Assignment 1 Report

2025-02-21

Exercise 2

We continue with the analysis of the values of crops for several farms in Iowa for which data was collected from three different counties of the state. Namely, we are interested in the effect (on the amount of crops) of the county in which the farm is located and the possible relation that the tenant might have with the landlord. In addition, we are also provided additional information on the size of each farm. A summary of the data is provided here (for the 4 variables, over the total 30 observations):

```
summary(data)
```

```
##  County      Crops            Size      Related
##  1:10   Min.   : 2490   Min.   : 90   no :15
##  2:10   1st Qu.: 5180   1st Qu.:156   yes:15
##  3:10   Median : 6308   Median :160
##         Mean   : 6733   Mean   :184
##         3rd Qu.: 8455   3rd Qu.:226
##         Max.   :11382   Max.   :320
```

a)

We are initially interested in analyzing the effect of *County* and *Related* on *Crops*, without taking *Size* into account, therefore the interest is firstly devoted to the following two-way ANOVA model:

$$Cr_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where Cr_{ijk} refers to the value of *Crops* of observation k (in *County* i and *Related* j), μ to the overall mean, α_i to the main effect of *County* i, β_j to the main effect of *Related* j, γ_{ij} to the interaction effect of levels i and j of *County* and *Related* (respectively) and e_{ijk} to the independent error of the model (of which normality will later be tested). Thus, by studying this (and the following) ANOVA model(s), we test (with a 0.05 significance level) the following hypotheses:

1. $H_{int} : \gamma_{ij} = 0$ for every (i, j) (no interaction between factors *County* and *Related*);
2. $H_{cou} : \alpha_i = 0$ for every i (no main effect of *County*);
3. $H_{rel} : \beta_j = 0$ for every j (no main effect of *Related*).

```
model_int = lm(Crops ~ County * Related, data=data); anova(model_int)
```

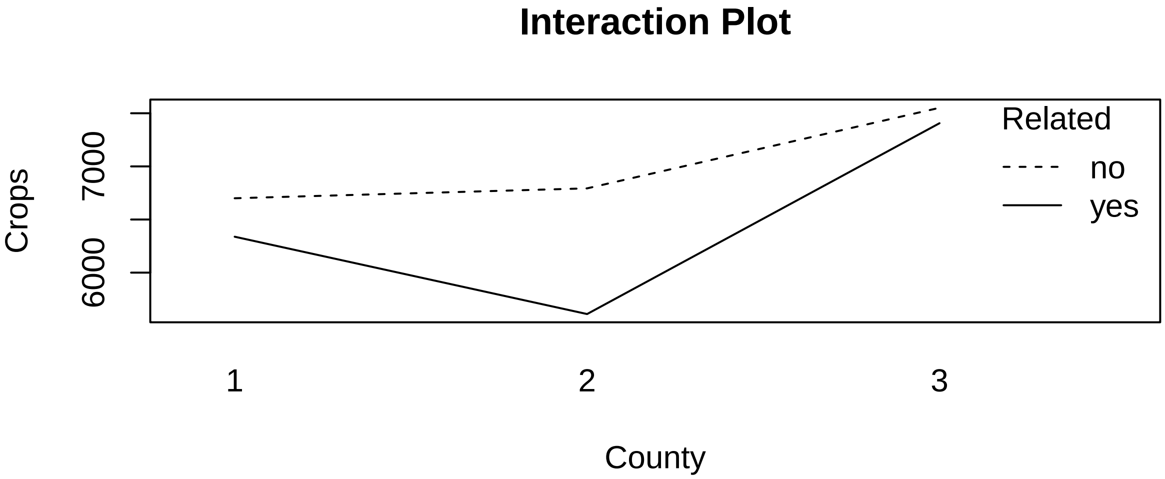
```
## Analysis of Variance Table
##
## Response: Crops
##                Df   Sum Sq Mean Sq F value Pr(>F)
## County          2 8.84e+06 4420721    0.76   0.48
## Related         1 2.38e+06 2378957    0.41   0.53
## County:Related  2 1.50e+06  748786    0.13   0.88
## Residuals      24 1.39e+08 5783578
```

Focusing on the last p-value, following the computation of the F-statistics reported on the *F value* column, we cannot reject H_{int} as $0.88 > 0.05$. We double check the statement by running an ANOVA model without interaction term and comparing it with our first model:

```
model = lm(Crops ~ County + Related, data=data); anova(model_int, model)
```

```
## Analysis of Variance Table
##
## Model 1: Crops ~ County * Related
## Model 2: Crops ~ County + Related
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1     24 1.39e+08
## 2     26 1.40e+08 -2  -1497573 0.13   0.88
```

The p-value is indeed the same (because of the use of a balanced design). For additional context, we plot the *Crops* amounts agains the *County* factor for the two *Related* levels:

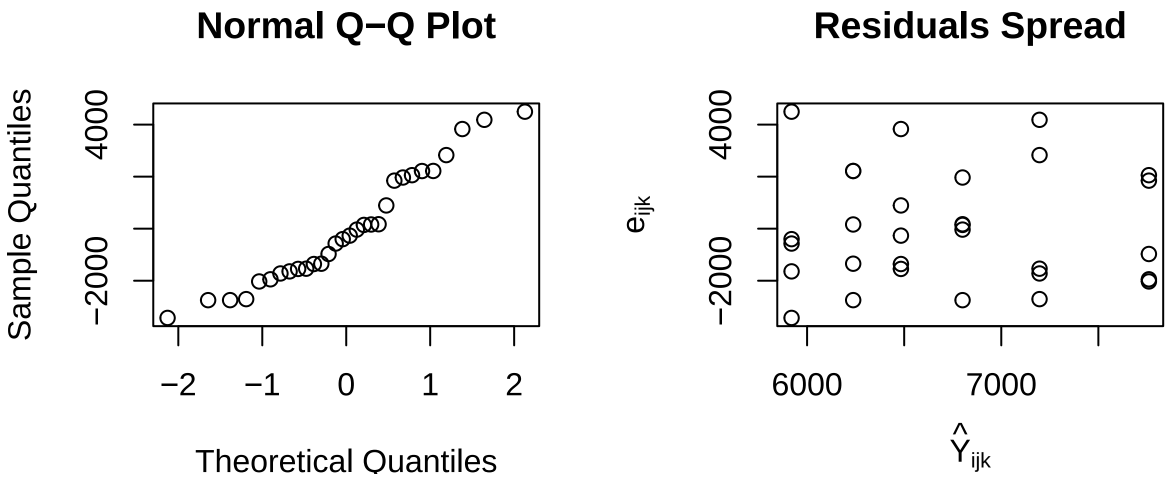


As expected, the lines do not differ significantly from a parallel behavior, signalling the high likelihood of lack of interaction. Therefore, in the following parts of the report, the model of interest will be the one without interaction:

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df   Sum Sq Mean Sq F value Pr(>F)
## County     2 8.84e+06 4420721    0.82   0.45
## Related    1 2.38e+06 2378957    0.44   0.51
## Residuals 26 1.40e+08 5396286
```

Analyzing this, the significance of the estimates for *County* and *Related* is not greatly impacted compared to the model with interaction and, thus, we again cannot reject neither H_{cou} nor H_{rel}, signalling the lack of sufficient evidence in our data needed to prove the existence of a main effect of *County* and *Related* on *Crops*. Nevertheless, we check the requirement of the model, namely the normality assumption for the model residuals e_{ijk} and that their spread does not change systematically with the fitted values $\hat{Y}_{ijk}$:



From the above plots, we cannot state the absence of the normality assumption, meaning the model, even if with not significant estimates, is at least valid (according to this observation). The last step of this first part of Exercise 2 consists in utilizing the estimates of our model: we want to estimate the crops for a typical farm in *County* 3 for which landlord and tenant are *not* related, referring to the following:

```
summary(model)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6801        848   8.017 1.70e-08
## County2         -317       1039  -0.305 7.62e-01
## County3          960       1039   0.924 3.64e-01
## Relatedyes      -563        848  -0.664 5.13e-01
```

The desired value can then be computed in the following way:

$$Crops_{cou=3,rel=no} = 6800.6 + 959.7 = 7760.3$$

b)

We proceed by including *Size* in our analysis, included in the model as numerical explanatory variable. In particular, we are interested in investigating whether the influence of *Size* on the *Crops* value is similar for all three counties and whether the influence of *Size* depends on the relation of landlord and tenant of the farm. Starting with the former, we formulate the following ANCOVA model:

$$Cr_{ik} = \mu + \alpha_i + \beta_j + \delta S_{ik} + \lambda_i S_{ik} + e_{ijk}$$

where S_{ik} refers to the *Size* of observation k (of *County* i), δ to its effect on *Crops* (Cr_{ik}) and λ_i to the interaction effect between *Size* of observation k (of *County* i) and factor *County* i. The corresponding null hypotheses then become:

1. $H_{c/s} : \lambda_i = 0$ for every i (no interaction between *County* and *Size*);
2. $H_{cou} : \alpha_i = 0$ for every i (no main effect of *County*).
3. $H_{rel} : \beta_j = 0$ for every j (no main effect of *Related*).

Since *Size* is only included as explanatory variable, we are not formally interested in its main effect and, thus, a null hypothesis about δ. To test $H_{s/c}$ we perform:

```
model_base_size <- lm(Crops ~ County + Related + Size, data=data)
model_interact_county <- lm(Crops ~ Related + County * Size, data = data)
anova(model_interact_county, model_base_size)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Crops ~ Related + County * Size
## Model 2: Crops ~ County + Related + Size
##   Res.Df      RSS Df Sum of Sq   F Pr(>F)
## 1     23 20277325
## 2     25 29805979 -2  -9528654 5.4  0.012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, the interaction effect (λ_i) is significant at the 0.05 significance value and $H_{s/c}$ can ultimately be rejected. Regarding H_{cou}, we explicitly run the following:

```
anova(model_interact_county)
```

```
## Analysis of Variance Table
##
## Response: Crops
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## Related      1 2.38e+06 2.38e+06    2.70   0.114
## County       2 8.84e+06 4.42e+06    5.01   0.016 *
## Size         1 1.10e+08 1.10e+08  125.33 8.7e-11 ***
## County:Size  2 9.53e+06 4.76e+06    5.40   0.012 *
## Residuals   23 2.03e+07 8.82e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and thus also reject H_{cou}, meaning that the data applied to this model offers sufficient evidence to reject $\alpha_i = 0$ for every i. H_{rel} cannot instead be rejected (p-value= 0.114).

Similarly, for the interaction between *Size* and the factor *Related*, the ANCOVA model can be expressed as follow:

$$Cr_{jk} = \mu + \alpha_i + \beta_j + \delta S_{jk} + \rho_j S_{jk} + e_{ijk}$$

where S_{jk} refers to the *Size* of observation k (of *Related* factor j), δ to its effect on *Crops* (Cr_{ik}) and ρ_j to the interaction effect between *Size* of observation k and factor *Related* j. The corresponding null hypotheses then become:

1. $H_{r/s} : \rho_j = 0$ for every j (no interaction between *Related* and *Size*);
2. $H_{cou} : \alpha_i = 0$ for every i (no main effect of *County*).
3. $H_{rel} : \beta_j = 0$ for every j (no main effect of *Related*).

Following the desogn applied above for *County* and its interaction with *Size*:

```
model_interact_related <- lm(Crops ~ County + Related * Size, data = data)
anova(model_interact_related, model_base_size)
```

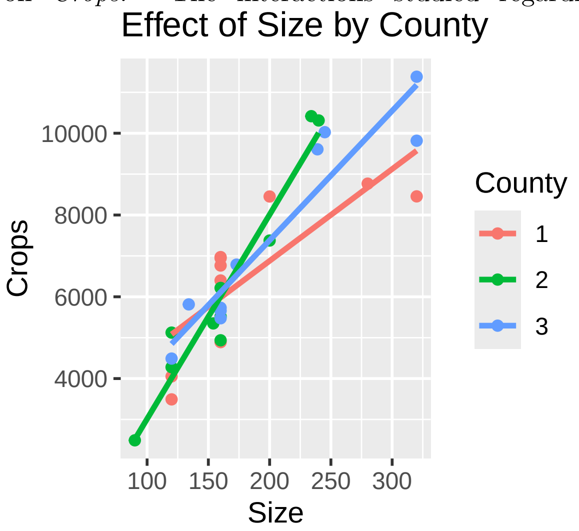
```
## Analysis of Variance Table
##
## Model 1: Crops ~ County + Related * Size
## Model 2: Crops ~ County + Related + Size
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1     24 28452313
## 2     25 29805979 -1  -1353666 1.14    0.3
```

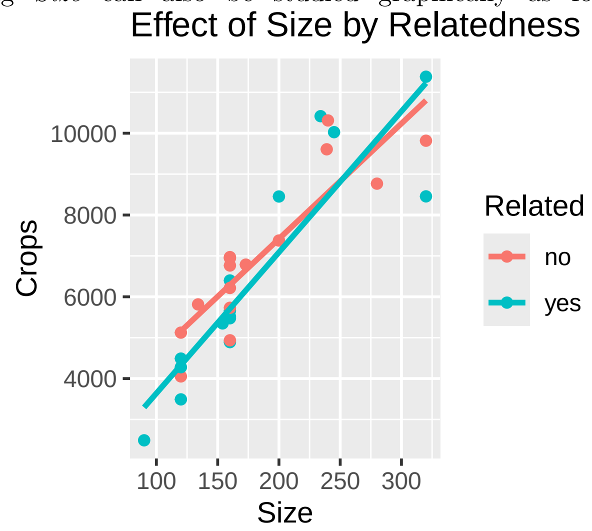
As the F test demonstrates, we cannot reject $H_{r/s}$ since its p-value it is larger than the significance desired. Printing the whole ANCOVA model, we can also state that we fail to reject H_{rel}:

```
anova(model_interact_related)
```

```
## Analysis of Variance Table
##
## Response: Crops
##              Df   Sum Sq  Mean Sq F value  Pr(>F)
## County        2 8.84e+06 4.42e+06    3.73   0.039 *
## Related       1 2.38e+06 2.38e+06    2.01   0.169
## Size          1 1.10e+08 1.10e+08   93.21 9.7e-10 ***
## Related:Size  1 1.35e+06 1.35e+06    1.14   0.296
## Residuals    24 2.85e+07 1.19e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

meaning that *Related* does not seem to have a significant effect on the value of *Crops* neither through a main nor a interaction (with *Size*) effect. On the other hand, H_{cou} can again be rejected (p-value= 0.039) on the basis of which we can state that *County* has a significant effect on *Crops*. The interactions studied regarding *Size* can also be studied graphically as follows:


Effect of Size by County


Effect of Size by Relatedness

It can then be stated that the graphics confirm the statistical tests above, as it can be seen how the effect of *Size* (the slope of the lines) differ more significantly in the *County* panel compared to the *Related* one.

The most appropriate model to study the data at hand is therefore the one including *County*, *Related*, *Size* and the interaction of the latter with *County*. ### c)

Observing the estimates of the model just mentioned:

```
summary(model_interact_county)$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2461.01     929.76   2.647 1.44e-02
## Relatedyes    -239.10     347.92  -0.687 4.99e-01
## County2      -4214.05    1447.24  -2.912 7.85e-03
## County3      -1284.81    1302.58  -0.986 3.34e-01
## Size            22.70       4.77   4.764 8.38e-05
## County2:Size    26.59       8.09   3.286 3.23e-03
## County3:Size     8.92       6.40   1.394 1.77e-01
```

we can study the effect of (1) *County*, (2) *Related* nad (3) *Size*:

1. *County* has a significant impact on *Crops*, such that farms in *County* 2 produce significantly less

Crops, while there is no significant difference between production of farms in *County* 1 and 3 (keeping everything else constant); this effect (related to *County* 2) is significantly contrasted the larger the farm is.

2. Being related to the landlord does not significantly impact the value of *Crops*.
3. Other than the already mentioned interaction effect with *County* factor 2, *Size* has a significant main positive effect on *Crops*: the larger the farm, the more it will produce (keeping everything else constant). ### d)

Lastly we are asked to predict the *Crops* for a farm from *County* 2 of *Size* 165, with related landlord and tenant. Looking at our chosen model, we compute:

$$Crops_{cou=2,rel=yes,s=165} = 2461.01 - 239.10 - 4214.05 + 22.70 * 165 + 26.59 * 165 = 6141.30$$

as confirmed by the follwing code, which also provides us the error variance:

```
farm <- data.frame(County = factor(2, levels = levels(data$County)),
                      Size = 165,
                      Related = "yes")
predicted_crops <- predict(model_interact_county, farm, interval="prediction")
predicted_crops
```

```
##    fit  lwr  upr
## 1 6141 4072 8210
```