INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO

January 27, 2020 **Final Exam 2019-20 (second date)** **A possible solution**

I

1. Given the total of $N=2501$ observations, but where the marginal (row/column) totals were *not* fixed in advance, the question may be answered using an independence test on this contingency table (two-dimensional table of count data), which has $a=3$ rows and $b=4$ columns. The Null Hypothesis is the hypothesis of independence, which assumes that the joint probability of an observation falling in any given table cell is the product of the marginal probabilities for the row and the column associated with the cell. In other words, $H_0 : \pi_{ij}=\pi_{i.}\times\pi_{.j}$, for all i and j. The Alternative Hypothesis H_1 is the negation of H_0: there exists at least one table cell for which $\pi_{ij}\neq\pi_{i.}\times\pi_{.j}$. Pearson's statistic is given by $X^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$. Its asymptotic distribution, if H_0 (independence) is true, is $\chi^2_{(a-1)(b-1)}$. We reject H_0 (at the $\alpha=0.05$ significance level) if $X^2_{calc}>\chi^2_{0.05(6)}=12.5916$.

2. The sample size is appropriate: we can use the asymptotic distribution. In fact, Cochran's criteria state that the asymptotic distribution for X^2 can be used if: (i) none of the estimated *expected* values $\hat{E}_{ij}$ is less than 1; and (ii) no more than 20% of the $\hat{E}_{ij}$ are less than 5. In order to check Cochran's criteria, we can choose the cell with the smallest expected value and see whether it is larger than 5 (**Note:** Cochran's criteria use the *expected* values $\hat{E}_{ij}$, and *not* the observed values O_{ij}). The cell with the smallest $\hat{E}_{ij}$ is the one in the row (Species) and column (Orientation) with the least observations. This is cell $(3,2)$, where $\hat{E}_{32}=\frac{N_{3.}\times N_{.2}}{N}=\frac{466\times366}{2501}=68.19512\gg5$. It is therefore safe to use the asymptotic distribution for Pearson's statistic.

3. The contribution of cell $(3,3)$ to the value of X^2_{calc} is $\frac{(O_{33}-\hat{E}_{33})^2}{\hat{E}_{33}}$. We have $O_{33}=243$ and $\hat{E}_{33}=\frac{N_{3.}\times N_{.3}}{N}=\frac{466\times484}{2501}=90.18153$. Therefore, the value of the term is 258.9608. This value is larger than the sum of the remaining 11 terms of the statistic (which is given in the question: 229.6256). Such a huge value is the result of a positive association: the observed number of individuals in this cell is much larger than would be expected under the independence hypothesis. The test statistic's value is $X^2_{calc}=488.5864$, and so we clearly reject the independence hypothesis (the sum of the 11 terms given in the question would already be sufficient to ensure rejection). This rejection is not unexpected: a visual inspection of the data table shows that the species *Zygophyllum simplex* clearly prefers South, unlike the other two species which prefer North.

II

1. This is a multiple linear regression with $n=109$ observations and $p=4$ predictors.

   (a) Since $R^2=0.7363$, the model explains 73.63% of the variance of the observed values of the response variable (`brix`). This is a reasnobaly good value.

   (b) What is being requested is a test on whether β_3 is *negative*. Without giving the benefit of the doubt to this hypothesis, we have $H_0 : \beta_3\geq0$ vs. $H_1 : \beta_3<0$. Since the borderline

value is $\beta_3 = 0$, the computed value of the test statistic is given in the question's output: $T_{calc} = -3.512$ [**Note:** the accompanying *p-value* is for a test with a two-sided (bilateral) critical region, and is therefore not useful here]. Given the nature of the hypotheses, the critical region for this test is one-sided (unilateral), and specifically it is the left-hand tail of the distribution. We reject H_0 if $T_{calc} = -3.512 < -t_{0.01(104)} = -2.362739$. Hence, we reject H_0 in favour of $H_1 : \beta_3 < 0$ and $b_3 = -0.61539$ may be considered significantly smaller than zero. The statement in the question is therefore legitimate.

(c) The plot has the values of the (internally) standardized residuals (R_i) on the vertical axis. In no case are their absolute values greater than 3 (although two are close). Thus, we cannot see any outlying observations. However, three observations have a large leverage (the values of which define the horizontal axis, measuring the degree to which each observation 'attracts' the fitted hyper-surface), bigger than 0.15, which is three times larger than the mean leverage $\overline{h} = \frac{p+1}{n} = 0.04587$. Among these observations, only one (observation 102) has a value of R_i far from zero. This means that its Cook's distance must be high (see on the formula sheet the expression for D_i). Its Cook's distance is close to the 0.5 threshold. Cook's distance is a measure of influence, that is, of the impact that excluding an observation will have on the fitted hyper-surface. It tends to be larger for points that are further away from the center of gravity of the scatterplot of n points in $\mathbb{R}^{p+1}$. Observation 102 is extreme in three of the predictor variables (it has the smallest yield and acidity, and the largest pH, among all $n = 109$ observations), and for the other two predictors it has values in one of the extreme quartiles (between the minimum value and the first quartile for grape weights and between the third quartile and the maximum value for the response variable `brix`). Observation 102 has, overall, a substantial an impact on the fitted model, and it should therefore be inspected with care.

2. The simple linear regression of `brix` (y) on `pH` (x).

(a) A partial F test is requested, to compare the full model from the previous question with the simple linear regression submodel (hence $k = 1$) of `brix` on `pH`. The Null Hypothesis of this test is that both models are the same, $H_0 : \mathcal{R}_c^2 = \mathcal{R}_s^2$. The Alternative Hypothesis is $H_1 : \mathcal{R}_c^2 > \mathcal{R}_s^2$. The test statistic may be written as $F = \frac{n-(p+1)}{p-k} \frac{R_c^2 - R_s^2}{1-R_c^2}$, whose distribution under H_0 is $F_{[p-k, n-(p+1)]}$. We reject H_0 if $F_{calc} > f_{0.05(3,104)} \approx 2.7$. To compute the value of the test statistic, it is necessary to know the submodel's coefficient of determination, R_s^2. Since the submodel is a simple linear regression, its coefficient of determination is the square of the linear correlation coefficient between the response and the predictor variables, which is given in the question. Thus, $R_s^2 = 0.8305^2 = 0.6897$. We have $F_{calc} = 6.1222$, so we reject H_0 at the $\alpha = 0.05$ significance level. The fitted submodel has a significantly worse fit than the full model.

(b) The formula sheet gives the expression for the leverage of an observation in a simple linear regression: $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}$. We know that $n = 109$; $x_{102} = 3.93$; $\overline{x} = 3.684495$; and $s_x^2 = 0.075136^2 = 0.005645418$. Hence, $h_{102,102} = 0.1080$, which is about half the corresponding value in the multiple linear regression model discussed above. However, the observation's Cook's distance is again close to the threshold 0.5. In fact, by the expression for D_i (see the formula sheet), $D_{102} = R_{102}^2 \cdot \frac{h_{102,102}}{1-h_{102,102}} \cdot \frac{1}{2} = 0.404$, which is relatively high.

3. The simple linear regression of `brix` (y) on `acidez` (x).

(a) This being a simple linear regression, the correlation coefficient between x and y is one of the square roots of the coefficient of determination. It must be the negative square

root, given the regression line's negative slope ($b_1 = -0.9263$), which indicates a decreasing relation. Thus, $r_{xy} = -\sqrt{R^2} = -\sqrt{0.1005} = -0.3170$.

(b) The goodness-of-fit test has as the Null Hypothesis $H_0 : \mathcal{R}^2 = 0$ (with $H_1 : \mathcal{R}^2 > 0$). The test statistic (for a simple linear regression) is $F = (n-2) \cdot \frac{R^2}{1-R^2}$, with distribution $F_{[1,n-2]}$ under H_0. The critical region is a one-sided right-hand region, with rejection of H_0 if $F_{calc} > f_{0.05(1,107)} \approx 3.94$. Now, $F_{calc} = 11.95497$, so we reject H_0, despite the very small value of R^2. This fact is not contradictory, because the goodness-of-fit test is only telling us that $R^2 = 0.1005$ is significantly different from zero, and not that the fitted model is necessarily good.

III

1. Since there is nothing that allows us to associate fields in different environments, this experimental design must be considered nested (hierarchical), with two factors: environment (dominant Factor A, with $a = 8$ levels) and fields (subordinate Factor B, where, within each environment there are $b_i = 9$ levels). This is a balanced design, with $n_c = 6$ repetitions for each of the $\sum_{i=1}^{a} b_i = 72$ experimental situations, giving a total of $n = 6 \times 72 = 432$ observations.

   Model equation: $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, where $i = 1, ..., 8$ indicates environment; $j = 1, ..., 9$ field (within environment); $k = 1, ..., 6$ repetition (for each experimental situation); Y_{ijk} indicates the yield in the k-th repetition in field j within environment i; ϵ_{ijk} is the corresponding random error. With the constraints $\alpha_1 = 0$ and $\beta_{1(i)} = 0$ for any i, μ_{11} represents the mean population yield for the first field in environment 1; α_i indicates the effect associated with environment i; and $\beta_{j(i)}$ indicates the effect of the j-th field within environment i.

   Distribution of the random errors: $\epsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$, for any i, j, k.

   Independent errors: $\{\epsilon_{ijk}\}_{i,j,k}$ are independent random errors.

2. There are two types of effects (of the factor environment and of the factor field). The summary table will therefore have three rows (one for each kind of effect and one row associated with residual variability). Two table values are given in the question: the Residual (Error) Mean Square, QMRE=2.2347 and the environment Sum of Squares, SQA=1666.2. The degrees of freedom are: $a - 1 = 7$ (Factor A); $\sum_{i=1}^{a}(b_i - 1) = 64$ (Factor B) and $n - \sum_{i=1}^{a} b_i = 432 - 72 = 360$ (Residual). Thus, we have $QMA = \frac{SQA}{a-1} = 238.0286$, hence $F^A_{calc} = \frac{QMA}{QMRE} = 106.5148$; $SQRE = \left(n - \sum_{i=1}^{a} b_i\right) \times QMRE = 804.492$. The Sum of Squares for the subordinate factor B results from the fact that $SQB(A) = SQT - (SQA + SQRE) = (n-1)\, s_y^2 - (1666.2 + 804.492) = 431 \times 6.05404 - 2470.692 = 2609.291 - 2470.692 = 138.5992$. Its Mean Square is $QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a}(b_i - 1)} = 2.165612$. Finally, the test statistic for the effects of the subordinate factor is $F^{B(A)}_{calc} = \frac{QMB(A)}{QMRE} = 0.969084$. Here is the full summary table:

Sources of Variation	df	Sums of Squares	Mean Squares	F_{calc}
Environment (Factor A)	7	1666.2	238.0286	106.5148
Field (Factor B(A))	64	138.5992	2.165612	0.969084
Residual	360	804.492	2.2347	—
Total	431	2609.291	–	–

3. There are two F tests of interest in this model, one for each factor's effects. In the test for environment effects, the hypotheses are $H_0 : \alpha_i = 0, \forall\ i$ and $H_1 : \exists i$, such that $\alpha_i \neq 0$. The test statistic is $F^A = \frac{QMA}{QMRE} \frown F_{\left[a-1\,,\,n-\sum_{i=1}^{a} b_i\right]}$, under H_0. The rejection rule at the $\alpha = 0.05$ significance level is to reject H_0 if $F_{calc} > f_{0.05(7,360)} \approx 2.02$. As $F^A_{calc} = 106.5148$, there is a very clear rejection of H_0, in other words, we clearly conclude that environment effects on yields exist. As for the test on field effects, the Null Hypothesis $H_0 : \beta_{j(i)} = 0$ for all fields (in all environments) is not rejected (H_1 was that there exist i, j such that $\beta_{j(i)} \neq 0$). The computed value of the statistic, $F^{B(A)} = 0.969084$, is less than 1, and therefore less than any tabulated value that could represent the borderline of a critical region (which for $\alpha = 0.05$, is $f_{0.05(64,360)} \approx 1.32$). Thus, we conclude that the variability of yields along the fields is not significant, once the variability along the environments that were studied is taken into account. The subordinate factor does not account for further significant variability.

4. Two population mean yields, in two different fields (from any environments) may be considered different (i.e., we may reject $\mu_{ij} = \mu_{i'j'}$ in favour of $\mu_{ij} \neq \mu_{i'j'}$) whenever we have the inequality $|\overline{y}_{ij.} - \overline{y}_{i'j'.}| > q_{\alpha(\sum_i b_i\,,\,n-\sum_i b_i)} \sqrt{\frac{QMRE}{n_c}}$. To compute the comparison term, we note that $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{2.2347}{6}} = 0.6102868$. Using the overall $\alpha = 0.05$ significance level, we have $q_{0.05(72,360)} = 5.939$ (value given in the question, since the parameter values for the Tukey distribution are very far away from those available in the tables). Thus, the significance threshold is $5.939 \times 0.6102868 = 3.624493$. The smallest sample mean yield for environment 2 is registered in field 1, and is $\overline{y}_{21.} = 4.873$. The largest mean yield is in field 6, and is $\overline{y}_{26.} = 8.617$. The difference between these two sample means is $8.617 - 4.873 = 3.744 > 3.624493$, and it is therefore a significant difference (although only just) for $\alpha = 0.05$. This conclusion seems contradictory with the result of the F test for field effects. Such a result is possible, since the theoretical results that underpin Tukey's tests and F tests are different. Besides, the difference that was now considered is only borderline significant (for $\alpha = 0.05$).

5. If nine types of fields had been previously defined, and in each environment fields of each type were selected, we would have a factorial experimental design, since each of the 8 environments would be combined with each of the nine types of fields. Since there are repetitions on each of the 72 resulting experimental situations, we can fit the two-way ANOVA model *with* interaction effects. This model's equation is $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, and it differs from the equation of the nested model in that the former terms $\beta_{j(i)}$ are now replaced by the sum of two terms: the field effects β_j (which correspond to the main effects of each of the $b = 9$ different types of fields, but with the constraint $\beta_1 = 0$, giving $b - 1 = 8$ such effects); and the interaction effects $(\alpha\beta)_{ij}$ which correspond to each experimental situation (with the constraints $(\alpha\beta)_{ij} = 0$ when $i = 1$ and/or $j = 1$, giving $(a-1)(b-1) = 56$ such effects).

IV

1. We have $y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$.

   (a) Thus, $1 - y = 1 - \frac{1}{1+\mathrm{e}^{-(c+d\,x)}} = \frac{\cancel{1}+\mathrm{e}^{-(c+d\,x)}-\cancel{1}}{1+\mathrm{e}^{-(c+d\,x)}} = \frac{\mathrm{e}^{-(c+d\,x)}}{1+\mathrm{e}^{-(c+d\,x)}}$. Dividing y by $1 - y$ gives:

$$\frac{y}{1-y} = \frac{\frac{1}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}}{\frac{\mathrm{e}^{-(c+d\,x)}}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}} = \frac{1}{\mathrm{e}^{-(c+d\,x)}} = \mathrm{e}^{c+d\,x}\ .$$

Taking logarithms, we get $\ln\left(\frac{y}{1-y}\right) = c + d\,x$, in other words, the *logit* of y is linearly related to the predictor x.

(b) The relative rate of change that is requested is the ratio $\frac{y'(x)}{y(x)}$. We must therefore calculate the derivative $y'(x)$. Now,

$$\begin{aligned} y'(x) &= [(1+\mathrm{e}^{-(c+dx)})^{-1}]' = (-1)[1+\mathrm{e}^{-(c+dx)}]^{-2}(1+\mathrm{e}^{-(c+dx)})' \\ &= (-1)[1+\mathrm{e}^{-(c+dx)}]^{-2}\mathrm{e}^{-(c+dx)}(-d) = \frac{d\,\mathrm{e}^{-(c+dx)}}{(1+\mathrm{e}^{-(c+dx)})^2}\ . \end{aligned}$$

Dividing by $y(x)$ gives the relative rate of change:

$$\frac{y'(x)}{y(x)} = \frac{\frac{d\,\mathrm{e}^{-(c+dx)}}{(1+\mathrm{e}^{-(c+dx)})^2}}{\frac{1}{1+\mathrm{e}^{-(c+dx)}}} = \frac{d\,\mathrm{e}^{-(c+dx)}}{1+\mathrm{e}^{-(c+dx)}} = d\,[1-y(x)]\ ,$$

taking into account the expression for $1-y(x)$ that was calculated above.

2. (a) The vector $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}}$ has a generic element $Y_i - \hat{Y}_i$, which is the residual for the i-th observation. In other words, $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} = \vec{\mathbf{E}}$ is the vector of residuals. The norm of any vector is the square root of the sum of squares of the vector's elements. Therefore, $\|(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}\|^2 = \|\vec{\mathbf{E}}\|^2 = \sum_{i=1}^{n} E_i^2 = SQRE$.

(b) If we multiply any matrix, on the right, by a vector, we get a linear combination of the columns of the matrix, whose coefficients are the vector's elements. Thus, the vector $\vec{\mathbf{1}}_n$, which is the first column of the matrix model $\mathbf{X}$, results from the product $\mathbf{Xv}$ with $\mathbf{v}^t = (1,0,0,...,0)$, i.e., the vector whose only non-zero element is a 1 in its first position. Thus, we have $\mathbf{H}\vec{\mathbf{1}}_n = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \cdot \mathbf{Xv} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X})}_{=\mathbf{I}}\mathbf{v} = \mathbf{Xv} = \vec{\mathbf{1}}_n$. (**Note:** In class and in the course notes, the fact that $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$ is shown in a different, but equally acceptable, way).

The product $\mathbf{H}\vec{\mathbf{1}}_n$ also defines a linear combination of the columns of matrix $\mathbf{H}$, with all coeffficients in this linear combination of the columns of $\mathbf{H}$ given by 1 (all elements of vector $\vec{\mathbf{1}}_n$ are 1). Hence, $\mathbf{H}\vec{\mathbf{1}}_n$ is the vector that results from adding all the columns in $\mathbf{H}$. In each position of the vector $\mathbf{H}\vec{\mathbf{1}}_n$ we have the sum of the elements in the corresponding row of $\mathbf{H}$. Since $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, all such sums are equal to 1.

(c) The mean of the observations in $\vec{\mathbf{Y}}$ may be calculated as $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\,\vec{\mathbf{1}}_n^t\vec{\mathbf{Y}}$, because the inner product of the vector $\vec{\mathbf{1}}_n$ with any other vector has the effect of adding up that vector's elements. In the same way, the mean of the fitted values $(\hat{Y}_i)$ results from considering $\overline{\hat{Y}} = \frac{1}{n}\,\vec{\mathbf{1}}_n^t\vec{\hat{\mathbf{Y}}} = \frac{1}{n}\,\vec{\mathbf{1}}_n^t\mathbf{H}\vec{\mathbf{Y}}$. But $\vec{\mathbf{1}}_n^t\mathbf{H} = (\mathbf{H}\vec{\mathbf{1}}_n)^t$, because $(\mathbf{H}\vec{\mathbf{1}}_n)^t = \vec{\mathbf{1}}_n^t\mathbf{H}^t$ and the matrix of orthogonal projections $\mathbf{H}$ is a symmetric matrix. Hence, $\overline{\hat{Y}} = \frac{1}{n}\,(\mathbf{H}\vec{\mathbf{1}}_n)^t\vec{\mathbf{Y}} = \frac{1}{n}\,(\vec{\mathbf{1}}_n)^t\vec{\mathbf{Y}} = \overline{Y}$.

(d) We have $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. Therefore, each fitted value $\hat{Y}_j$ is given by the corresponding element in the product $\mathbf{H}\vec{\mathbf{Y}}$. This is given by the inner product of row j of $\mathbf{H}$ with the vector of observations $\vec{\mathbf{Y}}$, that is, $\hat{Y}_j = \sum_{i=1}^{n} h_{ji}Y_i$. We saw in (b) that the sum of h_{ji} in any row j is 1, therefore $\sum_{i=1}^{n} h_{ji} = 1$. So $\hat{Y}_j$ is a weighted mean of all the observations Y_i, with weights given by the coefficients h_{ji}. The contribution of the observation Y_j towards its corresponding fitted valur $\hat{Y}_j$ has the weight h_{jj}, which is the leverage of observation Y_j.