Homework 1: Matrix Formulation and Maximum Likelihood

STAT 340: Applied Regression

Iris Data set

You may recall the famous iris data set, which was collected by Edgar Anderson in the 1930s to study the relationship between iris species (`setosa`, `versicolor`, `virginica`) and measurement variables: sepal length, sepal width, petal length, and petal width. **We would like to see if the three species have different sepal widths, on average.**

(1) Assuming that the necessary conditions are satisfied, write down the relevant linear model element-wise (the way you would have written this model before being introduced to matrix notation for linear models). The iris data set is built into R, so you may consult that for the relevant number of observations. Be sure to define what your explanatory variable(s) (e.g. x_i is only useful if you tell me what x_i represents).

```
data(iris)
```

(2) Translate the model from (1) into matrix form, showing your steps.

Note: If you are typing this, you may find the following shorthand useful.

- $\mathbf{1}_n$ is an $n \times 1$ vector of 1's
- $\mathbf{0}_n$ is an $n \times 1$ vector of 0's
- You can use dots so you don't have to type every element, as long as the structure is clear. For example,

$$\mathbf{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\epsilon^2 & \rho & \cdots & \rho \\ \rho & \sigma_\epsilon^2 & \cdots & \rho \\ \vdots & \rho & \ddots & \vdots \\ \rho & \rho & \cdots & \sigma_\epsilon^2 \end{bmatrix}$$

(3) Find X using R. You will find the function `model.matrix()` useful - see the help documentation (`?model.matrix`) for details. Store the result of `model.matrix()` as X.

```r
X <- model.matrix(Sepal.Width ~ Species, data=iris)
str(X)
```

```
##  num [1:150, 1:3] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:150] "1" "2" "3" "4" ...
##   ..$ : chr [1:3] "(Intercept)" "Speciesversicolor" "Speciesvirginica"
##  - attr(*, "assign")= int [1:3] 0 1 1
##  - attr(*, "contrasts")=List of 1
##   ..$ Species: chr "contr.treatment"
```

```r
class(X)
```

```
## [1] "matrix"
```

```r
solve(t(X)%*%X)%*%t(X)%*%iris$Sepal.Width
```

```
##                     [,1]
## (Intercept)        3.428
## Speciesversicolor -0.658
## Speciesvirginica  -0.454
```

(4) Find the maximum likelihood estimates for the coefficients in your model using

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

(5) Fit the model using the `lm()` function to verify your result in (4). If the results do not match, check your work in (2) and (3).

```r
summary(lm(Sepal.Width ~ Species, data=iris))
```

```
## 
## Call:
## lm(formula = Sepal.Width ~ Species, data = iris)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -1.128 -0.228  0.026  0.226  0.972 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        3.42800    0.04804  71.359  < 2e-16 ***
## Speciesversicolor -0.65800    0.06794  -9.685  < 2e-16 ***
## Speciesvirginica  -0.45400    0.06794  -6.683 4.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3397 on 147 degrees of freedom
## Multiple R-squared:  0.4008, Adjusted R-squared:  0.3926 
## F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```