Homework 2

Sarah Adilijiang

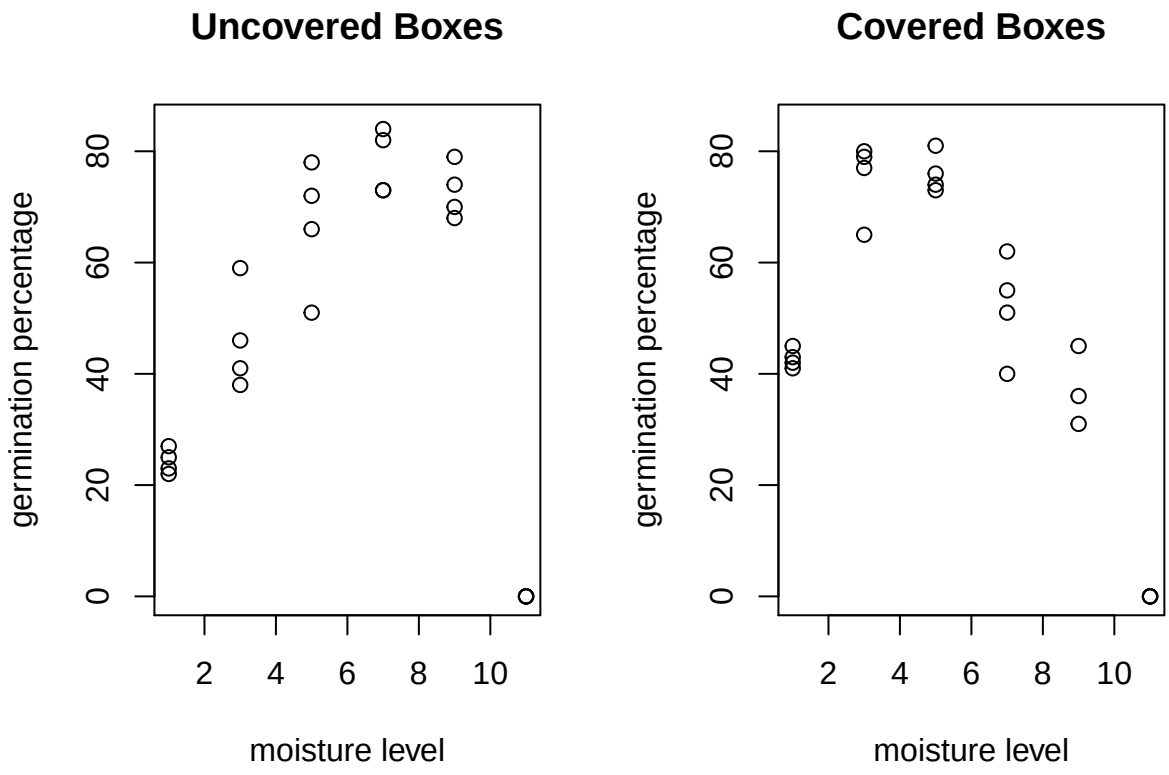
Problem 2

(a)

```
library(faraway)
data("seeds")
str(seeds)
```

```
## 'data.frame':    48 obs. of  3 variables:
##  $ germ    : num  22 41 66 82 79 0 25 46 72 73 ...
##  $ moisture: num  1 3 5 7 9 11 1 3 5 7 ...
##  $ covered : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# plot germination percentage ~ moisture level
par(mfrow = c(1,2))
plot(germ~moisture, data=seeds[seeds$covered=="no", ], ylim=c(0,85),
       main="Uncovered Boxes", xlab="moisture level", ylab="germination percentage")
plot(germ~moisture, data=seeds[seeds$covered=="yes", ], ylim=c(0,85),
       main="Covered Boxes", xlab="moisture level", ylab="germination percentage")
```



Answer:

(1) The relationship between germination percentage and moisture level is not monotonic. It looks like

quadratic, which increases first and then decreases. (2) For covered and uncovered boxes, the germination percentage reaches the maximum value at different moisture levels.

(b)

```
# add variable facotr "box"
seeds$box = as.factor((rep(1:8,rep(6,8))))  # = rep(c(1,2,3,4,5,6,7,8),c(6,6,6,6,6,6,6,6))
str(seeds)
```

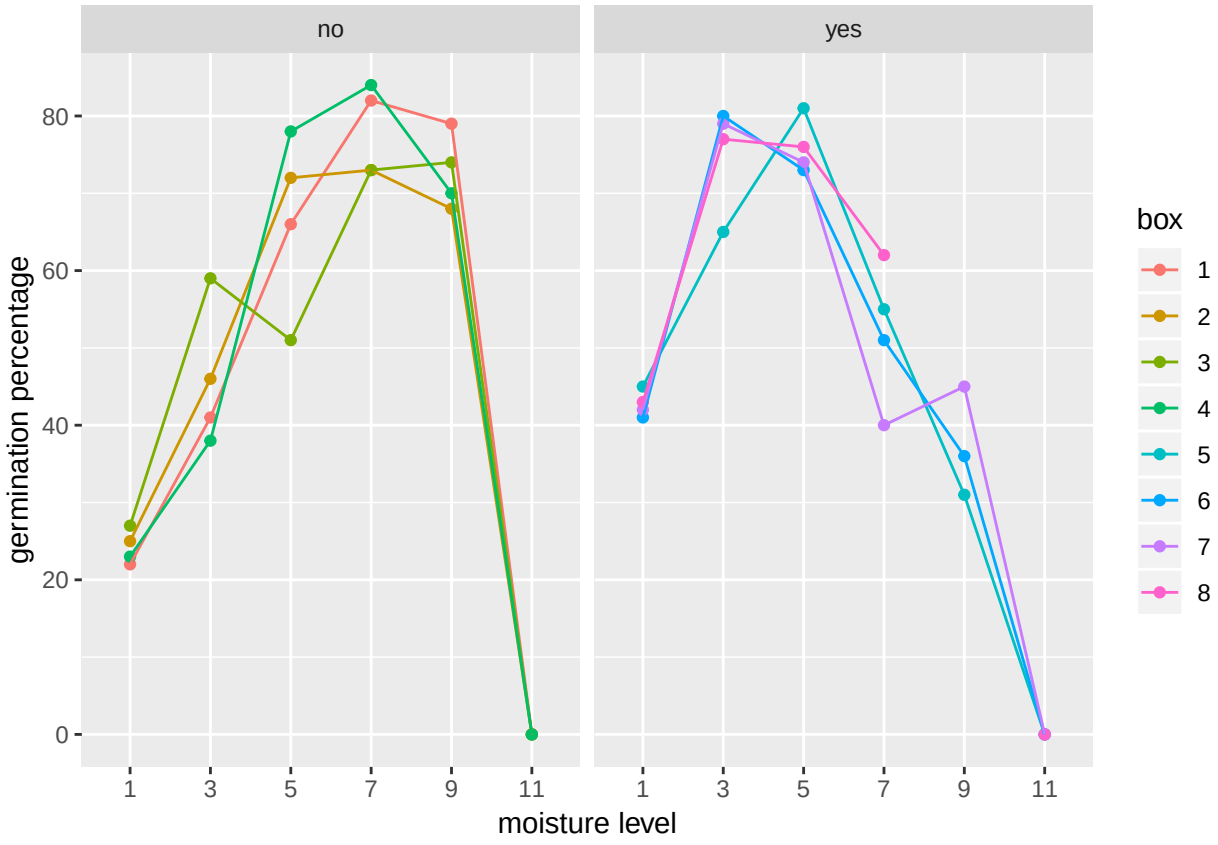
```
## 'data.frame':    48 obs. of  4 variables:
##  $ germ    : num  22 41 66 82 79 0 25 46 72 73 ...
##  $ moisture: num  1 3 5 7 9 11 1 3 5 7 ...
##  $ covered : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ box     : Factor w/ 8 levels "1","2","3","4",..: 1 1 1 1 1 1 2 2 2 2 ...
```

```
# plot germination percentage ~ moisture level, lining the same boxes
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
ggplot(data=seeds, aes(x=as.factor(moisture), y=germ, color=box)) +
    geom_point() + xlab("moisture level") + ylab("germination percentage") +
    geom_line(aes(group=box)) + facet_grid(~covered)
```

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



Answer:

There is no significant patterns for different boxes in the relationship between germination percentage and

moisture level. Thus there is no indication of a box effect.

Note: due to one missing value, there is no point connection for covered BOX8 at moisture level 9.

(c)

The generalized linear model (GLM) for binomial response here is:
likelihood: $P(germ_i = y_i|p_i) = \binom{100}{y_i} p_i^{y_i}(1-p_i)^{100-y_i},\ y_i = 0, 1, ..., 100$
linear predictor: $\eta_i = \beta_0 + \beta_1 moisture_i + \beta_2 1_{covered_i="yes"} + \beta_3 1_{box_i=2} + \beta_4 1_{box_i=3} + \beta_5 1_{box_i=4} + \beta_6 1_{box_i=5} + \beta_7 1_{box_i=6} + \beta_8 1_{box_i=7} + \beta_9 1_{box_i=8} + \epsilon_i$
link function (logit): $\eta_i = log\frac{p_i}{1-p_i}$, where $p_i = y_i/100$ is the percentage of germinated seeds in each box.

```
# modify the data frame
seeds$notgerm = 100-seeds$germ

# fit the logistic GLM model for binomial response
model.glm = glm(cbind(germ,notgerm)~moisture+covered+box, data=seeds, family=binomial)
summary(model.glm)
```

```
##
## Call:
## glm(formula = cbind(germ, notgerm) ~ moisture + covered + box,
##     family = binomial, data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.5285  -5.5046   0.7063   5.0465   7.9405
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.593872   0.098319   6.040 1.54e-09 ***
## moisture    -0.110487   0.008813 -12.537  < 2e-16 ***
## coveredyes   0.067297   0.123410   0.545    0.586
## box2        -0.041493   0.117609  -0.353    0.724
## box3        -0.041493   0.117609  -0.353    0.724
## box4         0.020724   0.117544   0.176    0.860
## box5        -0.157309   0.123521  -1.274    0.203
## box6        -0.129567   0.123476  -1.049    0.294
## box7        -0.136498   0.123486  -1.105    0.269
## box8              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1791.0  on 46  degrees of freedom
## Residual deviance: 1624.4  on 38  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1832.1
##
## Number of Fisher Scoring iterations: 5
```

```
# check the data design matrix of the model
model.matrix(model.glm)
```

```
##    (Intercept) moisture coveredyes box2 box3 box4 box5 box6 box7 box8
## 1            1        1          0    0    0    0    0    0    0    0
## 2            1        3          0    0    0    0    0    0    0    0
## 3            1        5          0    0    0    0    0    0    0    0
## 4            1        7          0    0    0    0    0    0    0    0
## 5            1        9          0    0    0    0    0    0    0    0
## 6            1       11          0    0    0    0    0    0    0    0
## 7            1        1          0    1    0    0    0    0    0    0
## 8            1        3          0    1    0    0    0    0    0    0
## 9            1        5          0    1    0    0    0    0    0    0
## 10           1        7          0    1    0    0    0    0    0    0
## 11           1        9          0    1    0    0    0    0    0    0
## 12           1       11          0    1    0    0    0    0    0    0
## 13           1        1          0    0    1    0    0    0    0    0
## 14           1        3          0    0    1    0    0    0    0    0
## 15           1        5          0    0    1    0    0    0    0    0
## 16           1        7          0    0    1    0    0    0    0    0
## 17           1        9          0    0    1    0    0    0    0    0
## 18           1       11          0    0    1    0    0    0    0    0
## 19           1        1          0    0    0    1    0    0    0    0
## 20           1        3          0    0    0    1    0    0    0    0
## 21           1        5          0    0    0    1    0    0    0    0
## 22           1        7          0    0    0    1    0    0    0    0
## 23           1        9          0    0    0    1    0    0    0    0
## 24           1       11          0    0    0    1    0    0    0    0
## 25           1        1          1    0    0    0    1    0    0    0
## 26           1        3          1    0    0    0    1    0    0    0
## 27           1        5          1    0    0    0    1    0    0    0
## 28           1        7          1    0    0    0    1    0    0    0
## 29           1        9          1    0    0    0    1    0    0    0
## 30           1       11          1    0    0    0    1    0    0    0
## 31           1        1          1    0    0    0    0    1    0    0
## 32           1        3          1    0    0    0    0    1    0    0
## 33           1        5          1    0    0    0    0    1    0    0
## 34           1        7          1    0    0    0    0    1    0    0
## 35           1        9          1    0    0    0    0    1    0    0
## 36           1       11          1    0    0    0    0    1    0    0
## 37           1        1          1    0    0    0    0    0    1    0
## 38           1        3          1    0    0    0    0    0    1    0
## 39           1        5          1    0    0    0    0    0    1    0
## 40           1        7          1    0    0    0    0    0    1    0
## 41           1        9          1    0    0    0    0    0    1    0
## 42           1       11          1    0    0    0    0    0    1    0
## 43           1        1          1    0    0    0    0    0    0    1
## 44           1        3          1    0    0    0    0    0    0    1
## 45           1        5          1    0    0    0    0    0    0    1
## 46           1        7          1    0    0    0    0    0    0    1
## 48           1       11          1    0    0    0    0    0    0    1
## attr(,"assign")
##  [1] 0 1 2 3 3 3 3 3 3 3
## attr(,"contrasts")
## attr(,"contrasts")$covered
## [1] "contr.treatment"
##
```

```
## attr(,"contrasts")$box
## [1] "contr.treatment"
```

Answer:

We can see that in the design matrix, there is a linear relationship between the following indicator covariates: $coveredyes = box5 + box6 + box7 + box8$, so $box8 = coveredyes - box5 - box6 - box7$, which makes it a redundant (linear dependent) indicator variable that occurs in the last. Thus there is an NA appearing for the BOX8 factor level due to the rank deficiency of the design matrix.

(d)

The linear predictor of the GLM submodel without "box" is:
linear predictor: $\eta_i = \beta_0 + \beta_1 moisture_i + \beta_2 1_{covered_i = "yes"} + \epsilon_i$

```
# fit a GLM submodel without "box"
model.glm_sub = glm(cbind(germ,notgerm)~moisture+covered, data=seeds, family=binomial)

# likelihood ratio test (LRT) test : Chi-square
LRT = deviance(model.glm_sub) - deviance(model.glm)
df = model.glm_sub$df.residual - model.glm$df.residual
p_val = 1 - pchisq(LRT, df); p_val
```

```
## [1] 0.884346
```

```
# or
anova(model.glm_sub, model.glm, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(germ, notgerm) ~ moisture + covered
## Model 2: cbind(germ, notgerm) ~ moisture + covered + box
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        44     1626.8
## 2        38     1624.4  6   2.3548   0.8843
```

Answer:

The p-value of the likelihood ratio test (Chi-square test) is 0.884346, thus we do not reject the null hypothesis, so the smaller model without variable "box" is preferred. Therefore, the box factor is not significant.

(e)

```
# aggregate the data set
seeds_agg = aggregate(seeds[,c(1,5)], by=list(seeds$moisture, seeds$covered),
                      FUN=sum, na.rm=TRUE)
colnames(seeds_agg)[c(1,2)] = c("moisture", "covered")
seeds_agg$total = seeds_agg$germ + seeds_agg$notgerm

# fit the model to the aggregated data set
model.glm_agg = glm(cbind(germ,notgerm)~moisture+covered, data=seeds_agg, family=binomial)
summary(model.glm_agg)
```

```
##
## Call:
## glm(formula = cbind(germ, notgerm) ~ moisture + covered, family = binomial,
##     data = seeds_agg)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.407   -9.285    1.211    8.812   13.441
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.581702   0.066863   8.700   <2e-16 ***
## moisture    -0.111056   0.008804 -12.615   <2e-16 ***
## coveredyes  -0.027603   0.059467  -0.464    0.643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1727.7  on 11  degrees of freedom
## Residual deviance: 1563.4  on  9  degrees of freedom
## AIC: 1631.9
##
## Number of Fisher Scoring iterations: 5
```

```
# model before aggregation
summary(model.glm_sub)
```

```
##
## Call:
## glm(formula = cbind(germ, notgerm) ~ moisture + covered, family = binomial,
##     data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.2035  -5.5103   0.5695   5.0362   8.0597
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.581702   0.066863   8.700   <2e-16 ***
## moisture    -0.111056   0.008804 -12.615   <2e-16 ***
## coveredyes  -0.027603   0.059467  -0.464    0.643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1791.0  on 46  degrees of freedom
## Residual deviance: 1626.8  on 44  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1822.5
##
## Number of Fisher Scoring iterations: 5
```

Answer:

Since the box factor is not significant, we can aggregate the 4 boxes that are with the same moisture level and the same coverage condition into one data point, thus changing total 48 observations into 12 observations.

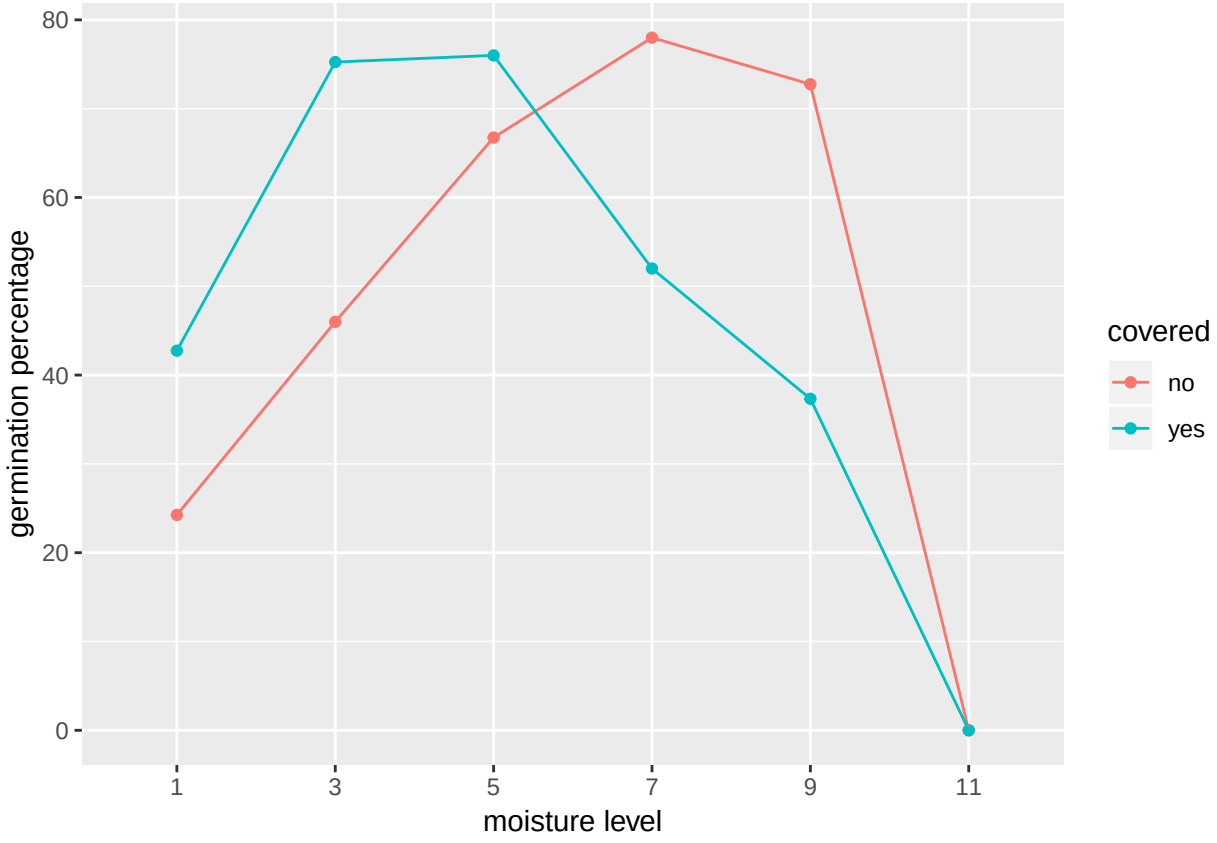
After aggregation, the estimated parameters did not change (also with the same significance result) because the data for each conditions are actually the same. However, the deviance and the degrees of freedom changed

since the number of observations has changed.

From now on, I will use the aggregated data for the following questions.

(f)

```
# plot germination percentage ~ moisture level
ggplot(data=seeds_agg, aes(x=as.factor(moisture), y=100*germ/total, color=covered)) +
    geom_point() + xlab("moisture level") + ylab("germination percentage") +
    geom_line(aes(group=covered))
```



Answer:

First, from the previous questions, we have shown that "box" is not a significant factor, so we remove it from the model.

Then, the plot shows that: (1) The relationship between germination percentage and moisture level is not monotonic. It looks like quadratic, which increases first and then decreases. So the linear predictor η cannot be just the linear combination of "moisture" and "covered", thus we can try adding the second-order term of the numeric variable "moisture" to the model. (2) For covered and uncovered boxes, the germination percentage reaches the maximum value at different moisture levels, which indicates that there might be an interaction between "moisture" and "covered".

Let's construct a new model adding these two terms. Now the linear predictor of the GLM submodel is: linear predictor: $\eta_i = \beta_0 + \beta_1 moisture_i + \beta_2 1_{covered_i="yes"} + \beta_3 moisture_i \times 1_{covered_i="yes"} + \beta_4 moisture_i^2 + \epsilon_i$

```
# fit a GLM submodel adding the two terms
model.glm_sub2 = glm(cbind(germ,notgerm)~moisture*covered+I(moisture^2),
```

```
                    data=seeds_agg, family=binomial)
summary(model.glm_sub2)
```

```
##
## Call:
## glm(formula = cbind(germ, notgerm) ~ moisture * covered + I(moisture^2),
##     family = binomial, data = seeds_agg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.6458   -4.0303   -0.2256    2.1640    8.5413
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.898092   0.143883  -20.14   <2e-16 ***
## moisture             1.401071   0.050946   27.50   <2e-16 ***
## coveredyes           1.657791   0.142531   11.63   <2e-16 ***
## I(moisture^2)       -0.118978   0.004114  -28.92   <2e-16 ***
## moisture:coveredyes -0.320169   0.023870  -13.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1727.68  on 11  degrees of freedom
## Residual deviance:  300.11  on  7  degrees of freedom
## AIC: 372.53
##
## Number of Fisher Scoring iterations: 6
```

```
# check the significance of the two terms with Chi-square test
anova(model.glm_sub2, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(germ, notgerm)
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                11    1727.68
## moisture          1   164.02        10    1563.67   <2e-16 ***
## covered           1     0.22         9    1563.45   0.6425
## I(moisture^2)     1  1064.33         8     499.12   <2e-16 ***
## moisture:covered  1   199.01         7     300.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

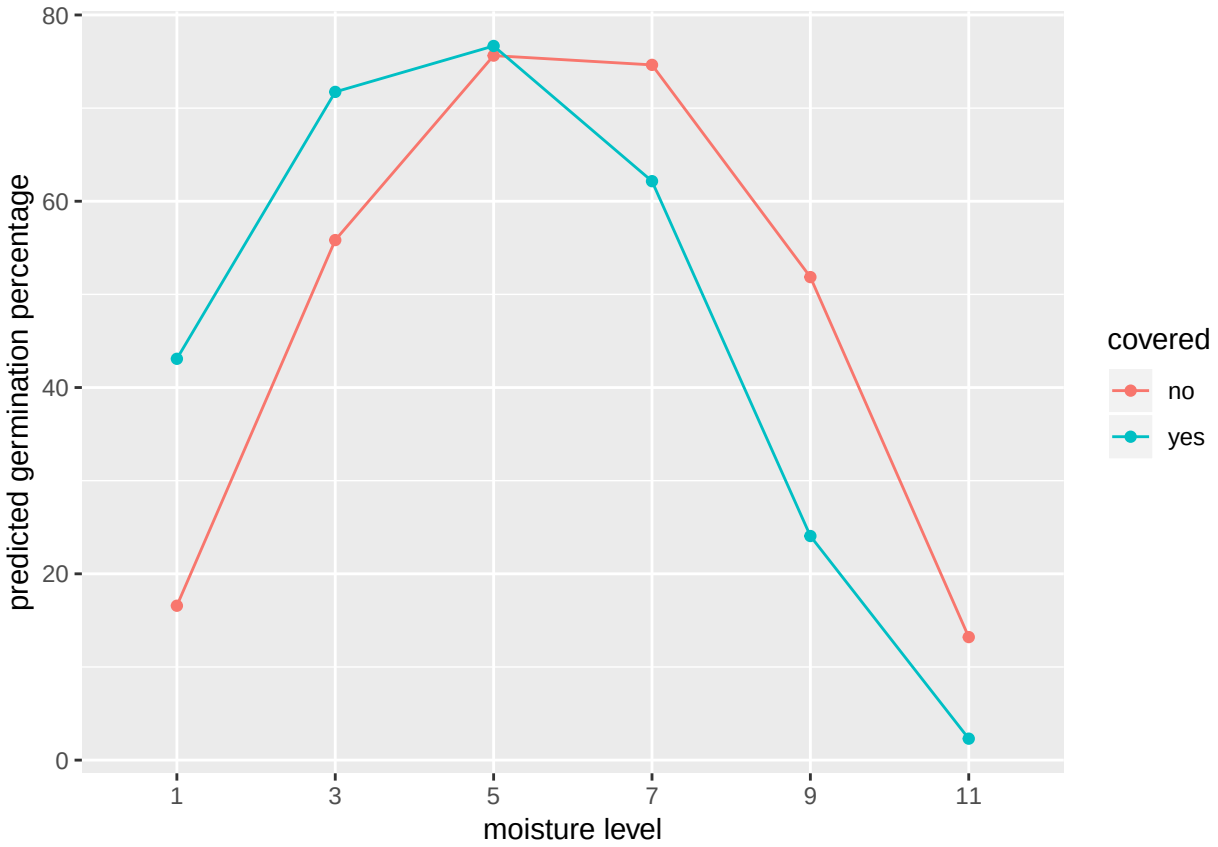
Answer:

The summary shows that all the predictors are now significant in this model. And the chi-square test for adding the two terms step by step also shows that both of the added terms are significant. Thus this is an

appropriate choice of model beyond main effects.

(g)

```
# add the predicted germination percentage (*100%)
seeds_agg$predgerm = 100 * predict(model.glm_sub2, type="response")

# plot predicted germination percentage ~ moisture level
ggplot(data=seeds_agg, aes(x=as.factor(moisture), y=predgerm, color=covered)) +
    geom_point() + labs(x="moisture level", y="predicted germination percentage") +
    geom_line(aes(group=covered))
```

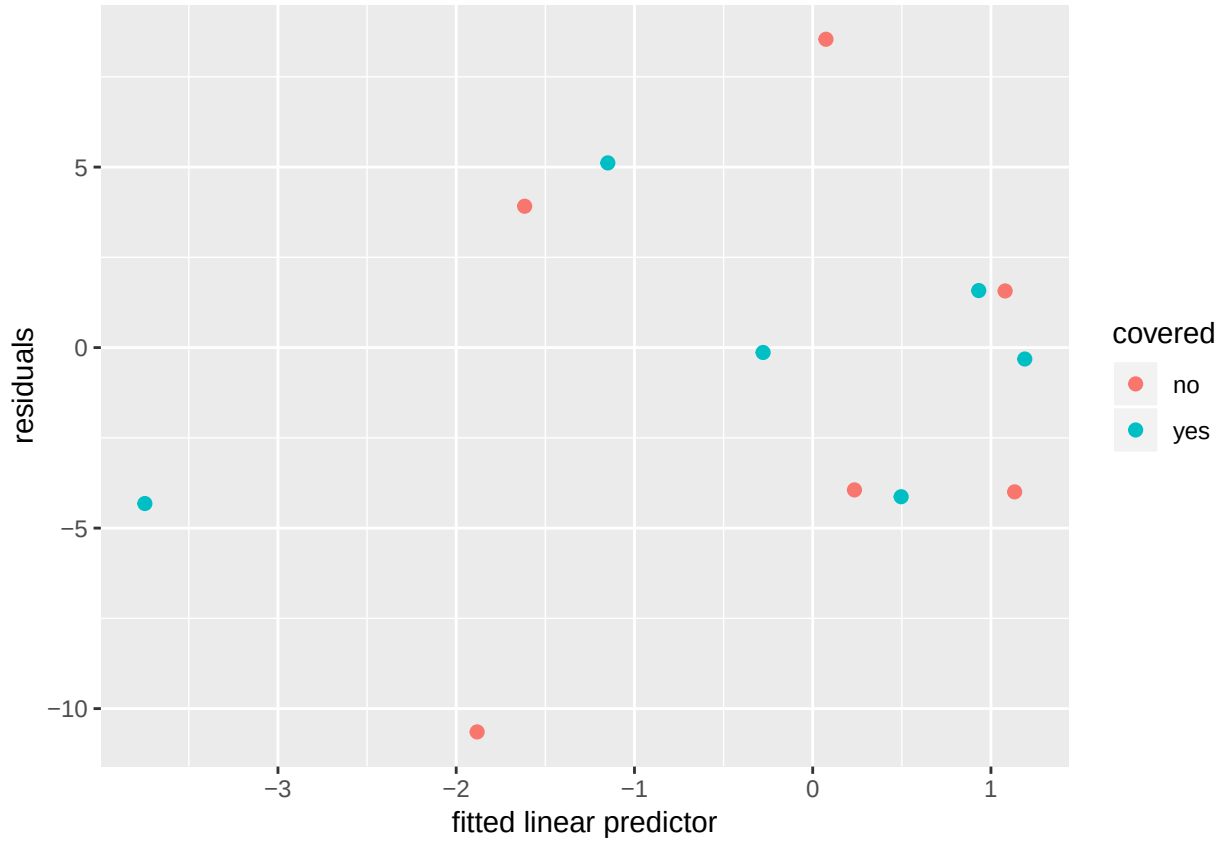


Answer:

For both covered and non-covered seeds, the predicted maximum germination percentage occurs both at moisture level 5.

(h)

```
seeds_agg$linpred = predict(model.glm_sub2)
residuals = residuals(model.glm_sub2)

# plot residuals ~ fitted values (linear predictor)
ggplot(seeds_agg, aes(x=linpred, y=residuals, color=covered)) +
    geom_point(size=2) + xlab("fitted linear predictor") + ylab("residuals")
```

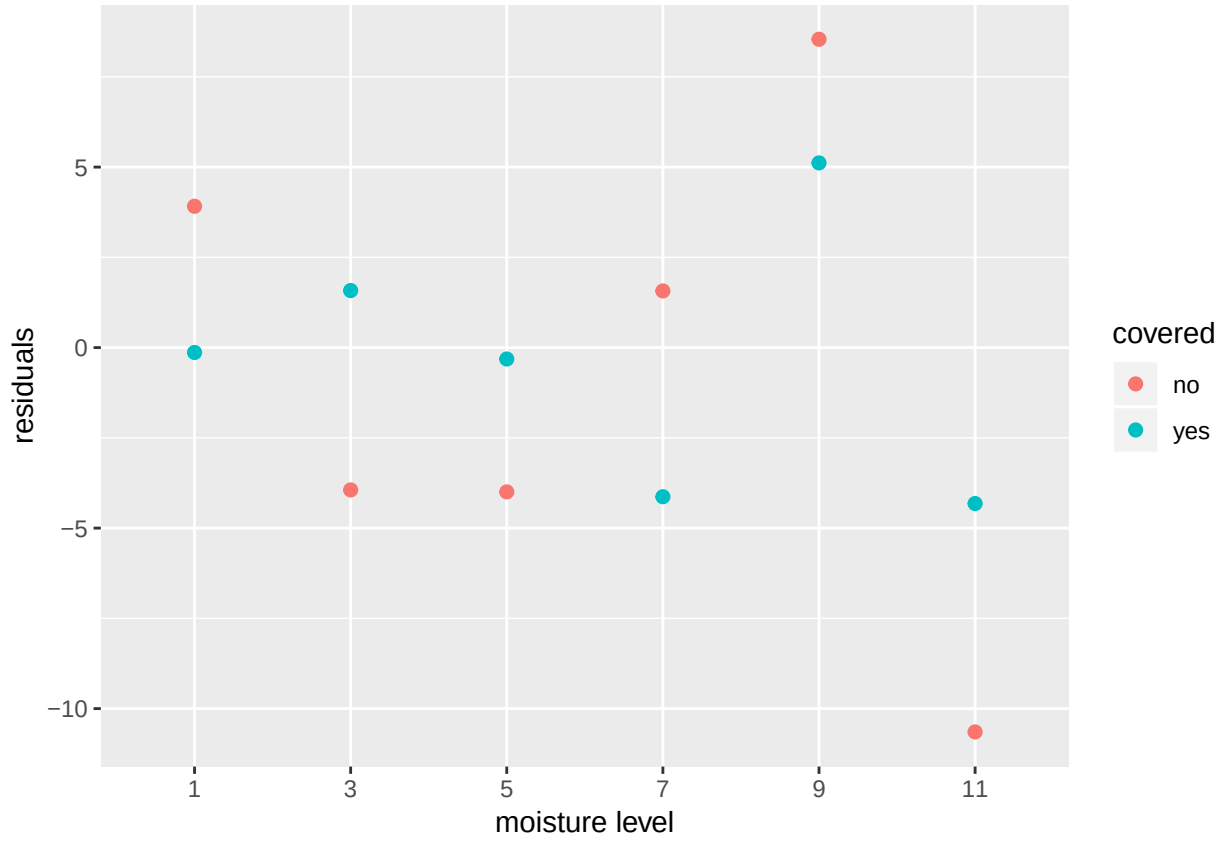


Answer:

The plot of residuals against the fitted linear predictor looks like normally distributed with constant variance (though the variance of residuals for uncovered boxes are larger that for covered boxes), and there is no significant patterns in the residuals. These all indicates that the assumptions of the model are not violated and the model structure is fine.

(i)

```
# plot residuals ~ moisture level
ggplot(seeds_agg, aes(x=as.factor(moisture), y=residuals, color=covered)) +
    geom_point(size=2) + xlab("moisture level") + ylab("residuals")
```



Answer:

In the plot of residuals against the moisture level, the variance of residuals for uncovered boxes are larger that for covered boxes. And in general, the varaince of residuals for both covered and uncovered boxes look slightly larger at larger moisture levels. These may indicate that the model fits better for covered boxes and at lower moisture levels.

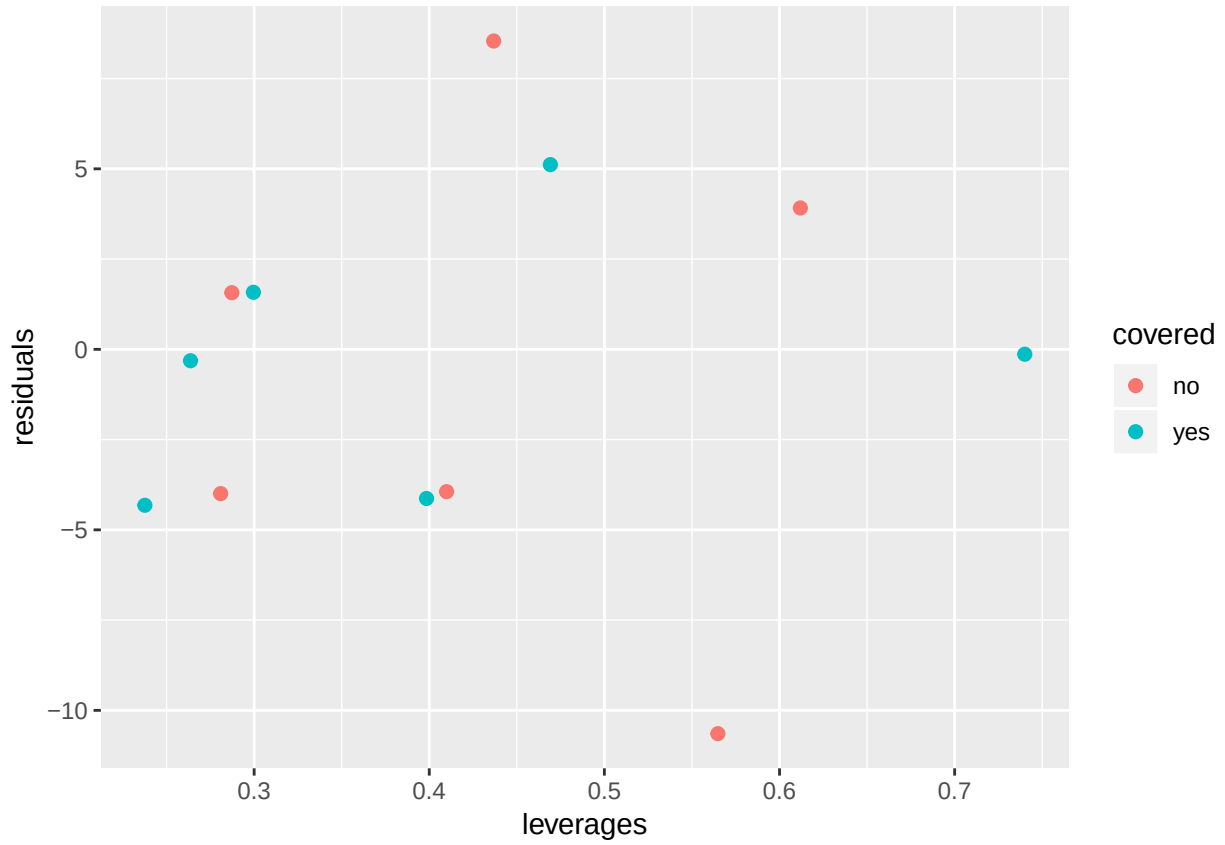
(j)

```
# W matrix
w = seeds_agg$total * seeds_agg$predgerm/100 * (1 - seeds_agg$predgerm/100)
W = diag(w)

# hat matrix
X = model.matrix(model.glm_sub2)
J = t(X) %*% W %*% X
H = W^(1/2) %*% X %*% solve(J) %*% t(X) %*% W^(1/2)

# individual leverages (i.e. hatvalues)
leverages = diag(H)    # or hatvalues(model.glm_sub) or influence(model.glm_sub)$hat
leverages
```

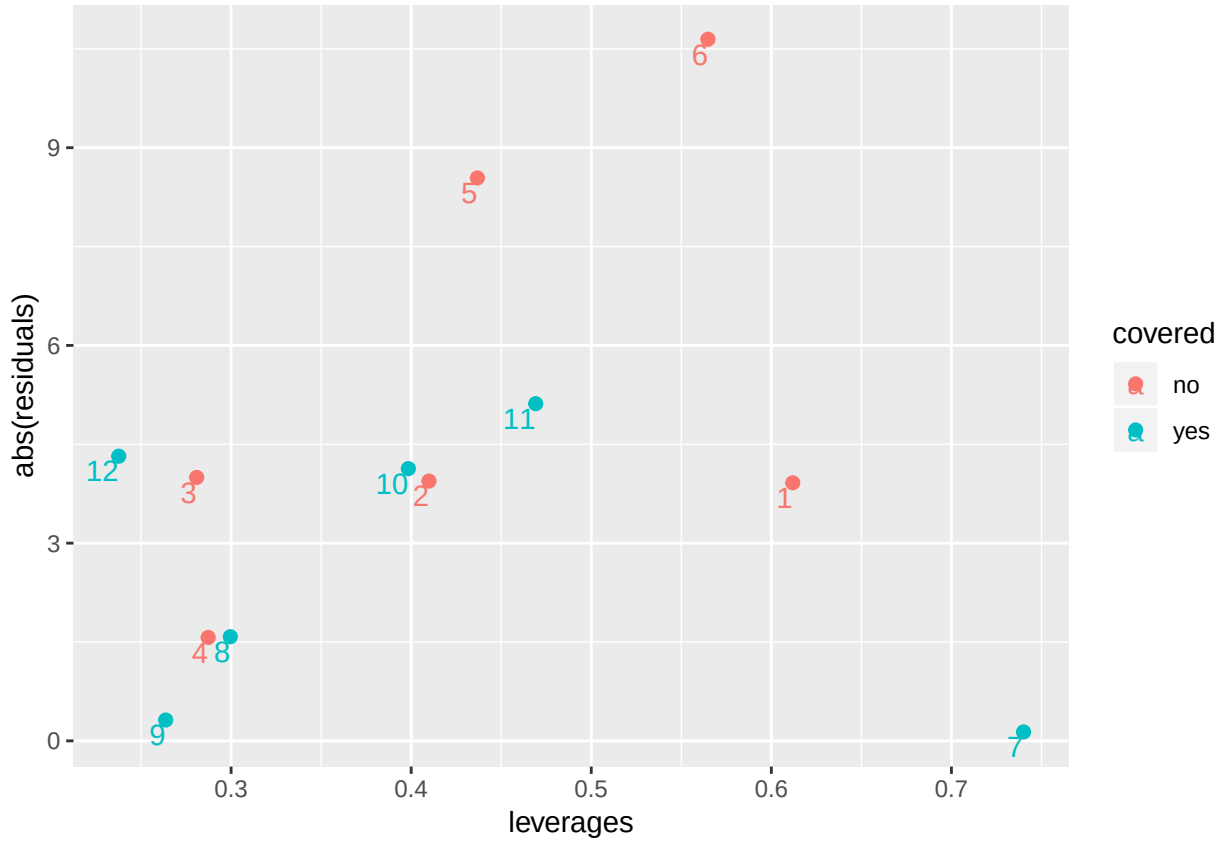
```
##  [1] 0.6119034 0.4098969 0.2808804 0.2872797 0.4367709 0.5648064 0.7400415
##  [8] 0.2995666 0.2636681 0.3984282 0.4691542 0.2376039
```

(k)

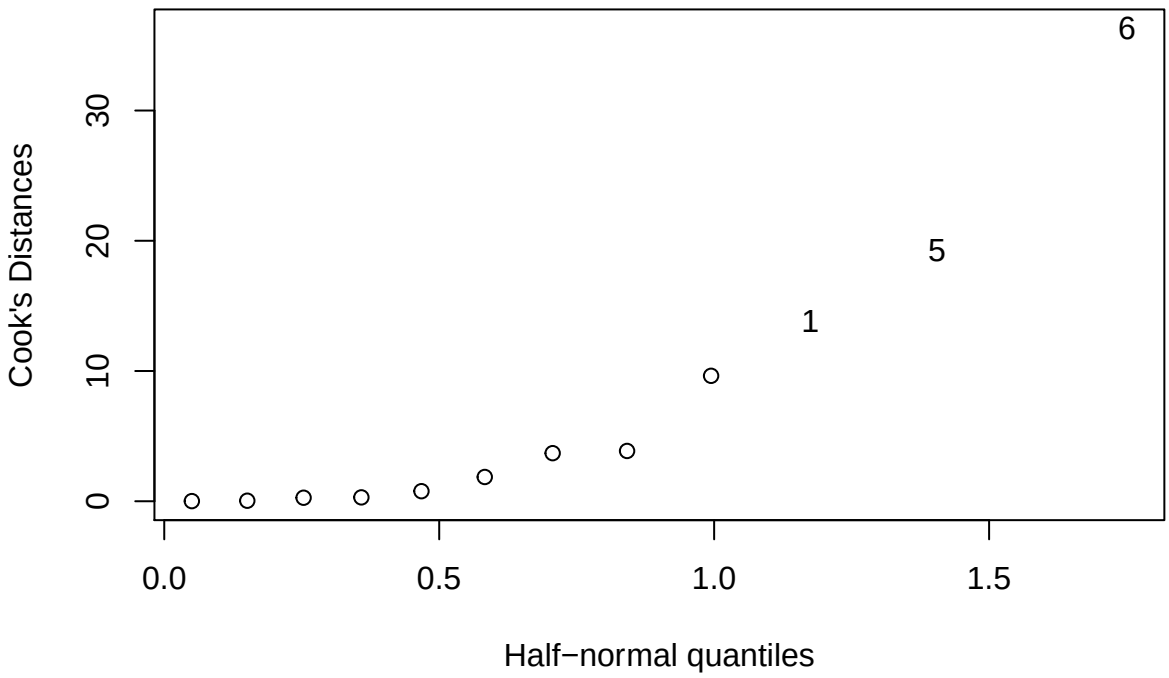
```r
# plot residuals ~ leverages
ggplot(seeds_agg, aes(x=leverages, y=residuals, color=covered)) +
    geom_point(size=2) + xlab("leverages") + ylab("residuals")
```




```r
# plot abs(residuals) ~ leverages
ggplot(seeds_agg, aes(x=leverages, y=abs(residuals), color=covered)) +
    geom_point(size=2) + geom_text(aes(label=row.names(seeds_agg)), hjust=1, vjust=1.2) +
    xlab("leverages") + ylab("abs(residuals)")
```




```
# find influential points with large Cook's Distance via half-normal plot
cook = cooks.distance(model.glm_sub2)
halfnorm(cook, ylab="Cook's Distances", nlab=3)
```




```
# locate observations with large Cook's Distances
seeds_agg[c(1,5,6), ]
```

```
##   moisture covered germ notgerm total predgerm     linpred
## 1        1      no   97     303   400 16.57575 -1.61599861
## 5        9      no  291     109   400 51.85805  0.07435624
## 6       11      no    0     400   400 13.20895 -1.88260859
```

Answer:

In the abs(residuals) against leverages plot, we see that the observations #5 and #6 have large values of residuals, and that observations #1 and #7 havve large values of leverages. However, though being an extreme value in the X range, the observation #7 has nearly zero residual value so it does not largely affect the model for fitting the model quite well. Therefore, we consider the observations #1, #5, and #6 as potential influential points. Then, in the halfnorm plot of the cook's distance, we confirm that these three observations have large cook's distance thus being influential points in the model. The data details of these three observations are also shown above, which are all in the category of uncovered boxes. This again indicates that the model fits not well for uncovered boxes.