Applied Psycholinguistics
ANCOVA

Hyunah Ahn

November 19, 2020

1 Pseudoreplication

The textbook discusses an experiment done on 6 rats (Snedecor & Cochran, 1980). They treated the rats three different ways and wanted to see if the different treatments led to varying levels of glycogen in the rats' livers.

```
rats <- read.csv("C:/Users/hyuna/OneDrive/Documents/01SNU/03GraduateSeminar/crawley/data/rats.csv")
rats

##    Glycogen Treatment Rat Liver
## 1       131         1   1     1
## 2       130         1   1     1
## 3       131         1   1     2
## 4       125         1   1     2
## 5       136         1   1     3
## 6       142         1   1     3
## 7       150         1   2     1
## 8       148         1   2     1
## 9       140         1   2     2
## 10      143         1   2     2
## 11      160         1   2     3
## 12      150         1   2     3
## 13      157         2   1     1
## 14      145         2   1     1
## 15      154         2   1     2
## 16      142         2   1     2
## 17      147         2   1     3
## 18      153         2   1     3
## 19      151         2   2     1
## 20      155         2   2     1
## 21      147         2   2     2
## 22      147         2   2     2
## 23      162         2   2     3
## 24      152         2   2     3
## 25      134         3   1     1
## 26      125         3   1     1
## 27      138         3   1     2
## 28      138         3   1     2
## 29      135         3   1     3
## 30      136         3   1     3
## 31      138         3   2     1
## 32      140         3   2     1
## 33      139         3   2     2
## 34      138         3   2     2
## 35      134         3   2     3
## 36      127          3   2     3
```

As you can see from the table above, the records from the first treatment are shown in the first twelve rows. The next twelve rows shows measurements from the second treatment, and the last twelve the third treatment. There are a total of 36 measurements from six rats. In each treatment, there were two rats, so, the first six rows of the first treatment shows rat number 1 and the latter six 2. The same goes for the second and third treatment but the numbers here only means that there were two rats in each condition, the numbers 1 and 2 do not indicate absolute identifiers for six rats. Lastly, the livers of each rat was cut up into three parts: left (1), center (2), and right (3). Then, from each part, two separate preparations were made. So, you can see in the liver column, the three parts of the liver is repeated twice each.

When the levels of each factor is labeld with a number, R might mistakenly think that the factors are continuous rather than categorical; to prevent this mistake, we will tell R that these are categorical values as below.

```
attach(rats)
Treatment<-factor(Treatment)
Rat<-factor(Rat)
Liver<-factor(Liver)
```

Now, remember I said they wanted to know the effect of treatments on the level of glycogen? Then, let's build a model as we learned in previous sessions.

```
model1<-aov(Glycogen~Treatment)
summary(model1)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment    2   1558   778.8    14.5 3.03e-05 ***
## Residuals   33   1773    53.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output table says that the effect of Treatment is significant with the F value of 14.5 and an extremely small p value of 0.00003!!! But this analysis has a problem.

We calculate the F value by dividing the mean square value of a factor of interest by that of errors (residuals). The mean square values are calcualted by dividing the sum of squares by degrees of freedoms. The degree of freedom for the Treatment factor is 2 because there were three different levels the Treatment. And the degree of freedom for errors should be n*(k-1), that is, the number of factors multiplied by the number of levels in each factor - 1. Then:

3 treatments x (2 rats - 1) = 3

The degree of freedom for residuals should be 3, not 33 as the output table above shows! What's going on? If one wrongfully assumes that the liver samples are from 12 different rats, one might calculate the degree of freedom for errors as below:

3 treatments x (12 rats -1) = 33

But this is wrong. As you remember, there were only 6 rats, from each rat's liver, we took three parts, and made two preparations from each rat. If the error degree of freedom gets larger like this for a wrong reason, the mean square value for the error term gets smaller, which makes the F value larger as a result. And we get significance for a wrong reason.

We can put these error terms in the ANOVA model as below.

```
model2<-aov(Glycogen~Treatment+Error(Treatment/Rat/Liver))
summary(model2)

##
## Error: Treatment
##           Df Sum Sq Mean Sq
## Treatment  2   1558   778.8
```

```
##
## Error: Treatment:Rat
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  3  797.7   265.9
##
## Error: Treatment:Rat:Liver
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 12    594    49.5
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 18    381   21.17
```

Now, the error degree of freedom should come only from the number of levels in a factor and the number of replications in each level. There were three levels in the factor ‘Treatment’ and in each level, there were two replications (repetitions) via rats, meaning each condition was repeatedly tested through two rats. Therefore, the error degree of freedom should be 3. From the output table, divide the mean of squares obtained from the factor ’Treatment’ by the mean of squares obtained from the error term considering only the rats (not three different liver parts and/or two different preparations).

Then, the true F value for this analysis should be:

$$\frac{778.8}{265.9} = 2.928921$$

```
qf(0.95, 2, 3)

## [1] 9.552094
```

Since the threshold of the F value when the factor df is 2 and the error df is 3 at the .05 significance level is 9.552094. The F value obtained above cannot be significant. If you would like the exact p value for the F value above, use the function below.

```
pf(2.92, 2, 3)

## [1] 0.8023014
```

As you can see, the percentile ranking of the F value 2.92 with 2 for the degree of freedom for the factor and 3 for the error degree of freedom is 0.80, which shows that the effect of Treatment is not significant at all. The variance of the glycogen level seems rather random.

You should be careful when you use ANOVA when measurements were made repeatedly from the same subjects so you can avoid inadvertently treating multiple observations from the same participant as independent from one another.

2 ANCOVA

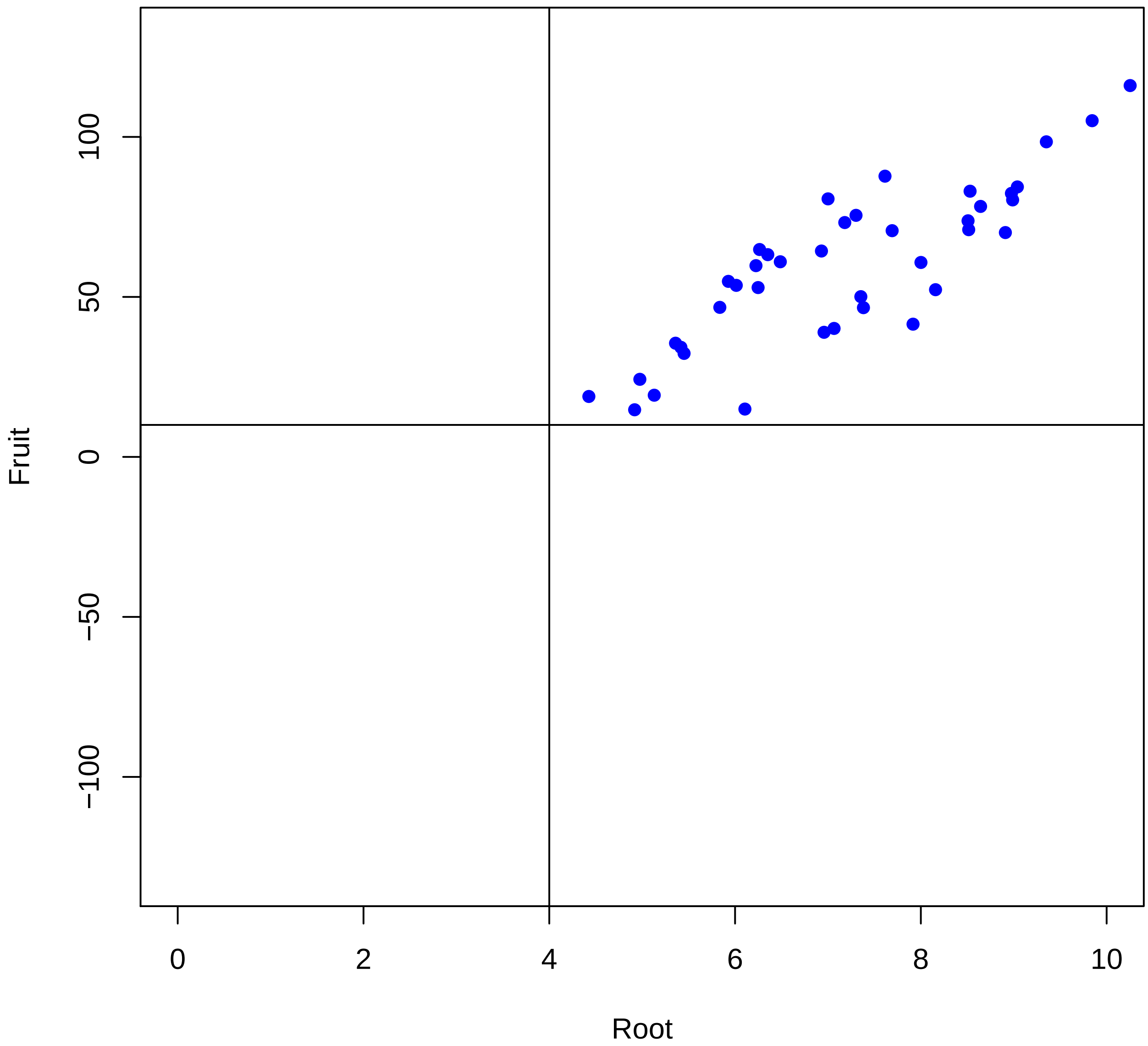
ANCOVA, in simple terms, is a combination of ANOVA and regression. ANOVA is for categorical independent factors whereas regression can have either categorical or continuous independent factors. For example, a person's weight can be dependent upon one's sex and age. Sex is a categorical factor whereas age is a continuous factor. Your weight covaries with your age. When you have to factor into covairance, you use ANCOVA, the analysis of covariance.

On p. 186, an example is given using data from an experiment that tested the weight of seeds from a plant by its size and by whether it was grazed or not. The size was measured by measuring the diameter of the topstock of the root.

```
compensation<-read.csv("C:/Users/hyuna/OneDrive/Documents/01SNU/03GraduateSeminar/crawley/data/ipomopsis.c
attach(compensation)
compensation

##      Root  Fruit  Grazing
## 1   6.225  59.77 Ungrazed
## 2   6.487  60.98 Ungrazed
## 3   4.919  14.73 Ungrazed
## 4   5.130  19.28 Ungrazed
## 5   5.417  34.25 Ungrazed
## 6   5.359  35.53 Ungrazed
## 7   7.614  87.73 Ungrazed
## 8   6.352  63.21 Ungrazed
## 9   4.975  24.25 Ungrazed
## 10  6.930  64.34 Ungrazed
## 11  6.248  52.92 Ungrazed
## 12  5.451  32.35 Ungrazed
## 13  6.013  53.61 Ungrazed
## 14  5.928  54.86 Ungrazed
## 15  6.264  64.81 Ungrazed
## 16  7.181  73.24 Ungrazed
## 17  7.001  80.64 Ungrazed
## 18  4.426  18.89 Ungrazed
## 19  7.302  75.49 Ungrazed
## 20  5.836  46.73 Ungrazed
## 21 10.253 116.05   Grazed
## 22  6.958  38.94   Grazed
## 23  8.001  60.77   Grazed
## 24  9.039  84.37   Grazed
## 25  8.910  70.11   Grazed
## 26  6.106  14.95   Grazed
## 27  7.691  70.70   Grazed
## 28  8.988  80.31   Grazed
## 29  8.975  82.35   Grazed
## 30  9.844 105.07   Grazed
## 31  8.508  73.79   Grazed
## 32  7.354  50.08   Grazed
## 33  8.643  78.28   Grazed
## 34  7.916  41.48   Grazed
## 35  9.351  98.47   Grazed
## 36  7.066  40.15   Grazed
## 37  8.158  52.26   Grazed
## 38  7.382  46.64   Grazed
## 39  8.515  71.01   Grazed
## 40  8.530  83.03   Grazed
```

The first columns shows the size (the diameter of the topstock of the root), the second column shows the weight of seeds (Fruit), and the last shows whether each plant was grazed or not. Forty different observations were made (on probably forty different plants). Let's plot the data.

```
plot(Fruit~Root, pch=16, col = "blue", xlim=c(0, 10), ylim=c(-130, 130))
abline(v=4, h=10)
```



Can you see the relationship between the size of Roots (Root) and the weight of seeds (Fruit)? Indeed, there seems to be a positive correlation between the two variables. Then, how about grazing?

```
plot(Fruit~Grazing, col="lightgreen")

## Warning in xy.coords(x, y, xlabel, ylabel, log):        NA
## Warning in min(x):  min      Inf
## Warning in max(x):  max      -Inf
## Error in plot.window(...):      'xlim'
```

Let's first learn how to read the box plots. The dark middle line in the middle of the box is the mean (and at the same time the second quartile of the data), the whiskers indicate the minimum value (on the bottom) and the maximum value (on the top) within each group. The bottom of the box is the first quartile of the data, which means any data points below the bottom of the box belongs to the bottom 25% of the data. The top of the box indicates the third quartile of the data, which, again, means that all data points below the third quartile are below the 75% cut point. The results are unexpected. We expected that ungrazed plants would have more seeds but the mean value seems to be higher on the grazed side. We can fit an anova model to see if the effect of 'Grazing' is significant.

```
model1<-aov(Fruit~Grazing)
summary(model1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Grazing      1   2910  2910.4   5.309 0.0268 *
## Residuals   38  20833   548.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output table shows that the p-value is less than .03. So, for now, we can have an interim conclusion that Grazing is a significant factor that determines the weight of seeds.

Since the size of the root is not categorical but continuous, let's fit a regression model.

```
model2<-lm(Fruit~Root)
summary(model2)

##
## Call:
## lm(formula = Fruit ~ Root)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.3844 -10.4447  -0.7574  10.7606  23.7556
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -41.286     10.723  -3.850 0.000439 ***
## Root          14.022      1.463   9.584  1.1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 38 degrees of freedom
## Multiple R-squared:  0.7073,Adjusted R-squared:  0.6996
## F-statistic: 91.84 on 1 and 38 DF,  p-value: 1.099e-11
```

The output shows that the size of the root is also a significant factor. But can we put both categorical and continuous factors in one `aov()` model?

```
model3<-aov(Fruit~Root*Grazing)
summary(model3)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## Root          1  16795   16795 359.968  < 2e-16 ***
## Grazing       1   5264    5264 112.832 1.21e-12 ***
## Root:Grazing  1      5       5   0.103     0.75
## Residuals    36   1680      47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What if we put it in a regression model?

```
model4<-lm(Fruit~Root*Grazing)
summary(model4)

##
## Call:
## lm(formula = Fruit ~ Root * Grazing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3177  -2.8320   0.1247   3.8511  17.1313
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -125.173     12.811  -9.771 1.15e-11 ***
## Root                    23.240      1.531  15.182  < 2e-16 ***
## GrazingUngrazed         30.806     16.842   1.829   0.0757 .  
## Root:GrazingUngrazed     0.756      2.354   0.321   0.7500    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.831 on 36 degrees of freedom
## Multiple R-squared:  0.9293,Adjusted R-squared:  0.9234 
## F-statistic: 157.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

Do they look similar but still different? Or should we try this?

```
summary.aov(model3)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Root          1  16795   16795 359.968  < 2e-16 ***
## Grazing       1   5264    5264 112.832 1.21e-12 ***
## Root:Grazing  1      5       5   0.103     0.75    
## Residuals    36   1680      47                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(model3)
```

```
## 
## Call:
## aov(formula = Fruit ~ Root * Grazing)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -17.3177  -2.8320   0.1247   3.8511  17.1313 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -125.173     12.811  -9.771 1.15e-11 ***
## Root                    23.240      1.531  15.182  < 2e-16 ***
## GrazingUngrazed         30.806     16.842   1.829   0.0757 .  
## Root:GrazingUngrazed     0.756      2.354   0.321   0.7500    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.831 on 36 degrees of freedom
## Multiple R-squared:  0.9293,Adjusted R-squared:  0.9234 
## F-statistic: 157.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
summary.aov(model4)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Root          1  16795   16795 359.968  < 2e-16 ***
## Grazing       1   5264    5264 112.832 1.21e-12 ***
## Root:Grazing  1      5       5   0.103     0.75    
## Residuals    36   1680      47                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(model4)
```

```
##
## Call:
## lm(formula = Fruit ~ Root * Grazing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3177  -2.8320   0.1247   3.8511  17.1313
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -125.173     12.811  -9.771 1.15e-11 ***
## Root                   23.240      1.531  15.182  < 2e-16 ***
## GrazingUngrazed        30.806     16.842   1.829   0.0757 .
## Root:GrazingUngrazed    0.756      2.354   0.321   0.7500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.831 on 36 degrees of freedom
## Multiple R-squared:  0.9293,Adjusted R-squared:  0.9234
## F-statistic: 157.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

Please note that ANOVA is just another type of general linear regression. Therefore, you produce the same results. It's just that the ANOVA output will give you ideas about how great (or little) the variances caused by the factors are in comparison to the variance cased by error. On the other hand, the regression output will tell you the intercept and slope of the regression line drawn by the data.

For now, let's focus on the ANOVA output only.

```
summary(model3)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## Root          1  16795   16795 359.968  < 2e-16 ***
## Grazing       1   5264    5264 112.832 1.21e-12 ***
## Root:Grazing  1      5       5   0.103     0.75
## Residuals    36   1680      47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see from the table above, both the covariate Root and the categorical factor Grazing have a significant effect on the weight of the seeds (Fruit) but the interaction term is not significant. Then, maybe we don't need the interaction term. But when you simplify a model, do not rely only on the p value of a term, run an `anova()` to compare models. A simpler model that does not include an interaction term will be a model of only main effects as in model 5.

```
model5<-aov(Fruit~Root + Grazing)
```

The numbers for Root remain the same while those for Grazing changed. Now, let's compare the model with an interaction term (model3) and the one without (model5).

```
anova(model3, model5)

## Analysis of Variance Table
##
## Model 1: Fruit ~ Root * Grazing
## Model 2: Fruit ~ Root + Grazing
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 1679.7
## 2     37 1684.5 -1   -4.8122 0.1031   0.75
```

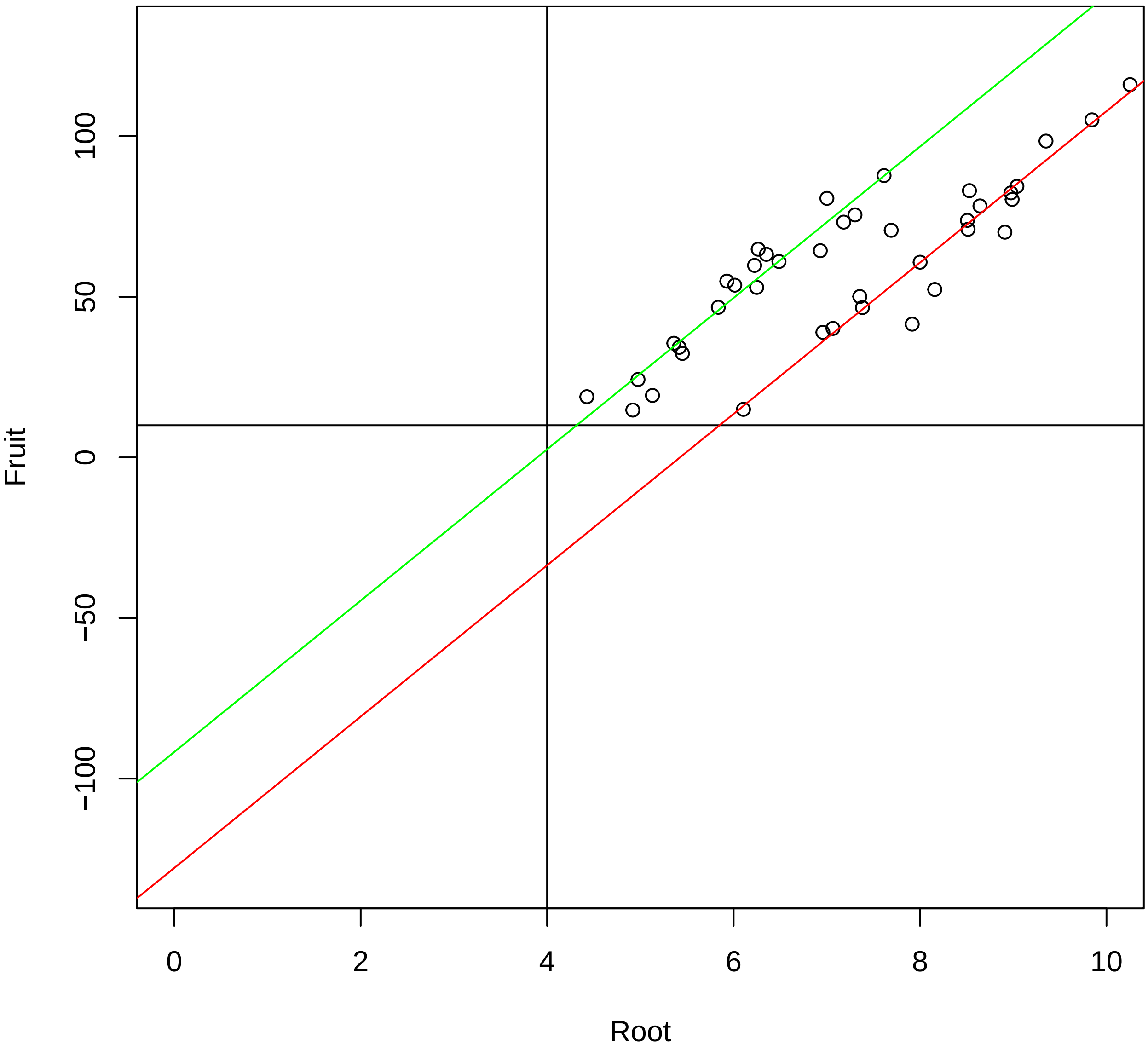
The F value is 0.1031 with the p value of 0.75. This means that the simpler model is not significantly different from the more complex model and you're safe to go ahead and reduce the model to a simpler one.

Now we've decided to use a simpler model, let's run a regression model to find out the intercept and slope for the model we're going to use.

```
model6<-lm(Fruit~Grazing+Root)
summary(model6)

##
## Call:
## lm(formula = Fruit ~ Grazing + Root)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1920  -2.8224   0.3223   3.9144  17.3290
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -127.829      9.664  -13.23 1.35e-15 ***
## GrazingUngrazed   36.103      3.357   10.75 6.11e-13 ***
## Root              23.560      1.149   20.51  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.747 on 37 degrees of freedom
## Multiple R-squared:  0.9291,Adjusted R-squared:  0.9252
## F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

Now, how will you interpret the ouput? It says the estimate for the `Intercept` is -127.829. The `Intercept` is the baseline when plants are grazed and its root is 0. The weight of the seeds will increase 36.103 when the same plant is ungrazed and for every unit of Root increase, the weight of the seeds will increase about 23.560.

```
plot(Fruit~Root, pch=21, bg=(1+as.numeric(Grazing)), xlim=c(0, 10), ylim=c(-130, 130))

## Warning in FUN(X[[i]], ...):         NA

abline(v=4, h=10)
abline(a=-127.829, b=23.56, col="red")
abline(a=-127.829+36.103, b=23.56, col="green")
```


Fruit
100
50
0
−50
−100
0
2
4
6
8
10
Root