formula = iris$Petal.Length ~ iris$Species)
##
## Coefficients:
##            (Intercept)  iris$Speciesversicolor   iris$Speciesvirginica
##                  1.462                   2.798                   4.090
```

```
contrasts(iris$Species)
```

```
##            versicolor virginica
## setosa              0         0
## versicolor          1         0
## virginica           0         1
```

The intercept is again the mean of the baseline (setosa)

```
mean(subset(iris, Species == 'setosa')$Petal.Length)
```

```
## [1] 1.462
```

The coefficients are the mean difference between the baseline and the cateogory that is being compared to the baseline

```
mean(subset(iris, Species == 'setosa')$Petal.Length - subset(iris, Species == 'versicolor')$Petal.Lengt
```

```
## [1] -2.798
```

```
mean(subset(iris, Species == 'setosa')$Petal.Length - subset(iris, Species == 'virginica')$Petal.Length
```

```
## [1] -4.09
```

Note, the dummy contrasts assumes we have a baseline level that we can compare the other levels with. This might be useful when thinking about treatment groups and control groups. However this is not always useful for other kinds of categorical variables. Instead we can use summed contrasts - which will allow us to compare the means for groups with the grand mean.

```
contrasts(two_species$Species) <- "contr.sum"
contrasts(two_species$Species)
```

```
##            [,1]
## setosa        1
## versicolor   -1
```

```
lm(two_species$Petal.Length ~ two_species$Species)
```

```
##
## Call:
## lm(formula = two_species$Petal.Length ~ two_species$Species)
##
## Coefficients:
##          (Intercept)  two_species$Species1
##                2.861                -1.399
```

The intercept (or the baseline we are comparing the group means to) in this case is the grand mean of petal length across species

```
mean(two_species$Petal.Length)
```

```
## [1] 2.861
```

The coefficient in this case is how much the average petal length of each species varies from the grand mean.

```
mean(two_species$Petal.Length) - mean((subset(two_species, Species == 'versicolor')$Petal.Length))
```

```
## [1] -1.399
```

```
mean(two_species$Petal.Length) - mean((subset(two_species, Species == 'setosa')$Petal.Length))
```

```
## [1] 1.399
```

When there are two conditions this is just half of the distance between the conditions we get with the 0,1 dummy coding. So if we wanted to have the same effect size we could set the contrasts to be -0.5 and 0.5

instead of 1 and -1.

Similarly looking at summed contrasts for three levels.

```
contrasts(iris$Species) <- "contr.sum"
contrasts(iris$Species)
```

```
##            [,1] [,2]
## setosa        1    0
## versicolor    0    1
## virginica    -1   -1
```

```
lm(iris$Petal.Length ~ iris$Species)
```

```
##
## Call:
## lm(formula = iris$Petal.Length ~ iris$Species)
##
## Coefficients:
##   (Intercept)  iris$Species1  iris$Species2
##         3.758         -2.296          0.502
```

```
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

```
mean(iris$Petal.Length) - mean((subset(iris, Species == 'setosa')$Petal.Length))
```

```
## [1] 2.296
```

```
mean(iris$Petal.Length) - mean((subset(iris, Species == 'versicolor')$Petal.Length))
```

```
## [1] -0.502
```

Note though this doesn't directly tell us value for virgincia, this should be negative sum of the other two. -(-2.296 + 0.502) = 1.794 link

Two predictors

Let us start by looking at two categorical predictors and without summed contrasts.

```
two_species <- subset(iris, Species != 'virginica')
two_species$Species <- factor(two_species$Species)
two_species$Sepal.Length.cat <- factor(ifelse(two_species$Sepal.Length > mean(two_species$Sepal.Length)

contrasts(two_species$Species)
```

```
##            versicolor
## setosa              0
## versicolor          1
```

```
#contrasts(two_species$Sepal.Length.cat) <- "contr.sum"
contrasts(two_species$Sepal.Length.cat)
```

```
##       short
## long      0
## short     1
```

```
lm(Petal.Length ~ Species + Sepal.Length.cat, data = two_species)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Species + Sepal.Length.cat, data = two_species)
##
## Coefficients:
##          (Intercept)     Speciesversicolor  Sepal.Length.catshort
##               1.8163                2.4909                -0.3937
```

Intercept is the baseline for both predictors (so setosa with long sepals). The coefficient for Species is the expected value when you keep the Sepal.length constant. The coefficient for Sepal.Length is the expected value when you keep the Species constant.

```
mean(subset(two_species, Species == 'setosa' & Sepal.Length.cat == 'long')$Petal.Length)  #Why is this
```

```
## [1] 1.42
```

```
short.versicolor.setosa <- mean(subset(two_species, Species == 'versicolor' & Sepal.Length.cat == 'shor

long.versicolor.setosa <- mean(subset(two_species, Species == 'versicolor' & Sepal.Length.cat == 'long')

mean(c(short.versicolor.setosa, long.versicolor.setosa))
```

```
## [1] 2.52447
```

```
setosa_short.long <- mean(subset(two_species, Species == 'setosa' & Sepal.Length.cat == 'short')$Petal.

versicolog_short.long <- mean(subset(two_species, Species == 'versicolor' & Sepal.Length.cat == 'short')

mean(c(setosa_short.long, versicolog_short.long))
```

```
## [1] -0.3611364
```

These values are not exactly identical to the parameters of the model but they are close. Is this because of error? — I think it needs to be weighted mean.

READ THIS: http://genomicsclass.github.io/book/pages/interactions_and_contrasts.html

According to this link the main effects should be something else: https://stats.stackexchange.com/questions/120030/interpretation-of-betas-when-there-are-multiple-categorical-variables/120035#120035

But that is not consistent with what I find.

```
mean(subset(two_species, Sepal.Length.cat == 'short' & Species == 'setosa')$Petal.Length) - mean(subset
```

```
## [1] 0.04666667
```

```
mean(subset(two_species, Sepal.Length.cat == 'long' & Species == 'versicolor')$Petal.Length) - mean(sub
```

```
## [1] 2.932273
```

```
mean(subset(two_species, Sepal.Length.cat == 'short')$Petal.Length) - mean(subset(two_species, Sepal.Le
```

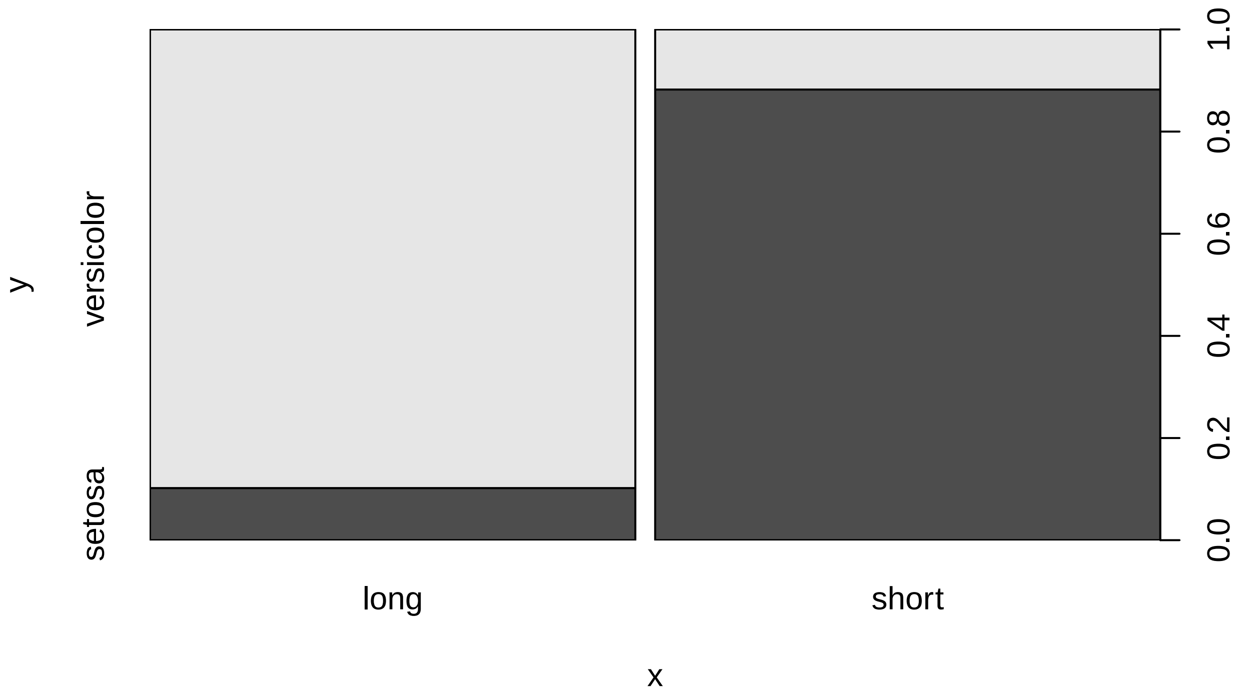
```
## [1] -2.337375
```

```
mean(subset(two_species, Species == 'versicolor')$Petal.Length) - mean(subset(two_species, Species == '
```

```
## [1] 2.798
```

Note when we run these predictors in an individual model, the coefficients of both the models are larger. But when they are in a model together, species is a much stronger predictor than sepal.length.cat. This suggests that most of the effect for sepal.length.cat is being driven by the confounding species difference. That is why we need an interaction.

```
plot(two_species$Sepal.Length.cat,two_species$Species)
```



But before that let us look at the model with summed contrasts

```
contrasts(two_species$Species) <- "contr.sum"
contrasts(two_species$Sepal.Length.cat) <- "contr.sum"
lm(Petal.Length ~ Species + Sepal.Length.cat, data = two_species)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Species + Sepal.Length.cat, data = two_species)
##
## Coefficients:
##      (Intercept)          Species1  Sepal.Length.cat1
##           2.8649           -1.2455             0.1968
```

```
contrasts(two_species$Sepal.Length.cat)
```

```
##       [,1]
## long     1
## short   -1
```

The intercept is the grand mean. Species is ??? Sepal.length.cat is ???

```
mean(two_species$Petal.Length)
```

```
## [1] 2.861
```

```
mean(two_species$Petal.Length) - mean(subset(two_species, Species == 'versicolor')$Petal.Length)
```

```
## [1] -1.399
```

```
mean(two_species$Petal.Length) - mean(subset(two_species, Sepal.Length.cat == 'short')$Petal.Length)
```

```
## [1] 1.145314
```

Interactions

```r
two_species <- subset(iris, Species != 'virginica')
two_species$Species <- factor(two_species$Species)
two_species$Sepal.Length.cat <- factor(ifelse(two_species$Sepal.Length > mean(two_species$Sepal.Length)

contrasts(two_species$Species)
```

```
##            versicolor
## setosa              0
## versicolor          1
```

```r
lm(Petal.Length ~ Sepal.Length.cat * Species, data = two_species)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length.cat * Species, data = two_species)
##
## Coefficients:
##                             (Intercept)
##                                 1.42000
##                    Sepal.Length.catshort
##                                 0.04667
##                       Speciesversicolor
##                                 2.93227
## Sepal.Length.catshort:Speciesversicolor
##                                -0.81561
```

Intercept: Both reference groups (i.e. long and setosa)

Sepal.Length.cat : The difference between short and long for the reference/ baseline species group - i.e. setosa Species: The difference between setosa and versicolor for the reference/ baseline sepal.length.cat group - i.e. long

So these cannot be the main effects.

Interaction: short versicolor - (long setosa + (short setosa - long setosa) + (long versicolor - long setosa))

= short versicolor + long setosa - short setosa - long versicolor = (short versicolor + long setosa) - (short setosa + long versicolor) = (short versicolor - short setosa) + (long setosa - long versicolor)

```r
mean(subset(two_species,Sepal.Length.cat == 'long' & Species == 'setosa')$Petal.Length)
```

```
## [1] 1.42
```

```r
mean(subset(two_species, Sepal.Length.cat == 'short' & Species == 'setosa')$Petal.Length) - mean(subset
```

```
## [1] 0.04666667
```

```r
mean(subset(two_species, Sepal.Length.cat == 'long' & Species == 'versicolor')$Petal.Length) - mean(subs
```

```
## [1] 2.932273
```

```r
mean(subset(two_species,Sepal.Length.cat == 'short' & Species == 'versicolor')$Petal.Length) - mean(subs
```

```
## [1] -0.8156061
```

Useful links: https://stats.stackexchange.com/questions/122246/interpretation-of-interaction-term/122251#122251

Questions: Should you always have an interaction term??

With summed contrasts:

```
contrasts(two_species$Sepal.Length.cat) <- "contr.sum"
contrasts(two_species$Species) <- "contr.sum"

lm(Petal.Length ~  Species * Sepal.Length.cat, data = two_species)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Species * Sepal.Length.cat, data = two_species)
##
## Coefficients:
##               (Intercept)                     Species1
##                    2.7056                      -1.2622
##         Sepal.Length.cat1  Species1:Sepal.Length.cat1
##                    0.1806                      -0.2039
```

```
contrasts(two_species$Sepal.Length.cat)
```

```
##       [,1]
## long     1
## short   -1
```

```
contrasts(two_species$Species)
```

```
##            [,1]
## setosa        1
## versicolor   -1
```

```
mean(two_species$Petal.Length)
```

```
## [1] 2.861
```

```
mean(two_species$Petal.Length) - mean(subset(two_species, Sepal.Length.cat == 'short')$Petal.Length)
```

```
## [1] 1.145314
```

```
mean(two_species$Petal.Length) - mean(subset(two_species, Species == 'versicolor')$Petal.Length)
```

```
## [1] -1.399
```

```
mean(two_species$Petal.Length) - mean(subset(two_species, Sepal.Length.cat == 'long' & Species == 'setos
```

```
## [1] 1.441
```

Interaction: short versicolor - (grand mean + (short setosa - grand mean) + (long versicolor - grand mean))

Continuous predictor

```
lm(Petal.Length ~ Sepal.Length, data = iris)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length, data = iris)
##
## Coefficients:
##  (Intercept)  Sepal.Length
##       -7.101         1.858
```

```r
lm(Petal.Length ~ Sepal.Length, data = two_species)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length, data = two_species)
##
## Coefficients:
##  (Intercept)  Sepal.Length
##       -7.180         1.835
```

This says that when there is a one unit difference in the sepal length between two irises, on average the difference in petal length is going to be 1.86. Note since this is an expected difference, we can't actually get the number directly from our data by subtracting the mean petal length of two irises with one unit difference in petal length.

```r
unique(iris$Sepal.Length)
```

```
##  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.4 4.8 4.3 5.8 5.7 5.2 5.5 4.5 5.3 7.0 6.4
## [18] 6.9 6.5 6.3 6.6 5.9 6.0 6.1 5.6 6.7 6.2 6.8 7.1 7.6 7.3 7.2 7.7 7.4
## [35] 7.9
```

```r
mean(subset(iris, Sepal.Length == 5.4)$Petal.Length - subset(iris, Sepal.Length == 4.4)$Petal.Length)
```

```
## [1] 0.7
```