Lecture 4 Multiple Regression

Ahmed Nadeem

2024-10-03

Multiple Regression

We're back to simulating the predictor variables directly, and then taking the response from those - this is a more typical way to simulate!

QUESTION: What parameters are you imagining in your simulations?

```r
library(MASS) ### this might be a package you need to install

##simulate multivariate normal data - this means that our parameters are related to each other
sigma <- matrix(c(1, 0.2, 0.1,
                       0.2, 1, 0.15,
                       0.1, 0.15, 1), nrow = 3) ## positive-definite symmetric matrix specifying the co

### This line gives us 3 parameters - how many do we want?
predictors<-mvrnorm(1000,rep(0,3), sigma) ###we are simulating the predictors here. What have we told t
cor.test(predictors[,1], predictors[,2])
```

```
##
##  Pearson's product-moment correlation
##
## data:  predictors[, 1] and predictors[, 2]
## t = 6.2111, df = 998, p-value = 7.712e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1325064 0.2518954
## sample estimates:
##       cor
## 0.1929148
```

QUESTION: How does changing sigma on line 22 change the correlations on line 28?

QUESTION: What response variable are you imagining? How is it distributed? QUESTION: What relationship are you simulating between your predictor variables and your response variable? QUESTION: How much error are you simulating? How much irreducible error do you expect in your response variable of interest?

```r
response<-1.5*predictors[,1]+0.75*predictors[,2]+0.05*predictors[,3]+rnorm(1000, 0, 1) ### change this 

cbind.data.frame(response, predictors)->data
names(data)<-c("tarsus_length", "food", "weather", "clouds") ### PUT INFORMATIVE NAMES HERE FOR WHAT YO
summary(data)
```

```
##  tarsus_length           food              weather               clouds
##  Min.   :-8.34813   Min.   :-3.42519   Min.   :-2.997460   Min.   :-3.25235
##  1st Qu.:-1.65779   1st Qu.:-0.71536   1st Qu.:-0.646392   1st Qu.:-0.72470
```

```
##  Median :-0.06497   Median :-0.03876   Median : 0.004969   Median :-0.02073
##  Mean   :-0.10740   Mean   :-0.06125   Mean   : 0.014677   Mean   :-0.02137
##  3rd Qu.: 1.31593   3rd Qu.: 0.61825   3rd Qu.: 0.750020   3rd Qu.: 0.62458
##  Max.   : 6.48741   Max.   : 3.21723   Max.   : 3.891603   Max.   : 3.25869
```

An assumption of multiple regression is that there isn't multicolinarity. We are breaking this assumption.

QUESTION: Explore what happens to your R^2 when you change the strength of correlation between the paramters.

```
lm(tarsus_length~food+weather+clouds, data=data)->model1
summary(model1)
```

```
##
## Call:
## lm(formula = tarsus_length ~ food + weather + clouds, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50218 -0.65198 -0.00443  0.61741  3.04026
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03100    0.02999  -1.034   0.3015
## food         1.41189    0.02986  47.289   <2e-16 ***
## weather      0.78642    0.02981  26.382   <2e-16 ***
## clouds       0.06839    0.03056   2.238   0.0254 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.946 on 996 degrees of freedom
## Multiple R-squared:  0.7851, Adjusted R-squared:  0.7844
## F-statistic:  1213 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: tarsus_length
##            Df  Sum Sq Mean Sq   F value  Pr(>F)
## food        1 2586.24 2586.24 2889.6395 < 2e-16 ***
## weather     1  665.62  665.62  743.7065 < 2e-16 ***
## clouds      1    4.48    4.48    5.0089 0.02544 *
## Residuals 996  891.43    0.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

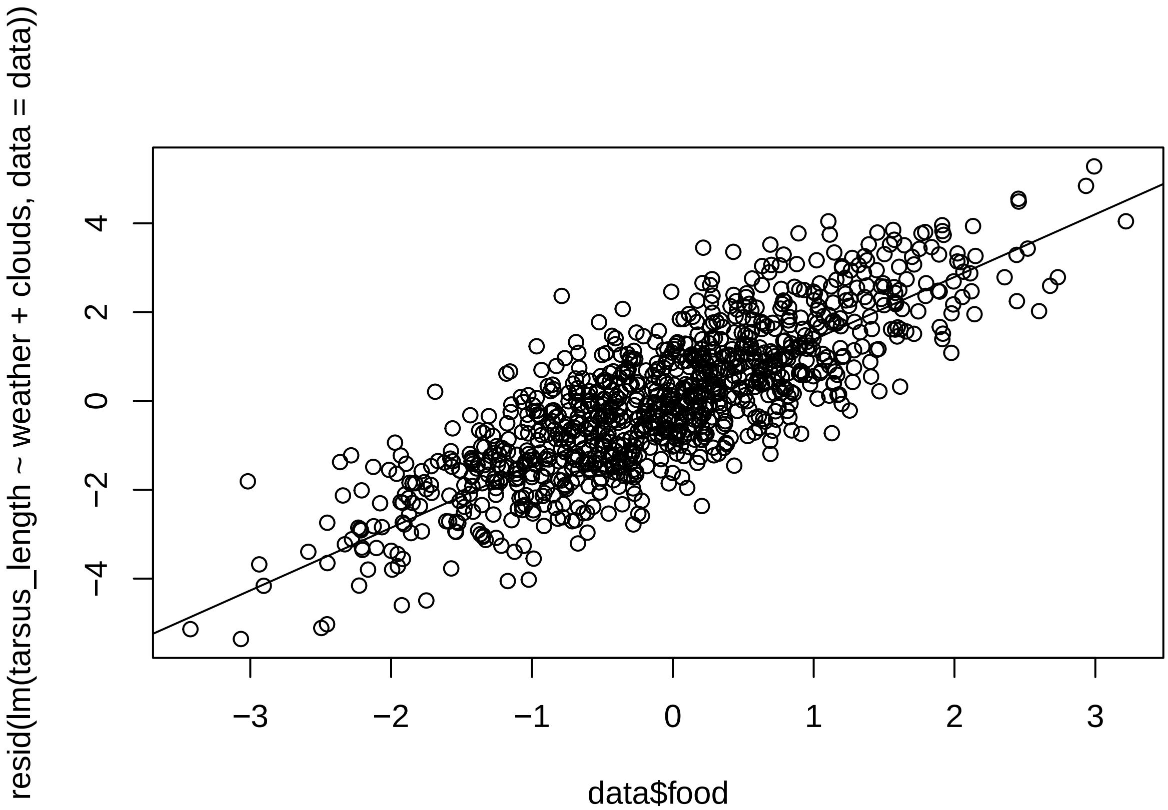
There are a couple of ways to plot multiple regressions. Some folks like 3D plots, so they can see the plan of the regression incorporating two predictor variables. This isn't my favourite, as it only works for two variables.

Here, I'm going to plot using partial residuals - essentially, we're going to plot the relationship of one variable after we've accounted for the other two.

```
plot(data$food, resid(lm(tarsus_length~weather+clouds, data=data))) ### Do this for each of the three p
abline(model1)
```

```
## Warning in abline(model1): only using the first two of 4 regression
```

```
## coefficients
```



QUESTION: Report the results of your multiple regression as you would in a results section of a paper.

```r
lm(tarsus_length~food*weather*clouds, data=data)->model2
summary(model2)
```

```
##
## Call:
## lm(formula = tarsus_length ~ food * weather * clouds, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44346 -0.66857 -0.00603  0.61241  2.98348
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.03960    0.03096  -1.279   0.2012
## food                 1.40903    0.03067  45.947   <2e-16 ***
## weather              0.78774    0.02990  26.344   <2e-16 ***
## clouds               0.06574    0.03119   2.108   0.0353 *
## food:weather         0.02765    0.02991   0.924   0.3555
## food:clouds          0.06462    0.03098   2.085   0.0373 *
## weather:clouds      -0.01405    0.02939  -0.478   0.6328
## food:weather:clouds  0.01913    0.03045   0.628   0.5300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9447 on 992 degrees of freedom
## Multiple R-squared:  0.7865, Adjusted R-squared:  0.785
## F-statistic: 522.2 on 7 and 992 DF,  p-value: < 2.2e-16
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: tarsus_length
##                      Df  Sum Sq Mean Sq   F value  Pr(>F)
## food                  1 2586.24 2586.24 2897.6834 < 2e-16 ***
## weather               1  665.62  665.62  745.7767 < 2e-16 ***
## clouds                1    4.48    4.48    5.0228 0.02524 *
## food:weather          1    1.65    1.65    1.8456 0.17460
## food:clouds           1    3.83    3.83    4.2895 0.03861 *
## weather:clouds        1    0.22    0.22    0.2427 0.62235
## food:weather:clouds   1    0.35    0.35    0.3947 0.52999
## Residuals           992  885.38    0.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

QUESTION: Do you have evidence of an interaction? How does an interaction change the p-values or R^2? Why is this the case?

An interaction is when the relationship between the response variable and a predictor variable changes depending on an other predictor variable.

QUESTION: Think of an interaction term!

```r
predictor_4<-rnorm(1000,0,1)

response2<-1.5*predictors[,1]+ifelse(predictors[,1]>0.4, 0.5, 0)*predictor_4+rnorm(1000, 0, 1) ### make

cbind.data.frame(data, response2, predictor_4)->data2

names(data2)<-c("tarsus_length", "food", "weather", "clouds", "fat_reserves", "density")
```

QUESTION: Describe in words what your interaction term is doing.

Below, I am testing a 4 way interaction. Explain why this might or might not be appropriate for your own simulated data.

```r
lm(fat_reserves~food*weather*clouds*density, data=data2)->model3
summary(model3)
```

```
##
## Call:
## lm(formula = fat_reserves ~ food * weather * clouds * density,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.78875 -0.70615 -0.00516  0.67672  3.04752
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.042892   0.033654  -1.274 0.202789
```

```
## food                         1.505935   0.033578  44.849  < 2e-16 ***
## weather                      0.027982   0.032506   0.861 0.389541
## clouds                      -0.025693   0.034176  -0.752 0.452368
## density                      0.156477   0.034334   4.558 5.83e-06 ***
## food:weather                 0.052169   0.032806   1.590 0.112110
## food:clouds                 -0.006892   0.034448  -0.200 0.841468
## weather:clouds               0.036986   0.032598   1.135 0.256815
## food:density                 0.124556   0.035965   3.463 0.000557 ***
## weather:density             -0.014340   0.033603  -0.427 0.669663
## clouds:density              -0.076900   0.035895  -2.142 0.032411 *
## food:weather:clouds          0.023650   0.035692   0.663 0.507733
## food:weather:density         0.057805   0.035557   1.626 0.104339
## food:clouds:density         -0.003113   0.035328  -0.088 0.929792
## weather:clouds:density      -0.001037   0.033489  -0.031 0.975303
## food:weather:clouds:density -0.001442   0.039053  -0.037 0.970549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 984 degrees of freedom
## Multiple R-squared:  0.7071, Adjusted R-squared:  0.7026
## F-statistic: 158.4 on 15 and 984 DF,  p-value: < 2.2e-16
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: fat_reserves
##                              Df  Sum Sq Mean Sq   F value    Pr(>F)
## food                          1 2422.51 2422.51 2321.9380 < 2.2e-16 ***
## weather                       1    0.60    0.60    0.5758 0.4481494
## clouds                        1    1.01    1.01    0.9671 0.3256543
## density                       1   26.13   26.13   25.0495 6.617e-07 ***
## food:weather                  1    4.82    4.82    4.6224 0.0318008 *
## food:clouds                   1    0.02    0.02    0.0203 0.8867404
## weather:clouds                1    1.42    1.42    1.3655 0.2428746
## food:density                  1   12.68   12.68   12.1509 0.0005123 ***
## weather:density               1    0.92    0.92    0.8850 0.3470547
## clouds:density                1    5.02    5.02    4.8074 0.0285710 *
## food:weather:clouds           1    0.44    0.44    0.4191 0.5175150
## food:weather:density          1    2.94    2.94    2.8160 0.0936487 .
## food:clouds:density           1    0.01    0.01    0.0108 0.9172638
## weather:clouds:density        1    0.00    0.00    0.0012 0.9724606
## food:weather:clouds:density   1    0.00    0.00    0.0014 0.9705493
## Residuals                   984 1026.62    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

QUESTION: Write a statistical methods paragraph and a statistical results paragraph. Make sure that everything in your statistical methods is reported in your results!