Análysis de Datos de Orquídeas

Ciro (Haozhe) Yu

2024-04-09

Data Import

- Import Data from the Tabla_de_medidas.xlsx data fila:

```
raworchids=read_excel("Data/Tabla_de_medidas.xlsx") %>% as_tibble()
```

Data Cleaning

Data Cleaning Steps for This Project Include:

- Nicknaming the Attributes to make them easier to code.
- Converting the light, species, and medium attributes into factors.
- Inserting in the Data for the *Cattleya maxima* in darkness and IBA concentration of 5 mg/L.
- Average the Observations at each root to generalize to each plant.
- Assign the higher root number to the plant if 2 root values were included incorrectly.

```
#Create Simpler Names
names(raworchids) <- c("es","med","p","nr","r","lr","l")

# Assign Values to Merged Data
orchids <-  raworchids %>% fill(es, med, p, nr, l) %>%
  mutate(l = as.factor(l)) %>%
  mutate(es = as.factor(es)) %>%
  mutate(med = as.factor(med))

# Remove Blank data (will create dummy data later)
orchids <-   filter(orchids,!grepl("N",p) | is.na(p))

# Sum values
orchidsum <-
  orchids %>%
  group_by(es, med, l, p) %>%
  summarise(rnum = mean(lr),nr1=last(nr),nr2=n())
```

```
## `summarise()` has grouped output by 'es', 'med', 'l'. You can override using
## the `.groups` argument.
```

```r
# Get correct Root Number Values
orchidsrn <- orchidsum %>% mutate(nr = ifelse(nr1 == 0, 0,
                                              ifelse(nr1 >= nr2, nr1, nr2))) %>%
  ungroup()

# Create Dummy/Assumed Data for 11, 0.5
orchidsfinal <- orchidsrn %>% select(-nr2, -nr1) %>%
  add_row(
    es = c(rep(as.factor(11), 19)),
    med = c(rep(as.factor(0.5), 19)),
    l = c(rep(as.factor("Oscuridad"), 11), rep(as.factor("Luz"), 8)),
    p = c(as.character(1:19)),
    rnum = 0,
    nr = 0
  ) %>%
  mutate(IBAconc = case_when(med == "P4" ~ 0,
                             med == '0.5' ~ 5,
                             med == '1' ~ 1))

orchidsfinal$med <- relevel(orchidsfinal$med,ref = "P4")
orchidsfinal$l <- relevel(orchidsfinal$l,ref = "Oscuridad")
```

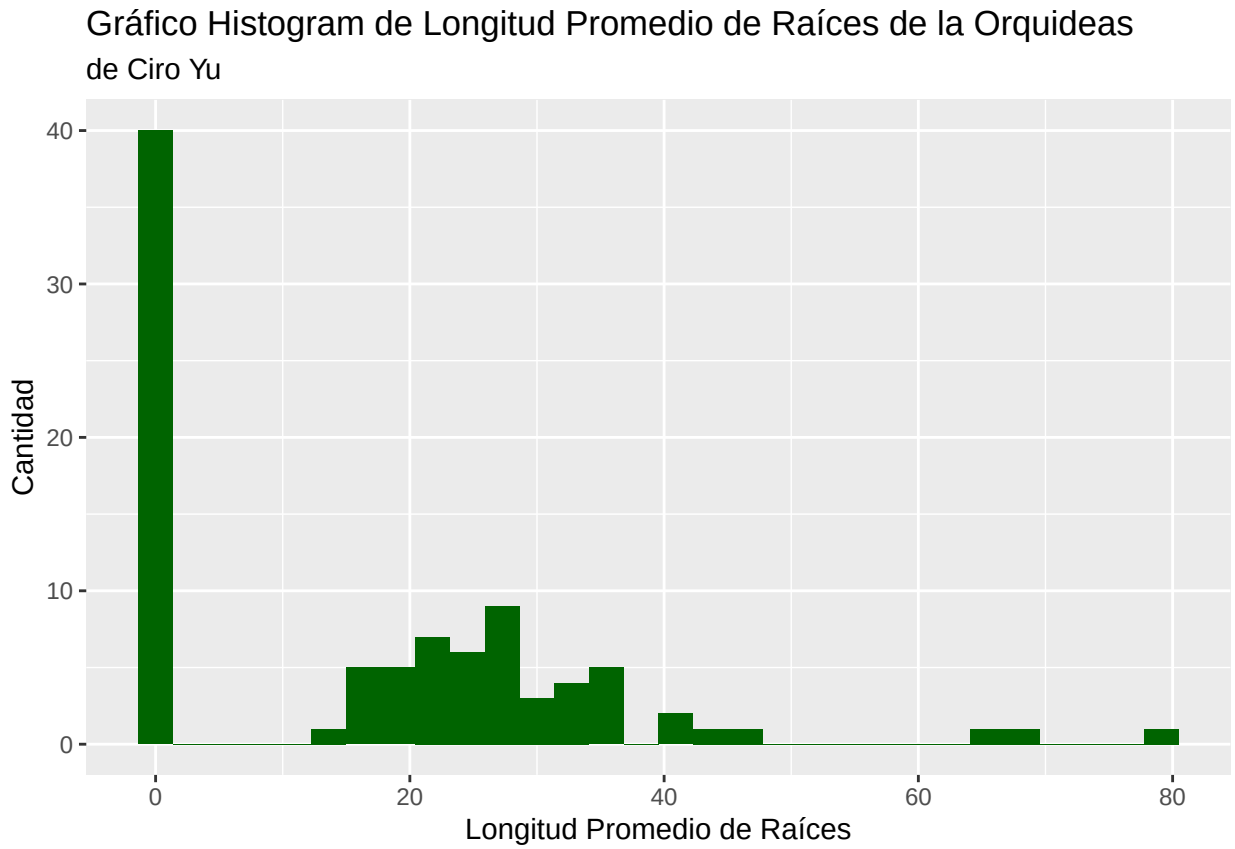
Data Analysis

Exploratory Data Analysis

We first plotted the distribution of the 2 analysis variables, Average Root Length and Number of Root.

```r
# Logidud de Raices
ggplot(orchidsfinal,aes(rnum)) +
  geom_histogram(fill="darkgreen") +
    labs(
    title = paste("Gráfico Histogram de Longitud Promedio de Raíces de la Orquideas"),
    subtitle = "de Ciro Yu"
  ) +
  xlab("Longitud Promedio de Raíces") +
  ylab("Cantidad")
```

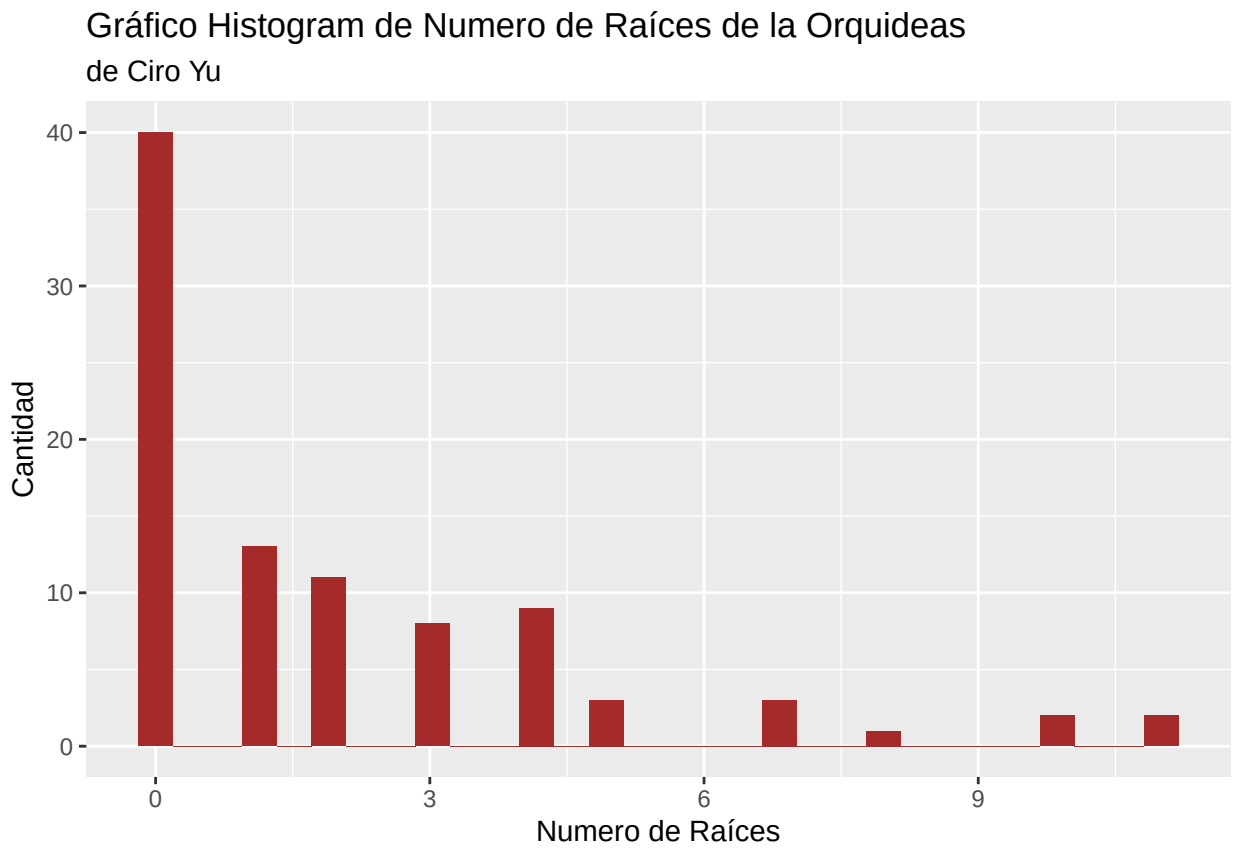
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see here that our data has a lot of 0s and a few large outlines on the right. While those points raise a concern, the central distribution of the roots looks approximately normal. Thus we continue with our analysis.

```
# Numero de Raices
ggplot(orchidsfinal,aes(nr)) +
  geom_histogram(fill="brown") +
    labs(
    title = paste("Gráfico Histogram de Numero de Raíces de la Orquideas"),
    subtitle = "de Ciro Yu"
  ) +
  xlab("Numero de Raíces") +
  ylab("Cantidad")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see here that the data does not look normal at all. There are a very large number of 0s and a big right tail of large numbers. As a result, we decided to use a GLM approach using models based on the poisson and negative binomial distributions. Further preliminary analysis (not shown) demonstrated that the negative binomial distribution was superior.

ANOVA Analysis

Data for Average Root Length was analyzed with a 2 Way ANOVA. Each species was analyzed as a separate experiment. The linear effects model below was fit to the data for Average Root Length. the model for this analysis is below:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

Where:

- Y_{ijk} = the average root length of the individual plant.
- α_i = the fixed effect of the media the plant was cultured in.
- β_j = the fixed effect of the condition (light vs darkness) the plant was grown in.
- $\alpha\beta_{ij}$ = the fixed effect of the interaction between condition and media.
- ϵ_{ijk} iid $N(0, \sigma^2)$ = The residual variance.

Meanwhile, the data for number of roots was analyzed with a Negative Binomial regression model. The model for this analysis is below:

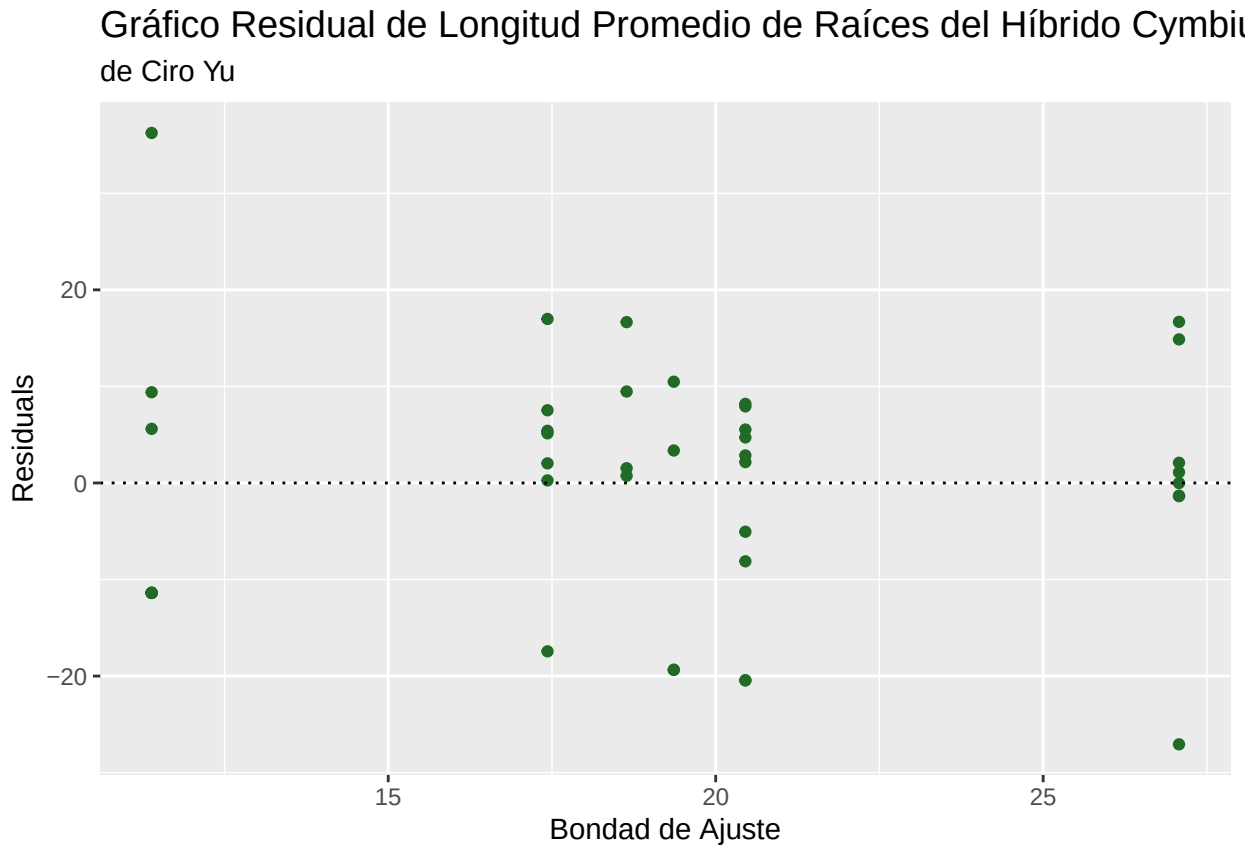
$$log(\mu) = \alpha_i + \beta_j + \alpha\beta_{ij}$$

- μ - the expected value of the number of roots
- α_i = the fixed effect of the media the plant was cultured in.
- β_j = the fixed effect of the condition (light vs darkness) the plant was grown in.
- $\alpha\beta_{ij}$ = the fixed effect of the interaction between condition and media.

Subexperiment 1: *Cymbidium* Hybrid (10)

Trait 1: Average Root Length Assumption Testing:

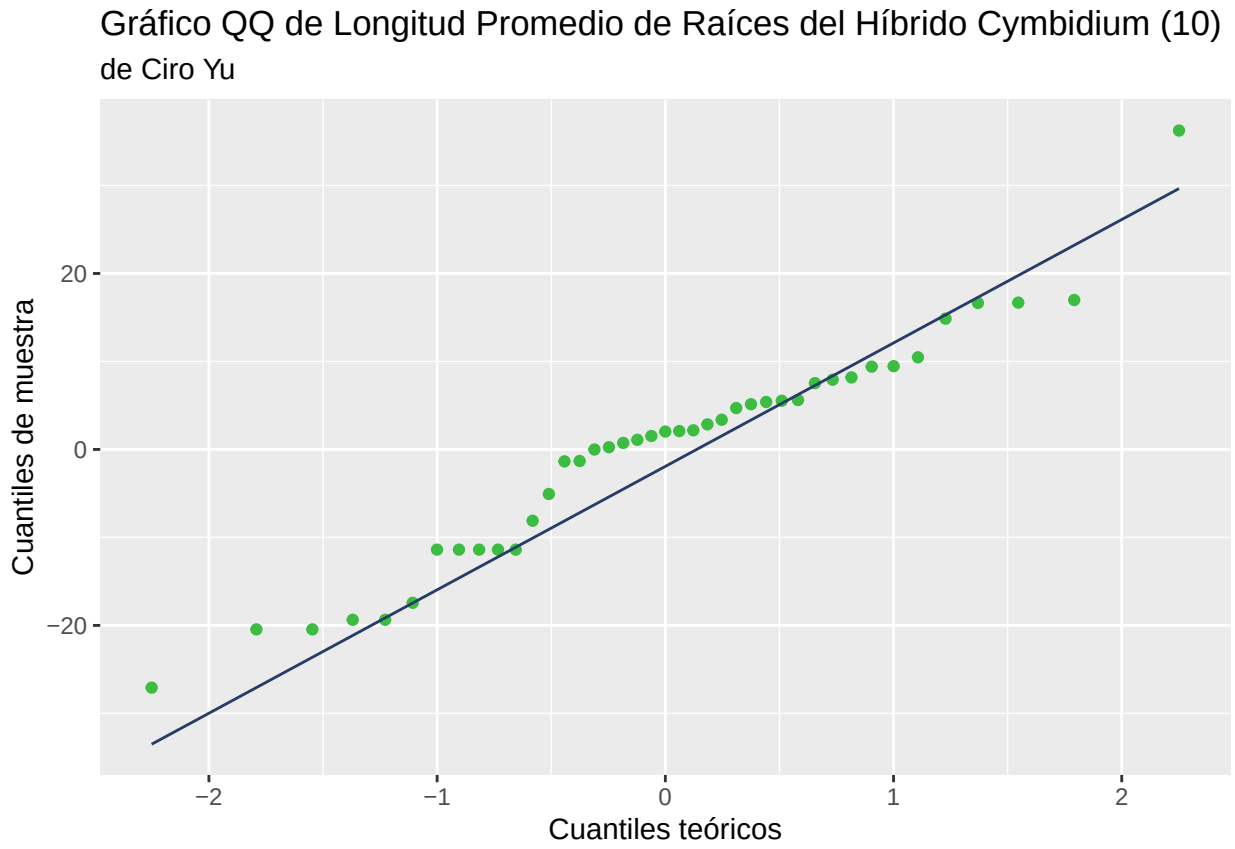
```
cym <-  orchidsfinal %>% filter(es=="10")
cymrnummIBA <- lm(rnum~IBAconc*l,cym)

#Residuals
cymrnummIBAr <- tibble(
  "fit" = cymrnummIBA$fitted.values,
  "resid" = cymrnummIBA$residuals) %>%
  add_column("rstandardized"=rstandard(cymrnummIBA))

ggplot(cymrnummIBAr,aes(x=fit,y=resid))+
  geom_jitter(color="#226b26")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Gráfico Residual de Longitud Promedio de Raíces del Híbrido Cymbiudium (10)"),
       subtitle = "de Ciro Yu")+
  ylab("Residuals")+
  xlab("Bondad de Ajuste")+
  theme(plot.title = element_text(size = 14))
```




```
bptest(cymrnummIBA,studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  cymrnummIBA
## BP = 2.4342, df = 3, p-value = 0.4873
```

```
ggplot(data = cymrnummIBAr, aes(sample = resid))+
  geom_qq( color="#3dbc42")+
  geom_qq_line( color="#273d67")+
  labs(
    title = paste("Gráfico QQ de Longitud Promedio de Raíces del Híbrido Cymbidium (10)"),
    subtitle = "de Ciro Yu"
  ) +
  xlab("Cuantiles teóricos")+
  ylab("Cuantiles de muestra")
```




```
shapiro.test(cymrnummIBA$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cymrnummIBA$residuals
## W = 0.95963, p-value = 0.1526
```

The residual plot lacks any patterns, indicating constant error variance. Additionally, the Breusch-Pagan test is not significant, further indicating constant error variance.

The residual plot lacks any patterns, which also indicates linearity.

The QQ plot is reasonably well fitted to the line, indicating normality of error variance. Additionally, the Shapiro-Wilk test is not significant, further indicating normality.

Independence is assumed due to the randomization of the experiment.

The assumptions for linear regression are met and we can proceed.

ANOVA Analysis

```
summary(cymrnummIBA)
```

```
##
## Call:
## lm(formula = rnum ~ IBAconc * l, data = cym)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.074 -11.388   2.028   7.536  36.242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20.453      3.717   5.502 2.96e-06 ***
## IBAconc        -1.813      1.221  -1.485   0.1459
## lLuz           -3.022      5.656  -0.534   0.5964
## IBAconc:lLuz    3.742      1.784   2.097   0.0429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.15 on 37 degrees of freedom
## Multiple R-squared:  0.1379, Adjusted R-squared:  0.06802
## F-statistic: 1.973 on 3 and 37 DF,  p-value: 0.1349
```

The Overall F value of the model is 1.9731668 with 3 degrees of freedom for the model and 37 degrees of freedom for the error. This gives us a total p value of 0.1348853. At $\alpha = 0.05$, p $> \alpha$, so we fail to reject the null hypothesis. That means for average root length, there is not enough evidence to support that either light or media influence average root length for our *Cymbidium* hybrid.

Trait 2: Root Number Assumption Testing:

```
dispersiontest(glm(nr~IBAconc*l,cym,family=poisson))
```

```
##
##  Overdispersion test
##
## data:  glm(nr ~ IBAconc * l, cym, family = poisson)
## z = 2.481, p-value = 0.006551
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.148793
```

From our exploratory analysis, we have already determined that the because the number of roots is an integer number, and that the distribution of the number of roots per plant fits a poisson or negative binomial distribution. Furthermore we can assume independence from randomization. Therefore, to evaluate which distribution is more appropriate, we run a dispersion test using `dispersiontest`. Since the p value is so small, we conclude that there is evidence for overdispersion, and that therefore a negative binomial is more appropriate.

Negative Binomial Analysis

```
cymnrm <- glm.nb(nr~IBAconc*l,cym)
cymchisq <- 1-pchisq(cymnrm$null.deviance - cymnrm$deviance , cymnrm$df.null - cymnrm$df.residual)
```

By comparing the null and residual devance using a chi square test, we obtain a p value of 0.0018123. As α = 0.05, p $< \alpha$, and we reject the null hypothesis and conclude that knowing the light and IBA concentration is helpful in predicting the number of roots and orchid has.

```r
summary(cymnrm)
```

```
##
## Call:
## glm.nb(formula = nr ~ IBAconc * l, data = cym, init.theta = 1.876013937,
##     link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.3752     0.2552   5.389 7.09e-08 ***
## IBAconc       -0.4263     0.1283  -3.323  0.00089 ***
## lLuz          -0.1073     0.3890  -0.276  0.78274
## IBAconc:lLuz   0.3635     0.1582   2.298  0.02155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.876) family taken to be 1)
##
##     Null deviance: 60.180  on 40  degrees of freedom
## Residual deviance: 45.175  on 37  degrees of freedom
## AIC: 175.51
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.876
##           Std. Err.:  0.797
##
##  2 x log-likelihood:  -165.514
```

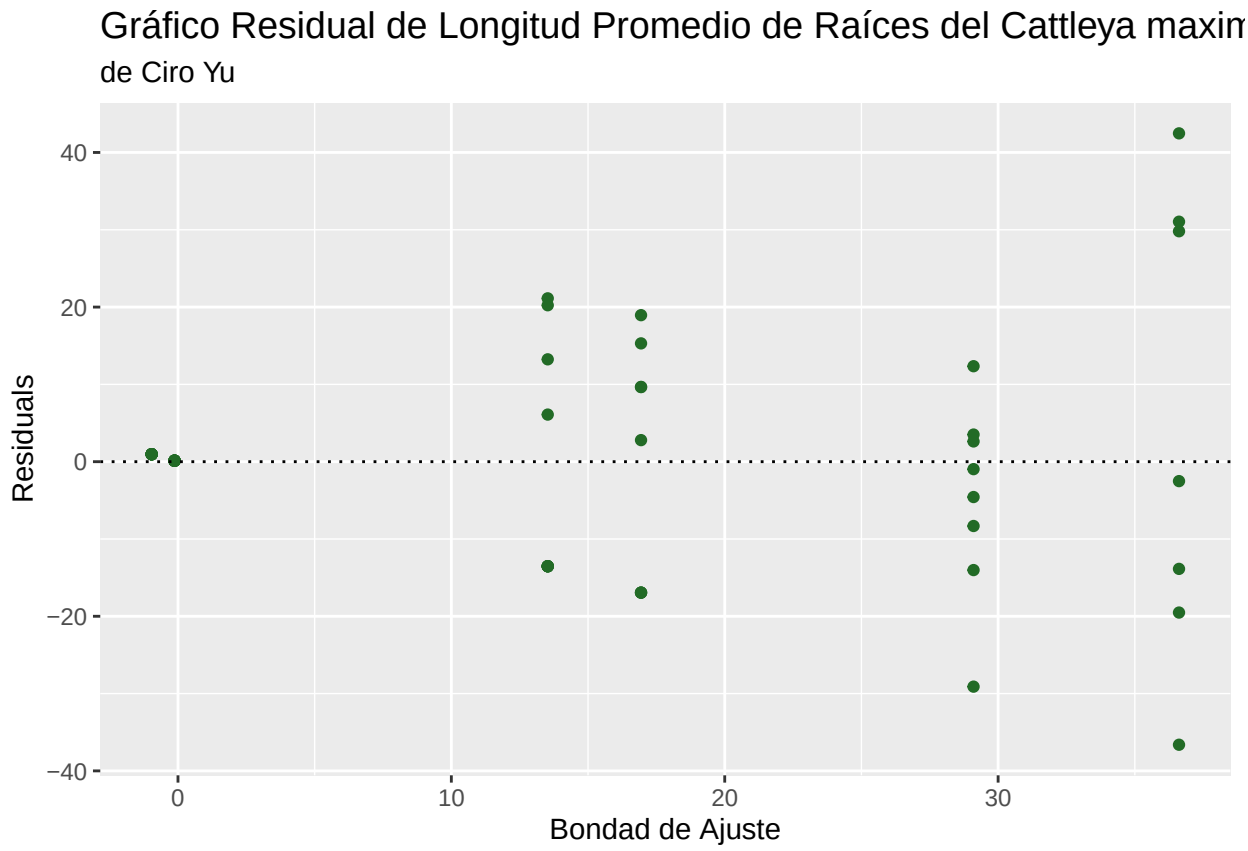
Here we see the terms for IBA concentration and the interaction term are statistically significant. This indicates that there is a plausible relationship between IBA concentration and the number of roots an orchid has as well as the interaction between the IBA concentration and the presence of light for the number of roots and orchid has.

Subexperiment 2: Cattleya Maxima (11)

Trait 1: Average Root Length Assumption Testing:

```r
cat <-  orchidsfinal %>% filter(es=="11")
catrnummIBA <- lm(rnum~IBAconc*l,cat)

#Residuals
catrnummIBAr <- tibble(
  "fit" = catrnummIBA$fitted.values,
  "resid" = catrnummIBA$residuals) %>%
  add_column("rstandardized"=rstandard(catrnummIBA))

ggplot(catrnummIBAr,aes(x=fit,y=resid))+
  geom_jitter(color="#226b26")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Gráfico Residual de Longitud Promedio de Raíces del Cattleya maxima (11)"),
```

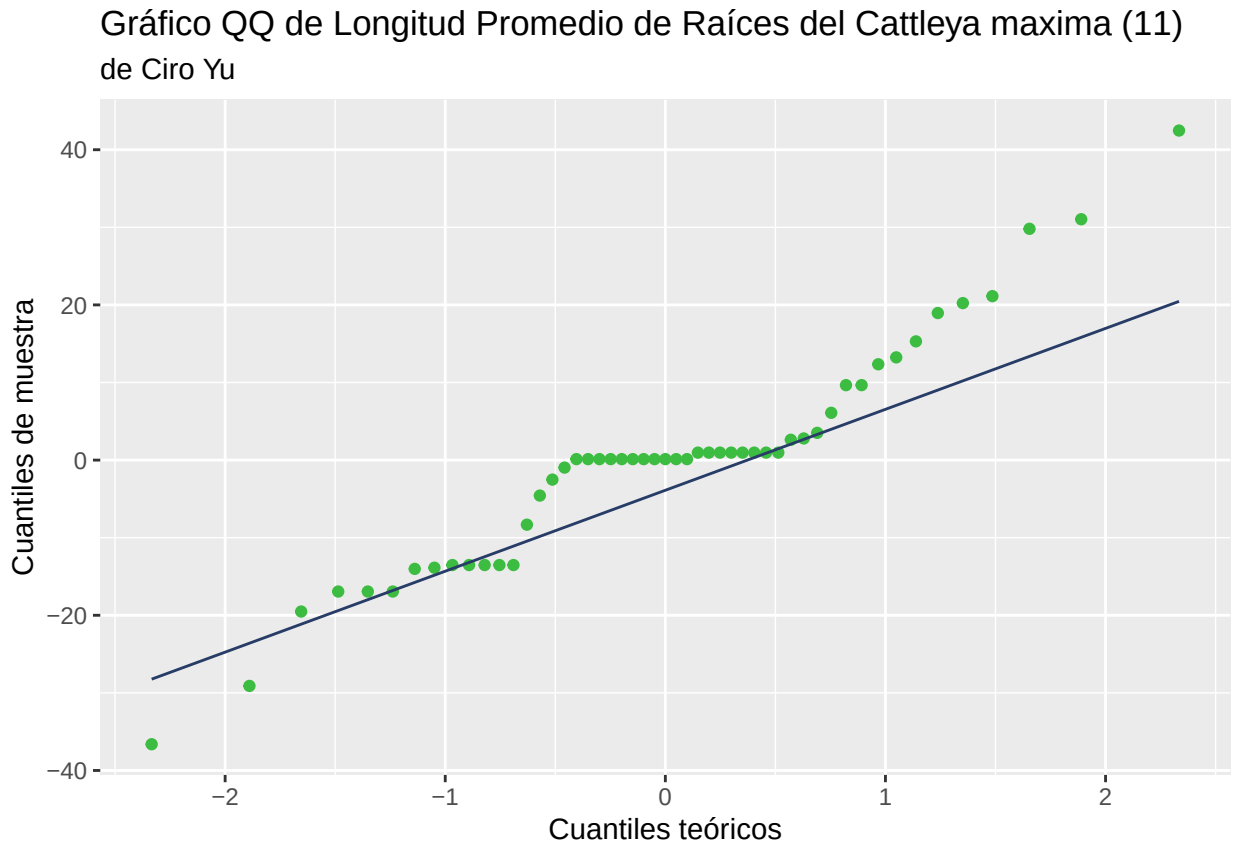
```
      subtitle = "de Ciro Yu")+
  ylab("Residuals")+
  xlab("Bondad de Ajuste")+
  theme(plot.title = element_text(size = 14))
```




```
bptest(catrnummIBA,studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  catrnummIBA
## BP = 26.198, df = 3, p-value = 8.669e-06
```

```
ggplot(data = catrnummIBAr, aes(sample = resid))+
  geom_qq( color="#3dbc42")+
  geom_qq_line( color="#273d67")+
  labs(
    title = paste("Gráfico QQ de Longitud Promedio de Raíces del Cattleya maxima (11)"),
    subtitle = "de Ciro Yu"
  ) +
  xlab("Cuantiles teóricos")+
  ylab("Cuantiles de muestra")
```




```
shapiro.test(catrnummIBA$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  catrnummIBA$residuals
## W = 0.93915, p-value = 0.01136
```

Unfortunately in this case, our assumptions are violated.

- The residual plot shows a clear funnel shape, indicate nonconstant error variance (Heteroscedasticity). Additionally, the Breusch-Pagan test is very significant
- The QQ plot is not well fitted to the line in the middle section and ends, indicating nonnormality of error variance. Additionally, the Shapiro-Wilk test is significant.
- Independence is still assumed due to the randomization of the experiment.

A log, log-log, and box cox transformations were tried with this data, but none reduced Heteroscedasticity or nonormality significantly. A 2 part model was also attempted, but the results of said model were not significant. Thus, the linear model continued to be used.

```
summary(catrnummIBA)
```

```
##
## Call:
## lm(formula = rnum ~ IBAconc * l, data = cat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.617 -10.924   0.126   3.148  42.473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.939      4.061   4.172 0.000129 ***
## IBAconc        -3.413      1.275  -2.677 0.010201 *
## lLuz           19.679      5.938   3.314 0.001778 **
## IBAconc:lLuz   -4.103      1.924  -2.133 0.038215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 47 degrees of freedom
## Multiple R-squared:  0.4732, Adjusted R-squared:  0.4396
## F-statistic: 14.07 on 3 and 47 DF,  p-value: 1.123e-06
```

The overall p value is significant at $\alpha = 0.05$. The values for IBA concentration, light, and their interactions were all significant, with Light having a positive term and the other 2 being negative. That means that on average, an increase in 1 mg/L of IBA in the media will cause a 3.413 mm decrease in average root length for the orchid. In the light condition, this increases to -7.523 mm per mg/L, but from a higher inital value.

Trait 2: Root Number Assumption Analysis

```
dispersiontest(glm(nr~IBAconc*l,cat,family=poisson))
```

```
##
##  Overdispersion test
##
## data:  glm(nr ~ IBAconc * l, cat, family = poisson)
## z = 1.7242, p-value = 0.04234
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##    2.27231
```

As with the *Cymbidium* data, our p value for overdispersion is under 0.05. This indicates that there is statistically significant evidence that the data is overdispered, and a negative binomial is appropriate.

Negative Binomial Analysis

```
catnrm <- glm.nb(nr~IBAconc*l,cat)
catchisq <- 1-pchisq(catnrm$null.deviance - catnrm$deviance , catnrm$df.null - catnrm$df.residual)
catchisq
```

```
## [1] 4.175958e-08
```

By comparing the null and residual deviance using a chi square test, we obtain a p value of 4.1759583×10^{-8}. As $\alpha = 0.05$, p $< \alpha$, and we reject the null hypothesis and conclude that knowing the light level or the IBA concentration is helpful in predicting the number of roots and orchid has.

```
summary(catnrm)
```

```
##
## Call:
## glm.nb(formula = nr ~ IBAconc * l, data = cat, init.theta = 1.115983047,
##     link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.9057     0.3678   2.463   0.0138 *
## IBAconc       -1.0530     0.4286  -2.457   0.0140 *
## lLuz           0.5165     0.4978   1.038   0.2995
## IBAconc:lLuz   0.2947     0.4951   0.595   0.5517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.116) family taken to be 1)
##
##     Null deviance: 73.681  on 50  degrees of freedom
## Residual deviance: 36.482  on 47  degrees of freedom
## AIC: 136.72
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.116
##           Std. Err.:  0.482
##
##  2 x log-likelihood:  -126.724
```

```
save(cat,file="Data/cat.Rdata")
save(cym,file="Data/cym.Rdata")
```

Looking at our table, we see that the term for the IBA concentration is significant at $\alpha = 0.05$. That means that we have statistical evidence to conclude that IBA concentration of the media has an effect on the number of roots for an orchid. Since the value of the estimated coefficient is negative, an increase in IBA concentration seems to decrease the number of roots an orchid has.As the coefficients for light and interaction between light and IBA concentration were not significant, a reduced model was fitted, but 2x log likelihood increased while the AIC did not decrease, so we went with the full model for a better fit.