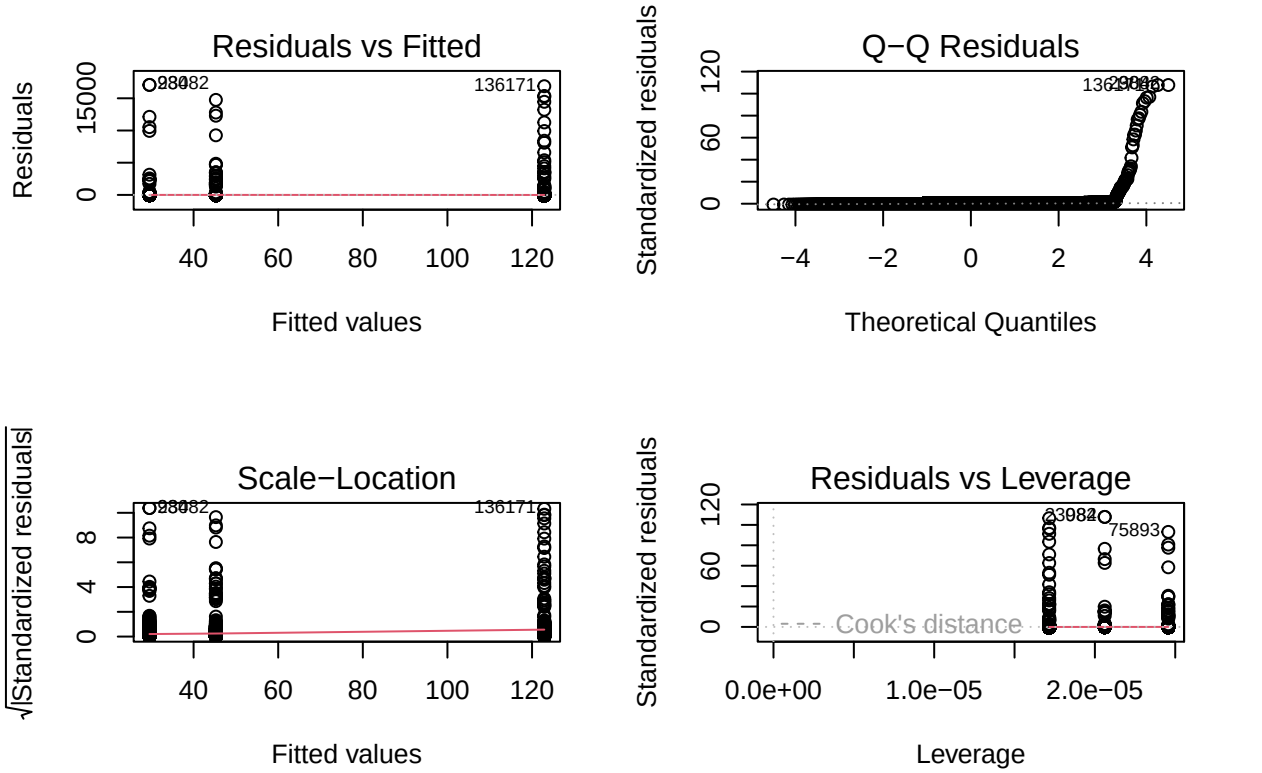
```
par(mfrow = c(2,2))
plot(mod2)
```



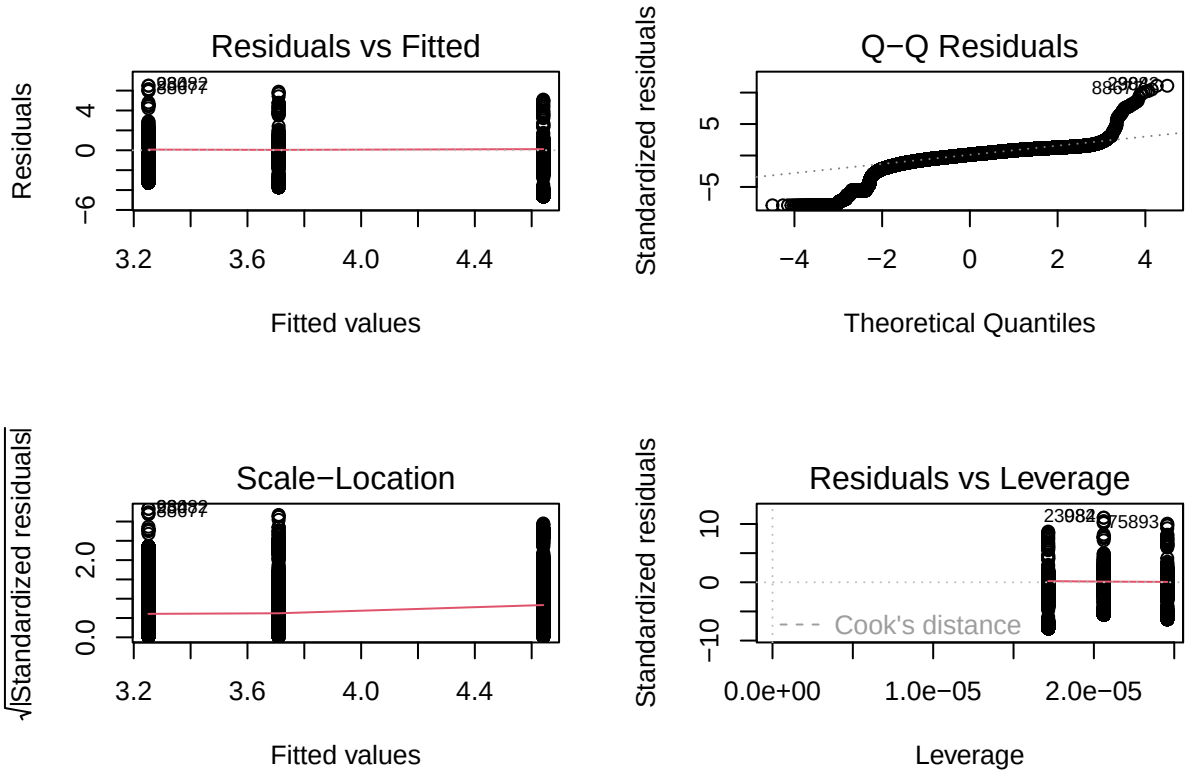
CONCLUSION: The residuals seem to have similar variances in the **Residual vs Fitted** plot, but we notice that it is not a striaght line in the **QQ Residuals** plot which means we must use a transformation in order to ensure normality.

d) Regardless if you think the residuals are normal or not, log transform `yield_bushels_per_acre`, re-run the ANOVA, re-check the assumptions. Do your transformed residuals meet the normality and variance assumptions (yes/no)? Must comment on the shape of the "Residual vs Fitted" and "Q-Q Residuals" graphs for full credit.

```
crop$yield_bushels_per_acre <- log(crop$yield_bushels_per_acre+1)


mod2<- aov(yield_bushels_per_acre~species, data=crop)

par(mfrow = c(2,2))
plot(mod2)
```



CONCLUSION: The log transformation met the variance assumption pretty well in the **Residuals vs Fitted** plot and it is more of a linear line in the **Q-Q Residuals** plot which can indicates normally distributed residuals.

e) Regardless if you think the transformed residuals meet the ANOVA assumptions, please interpret the ANOVA model. You must state the null and alternative hypotheses of the ANOVA model. And include the degrees of freedom, F-value, p-value, if you reject or fail the null hypothesis. After reporting these values, in 1-2 sentences discuss what this means about the different crop species and if they have similar or dissimilar crop yields.

```
summary(mod2)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## species          2  53743   26872   78234 <2e-16 ***
## Residuals   147497  50662       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSION: Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : Not\ all\ means\ are\ equal$$

The **DF** are 2 for species and 147497 for the residuals. The **F-value** is 37590 The **P-value** is 2e-16. Therefore, we reject the null hypothesis which means that there is a significant difference between the crop yields among the species.

f) Based on the result from the ANOVA model, should you run a post-hoc test (e.g. Tukey-Kramer)? Say yes or no and explain your answer.

ANSWER: You should run a post-hoc test because we reject the null hypothesis, so there exists a difference in atleast one of the groups and the post-hoc test will help us find it.

g) Regardless of your answer to e, run a Tukey-Kramer post-hoc test using the glht() and interpret the summary of the post-hoc test.

```
post_hoc <- glht(mod2, linfct = mcp(species= "Tukey"))
summary(post_hoc)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = yield_bushels_per_acre ~ species, data = crop)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## soybeans - corn == 0  -1.389133   0.003602  -385.7   <2e-16 ***
## wheat - corn == 0     -0.931075   0.003786  -245.9   <2e-16 ***
## wheat - soybeans == 0  0.458059   0.003939   116.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

CONCLUSION: The post_hoc test showed that there is a statiscally signficant different among all species since $p < 0.05$ for all 3.

h) Regardless of your answer to e, print out the compact letter display of the Tukey-Kramer results using the cld(). For each species, report their letter

```
cld(post_hoc)
```

```
##     corn soybeans    wheat
##      "a"      "b"      "c"
```

CONCLUSION: We see that each species has a different letter which indicates that there is a significant difference among all the species because if they were not significnatly different, the groups would share the same letter.

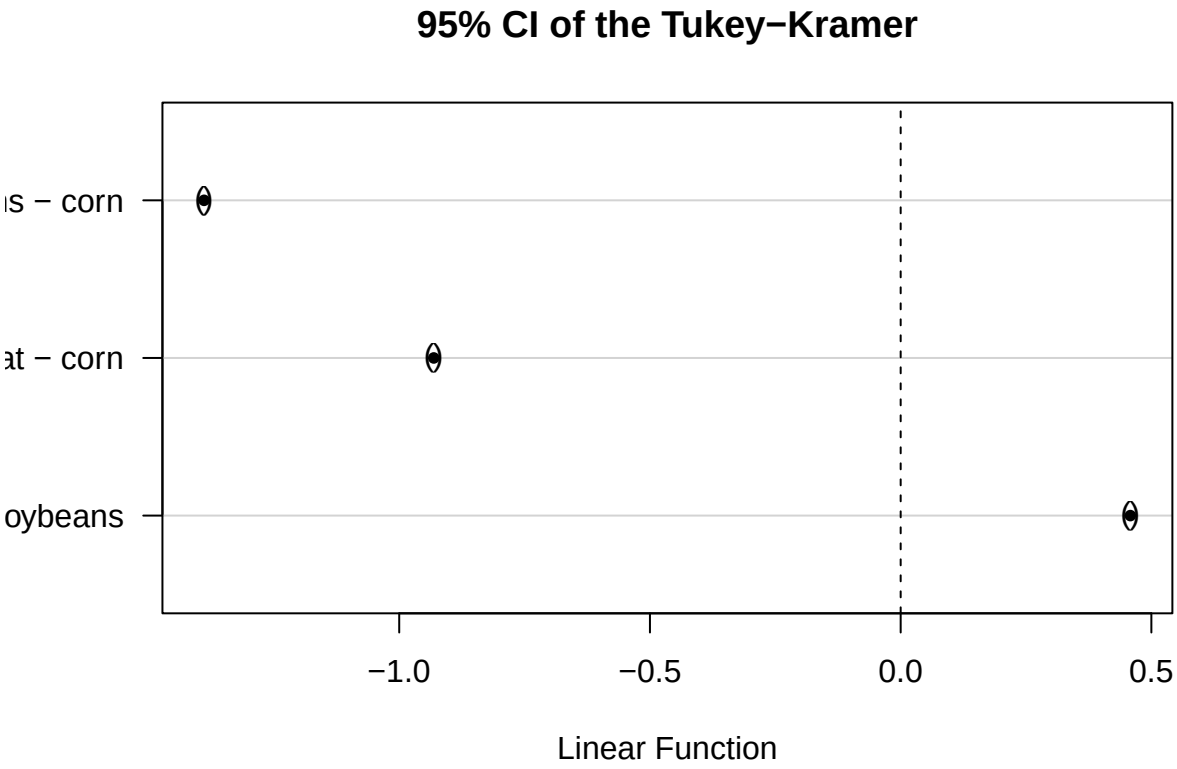
i) Regardless of your answer to e, plot the 95% CI of the Tukey-Kramer post-hoc results using plot(). Comment if any 95% CI group overlaps with the 0 dotted line and what does it mean if a 95% CI group doesn't overlap the 0 dotted line.

```
confint(post_hoc)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = yield_bushels_per_acre ~ species, data = crop)
##
## Quantile = 2.3423
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                         Estimate lwr     upr
```

```
## soybeans - corn == 0  -1.3891  -1.3976 -1.3807
## wheat - corn == 0     -0.9311  -0.9399 -0.9222
## wheat - soybeans == 0  0.4581   0.4488  0.4673
```

```
plot(post_hoc, main = "95% CI of the Tukey-Kramer")
```



CONCLUSION: None of the groups overlap with the 0 dotted line which means that all the groups are different.

j) Based on answers from f-h, which crop species has the highest yield and why do you say that?

CONCLUSION: Corn has the highest yield because it had the greatest difference between soybeans and wheat in the Tukey-Kramer results.

End