ST553 HW2

Nick Sun

April 14, 2019

Question 1

Our question here is **Does the type of fertilizer affect lettuce production?**

We have 20 plots which constitute our experimental units. The nitrogen treatments are:

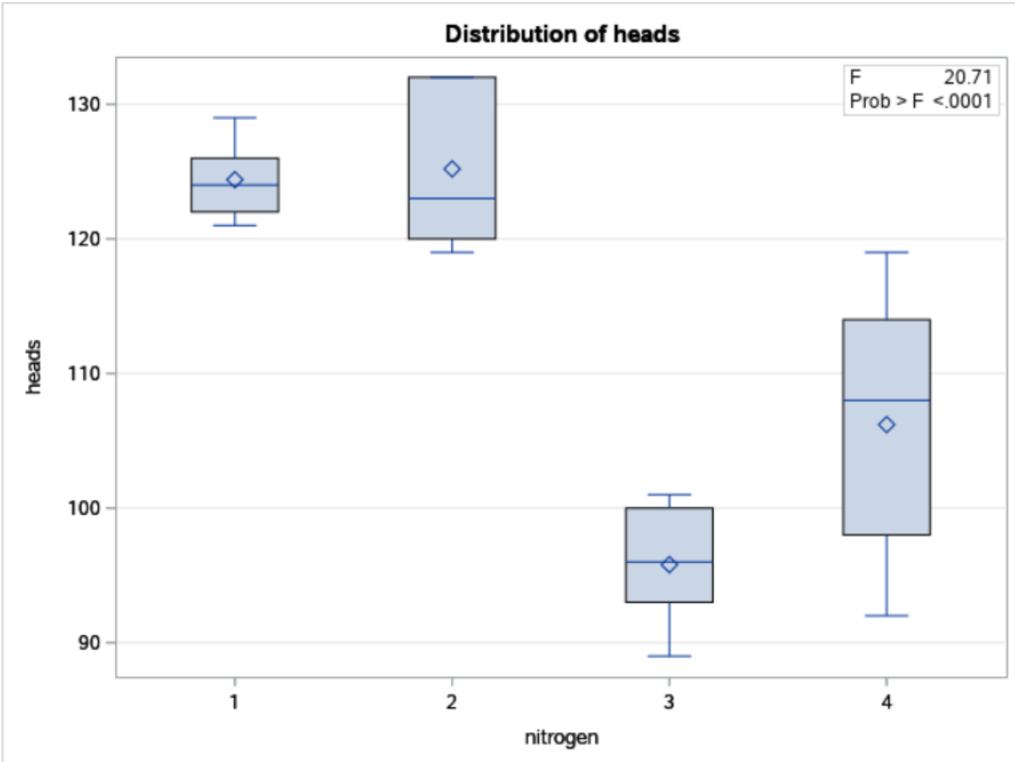
- (1) Blood meal
- (2) cottonseed meal
- (3) ammonium nitrate
- (4) urea

We have five replicates of each treatment. The recorded response is the number of heads of lettuce which grew in each plot.

We can read the data in and perform a preliminary analysis using `PROC GLM` which produces this ANOVA table and associated boxplot. We will also use contrasts to test different treatments against one another. To avoid data snooping, we will define our contrasts first. We will look at three contrasts:

- The organic treatments vs. the chemical treatments ($H_0 : \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4 = 0$)
- Blood meal vs cottonseed meal ($H_0 : \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + 0\mu_3 + 0\mu_4 = 0$)
- Ammonium nitrate vs urea ($H_0 : 0\mu_1 + 0\mu_2 - \frac{1}{2}\mu_3 + \frac{1}{2}\mu_4 = 0$)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3104.200000	1034.733333	20.71	<.0001
Error	16	799.600000	49.975000		
Corrected Total	19	3903.800000			



The low p-value in the F-test tells us that our full model where each treatment group has its own calculated estimate $\bar{y}_{i.}$ performs significantly better than the null model with only the grand mean as its estimate for all responses. The boxplots corroborate this; the nitrogen groups appear to produce different amounts of lettuce. In particular, treatments 1 and 2 seem very different from 3 and 4. So to answer our overall research questions, **yes, the fertilizers affect the lettuce yield**.

We can now explore the contrasts we predefined. SAS outputs include the point estimate and the confidence interval.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
1,2 vs 3,4	-47.6000000	6.32297398	-7.53	<.0001	-61.0041061	-34.1958939
1 vs 2	-0.8000000	4.47101778	-0.18	0.8602	-10.2781343	8.6781343
3 vs 4	10.4000000	4.47101778	2.33	0.0335	0.9218657	19.8781343

Assuming $\alpha = .05$, these contrasts tell us that Treatments 1 and 2 are significantly different from Treatments 3 and 4. The effect difference between Treatments 1 and 2 and Treatments 3 and 4 is that Treatments 1 and 2 will have around 23.8 more heads of lettuce. Within these subgroups, Treatments 1 and 2 are not significantly different from each other. Treatments 3 and 4 are significantly different from each other. The effect size difference is that urea produces on average 10.4 more heads of lettuce than ammonium nitrate.

For farmers interested in finding the nitrogen treatment which produces the most lettuce, our experiment suggests that treatments 1 or 2 have a similar effect which is significantly greater than either treatment 3 or 4.

Question 2

This question deals with an experiment involving treatments of "fines" in paper pulp. There are five different treatment levels with three replicates apiece:

- 0%
- 10%
- 20%
- 30%
- 40%

The response variable is the tensile index of the paper made from the pulp.

a.

If we define y_{ij} as the jth observation in the ith treatment group, the cell means parameterization uses the follwoing model:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i is the mean for the ith treatment group and ϵ_{ij} is an error term associated with the ijth observation. The necessary assumptions for this model is that the errors are independent and identically distributed $N(0, \sigma^2)$.

Our estimates $\hat{y_{ij}}$ will be the samples means for each treatment group.

b.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	92.5176390	23.1294097	4.68	0.0218
Error	10	49.4393831	4.9439383		
Corrected Total	14	141.9570221			

The error sum of squares for the cell means parameterization model in part (a) is 49.44 and the error sum of squares for the equal means model is 141.96

c.

Our polynomial regression parameterization takes the form:

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_{ij}$$

where y_{ij} is the jth observation in the ith treatment group, x_i is the fines percentage in the paper pulp treatment, and β_k is the regression parameter for the x_i^k variable.

The ANOVA table is:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	92.5176390	23.1294097	4.68	0.0218
Error	10	49.4393831	4.9439383		
Corrected Total	14	141.9570221			

d.

The Type I and Type III Sum of Squares are given in the following table:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
fines	1	25.86293751	25.86293751	5.23	0.0452
fines2	1	47.99390229	47.99390229	9.71	0.0110
fines3	1	10.23813252	10.23813252	2.07	0.1807
fines4	1	8.42266665	8.42266665	1.70	0.2210

Source	DF	Type III SS	Mean Square	F Value	Pr > F
fines	1	0.10536656	0.10536656	0.02	0.8868
fines2	1	3.38257656	3.38257656	0.68	0.4274
fines3	1	6.69934763	6.69934763	1.36	0.2714
fines4	1	8.42266665	8.42266665	1.70	0.2210

e.

Type I Sum of Squares are obtained sequentially. The full models include all predictors up to a particular polynomial degree and the reduced models include all of the same predictors except for the highest degree polynomial in the full model.

	Full Model	Reduced Model
Fines	Fines	Null model
Fines^2	Fines + Fines^2	Fines
Fines^3	Fines + Fines^2 + Fines^3	Fines + Fines^2
Fines^4	Fines + Fines^2 + Fines^3 + Fines^4	Fines + Fines^2 + Fines^3

Type III Sum of Squares are obtained partially. The full models include all predictors. The reduced models include all predictors except the predictor of interest.

	Full Model	Reduced Model
Fines	Fines + Fines^2 + Fines^3 + Fines^4	Fines^2 + Fines^3 + Fines^4
Fines^2	Fines + Fines^2 + Fines^3 + Fines^4	Fines + Fines^3 + Fines^4
Fines^3	Fines + Fines^2 + Fines^3 + Fines^4	Fines + Fines^2 + Fines^4
Fines^4	Fines + Fines^2 + Fines^3 + Fines^4	Fines + Fines^2 + Fines^3

f.

Using `R`'s `anova()` function, we get the following table.

Table 3: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
12	68.1	NA	NA	NA	NA
10	49.44	2	18.66	1.887	0.2017

The p-value of .2017 tells us that we fail to reject the null hypothesis that the quadratic fit is appropriate for this data.

Question 3

a.

From pg 69, the t statistic with N-g degrees of freedom is given as follows:

$$t = \frac{w\left[\bar{y_{i.}}\right] - \delta}{\sqrt{MS_E}\sqrt{\sum_{i=1}^{g}\frac{w_i^2}{n_i}}}$$

where w_i are the contrast coefficients, $\bar{y_{i.}}$ are the sample means and n_i are the sample sizes. δ is the contrast value under the null hypothesis which is usually set to 0. Oehlert uses $w[\bar{y_{i.}}]$ to refer to the sum $\sum_{i=1}^{g} w_i y_{i.}$.

b.

We know from ST561 that the square of the t-statistic with k degrees of freedom is equivalent to an F statistic with 1 and k degrees of freedom.

$$t^2 = \left(\frac{w\left[\bar{y_{i.}}\right] - \delta}{\sqrt{MS_E}\sqrt{\sum_{i=1}^{g}\frac{w_i^2}{n_i}}}\right)^2 = \frac{w[\bar{y_{i.}}] - \delta}{MS_E\sum_{i=1}^{g}\frac{w_i^2}{n_i}}$$

We know that $MS_E = \frac{SS_E}{N-g}$ and $SS_W = \frac{w[\bar{y_{i.}}]}{\sum_{i=1}^{g}\frac{w_i^2}{n_i}}$. Therefore, we can rewrite our t^2 quantity as $t^2 = \frac{SS_W/1}{SS_E/N-g}$ which is indeed an F distribution with 1 and N-g degrees of freedom.

c.

$$SS_W = t^2 MSE = F\frac{SS_E}{N-g}$$