

As always, not every possible way of expressing an appropriate answer is included in the following solutions. A point value for each part of each question is shown in parentheses. You should have a score for each part of each question written on your paper, even if it is zero. Your score for the final exam is written at the bottom of the last page of your paper. Check it to make sure that your total score was properly recorded. Bring any irregularities to the attention of the instructor.

- 1.(a) (6 points) Assume that $\beta_3 > 0$. Then, β_0 is the lower limit of the expected (or mean) 15 day weight gain. but it is approached as level of phenylalanine tend to minus infinity which is unrealistic. $\beta_0 + \beta_1$ is the upper limit on the expected 15 day weight gain and it is approached as the phenylalanine concentration becomes very large (which is also unrealistic). Nevertheless, β_1 is the difference between the upper and lower limits of the curve. When, the percentage of phenylalanine is $X = \beta_2/\beta_3$, the curve is at its inflection point where the rate of increase in weight gain is the largest. At this level of phenylalanine the expected 15 day weight gain is $\beta_0 + 0.5\beta_1$. So, you can interpret $0.5\beta_1$ as the difference between the upper limit to the mean response and the mean response at the inflection point of the curve. This also provides a way to interpret β_2 . You could also consider the expected 15 day weight gain (it is actually a weight loss) for a diet completely depleted of phenylalanine ($X = 0$) and the expected 15 day weight gain for a diet consisting of 100% phenylalanine (which may be unattainable).

$$(b) (6 points) b_1 \pm t_{(27-4),.025} S_{b_1} \quad \Rightarrow \quad 57.82 \pm t_{(23),.025} \sqrt{18.92}$$

- (c) (8 points) Set the formula for the mean 15 day weight gain equal to zero and solve for the level of phenylalanine.

$$0 = \beta_0 + \frac{\beta_1}{1 + \exp(\beta_2 - \beta_3 X_0)} \quad \Rightarrow \quad X_0 = \frac{1}{\beta_3} (\beta_2 - \log(-1 - \beta_1/\beta_0)).$$

Evaluate this formula at the $\mathbf{b} = (-6.42, 57.82, 3.57, 4.79)^T$ to obtain the estimate

$$\hat{X}_0 = h(\mathbf{b}) = \frac{1}{4.79} (3.57 - \log(-1 + 57.82/6.42)) = 0.311.$$

Then, a large sample approximation to the standard error of the estimated mean for 15 weight gain is the square root of $\hat{G} * V * \hat{G}^T$, where V is the estimated covariance matrix for the parameter estimates shown at the top of page 2 of the exam and

$$\begin{aligned} G &= \begin{bmatrix} \frac{\partial h(\beta)}{\partial \beta_0} & \frac{\partial h(\beta)}{\partial \beta_1} & \frac{\partial h(\beta)}{\partial \beta_2} & \frac{\partial h(\beta)}{\partial \beta_3} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta_1}{\beta_3 \beta_0 (\beta_1 + \beta_0)} & \frac{1}{\beta_3 (\beta_1 + \beta_0)} & \frac{1}{\beta_3} & \frac{-\beta_2 + \log(-1 - (\beta_1/\beta_0))}{\beta_3^2} \end{bmatrix}. \end{aligned}$$

Evaluating G at $\mathbf{b} = (-6.42, 57.82, 3.57, 4.79)^T$, we have $\hat{G} = [-0.03658 \quad -0.01776 \quad -0.01455 \quad -0.00406]$ and $S_{X_0} = \sqrt{\hat{G} * V * \hat{G}^T} = 0.04848$. Then, an approximate 95 percent confidence interval for the level of phenylalanine that produces no growth is

$$X_0 \pm t_{(23),.05} \sqrt{\hat{G} * V * \hat{G}^T} \quad \Rightarrow \quad 0.3110 \pm (2.069)(0.04848) \quad \Rightarrow \quad (0.21, 0.41).$$

You were not expected to obtain the final numerical solution during the exam. Many students demonstrated that they could not properly derive the formulas for the required partial derivatives. Learning to use the `deriv3()` function in S-PLUS should be a beneficial aid to those students in the future. Furthermore, many students used $\log(-1 - a) = -\log(1 + a)$ in their derivation of the partial derivatives, which is not correct. This mistake was not marked on every paper.

- (d) (8 points) There were a variety of answers for this part of the exam. Many students did not receive full credit because they did not clearly state how the resampling would be done or they did not clearly describe the statistic they would evaluate for each bootstrap sample. The expected solution is as follows:

- step 1 To replicate this experiment we must have three observations at each of the nine levels of phenylalanine used in this study for each bootstrapped sample. This can be done by fitting the proposed model to the original data and computing the 27 residuals. Obtain a set of approximations to the random errors, $e_1^*, e_2^*, \dots, e_{27}^*$, by centering the residuals at zero and rescaling the i -th centered residual by dividing by the square root of the i -th diagonal element of $I - D(D^T D)^{-1} D^T$, where D is the matrix of first partial derivatives of $f(\mathbf{b}) = \beta_0 + \frac{\beta_2}{1 + \exp(\beta_2 - \beta_3 X)}$ evaluated at $\mathbf{b} = (-6.42, 57.82, 3.57, 4.79)^T$ and the values of X used in this experiment.
- step 2 Using simple random sampling with replacement, take a sample of 27 values from $e_1^*, e_2^*, \dots, e_{27}^*$. Add the values from this sample to the estimated means obtaining from fitting the model to the original sample to obtain a bootstrap sample of 27 observations.
- step 3 Fit the model to the bootstrap sample and to obtain a new set of parameter estimates, and use the new set of parameter estimates to obtain an estimate of the mean 15 day weight gain with a 1.0 percent level of phenylalanine.
- step 4 Repeat steps 2 and 3 to obtain a large number (say $B=5000$) bootstrapped estimates of the mean 15 day weight gain with a 1.0 percent level of phenylalanine.
- step 5 Construct an approximate $(1 - \alpha) \times 100\%$ confidence interval by by evaluating the $\alpha/2$ and $1 - \alpha/2$ percentiles of the bootstrapped estimates in step 4.

You could have proposed using simple random sampling with replacement to select a new set of three observed weight gains at each of the 9 levels of phenylalanine used in this study. This preserves the model matrix from one bootstrap sample to the next. The suggestion to apply simple random sampling to the original data set to select a new set of 27 observations received partial credit because it does not preserve the model matrix (design) across possible bootstrap samples. Some students decided to use only the three observations from the original sample taken at $X = 1.0$. They computed $B = 5000$ bootstrap sample of size 3. This solution received partial credit. It is a consistent procedure as the number of observations at $X = 1.0$ becomes large in the original sample, and it is not biased by incorrect specification of the form of the model that relates growth to the level of phenylalanine (if someone had discussed this I may have awarded full credit), but it ignores potential information provided by the other observations in the original sample. With only three observations at $X = 1.0$ and evidence in the residual plots that the proposed model is reasonable, it is advantageous (reducing variability in the estimation of mean 15 day growth at $X = 1.0$ and the length of the confidence interval) to take advantage of the information in the other 24 observations in the original data set.

- (e) (6 points) The plot of the residuals versus the fitted values and the plot of the residuals versus the percentage of phenylalanine do not reveal any trends, so there is not obvious defect (or bias) in the model for the mean 15 day weight gain. These plots also suggest that the error variance is nearly homogeneous across the levels of phenylalanine used in this experiment. The normal probability plot in the lower right corner of the page indicates that the random errors are approximately normally distributed.
- (f) (8 points) Consider a lack-of-fit test. The residual plots indicate the the random errors are approximately normally distributed with nearly homogeneous variances. Consequently, the following statistic should have an approximate central F-distribution if the proposed model is correct. Fit two models to the data:

$$\text{Model A: } Y_{ij} = \beta_0 + \frac{\beta_2}{1 + \exp(\beta_2 - \beta_3 X_i)} + \epsilon_{ij}$$

$$\text{Model B; } Y_{ij} = \mu_i + \epsilon_{ij}$$

Model B fits a different mean weight gain at each of the nine levels of phenylalanine used in this study. It is the most general model that can be fit to these data. Compute the sum of squared residuals (RSS) for each model and conclude that model B is a significant improvement over model A if

$$F = \frac{(RSS_{\text{model A}} - RSS_{\text{model B}})/5}{RSS_{\text{model B}}/18} > F_{(5,18), .05}.$$

If you assume $\epsilon_{ij} \sim NID(0, \sigma^2)$, you can use maximum likelihood estimation to fit the models to the data and obtain a test from the large sample chi-square approximation to the natural logarithm of the ration of likelihoods multiplied by -2. The F-test generally provides better control over the Type I error level in small samples. It is not appropriate to REML log-likelihoods in this case. Why?

2.(A) (9 points) $df_{fields} = 9$, $df_{plotswithinfields} = 18$, $df_{error} = 81$. Formulas for estimates of variance components are:

$$\begin{aligned}\hat{\sigma}_e^2 &= MS_{error} \\ \hat{\sigma}_p^2 &= \frac{MS_{plotswithinfields} - MS_{error}}{4} \\ \hat{\sigma}_f^2 &= \frac{MS_{fields} - MS_{plotswithinfields}}{12}\end{aligned}$$

(B) (6 points) Reject the null hypothesis of no interaction between the levels of the irrigation factor and time if

$$F = \frac{MS_{interaction}}{MS_{error}} > F_{(6,81), .05}.$$

(C) (12 points) $Var(\bar{Y}_{\bullet 24} - \bar{Y}_{\bullet 34}) = \frac{\sigma_e^2 + \sigma_p^2}{5}$. An unbiased estimate of this variance is

$$S_{\bar{Y}_{\bullet 24} - \bar{Y}_{\bullet 34}}^2 = \frac{MS_{plotswithinfields} + 3MS_{error}}{20}.$$

Then, an approximate 95 percent confidence interval is

$$(\bar{Y}_{\bullet 24} - \bar{Y}_{\bullet 34}) \pm t_{(v), .025} \sqrt{\frac{MS_{plotswithinfields} + 3MS_{error}}{20}}$$

where

$$v = \frac{(MS_{plotswithinfields} + 3MS_{error})^2}{\frac{(MS_{plotswithinfields})^2}{18} + 9\frac{(MS_{error})^2}{81}}$$

are the Satterthwaite degrees of freedom.

(D) (10 points, 2 for each statement) Only statement (ii) is false. With respect to item (iv), REML estimates of variance components are the same as the method of moment estimators obtained from the expectations of the mean squares in the ANOVA table in balanced experiments like this study when they are positive. Hence, item (iv) was scored as true.

(E) (6 points) This is a testable hypothesis that can be written as

$$\begin{aligned}0 &= E(Y_{i31}) - E(Y_{i34}) = \tau_1 + \gamma_{31} - \tau_4 - \gamma_{34} \\ 0 &= E(Y_{i32}) - E(Y_{i34}) = \tau_2 + \gamma_{32} - \tau_4 - \gamma_{34} \\ 0 &= E(Y_{i33}) - E(Y_{i34}) = \tau_3 + \gamma_{33} - \tau_4 - \gamma_{34}\end{aligned}$$

Write the model as $\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon}$, where \mathbf{u} is a vector of random field and plot effects, Z is a matrix of ones and zeros that identifies the plot and field in which each observation was made, X is a model matrix for the fixed effects, and

$$\boldsymbol{\beta} = (\mu \ \alpha_1 \ \alpha_2 \ \alpha_3 \ \tau_1 \ \tau_2 \ \tau_3 \ \tau_4 \ \gamma_{11} \ \gamma_{12} \ \gamma_{13} \ \gamma_{14} \ \gamma_{21} \ \gamma_{22} \ \gamma_{23} \ \gamma_{24} \ \gamma_{31} \ \gamma_{32} \ \gamma_{33} \ \gamma_{34})^T$$

Then, the null hypothesis can be expressed as $\mathbf{0} = C\boldsymbol{\beta}$ where

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

Let \mathbf{b} be any solution to the normal equations. Then, an F-test is performed by rejecting the null hypothesis if

$$F = \frac{\mathbf{b}^T C^T (C^T (X^T X)^{-1} C)^{-1} C \mathbf{b}}{MS_{error}} > F_{(3,81)} < .05.$$

This formula does not use $Var(\mathbf{Y}) = ZGZ^T + \sigma_e^2 I$, where $G = Var(\mathbf{u})$, because of the balanced experiment that was used and the model that was proposed, although this may not be obvious at first glance. This question should not have appeared on this exam. Some students gave an alternative answer using using a set of three t-tests for the three contrasts that are set to zero under the null hypothesis.

- 3.(a) (6 points) The models in problems 2 and 3 have the same fixed effects parts, so they impose the same model on the expectations of the 15 day weight gains. The difference between the two models is in the covariance structures. Both models contain random field effects. In both models observations taken from different fields are independent, and observations taken from the same field are correlated. The main difference is in the covariance structure for the four observations taken at different time points from a single plot. For the model in problem 2, every observation has variance $\sigma_e^2 + \sigma_p^2 + \sigma_f^2$ and the correlation between any two observations from the same plot is $(\sigma_p^2 + \sigma_f^2)/(\sigma_e^2 + \sigma_p^2 + \sigma_f^2)$. The correlation between any two observations taken from different plot in the same field is $\sigma_f^2/(\sigma_e^2 + \sigma_p^2 + \sigma_f^2)$. The model in problem 3 allows the variances of the observations to differ across time points. It also allows correlations between observations taken in the same plot to decrease as time points become farther apart. You can work out the details.
- (b) (8 points) Since the model from problem 2 is not a special case of the model from problem 3, you cannot necessarily use the large sample chi-square approximation to the distribution of $-2\log(\text{ratio of likelihoods})$ to compare the two models. You could use the large sample chi-square approximation to the distribution of $-2\log(\text{ratio of likelihoods})$ to compare each model to a slightly more complex model that has an arbitrary covariance matrix for the four observations taken on a single plot. You could also consider AIC or BIC values for full likelihoods or REML likelihoods.

Exam scores out of a possible 100 points are displayed below as a stem-leaf display.

```

8 | 5 6
8 | 0 1 4
7 | 5 7 7 7 7 8 9
7 | 0 0
6 | 6 6 6 6 6 7 7 8 8 8 9 8 9
6 | 1 1 1 1 1 2 3 3 3 3 4 4 4 4 4
5 | 5 6 7 7 8 8 8 8 9 9 9
5 | 0 1 1 1 2 2 4 4
4 | 7 8 8 8
4 | 0 2 2
3 | 8 9
3 |

```