Comparing Crop Yield of Species of Barley Plant

Name: Leigh West

Exploratory Analysis

An interaction plot of a factorial experiment investigating the yield of 3 different varieties of barley compared at 3 different row spacings is depicted in Figure 1 (below).

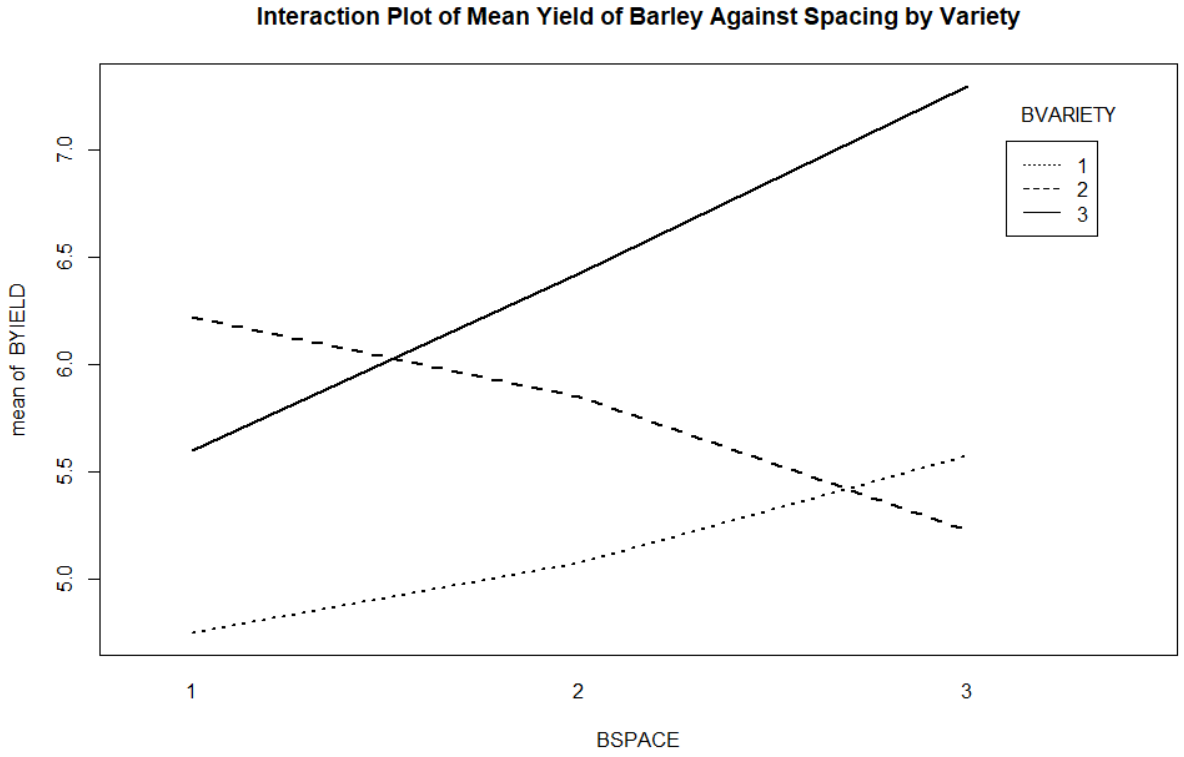

Figure 1: An interaction plot of mean barley yield against spacing and by variety.

Figure 2 (below) plots the mean yield and the respective confidence intervals for each block on the left. The right plot in Figure 2 depicts the mean barley yield and their respective confidence intervals for each variety at the three levels of spacing.

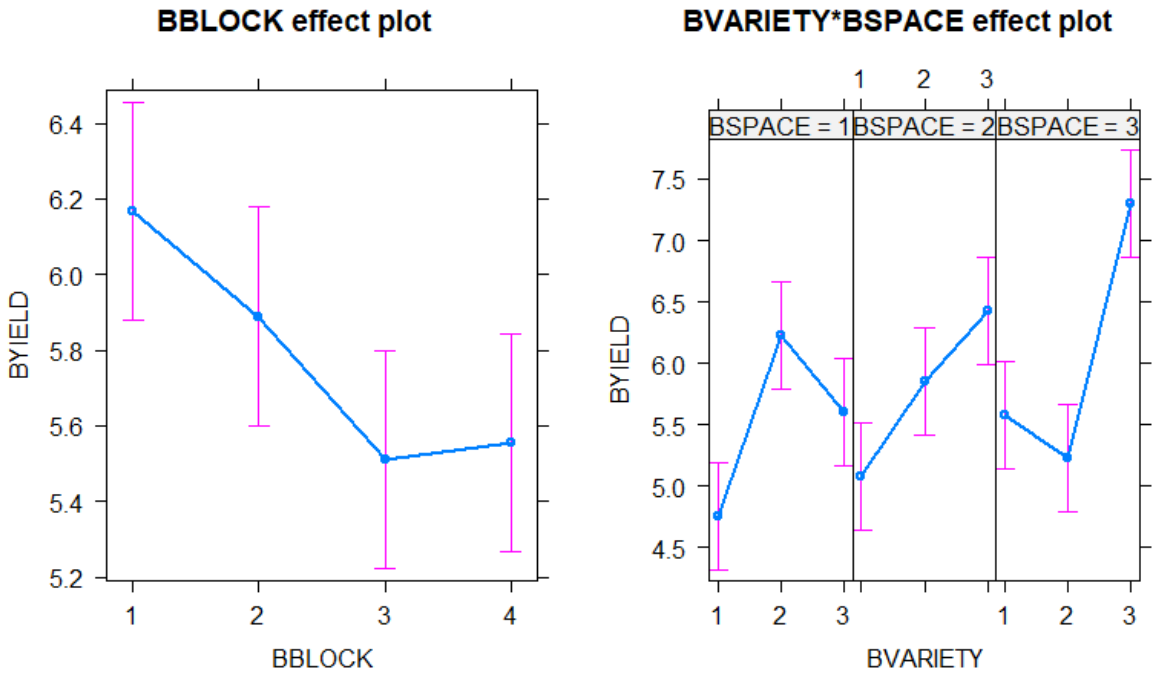

Figure 2: Mean barley yield and 95% CI for each block (left) and for each variety by spacing level (right)

Considering the Statistical Model

The interaction plot in Figure 1 is indicative of an interaction because the trend lines are not parallel. The gradient of the trend line for variety 3 appears to increase at a constant rate as we move from spacing level 1 to 2, and from 2 to 3. The trend line for variety 1 also increases as we move from spacing level 1 to 2 (although not as sharply as variety 3), and the gradient increases marginally as we move from level 2 to 3. However, these varieties contrast with variety 2, which achieves its peak mean yield with spacing level 1, and which decreases as we move to level 2, and decreases once again as we move to level 3.

The interaction plot in Figure 1 also suggests that the peak yield for variety 1 is achieved with level 3 spacing, and for variety 2 it is achieved with level 1 spacing. The interaction plot indicates that overall, the highest yield is achieved with variety 3 at spacing level 3.

The left plot in Figure 2 depicts the mean barley yield by block. There does not appear to be a significant difference in yields between blocks 2, 3 and 4 because there is a non-trivial amount of overlap among the confidence intervals. An overlap in confidence intervals also exists between blocks 1 and 2, however there does not appear to be any overlap in the confidence intervals between blocks 1 and 3, and blocks 1 and 4. The plot indicates this difference is significant.

The right plot in Figure 2 depicts the mean barley yield by variety and spacing. The effects plot indicates that the mean barley yield for variety 3 under the 3rd level of row spacing is significantly greater than all

other combinations, except potentially variety 3 in row space level 2. Under row spacing level 1, it appears that variety 2 returns a yield significantly greater than variety 1.

From Figure 2, there appears to be a significant blocking effect. In addition, per Figure 1, there appears to be a significant interaction between row spacing and plant variety. Accordingly, an interaction model which includes a blocking effect, generalised to the following form is likely appropriate:

$$E(y) = \beta_0 + \beta_1 Block2 + \beta_2 Block3 + \beta_3 Block4 + \beta_4 BVariety2 + \beta_5 BVariety3 + \beta_6 BSpace2 + \beta_7 BSpace3 + \beta_8 BVariety2{:}\,BSpace2 + \beta_9 BVariety2{:}\,BSpace3 + \beta_{10} BVariety3{:}\,BSpace2 + \beta_{11} BVariety3{:}\,BSpace3$$

Determining the Statistical Model

Variable screening was performed in R using backward stepwise regression, commencing with the interaction model and the lower model set to the null model. Per Table 1, the interaction model produced an Akaike's Information Criteria (AIC) of -52.99. Removing the interaction term would have increased the AIC to -23.87, and similarly, removing the blocking term would have increased the AIC to -23.87. Hence, backward stepwise regression did not progress beyond the initial step.

Table 1: Backward Stepwise Regression (First and Only Step)

```
First Step

Start: AIC=-52.99
BYIELD ~ BBLOCK + BVARIETY * BSPACE
                    Df   Sum of Sq       RSS      AIC
<none>                                 4.2411  -52.993
- BBLOCK             3      2.5564    6.7975  -42.011
- BVARIETY:BSPACE    4      7.6544   11.8956  -23.865
```

The results of the analysis of variance of the final stepwise model is provided in Table 2 (below). From this table we note that the interaction term BVARIETY:BSPACE is significant (p-value ≈ 0), and we conclude that the change in mean yield with a change in row spacing is not the same for each plant variety.

Table 2: Analysis of Variance of the Final Stepwise Model

```
Analysis of Variance Table

Response: BYIELD
                 Df    Sum Sq   Mean Sq   F value      Pr(>F)
BBLOCK            3    2.5564    0.8521    4.8221     0.00912
BVARIETY          2   10.2739    5.1369   29.0694   3.872e-07
BSPACE            2    1.5506    0.7753    4.3872     0.02377
BVARIETY:BSPACE   4    7.6544    1.9136   10.8289   3.679e-05
Residuals        24    4.2411    0.1767
```

An excerpt of the summary of the coefficients of regression coefficients of the final stepwise model produced in R is provided in Table 3 (below). The data indicates that the final stepwise model is useful (F-statistic = 11.34 on 11, 24 df, p-value ≈ 0) and explains approximately 76% of the observed variability in mean barley yield.

Table 3: Table of Regression Coefficients (Final Stepwise Model)

```
Residual standard error: 0.4204 on 24 degrees of freedom
Multiple R-squared: 0.8386, Adjusted R-squared:  0.7646
F-statistic: 11.34 on 11 and 24 DF,  p-value: 5.173e-07
```

A table of means of the final model is provided below in Table 4. As indicated in the effect plot in Figure 2, the highest yield appears to occur in variety 3 with row spacing level 3.

Table 4: Table of Means (Final Stepwise Model)

```
BBLOCK effect
BBLOCK
       1        2        3        4
6.166667 5.888889 5.511111 5.555556

BVARIETY*BSPACE effect
       BSPACE
BVARIETY       1      2      3
        1  4.750  5.075  5.575
        2  6.225  5.850  5.225
        3  5.600  6.425  7.300
```

From the table of coefficients and their 95% confidence intervals in Table 5 (below), we can conclude that the main differences in interaction means occur for varieties 2 and 3 at space level 3. This is consistent with the effects plot in Figure 2, where the confidence interval for variety 3 at space level 3 does not overlap with variety 3 at space level 1 or 2. Accordingly, we can conclude that the change in mean yield of variety 3 is significantly greater as we move from space level 1 to 3, than it is for variety 1. The confidence interval of the interaction between variety 2 and space level 3 is entirely negative. Therefore, we can conclude that the change in mean yield of variety 2 is significantly lower than variety 1, as we move from space level 1 to space level 3.

Table 5: Table of Coefficients and 95% Confidence Intervals (Final Stepwise Model)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	2.5%	97.5%
(Intercept)	5.1361	0.2427	21.162	< 2e-16	4.64	5.64
BBLOCK2	-0.2778	0.1982	-1.402	0.173792	-0.69	0.13
BBLOCK3	-0.6556	0.1982	-3.308	0.002953	-1.07	-0.25
BBLOCK4	-0.6111	0.1982	-3.084	0.005081	-1.02	-0.20
BVARIETY2	1.4750	0.2972	4.962	4.58e-05	0.86	2.09
BVARIETY3	0.8500	0.2972	2.860	0.008642	0.24	1.46
BSPACE2	0.3250	0.2972	1.093	0.285088	-0.29	0.94
BSPACE3	0.8250	0.2972	2.775	0.010511	0.21	1.44
BVARIETY2:BSPACE2	-0.7000	0.4204	-1.665	0.108877	-1.57	0.17
BVARIETY3:BSPACE2	0.5000	0.4204	1.189	0.245909	-0.37	1.37
BVARIETY2:BSPACE3	**-1.8250**	**0.4204**	**-4.341**	**0.000222**	**-2.69**	**-0.96**
BVARIETY3:BSPACE3	**0.8750**	**0.4204**	**2.081**	**0.048227**	**0.01**	**1.74**

Diagnostic Analysis

The assumptions of the linear model are that the residuals are independent, normally distributed, centred around 0 and have constant variance. From the residuals vs fitted plot (upper left of Figure 3), we note that the residuals appear randomly scattered around 0, with 3 observations having relatively large residuals (observations 9, 29 and 35). These observations are potential outliers. The scale location plot (bottom left of Figure 3) does show a slight positive trend, indicating a slight increase in variance.

The normal Q-Q plot is depicted in the upper right corner of Figure 3. The majority of our observations tend to track along the diagonal line, with only slight deviations at the tail. Observations 9, 29 and 35, which appeared as potential outliers in the residuals vs fitted plot, appear within two standardised residuals and are accordingly not regarded as outliers. The data appears to be normally distributed. Moreover, the Shapiro-Wilk test returns a p-value of 0.08 > 0.05, and hence there is no evidence to suggest the residuals are not normally distributed.

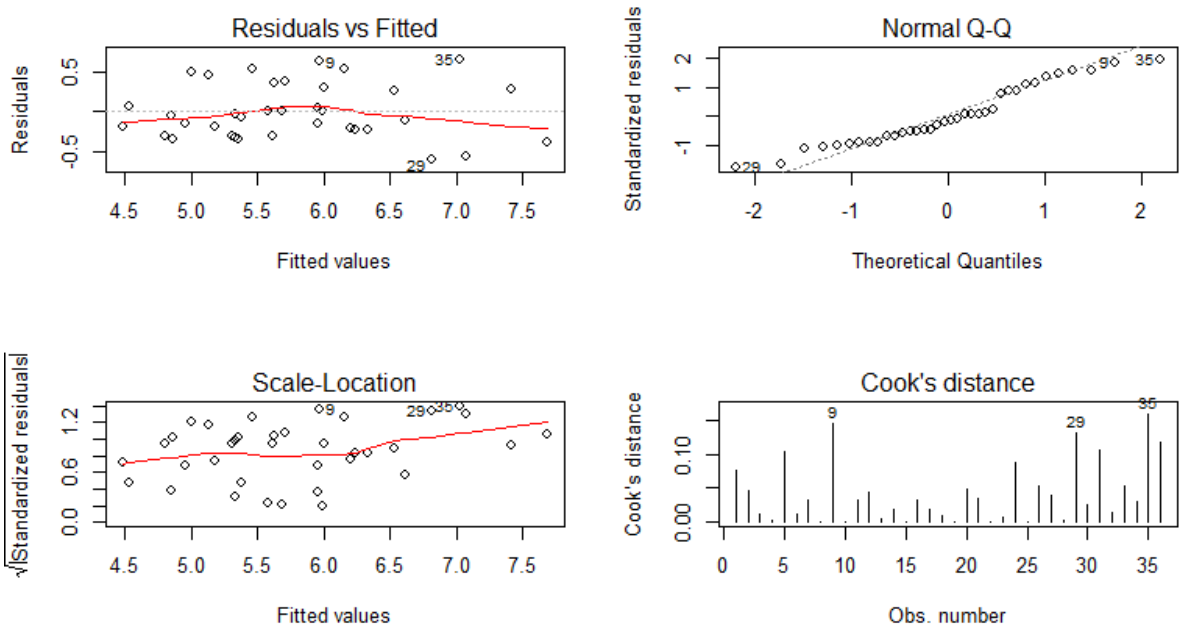

Figure 3: Diagnostic plots of the Final Stepwise Model

The Cook's distance plot is depicted in the lower left of Figure 3. Observation 35 appears to have the largest Cook's distance of approximately 0.15, which lies at approximately 0.07th percentile of the $F_{12,24}$ distribution. This observation is not influential. Note also that as observations 9 and 29 have Cook's distances corresponding to less than 0.15, they can also be regarded as not influential.

In summary, despite evidence of a slight increase in variance in the scale location plot, it appears the model assumptions are reasonable and free of outliers and influential observations. Our model appears appropriate.

Concluding Remarks

This paper has considered a potential model to evaluate barley yield of 3 different varieties at 3 different row spacings. The exploratory analysis suggested that an interaction effect existed between variety and row spacing, and accordingly, that an interaction model would best fit this data. An effects plot indicated that the highest yield would be achieved with variety 3 in row space level 3. The multiple regression analysis (incorporating all predictors and their interactions) using stepwise backward selection produced the following interaction model:

$$E(Yield) = 5.14 - 0.28Block2 - 0.66Block3 - 0.61Block4 + 1.47BVariety2 + 0.85BVariety3 + 0.33BSpace2 + 0.83BSpace3 - 0.7BVariety2{:}\,BSpace2 + 0.5BVariety2{:}\,BSpace3 - 1.83BVariety3{:}\,BSpace2 + 0.88BVariety3{:}\,BSpace3$$

The analysis of regression coefficients indicates that this model is useful (F-statistic = 11.34 on 11, 24 df, p-value ≈ 0) and explains approximately 76% of the observed variability in mean barley yield. The mean barley yield for variety 1 at row space level 1 was 4.75, and the mean yield was greatest for variety 3 at row space level 3, at 7.3. The coefficient of the interaction term variety 3 and space 3 is positive and significant, hence we can confirm that this difference is significant. In addition, given there is no apparent overlap in the confidence intervals of variety 3 at space levels 2 and 3, we can conclude that variety 3 at row space level 3 produces the highest yield of barley overall. Diagnostic analysis revealed that despite evidence of a slight increase in residual variance, it appears the assumptions of the linear model are reasonable and free of outliers and influential observations. Our model appears useful.