

Exercise 2: Crops

Exercise 2: Crops

This data comes from a sample of farms from three counties in Iowa. We want to know how the factors of the county and whether the farmer is related to the landlord of the farmland is related to the total crop yield of the farms.

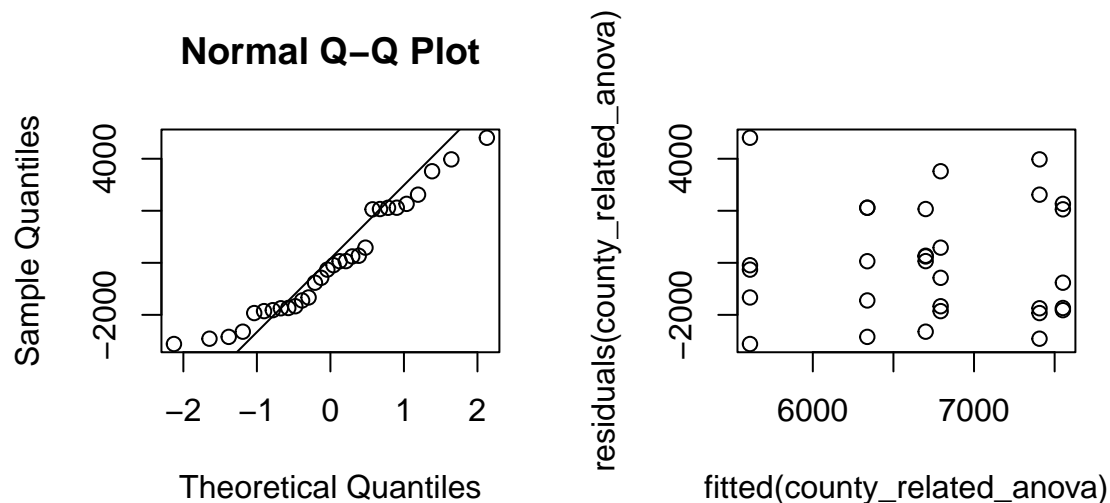
Part A

Firstly, we perform a two-way ANOVA on the Count, Related and Crops columns, which gives

```
anova(county_related_anova)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df      Sum Sq Mean Sq F value Pr(>F)
## Related    1   2378957 2378957   0.4113 0.5274
## County     2    8841441 4420721   0.7644 0.4766
## Related:County 2    1497573  748786   0.1295 0.8792
## Residuals  24 138805865 5783578
```

The p-value here is given by the $\text{Pr}(>F)$ column. For the Related and County factors separately, the p-value is not below 0.05, which means that a linear relation between these factors and the crop yield cannot be conclusively established. This is also true for the interaction between County and Related. As the two-way ANOVA assumes the data is normally distributed, we have to assess the normality of the data.



The left Q-Q plot of the residuals of the ANOVA places the point in roughly a straight line, which implies that this data is normally distributed.

The right plot shows how the spread of the residuals is roughly equal for all values. This implies that the underlying data is normally distributed.

```
summary(county_related_anova)

##
## Call:
## lm(formula = Crops ~ Related * County, data = crops_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3120.4 -1744.7  -176.9   2064.2  4806.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6700.0     1075.5   6.230 1.94e-06 ***
## Related1         -362.0     1521.0  -0.238   0.814
## County2           93.0     1521.0   0.061   0.952
## County3          851.2     1521.0   0.560   0.581
## Related1:County2  -820.6     2151.0  -0.381   0.706
## Related1:County3   217.0     2151.0   0.101   0.920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2405 on 24 degrees of freedom
## Multiple R-squared:  0.08393,    Adjusted R-squared:  -0.1069
## F-statistic: 0.4398 on 5 and 24 DF,  p-value: 0.8163
```

This summary shows that the average farm in county 1 of which the farmer is not related to the landlord is 6700. The average for a farmer in county 3 would then be

```
6700.0 + 851.0
```

```
## [1] 7551
```

This seems plausible.

Part B

```
size_anova <- lm(Crops ~ Size, data = crops_frame)
ancova_county_lm <- lm(Crops ~ Size + County, data = crops_frame)
ancova_related_lm <- lm(Crops ~ Size + Related, data = crops_frame)
```

Now we want to take the size of the farm into account.

```
drop1(ancova_county_lm, test = "F")

## Single term deletions
##
## Model:
## Crops ~ Size + County
##      Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                 31187313 423.63
## Size   1 111495081 142682394 467.25 92.9504 4.513e-10 ***
## County 2   767179 31954491 420.36  0.3198   0.7291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ancova_county_lm)
```

```
##
## Call:
## lm(formula = Crops ~ Size + County, data = crops_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2411.8  -609.7  -157.8   897.1  1970.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  635.337    701.699   0.905   0.374
## Size          31.976     3.317   9.641 4.51e-10 ***
## County2       328.624    494.359   0.665   0.512
## County3       348.950    493.878   0.707   0.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1095 on 26 degrees of freedom
## Multiple R-squared:  0.7942, Adjusted R-squared:  0.7704
## F-statistic: 33.44 on 3 and 26 DF,  p-value: 4.477e-09
```

This ANCOVA shows that the size of the farm is strongly correlated with the yield of a farm, which can be seen in the low p-value of $4.5 \cdot 10^{-10}$. The county does not appear to have a significant effect here, as the p-value is high. The data for the size and related status looks similar:

```
drop1(ancova_related_lm, test = "F")
```

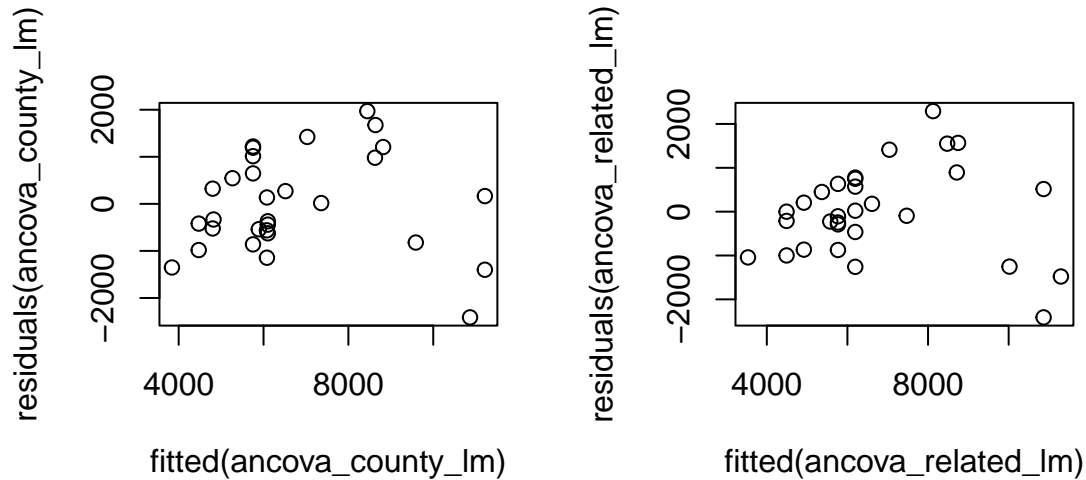
```
## Single term deletions
##
## Model:
## Crops ~ Size + Related
##      Df Sum of Sq      RSS      AIC F value    Pr(>F)
## <none>             30573906 421.03
## Size      1 118570972 149144879 466.58 104.7107 8.646e-11 ***
## Related  1  1380585  31954491 420.36   1.2192   0.2793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ancova_related_lm)
```

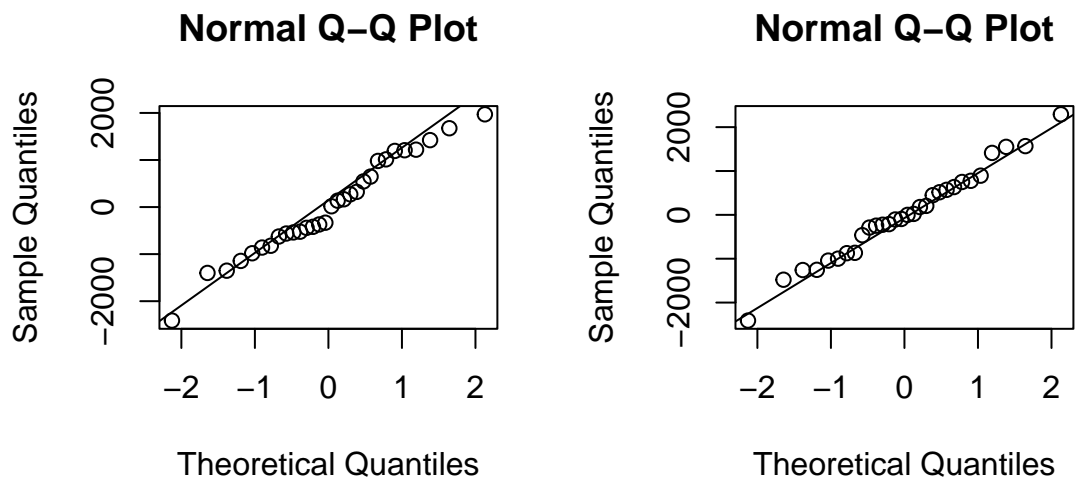
```
##
## Call:
## lm(formula = Crops ~ Size + Related, data = crops_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2410.4  -765.2   -47.1   618.2  2292.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1092.777    640.632   1.706   0.0995 .
## Size          31.884     3.116  10.233 8.65e-11 ***
## Related1     -429.286    388.785  -1.104   0.2793
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1064 on 27 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.7833
## F-statistic: 53.41 on 2 and 27 DF,  p-value: 4.128e-10
```

While the correlation between crop yield and farm size is significant, the related status does not appear to have an effect. These results lead us to conclude that the effect of county and relation on the influence of size is negligible.



As ANCOVA assumes that the underlying data is normally distributed, we will have to check for normality. Both of the above residual plots show no obvious correlation in the variance of the residuals, which implies a normal distribution.



The Q-Q plots above show the points roughly in a straight line. This implies that the underlying data is normally distributed.

Part C