Q1. #EMLR 3.2 a-f 2. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded and can be found in the turtle dataset.

```
library(faraway)
library(ggplot2)
library(dplyr)
```

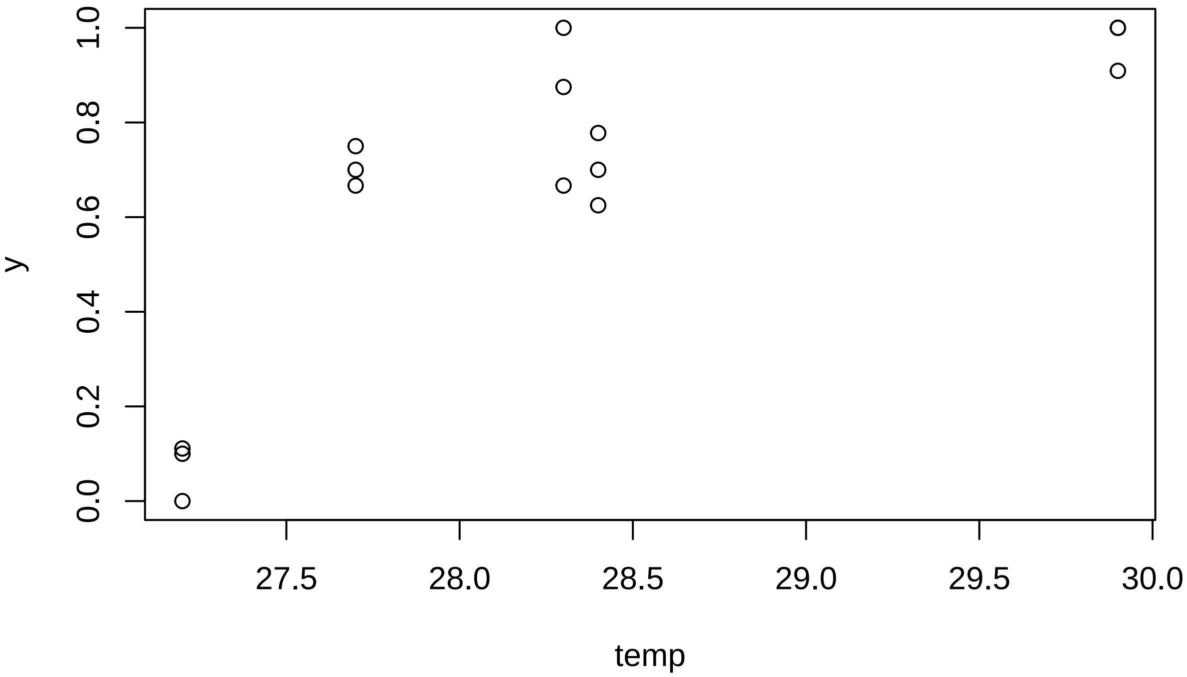
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data(turtle, package = "faraway")
head(turtle)
```

```
##   temp male female
## 1 27.2    1      9
## 2 27.2    0      8
## 3 27.2    1      8
## 4 27.7    7      3
## 5 27.7    4      2
## 6 27.7    6      2
```

(a) Plot the proportion of males against the temperature. Comment on the nature of the relationship.

```
turtle$y <- ifelse(turtle$female==0,1,(turtle$male)/(turtle$male+turtle$female))
attach(turtle, warn.conflicts = FALSE)
plot(x=temp, y)
```

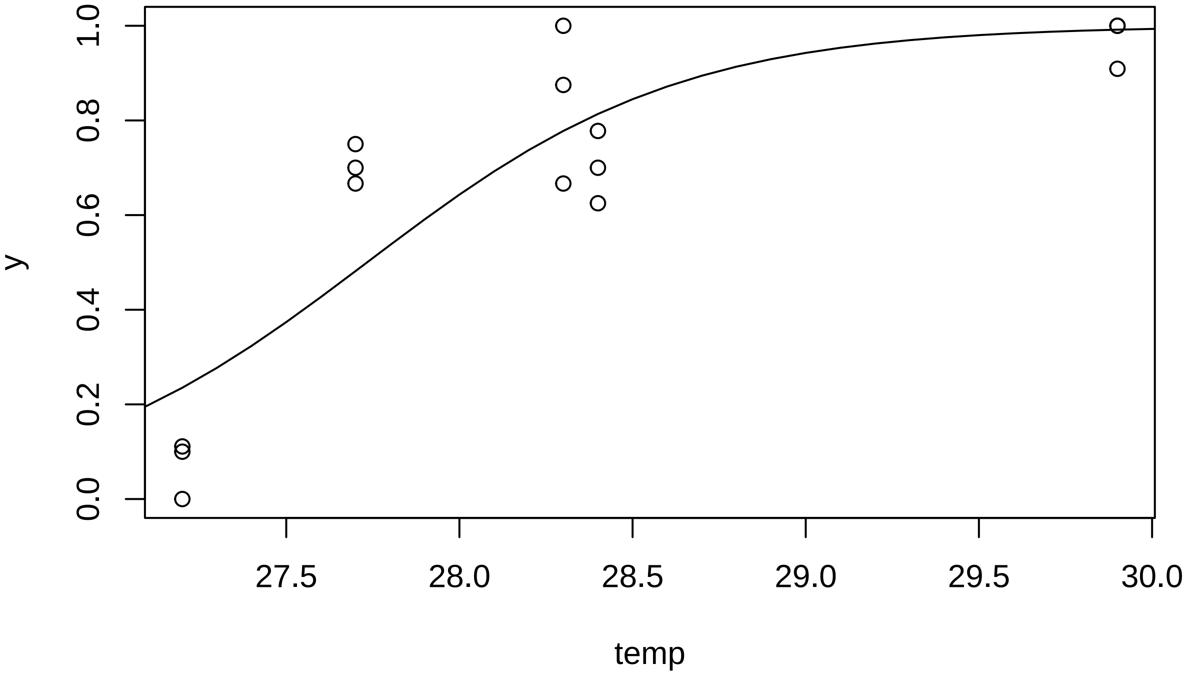


It shows the increasing relationship between temp and y. (b) Fit a binomial response model with a linear term in temperature. Does this model fit the data?

```
tMod <- glm(cbind(male, female) ~temp, family="binomial")
summary(tMod)
```

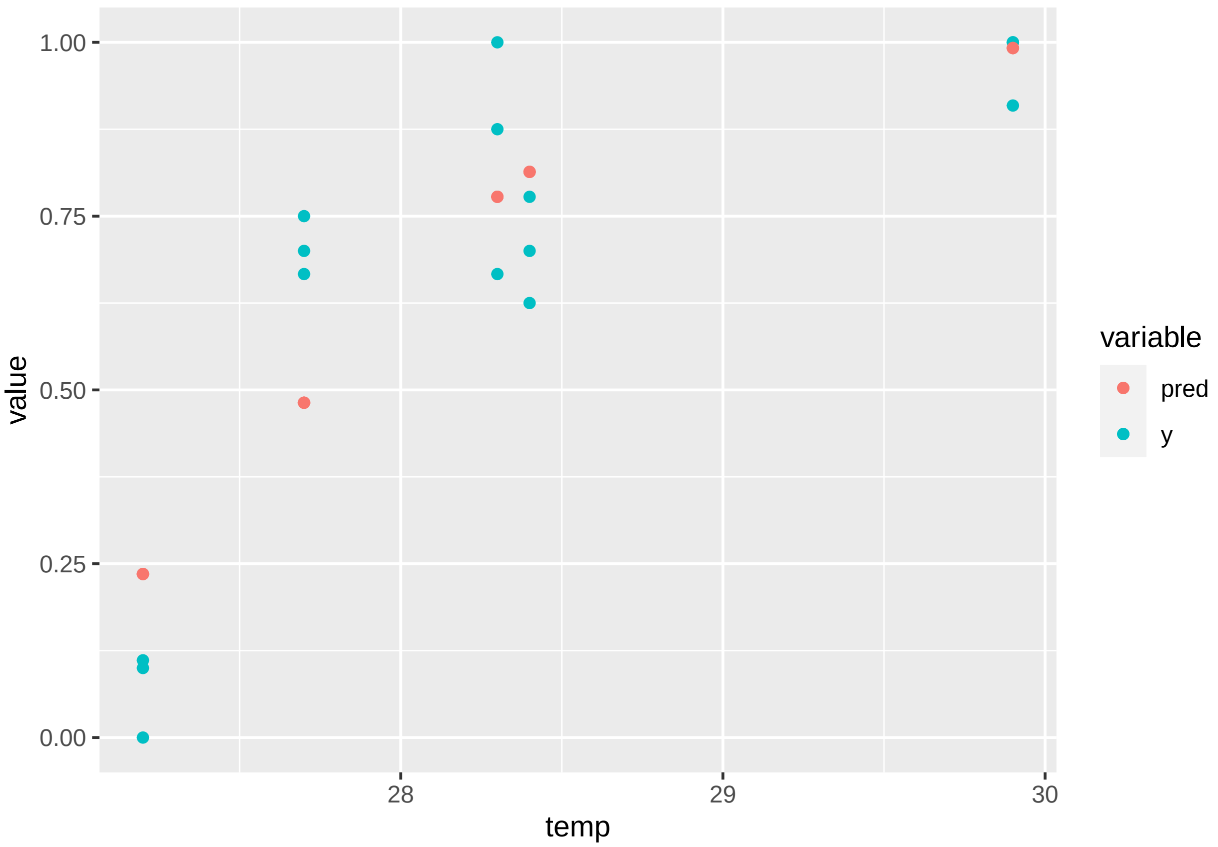
```
##
## Call:
## glm(formula = cbind(male, female) ~ temp, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0721  -1.0292  -0.2714   0.8087   2.5550
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
## temp          2.2110     0.4309   5.132 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 24.942  on 13  degrees of freedom
## AIC: 53.836
##
## Number of Fisher Scoring iterations: 5
```

```
x <- seq(27,30.2,0.1)
plot(x=temp, y)
lines(x, ilogit(-61.3183+(2.2110*x)))
```



```
turtle$pred <- ilogit(-61.3183+(2.2110*temp))
attach(turtle, warn.conflicts = FALSE)
```

```r
ggplot(data=turtle, aes(y=value,x=temp,colour = variable)) +
    geom_point(aes(y=y, col="y")) +
    geom_point(aes(y=pred, col="pred"))
```



It fits well since the significance level is small enough for each coefficient and p-value is smaller than .05.

(c) Is this data sparse? Since the number of observations is 15, yes, there is main one predictor as ratio of male/female.

(d) Check for outliers. There is no strong outliers from the above plot.

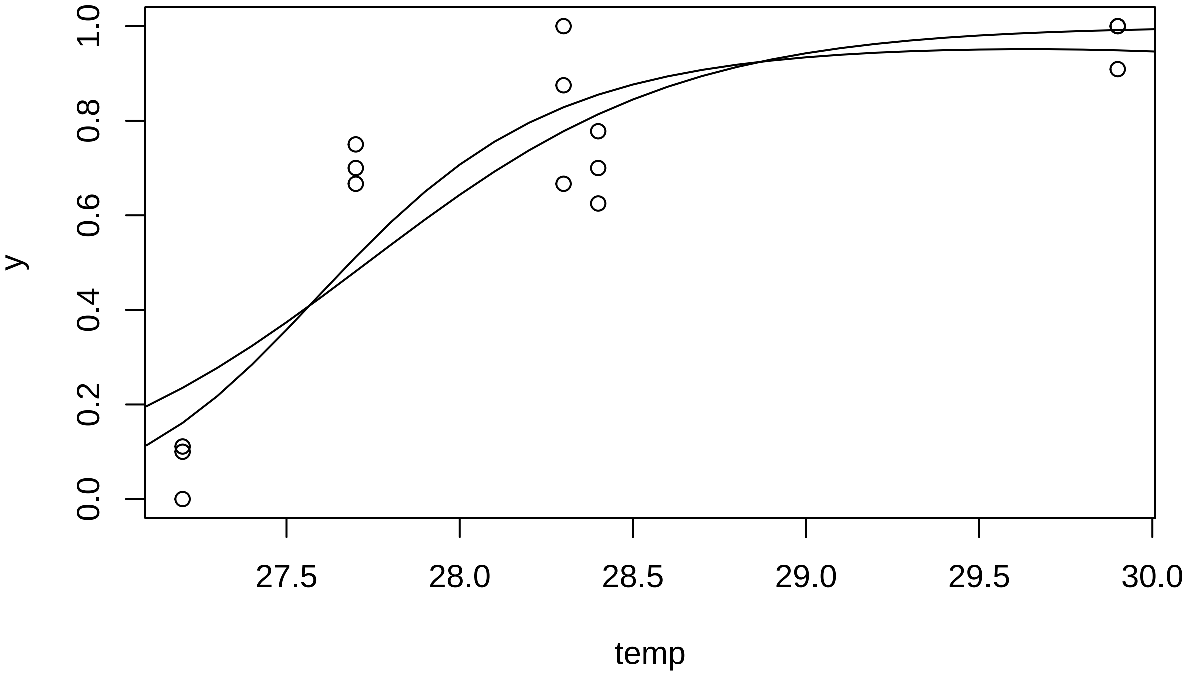
(e) Compute the empirical logits and plot these against temperature. Does this indicate a lack of fit? From the empirical logits and the plot above in (b), they show that the fitting can be improved to be better.

(f) Add a quadratic term in temperature. Is this additional term a significant predictor of the response. Does the quadratic model fit the data?

```r
tMod <- glm(cbind(male, female) ~temp + I(temp^2), family="binomial")
summary(tMod)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp + I(temp^2), family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6703  -0.8875  -0.4194   0.9481   2.2198
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -677.5950   268.7984  -2.521   0.0117 *
## temp          45.9173    18.9169   2.427   0.0152 *
## I(temp^2)     -0.7745     0.3327  -2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 20.256  on 12  degrees of freedom
## AIC: 51.15
##
## Number of Fisher Scoring iterations: 4
```

```
plot(x=temp, y)
lines(x, ilogit(-61.3183+(2.2110*x)))
lines(x, ilogit(-677.595 +(45.9173*x)- (.7745 * x^2)))
```



Yes.

The square term x^2 has the p-value as 0.0199, which is less than .05. So, this is significant to use. From using the plotting, we can see that the more closer modeling can be possible.

Q2. #EMLR 3.3

A biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. Four boxes were covered and four left uncovered. The process was repeated at six different moisture levels

```
library(faraway)
data(seeds)
head(seeds)
```

```
##   germ moisture covered
## 1   22        1      no
## 2   41        3      no
## 3   66        5      no
## 4   82        7      no
## 5   79        9      no
## 6    0       11      no
```

(a) Plot the germination percentage against the moisture level on two side-by-side plots according to the coverage of the box. What relationship do you see?

```
seeds$c01 <- factor(ifelse(seeds$covered == "yes", 1, 0))
seeds
```

```
##    germ moisture covered c01
## 1    22        1      no   0
## 2    41        3      no   0
## 3    66        5      no   0
## 4    82        7      no   0
## 5    79        9      no   0
## 6     0       11      no   0
## 7    25        1      no   0
## 8    46        3      no   0
## 9    72        5      no   0
## 10   73        7      no   0
## 11   68        9      no   0
## 12    0       11      no   0
## 13   27        1      no   0
## 14   59        3      no   0
## 15   51        5      no   0
## 16   73        7      no   0
## 17   74        9      no   0
## 18    0       11      no   0
## 19   23        1      no   0
## 20   38        3      no   0
## 21   78        5      no   0
## 22   84        7      no   0
## 23   70        9      no   0
## 24    0       11      no   0
## 25   45        1     yes   1
## 26   65        3     yes   1
## 27   81        5     yes   1
## 28   55        7     yes   1
## 29   31        9     yes   1
## 30    0       11     yes   1
## 31   41        1     yes   1
## 32   80        3     yes   1
## 33   73        5     yes   1
## 34   51        7     yes   1
## 35   36        9     yes   1
## 36    0       11     yes   1
## 37   42        1     yes   1
## 38   79        3     yes   1
## 39   74        5     yes   1
## 40   40        7     yes   1
## 41   45        9     yes   1
## 42    0       11     yes   1
## 43   43        1     yes   1
## 44   77        3     yes   1
## 45   76         5     yes   1
## 46   62         7     yes   1
## 47   NA         9     yes   1
## 48    0        11     yes   1
```

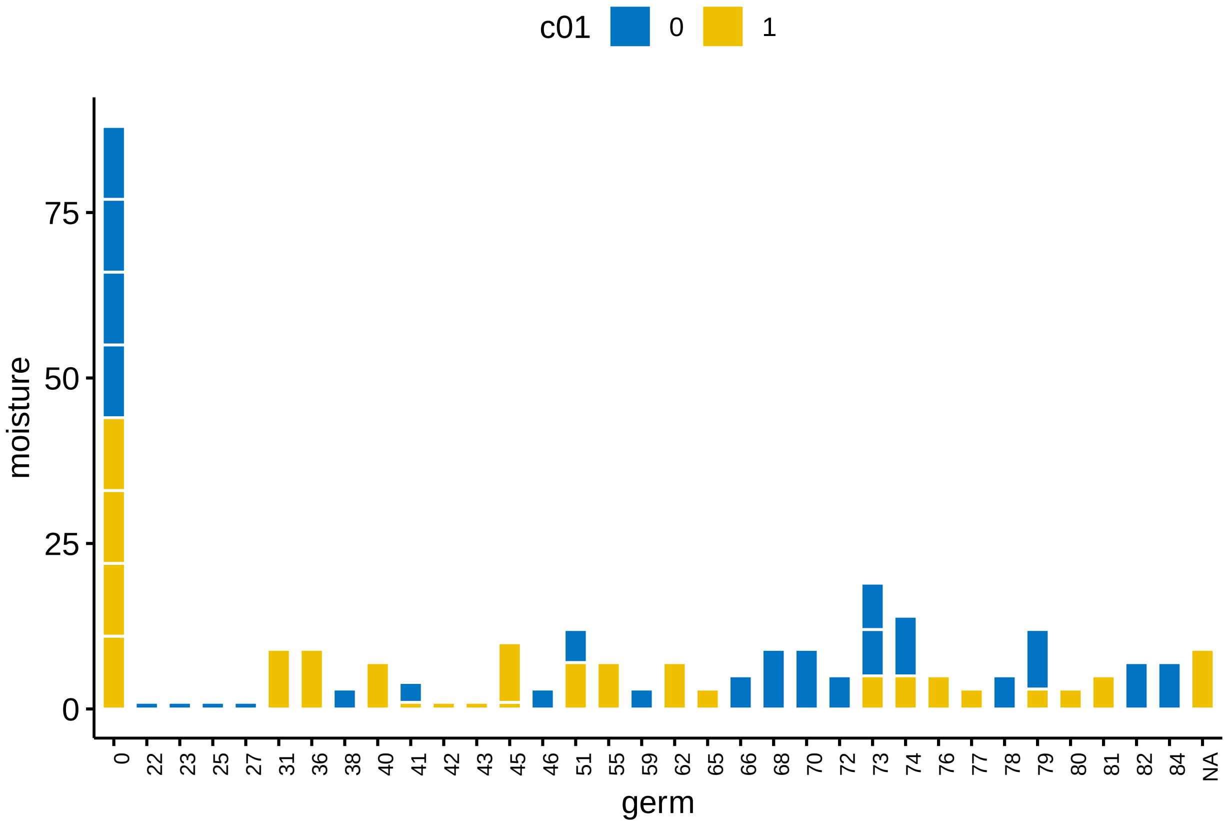
```
library(cowplot)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
```

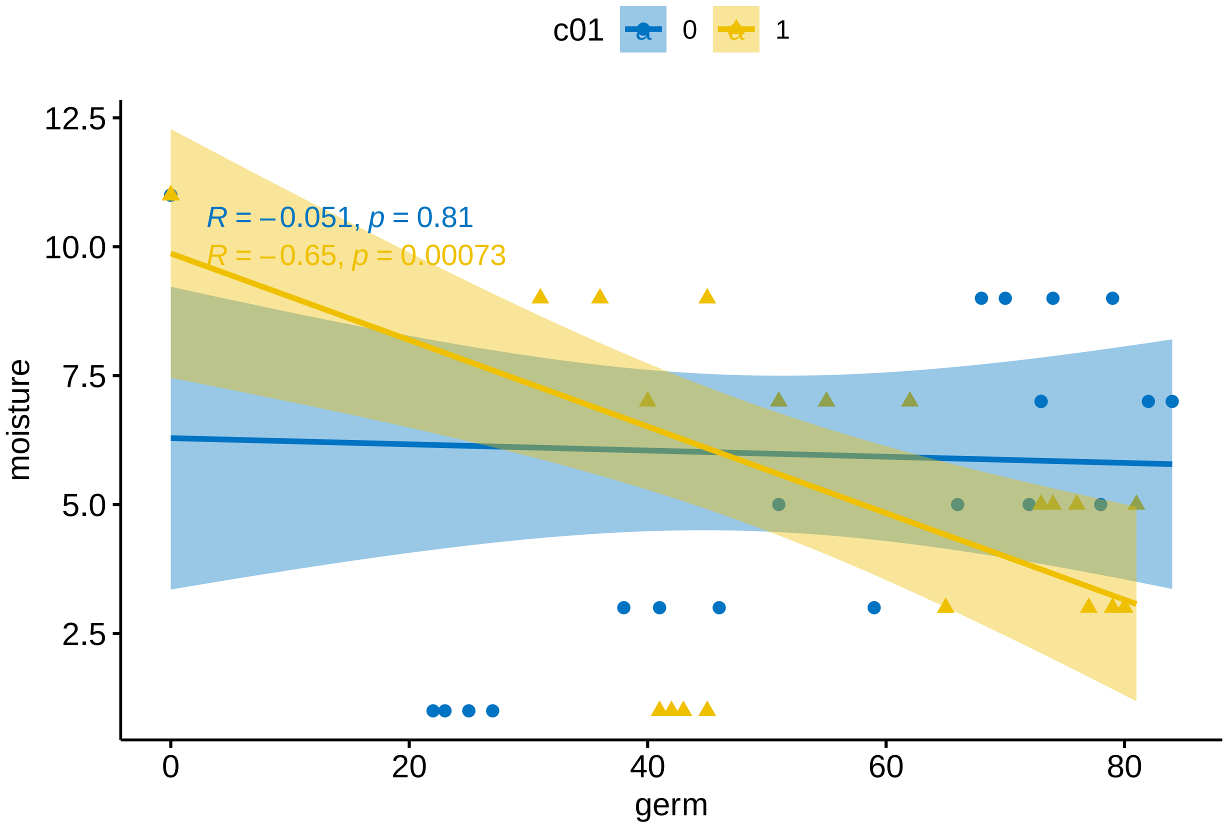
```
##
## Attaching package: 'ggpubr'
##
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
# Bar plot (bp)
bp <- ggbarplot(seeds, x = "germ", y = "moisture",
          fill = "c01",                # change fill color by cyl
          color = "white",             # Set bar border colors to white
          palette = "jco",             # jco journal color palett. see ?ggpar
                    # Sort the value in ascending order
          sort.by.groups = TRUE,       # Sort inside each group
          x.text.angle = 90            # Rotate vertically x axis texts
          )
bp + font("x.text", size = 8)
```



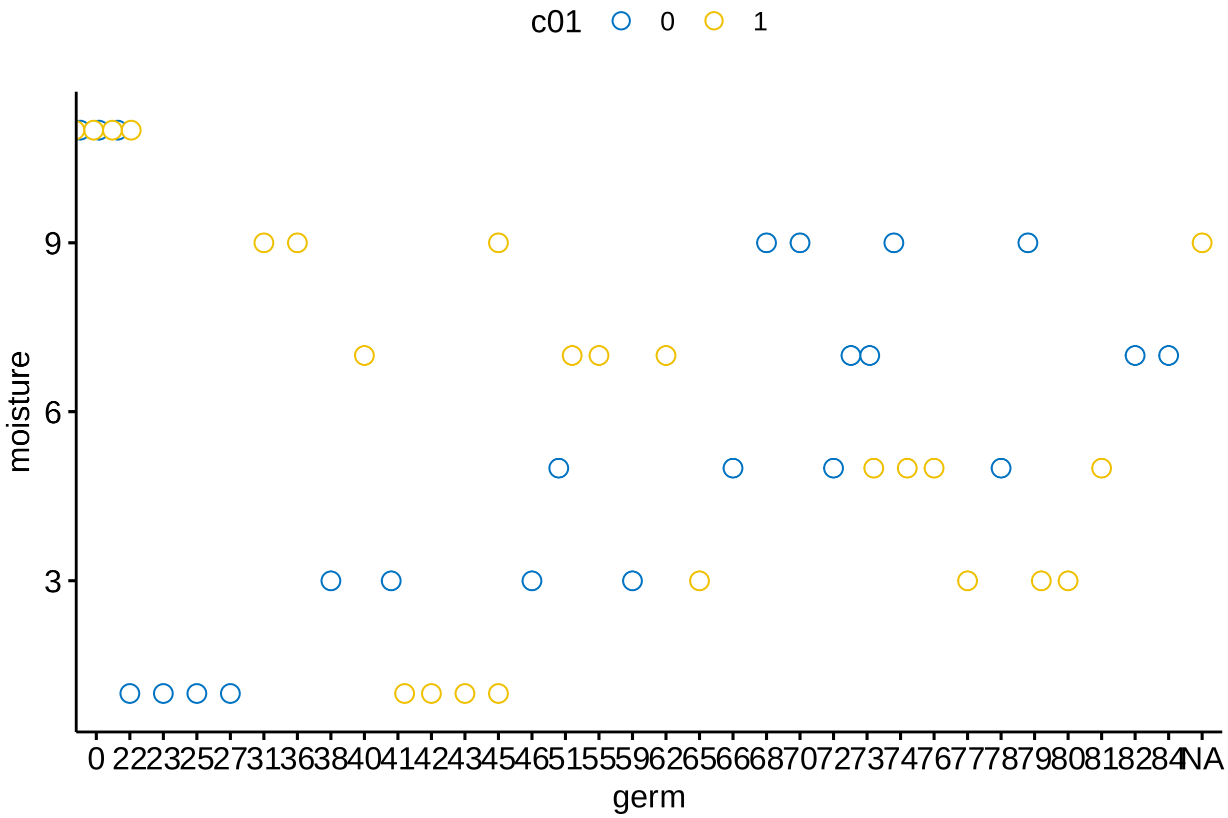
```
# Scatter plots (sp)
sp <- ggscatter(seeds, x = "germ", y = "moisture",
                add = "reg.line",                # Add regression line
                conf.int = TRUE,                 # Add confidence interval
                color = "c01", palette = "jco", # Color by groups "cyl"
                shape = "c01"                    # Change point shape by groups "cyl"
                )+
  stat_cor(aes(color = c01), label.x = 3)        # Add correlation coefficient
sp
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing non-finite values (stat_cor).

## Warning: Removed 1 rows containing missing values (geom_point).
```



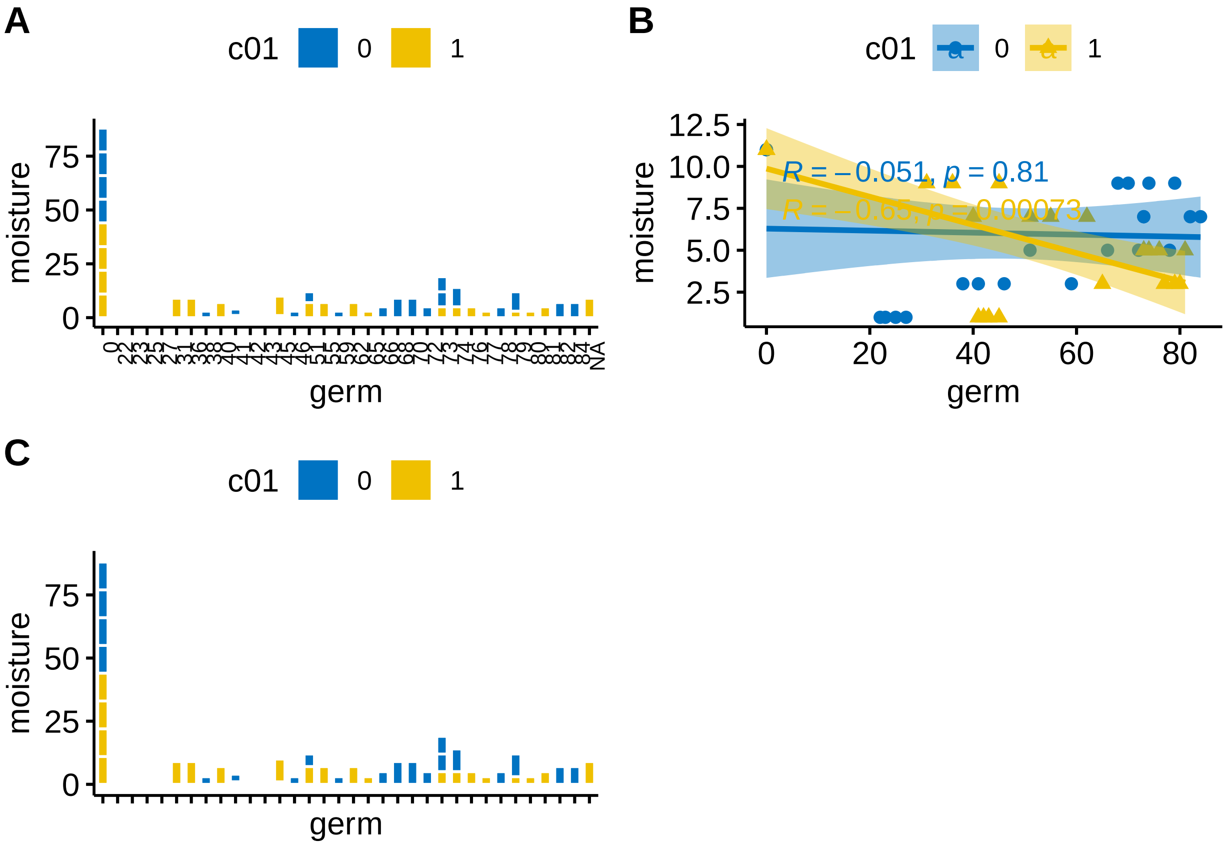
```
# Dot plot (dp)
dp <- ggdotplot(seeds, x = "germ", y = "moisture",
                color = "c01", palette = "jco")
dp
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



```
ggarrange(bp + font("x.text", size = 8), sp, bp + rremove("x.text"),
          labels = c("A", "B", "C"),
          ncol = 2, nrow = 2)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing non-finite values (stat_cor).

## Warning: Removed 1 rows containing missing values (geom_point).
```



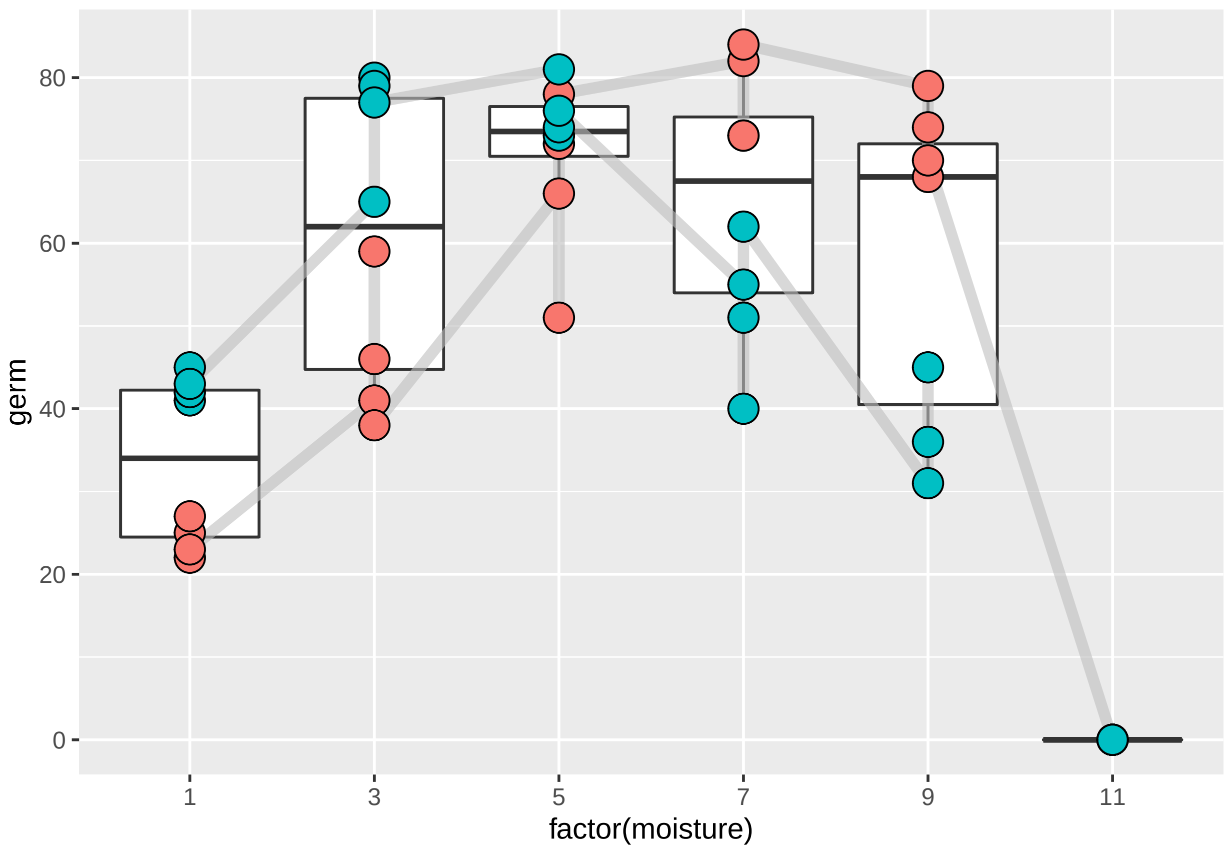
From the scatter plots, we know that the covered one(c01 = 1) has meaningful regression line(decreasing line) between germ and moisture.

(b) Create a new factor describing the box (the data are ordered in blocks of 6 observations per box). Add lines to your previous plot that connect observations from the same box. Is there an indication of a box effect?

```
# Plot boxplot using ggplot function
# diamonds dataset used here is inbuilt in the R Language
plot <- ggplot(seeds, aes(x=factor(moisture), y=germ), fill = c01)+
     geom_boxplot()+
     theme( legend.position = "none" ) +
  geom_boxplot()+

  # geom_line() joins the paired datapoints
  # color and size parameters are used to customize line
  geom_line(aes(group = c01), size=2, color='gray', alpha=0.6)+

  # geom_point() is used to make points at data values
  # fill and size parameters are used to customize point
  geom_point(aes(fill=c01,group=c01),size=5,shape=21)

# print boxplot
plot
```

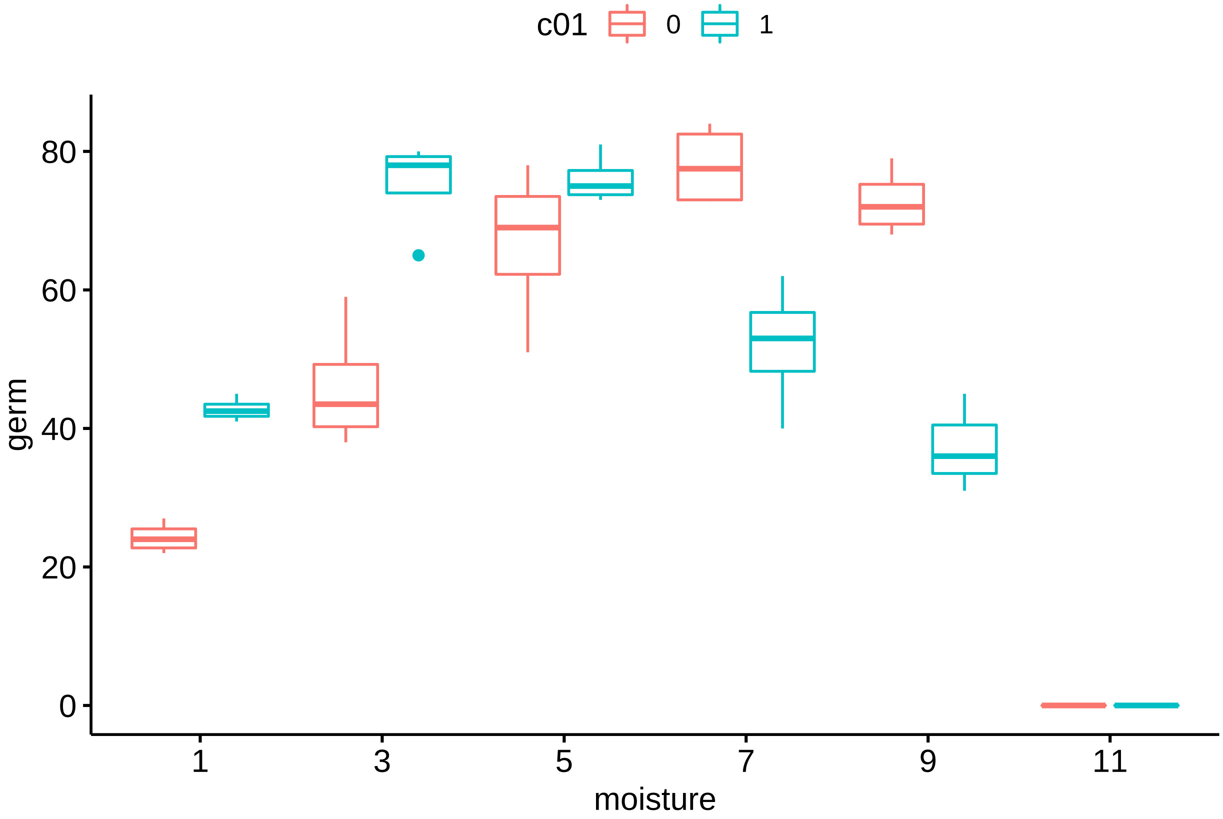
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
## Removed 1 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



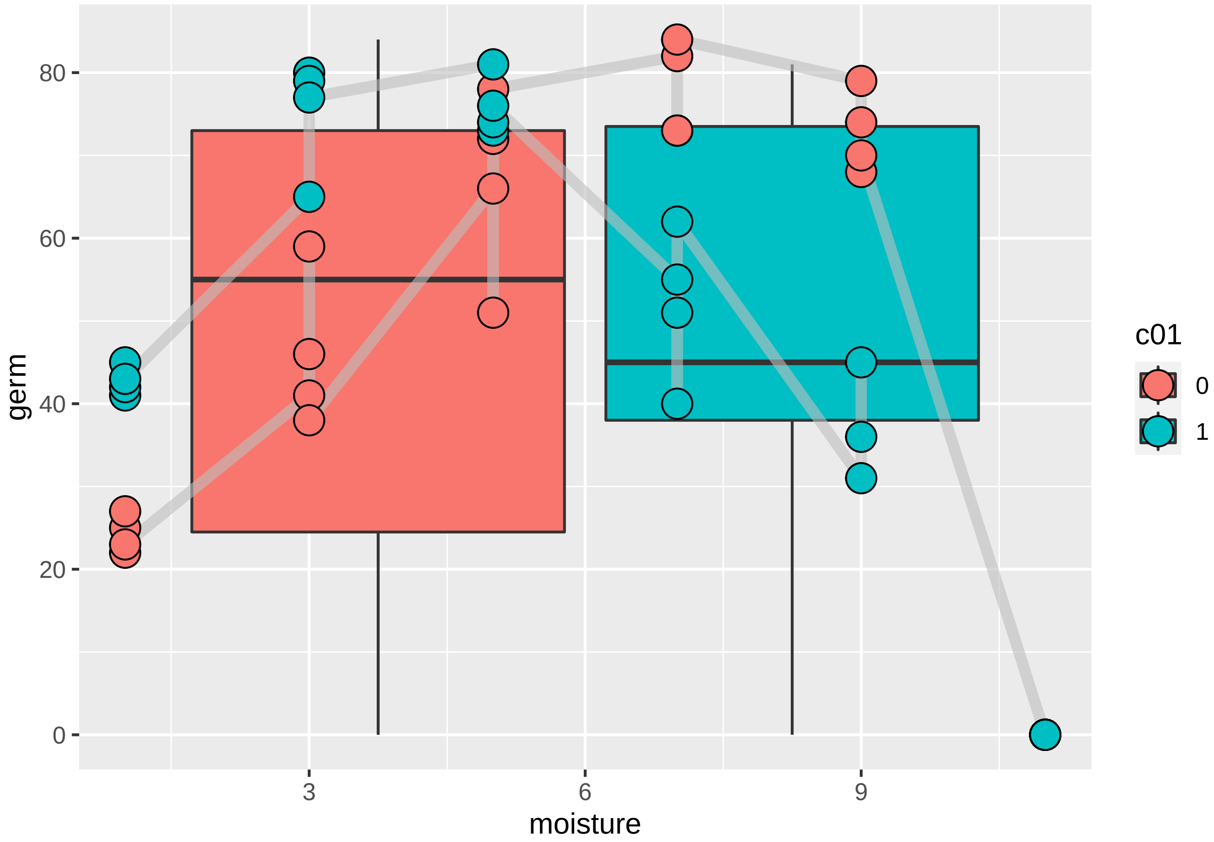
```
# Box plot (bp)
bxp <- ggboxplot(seeds, x = "moisture", y = "germ",
                 color = "c01", sort.val = "asc",           # Sort the value in ascending order
          sort.by.groups = TRUE)      # Sort inside each group
bxp
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



```
# create plot using ggplot() and geom_boxplot() functions
ggplot(seeds, aes(moisture, germ, fill=c01)) +
  geom_boxplot()+

  # geom_line() joins the paired datapoints
  # color and size parameters are used to customize line
  geom_line(aes(group = c01), size=2, color='gray', alpha=0.6)+

  # geom_point() is used to make points at data values
  # fill and size parameters are used to customize point
  geom_point(aes(fill=c01,group=c01),size=5,shape=21)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
## Removed 1 rows containing missing values (geom_point).
```



I created c01 column, according to the covered. For the number “1”(=green) in c01 column, the boxplot line above shows the decreasing line when we follow the increasing moisture>3.

(c) Fit a binomial response model including the coverage, box and moisture predictors. Use the plots to determine an appropriate choice of model.

```
seeds$germ2 <- seeds$germ/100
#select only c01=1 rows.
sd2 <- seeds %>% filter(c01 == 1)
head(sd2)
```

```
##   germ moisture covered c01 germ2
## 1   45        1     yes   1  0.45
## 2   65        3     yes   1  0.65
## 3   81        5     yes   1  0.81
## 4   55        7     yes   1  0.55
## 5   31        9     yes   1  0.31
## 6    0       11     yes   1  0.00
```

```
#Fit a binomial response model

head(sd2)
```

```
##   germ moisture covered c01 germ2
## 1   45        1     yes   1  0.45
## 2   65        3     yes   1  0.65
## 3   81        5     yes   1  0.81
## 4   55        7     yes   1  0.55
## 5   31        9     yes   1  0.31
## 6    0       11     yes   1  0.00
```

```
glm1 <- glm(germ2 ~ moisture, family = binomial, data = sd2)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
#select only c01=0 rows.
sd3 <- seeds %>% filter(c01 == 0)

glm2 <- glm(germ2 ~ moisture, family = binomial, data = sd3)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
#Test the significance of each variable using a likelihood ratio test.
drop1(glm1, test = "Chi")
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
## Single term deletions
##
## Model:
## germ2 ~ moisture
##          Df Deviance    AIC    LRT Pr(>Chi)
## <none>        5.3169 33.894
## moisture  1   8.2298 34.807 2.9129  0.08787 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(glm2, test = "Chi")
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
## Single term deletions
##
## Model:
## germ2 ~ moisture
##          Df Deviance    AIC      LRT Pr(>Chi)
## <none>        9.6594 37.802
## moisture  1   9.6797 35.822 0.020269   0.8868
```

For the both models, P-value is not less than 0.05. So they are not proper. But I can choose model1 since the p-value is much less than the one of model2.

(d) Test for the significance of a box effect in your model. Repeat the same test but using the Pearson's Chi-squared statistic instead of the deviance.

```
#' ## 3. Pearson Chi-squared
pearson.chi = sum(residuals(glm1,type="pearson")^2)
pearson.chi
```

```
## [1] 4.586438
```

```
1-pchisq(pearson.chi, glm1$df.residual)
```

```
## [1] 0.9999358
```

```
deviance(glm1)
```

```
## [1] 5.316907
```

```
1-pchisq(deviance(glm1), glm1$df.residual)
```

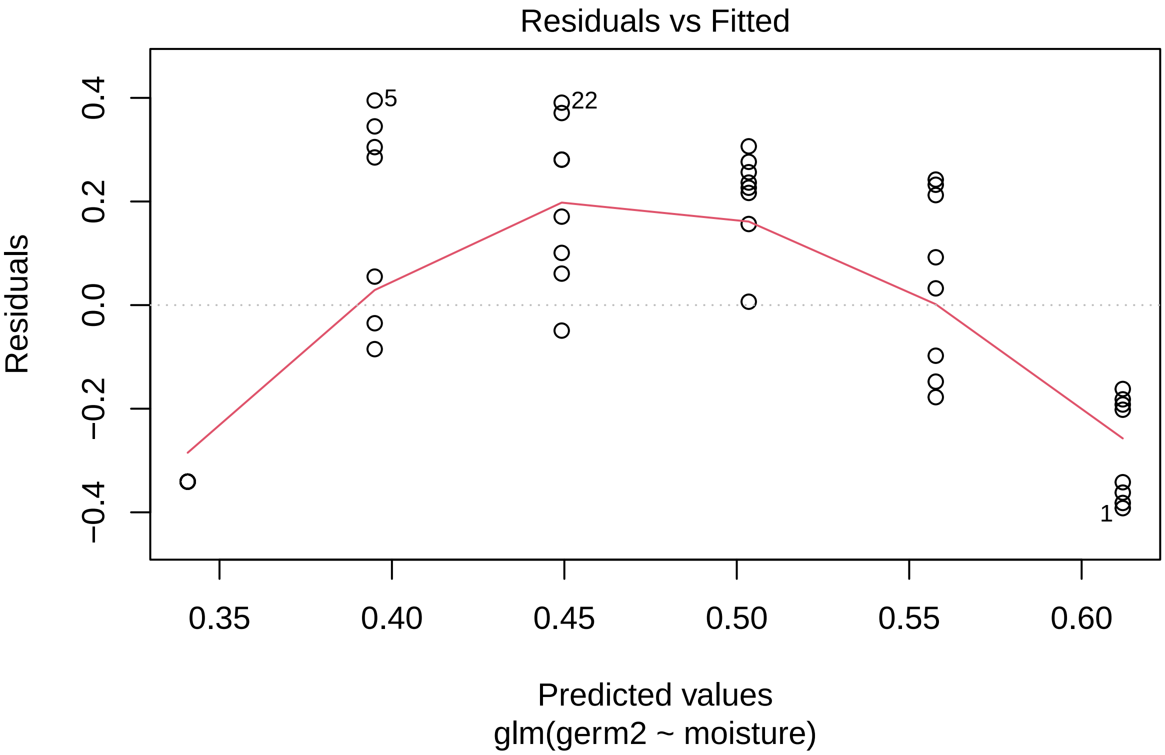
```
## [1] 0.9997813
```

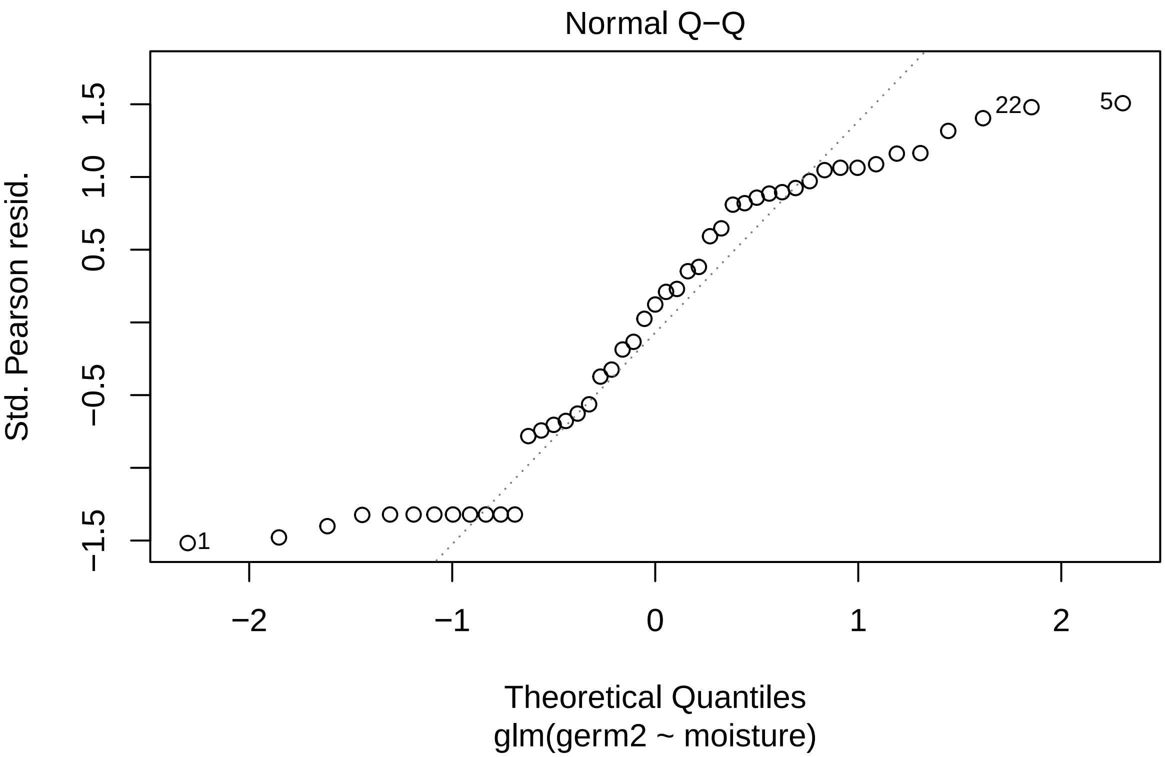
Pearson's Chi-squared statistic has error 0.2675654.

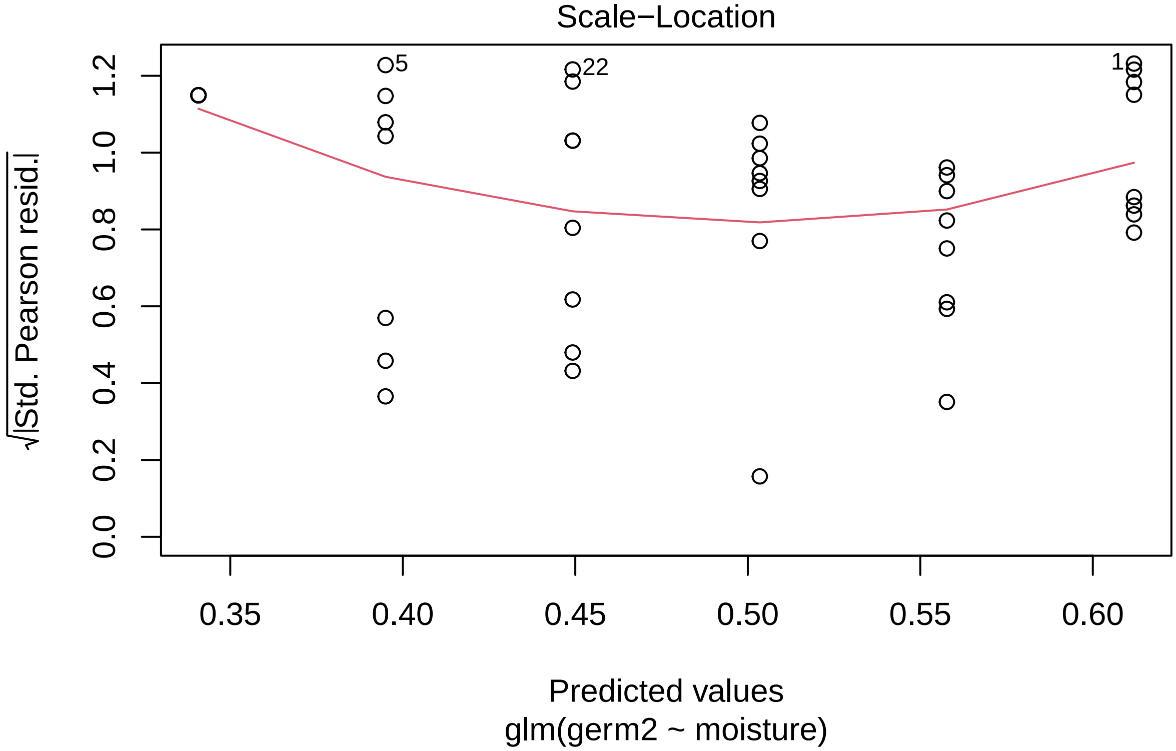
(e) At what value of moisture does the predicted maximum germination occur for noncovered boxes? For covered boxes? From the lined box plot in (b), uncovered one with moisture = 7 makes maximum germination.

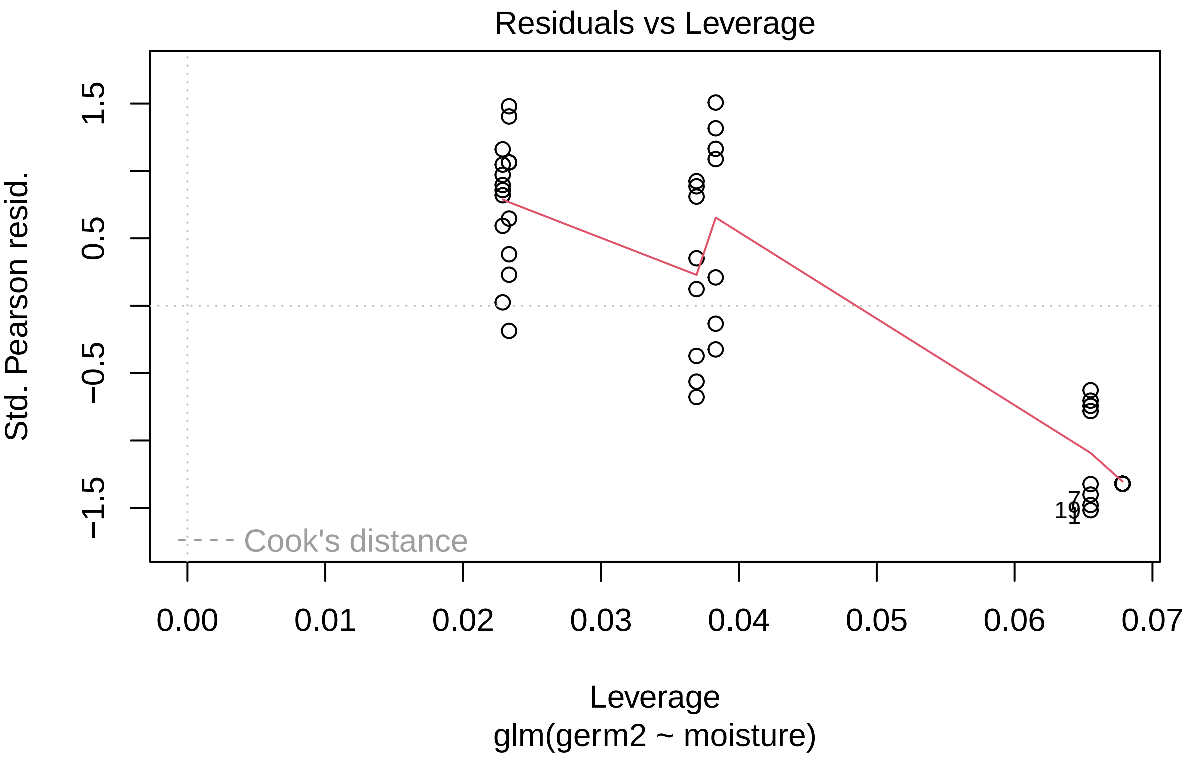
(f) Produce a plot of the residuals against the fitted values and interpret.

```
# Scatter plots (sp)
plot(glm(germ2~moisture,data=seeds))
```




Normal Q−Q
Std. Pearson resid.
−1.5
−0.5
0.5
1.0
1.5
−2
−1
0
1
2
1
22
5
Theoretical Quantiles
glm(germ2 ~ moisture)


Scale−Location
√|Std. Pearson resid.|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.35
0.40
0.45
0.50
0.55
0.60
5
22
1
Predicted values
glm(germ2 ~ moisture)



Residuals are small enough in the middle range of predicted values. Therefore, the prediction of glm would be good enough in the middle range of the predicted values.

Q3. #EMLR 7.1 a-f

1. The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status (SES); school type; chosen high school program type; scores on reading, writing, math, science, and social studies.We want to determine which factors are related to the choice of the type of program — academic, vocational or general —that the students pursue in high school. The response is multinomial with three levels.

```
rm(list = ls())
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(dev = 'pdf')
knitr::opts_chunk$set(cache=TRUE)
knitr::opts_chunk$set(tidy=TRUE)
knitr::opts_chunk$set(prompt=FALSE)
knitr::opts_chunk$set(fig.height=5)
knitr::opts_chunk$set(fig.width=7)
knitr::opts_chunk$set(warning=FALSE)
knitr::opts_chunk$set(message=FALSE)
knitr::opts_knit$set(root.dir = ".")
#install.packages('glmtoolbox', dependencies = TRUE, repos='http://cran.rstudio.com/')

library(latex2exp)
library(pander)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:pander':
##
##     wrap
```

```
## The following object is masked from 'package:faraway':
##
##     happy
```

(a) Make a table showing the proportion of males and females choosing the three different programs. Comment on the difference. Repeat this comparison but for SES rather than gender.

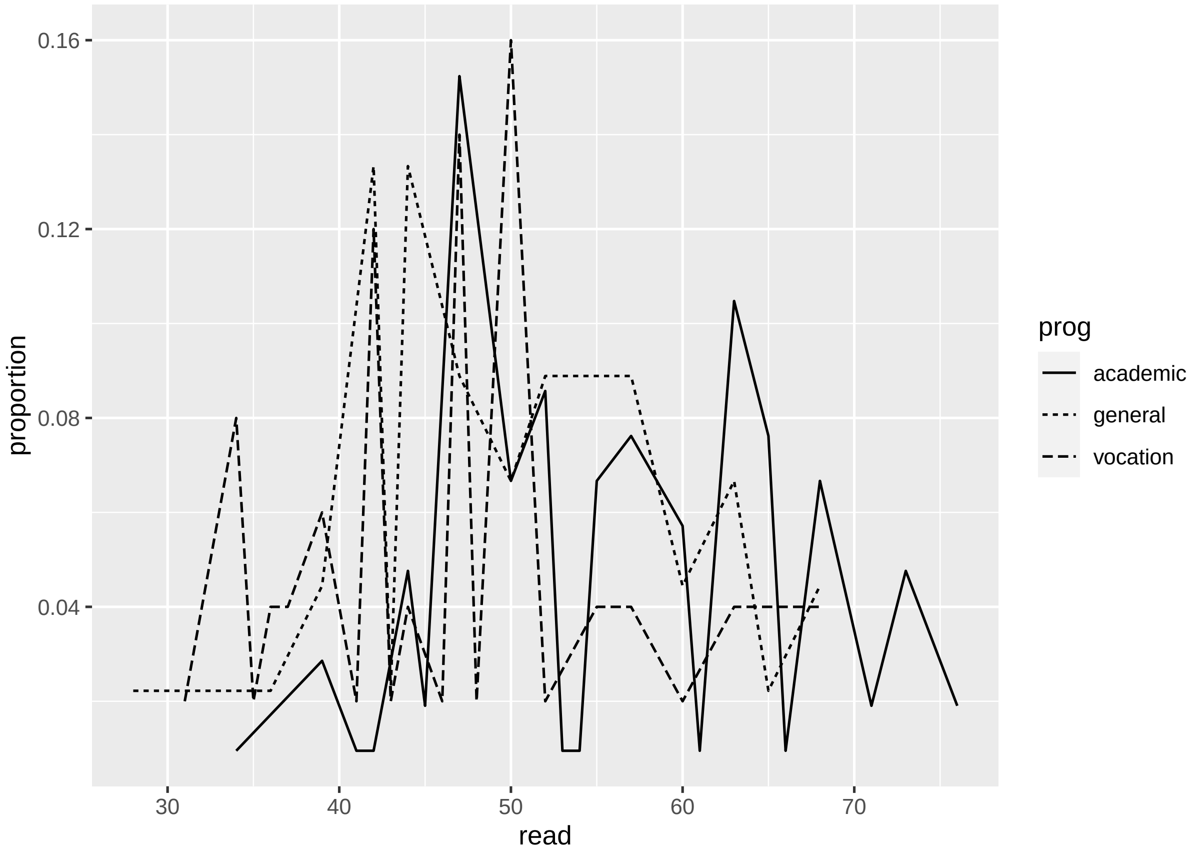
```
##    id gender  race    ses schtyp     prog read write math science socst
## 1  70   male white    low public  general   57    52   41      47    57
## 2 121 female white middle public vocation   68    59   53      63    61
## 3  86   male white   high public  general   44    33   54      58    31
## 4 141   male white   high public vocation   63    44   47      53    56
## 5 172   male white middle public academic   47    52   57      53    61
## 6 113   male white middle public academic   44    52   51      63    61
```

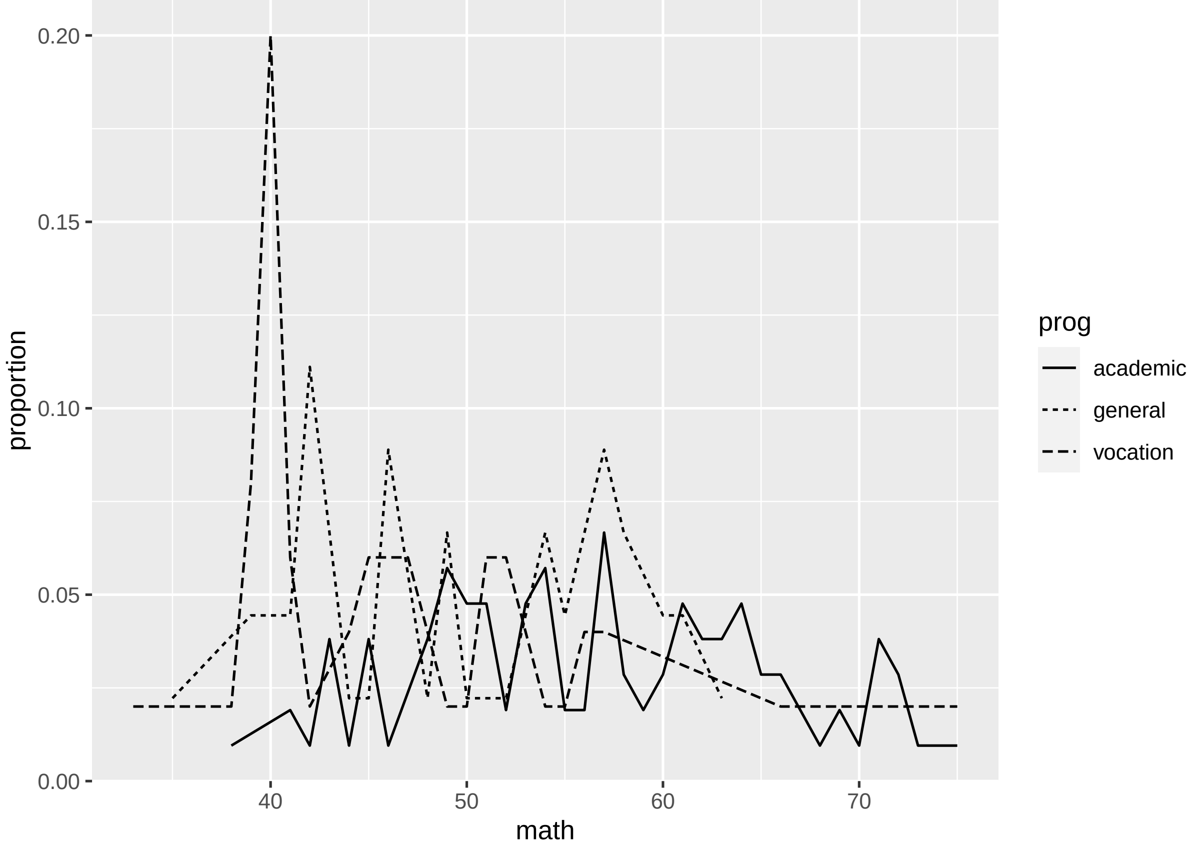
```
##           gender
## prog       female male
##   academic     58   47
##   general      24   21
##   vocation     27   23
```

```
##           ses
## prog       high low middle
##   academic   42  19     44
##   general     9  16     20
##   vocation    7  12     31
```

Generally, on the first table in above, the number of female is more than the number of man. Also, there are more difference in the academic one than other two categories. For the secone table, we have middle level mostly on the ses, the biggest difference happends on the vocation categry.

(b) Construct a plot like the right panel of Figure 7.1 that shows the relationship between program choice and reading score. Comment on the plot. Repeat for math in place of reading.

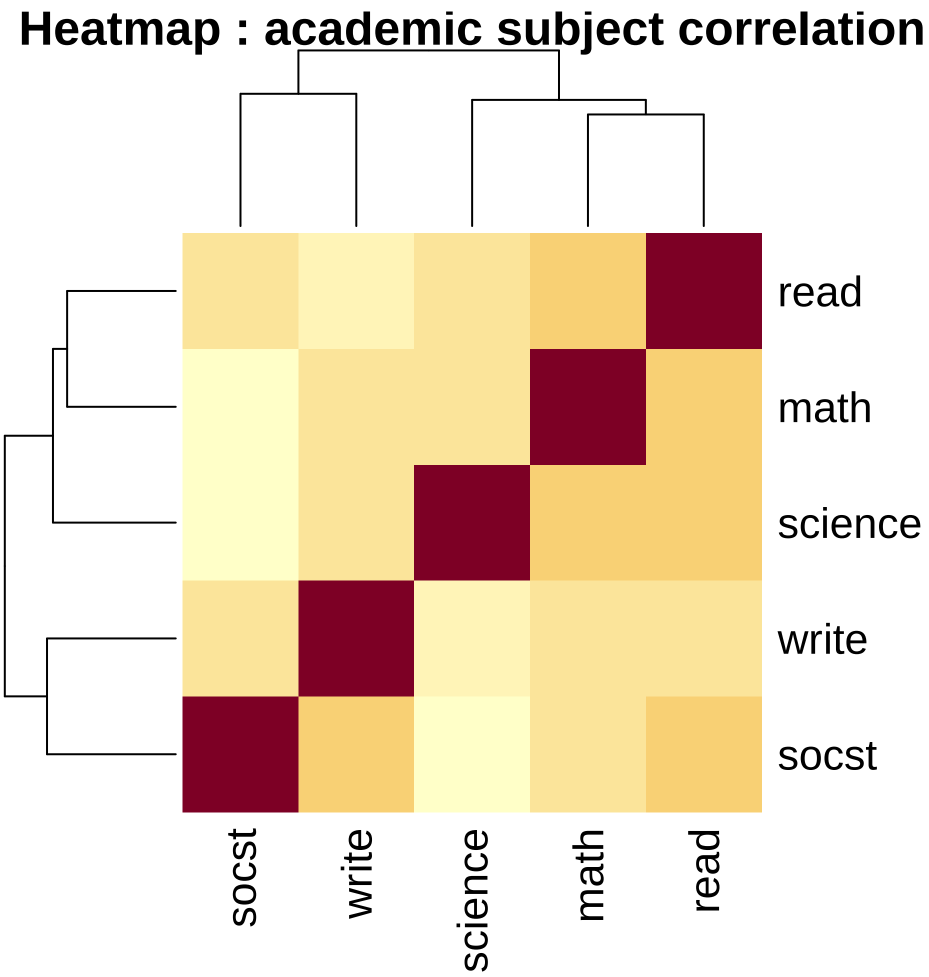




For the fist graph above, the vocation category has the highest proportion. But, in general, the academic category records overalll constantly fair middle-high proportion. Forthe second graph above, the general category records the highesrt propertion and the second highest record of proportion come from the genearal category.

(c) Compute the correlation matrix for the five subject scores.

	read	write	math	science	socst
read	1	0.5968	0.6623	0.6302	0.6215
write	0.5968	1	0.6174	0.5704	0.6048
math	0.6623	0.6174	1	0.6307	0.5445
science	0.6302	0.5704	0.6307	1	0.4651
socst	0.6215	0.6048	0.5445	0.4651	1



(d) Fit a multinomial response model for the program choice and examine the fitted coefficients. Of the five subjects, one gives unexpected coefficients. Identify this subject and suggest an explanation for this behavior.

```
## # weights:  21 (12 variable)
## initial  value 219.722458
## iter  10 value 173.294170
## final  value 164.975567
## converged
## Call:
## multinom(formula = prog ~ socst + write + science + math + read,
##     data = df)
##
## Coefficients:
##          (Intercept)       socst       write    science        math        read
## general     4.804643 -0.03207906 -0.02299500 0.09359693 -0.09837739 -0.04558434
## vocation    9.405865 -0.06751481 -0.03743141 0.05902499 -0.11579827 -0.03583817
##
## Std. Errors:
##          (Intercept)      socst      write    science       math       read
## general     1.498368 0.02567748 0.02958054 0.02985794 0.03329584 0.02958174
## vocation    1.621442 0.02642707 0.02955005 0.03026686 0.03639133 0.03205540
##
## Residual Deviance: 329.9511
## AIC: 353.9511
```

For the fitted coeffieicnts for the five subjects, we can see the only science subject has the negative coefficients on the general and vocation cateries. We can interpret this as for the following formula,

#(((Check again))) ln(P(prog = general)/P(prog = academic)) has the linear line with the science coefficeints as the 0.09359693. And, ln(P(prog = vocation)/P(prog = academic)) has the linear line with the science coefficients as the 0.05902499.

(e) Construct a derived variable that is the sum of the five subject scores. Fit a multinomial model as before except with this one sum variable in place of the five subjects separately. Compare the two models to decide which should be preferred.

```
## # weights:  33 (20 variable)
## initial  value 219.722458
## iter  10 value 167.158173
## iter  20 value 164.141699
## final  value 164.130704
## converged
```

```
## Call:
## multinom(formula = prog ~ id + gender + race + ses + schtyp +
##     sum.subject, data = df.reduced)
##
## Coefficients:
##          (Intercept)           id  gendermale  raceasian racehispanic racewhite
## general     3.227335 -0.003708235  0.24883040  1.0243408   -0.5484976  1.060033
## vocation    7.112010 -0.003220142 -0.09614882 -0.6015843   -0.1937564  1.098265
##              seslow sesmiddle schtyppublic sum.subject
## general  1.0593830 0.6350558    0.3875245 -0.02052599
## vocation 0.2517821 1.1874930    1.8098161 -0.04125543
##
## Std. Errors:
##          (Intercept)          id gendermale raceasian racehispanic racewhite
## general     1.798815 0.006823237  0.3941480 0.9439661    0.8799224 0.8740777
## vocation    2.157426 0.007659938  0.4364287 1.3769618    0.8411264 0.8970833
##              seslow sesmiddle schtyppublic sum.subject
## general  0.5664146 0.4789630    0.6826598 0.005976099
## vocation 0.6797684 0.5566371    0.9568939 0.007225491
##
## Residual Deviance: 328.2614
## AIC: 368.2614
```

For the comparing two models from aboce two summary, we consider the Residual Deviance and AIC values. If model fits well, it has small deviance and small AIC. So from this comparison, model in (d) will be better in the view of Residual Deviance value and model in (e) will be better in the view of AIC value.

The single subject variable is much begger than the s.e. for the combined subject variable.

(f) Use a stepwise method to reduce the model. Which variables are in your selected model?

```
## Call:
## multinom(formula = prog ~ ses + schtyp + sum.subject, data = df.reduced)
##
## Coefficients:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     2.593944 0.8078324 0.5808536    0.5594952 -0.01635887
## vocation    6.372051 0.1330839 1.1517240    1.8490860 -0.03681150
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     1.587502 0.5386033 0.4720925    0.5219044 0.005422494
## vocation    1.877764 0.6468558 0.5465572    0.7974692 0.006553295
##
## Residual Deviance: 336.0554
## AIC: 356.0554
```

From the above, we can choose the three variables ses, schtyp an d sum.subjec are chosen variables.

HW4

#3 Since $p_{ij} = P(Y_{ij} = 1)$,

$$\sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{e^{\eta_{ij}}}{1+\sum_{k=2}^{J} e^{\eta_{ik}}}$$

$$= \frac{1+\sum_{k=2}^{J} e^{\eta_{ik}}}{1+\sum_{k=2}^{J} e^{\eta_{ik}}}$$

$$= 1$$